Multiview Learning with Sparse and Unannotated data

Tanmoy Mukherjee

June 2019

Submitted in partial fulfilment of the requirements of the Degree of Doctor of Philosophy

Abstract

Obtaining annotated training data for supervised learning, is a bottleneck in many contemporary machine learning applications. The increasing prevalence of multi-modal and multi-view data creates both new opportunities for circumventing this issue, and new application challenges. In this thesis we explore several approaches to alleviating annotation issues in multi-view scenarios.

We start by studying the problem of zero-shot learning (ZSL) for image recognition, where class-level annotations for image recognition are eliminated by transferring information from text modality instead. We next look at cross-modal matching, where paired instances across views provide the supervised label information for learning. We develop methodology for unsupervised and semi-supervised learning of pairing, thus eliminating the need for annotation requirements.

We first apply these ideas to unsupervised multi-view matching in the context of bilingual dictionary induction (BLI), where instances are words in two languages and finding a correspondence between the words produces a cross-lingual word translation model. We then return to vision and language and look at learning unsupervised pairing between images and text. We will see that this can be seen as a limiting case of ZSL where text-image pairing annotation requirements are completely eliminated.

Overall these contributions in multi-view learning provide a suite of methods for reducing annotation requirements: both in conventional classification and cross-view matching settings.

Contents

Acknowledgement

I would like to express my sincere appreciation to my supervisor, Timothy Hospedales. Tim is an ideal teacher, mentor and friend who has guided me and who I still look upto when asking difficult questions. I would like to take this opportunity to thank Tao Xiang and Matthew Purver who took up being my primary supervisor when Tim moved to Edinburgh. I have been a part of amazing groups in Edinburgh and London and I would like to thank each and every member who has taught me something. I would like to thank my family for being supportive. And last but not the least, Fede this thesis is for you.

Chapter 1

Introduction

1.1 Background

Recent decades have seen increasing amounts of data being collected across industrial, scientific and social applications —and a corresponding drive to develop innovative data analysis methods. Data in this digital age is continuously evolving and comes through multiple channels or is collected from diverse domains, for example, images are typically associated with description and tags, videos contain audio and visual signals, a given web page has the textual content of the page and the anchor text linking to other web pages. The multi-modality of this digital data puts a strain on traditional learning algorithms due to their inability to exploit the different views they arrive in. While each of the input modalities exhibits different properties or lies in different heterogenous spaces, the information content in multiple modalities maybe associated with each other. For example, a wikipedia article often can be represented in text vector space but also contains hyperlinks to be modeled in graph space.

Many popular machine learning tasks ranging from classification to regression can benefit if multiple views of the data can be integrated. Furthermore, there is an increasing realisation that important societal applications ranging from healthcare, multimedia, visual recognition etc. can immensely benefit from comparing data which exists in multiple views [Ding et al., 2019]. Multi-view learning aims to model all the available views present and improve the learning performance.

Most standard supervised learning algorithms require annotated data which can prove to be a bottleneck in building scalable systems. Multi-view data or multi-modal data can circumvent this issue by taking advantage of the other modality to replace conventional manual annotation. Zero shot learning (ZSL) promises to reduce the annotation burden in visual recognition by borrowing from text representation which are usually available in abundance. In this thesis, we explore various ways to reduce annotation cost with zero-shot learning. Existing ZSL methods contribute by proposing new vector embeddings for text/image or new cross-modal mapping methods. Differently, in this thesis we contribute to ZSL by studying distributions rather than conventional vector embeddings of images and text. For this task we develop a new cross-modal matching objective function and the results show improved performance vs vector embeddings. We also show how distribution embeddings can model intra-class variability and how this feature enables meaningful conjunction-based image query.

A common assumption held in multi-view learning algorithms is that in the training data the views are *paired*, which means for every example in one view, the corresponding example in the other view should be known. However this assumption is often violated in real world situations. We provide some relevant examples. Standard neural machine translation [Artetxe et al., 2018, Lample et al., 2017] tasks require the presence of large parallel corpora which are difficult to build and might be non-existent for low-resource languages. Image captioning models [Karpathy and Fei-Fei, 2017] require the presence of corresponding captions with their images but collecting such labeled corpus might be unavailable. The question we would like to answer is it possible to learn meaningful representation in data scarce or unpaired settings ?

In unsupervised cross-modal matching, existing methods are based on kernelized sorting (KS) [Quadrianto et al., 2009] or the recently proposed CycleGAN architecture [Zhu et al., 2017]. In this thesis, we adopt the kernelized sorting line of work through statistical dependency measures and extend them with end-to-end deep learning. We show that this end-to-end learning outperforms classic shallow KS methods, while being easier to use that recent GAN methods. We first look at bilingual dictionary induction, where instances are words in two languages and learning their pairing produces a cross-lingual word translation model. We finally return to vision and language and look at learning unsupervised pairing between images and text. We will see that this can be seen as a limiting case of zero-shot learning where text-image pairing annotation requirements are completely eliminated.

1.1.1 Thesis Goals and Layout

In this thesis, we explore the following research questions:

Q1: How well can text description of categories be used to eliminate labelling requirements for supervised learning of recognition ? Specifically: can we define a probabilistic embedding of images and text over conventional vector space embeddings to reduce annotation requirements with text?

Q2: Can we learn to pair or associate elements in sets of vectors defined in heterogenous views which might be from same or different modalities ? Particularly can we do so without resorting to unstable adverserial learning?

Q3: To what extent can such unsupervised pairing algorithms be used to perform unsupervised learning of cross-lingual word translation and image-text matching?

The remainder of this thesis consists of seven chapters

Chapter 2 We present a background on multi-modal learning and discuss various technical challenges associated with it.

Chapter 3 We study zero-shot learning through text → image transfer via word-vector. We present the first distribution-embedding approach to this task and explore its benefits compared to standard vector-embedding approaches. This chapter corresponds to work published in (Mukherjee et al, EMNLP 16)

Chapter 4 We introduce the problem of unsupervised matching across heterogenous views. We introduce Canonical Correlation Analysis (CCA) [Hotelling, 1936a] and its related unsupervsied matching models.

Chapter 5 We study the problem of bilingual dictionary induction and provide one of the first purely unsupervised induction methods using Deep Squared Mutual Information (SMI) as a metric for pairing. Compared to other GAN-based approaches ours is much more stable to train. This chapter corresponds to work published in (Mukherjee et al, EMNLP 18)

Chapter 6 We next apply the same Deep SMI approach for image-text pairing and study unsupervised captioning and unsupervised classifier learning applications. The latter can be seen as an extreme form of ZSL where even the source class annotation requirements are removed.

Chapter 7 We finally conclude by summarising and discussing our contribution as well as potential future work.

Chapter 2

Background

In this chapter we present a background of multi-modal learning. We look at various popular methods for multi-modal data analysis. We finally discuss some background on Zero-shot learning.

2.1 Multi-modal learning

Information in real world is inherently multimodal in nature- we see objects, hear sound, smell odours and so on. The common notion of modality can be affiliated with a unified bundle of sensation from multiple sensory modalities. A research problem is hence characterized as multimodal when multiple sensory modalities like vision ,sound, touch are involved [Baltrušaitis et al., 2017]. A logical representation of objects combining various modalities allows for a meaningful perceptual experience.

To make truly intelligent machines, artificial intelligence needs to narrow the heterogeneity gap among the various multimodal signals being generated. End-to end speech recognition [Graves and Jaitly, 2014, Oord et al., 2016], neural machine translation [Bahdanau et al., 2014, Vaswani et al., 2017], image captioning are some of the examples where multimodal data is extensively used.

Multimodal data analysis brings in some unique challenges and some opportunities given the heterogenous nature of data. The underlying motivation to use multimodal data is that complementary information could be extracted from each of the modalities considered, giving a unique and comprehensive view and is generally more informative than unimodal data. For example, early research in speech recognition showed that visual modality provides valuable information on lip motion and articulation of the mouth, thus helping to improve speech recognition [Guo et al., 2019]. Learning from multimodal data sources offers the possibility to learn from multiple corresponding sources and offers a deep understanding to the natural phenomenon. We list and review some technical challenges associated with multimodal data. Our list consists of the following challenges:-

Information Fusion methods combine information from two or multiple separate modalities in making a single decision or prediction. The foundation of information fusion was laid in the beginning of the 20th century [Hotelling, 1936a, Cattell, 1944]. Further research in the early 1970s came with the formulation of multiset canonical correlation analysis [Kettering, 1971], parallel factor analysis (PARAFAC) [Harshman] and other tensor decomposition tools [Tucker, 1966]. However most of these techniques have remained confined in the field of chemometrics and psycometrics, the communities where they first evolved. The late 20th century sees a lot more technological advances with growing availibility of data sources and domains, leading to interest in exploiting the resources efficiently. These resources are multiview, multirelational, multimodal in nature and span the areas of social, health, electronic, manufacturing and thus the drive to develop tools for analytical understanding is high and relevant outside of academia. Information fusion remains one of the most popular tools due to its relevance in providing a unified picture and global view; improving decision making process, exploratory research, identifying common versus distinctive elements across the modalities and in general providing knowledge which could be utilized for various processes. Despite the popularity and the massive amount of research conducted [Khaleghi et al., 2013, Shivappa et al., 2010, Turk, 2014, Biessmann et al., 2011, Stathaki, 2008, Mitchell, 2012], the process of collectively learning from multiple sources is still at its earlier stages. Data fusion is a challenging task and raises several questions, conceptual and technical.

Earlier work in information fusion [Atrey et al., 2010, Khaleghi et al., 2013] has spanned different research communities and the matter has been throughly investigated. Depending on the stage of fusion, data fusion can be roughly categorized into *early fusion* or *late fusion*. Early fusion focuses on the best way to combine input features from multiple data sources either by removing correlations between the modalities and representing the fused data in a lower dimensional subspace. Earlier techniques that concentrated on these objectives include principal compoenent analysis (PCA), independent component analysis (ICA) and canonical correlation analysis (CCA). The training pipeline usually is usually simple as it requires a single model but well engineered features from the modalities so that they align or their semantics can be represented well.

Late fusion focusses on using the decisions seperately made by each of the machine learning models by using ensemble models. Late fusion allows the use of different models on different modalities, thus allowing freedom and flexibility in handling missing modalities. We provide some relevant examples of information fusion in multimodal settings. In computer vision, RGB-D (RGB-depth) along with multi-view images is used to generate effective features. In [Eitel et al., 2015], the feature vectors obtained from fully connected (FC) layer of two seperate CNNs are combined to generate joint features for RGB-D. In [Gupta et al., 2014], the performance of RGB-D fusion improved the effective encoding scheme for depth image. In [Li et al., 2017], multi-level fusion was proposed to learn multimodal features for semantic segmentation. Other areas where multimodal fusion have been succesfully applied is multimodal scene understanding [Hospedales and Vijayakumar, 2008], understanding brain functionality [Nunez and

Silberstein, 2000, Horwitz and Poeppel, 2002, Biessmann et al., 2011] using EEG and fMRI data, environmental studies [Stathaki, 2008, Yokoya et al., 2011, Vivone et al., 2015]. We refer the reader to [Lahat et al., 2015] for further reading.

Alignment Humans have a remarkable ability to spot analogies, or translate (mapping) information from one modality to another. This ability has been shown to be a fundamental ingredient of human intelligence and creativity [Gentner, 1983, Gentner and Forbus, 2011, Hummel and Holyoak, 1997, Lovett et al., 2009]. Alignment involves using one modality, termed as *source* or *base* to better understand the second modality known as *target.* The task of flexibly mapping between domains remains a challenge for machines. Classical or symbolic AI systems are ill-equipped and lack the flexibility to extend relations from source to target domains especially across domains previously unknown. With the availability of multimodal datasets, alignment has been particularly studied by the vision, natural language processing and speech community. Some examples include image captioning where one might want to find a correspondence between image regions and captions [Karpathy and Fei-Fei, 2017], aligning movies to script [Zhu et al., 2015], alignment of movie script to videos [Bojanowski et al., 2013, Alayrac et al., 2016], style transfer [Hoshen and Wolf, 2018a], unsupervised learning of word translations [Conneau et al., 2017b] and cross-modal alignment of speech and text [Chung et al., 2018].

Alignment can be broadly categorized into *unsupervised* and *supervised* algorithms. Unsupervised algorithms operates with no label correspondence between the two domains while supervised methods have access to them. We briefly review the two categories.

Unsupervised multimodal alignment arises when no direct correspondence between the two modalities exist. Consider the example of bilingual lexicon induction for machine translation systems where one needs to recover an alignment between two sentences. Some of the earliest works on unsupervised alignment were motivated by applications in measuring similarities between biological sequences or alignment for statistical machine translation systems. To aid the task, certain constraints are put on the alignment objective such as temporal ordering or existence of similarity metric [Baltrušaitis et al., 2017].

Dynamic time warping (DTW) [Kruskal, 1983, **?**] is one of the algorithms for measuring similarity between sequences. DTW measures the similarity between two time sequences and calculates an optimal score with certain restrictions in place. DTW has been extended to multimodal alignment by handcrafting similarity metrics between the modalities for example, in [Miró et al., 2014] defined a hand crafted similarity metric between graphemes and phonemes, [Tapaswi et al., 2014] define a similarity metric between visual scenes and sentences to align TV shows to plots.

Both techniques for unsupervised alignment, DTW and graphical models, place certain restrictions on the alignment e.g temporal consistency, no large jumps. While DTW based alignments allow the latent similarity metric and alignment to be jointly learnt, graphical models based approaches require expert knowledge for such construction [Baltrušaitis et al., 2017].

Alignment of embedding spaces or high dimensional vectors is also popular across problems in natural language processing, computer vision and speech. When dealing with unstructured sets of high dimensional points, it is common to provide supervision in forms of anchor points [Mikolov et al., 2013a, Xing et al., 2015]. Recently, unsupervised alignment approaches have obtained compelling results by framing the problem as a distance minimization between distributions either by adversarial training [Conneau et al., 2017b, Zhang et al., 2017c] or by non-adversarial techniques [Hoshen and Wolf, 2018b, Mukherjee et al., 2018]. Unsupervised alignent approaches have also found success in unsupervised domain alignment [Benaim and Wolf, 2017, Hoshen, 2018] and neural style transfer [Liao et al., 2017].

Supervised alignment methods rely on paired aligned instances or some form of supervisory signal or access to explicit alignment between instances. In word translation [Mikolov et al., 2013c] propose to use a seed dictionary of 5000 words for cross-lingual word translation. Many of the supervised algorithms take inspiration from unsupervised alignment. [Bojanowski et al., 2014, 2015] propose a method similar to canonical time warping and augment it with supervisory aligned signals for model training. [Plummer et al., 2018, 2015] uses CCA to find latent space where image regions align to phrases.

The growing availibility of aligned language and vision datasets [Mao et al., 2016, Plummer et al., 2015] has allowed deep learning algorithms to gain popularity. [Mao et al., 2016] used a CNN to model visual data and LSTM language model to evaluate the matching between an image region and referring expression. [Chan et al., 2016] model consists of an encoder RNN network named *listener* and decoder RNN network named *speller* which are trained jointly to map low level speech signals to output utternaces.

Translation A major challenge in multiview multimodal learning is concerned with translating from one domain or view to the other so that the semantics of each domain is preserved. Translation is a widely studied problem in multimodal learning with applications in caption generation [Karpathy and Fei-Fei, 2017, Vinyals et al., 2015], video captioning [Krishna et al., 2017], image to image translation [Zhu et al., 2017], cross-modal retrieval [Rasiwasia et al., 2010]. As can be seen by the noticable efforts of computer vision and natural language processing communities in generating large scale aligned datasets, multimodal translation is a problem of growing interest. Popular problems include video and image captioning [Venugopalan et al., 2015, Vinyals et al., 2015], image to image translation [Zhu et al., 2017, Isola et al., 2017], style transfer in text [Shen et al., 2017, Mueller et al., 2017]. While there are multiple approaches for multimodal translation, we broadly categorize them as *combination* based approaches and *generative* approaches. Combination based approaches are motivated by the fact that modalities often have common structure and syntax and form a model dictionary which can be further exploited for domain translation. Most of the rules for combinations are hand crafted or based on heuristics [Baltrušaitis et al., 2017]. [Kuznetsova et al., 2012] use a two stage approach: Firstly, they use a retrieval framework to retrieve candidate phrases. Secondly, generate a coherent description using integer linear programming (ILP) formulation. [Gupta et al., 2012] first retrieve a set of $k-$ candidate images similar to source images and then use phrases collected from a dictionary to generate a target sentence. Combination based models are flexible and generate translations but are restrictive to include the presence of a large dictionary, often making them expesnive to make inference.

Generative models for translation require understanding the source modality to generate a consistent and meaningful target modality. Due to the large space of possible correct answers, these models can be quite challenging. Earlier approaches relied on pre-defined grammar or template based models to generate a modality. [Kojima et al., 2002] proposed a system to determine human behaviour from videos and used a template based system to generate a description. Babytalk [Guadarrama et al., 2013] extracted triples of the order *subject,object,predicate* and combine with a conditional random field (CRF) to generate a sentence. [Li et al., 2011] take a two step approach where the first step selects candidate phrases useful for description and a second phase of fusion which finds an optimal and compatible set of phrases using dynamic programming. [Mitchell et al., 2012] use a tree-generating process rather than a template based process similar to a tree substitution grammar, which allows for descriptions to be syntactically well formed. [Kulkarni et al., 2013] which given an image generates triple *subject*, *object* and *predicate* that is used with conditional random fields to generate sentences. [Yang et al., 2011] use object detectors and scenes from an input image, estimating a quadrupulet structure of object, actions, scene and propostion which is used with a HMM graphical model. An advantage of generative models using syntax is that they are likely to generate logically correct and meaningful sentences. However the use of complex pipelines severly limits them.

Deep learning generative models are a recent addition used for multimodal translation. The popular architecture is the encoder-decoder model where an encoder is used to model the source modality and the decoder is used to generate the target modality all in single pass. The encoder-decoder architecture popular in neural machine translation [Cho et al., 2014], have also been used in image captioning [Mao et al., 2015, Vinyals et al., 2015], video description [Rohrbach et al., 2015, Venugopalan et al., 2015]. Popular encoders to model the source modality include using RNNs to model acoustic features [Prabhavalkar et al., 2017]. For words or sentences they are mostly encoded using distributional semantic models [Mikolov et al., 2013b, Pennington et al., 2014]. Images are mostly encoded using CNN or their variants [Krizhevsky et al., 2012, Simonyan and Zisserman, 2015b, He et al., 2016]. The step of decoding also uses an RNN or an LSTM using the encoded representation as the initial hidden state. Various extensions and strategies have been discussed in the literature to aid in the translation process [Venugopalan et al., 2015, Rohrbach et al., 2015]. A problem generally encountered using an RNN is that the model has to generate a description, image or sound from a single vector which is impoverished to handle long range dependencies. It was observed

[1] that reversing the source sentence i.e feeding it backwards to the encoder produces significantly better results as it shortens the path from the decoder to the relevant parts of the encoder. Similar results were observed when feeding the input sentence twice to the encoder to better memorize things. The advantage of the attention mechanism is that one can avoid such hacks and allow the decoder to attend to different parts of the source sentence at each step of the output generation. Attention based models have been succesfully applied in neural machine translation [Bahdanau et al., 2014], neural caption generation [Xu et al., 2015], video description [Yao et al., 2015].

2.2 Zero-Shot learning

Recent major progress in visual recognition has been driven by training complex models using a large number of annotated training examples. Scaling this paradigm to many categories is not feasible due to the need to collect and to annotate many examples of every category to recognize. Zero shot recognition provides a paradigm to eliminate the need to annotate each new category, once a certain number of background categories has been learned. Specifically, it does this by cross-modal transfer from language. The idea is to use a limited set of training data to learn a lingusitic-visual mapping; and then apply the induced function to map vectors representing novel entities unseen during training to the visual domain or a shared embedding, thus enabling recognition in the absence of visual training examples. As discussed earlier in Sec 2.1, this can be seen as a special limiting case of *translation*: learn a translation model from language description to a visual classifier given a set of aligned language description and visual examples. An inverse translation is also feasible where one learns a mapping function from image to language description and then match in language domain. The task of ZSL was originally tested for neural decoding [Palatucci et al., 2009b, Mitchell et al., 2008], mapping fMRI activations to word vectors, and then applying it to the brain signal of a concept outside the training set, in order to read from the brain. More recently, it has generated big impact in visual recognition [Lampert et al., 2014, Socher et al., 2013a, Lazaridou et al., 2014] due to the potential for leveraging language to help visual recognition scale: to many categories without work intensive image annotation, or to fine-grained/rare categories where extensive training data may simply be unavailable.

Distributed semantic models (DSMs) typically generate *vector* embeddings of words, and hence existing zero-shot methods mostly focus on establishing a cross-modal mapping between DSMs of category *name*, and visual examples of that category. However, such vector representation have limited expressivity providing no notion of various intra-class variances. Point vectors are compared using a series of operations comprising of dot product, cosine distance or Euclidean distance which are incapable to represent assymetric or hierarchical relationships. In this work we represent visual and linguistic vectors as *Gaussian distribution* [Vilnis and McCallum, 2015]. Representing words as

[1] http://www.wildml.com/2016/01/attention-and-memory-in-deep-learning-and-nlp/

distributions was initially done by [Vilnis and McCallum, 2015] where the mean vector represents the semantics and the covariance describes the uncertainty in the meanings. Our proposed distribution-based approach provides a representation of intra-class variability that improves zero-shot recognition, allows more meaningful retrieval by multiple keywords, and also produces better point-estimates of word vectors.

2.3 Supervised and Unsupervised Pairing

Much of the success of deep learning can be attributed to big datasets annotated with explicit correspondence between the modalities. Learning correspondence between data is a fundermental building block of many applications which can be used to sort, align and rank data. Given data from two sources, the problem of learning correspondence finds applications in multi-modal settings. In image captioning [Karpathy and Fei-Fei, 2017] , images are usually accompanied with descriptions. Neural machine translation [Bahdanau et al., 2014] expects a parallel corpus of source and target language.

Most successful methods heavily rely on cross-lingual supervision in the form of translation dictionaries [Mikolov et al., 2013a, Vulic and Korhonen, 2016] or sentence aligned corpus to derive bilingual word vectors which now have a notion of word association between the corpus [Gouws et al., 2015, Luong et al., 2015]. However to collect or assume the presence of sentence aligned or parallel corpus is quite an unreasonable assumption in real-world settings. This leads us into exploring the possibility of learning explicit correspondence without any form of supervision.

Learning correspondence across domains is also relevant in computer vision. Image matching is a long standing problem in computer vision with several applications ranging from scene recognition to optical flow estimation [Forsyth and Ponce, 2002, Szeliski, 2010]. Most notable image matching have been based on feature matching or pixel based matching. Earlier approaches were based on using descriptors such as SIFT [Lowe, 2004] or HOG [Dalal and Triggs, 2005]. With recent advances in deep learning especially generative adversarial networks (GAN) [Goodfellow et al., 2014], the problem of image to image translation has gained importance provided it receives paired data. [Isola et al., 2016] uses a GAN where the discriminator receives a pair of images where one image is the source image and the other image is the paired image or generated image (fake pair). The link between the source and target is further strengthened by the U-net architecture [Ronneberger et al., 2015]. While learning correspondence across domains require sample sets of supervision in the form of bilingual dictionaries [Mikolov et al., 2013c] for cross-lingual transfer of word embeddings or matching pairs of images [Isola et al., 2016] for style transfer, a question which needs to be asked is whether can such a mapping or correspondence can be learnt without sample pairs or presence of any supervision ?

Recent progress in GANs [Goodfellow et al., 2014] has led to major developments in image to image translation techniques and it comes as no surprise that the state of the art translation is employed by variants of GANS. The most popular of them has been CycleGAN [Zhu et al., 2017] which employs the cycle consistency as a constraint. Other variants include DiscoGAN [Kim et al., 2017],DualGAN [Yi et al., 2017] which include additional constraints. In natural language processing, unsupervised transfer of monolingual word embeddings has been gaining attention especially through adverserial techniques. [Zhang et al., 2017d] adopt GAN to transform from a source monolingual embedding to target monolingual embedding. [Conneau et al., 2017a] use an improved adversarial training along with a refinement procedure for cross-lingual word mapping. More recent works use the cyclic consistency of CycleGAN into back translation loss and adopt the sinkhorn distance into the objective function [Xu et al., 2018].

In this thesis we address unsupervised learning of cross-modal pairing. Unlike these other approaches we do not use GAN or adverserial training, which makes our approach easier and more stable to train.

Chapter 3

Zero Shot Learning with Gaussian Category Embeddings

3.1 Introduction

Learning vector representations of word meaning is a topical area in computational linguistics. Based on the distributional hypothesis – that words in similar context have similar meanings – distributed semantic models (DSM)s build vector representations based on corpus-extracted context. DSM approaches such as topic models [Blei et al., 2003], and more recently neural networks [Collobert et al., 2011, Mikolov et al., 2013c] have had great success in a variety of lexical and semantic tasks [Arora et al., 2015, Schwenk, 2007].

However despite their successes, classic DSMs are severely impoverished compared to humans due to learning solely from word co-occurrence without grounding in the outside world. This has motivated a wave of recent research into multi-modal and cross-modal learning that aims to *ground* DSMs in non-linguistic modalities [Bruni et al., 2014, Kiela and Bottou, 2014, Silberer and Lapata, 2014]. Such multi-modal DSMs are attractive because they learn richer representations than language-only models (e.g., that bananas are *yellow* fruits [Bruni et al., 2012b]), and thus often outperform language only models in various lexical tasks Bruni et al. [2012a].

In this thesis, we focus on a key unique and practically valuable capability enabled by cross-modal DSMs: that of zero-shot learning (ZSL). Zero-shot recognition aims to recognise visual categories in the absence of any training examples by cross-modal transfer from language. The idea is to use a limited set of training data to learn a linguistic-visual mapping and then apply the induced function to map images from novel visual categories (unseen during training) to a linguistic embedding: thus enabling recognition in the absence of visual training examples. ZSL has generated big impact

[Lampert et al., 2009a, Socher et al., 2013a, Lazaridou et al., 2014] due to the potential of leveraging language to help visual recognition scale to many categories without labor intensive image annotation.

DSMs typically generate *vector* embeddings of words, and hence ZSL is typically realised by variants of vector-valued cross-modal regression. However, such vector representations have limited expressivity – each word is represented by a point, with no notion of intra-class variability. In this paper, we consider ZSL in the case where both visual and linguistic concepts are represented by *Gaussian distribution* embeddings. Specifically, our Gaussian-embedding approach to ZSL learns concept distributions in both domains: Gaussians representing individual words (as in [Vilnis and McCallum, 2015]) and Gaussians representing visual concepts. Simultaneously, it learns a cross-domain mapping that warps language-domain Gaussian concept representations into alignment with visual-domain concept Gaussians. Some existing vector DSM-based cross-modal ZSL mappings [Akata et al., 2013, Frome et al., 2013a] can be seen as special cases of ours where the within-domain model is pre-fixed as vector corresponding to the Gaussian means alone, and only the cross-domain mapping is learned. Our results show that modeling linguistic and visual concepts as Gaussian distributions rather than vectors can significantly improve zero-shot recognition results.

3.2 Related Work

3.2.1 Distributed Semantic Models

Finding good representation of words which convey meaning is an important research direction in cognitive science. Distributed semantic models (DSM) motivated by *distributional hypothesis* [Harris, 1954] have a long history in cognitive science, psycology and linguistics [Firth, 1957, Miller and Charles, 1991, Wittgenstein, 1953]. Contemporary vector space representations are generated by word context, with the assumption that word similarity is then reflected by geometric similarity of their context vectors. DSM are typically represented through vector space models (VSM) where the word tokens are represented as a vector in high dimensional space. The earliest application of vector based models was explored in Information Retrieval where a document was represented by *vectors* [Salton et al., 1975] with the whole vocabulary represented as dimensions. The weights of individual tokens were either computed using the frequency of their appearance or normalized frequencies. Vector based representation have been applied in various applications ranging from information retrieval [Lee et al., 1997], text classification [Soucy and Mineau, 2005] to sentiment analysis [Turney, 2002]. [Turney and Pantel, 2010] provide a comprehensive survey for vector space models of meaning and its applications in various language domains. Later deep learning based approaches have been exploited for learning low dimensional representations of natural language text popularly called as *word embeddings.* These word embeddings have been attractive and have been applied in multiple NLP downstream applications [Zou et al., 2013, Kim, 2014, Weiss et al., 2015].

In this thesis, we are particularly interested in using DSMs to bridge linguistic and visual modalities, so we focus specifically on multi-modal DSMs.

3.2.2 Multi-modal semantics

Computational linguistic models of meaning that rely of contextual information provide a good approximation to word meaning, since semantically similar words tend to have similar contextual distributions. Distributional semantic models use vectors to keep track of the contexts in which target terms appear in a large corpus as proxies for meaning representations, and apply geometric techniques to these vectors to measure the similarity or relatedness of to the corresponding words [Allen et al., 2019, Allen and Hospedales, 2019].

Distributional semantic models (DSM) have been criticized in that they represent the meaning of a word solely by connection with other words in a corpus. There is increasing realisation that meaning of a word is not only acquired from linguistic environment but is essentially *grounded* to the external world through multiple channels [Landau et al., 1998].

Multi-modal semantics are motivated from human *concept acquisition* where learned linguistic representations are grounded in other modalities such as vision – as well as obtaining better representations to improve performance on linguistic tasks, and developing cross-modal mappings. To address the *grounding* problem, and enrich concept vectors with visual information, early studies simply *concatenated* conventional uni-modal linguistic DSM representations with uni-modal visual representations (e.g., gradient histograms such as SIFT) from corresponding image categories [Bruni et al., 2012a,b, Kiela and Bottou, 2014]. This improved the representation and resulting performance on a variety of tasks but did not provide a truly integrated and synergistically learned multi-modal representation. Thus more recent studies have focused on jointly learning multi-modal models, for example with multi-modal auto encoders [Silberer and Lapata, 2012] or Boltzmann machines [Srivastava and Salakhutdinov, 2012], multi-modal skip-gram models [Lazaridou, 2015], deep embeddings [Frome et al., 2013b] and dependency tree recursive neural networks [Socher et al., 2014]. These models have been shown to be successful in various concept learning tasks [Silberer and Lapata, 2014].

3.2.3 Zero Shot Learning

An exciting and practically valuable property of learning multi-modal semantics is the ability to do zero-shot learning [Palatucci et al., 2009a]. Applied across language and vision domains, ZSL corresponds to the ability to recognise a visual category without requiring *any* annotated examples, let alone the extensive sets typically required for state of the art supervised learning. ZSL has generated extensive interest in both computational linguistic [Lazaridou et al., 2014], machine learning [Palatucci et al., 2009a, Frome et al., 2013b] and computer vision [Fu et al., 2014] communities. Language-driven ZSL is typically realised by learning text (e.g., DSM vector) and visual domain (e.g., CNN activation) representations using an auxiliary dataset and mapping them into a common embedding. Then at test time, given the name of a previously (visually) unseen category, its DSM vector and thus its visual embedding can be generated, allowing it to be matched (e.g., using nearest-neighbour) to images for recognition. Thus ZSL can be seen as a form of cross-modal knowledge transfer from language to vision [Socher et al., 2013b]. The simplest way to realise ZSL is to generate fixed and independent linguistic and visual representations, and then learn a mapping between them [Lazaridou et al., 2014, Fu et al., 2014, Socher et al., 2013b]. However, reflecting the same research progression in broader multi-modal semantics, more sophisticated approaches have also been proposed that simultaneously learn both representations and the mapping between them [Frome et al., 2013b] where such joint multi-modal learning is typically more effective.

3.3 Methodology

3.3.1 Background

Vector Word Embeddings In a typical setup for unsupervised learning of word-vectors, we observe a sequence of tokens $\{w_i\}$ and their context words $\{c(w)_i\}$. The goal is to map each word w to a d-dimensional vector e_w reflecting its distributional properties. Popular skip-gram and CBOW models [Mikolov et al., 2013c], learn a matrix $W \in \mathbb{R}^{|V| \times d}$ of word embeddings for each of V vocabulary words ($e_w = W_{(w,:)}$) based on the objective of predicting words given their contexts.

Another way to formalise a word vector representation learning problem is to search for a representation W so that words w have high representational similarity with context words $c(w)$, and low similarity with representations of words not in context $\neg c(w)$. This could be expressed as optimisation of max-margin loss J; requiring that each word w's representation e_w is more similar to that of context words e_p than non-context words e_n by a margin δ

$$J(W) = \sum_{w, w_p \in c(w), w_n \in \neg c(w)} \max(0, \delta - E(e_w, e_{w_p}) + E(e_w, e_{w_n})) \tag{3.1}$$

where similarity measure $E(\cdot, \cdot)$ is a distance in $\mathbb{R}^d$ space such as cosine or euclidean.

Gaussian Word Embeddings Vector-space models are successful, but have limited expressivity in terms of modelling the variance of a concept, or asymmetric distances between words, etc. This has motivated recent work into *distribution*-based embeddings [Vilnis and McCallum, 2015]. Rather than learning word-vectors e_w, the goal here is now to learn a distribution for each word, represented by a per-word mean μ_w and covariance Σ_w.

In order to extend word representation learning approaches such as Eq. (3.1) to learning Gaussians, we need to replace vector similarity measure $E(\cdot,\cdot)$ with a similarity measure for Gaussians. We follow [Vilnis and McCallum, 2015] in using the inner product between distributions f and g – the probability product kernel [Jebara et al., 2004].

$$E(f,g) = \int_{x\in\mathbb{R}^n} f(x)g(x). \tag{3.2}$$

The probability product kernel (PPK) has a convenient closed form in the case of Gaussians:

$$\begin{aligned} E(f,g) &= \int_{x\in\mathbb{R}^n} \mathcal{N}(x;\mu_f,\Sigma_f)\mathcal{N}(x;\mu_g,\Sigma_g)dx \\ &= \mathcal{N}(0;\mu_f-\mu_g,\Sigma_f+\Sigma_g) \end{aligned} \tag{3.3}$$

where μ_f,μ_g are the means and Σ_f,Σ_g are the covariances of the probability distribution f and g.

3.3.2 Gaussian models of images and text

Distributed representation of word embeddings has shown the ability to capture semantic and syntactic relationships [Mikolov et al., 2013c, Pennington et al., 2014]. However due to their inability to model uncertainty we represent words as distributions [Vilnis and McCallum, 2015].

Given a pre-trained set of word embeddings which would represent the means, we describe a simple procedure to construct the empirical covariances motivated by [Vilnis and McCallum, 2015]. For a word w and its context represented by $\{c(w)_i\}$ and window size W, the variance is

$$\Sigma_w = \frac{1}{W}\sum_{i=1}^{W}(c(w)_i - w)^T(c(w)_i - w) \tag{3.4}$$

3.3.3 Cross-Modal Distribution Mapping

Gaussian models of words can be learned as described in Sec 3.3.2, and that Gaussian models of image categories can be trivially obtained by maximum likelihood. The central task is therefore to establish a mapping between word-and image-Gaussians, which will be of different dimensions d_w and d_x.

We aim to find a projection matrix $A \in \mathbb{R}^{d_x\times d_w}$ such that a word w generates an image vector as $e_x = Ae_w$. Working with distributions, this implies that we have $\mu_x = A\mu_w$ and $\Sigma_x = A\Sigma_w A^T$. We can now evaluate the similarity of concept distributions across modalities. The similarity between image-and text-domain Gaussians f and g is:

$$E(f,g) = \mathcal{N}(0;\mu_f - A\mu_g, \Sigma_f + A\Sigma_g A^T) \tag{3.5}$$

Using this metric, we can train our cross-modal projection A via the cross-domain loss:

$$J(A) = \sum_{f,g \in P, k \in N} \max(0, \delta - E(f,g) + E(f,k)) \tag{3.6}$$

where P is the set of matching pairs that should be aligned (e.g., the word Gaussian 'plane' and the Gaussian of plane images) and N is the set of mismatching pairs that should be separated (e.g., 'plane' and dogs images). This can be optimised with SGD using the gradient:

$$\begin{aligned} \frac{\partial E}{\partial A} &= \frac{1}{2}((\Sigma_f + A\Sigma_g A^T)^{-1} A(\Sigma_g + \Sigma_g^T)) \\ &+ ((\mu_g^T(\Sigma_f + A\Sigma_g A^T)^{-1}(\mu_f - A\mu_g) \\ &+ (\mu_f - A\mu_g)^T(\Sigma_f + A\Sigma_g A^T)^{-1}\mu_g^T \\ &+ (\mu_f - A\mu_g)^T(\Sigma_f + A\Sigma_g A^T)^{-1} \\ &A^T(\Sigma_g + \Sigma_j^T)(\Sigma_f + A\Sigma_g A^T)^{-1}(\mu_f - A\mu_g)) \end{aligned} \tag{3.7}$$

3.3.4 Joint Representation and Mapping

The cross-domain mapping A can be learned by picking an energy function (Eq. 3.5), a loss function (max-margin) (Eq. 3.6) and a set of positive and negative training pairs. It is also possible to simultaneously learn the mapping along with the text and image-domain gaussians ($\{\mu_f, \Sigma_f\}^{text}, \{\mu_g, \Sigma_g\}^{img}$) by optimising the sum of three coupled losses: Eq. 3.1 with Eq. 3.3, Eq. 3.6 and max-margin image-classification using Gaussians. We found jointly learning the image-classification Gaussians did not bring much benefit over the MLE Gaussians, so we only jointly learn the text Gaussians and cross-domain mapping. Algorithm 3.3.4 summarizes the training procedure.

Algorithm 1 Algorithm for Cross-Modal Training

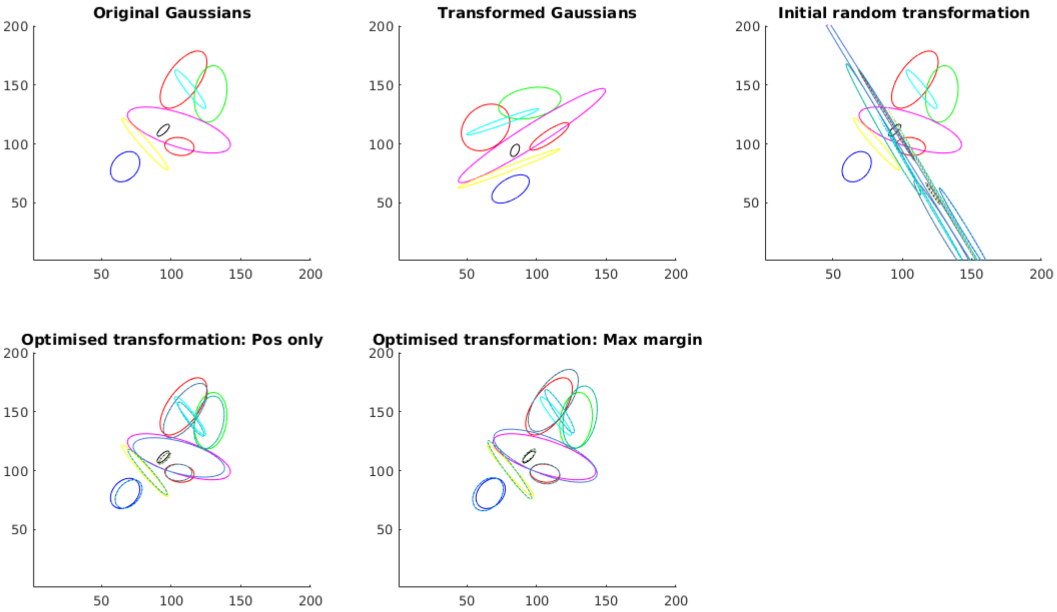
```
procedure TRAINING($D_s, D_{text}$) // $D_s$ is cross-modal annotated category name and image pairs,$D_{text}$ is a text-corpus
    Initialize : $\{\mu_g, \Sigma_g\}$,A
    Train: $\{\mu_f, \Sigma_f\}$ by MLE
    while not converged
        Sample $f, g, k \sim D_s, w_p, w_n \sim D_{text}$
        Gradient step on Eq 3.1+ Eq 3.6
end procedure
```

3.3.5 Synthetic Data

We try to simulate a model where we are initially provided some word vectors and then transform them to an image vectors. Now given the original word vectors and image word vectors, can we recover the matrix used for the transformation ?

We simulate some data from normal distribution. We project this data through a projection matrix A and get another transformed matrix.This will be 2-D simulations of word vectors and image vectors. Now using Eq 3.3 can we recover the transformation matrix A.

Figure 3.1: Simulation from synthetic gaussian distributions



In Fig 3.1 the top left corner represents the original gaussian and the centre image represents the transformation by using a projection matrix A. The top right image just projects the image vectors through some random transformation matrix. The bottom left image represents by just using Eq 3.3 i.e without using the max-margin framework. The bottom right uses the complete framework. One can observe that we are able to recover the original word vectors (approximately).

3.3.6 Application to Zero-Shot Recognition

Once the text-domain Gaussians and cross-domain mapping have been trained for a set of known words/classes, we can use the learned model to recognise any novel/unseen but name-able visual category w as follows:

Algorithm 2 Algorithm for Zero Shot Recognition

```
procedure TESTING(x,{$\mu_w$,$\Sigma_w$})//Input test image and the set of known categories subscripted by $w$.
    for each target category $w$
    $p(x|w) \propto \mathcal{N}(x|A\mu_w, A\Sigma_w A^T)$
    end for
    return $w^* = \text{argmax}(p(x|w))$ //Return the ML category.
end procedure
```

3.3.7 Contextual Query

To illustrate our approach, we also experiment with a new variant of the ZSL setting. In conventional ZSL, a novel word can be matched against images by projecting it into image space, and sorting images by their distance to the word (vector), or likelihood under the word (Gaussian). However, results may be unreliable when used with polysemous words, or words with large appearance variability. In this case we may wish to enrich the query with contextual words that disambiguate the visual meaning of the query. With regular vector-based queries, the typical approach is to sum the word-vectors. For example: For contextual disambiguation of polysemy, we may hope that vec('bank')+vec('river') may retrieve a very different set of images than vec('bank')+vec('finance'). For specification of a specific subcategory or variant, we may hope that vec('plane')+vec('military') retrieves a different set of images than vec('plane')+vec('passenger'). Fig 3.2 illustrates the contextual concept with the plane example where we can see that different intersection (Eq 3.3) between word Gaussians map to different regions in image space.

By using distributions rather than vectors, our framework provides a richer means to make such queries that accounts for the intra-class variability of each concept. Consider an example of a contextual query represented by two words. When each word is represented by a Gaussian with means μ_1 and μ_2, and covariances of Σ_1 and Σ_2 respectively, a two-word query can be represented by their product, which is the new Gaussian with mean and covariance

$$\mu = \frac{\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2}{\Sigma_1^{-1} + \Sigma_2^{-1}}$$

$$\Sigma = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1})$$

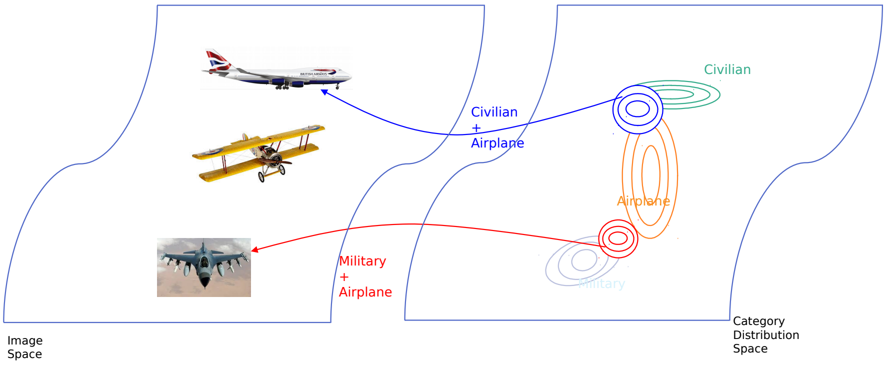


Figure 3.2: Schematic illustration of contextual query. Querying the conjunction of different words is achieved by the product their corresponding Gaussians and mapping the Gaussian intersection to image space for retrieval

3.4 Experiments

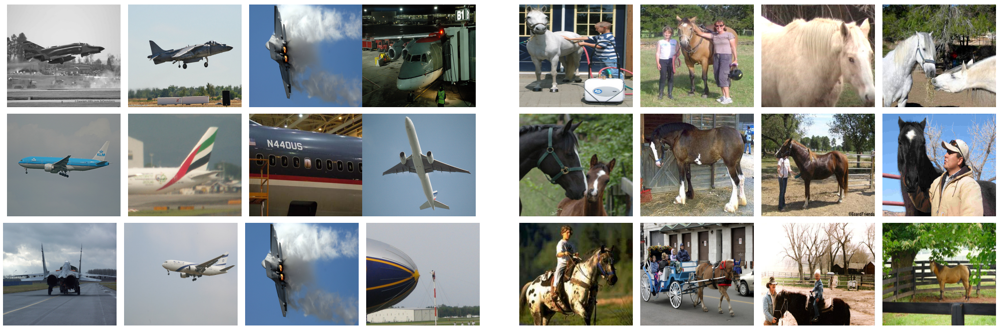
3.4.1 Datasets and Settings

Datasets: We evaluate our method [1] using the main Animals with Attributes (**AWA**) and **ImageNet1K** benchmarks. To extract visual features we use the VGG-16 CNN [Simonyan and Zisserman, 2015a] to extract a $d_x = 4096$ dimensional feature for each image. To train the word Gaussian representation, we use a combination of UkWAC Ferraresi et al. [2008] and Wikipedia corpus of 25 million tokens, and learn a $d_w = 100$ dimensional Gaussian representation with spherical covariance. We set our margin parameter to $\delta = 1$. We use mini-batch stochastic gradient descent with learning rate set at $1e - 3$ and batch size to 128.

Settings: Our zero-shot setting involves training a visual recogniser (i.e., our mapping A) on a subset of classes, and evaluating it on a disjoint subset. For AWA, we use the standard 40/10 class split [Lampert et al., 2009a], and for ImageNet we use a standard 800/200 class split [Mensink et al., 2012].

Competitors: We implement a set of representative alternatives for direct comparison with ours on the same visual features and text corpus. These include: cross-modal linear regression (LinReg, [Dinu et al., 2015]), non-linear regression (NLinReg, [Lazaridou

[1] Code and datasets kept at `http://bit.ly/2cI64Zf`



(a) Top: ‘Military’+‘Plane’ (Gaussian), Middle: ‘Passenger’+‘Plane’ (Gaussian), Bottom: ‘Passenger’+’Plane’ (Vector)

(b) Top: ‘White’+‘Horse’ (Gaussian), Middle: ‘Black’+‘Horse’ (Gaussian), Bottom: ‘Black’+’Horse’ (Vector)

Figure 3.3: Qualitative visualisation of zero-shot query with context words.

	Vector space models				Ours
Dataset	LinReg	NLinReg	CME	ES-ZSL	Gaussian
AWA	44.0	48.4	43.1	58.2	65.4

Table 3.1: Zero-shot recognition results on AWA (% accuracy).

et al., 2014, Socher et al., 2013a]), ES-ZSL [Romera-Paredes and Torr, 2015], and a max-margin cross-modal energy function method (CME, Akata et al. [2013], Frome et al. [2013a]). Note that the CME strategy is the most closely related to ours in that it also trains a $d_x \times d_w$ matrix with max-margin loss, but uses it in a bilinear energy function with vectors $E(x, y) = x^T Ay$; while our energy function operates on Gaussians.

3.4.2 Results

Table 3.1 compares our results on the AWA benchmark against alternatives using the same visual features, and word vectors trained on the same corpus. We observe that: (i) Our Gaussian-embedding obtains the best performance overall. (ii) Our method outperforms CME which shares an objective function and optimisation strategy with ours, but operates on vectors rather than Gaussians. This suggests that our new distribution rather than vector-embedding does indeed bring significant benefit.

A comparison to published results obtained by other studies on the same ZSL splits is given in Table 3.2, where we see that our results are competitive despite exploitation of supervised embeddings such as attributes [Fu et al., 2014], or combinations of embeddings [Akata et al., 2013] by other methods.

We next demonstrate our approach qualitatively by means of the contextual query idea introduced in Sec 3.3.7. Fig. 3.3 shows examples of how the top retrieved images differ intuitively when querying ImageNet for zero-shot categories ‘plane’ and ‘horse’ with different context words. To ease interpretation, we constrain the retrieval to the true target class, and focus on the effect of the context word. Our learned Gaussian method retrieves more relevant images than the word-vector sum baseline. E.g., with

ImageNet	
ConSE [Norouzi et al., 2014]	28.5%
DeVISE [Frome et al., 2013a]	31.8%
Large Scale Metric. [Mensink et al., 2012]	35.7%
Semantic Manifold. [Fu et al., 2015b]	41.0%
Gaussian Embedding	45.7%
AwA	
DAP (CNN feat) [Lampert et al., 2009a]	53.2%
ALE [Akata et al., 2013]	43.5%
TMV-BLP [Fu et al., 2014]	47.1%
ES-ZSL [Romera-Paredes and Torr, 2015]	49.3%
Gaussian Embedding	65.4%

Table 3.2: Comparison of our ZSL results with state of the art.

the Gaussian model all of the top-4 retrieved images for Passenger+Plane are relevant, while only two are relevant with the vector model. Similarly, the retrieved black horses are more clearly black.

3.4.3 Further Analysis

To provide insight into our contribution, we repeat the analysis of the AWA dataset and evaluate several variants of our full method. These use our features, and train the same cross-domain max-margin loss in Eq 3.6, but vary in the energy function and representations used. Variants include: (i) Bilinear-WordVec: Max-margin training on word vector representations of words and images with a bilinear energy function. (ii) Bilinear-MeanVec: As before, but using our Gaussian means as vector representations in image and text domains. (iii) PPK-MeanVec: Train the max-margin model with Gaussian representation and PPK energy function as in our full model, but treat the resulting means as point estimates for conventional vector-based ZSL matching at testing-time. (v) PPK-Gaussian: Our full model with Gaussian PPK training and testing by Gaussian matching.

From the results in Table 3.3, we make the observations: (i) Bilinear-MeanVec outperforming Bilinear-WordVec shows that cross-modal (Sec 3.3.4) training of word Gaussians learns better point estimates of words than conventional word-vector training, since these only differ in the choice of vector representation of class names. (ii) PPK-Gaussian outperforming PPK-MeanVec shows that having a model of intra-class variability (as provided by the word-Gaussians) allows better zero-shot recognition, since these differ only in whether covariance is used at testing time.

3.4.4 Discussion

Existing visual semantic methods model texts and images as vectors in the semantic space. As pointed out , the popular DSM based word embeddings severely lack representation capability. In this work, we explore the case where images and texts are

AwA	
Bilinear-WordVec	43.1%
Bilinear-MeanVec	52.2%
PPK-MeanVec	52.6%
PPK-Gaussian	65.4%

Table 3.3: Impact of training and testing with distribution rather than vector-based representations

represented as distributions. Our visual-linguistic mapping is able to learn cross-domain mapping by aligning language domain Gaussian concepts to visual-domain Gaussian concepts.

Our approach models intra-class variability in both images and text. For example, the variability in visual appearance of military versus passenger planes, and the variability in context according to whether the word 'plane' is being used in a military or civilian sense. Given distribution-based representations in each domain, we find a cross-modal map that warps the two distributions into alignment.

Concurrently with our work, [Ren et al., 2016] present a related study on distribution-based visual-text embeddings. Methodologically, they benefit from end-to-end learning of deep features as well as cross-modal mapping, but they only discriminatively train word covariances, rather than jointly training both means and covariances as we do.

With regards to efficiency, our model is fast to train if fixing pre-trained word-Gaussians and optimising only the cross-modal mapping A. However, training the mapping jointly with the word-Gaussians comes at the cost of updating the representations of all words in the dictionary, and is thus much slower.

In terms of future work, an immediate improvement would be to generalise our Gaussian embeddings to model concepts as mixtures of Gaussians or other exponential family distributions [Rudolph et al., 2016, Chen et al., 2015]. This would for example, allow polysemy to be represented more cleanly as a mixture, rather than as a wide-covariance Gaussian as happens now. We would also like to explore distribution-based embeddings of sentences/paragraphs for class description (rather than class name) based zero-shot recognition [Reed et al., 2016]. Finally, besides end-to-end deep learning of visual features, going beyond our linear mapping A, and training non-linear cross-modal mappings is also of interest.

3.5 Conclusion

In this chapter, we advocate using distribution-based embeddings of text and images when bridging the gap between vision and text modalities. Instead of modelling text and images as vectors as commonly practised, we advocate modeling them as distributions. We focus on the unique ability of zero-shot learning showing improved results. Our distribution-based approach provides a representation of intra-class variability that improves zero-shot recognition, allows more meaningful retrieval by multiple keywords, and also produces better point-estimates of word vectors.

An improvement to the above model would be to model concepts as mixtures of Gaussians or exponential families [Rudolph et al., 2016, Chen et al., 2015]. Words with flexible structure will be able to capture subtle word meanings and advance the state of the art. Distributions are naturally able to represent that words do not have single precise meanings but are naturally able to capture multiple semantic information. Other future work will include exploring better and efficient training mechanism, hyperbolic geometry and optimal transport based models.

Chapter 4

Unsupervised Pairing of Multi-modal data

4.1 Introduction

An implicit assumption in modern machine learning algorithms is presence of paired data. Succesful applications include image captioning [Vinyals et al., 2015, Karpathy and Fei-Fei, 2017], neural machine translation [Bahdanau et al., 2014] and end-to-end speech recognition [Graves and Jaitly, 2014]. Most algorithms require similarity measures between domains or association between domains. If such an information is provided, one can obtain a mapping function from one domain to another. However providing such annotations and correspondence can be quite expensive and proves to be a bottleneck in developing truly intelligent agents. In this chapter we introduce methodologies to learn cross-modal mappings from unpaired data.

The classical method to study paired samples has been dominated by Canonical Correlation Analysis (CCA) [Hotelling, 1936b], a classical yet powerful tool. CCA links two sources by maximizing the correlation between the sources or the views. CCA has been studied and generalized to add regularization [Mardia et al., 1979], kernelized [Lai and Fyfe, 2000, Schölkopf et al., 1999]. With the excitement around deep learning, Deep CCA has been developed [Andrew et al., 2013] and showed promise in multi-modal applications [Wang et al., 2015, Zeng et al., 2018].

While CCA and its variants have enjoyed success, they require access to paired samples or representations in the respective domains. Recently, work has emerged which studies mapping these domains in an unsupervised way. They have been applied in learning domain mapping [Hoshen, 2018], image to image translation [Kim et al., 2017, Zhu et al., 2017] and bilingual lexicon induction [Conneau et al., 2017b, Haghighi et al., 2008].

In this chapter, we introduce some background material explaining prior unsupervised matching algorithms, and providing some background methodology that we will exploit later. We start by studying CCA and its unsupervised variant of Matching CCA [Haghighi et al., 2008]. This will be useful in studying bilingual lexicon induction which we will focus in Chapter 5. and unsupervised matching of images and text which will be a focal point in Chapter 6.

4.2 Canonical Correlation Analysis

Canonical Correlation Analysis is a classical method for dimensionality reduction for two paired data sources which finds a subspace that maximizes the correlation between the data sources. Let dataset $\mathcal{D}$ contain two sets of vectors from two sources $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{n}$ where $\boldsymbol{x} \in \mathbb{R}^{d1}$ and $\boldsymbol{y} \in \mathbb{R}^{d2}$. Both the datasets are assumed to be centered which can be achieved by subtracting the sample mean from each sample. CCA finds two bases $\mathbf{w}_x$ and $\mathbf{w}_y$ such that their correlation is maximized as

$$\max_{(\mathbf{w}_x, \mathbf{w}_y)} \frac{\mathbf{w}_x^T \mathbf{C}_{XY} \mathbf{w}_y}{\sqrt{\mathbf{w}_x^T \mathbf{C}_{XX} \mathbf{w}_x} \sqrt{\mathbf{w}_y^T \mathbf{C}_{YY} \mathbf{w}_y}} \tag{4.1}$$

where $\mathbf{C}_{XX}$,$\mathbf{C}_{YY}$ and $\mathbf{C}_{XY}$ are the covariance matrices respectively. The algorithm for obtaining these transformations is summarized in the following steps

Algorithm 3 CCA-Projection

```
**INPUT:** $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n] \in \mathbb{R}^{d1}$ and $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_n] \in \mathbb{R}^{d2}$ with
$\mathbb{E}[X] = \mathbb{E}[Y] = 0$ , dimension $m \leq min(d_1, d_2)$
**OUTPUT :** CCA projection $\mathbf{w}_x \in \mathbb{R}^{d_1 \times m}$ and $\mathbf{w}_y \in \mathbb{R}^{d_2 \times m}$
Calculate covariance matrices $\mathbf{C}_{XX} = \sum_i \mathbf{x}_i \mathbf{x}_i^T = \frac{1}{n}\mathbf{x}\mathbf{x}^T$ $\mathbf{C}_{YY} = \sum_i \mathbf{y}_i \mathbf{y}_i^T = \frac{1}{n}\mathbf{y}\mathbf{y}^T$ and $\mathbf{C}_{XY} = \sum_i \mathbf{x}_i \mathbf{y}_i^T = \frac{1}{n}\mathbf{x}\mathbf{y}^T$
$\Omega \leftarrow \mathbf{C}_{XX}^{-\frac{1}{2}} \mathbf{C}_{XY} C_{YY}^{-\frac{1}{2}}$
$[U \Sigma V^T] = SVD(\Omega)$
**Return** $\mathbf{w}_x = \mathbf{C}_{XX}^{-1/2}[u_1, .., u_m]$ and $\mathbf{w}_y = \mathbf{C}_{YY}^{-1/2}[v_1, .., v_m]$
```

Cross-modal retrieval aims to flexibly retrieve objects across unfamiliar heterogeneous modalities. When two modalities have a natural correspondence, the cross-modal retrieval reduces to a classical retrieval problem. CCA aims to bridge the gap by maximising the pairwise correlations between two sets of heterogenous data. Under this approach, CCA learns two linear projections $\mathbf{w}_x$ and $\mathbf{w}_y$

$$\mathbf{w}_x : \mathbf{X} \rightarrow Z^X$$

and

$$\mathbf{w}_y : \mathbf{Y} \rightarrow Z^Y$$

to map $\mathbf{X}$ and $\mathbf{Y}$ onto Z^X and Z^Y respectively. The resulting intermediate subspace is a compact, efficient representation of both modalities that possess a natural correspondence. During cross-modal retrieval, one can project a text query $T_q \in \mathbb{R}^T$ or image query $I_q \in \mathbb{R}^I$ with a projection. With an appropriate choice of distance function $d(Z^X, Z^Y)$ one can now flexibly retrieve and match by simple nearest neighbour calculation.

4.3 From supervised pairing to unsupervised pairing

The CCA algorithm is predicated on the assumption of paired training data. However, many real-world scenarios of high economic impact arise where no correspondence or alignment is provided [Haghighi et al., 2008, Kim et al., 2017]. To adapt CCA in an unsupervised setting, [Haghighi et al., 2008] proposed Matching CCA. The goal of Matching CCA is twofold (i)To compute the shared subspace as per regular CCA (ii)To compute the unknown correspondence between instances in the two views, which is assumed given in regular CCA. Since these two quantities are independent, Matching CCA resorts to coordinate-descent alternating optimisation stratergy. It iterates between solving for the best correspondence, assuming a given subspace using the Munkres algorithm and finding the best subspace, assuming a given correspondence, using vanilla CCA as a subroutine.

Similar to CCA, Matching CCA assumes a dataset $\mathcal{D}$ containing two sets of unpaired vectors from two sources $\mathcal{D} = (\{\boldsymbol{x}_i\}_{i=1}^n, \{\boldsymbol{y}_j\}_{j=1}^n)$ where $\boldsymbol{x} \in \mathbb{R}^{d_1}$ $\boldsymbol{y} \in \mathbb{R}^{d_2}$. The mapping is based on the assumption that the correct matching is the one that best captures the correlation between the two sets. Formally we can state the problem as

$$\max_{\mathbf{w_x}, \mathbf{w_y}, \pi} \operatorname{corr}(\mathbf{X}\mathbf{w}_x, \mathbf{Y}\pi\mathbf{w}_y)$$

where w_x and w_y are linear projections similar to supervised CCA and π is a permutation function over $\{1 \cdots n\}$. Algorithm 4 describes the procedure in brief.

Algorithm 4 Matching CCA

```
Repeat
    //E-step
    $\pi \leftarrow \mathbf{MUNKRES}_{\pi \in \Pi}(\mathbf{w}_x\mathbf{X}, \mathbf{w}_y\mathbf{Y}\pi)$
    //M-step
    $\mathbf{w}_x, \mathbf{w}_y = CCA(\mathbf{X}, \mathbf{Y}\pi)$
until convergence
Return $\mathbf{w}_x, \mathbf{w}_y, \pi$
```

4.4 Cross Domain Object Matching

The task of cross domain object matching (CDOM) is to determine *correspondence* between sets of objects such as mapping of cross lingual word embeddings, point clouds etc. CDOM is formulated as finding a correspondence between pair of objects between different domains. The goal of CDOM can be written as follows : Given two sets of samples $\{x_i\}_{1=1}^n$ and $\{y_i\}_{i=1}^n$, find a mapping that matches them well. Thats is, we would like to find a correspondence function $\pi \in \Pi$

$$\Pi \in \{0,1\}^{n \times n}, \Pi \mathbf{1}_n = \mathbf{1}_n, \Pi^T \mathbf{1}_n = \mathbf{1}_n$$

where $\mathbf{1}_n$ is the n dimensional vector of all ones. We seek to find a permutation

$$Z(\Pi) = \{(x_i, y_{\pi(i)})\}$$

The optimal permutation matrix Π^* is obtained by maximizing the dependence criterion between the two sets of objects $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$:

$$\Pi^* = \arg\max_{\Pi} D(Z(\Pi))$$

In this thesis we are concerned with D which maximizes the dependency between the two sets of variables. In particular, we will look at Kernel Target Alignment (KTA) [Cristianini et al., 2002] and Squared Mutual Information (SMI) [Yamada et al., 2015] which we cover in the following section.

4.5 Dependence Estimation

In Section 4.3, we briefly highlighted how CCA can be adapted towards learning when data is unpaired. While Matching CCA shows a promising direction, it is highly limited as it works with only linear dependence. In this thesis we explore more general statistical dependency measure statistical dependency approaches to matching. To extend CCA with non-linear dependence, non-linear extensions have been proposed. Initially based on neural networks [Hsieh, 2000], using kernel methods [Bach and Jordan, 2001] has become a promising approach for extracting complex non-linear relationships. In this thesis, we explore the use of two alternatives , the unnormalized kernel-target alignment [Cristianini et al., 2002] and squared-loss mutual information (SMI) [Yamada et al., 2015].

Unnormalized Kernel Target Alignment (uKTA) measures the similarity between two kernel functions. The similarity function is given by

$$\mathrm{uKTA}(\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n) = \mathrm{tr}\left(\boldsymbol{K}\boldsymbol{L}\right),$$

where $tr(.)$ is the trace operator, $\boldsymbol{K}$ and $\boldsymbol{L}$ are the Gram matrices for $\boldsymbol{x}$ and $\boldsymbol{y}$ respectively. This similarity function takes large value if the Gram matrices $\boldsymbol{K}$ and $\boldsymbol{L}$ are similar, and a small value if they are not similar. Note that, in the original KTA, we have the normalization term. However, this makes the optimization hard, and thus we employ the unnormalized variant of KTA. Moreover, uKTA can be regarded as a non-centered variant of HSIC [Gretton et al., 2005].

Kernelized Sorting with Kernel Target Alignment Kernelized sorting [Quadrianto et al., 2009] refers to the problem of finding correspondence across two different domains by maximizing the dependency measure. Specifically we look at unnormalized kernel target alignment [Cristianini et al., 2002]. KS-uKTA is formulated as

$$\max_{\Pi} \mathrm{uKTA}(Z(\Pi)) \tag{4.2}$$

where

$$\mathrm{uKTA}(Z(\Pi)) = \mathrm{tr}\left(\boldsymbol{K}\Pi^T\boldsymbol{L}\Pi\right) \tag{4.3}$$

The solution of Eq 4.3 requires solving a *quadratic assignment problem* which is known to be NP-hard. Existing quadratic solvers are not practical as they have multiple tuning parameters.

An alternative to solve Eq 4.3 is based on *linear assignment problem* (LAP). [Quadrianto et al., 2009] proposed to use LAP while solving the KS-HSIC formulation.

Optimization Optimizing Eq 4.3 requires minimizing a lower bound. Since the equation is convex in Π (Lemma 7 of [Quadrianto et al., 2009]), we minimize the lower bound using convex concave procedure (CCCP) [Yuille and Rangarajan, 2002]. The CCCP procedure involves minimizing the difference of two functions $f(x) = g(x) - h(x)$ where g is a convex function and h is a concave function. A lower bound of f is estimated by

$$f(x) \geq g(x_0) + \langle x - x_0, \partial_x g(x_0)\rangle - h(x)$$

For a value $\hat{\Pi}$ we rewrite the function g as

$$g(\hat{\Pi}) = \mathrm{tr}(\boldsymbol{K}\hat{\Pi^T}\boldsymbol{L}\hat{\Pi})$$

Invoking $\partial_A \mathrm{tr}(ABA^TC) = CAB + C^TAB^T$, we know that $\partial_\Pi \mathrm{tr}\boldsymbol{K}\Pi^T\boldsymbol{L}\Pi = \boldsymbol{K}\Pi\boldsymbol{L} + \boldsymbol{K}^T\Pi\boldsymbol{L}^T$

By rearranging the values, we find the lower bound as

$$\begin{aligned} f(\Pi) &\geq \mathrm{tr}(\boldsymbol{K}\hat{\Pi}^T\boldsymbol{L}\hat{\Pi}) + \langle \Pi - \hat{\Pi}, \partial_\Pi \mathrm{tr}\boldsymbol{K}\Pi^T\boldsymbol{L}\Pi\rangle \\ &= \mathrm{tr}(\boldsymbol{K}\hat{\Pi}^T\boldsymbol{L}\hat{\Pi}) + \langle \Pi - \hat{\Pi}, 2\mathrm{tr}(\boldsymbol{K}\hat{\Pi}\boldsymbol{L})\rangle \\ &= \mathrm{tr}(\boldsymbol{K}\hat{\Pi}^T\boldsymbol{L}\hat{\Pi}) + \mathrm{tr}(\boldsymbol{K}\hat{\Pi}(\Pi - \hat{\Pi})^T\boldsymbol{L}) \\ f(\Pi) &\geq \mathrm{tr}(\boldsymbol{K}\Pi^T\boldsymbol{L}\hat{\Pi}) \end{aligned}$$

To update the permutation matrices, a line search method is adopted which is used to yield successive permutation matrices [Quadrianto et al., 2009]

$$\Pi^{new} = (1-\eta)\Pi^{old} + \eta \arg\max_{\Pi} \operatorname{tr}(\boldsymbol{K}\Pi\boldsymbol{L}\Pi^{old}) \tag{4.4}$$

where η is the step size. The second term in Eq 4.4 is the well known linear assignment problem (LAP) which can be solved by the Hungarian algorithm [Kuhn and Yaw, 1955].

The success of the iterative procedure to obtain an optimal solution is dependent on the choice of initial conditions. [Quadrianto et al., 2009] proposed to sort the elements of the kernel matrices $\boldsymbol{K}$ and $\boldsymbol{L}$ i.e matching can be achieved by sorting the elements x and y. However in practise the kernels need to be of rank 1 which is difficult to achieve. To alleviate this situation, [Quadrianto et al., 2009] suggested to use the principal eigen vectors to match the initial kernel matrices.

The uKTA offers the advantage of being distribution-free but is sensitive to the choice of the kernel. [Yamada and Sugiyama, 2011] suggests using a Gaussian kernel with the width set to the median distance between the samples.

Now that we introduced the u-KTA based cross-modal matching, we look at kernelized sorting based on SMI.

Squared Mutual Information Mutual information (MI) represents the statistical independence between two random variables [Cover and Thomas, 2006, Shannon, 2001] is used in a plethora of machine learning applications and has recently found its way in deep learning [Zhao et al., 2017, Belghazi et al., 2018]. The mutual information between two random variables $\mathbf{X}$ and $\mathbf{Y}$

$$MI(\mathbf{X}, \mathbf{Y}) = \iint p(\boldsymbol{x}, \boldsymbol{y}) \log \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y},$$

where $p(\boldsymbol{x}, \boldsymbol{y})$ is the joint probability distribution of $\mathbf{X}$ and $\mathbf{Y}$, and $p(\boldsymbol{x})$ and $p(\boldsymbol{y})$ are the marginal probabilities of $\mathbf{X}$ and $\mathbf{Y}$ respectively.

Estimation of mutual information from data challenges has been proven to be notoriously hard. Nonparametric density estimation tools like kernel density estimation (KDE) [Fraser and Swinney, 1986] or histogram based approaches have been applied [Darbellay and Vajda, 1999], however these methods strongly are influenced by the curse of dimensionality and could be unreliable in higher dimensions.

Approximation of MI via estimation of the density ratio $\frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}$ has recently been proposed [Suzuki et al., 2008] which is based on the KL-divergence approximation via direct density-ratio estimation [Sugiyama et al., 2008, Nguyen et al., 2008, Sugiyama et al., 2012]. An advantage of this method is that it does not involve estimating of the joint distribution $p(x, y)$ or the marginals $p(x)$ and $p(y)$. However the presence of a log-term is rather computationally expensive.

To bypass these problems, a variant of MI called Squared mutual information (SMI) has been proposed. SMI between two random variables is defined as [Suzuki and Sugiyama, 2010]

$$\mathrm{SMI} = \iint \left(\frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} - 1 \right)^2 p(\boldsymbol{x})p(\boldsymbol{y})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y},$$

which is the Pearson divergence [Pearson, 1900] from $p(\boldsymbol{x},\boldsymbol{y})$ to $p(\boldsymbol{x})p(\boldsymbol{y})$. The SMI is an f-divergence [Ali and Silvey, 1966] i.e it is a non-negative measure and is zero only if the random variables are independent. The SMI is more attractive to use than MI as (i) It can be shown to hold optimal non-parametric convergence rates [Sugiyama et al., 2012] (ii) the SMI can be estimated by solving a set of linear equations (iii) The SMI estimator is also known to be robust against outliers [Sugiyama et al., 2012].
Estimation of SMI To estimate SMI, a direct density ratio estimation approach is useful. The key idea is to approximate the true density ratio

$$r(\boldsymbol{x},\boldsymbol{y}) = \frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}$$

without estimating the densities $p(\boldsymbol{x},\boldsymbol{y})$, $p(\boldsymbol{x})$ and $p(\boldsymbol{y})$.

In LSMI, the ratio $r(\boldsymbol{x},\boldsymbol{y})$ is directly modeled by the linear model:

$$r(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\alpha}) = \sum_{\ell=1}^{n} \alpha_\ell K(\boldsymbol{x}_\ell,\boldsymbol{x})L(\boldsymbol{y}_\ell,\boldsymbol{y}),$$

where $\boldsymbol{\alpha} = [\alpha_1,\dots,\alpha_n]^\top \in \mathbb{R}^n$ is the model parameter, n is the number of basis functions.

Then, the model parameter is given by minimizing the error between true density-ratio and its model:

$$J(\boldsymbol{\alpha}) = \iint \left(\frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} - r(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\alpha}) \right)^2 p(\boldsymbol{x})p(\boldsymbol{y})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}.$$

By approximating the loss function by samples, the parameter $\boldsymbol{\alpha}$ is learned by solving the following optimization problem [Suzuki and Sugiyama, 2010]

$$\min_{\boldsymbol{\alpha}} \left[\frac{1}{2}\boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}}\boldsymbol{\alpha} - \boldsymbol{\alpha}^\top \widehat{\boldsymbol{h}} + \frac{\lambda}{2}||\boldsymbol{\alpha}||_2 \right] \tag{4.5}$$

where

$$\widehat{\boldsymbol{H}} = \frac{1}{n^2}(\boldsymbol{K}\boldsymbol{K}^\top) \circ (\boldsymbol{L}\boldsymbol{L}^\top), \quad \widehat{\boldsymbol{h}} = \frac{1}{n}(\boldsymbol{K} \circ \boldsymbol{L})\boldsymbol{1}_n,$$

$\lambda \geq 0$ is a regularization parameter and $\circ$ is the elementwise product and $\boldsymbol{1}_n$ is the

n-dimensional vector whose element are all ones. Differentiating Eq 4.5 with respect to $\boldsymbol{\alpha}$ and equating it to zero, we obtain an optimal solution

$$\widehat{\boldsymbol{\alpha}} = \left(\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}_n\right)^{-1} \widehat{\boldsymbol{h}},$$

where $\boldsymbol{I}_n$ is the $n \times n$ dimensional identity matrix. Then, the estimator of SMI can be given by [Yamada et al., 2015, Yamada and Sugiyama, 2011]

$$\widehat{\mathrm{SMI}}(\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n) = \frac{1}{2n} \mathrm{tr}\left(\mathrm{diag}\left(\widehat{\boldsymbol{\alpha}}\right) \boldsymbol{K}\boldsymbol{L}\right) - \frac{1}{2}, \tag{4.6}$$

where $\mathrm{diag}\left(\boldsymbol{\alpha}\right) \in \mathbb{R}^{n \times n}$ is the diagonal matrix whose diagonal elements are $\boldsymbol{\alpha}$. We can see that uKTA is a special case of SMI. Specifically, if we set $\widehat{\boldsymbol{\alpha}} = \mathbf{1}_n$, SMI boils down to uKTA.

Kernelized sorting with SMI The objective of kernelized sorting is to find a mapping between two sets of samples $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{y}_i\}_{i=1}^n$ so that they can be matched. Let π be the permutation function over $\{1, .., n\}$ and let Π be the permutation matrix indicator matrix i.e

$$\Pi \in \{0, 1\}^{n \times n} \quad \Pi \mathbf{1}_n = \mathbf{1}_n \quad \text{and} \quad \Pi \mathbf{1}_n^T = \mathbf{1}_n$$

The optimal permutation is obtained by maximizing the dependency measure SMI between the two sets $\boldsymbol{X}$ and $\boldsymbol{Y}\Pi$ given by

$$\widehat{\mathrm{SMI}}(\{(\boldsymbol{X}, \boldsymbol{Y}\Pi)\}) = \frac{1}{2n} \mathrm{tr}\left(\mathrm{diag}\left(\boldsymbol{\alpha}_\Pi\right) \boldsymbol{K}\Pi^T \boldsymbol{L}\Pi\right) - \frac{1}{2}, \tag{4.7}$$

and the permutation is given by

$$\Pi^* = \arg\max_{\Pi} \mathrm{SMI}(\boldsymbol{X}, \boldsymbol{Y}\Pi) \tag{4.8}$$

We summarize the steps in Algorithm 5

Algorithm 5 Algorithm for optimizing Π

```
1: Initialize $\Pi$ using eigenvalue based initialization
2:     //Dependence Estimation i.e obtain an SMI estimator given $\Pi$
3:     $\widehat{\mathrm{SMI}}(\{(\boldsymbol{x}_i, \boldsymbol{y}_{\pi(i)})\}_{i=1}^n) = \frac{1}{2n}\mathrm{tr}\left(\mathrm{diag}\left(\widehat{\boldsymbol{\alpha}_\pi}\right) \boldsymbol{K}\Pi^T \boldsymbol{L}\Pi\right) - \frac{1}{2}$
4:     //Dependence Maximization i.e Obtain a permutation matrix alignment $\Pi$ given $\widehat{\mathrm{SMI}}$
5:     $\Pi^* = \arg\max_\Pi \widehat{\mathrm{SMI}}(\boldsymbol{X}, \boldsymbol{Y}\Pi)$
6: Alternate between Step 3 and Step 5
7: Return $\Pi$
```

Similar to kernelized sorting with uKTA, we adopt a line search procedure to update the permutation matrix as

$$\Pi^{new} = (1-\eta)\Pi^{old} + \eta \arg\max_{\Pi} \mathrm{tr}(\Pi^T \boldsymbol{L}\Pi^{old}\alpha_{\Pi}\boldsymbol{K}) \tag{4.9}$$

The second term in Eq 4.9 is a linear assignment problem which can be solved using the Hungarian method.

4.6 Optimal Transport

In Sec 4.5, we discussed KTA and SMI which uses the Hungarian algorithm [Kuhn and Yaw, 1955] to search for the best cross-modal pairing. In this section, we introduce an alternative in the form of Sinkhorn algorithm, which is underpinned by the notion of Optimal Transport.

Optimal Transport (OT) plays a natural role spanning across multiple problems in machine learning problems in realizing correspondence between sets which could exist between words or across objects in different images. OT poses the problem of finding a correspondence between two probability masses by elegantly formulating as finding the transportation matrix which minimizes distance between the probability masses.

Consider two sets of embeddings, $X = \{\boldsymbol{x}_i\}_{i=1}^n$ and $Y = \{\boldsymbol{y}_j\}_{j=1}^m$ where $\boldsymbol{x}_i \in \mathbb{R}^{d_x}$ and $\boldsymbol{y}_j \in \mathbb{R}^{d_y}$ are the source and target respectively. Specifically we assume two empirical distributions over these embeddings:

$$\mu = \sum_{i=1}^{n} \boldsymbol{p}_i \delta_{\boldsymbol{x}_i} \quad \nu = \sum_{j=1}^{m} \boldsymbol{q}_j \delta_{\boldsymbol{y}_j} \tag{4.10}$$

where $\boldsymbol{p}$ and $\boldsymbol{q}$ are probability weight vectors for each point usually set to uniform i.e $\mathbf{p}_i = 1/n$ and $\mathbf{q}_j = 1/m.$ and $\delta_{\boldsymbol{x}_i}$ and $\delta_{\boldsymbol{y}_j}$ is the dirac at point x_i and y_j, inuitively representing a unit of mass concentred at the locations. We seek to find a transportation map T realizing

$$\inf_T \left\{ \int_{\mathcal{X}} c(\mathbf{x}, T(\mathbf{x})) d\mu(\mathbf{x}) | T\#\mu = \nu \right\} \tag{4.11}$$

where the cost matrix $c(\mathbf{x}, T(\mathbf{x}))$ contains the cost of transport $\mathbf{x}$ and $T\#\mu = \nu$ usually called *push forward operator* [Peyré and Cuturi, 2019] maps the source points to the target. While the existance of such a map maybe non-existant, a commonly used practise is to relax to Kantorovich's formulation [Peyré and Cuturi, 2019, Alvarez-Melis and Jaakkola, 2018]. Kantorovich's formulation seeks to minimize the set of transportation plans which is a polytope:

$$\Pi(p,q) = \{\pi \in \mathbb{R}^{n\times m} | \pi \mathbf{1}_n = \mathbf{p}, \pi^T \mathbf{1}_m = \mathbf{q}\} \tag{4.12}$$

The set of all cost matrices is denoted by $\mathbf{C} \in \mathbb{R}^{n\times m}$ i.e $C_{ij} = ||\boldsymbol{x}_i - \boldsymbol{y}_j||^2$. The total cost incurred by π is $\langle \pi, \mathbf{C} \rangle = \sum_{ij} \pi_{ij} C_{ij}$. Thus the discrete optimal transport consists of finding a plan π that solves

$$\min_{\pi \in \Pi(p,q)} \langle \pi, \boldsymbol{C} \rangle \tag{4.13}$$

Eq 4.13 is a linear problem that can be solved by interior point methods in cubic time complexity [Peyré and Cuturi, 2019, Alvarez-Melis and Jaakkola, 2018]. More recently [Cuturi, 2013] proposed adding a entropic regularization term which yields an efficient optimization and often better empirical results.

$$\min_{\pi \in \Pi(p,q)} \langle \pi, \boldsymbol{C} \rangle - \lambda H(\pi) \tag{4.14}$$

The solution of Eq 4.14 has the form $\pi^* = \operatorname{diag}(a)\,\mathbf{K}\,\operatorname{diag}(b)$ where $\mathbf{K}$ called the Gibbs kernel is associated to the cost matrix $\mathbf{C}$ with $\mathbf{K} = \exp^{-\frac{\mathbf{C}}{\lambda}}$ and can be obtained efficiently via the Sinkhorn-Knopp procedure, a matrix scaling procedure [Nemirovski and Rothblum, 1999] which iteratively calculates:

$$\mathbf{a} \leftarrow \mathbf{p} \oslash \mathbf{K}\mathbf{b} \quad \text{and} \quad \mathbf{b} \leftarrow \mathbf{q} \oslash \mathbf{K}^T\mathbf{a} \tag{4.15}$$

where $\oslash$ denotes entry-wise division. Algorithm 6 summaries the steps for obtaining a transportation matrix.

Algorithm 6 Sinkhorn iterations to learn a transportation matrix

```
Input: Unpaired Data $\{\boldsymbol{X}_i\}, \{\boldsymbol{Y}_i\}$. Params: $\lambda$, probability vectors $\mathbf{p}$ and $\mathbf{q}$
//Compute cost matrix $C_{ij} = ||x_i - y_j||^2$
    $\mathbf{a} \leftarrow \mathbf{1} \quad \mathbf{K} \leftarrow \exp\{-\mathbf{C}/\lambda\}$
while not converged
//Sinkhorn iterations of Eq 4.15
    $\mathbf{a} \leftarrow \mathbf{p} \oslash \mathbf{K}\mathbf{b}, \mathbf{b} \leftarrow \mathbf{q} \oslash \mathbf{K}^T\mathbf{a}$
    $\boldsymbol{\pi} \leftarrow \operatorname{diag}(a)\,\mathbf{K}\,\operatorname{diag}(b)$
end while
Output: Transportation Matrix $\boldsymbol{\pi}$.
```

4.7 Conclusions

In this chapter we introduced prior work on supervised and unsupervised pairing. We first introduced the classic CCA algorithm which assumes paired training data. We then introduced matching CCA, which extends CCA from supervised to unsupervised pairing. Finally we introduced SMI and KTA, which provides the objective for more advanced kernelised sorting style unsupervised pairing. We also introduce the Sinkhorn Algoritm which will be later used in Chapter 6 to learn a permutation matrix as an alternative to the standard Hungarian algorithm. However all this work assumes that the input data representation X and Y are given and fixed. In the following chapters we investigate joint representation learning (of X and Y) and unsupervised pairing using the statistical dependency measures introduced in this chapte and the non-deep version of Sinkhorn algorithm.

Chapter 5

Unsupervised Word Translation without Adversaries

5.1 Introduction

Translating words between languages, or more generally inferring bilingual dictionaries, is a long-studied research direction with applications including machine translation [Lample et al., 2017], multilingual word embeddings [Klementiev et al., 2012], and knowledge transfer to low resource languages [Guo et al., 2016]. Research here has a long history under the guise of decipherment [Knight et al., 2006]. Current contemporary methods have achieve effective word translation through theme-aligned corpora [Gouws et al., 2015], or seed dictionaries [Mikolov et al., 2013a].

[Mikolov et al., 2013a] showed that monolingual word embeddings exhibit isomorphism across languages, and can be aligned with a simple linear transformation. Given two sets word vectors learned independently from monolingual corpora, and a dictionary of seed pairs to learn a linear transformation for alignment; they were able to estimate a complete bilingual lexicon. Many studies have since followed this approach, proposing various improvements such as orthogonal mappings [Artetxe et al., 2016] and improved objectives [Lazaridou et al., 2015b].

Obtaining aligned corpora or bilingual seed dictionaries is nevertheless not straightforward for all language pairs. This has motivated a wave of very recent research into *unsupervised* word translation: inducing bilingual dictionaries given only monolingual word embeddings [Conneau et al., 2017b, Zhang et al., 2017b,a, Artetxe et al., 2017]. The most successful have leveraged ideas from Generative Adversarial Networks (GANs) [Goodfellow et al., 2014]. In this approach the generator provides the cross-modal mapping, taking embeddings of dictionary words in one language and 'generating' their translation in another. The discriminator tries to distinguish between this 'fake' set of translations and the true dictionary of embeddings in the target language. The two play a competitive game, and if the generator learns to fool the discriminator, then its cross-modal mapping should be capable of inducing a complete dictionary, as per [Mikolov et al., 2013a].

Despite these successes, such adversarial methods have a number of well-known drawbacks [Arjovsky et al., 2017]: Due to the nature of their min-max game, adversarial training is very unstable, and they are prone to divergence. It is extremely hyper-parameter sensitive, requiring problem-specific tuning. Convergence is also hard to diagnose and does not correspond well to efficacy of the generator in downstream tasks [Hoshen and Wolf, 2018c].

In this chapter, we propose an alternative statistical dependency-based approach to unsupervised word translation. Specifically, we propose to search for the cross-lingual word pairing that maximizes statistical dependency in terms of squared loss mutual information (SMI) [Yamada et al., 2015, Suzuki and Sugiyama, 2010]. Compared to prior statistical dependency-based approaches such as Kernelized Sorting (KS) [Quadrianto et al., 2009] we advance: (i) through use of SMI rather than their Hilbert Schmidt Independence Criterion (HSIC) and (ii) through jointly optimising cross-modal pairing with representation learning within each view. In contrast to prior work that uses a fixed representation, by non-linearly projecting monolingual world vectors before matching, we learn a new embedding where statistical dependency is easier to establish. Our method: (i) achieves similar unsupervised translation performance to recent adversarial methods, while being significantly easier to train and (ii) clearly outperforms prior non-adversarial methods.

5.2 Proposed model

5.2.1 Deep Distribution Matching

Let dataset $\mathcal{D}$ contain two sets of unpaired monolingual word embeddings from two languages $\mathcal{D} = (\{\boldsymbol{x}_i\}_{i=1}^n, \{\boldsymbol{y}_j\}_{j=1}^n)$ where $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$. Let π be a permutation function over $\{1, 2, \dots, n\}$, and $\boldsymbol{\Pi}$ the corresponding permutation indicator matrix: $\boldsymbol{\Pi} \in \{0,1\}^{n \times n}, \boldsymbol{\Pi}\boldsymbol{1}_n = \boldsymbol{1}_n,$ and $\boldsymbol{\Pi}^\top \boldsymbol{1}_n = \boldsymbol{1}_n$. where $\boldsymbol{1}_n$ is the n-dimensional vector with all ones. We aim to optimize for both the permutation $\boldsymbol{\Pi}$ (bilingual dictionary), and non-linear transformations $\boldsymbol{g_x}(\cdot)$ and $\boldsymbol{g_y}(\cdot)$ of the respective wordvectors, that maximize statistical dependency between the views. While regularising by requiring the original word embedding information is preserved through reconstruction using decoders $\boldsymbol{f_x}(\cdot)$ and $\boldsymbol{f_y}(\cdot)$. Our overall loss function is:

$$
\begin{aligned}
&\underset{\boldsymbol{\Theta}_x,\boldsymbol{\Theta}_y,\boldsymbol{\Pi}}{min} \underbrace{\Omega(\mathcal{D};\boldsymbol{\Theta}_{\boldsymbol{x}},\boldsymbol{\Theta}_{\boldsymbol{y}})}_{Regularizer} - \underbrace{\lambda D_{\boldsymbol{\Pi}}(\mathcal{D};\boldsymbol{\Theta}_{\boldsymbol{x}},\boldsymbol{\Theta}_{\boldsymbol{y}})}_{Dependency},\\
&D_{\boldsymbol{\Pi}}(\mathcal{D};\boldsymbol{\Theta}_{\boldsymbol{x}},\boldsymbol{\Theta}_{\boldsymbol{y}}) = D_{\boldsymbol{\Pi}}(\{\boldsymbol{g}_{\boldsymbol{x}}(\boldsymbol{x}_i),\boldsymbol{g}_{\boldsymbol{y}}(\boldsymbol{y}_{\pi(i)})\}_{i=1}^{n}),\\
&\Omega(\mathcal{D};\boldsymbol{\Theta}_{\boldsymbol{x}},\boldsymbol{\Theta}_{\boldsymbol{y}}) = \sum_{i=1}^{n} \boldsymbol{x}_i - \boldsymbol{f}_{\boldsymbol{x}}(\boldsymbol{g}_{\boldsymbol{x}}(\boldsymbol{x}_i))_2^2\\
&\qquad\qquad\qquad\qquad + \boldsymbol{y}_i - \boldsymbol{f}_{\boldsymbol{y}}(\boldsymbol{g}_{\boldsymbol{y}}(\boldsymbol{y}_i))_2^2\\
&\qquad\qquad\qquad\qquad + R(\boldsymbol{\Theta}_{\boldsymbol{x}}) + R(\boldsymbol{\Theta}_{\boldsymbol{y}}).
\end{aligned}
\tag{5.1}
$$

where $\boldsymbol{\Theta}$s parameterize the encoding and reconstruction transformations, $R(\cdot)$ is a regularizer (e.g., ℓ_2-norm and ℓ_1-norm), and $D_{\boldsymbol{\Pi}}(\cdot,\cdot)$ is a statistical dependency measure. Crucially compared to prior methods such as matching CCA [Haghighi et al., 2008], dependency measures such as SMI do not need comparable representations to get started, making the bootstrapping problem less severe.

5.2.2 Dependence Estimation

Dependence estimation is a fundermental property to study the relation between two random variables in staitstics. Familiar examples of dependent phenomenon are correlation between the height of parents and offsprings, correlation between the price of goods and supply of product.

As discussed in Chapter 4, mutual information is a popular mechanism to study the independence of two random variables. Our focus in this thesis and specifically in this chapter is Squared-Loss Mutual Information (SMI). SMI between two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as [Suzuki and Sugiyama, 2010]:

$$
\mathrm{SMI} = \iint \left(\frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} - 1\right)^2 p(\boldsymbol{x})p(\boldsymbol{y})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y},
$$

which is the Pearson divergence [Pearson, 1900] from $p(\boldsymbol{x},\boldsymbol{y})$ to $p(\boldsymbol{x})p(\boldsymbol{y})$. The SMI is an f-divergence [Ali and Silvey, 1966]. That is, it is a non-negative measure and is zero only if the random variables are independent.

5.2.3 Optimization of parameters

To initialize $\boldsymbol{\Theta}_{\boldsymbol{x}}$ and $\boldsymbol{\Theta}_{\boldsymbol{y}}$, we first independently estimate them using autoencoders. Then we employ an alternative optimization on Eq. 5.1 for $(\boldsymbol{\Theta}_{\boldsymbol{x}},\boldsymbol{\Theta}_{\boldsymbol{y}})$ and $\boldsymbol{\Pi}$ until convergence. We use 3 layer MLP neural networks for both $\boldsymbol{f}$ and $\boldsymbol{g}$. Algorithm 7 summarises the steps.

Algorithm 7 SMI-based unsupervised word translation

```
Input: Unpaired word embeddings $\mathcal{D} = (\{\boldsymbol{x}_i\}_{i=1}^n, \{\boldsymbol{y}_j\}_{j=1}^n)$.
1: Init: weights $\boldsymbol{\Theta}_{\boldsymbol{x}}$, $\boldsymbol{\Theta}_{\boldsymbol{y}}$, permutation matrix $\boldsymbol{\Pi}$.
2: while not converged
3: Update $\boldsymbol{\Theta}_{\boldsymbol{x}}, \boldsymbol{\Theta}_{\boldsymbol{y}}$ given $\boldsymbol{\Pi}$: Backprop Eq (5.2).
4: Update $\boldsymbol{\Pi}$ given $\boldsymbol{\Theta}_{\boldsymbol{x}}, \boldsymbol{\Theta}_{\boldsymbol{y}}$: LSOM Eq (5.3).
5: Output: Permutation Matrix $\boldsymbol{\Pi}$. Params $\boldsymbol{\Theta}_{\boldsymbol{x}}$, $\boldsymbol{\Theta}_{\boldsymbol{y}}$.
```

Optimization for $\boldsymbol{\Theta}_{\boldsymbol{x}}$ and $\boldsymbol{\Theta}_{\boldsymbol{y}}$ With fixed permutation matrix $\boldsymbol{\Pi}$ (or π), the objective function

$$\min_{\boldsymbol{\Theta}_{\boldsymbol{x}}, \boldsymbol{\Theta}_{\boldsymbol{y}}} \Omega(\mathcal{D}; \boldsymbol{\Theta}_{\boldsymbol{x}}, \boldsymbol{\Theta}_{\boldsymbol{y}}) - \lambda D_{\boldsymbol{\Pi}}(\mathcal{D}; \boldsymbol{\Theta}_{\boldsymbol{x}}, \boldsymbol{\Theta}_{\boldsymbol{y}}) \tag{5.2}$$

is an autoencoder optimization with regularizer $D_{\boldsymbol{\Pi}}(\cdot)$, and can be solved with backpropagation.

Optimization for $\boldsymbol{\Pi}$ To find the permutation (word matching) $\boldsymbol{\Pi}$ that maximizes SMI given fixed encoding parameters $\boldsymbol{\Theta}_{\boldsymbol{x}}, \boldsymbol{\Theta}_{\boldsymbol{y}}$, we only need to optimize the dependency term $D_{\boldsymbol{\Pi}}$ in Eq. 5.1. We employ the LSOM algorithm [Yamada et al., 2015]. The estimator of SMI for samples $\{\boldsymbol{g}_{\boldsymbol{x}}(\boldsymbol{x}_i), \boldsymbol{g}_{\boldsymbol{y}}(\boldsymbol{y}_{\pi(i)})\}_{i=1}^n$ encoded with $\boldsymbol{g}_x, \boldsymbol{g}_y$ is:

$$\widehat{\mathrm{SMI}} = \frac{1}{2n} \mathrm{tr}\left(\mathrm{diag}\left(\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\Theta},\boldsymbol{\Pi}}\right) \boldsymbol{K}_{\boldsymbol{\Theta}_{\boldsymbol{x}}} \boldsymbol{\Pi}^\top \boldsymbol{L}_{\boldsymbol{\Theta}_{\boldsymbol{y}}} \boldsymbol{\Pi}\right) - \frac{1}{2}.$$

Which leads to the optimization problem:

$$\begin{aligned} \max_{\boldsymbol{\Pi} \in \{0,1\}^{n \times n}} \quad & \mathrm{tr}\left(\mathrm{diag}\left(\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\Theta},\boldsymbol{\Pi}}\right) \boldsymbol{K}_{\boldsymbol{\Theta}_{\boldsymbol{x}}} \boldsymbol{\Pi}^\top \boldsymbol{L}_{\boldsymbol{\Theta}_{\boldsymbol{y}}} \boldsymbol{\Pi}\right) \\ \text{s.t.} \quad & \boldsymbol{\Pi} \boldsymbol{1}_n = \boldsymbol{1}_n, \boldsymbol{\Pi}^\top \boldsymbol{1}_n = \boldsymbol{1}_n. \end{aligned} \tag{5.3}$$

Since the optimization problem is NP-hard, we iteratively solve the relaxed problem [Yamada et al., 2015]:

$$\boldsymbol{\Pi}^{\mathrm{new}} = (1-\eta)\boldsymbol{\Pi}^{\mathrm{old}} + \eta \arg\max_{\boldsymbol{\Pi}} \mathrm{tr}\left(\mathrm{diag}\left(\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\Theta},\boldsymbol{\Pi}}\right) \boldsymbol{K}_{\boldsymbol{\Theta}_{\boldsymbol{x}}} \boldsymbol{\Pi}^\top \boldsymbol{L}_{\boldsymbol{\Theta}_{\boldsymbol{y}}} \boldsymbol{\Pi}^{old}\right)$$

where $0 < \eta \leq 1$ is a step size. The optimization problem is a *linear assignment problem* (LAP). Thus, we can efficiently solve the algorithm by using the *Hungarian method* Kuhn [1955]. To get discrete $\boldsymbol{\Pi}$, we solve the last step by setting $\eta = 1$. Intuitively, this can be seen as searching for the permutation $\boldsymbol{\Pi}$ for which the data in the two (initially unsorted views) have a matching within-view affinity (gram) matrix, where matching is defined by maximum SMI.

5.2.4 Cross-Domain Similarity Local Scaling (CSLS)

Most existing cross-lingual systems view translation as a retrieval of the nearest neighbors from source word embeddings in a shared common embedding space based on cosine similarity. However in higher dimensions, a common problem is encountered known as the *hubness* problem [Radovanovic et al., 2009, 2010] i.e in higher dimensions a phenomenon is observed that an object is the nearest neighbour of multiple objects while other objects dubbed *antihub* are not nearest neighbors to any object. [Lazaridou et al., 2015a] proposed to use a corrected neighbour retrieval method to mitigate hubness. [Smith et al., 2017] propose a similar strategy by inverting the softmax for finding the translation of target words rather than source words [Ruder, 2017]. In this work, we adopt the Cross-domain Similarity Local Scaling (CSLS) [Conneau et al., 2017b]. Given two embeddings $\boldsymbol{x}$ and $\boldsymbol{y}$, CSLS can be computed by

$$CSLS(x,y) = 2cos(\boldsymbol{x},\boldsymbol{y}) - \frac{1}{k}\sum_{\boldsymbol{y}'\in\mathcal{N}_Y(\boldsymbol{x})} cos(\boldsymbol{x},\boldsymbol{y}') - \frac{1}{k}\sum_{\boldsymbol{x}'\in\mathcal{N}_X(\boldsymbol{y})} cos(\boldsymbol{x}',\boldsymbol{y})) \tag{5.4}$$

where $\mathcal{N}_Y(\boldsymbol{x})$ is the set of k nearest neighbors of the point $\boldsymbol{x}$ in the set of target word vectors $\boldsymbol{y}$ and **cos** is the cosine similarity.

Once we obtain the parameters $\boldsymbol{\Theta_x}$, $\boldsymbol{\Theta_y}$ and Π from Algorithm 7 we revise the CSLS score as

$$CSLS(\boldsymbol{\Theta_x x}, \boldsymbol{\Theta_y y}) = 2\cos(\boldsymbol{\Theta_x x}, \boldsymbol{\Theta_y y}) - r^t(\boldsymbol{\Theta_x x}) - r^s(\boldsymbol{\Theta_y y}) \tag{5.5}$$

where r^t is the mean cosine similarity of a target word to its neighbourhood defined as $r^t(\boldsymbol{\Theta_x x}) = \frac{1}{k}\sum_{\boldsymbol{y}'\in\mathcal{N}_Y(\boldsymbol{\Theta_x x})}\cos(\boldsymbol{\Theta_x x}, \boldsymbol{y}')$

5.2.5 Regression based bilingual mapping

We start with our dataset $\mathcal{D} = (\{\boldsymbol{x}_i\}_{i=1}^n, \{\boldsymbol{y}_j\}_{j=1}^n)$ where $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$. A linear regression [Montgomery et al., 2006] model learns a linear mapping $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ between the word vectors that minimizes the discrepancy between mapped word vectors of the source language and word vector language:

$$\min_{\boldsymbol{W}\in\mathbb{R}^{d\times d}} \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{W}\boldsymbol{x}_i, \boldsymbol{y}_i) \tag{5.6}$$

where ℓ is the commonly used square loss i.e $\ell_2(\boldsymbol{x},\boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}||^2$

By plugging this $\boldsymbol{W}$ in the CSLS objective Eq 5.4 where $\cos(\boldsymbol{W}\boldsymbol{x}_i, \boldsymbol{y}_i) = \boldsymbol{x}_i^T\boldsymbol{W}^T\boldsymbol{y}_i$ we can rewrite the new objective as

	MUSE Dataset						BLI Datasets					
Methods	es-en	en-es	it-en	en-it	zh-en	en-zh	es-en	en-es	it-en	en-it	zh-en	en-zh
TM Mikolov et al. [2013a]	5.6	4.8	5.2	4.8	2.6	1.8	3.2	2.9	4.6	4.2	3.2	2.0
CCA Faruqui and Dyer [2014]	6.1	5.6	5.8	5.2	3.1	2.3	5.3	5.0	4.6	4.1	3.2	2.9
MCCA Haghighi et al. [2008]	5.7	5.1	5.4	4.8	3.0	2.2	2.9	2.5	4.2	4.1	2.8	1.9
KS Quadrianto et al. [2009]	8.3	7.4	6.3	5.7	4.8	3.2	9.6	8.9	8.2	7.3	3.7	3.5
Self-Training Artetxe et al. [2017]	12.4	12.2	10.7	10.2	5.8	5.6	15.8	14.5	13.7	12.7	14.8	13.4
EMDOT Zhang et al. [2017b]	72.4	71.8	72.8	72.6	32.8	31.7	29.3	31.2	25.6	28.4	24.2	27.8
W-GAN Zhang et al. [2017b]	78.2	77.4	75.3	74.8	38.6	37.5	23.4	26.7	24.0	25.3	21.2	22.8
GAN-NN Conneau et al. [2017b]	69.8	71.3	72.1	71.5	41.3	40.2	21.4	24.3	22.7	23.2	21.3	21.8
Deep-SMI (Ours)	75.9	80.6	75.7	75.2	38.5	38.1	27.3	28.2	25.7	26.4	22.5	22.3
Deep-SMI-CSLS	79.2	84.5	78.8	78.5	43.7	42.8	28.6	29.3	26.7	28.2	23.2	24.7

Table 5.1: Unsupervised word translation on MUSE and BLI datasets. Precision @ 1 metric. Top group: Conventional methods. Middle group: Adversarial methods. Bottom group: Our methods. Language codes zh=Chinese,en=English,es=Spanish,it=Italian

	MUSE Dataset						BLI Datasets					
Methods	es-en	en-es	it-en	en-it	zh-en	en-zh	es-en	en-es	it-en	en-it	zh-en	en-zh
TM Mikolov et al. [2013a]	32.6	30.1	34.3	33.6	32.4	31.2	28.2	32.1	29.2	32.1	28.5	27.4
CCA Faruqui and Dyer [2014]	27.3	27.1	25.4	24.2	23.1	20.2	25.8	28.3	24.3	25.1	19.2	22.8
MCCA Haghighi et al. [2008]	26.3	25.8	22.7	21.3	24.5	23.8	24.2	26.1	17.6	19.2	18.4	21.6
KS Quadrianto et al. [2009]	34.5	32.6	35.2	33.8	34.3	33.2	27.5	29.1	34.3	32.1	20.0	23.2
Self-Training Artetxe et al. [2017]	35.8	31.4	36.0	34.6	34.3	33.0	27.8	29.8	39.7	33.8	23.6	21.4
EMDOT Zhang et al. [2017b]	78.2	76.3	75.0	74.6	33.2	32.0	30.2	28.4	31.7	30.3	29.3	28.7
W-GAN Zhang et al. [2017b]	81.2	80.5	77.2	75.1	39.0	38.2	28.6	27.9	33.7	29.5	36.7	34.4
GAN-NN Conneau et al. [2017b]	74.8	72.3	74.3	72.5	43.2	42.7	22.8	26.1	27.9	27.1	24.2	23.6
Deep-SMI (Ours)	80.6	75.9	78.2	76.7	45.7	44.6	38.5	37.6	42.3	38.2	29.2	27.4
Deep-SMI-CSLS	84.5	79.2	79.7	78.7	42.3	44.4	28.6	29.3	26.7	28.2	23.2	24.7

Table 5.2: Semi-supervised word translation on MUSE and BLI using 500 seed pair initial dictionary. Precision @ 1 metric. Top group: Conventional methods. Middle group: Adversarial methods. Bottom group: Our methods.

$$CSLS(\boldsymbol{W}\boldsymbol{x}, \boldsymbol{y}) = 2\cos(\boldsymbol{W}\boldsymbol{x}, \boldsymbol{y}) - \frac{1}{k}\sum_{\boldsymbol{y}_j \in \mathcal{N}_Y(\boldsymbol{W}\boldsymbol{x}_i)} \boldsymbol{x}_i \boldsymbol{W}^T \boldsymbol{y}_j - \frac{1}{k}\sum_{\boldsymbol{W}\boldsymbol{x}_j \in \mathcal{N}_X(\boldsymbol{y}_i)} \boldsymbol{x}_j \boldsymbol{W}^T \boldsymbol{y}_i \tag{5.7}$$

5.3 Experiments

In this section, we evaluate the efficacy of our proposed method against various state of the art methods for word translation.

Implementation Details Our autoencoder consists of two layers with dropout and a *tanh* non-linearity. The encoding layers consists of $300 - 250 - 200$. We use polynomial kernel to compute the gram matrices $\boldsymbol{K}$ and $\boldsymbol{L}$. For all pairs of languages, we fix the number of training epochs to 20. All the word vectors are ℓ_2 unit normalized. For CSLS we set the number of neighbors to 10. For optimizing $\boldsymbol{\Pi}$ at each epoch, we set the step size $\eta = 0.75$ and use 20 iterations. For the regularization $R(\boldsymbol{\Theta})$, we use the sum of the Frobenius norms of weight matrices and the regularization parameter λ is 0.75. We train $\boldsymbol{\Theta}$ using full batch gradient-descent, with learning rate 0.05.

Datasets We performed experiments on the publicly available English-Italian, English-Spanish and English-Chinese datasets released by [Dinu and Baroni, 2015, Zhang et al., 2017b, Vulic and Moens, 2013]. We name this collective set of benchmarks BLI. We also conduct further experiments on a much larger recent public benchmark, MUSE [Conneau et al., 2017b][1]. The words have a vocabulary of $200,000$ and dimensions of 300. The smaller BLI dataset maintains $20,000$ words and dimensions of 50.

Setting and Metrics We evaluate all methods in terms of Precision@1, following standard practice. We note that while various methods in the literature were initially presented as fully supervised [Mikolov et al., 2013a], semi-supervised (using a seed dictionary) [Haghighi et al., 2008], or unsupervised [Zhang et al., 2017b], most of them can be straightforwardly adapted to run in any of these settings. Therefore we evaluate all methods both in the unsupervised setting in which we are primarily interested. These methods are transductive in nature as we allow the test words from the bilingual dictionary to be a part of the initial monolingual word embedding. Based on prior work of [Glavaš et al., 2019, Conneau et al., 2017b] dictionaries are created of size $5K$ if it is to be used for training and $1.5K$ to be used for testing. Furthermore smaller dictionaries of 500 words are maintained to test for semi-supervised examples.

Testing To compare our method to other baselines, we need to fix the monolingual embeddings and evaluation dictionary. For that reason, we decided to use the monolingual embedding and evaluation dictionary from MUSE [Conneau et al., 2017b]. Once Alg 7 is run over the embeddings, the returned $\mathbf{\Pi}$ and parameters $\mathbf{\Theta}_{\boldsymbol{x}}$, $\mathbf{\Theta}_{\boldsymbol{y}}$ are used to align the monolingual embeddings. The induced embeddings are evaluated with retrieval methods (standard nearest neighbor and CSLS).

Competitors: Non-Adversarial In terms of competitors that, like us, do not make use of GANs, we evaluate: **Translation Matrix** Mikolov et al. [2013a], which alternates between estimating a linear transformation by least squares and matching by nearest neighbour (NN). **Multilingual Correlation** [Faruqui and Dyer, 2014], and **Matching CCA** [Haghighi et al., 2008], which alternates between matching and estimating a joint linear subspace. **Kernelized Sorting** [Quadrianto et al., 2009], which directly uses HSIC-based statistical dependency to match heterogeneous data points. **Self Training** Artetxe et al. [2017] A recent state of the art method that alternate between estimating an orthonormal transformation, and NN matching.

Competitors: Adversarial In terms of competitors that do make use of adversarial training, we compare: **W-GAN** and **EMDOT** [Zhang et al., 2017b] make use of adversarial learning using Wasserstein GAN and Earth Movers Distance respectively. **GAN-NN** [Conneau et al., 2017b] uses adversarial learning to train an orthogonal

[1] https://github.com/facebookresearch/MUSE/

transformation, along with some refinement steps and an improvement to the conventional NN matching procedure called 'cross-domain similarity local scaling' (CSLS). Since this is a distinct step, we also evaluate our method with CSLS.

We use the provided code for GAN-NN and Self-Train, while re-implementing EDOT/W-GAN to avoid dependency on theano.

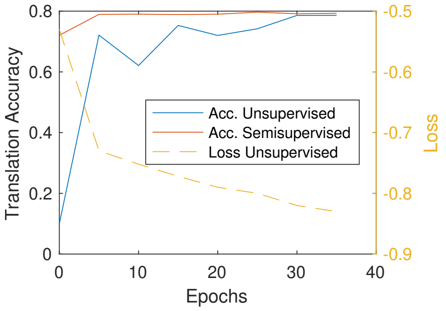


Figure 5.1: Training process of Deep-SMI

5.3.1 Results

Fully Unsupervised Table 5.1 presents comparative results for unsupervised word translation on BLI and MUSE. From these we observe: (i) Our method (bottom) is consistently and significantly better than non-adversarial alternatives (top). (ii) Compared to adversarial alternatives Deep-SMI performs comparably.

All methods generally perform better on the MUSE dataset than BLI. These differences are due to MUSE being a significantly larger dataset than BLI, benefitting methods that can exploit a large amount of training data. In the ground-truth annotation, BLI contains 1-1 translations while MUSE contains more realistic 1-many translations (if any correct translation is picked, a success is counted), making it easier to reach a higher score. We would like to highlight that generally MUSE dictionaries have certain entries which allow for 1-N translations. While training, we enforce our permutation matrix to be 1-1 while during testing we rely on the NN or CSLS for retrieval.

Semi-supervised The first experiment studied fully unsupervised learning. However it is often the case that at least a small set of frequent words will have known translations. This leads to a semi-supervised learning scenario where we wish to learn a complete bilingual dictionary based on a small matched set and a large unmatched set of words.

Results using a 500-word bilingual seed dictionary are presented in Table 5.2.

From these we observe: (i) The conventional methods' performances (top) jump up, showing that they are more competitive if at least some sparse data is available. (ii) Deep-SMI performance also improves, and still outperforms the classic methods significantly overall. (iii) Again, we perform comparably to the GAN methods.

	MUSE Dataset			
Methods	CSLS	Reconstruction	es-en	en-es
KTA	X	-	36.56	29.85
SMI	X	-	37.80	30.73
SMI	✓	-	40.49	32.26
Deep-KTA	X	✓	68.06	61.48
Deep-KTA	✓	✓	71.32	63.70
Deep-SMI	X	X	57.71	49.63
Deep-SMI	✓	X	63.16	52.59
Deep-SMI	X	✓	75.90	80.62
Deep-SMI	✓	✓	79.2	84.5

Table 5.3: Ablation study on MUSE dataset with Shallow and Deep version

5.3.2 Further Experiments

Ablation Study We next perform some ablation studies on the different components of the model (CSLS post processing and auto encoder reconstruction loss). Our experiments on English-Spanish pair on the MUSE dataset are presented in Table 5.3. From the results we can see that: (i) CSLS makes a quite a consistent improvement in performance compared to vanilla NN matching across a variety of settings, (ii) Using the reconstruction loss is very important to make our idea of joint deep presentation learning and pairing perform well. This is understandable, because without this regulariser in Eq 5.1, statistical dependency can be improved for an arbitrary pairing Π by learning degenerate representations such as mapping paired words to matching 1-hot vectors.

Qualitative Analysis Figure 5.1 shows the convergence process of Deep-SMI. From this we see that: (i) Unlike the adversarial methods, our objective (Eq. (5.1)) improves smoothly over time, making convergence much easier to assess. (ii) Unlike the adversarial methods, our accuracy generally mirrors the model's loss. In contrast, the various losses of the adversarial approaches do not well reflect translation accuracy, making model selection or early stopping a challenge in itself. Please compare our Figure 5.1 with Fig 3 in [Zhang et al., 2017b], and Fig 2 in [Conneau et al., 2017b].

There are two steps in our optimization: matching permutation $\mathbf{\Pi}$ and representation weights $\mathbf{\Theta}$. Although this is an alternating optimization, it is analogous to an EM-type algorithm optimizing latent variables ($\mathbf{\Pi}$) and parameters ($\mathbf{\Theta}$). While local minima are a risk, every optimisation step for either variable reduces our objective Eq. (5.1). There is no min-max game, so no risk of divergence as in the case of adversarial GAN-type methods.

Our method can also be understood as providing an unsupervised *Deep*-CCA type model for relating heterogeneous data across two views. This is in contrast to the recently proposed unsupervised shallow CCA [Hoshen and Wolf, 2018c], and conventional supervised Deep-CCA [Chang et al., 2018] that requires paired data for training; and using SMI rather than correlation as the optimisation objective.

5.4 Discussion

In this chapter we studied the problem of unsupervised word translation. The current model is inductive i.e test data is not used for training. But it could be studied in a transductive framework to improve results. Since SSL is generally better than supervised lower bound, this is expected to work.

Our permutation matrix is assumed to be 1-1 during training but this is a generally hard and strong constraint and can be relaxed in future work in order for this method to actually be useful in practice.

5.5 Conclusion

We have presented an effective approach to unsupervised word translation that performs comparably to adversarial approaches while being significantly easier to train and diagnose; as well as outperforming prior non-adversarial approaches.

Chapter 6

Unsupervised Learning in Vision and Language

6.1 Introduction

Learning representations from multi-modal data is a widely relevant problem setting in many applications of machine learning and pattern recognition. In computer vision it arises in tagging [Feng et al., 2014, Gong et al., 2013], cross-view [Gong et al., 2014, Kan et al., 2016] and cross-modal [Ouyang et al., 2016] learning. It is particularly relevant at the border between vision and other modalities, for example audio-visual speech classification [Ngiam et al., 2011] and describing images and videos [Coyne and Sproat, 2001, Guadarrama et al., 2013, Gupta et al., 2012, Krishnamoorthy et al., 2013, Ordonez et al., 2011] in the case of audio and text respectively. The wide applicability of multi-modal representation learning has motivated the study of numerous cross-modal learning methods including Canonical Correlation Analysis (CCA) [Hardoon et al., 2004, Hotelling, 1936b] and Kernel CCA [Bach and Jordan, 2003]. Progress has further accelerated recently with the contribution of large parallel datasets [Lin et al., 2014b, Young et al., 2014], which have permitted the application of deep multi-modal models such as DeepCCA [Andrew et al., 2013] and other two branch deep networks to tasks such as image-caption matching [Wang et al., 2017] and zero-shot learning [Frome et al., 2013b]. Nevertheless a pervasive limitation of all these methods is that they are fully supervised methods in the sense that they require *paired* training data to learn the cross-modal mapping or embedding space. However, in many applications paired data may be relatively sparse compared to unpaired data, in which case semi-supervised cross-modal learning methods would be beneficial to exploit the abundant unpaired data. Moreover, in some cases it may be desirable to learn from pools of data in each modality which are completely unpaired, necessitating unsupervised cross-modal learning.

In this paper we address the task of cross-modal learning from partially or completely *unpaired* data. There have been only a few prior attempts to address inferring pairings from partially or completely unpaired data. These include Kernelized sorting (KS) [Djuric et al., 2012, Jebara, 2004, Quadrianto et al., 2009], least-square object matching (LSOM) [Yamada et al., 2015, Yamada and Sugiyama, 2011], and matching CCA (MCCA) [Haghighi et al., 2008]. However these existing algorithms are all *non-deep* approaches and thus may not perform well on challenging complex data where representation learning is important, such as images and text. We introduce Deep Matching Autoencoders (DMAE), which to our knowledge provides the first deep representation learning approach to unpaired cross-modal learning.

Our DMAE method employs auto-encoders in both data views, which are learned by minimizing reconstruction error as usual. We further introduce a latent alignment matrix to model the unknown pairing between views, which we optimize using cross-modal dependency measures kernel target alignment (KTA) [Cristianini et al., 2002] and squared-loss mutual information (SMI) [Yamada et al., 2015]. With this framework we simultaneously learn the autoencoding representation and the cross-view pairing. In this way the representation is trained to support cross-view matching. During training the learned representation improves as cross-view matching is progressively disambiguated, and cross-modal items are paired more accurately as the learned representation progressively improves.

Our proposed framework elegantly spans the spectrum from fully supervised to fully unsupervised cross-modal learning. The *fully supervised* case corresponds to conventional cross-modal learning, where it is an alternative to DeepCCA [Andrew et al., 2013] or two branch matching nets [Wang et al., 2017], except that we use a statistical dependency-based rather than correlation or ranking-based loss. More interestingly, our approach is effective for *semi-supervised* learning (only subset of pairings available), and we show that it is able to better exploit unlabeled multi-modal data to improve performance compared to alternatives such as matching CCA [Haghighi et al., 2008]. Most interestingly, DMAE is effective for *semi-supervised* cross-modal learning where partial pairings are given. We demonstrate this capability by introducing and solving a novel task termed unsupervised classifier learning (UCL).

In the UCL task we assume a pool of unlabelled images are given along with a pool of category embeddings (e.g., word-vectors) that describe the images in the pool. However it is *unsupervised* in that no pairings between images and categories are given. This task corresponds to an application where we have a pool of images and we have some idea of the classes likely to be represented in those images; but no specific class-image pairings. Based on these inputs alone we can train classifiers to recognise the categories represented in the category embedding pool. Like the classic clustering problem, this task is unsupervised in that there is no supervisory pairing given. However like the conventional supervised learning setting, UCL produces classifiers for specific nameable image categories as an output. This task can be seen as an extreme version of zero-shot learning [Lampert et al., 2009b, Tsai et al., 2017], where there is *no* auxiliary set with image + class embedding pairs available to learn an image-category embedding mapping. The image-category mapping must be learned in an entirely unsupervised

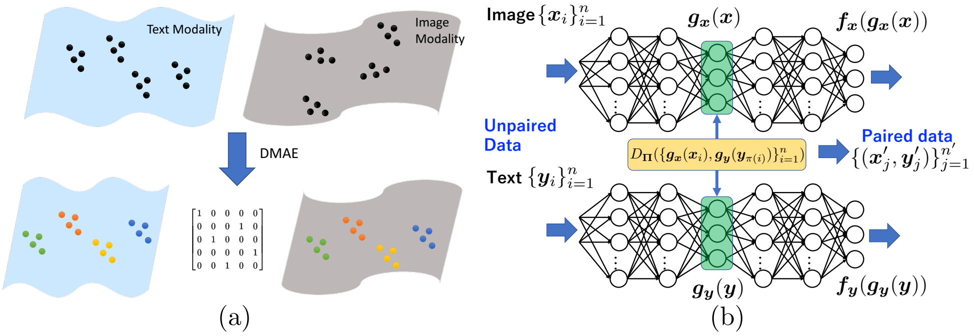

Text Modality
Image Modality
DMAE
(a)
Image $\{\boldsymbol{x}_i\}_{i=1}^n$
$\boldsymbol{g_x}(\boldsymbol{x})$
$\boldsymbol{f_x}(\boldsymbol{g_x}(\boldsymbol{x}))$
Unpaired Data
$D_{\boldsymbol{\Pi}}(\{\boldsymbol{g_x}(\boldsymbol{x}_i), \boldsymbol{g_y}(\boldsymbol{y}_{\pi(i)})\}_{i=1}^n)$
Paired data $\{(\boldsymbol{x}'_j, \boldsymbol{y}'_j)\}_{j=1}^{n'}$
Text $\{\boldsymbol{y}_i\}_{i=1}^n$
$\boldsymbol{g_y}(\boldsymbol{y})$
$\boldsymbol{f_y}(\boldsymbol{g_y}(\boldsymbol{y}))$
(b)

Figure 6.1: (a) Multimodal learning from unpaired data problem setting. DMAE inputs a set of unpaired instances in each view and learns both a permutation matrix associating objects across views and a new representation for each with maximum statistical dependency. (b) Architecture and dataflow schematic of DMAE.

way.

Our contributions are summarized as follows: (i) We propose DMAE, a cross-view learning and matching framework that elegantly spans supervised, semi-supervised and unsupervised cross-modal learning. (ii) We introduce and provide a first solution to the novel problem of unsupervised classifier learning.

6.2 Related Work

Many modern digital events are inherently multimodal in nature, i.e a video or image that you favourite is followed with a caption, a tag or comment. In most supervised multi-modal learning setups, it is a privilege to have access to paired data (i.e., $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$). For example where $\boldsymbol{x}$ is a vector of image and $\boldsymbol{y}$ is a vector of text. In unsupervised multi-modal learning setup, we can only access to unpaired data $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{y}_j\}_{j=1}^n$. The semi-supervised setup is the obvious mixture of the supervised and unsupervised setup.

Supervised multi-modal learning The most established supervised multi-modal learning algorithm is canonical correlation analysis (CCA) [Hotelling, 1936b], which learns a linear projection of features in two views such that are maximally correlated in a common latent space. CCA has been studied extensively and has a number of useful properties Hardoon et al. [2004]. In particular, the optimal linear projection mapping can be obtained by solving an eigenvalue decomposition. It has also been extended to the non-linear case via kernelization (KCCA) [Bach and Jordan, 2003]. The huge success of deep neural network (DNN) in computer vision and NLP has inspired many deep multi-modal learning algorithms including DeepCCA [Andrew et al., 2013], multimodal deep autoencoders (DAEs) [Feng et al., 2014, Ngiam et al., 2011], and two branch matching or ranking networks [Wang et al., 2017]. DeepCCA [Andrew et al., 2013] shares the correlation maximizing objective with classic CCA, but learns a non-linear projection via deep neural networks. It has been shown to outperform linear CCA and its non-linear KCCA extension. In multi-modal DAEs [Ngiam et al., 2011] multi-modal autoencoders are trained with a shared hidden layer. More generally paired data has been used to train two branch DNNs to learn view-invariant embeddings for example via a learning to rank [Frome et al., 2013b, Wang et al., 2017] objective.

In contrast to these Euclidean-based metrics, statistical dependency-based measures, namely Hilbert-schmidt independence criterion (HSIC) [Gretton et al., 2005] have hardly been studied as objectives for multi-modal learning. One example is HSIC-CCA [Chang et al., 2013], which learned a CCA type architecture but with HSIC rather than correlation objective. However, the above supervised algorithms – particularly the deep learning ones – require a large number of *paired* samples to learn an effective cross-modal embedding.

Unsupervised multi-modal learning The desirability of learning from more widely available unpaired data has motivated some research into the harder problem of unsupervised cross-modal learning by introducing latent variables for cross-view pairing. An early approach was Matching CCA [Haghighi et al., 2008]. It alternates between learning a joint embedding with CCA, and solving a bipartite matching problem to associate the unpaired data. Unlike statistical dependency measures, CCA's correlation-based objective requires comparable embeddings to estimate a match. So Matching CCA can never bootstrap itself if initialised with completely random embeddings and no pairing information at all. Indeed it was only shown to work when used with a seed of paired samples for bootstrapping [Haghighi et al., 2008] – i.e., in the semi-supervised setting. Probabilistic latent variable approaches have also been proposed to match across-views [Iwata et al., 2013], however this was only demonstrated to work on toy problems. Both of these are limited to linear projections.

To handle non-linearity in unsupervised multi-modal learning, kernel based approaches were proposed including Kernelized sorting (KS) [Djuric et al., 2012, Jebara, 2004, Quadrianto et al., 2009] and least-squared object matching [Yamada et al., 2015, Yamada and Sugiyama, 2011]. In KS, unpaired data are matched by maximizing HSIC, and it outperforms MCCA on NLP tasks [Jagarlamudi et al., 2010]. In LSOM, squared-loss mutual information (SMI) is used as a dependence measure, and it was shown to outperform the HSIC-based KS. However, both KS and LSOM are non-deep methods, so may not perform well for image and text data where representation learning is beneficial. In this paper we leverage HSIC and SMI-based objectives for learning representations for matching in a deeper context. In early work, [de Sa, 1993] showed that a disagreement cue can also be used to learn from complementary views, however note that despite the title, this method requires *paired* data and so is supervised in our context.

Visual Description with Natural Language Generating or matching natural language descriptions for images and videos has recently become a popular topic in cross-modal learning in the last five years [Ordonez et al., 2011]. A common approach is to learn an image embedding (e.g., CNN), a text representation (e.g., Bag of Words or LSTM [Hochreiter and Schmidhuber, 1997]) and then map these into a common latent space via two-branch deep networks [Klein et al., 2015, Wang et al., 2017, Yan and Mikolajczyk, 2015]. In this latent space, images or videos and associated text descriptions can be matched: supporting annotation or retrieval applications. Our proposed DMAE solves supervised image captioning comparably well to the state of the art methods. But unlike prior approaches it can be generalized to the semi-supervised and unsupervised case for exploiting unpaired data.

Zero-shot learning Our DMAE approach is related to ZSL methods in that it can be applied to learn cross-modal embeddings between images and category vectors, and hence it can also be used as a classifier for novel classes. However it has a few crucial benefits: (i) It can be learned in a semi-supervised way, which encompasses the transductive [Fu et al., 2015a, Tsai et al., 2017] and semi-supervised [Tsai et al., 2017] variants of ZSL. (ii) More interestingly, it can also be learned in an entirely unsupervised way – requiring *no paired samples at all*; unlike all existing ZSL methods. We term this specific problem setting unsupervised classifier learning (UCL).

A recent ZSL method ReViSE [Tsai et al., 2017] is related to ours in that it can also benefit from the semi-supervised learning setting via a MMD-based domain adaptation loss. However ReViSE is engineered specifically for ZSL. In contrast our DMAE is a general cross-modal learner, and can address the completely unsupervised setting unlike ReViSE.

6.3 Deep Matching Autoencoders

We introduce our cross-domain object matching methodology, Deep Matching Autoencoders (Figure 6.1 (a)) from the unsupervised learning perspective where no paired training data is assumed. From here semi-supervised and supervised variants are a straightforward special case. For simplicity we also assume an equal number of samples in each view, but this can be relaxed in practice.

6.3.1 Multi-View Autoencoders

Consider two unpaired sets of samples, $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{y}_i\}_{i=1}^n$, where $\boldsymbol{x} \in \mathbb{R}^{d_x}$ and $\boldsymbol{y} \in \mathbb{R}^{d_y}$. For example, $\boldsymbol{x}$ is a *feature vector* extracted from an image and $\boldsymbol{y}$ is a vector representation of a text. We assume a heterogeneous setup; the dimensionality of $\boldsymbol{x}$ and $\boldsymbol{y}$ are completely different.

Let us denote the autoencoders of $\boldsymbol{x}$ and $\boldsymbol{y}$ as

$$\boldsymbol{f_x}(\boldsymbol{g_x}(\boldsymbol{x}; \boldsymbol{\Theta_x}); \boldsymbol{\Theta_x}), \quad \boldsymbol{f_y}(\boldsymbol{g_y}(\boldsymbol{y}; \boldsymbol{\Theta_y}); \boldsymbol{\Theta_y}),$$

where $\boldsymbol{g}(\cdot)$ and $\boldsymbol{f}(\cdot)$ are encoder and decoder functions, with parameters. $\boldsymbol{\Theta_x}$ and $\boldsymbol{\Theta_y}$. Our motivation is to learn comparable representation embeddings $\boldsymbol{g}_x(\cdot)$ and $\boldsymbol{g}_y(\cdot)$ given no paired training data. This is a significantly harder problem than other multi-modal autoencoder approaches that rely on paired data. [Chandar et al., 2016, Ngiam et al., 2011]

6.3.2 Learning from Unpaired Data

To learn from unpaired data we introduce a permutation matrix to represent the unknown correspondence between data items in two views [Quadrianto et al., 2009, Yamada and Sugiyama, 2011, Yamada et al., 2015]. Let π be an permutation function over $\{1, 2, \ldots, n\}$, and let $\boldsymbol{\Pi}$ be the corresponding permutation indicator matrix:

$$\boldsymbol{\Pi} \in \{0,1\}^{n\times n}, \boldsymbol{\Pi}\boldsymbol{1}_n = \boldsymbol{1}_n, \text{and } \boldsymbol{\Pi}^\top \boldsymbol{1}_n = \boldsymbol{1}_n,$$

where $\boldsymbol{1}_n$ is the n-dimensional vector with all ones. Then, we consider the following optimization problem:

$$\min_{\boldsymbol{\Theta_x},\boldsymbol{\Theta_y},\boldsymbol{\Pi}} \sum_{i=1}^{n} ||\boldsymbol{x}_i - \boldsymbol{f}_x(\boldsymbol{g_x}(\boldsymbol{x}_i))||_2^2 + ||\boldsymbol{y}_i - \boldsymbol{f}_y(\boldsymbol{g_y}(\boldsymbol{y}_i))||_2^2 - \lambda D_\Pi(\{\boldsymbol{g_x}(\boldsymbol{x}_i), \boldsymbol{g_y}(\boldsymbol{y}_{\pi(i)})\}_{i=1}^n) \tag{6.1}$$

where we simultaneously optimise autoencoders ($\boldsymbol{\Theta_x}$ and $\boldsymbol{\Theta_y}$) as well as the cross-domain match ($\boldsymbol{\Pi}$) with tradeoff parameter λ. The key component here is the function $D_{\boldsymbol{\Pi}}(\cdot,\cdot)$ which is a non-negative statistical dependence measure between the x and y views. $D_{\boldsymbol{\Pi}}(\cdot,\cdot)$ needs to be a measure which does not require comparable representations *a priori* in order to enable learning to get started.

6.3.3 Dependence Measures

The statistical dependence measure is the crucial component in achieving our goal. In this paper, we explore two alternatives: the squared-loss mutual information (SMI) introduced in Sec 4.5 [Suzuki and Sugiyama, 2010, Yamada et al., 2015, Yamada and Sugiyama, 2011] and the unnormalized kernel target alignment (KTA) [Cristianini et al., 2002]. Note that SMI is an independence measure. However, since we want to make $\boldsymbol{\Theta_x}$ and $\boldsymbol{\Theta_y}$ generate similar representations, we use SMI as a dependence measure.

$p(\boldsymbol{x}, \boldsymbol{y})$ to $p(\boldsymbol{x})p(\boldsymbol{y})$. The SMI is an f-divergence [Ali and Silvey, 1966] that is it is a non-negative measure and is zero only if the random variables are independent.

To estimate SMI we take a direct density ratio estimation approach [Suzuki and Sugiyama, 2010]. This leads [Yamada et al., 2015, Yamada and Sugiyama, 2011] to the estimator:

$$\widehat{\text{SMI}}(\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n) = \frac{1}{2n}\text{tr}\left(\text{diag}\left(\widehat{\boldsymbol{\alpha}}\right)\boldsymbol{K}\boldsymbol{L}\right) - \frac{1}{2},$$

where tr$(\cdot)$ is the trace operator, $\boldsymbol{K}$ is the Gram matrix for $\boldsymbol{x}$ and $\boldsymbol{L}$ is the Gram matrix for $\boldsymbol{y}$, and $\widehat{\boldsymbol{\alpha}}$ is the model parameter written by [Suzuki and Sugiyama, 2010]

$$\widehat{\boldsymbol{\alpha}} = \left(\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}_n\right)^{-1} \widehat{\boldsymbol{h}}, \quad \widehat{\boldsymbol{H}} = \frac{1}{n^2}(\boldsymbol{K}\boldsymbol{K}^\top) \circ (\boldsymbol{L}\boldsymbol{L}^\top), \quad \widehat{\boldsymbol{h}} = \frac{1}{n}(\boldsymbol{K} \circ \boldsymbol{L})\boldsymbol{1}_n.$$

Here λ is a regularizer, and we use the Gaussian kernel:

$$\boldsymbol{K}_{ij} = \exp\Big(- \frac{||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2}{2\boldsymbol{\sigma}_x^2}\Big), \boldsymbol{L}_{ij} = \exp\Big(- \frac{||\boldsymbol{y}_i - \boldsymbol{y}_j||_2^2}{2\boldsymbol{\sigma}_y^2}\Big)$$

where $\sigma_x > 0$ and $\sigma_y > 0$ are the Gaussian width. Given un-aligned data which depends on a permutation matrix $\boldsymbol{\Pi}$ with respect to $\boldsymbol{y}$, we can write SMI as [Yamada et al., 2015, Yamada and Sugiyama, 2011]

$$\widehat{\text{SMI}}(\{(\boldsymbol{x}_i, \boldsymbol{y}_{\pi(i)})\}_1^n) = \frac{1}{2n}\text{tr}\left(\text{diag}\left(\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\Pi}}\right) \boldsymbol{K}\boldsymbol{\Pi}^\top \boldsymbol{L}\boldsymbol{\Pi}\right) - \frac{1}{2}, \tag{6.2}$$

where $\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\Pi}}$ is computed by using $\{(\boldsymbol{x}_i, \boldsymbol{y}_{\pi(i)})\}_{i=1}^n$. If we set $\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\Pi}} = \boldsymbol{1}_n$ and ignore constants of Eq.(6.2), SMI boils down to an unnormalized variant of the kernel target alignment (uKTA) [Cristianini et al., 2002]:

$$\text{uKTA}(\{(\boldsymbol{x}_i, \boldsymbol{y}_{\pi(i)})\}_{i=1}^n) = \text{tr}\left(\boldsymbol{K}\boldsymbol{\Pi}^\top \boldsymbol{L}\boldsymbol{\Pi}\right). \tag{6.3}$$

This similarity function takes large value if the Gram matrices $\boldsymbol{K}$ and $\boldsymbol{L}$ are similar, and a small value if they are not similar. Note that, in the original KTA, we have the normalization term. However, this makes the optimization hard, and thus we employ the unnormalized variant of KTA. Moreover, uKTA can be regarded as a non-centered variant of HSIC [Gretton et al., 2005].

6.3.4 Optimization

For initializing $\boldsymbol{\Theta_x}$ and $\boldsymbol{\Theta_y}$, we first independently estimate $\boldsymbol{\Theta_x}$ and $\boldsymbol{\Theta_y}$ by using autoencoders. Then we employ an alternative optimization for learning $\boldsymbol{\Theta_x}$ and $\boldsymbol{\Theta_y}$ and $\boldsymbol{\Pi}$ together. We optimize $\boldsymbol{\Theta_x}$ and $\boldsymbol{\Theta_y}$ with fixed $\boldsymbol{\Pi}$ (intuition: learn a representation that maximizes statistical dependency, while preserving reconstruction), and then optimize $\boldsymbol{\Pi}$ with fixed $\boldsymbol{\Theta_x}$ and $\boldsymbol{\Theta_y}$ (intuition: find the cross-domain matches that maximize statistical dependency). This alternation is continued until convergence. We summarize the steps in Algorithm 8

Optimization for $\boldsymbol{\Theta_x}$ and $\boldsymbol{\Theta_y}$ With fixed permutation matrix $\boldsymbol{\Pi}$ (or π), the overall DMAE objective function is written as:

$$\min_{\boldsymbol{\Theta_x}, \boldsymbol{\Theta_y}} \sum_{i=1}^n ||\boldsymbol{x}_i - \boldsymbol{f}_x(\boldsymbol{g_x}(\boldsymbol{x}_i))||_2^2 + ||\boldsymbol{y}_i - \boldsymbol{f}_y(\boldsymbol{g_y}(\boldsymbol{y}_i))||_2^2 - \lambda D_\Pi(\{\boldsymbol{g_x}(\boldsymbol{x}_i), \boldsymbol{g_y}(\boldsymbol{y}_{\pi(i)})\}_{i=1}^n)$$

This problem is within-view autoencoder learning with the additional objective that the representation should maximize statistical dependency between the views. This can be solved by backpropagation, by differentiating the dependency measures in Eqs. 6.3 or 6.2 with respect to $\boldsymbol{\Theta_x}$ and $\boldsymbol{\Theta_y}$.

Optimizing $\boldsymbol{\Pi}$ (SMI) For optimizing $\boldsymbol{\Pi}$, we employ a regularized variant of LSOM [Yamada et al., 2015, Yamada and Sugiyama, 2011]. Given our autoencoder representations $\{\boldsymbol{g_x}(\boldsymbol{x}_i), \boldsymbol{g_y}(\boldsymbol{y}_{\pi(i)})\}_{i=1}^n$, the empirical estimate of SMI is:

$$\mathrm{SMI} = \frac{1}{2n}\mathrm{tr}\left(\mathrm{diag}\left(\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\Theta}}\right)\boldsymbol{K}_{\boldsymbol{\Theta_x}}\boldsymbol{\Pi}^\top \boldsymbol{L}_{\boldsymbol{\Theta_y}}\boldsymbol{\Pi}\right) - \frac{1}{2}, \tag{6.4}$$

where

$$[\boldsymbol{K}_{\boldsymbol{\Theta}_x}]_{ij} = \exp\Big(-\frac{||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2}{2\boldsymbol{\sigma}_x^2}\Big), [\boldsymbol{L}_{\boldsymbol{\Theta}_y}]_{ij} = \exp\Big(-\frac{||\boldsymbol{y}_i - \boldsymbol{y}_j||_2^2}{2\boldsymbol{\sigma}_y^2}\Big)$$

The optimization problem can then be written as:

$$\max_{\boldsymbol{\Pi}\in\{0,1\}^{n\times n}} \quad \mathrm{tr}\left(\mathrm{diag}\left(\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\Theta}}\right)\boldsymbol{K}_{\boldsymbol{\Theta_x}}\boldsymbol{\Pi}^\top \boldsymbol{L}_{\boldsymbol{\Theta_y}}\boldsymbol{\Pi}\right) \quad \mathrm{s.t.} \quad \boldsymbol{\Pi}\boldsymbol{1}_n = \boldsymbol{1}_n, \boldsymbol{\Pi}^\top\boldsymbol{1}_n = \boldsymbol{1}_n.$$

This is a quadratic assignment programming problem and is NP-hard. Thus, to solve for the permutation matrix $\boldsymbol{\Pi}$ efficiently, we solve a relaxed version of the problem with the regularization based optimization technique [Djuric et al., 2012]:

$$\begin{aligned}\max_{\boldsymbol{\Pi}} \quad & \mathrm{tr}\left(\mathrm{diag}\left(\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\Theta}}\right)\boldsymbol{K}_{\boldsymbol{\Theta_x}}\boldsymbol{\Pi}^\top \boldsymbol{L}_{\boldsymbol{\Theta_y}}\boldsymbol{\Pi}\right) - \lambda_{\boldsymbol{\Pi}}\sum_{k=1}^n \Big((\sum_{\ell=1}^n \boldsymbol{\Pi}_{k\ell} - 1)^2 + (\sum_{\ell=1}^n \boldsymbol{\Pi}_{\ell k} - 1)^2\Big) \\ \mathrm{s.t.} \quad & \boldsymbol{\Pi}_{k\ell} \geq 0, \mathrm{for}\ k, \ell \in \{1, 2, \ldots, n\},\end{aligned}$$

where $\lambda_{\boldsymbol{\Pi}} \geq 0$ is a regularizer and $\boldsymbol{\Pi}$ is optimized with gradient ascent.

Optimizing $\boldsymbol{\Pi}$ (KTA) For optimizing $\boldsymbol{\Pi}$, we employ a kernelized sorting Djuric et al. [2012], Quadrianto et al. [2009] strategy. The empirical estimate of uKTA using $\{\boldsymbol{g_x}(\boldsymbol{x}_i), \boldsymbol{g_y}(\boldsymbol{y}_{\pi(i)})\}_{i=1}^n$ is

$$\mathrm{uKTA} = \mathrm{tr}\left(\boldsymbol{K}_{\boldsymbol{\Theta_x}}\boldsymbol{\Pi}^\top \boldsymbol{L}_{\boldsymbol{\Theta_y}}\boldsymbol{\Pi}\right). \tag{6.5}$$

This is again a quadratic assignment programming problem and is NP-hard. Thus, we solve a relaxed version of this problem which is convex in nature [Djuric et al., 2012]:

$$\min_{\boldsymbol{\Pi}\in[0,1]^{n\times n}} ||\boldsymbol{K}_{\boldsymbol{\Theta}_x}\boldsymbol{\Pi}^T - (\boldsymbol{L}_{\boldsymbol{\Theta}_y}\boldsymbol{\Pi})^T||_F^2 \quad \mathrm{s.t.} \quad \boldsymbol{\Pi}\boldsymbol{1}_n = \boldsymbol{1}_n, \boldsymbol{\Pi}^\top\boldsymbol{1}_n = \boldsymbol{1}_n.$$

This problem is convex with respect to $\boldsymbol{\Pi}$, and thus, we can obtain a globally optimal solution for this sub-problem. To efficiently estimate the permutation matrix, we solve the following problem by using gradient descent. [Djuric et al., 2012]:

$$\min_{\mathbf{\Pi}} ||\boldsymbol{K}_{\mathbf{\Theta}_x}\mathbf{\Pi}^T - (\boldsymbol{L}_{\mathbf{\Theta}_y}\mathbf{\Pi})^T||_F^2 + \lambda_{\mathbf{\Pi}} \sum_{k=1}^{n} \left((\sum_{\ell=1}^{n} \mathbf{\Pi}_{k\ell} - 1)^2 + (\sum_{\ell=1}^{n} \mathbf{\Pi}_{\ell k} - 1)^2 \right)$$
$$\text{s.t.} \mathbf{\Pi}_{k\ell} \geq 0, \text{for } k, \ell \in \{1, 2, \dots, n\},$$

Algorithm 8 Learning algorithm for DMAE-SMI

```
Input: Unpaired Data $\{\boldsymbol{X}_i\}, \{\boldsymbol{Y}_i\}$. Params: $\lambda$, $\sigma_x$ and $\sigma_y$.
Init: weights $\mathbf{\Theta}_{\boldsymbol{X}}$, $\mathbf{\Theta}_{\boldsymbol{Y}}$, alignment matrix $\mathbf{\Pi}$.
while not converged
Update $\mathbf{\Theta}_{\boldsymbol{X}}$, $\mathbf{\Theta}_{\boldsymbol{Y}}$ with backprop on Eq 6.1. Fix $\mathbf{\Pi}$.
Update $\mathbf{\Pi}$ with gradient ascent. Fixing $\mathbf{\Theta}_{\boldsymbol{X}}$, $\mathbf{\Theta}_{\boldsymbol{Y}}$.
end while
Output: Pairing Matrix $\mathbf{\Pi}$. Encoders $\mathbf{\Theta}_{\boldsymbol{X}}$, $\mathbf{\Theta}_{\boldsymbol{Y}}$.
```

6.3.5 Generalizations

Learning from Paired and Unpaired Data In the previous section we introduced our method assuming no paired data was available (unsupervised) case. We next explain our method in the case of some paired data (semi-supervised) case. Denote the paired data as $\{(\boldsymbol{x}'_j, \boldsymbol{y}'_j)\}_{j=1}^{n'}$ and unpaired data as $\{\boldsymbol{x}_i\}_{i=1}^{n}$ and $\{\boldsymbol{y}_i\}_{i=1}^{n}$ $(n' < n)$. Then, the semi-supervised variant of DMAE is as follows:

With fixed permutation matrix $\mathbf{\Pi}$ (or π), the objective function is:

$$\begin{aligned}\min_{\mathbf{\Theta}_{\boldsymbol{x}}, \mathbf{\Theta}_{\boldsymbol{y}}} & \sum_{i=1}^{n} ||\boldsymbol{x}_i - \boldsymbol{f}_x(\boldsymbol{g}_x(\boldsymbol{x}_i))||_2^2 + ||\boldsymbol{y}_i - \boldsymbol{f}_y(\boldsymbol{g}_y(\boldsymbol{y}_i))||_2^2 \\ & + \sum_{j=1}^{n'} ||\boldsymbol{x}_j^{'} - \boldsymbol{f}_x(\boldsymbol{g}_x(\boldsymbol{x}_j^{'}))||_2^2 + ||\boldsymbol{y}_j^{'} - \boldsymbol{f}_{\boldsymbol{y}}(\boldsymbol{g}_y(\boldsymbol{y}_j^{'}))||_2^2 \\ & - \lambda \left(D_{\mathbf{\Pi}}(\boldsymbol{g}_x(\boldsymbol{x}_i), \boldsymbol{g}_y(\boldsymbol{y}_{\pi(i)})_{i=1}^{n}) + D(\boldsymbol{g}_x(\boldsymbol{x}_j), \boldsymbol{g}_y(\boldsymbol{y}_j)_{j=1}^{n'}) \right)\end{aligned}$$

This is optimized for $\mathbf{\Theta}_{\boldsymbol{x}}$ and $\mathbf{\Theta}_{\boldsymbol{y}}$ with backpropagation as before, fixing $\mathbf{\Pi}$. Then with a given $\mathbf{\Theta}_{\boldsymbol{x}}$ and $\mathbf{\Theta}_{\boldsymbol{y}}$, we can optimize $\mathbf{\Pi}$ as:

$$\max_{\mathbf{\Pi}} \quad D_{\mathbf{\Pi}}(\{\boldsymbol{g}_{\boldsymbol{x}}(\boldsymbol{x}_i), \boldsymbol{g}_{\boldsymbol{y}}(\boldsymbol{y}_{\pi(i)})\}_{i=1}^{n}) - \lambda_{\mathbf{\Pi}} \sum_{k=1}^{n} \left((\sum_{\ell=1}^{n} \mathbf{\Pi}_{k\ell} - 1)^2 + (\sum_{\ell=1}^{n} \mathbf{\Pi}_{\ell k} - 1)^2 \right)$$
$$\text{s.t.} \quad \mathbf{\Pi}_{k\ell} \geq 0, \text{for } k, \ell \in \{1, 2, \dots, n\}.$$

Fully Supervised Case The fully supervised case is a trivial extension of the above. In this case $n = 0$, $\mathbf{\Pi}$ is given, and we only need to optimize $\mathbf{\Theta}_x$ and $\mathbf{\Theta}_y$ for matching criterion $D_{\mathbf{\Pi}}(\cdot, \cdot)$.

6.4 Optimal Transport & Sinkhorn Matching

Thus far we have introduced DMAE-SMI as our main method for cross-modal matching. This algorithm benefits from joint representation learning and matching, however depends on the Hungarian algorithm. Hungarian algorithm is known to be of cubic time complexity and thus prevents us to scale to large multi-view datasets. In this section we briefly introduce an alternative cross-modal matching algorithm based on the notion of optimal transport, as realised by the Sinkhorn algorithm.

Aligning two high dimensional points is a fundermental problem in machine learning with applications in natural language processing such as word translation [Alvarez-Melis and Jaakkola, 2018, Artetxe et al., 2016] to problems in computer vision such as point set registration [Cootes et al., 1995]. These approaches assumed certain geometric constraints and provided reasonable success. Optimal Transport (OT) [Peyré and Cuturi, 2019] provides an elegant framework to compare high dimensional probability spaces. It provides a well-founded, geometrically well driven approach to realize the alignment between objects such as words in different languages. The Sinkhorn algorithm (Sec 4.6 Alg 6) can be used to estimate effective cross-modal matching between two sets of data points that can be compared directly to define a reasonable cost matrix. In the multi-view case data cannot be compared directly to define a cost matrix, so applying Sinkhorn to this problem requires jointly learning a shared embedding or cross-modal mapping. We explore an optimization scheme that iterates between: (i) updating a cross-modal mapping in the form of a linear regression, conditional on the currently estimated matching, and (ii) updating Sinkhorn-based matching conditional on the current cross-modal regression. The resulting procedure, shown in Algorithm 9, thus jointly estimates both matching and cross-modal mapping. In the experiments we compare this Sinkhorn-based alternative to our DMAE-SMI.

Algorithm 9 Procedure to learn a transportation matrix

```
Input: Unpaired Data $\{\boldsymbol{X}_i\}, \{\boldsymbol{Y}_i\}$. Params: $\lambda$, probability vectors $\mathbf{p}$ and $\mathbf{q}$, regularizer $\nu$
//Compute weight matrix $W = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y$
//Compute cost matrix $\boldsymbol{C}_{ij} = ||W\boldsymbol{x}_i - \boldsymbol{y}_j||^2$
    $\mathbf{a} \leftarrow \mathbf{1} \quad \mathbf{K} \leftarrow \exp\{-\mathbf{C}/\lambda\}$
while not converged
//Sinkhorn iterations of Eq 4.15
    $\mathbf{a} \leftarrow \mathbf{p} \oslash \mathbf{K}\mathbf{b}, \mathbf{b} \leftarrow \mathbf{q} \oslash \mathbf{K}^T\mathbf{a}$
    $\boldsymbol{\pi} \leftarrow \mathrm{diag}(a)\,\mathbf{K}\,\mathrm{diag}(b)$
    $W \leftarrow W - \nu\frac{\partial}{\partial W}(\boldsymbol{X}W - \pi\boldsymbol{Y})^2$
end while
Output: Transportation Matrix $\boldsymbol{\pi}$,$W$.
```

6.5 Experiments

We evaluate our contributions with two sets of experiments: image-caption matching (Section 6.5.1) and classifier learning (Section 6.5.2).

Datasets We evaluate our method on the well known Microsoft COCO dataset [Lin et al., 2014a] and the Flickr30k dataset [Plummer et al., 2015]. Flickr30k has 30000 standard training images. We use identical training , testing splits of [Karpathy and Fei-Fei, 2017, Faghri et al., 2018]. As explained in [Faghri et al., 2018], there is a set of $30,504$ validation images that are generally included in the training process have been left out of this split. The results are reported on testing on the full $5K$ test images.

Settings We use a standard SGD optimizer. The number of encoding and decoding layers were set to 3. The encoding layer consists of $1000-300-50$ and *tanh* was used for activation (See Figure 6.1 (b) for the model architecture). The regularization parameter were set to $\lambda = 0.7$, $\lambda_{\mathbf{\Pi}} = 1.0$, and the kernel parameters $\sigma_x^2 = 2.5$ and $\sigma_y^2 = 0.5$ for all experiments. The learning rate was set at $1e-3$.

Alternatives: Supervised For supervised learning, we evaluate the following baselines. **DeepCCA:** CCA with deep architecture [Andrew et al., 2013]. **Two-way Nets:** Two way nets use pre-trained VGG networks followed by fully connected layers (FC) and ReLU nonlinearities [Wang et al., 2016, 2017]. Captioning only. **ReViSE:** uses autoencoders for each modality, minimizing the reconstruction loss for each modality and the maximum mean discrepancy between them [Tsai et al., 2017].

Alternatives: Semi-supervised We evaluate our proposed **DMAE-uKTA** and **DMAE-SMI** methods against the following alternatives for unpaired data learning: **MCCA:** Matching CCA [Haghighi et al., 2008] for learning from paired and unpaired data across multiple views. **Shallow-KTA** and **Shallow-SMI** which are the non deep version are evaluated to learn from paired and unpaired data.

6.5.1 Image-Sentence/Sentence-Image Retrieval

Benchmark Details We evaluate Image→Sentence and Sentence→Image retrieval using the widely studied Flickr30K [Young et al., 2014] and MS-COCO [Lin et al., 2014b] datasets. Flickr30K consists of 31,783 images accompanied by descriptions. The larger MS-COCO dataset [Lin et al., 2014b] consists of 123,000 images, along accompanied by descriptions. Each dataset has 1000 testing images. Flickr30K has 5000 test sentences and COCO has 1000. To compare the methods, we use the evaluation metrics proposed in [Karpathy and Fei-Fei, 2017]: Image-text and text-image matching performance quantified by Recall@$K = \{1, 5\}$. We encode each image with $4096d$ VGG-19 deep feature [Simonyan and Zisserman, 2014] and a $300d$ word-vector [Mikolov et al., 2013d] average for each sentence.

Table 6.1: Fully supervised image-sentence matching results on Flickr30K and MS-COCO. [1]. Our implementation of ReViSE^b variant (reconstruction loss and MMD). Top block: Prior methods. Middle block: Ablations of our method. Bottom block: Our methods.

Flickr30K				
	Image-to-Text		Text-to-Image	
Approach	R@1	R@5	R@1	R@5
DeepCCA [Andrew et al., 2013]	29.3	57.4	28.2	54.7
Two-way nets [Wang et al., 2017]	49.8	67.5	36.0	55.6
ReViSE ^{b} [Tsai et al., 2017] [2]	34.7	63.2	29.2	58.0
MCCA	4.3	5.7	3.1	8.4
DMAE-SMI	20.7	22.6	10.5	11.4
DMAE-uKTA	20.2	22.4	10.4	11.2
MS-COCO				
DeepCCA [Andrew et al., 2013]	40.2	68.7	27.8	58.9
Two-way nets [Wang et al., 2017]	**55.8**	75.2	**39.7**	63.3
ReViSE [Tsai et al., 2017]	51.8	76.3	38.7	64.2
MCCA	12.8	13.6	7.2	8.3
DMAE-SMI	19.8	20.9	17.6	21.2
DMAE-uKTA	19.0	20.7	17.0	21.1

Table 6.2: Semi-supervised and unsupervised image-sentence retrieval results on Flickr30K and MS-COCO. The metric used is R@1 and retrieval is done directly and using regression methods. Chance value is 0.1%

Flickr30K

	Supervised										Un/Semi-Supervised									
	MCCA		shallow KTA		shallow SMI		DMAE-uKTA		DMAE-SMI		MCCA		shallow KTA		shallow SMI		DMAE-uKTA		DMAE-SMI	
Labels	I→T	T→I	I→T	T→I	I → T	T → I	I →T	T → I	I → T	T→I	I→T	T→I	I→T	T→I	I→T	T→I	I→T	T→I	I→T	I→T
0% (Direct)	-	-	-	-	-	-	-	-	-	-	0.0	0.0	-	-	-	-	0.1	0.1	0.2	0.1
0% (Regression)	-	-	-	-	-	-	-	-	-	-	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.2	0.1
20% (Direct)	0.1	0.0	-	-	-	-	0.5	0.6	0.5	0.6	0.1	0.0	-	-	-	-	0.8	0.6	1.2	0.6
20% (Regression)	0.1	0.1	0.1	0.1	0.1	0.1	0.6	0.6	0.6	0.6	0.1	0.2	0.2	0.0	0.2	0.0	0.9	0.6	1.3	0.6
40% (Direct)	0.9	0.5	-	-	-	-	4.3	3.4	4.5	3.4	0.9	0.6	-	-	-	-	4.5	3.3	4.8	3.4
40% (Regression)	0.9	0.6	4.5	3.3	4.5	3.3	4.7	3.4	4.7	3.4	1.0	0.6	4.3	3.4	4.5	3.4	4.5	3.3	5.0	3.4

MS-COCO

	Supervised										Un/Semi-Supervised									
	MCCA		shallow KTA		shallow SMI		DMAE-uKTA		DMAE-SMI		MCCA		shallow KTA		shallow SMI		DMAE-uKTA		DMAE-SMI	
Labels	I→T	T→I	I→T	T→I	I → T	T → I	I →T	T → I	I → T	T→I	I→T	T→I	I→T	T→I	I→T	T→I	I→T	T→I	I→T	I→T
0% (Direct)	-	-	-	-	-	-	-	-	-	-	0.0	0.0	-	-	-	-	0.1	0.1	0.3	0.1
0% (Regression)	-	-	-	-	-	-	-	-	-	-	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.1	0.3	0.1
20% (Direct)	0.0	0.0	-	-	-	-	0.4	0.3	0.5	0.3	0.1	0.2	-	-	-	-	0.4	0.3	0.8	0.5
20% (Regression)	0.1	0.0	0.1	0.0	0.1	0.1	0.4	0.3	0.5	0.3	0.1	0.2	0.1	0.2	0.2	0.2	0.5	0.3	1.0	0.5
40% (Direct)	0.7	0.3	-	-	-	-	3.4	2.6	3.6	2.7	0.8	0.4	-	-	-	-	3.6	2.7	3.8	2.9
40% (Regression)	0.9	0.3	0.9	0.4	1.2	0.5	3.5	2.8	3.8	2.8	1.2	0.5	1.2	0.6	1.2	0.6	3.7	2.8	3.8	2.9

Supervised Learning We first evaluate our methods against prior state of the art in Image-Sentence matching in the standard supervised learning setting. From the results in Table 6.1 we make the following observations: (i) Our SMI provides a slightly better objective for our method than uKTA, this is expected since as we saw uKTA is a special case of SMI. (ii) Overall our approach is not comparable to state of the art captioning algorithms such as [Wang et al., 2016, 2017]. (iii) However unlike these, our method is general purpose designed for captioning. Nevertheless we outperform alternative general purpose methods such as MCCA.

Semi-supervised and Unsupervised Learning In the second experiment we investigate whether it is possible to learn captioning from partially paired or unpaired data. For the results in Table 6.2 the left (Supervised) block uses only the specified % of labeled data, and the right (Un/Semi-supervised) block uses both labeled and the available unlabeled data. We make the following observations: (i) This task is clearly significantly harder as all methods struggle with reduced data annotation. In particular, in the unsupervised case, only DMAE-SMI performs slightly above chance (0.1%) in the I→T condition. (ii)Semi-supervised learning here is also challenging. Comparing the left and right column groups, we can see that only DMAE-SMI based SSL sometimes improves on the supervised baseline (e.g Flickr I→T in the 20% condition). (iii)Comparing direct NN matching vs matching via using the estimated pairing to train a cross-modal regression model, we can see that using the regression based approach tends to improve performance slightly.

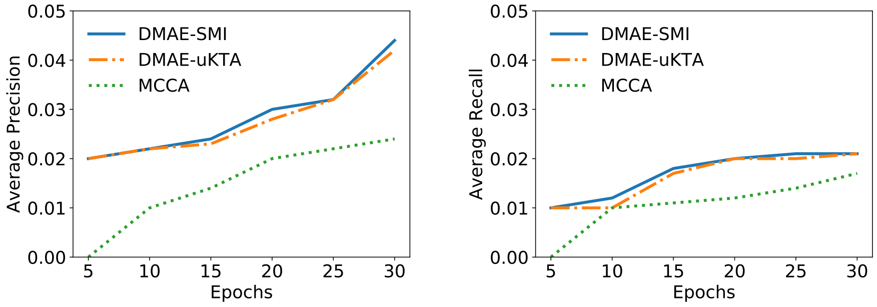
Discussion Unsupervised image captioning is a difficult, challenging and a real problem. While recently, unsupervised word translation [Conneau et al., 2017b] and unsupervised neural translation [Lample et al., 2017] map source and target language in similar space so that words across different languages can be aligned, the same does not hold for image captioning datasets. Some recent work [Kim et al., 2019] has managed to obtain some impressive results with semi-supervision. We hope future work can build towards this direction.

6.5.2 Unsupervised Classifier Learning

We consider training a classifier given a stack of images and stack of category embeddings (we use word vectors) that describe the categories covered by images in the stack. This is the 'unsupervised classifier learning' problem when there are *no annotated images*, so no pairings given. If the category labels of *some* images are unknown, and all categories have at least one annotated image, this a semi-supervised learning problem. In the case where the category labels of *some* images are unknown and some categories have no annotated images, this is a zero-shot learning problem. If category labels of all images are known (all pairings given), this is the standard supervised learning problem. Our framework can apply to all of these settings, but as fully supervised and zero-shot learning are well studied, we focus on the unsupervised and semi-supervised variants.

Benchmark Details We evaluate our approach on AWA [Lampert et al., 2009b]. As category embeddings, we use $300d$ word-vectors [Mikolov et al., 2013d]. For image features we use $4096d$ VGG-19 [Simonyan and Zisserman, 2014] features for AwA, Thus for AwA, image data is a $4096 \times n$ stack of n images, and category domain data is a $300 \times m$ stack of $m = 50$ word vectors. Unsupervised DMAE learns a joint embedding and the association matrix $\mathbf{\Pi} \in \{0, 1\}^{n \times n}$ that pairs images with categories where we duplicated the text vectors to ensure $\mathbf{\Pi}$ is square. The learned $\mathbf{\Pi}$ should ideally match the 1-hot label matrix that would normally be given as a target in supervised learning.

Settings We consider the unsupervised and semi-supervised in which only partial or no paired data are given for training.



(a) Precision.

(b) Recall.

Figure 6.2: Evolution of label matrix $\mathbf{\Pi}$ prediction accuracy during unsupervised classifier learning.

Metrics To fully diagnose the performance, we evaluate the following metric: (i) Matching accuracy. The accuracy of predicted $\widehat{\mathbf{\Pi}}$ on the training split compared to the ground-truth $\mathbf{\Pi}$ as quantified by average precision and average recall.

Results: Unsupervised Matching In the unsupervised classifier learning setting, it is a non-trivial achievement to correctly estimate associations between images and categories better than chance since we have no annotated pairings, and the heterogeneous domains are not a priori comparable. To quantify the accuracy of pairing, we compare estimated $\widehat{\mathbf{\Pi}}$ and true $\mathbf{\Pi}$ and compute compute the precision and recall by class. After learning DMAE-SMI on AwA we obtain a precision of 0.042 and recall of 0.021 averaged over all 50 classes given *no prior pairings* to start with.

To see how the accuracy of $\mathbf{\Pi}$ estimation changes during learning, we visualise the mean precision and recall over learning iterations in Figure 6.2. We can see that: (i) Precision and recall rise monotonically over time before eventually asymptoting. (ii) DMAE-SMI performance grows faster and converges to a higher point than the alternatives.

Results: Testing Accuracy To complete the evaluation of the actual learned classifier, we next assume that the estimated $\widehat{\mathbf{\Pi}}$ label matrix is correct, and use these labels to train a SVM classifier, which is then evaluated on the testing split of each dataset. The results for AwA are shown in Table 6.3. The L-U-T splits listed define different supervised (all training data pairs given), semi-supervised (some training pairs given) and unsupervised (no training pairs given) experimental conditions. For example 40-0-60 in AWA is supervised setting with 40% paired images for training and 60% for testing, while 20-20-60 is semi-supervised with 20% paired images and 20% unpaired images for training, and 60% testing images. We use the SVM classifier provided in [Pedregosa et al., 2011].

From the results we can see that (i) In the unsupervised 0-40-60 condition, all the shallow models perform at chance level (2%) while DMAE-SMI and Sinkhorn perform

Table 6.3: Classification accuracy on AWA test sets. The data is split is denoted as L-U-T, specifying the amount of (L)abeled training, (U)nlabeled training, and (T)esting data. AWA is 50-way classification so chance is 2% and the given L-U-T split is in %.

AwA						
L-U-T	MCCA	KTA	SMI	DMAE-uKTA	DMAE-SMI	Sinkhorn
0-40-60	0.02	0.02	0.02	0.03	0.03	0.04
20-0-60	0.82	0.82	0.82	0.82	0.82	0.82
20-20-60	0.82	0.84	0.84	0.84	0.85	0.87
40-0-60	0.88	0.88	0.88	0.88	0.88	0.88

above chance. (ii)Using labeled data for supervised learning brings a dramatic increase in performance as seen in 20-0-60 and 40-0-60 conditions. (iii)The 20-20-60 condition is the semi-supervised learning condition which aims to bridge the gap between the lower (20-0-60) and upper (40-0-60) bound supervised learning conditions. We can see that most methods bring some improvement from semi-supervised learning with DMAE-SMI and Sinkhorn performing best.

Discussion The current paradigm for building visual recognition systems requires laborious and costly per-image annotation. The above proof of concept demonstration of UCL with some labeled information shows a promising direction. In future by providing a pool of images and a list of wordvectors describing categories likely to be contained therein, visual recognition models of a reasonable degree of reliability could be learned with by providing some labeled information. However while promising results are obtained for SSL setting, more work is required to realise this vision for unsupervised setting. This could provide significant time and cost savings in many potential application scenarios like cross domain alignment [Yuan et al., 2020] which could reduce cost in automatic caption generation systems.

6.5.3 Further Analysis

Runtime and Complexity Our DMAE implementation is full batch for simplicity and accuracy. This means the required pairwise matching problem includes a $O(n^2)$ cost term. This limits scalability, but is not unexpected, and widely shared by many other matching algorithms. There are approximation routes to alleviate this. For example divide and conquer [Lyzinski et al., 2015] minibatch-based training reduces the cost to $O(Bn_b^2 + nB)$ for n instances, B minibatches and $n_b \ll n$ instances per minibatch.

6.5.4 Discussion

Limitations In this chapter we discussed some ways to learn unsupervised alignment between images and text. However there remains some limitations among the described algorithms which we highlight below

- The current model assumes our permutation matrix to be square and 1-1 which could be made more realistic using 1-many matching. For UCL, we currently stack the word vectors to ensure the square nature of the $\mathbf{\Pi}$ matrix to make it square which is inelegant method. A more realistic vision is to relax and make it rectangular which will allow for more realistic applications. To handle two different sequences, [Yamada et al., 2015] padded the sequences with zeros to ensure a square cost matrix. One can potentially use this idea in the current framework.

- The current DMAE-SMI algorithm has a clean objective for end-to-end matching and deep representation learning, but suffers from relying on the Hungarian algorithm. Meanwhile Sinkhorn algorithm provides effective matching, but currently only addresses cross-modal matching through an alternating optimization heuristic, and a simple linear cross-modal mapping. The main future task is to thus define an single objective for end-to-end optimization of both matching (using Sinkhorn) and deep representation learning.

- While we have explored the Sinkhorn algorithm, we havent explored the utility of different cost metrics. Commonly used metrics like Euclidean distance might not be the appropriate choice and other non-euclidean metrics could offer an elegant choice.

- The quality of the permutation matrix in Algorithm 9 depends on the choice of initial weight matrix W. Some future work would be to learn a better weight matrix with convergence guarantee.

- Gromov-Wasserstein [Mémoli, 2011] provides an elegant framework to compare two hetergenous metric spaces and provide a transportation cost to move from one space to the other. We would like to explore this in future work.

6.6 Conclusion

We proposed *Deep Matching Autoencoders* (DMAE), as an application of our previous cross-lingual matching idea to match between image and language modalities.

Conceptually DMAE elegantly spans unsupervised, semi-supervised, fully-supervised and zero-shot settings. However in practice our results were weaker than in the case of language. This may be because the intra-domain simliarities that our mehod aligns are less consistent between vision and language than they are between different languages, and thus harder to align. We have also shown how the Sinkhorn algorithm can lead to improved results with some seed labelled data. In future we will try to improve these results by combining sinkhorn-based alignment algorithm to replace the iterative Hungarian algorithm-based solution that we evaluated here.

Chapter 7

Conclusions and Outlook

7.1 Conclusion

In this thesis, we thoroughly studied the problems related with reducing annotation requirement through multi-view learning. With the increasing demand on creating intelligent systems, obtaining or creating annotated data proves to be a gridlock in many contemporary machine learning systems. Motivated by these ambition, the thesis looked at three different problems.

- In Chapter 3 we studied Zero shot learning through text-image transfer. We present the first distribution-embedding approach to category names and showed the benefit of using this approach compared to the traditional setting of vector based embeddings.
- Chapter 4 discussed varying degrees of metrics for aligning different views of a dataset.
- In Chapter 5, we discussed bilingual dictionary induction and looked into aligning monolingual word embeddings. We extended the existing SMI-based measures for unsupervised pairing to an end-to-end deep learning setting and demonstrated improved dictionary induction performance as a result.
- In Chapter 6, we applied our ideas to vision and language problems including captioning and introduced the novel problem of unsupervised classifier learning.

We look at some possible future directions. Specific highlights are summarised as

7.2 Limitations and Future work

In this section we discuss the possibilities of some limitations of our methods and some new future directions

7.2.1 Multi-sense probabilistic embeddings

Traditional distributional semantic models (DSM's) derive the meaning of a word solely based on co-occurance of words in a text. An exciting opportunity arises in DSM's in infusing visual information with text corpora. Most of the existing work represent word or image as vectors. Inspired by [Vilnis and McCallum, 2015], we proposed to represent images and text as distributions. In Chapter 3, we extended this framework towards zero shot learning. The current model has unimodal structure due to the gaussian assumption on embeddings. This assumption is problematic in the case of polysemous words. Many words have different senses based on the contextual surrounding. For example the word 'apple' could mean a fruit or could also represent the incorporation. A possible solution to ease this problem is to represent words as Gaussian mixture models (GMM's) where each sense is represented as a gaussian component. For future work, another worthy direction is to look at embedding distributions in a Wasserstein spaces. Wasserstein spaces provide probability distributions with an optimal transport metric which measures the distance traveled in transporting the mass in one distribution to the other. Recent work has shown that Wasserstein spaces offer more flexibility and are able to model complex relationships where Euclidean spaces fail [Frogner et al., 2019].

7.2.2 Optimal Transport

The DMAE algorithm comprises of a representation learning framework along with a dependency matching framework. The dependency measure framework relies on the Hungarian algorithm which is known to not scale to large datasets. Meanwhile we discussed the Sinkhorn algorithm which provides effective matching, but currently only addresses cross-modal matching through an alternating optimization heuristic, and a simple linear cross-modal mapping. In future, we would like to build our representation learning framework along with the Sinkhorn algorithm hoping to build more scalable models.

Another interesting direction is to build on top of the Gromov Wasserstein distance [Mémoli, 2011] which can be used to compare different metrics in different spaces. While it has been popularly used in the computer graphics community in shape matching, it holds a lot of promise in problems involving multimodal data.

Almost all problems in finding correspondence in optimal transport depends on the definition of a *transportation cost.* To design a reliable function is often difficult and practitioners often resort to using hand crafted measures or Euclidean distance. In future work, we could take a step towards learning appropriate cost function.

7.2.3 Graph Matching

In Chapters 4,5,6, we introduced an assignment matrix which was usually for simplicity assumed to be square and usually assumes one-to-one mapping. Graph Matching for shape comparision or network analysis is an exciting area in network modelling. Among the known algorithms, a class of problems called inexact graph matching is tailored specifically to real-world graph representations. This class of methods allows for a less strict correspondence of the graph vertices, allowing for many-to-many graph matching. Combined with optimal transport they open an interesting line of research direction which can be further explored.

Bibliography

Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *Computer Vision and Pattern Recognition*, 2013. doi: 10.1109/CVPR.2013.111.

J.-B. Alayrac, P. Bojanowski, N. Agrawal, J. Sivic, I. Laptev, and S. Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(1):131–142, 1966. ISSN 00359246. URL http://www.jstor.org/stable/2984279.

C. Allen and T. Hospedales. Analogies explained: Towards understanding word embeddings. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 223–231, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/allen19a.html.

C. Allen, I. Balazevic, and T. Hospedales. What the vec? towards probabilistically grounded embeddings. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7465–7475. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8965-what-the-vec-towards-probabilistically-grounded-embeddings.pdf.

D. Alvarez-Melis and T. Jaakkola. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1214.

G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1247–1255, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL http://proceedings.mlr.press/v28/andrew13.html.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017.

S. Arora, Y. Li, Y. Liang, T. Ma, and A. Risteski. Random walks on context spaces: Towards an explanation of the mysteries of semantic word embeddings. *CoRR*, abs/1502.03520, 2015.

URL http://arxiv.org/abs/1502.03520.

M. Artetxe, G. Labaka, and E. Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas, November 2016. Association for Computational Linguistics. URL https://aclweb.org/anthology/D16-1250.

M. Artetxe, G. Labaka, and E. Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *ACL*, 2017.

M. Artetxe, G. Labaka, E. Agirre, and K. Cho. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*, April 2018.

P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379, Nov 2010. ISSN 1432-1882. doi: 10.1007/s00530-010-0182-0. URL https://doi.org/10.1007/s00530-010-0182-0.

F. R. Bach and M. I. Jordan. Kernel independent component analysis. Technical report, USA, 2001.

F. R. Bach and M. I. Jordan. Kernel independent component analysis. *JMLR*, 3:1–48, Mar. 2003. ISSN 1532-4435.

D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate, 2014. URL http://arxiv.org/abs/1409.0473. cite arxiv:1409.0473Comment: Accepted at ICLR 2015 as oral presentation.

T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *arXiv preprint arXiv:1705.09406*, 2017.

M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm. Mutual information neural estimation. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 531–540, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/belghazi18a.html.

S. Benaim and L. Wolf. One-sided unsupervised domain mapping. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 752–762, USA, 2017. Curran Associates Inc. ISBN 978-1-5108-6096-4. URL http://dl.acm.org/citation.cfm?id=3294771.3294843.

F. Biessmann, S. Plis, F. C. Meinecke, T. Eichele, and K. Muller. Analysis of multimodal neuroimaging data. *IEEE Reviews in Biomedical Engineering*, 4:26–58, 2011. ISSN 1937-3333. doi: 10.1109/RBME.2011.2170675.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

P. Bojanowski, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic. Finding actors and actions in movies. In *Proc. ICCV*, 2013.

P. Bojanowski, R. Lajugie, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic. Weakly supervised action labeling in videos under ordering constraints. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 628–643, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

P. Bojanowski, R. Lajugie, E. Grave, F. Bach, I. Laptev, J. Ponce, and C. Schmid. Weakly-supervised alignment of video with text. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 4462–4470, Washington, DC, USA, 2015. IEEE Computer Society. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.507. URL `http://dx.doi.org/10.1109/ICCV.2015.507`.

E. Bruni, G. Boleda, M. Baroni, and N.-K. Tran. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, pages 136–145, 2012a. URL `http://dl.acm.org/citation.cfm?id=2390524.2390544`.

E. Bruni, J. Uijlings, M. Baroni, and N. Sebe. Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. In *ACM Multimedia*, 2012b.

E. Bruni, N. K. Tran, and M. Baroni. Multimodal distributional semantics. *J. Artif. Int. Res.*, 49(1):1–47, Jan. 2014. ISSN 1076-9757. URL `http://dl.acm.org/citation.cfm?id=2655713.2655714`.

R. B. Cattell. "parallel proportional profiles" and other principles for determining the choice of factors by rotation. *Psychometrika*, 9(4):267–283, Dec 1944. doi: 10.1007/BF02288739. URL `https://doi.org/10.1007/BF02288739`.

W. Chan, N. Jaitly, Q. Le, and O. Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964, March 2016. doi: 10.1109/ICASSP.2016.7472621.

S. Chandar, M. M. Khapra, H. Larochelle, and B. Ravindran. Correlational neural networks. *Neural Computation*, 28(2):257–285, February 2016.

B. Chang, U. Kruger, R. Kustra, and J. Zhang. Canonical correlation analysis based on hilbert-schmidt independence criterion and centered kernel target alignment. In *ICML*, 2013. URL `http://jmlr.csail.mit.edu/proceedings/papers/v28/chang13.pdf`.

X. Chang, T. Xiang, and T. M. Hospedales. Scalable and effective deep CCA via soft decorrelation. In *CVPR*, 2018.

X. Chen, X. Qiu, J. Jiang, and X. Huang. Gaussian mixture embeddings for multiple word prototypes. *arXiv preprint arXiv:1511.06246*, 2015.

K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-4012. URL `https://www.aclweb.org/anthology/W14-4012`.

Y.-A. Chung, W.-H. Weng, S. Tong, and J. Glass. Unsupervised cross-modal alignment of speech and text embedding spaces. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 7365–7375, USA, 2018. Curran Associates Inc. URL `http://dl.acm.org/citation.cfm?id=3327757.3327837`.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, Nov. 2011. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=1953048.2078186`.

A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou. Word translation without parallel data. *CoRR*, abs/1710.04087, 2017a. URL `http://arxiv.org/abs/1710.04087`.

A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017b.

T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38 – 59, 1995. ISSN 1077-3142. doi: https://doi.org/10.1006/cviu.1995.1004. URL `http://www.sciencedirect.com/science/article/pii/S1077314285710041`.

T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, New York, NY, USA, 2006. ISBN 0471241954.

B. Coyne and R. Sproat. Wordseye: An automatic text-to-scene conversion system. In *SIGGRAPH*, 2001. doi: 10.1145/383259.383316. URL `http://doi.acm.org/10.1145/383259.383316`.

N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. S. Kandola. On kernel-target alignment. In *NIPS*, 2002.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2292–2300. Curran Associates, Inc., 2013. URL `http://papers.nips.cc/paper/4927-sinkhorn-distances-lightspeed-computation-of-optimal-transport.pdf`.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, June 2005. doi: 10.1109/CVPR.2005.177.

G. A. Darbellay and I. Vajda. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45(4):1315–1321, May 1999. doi: 10.1109/18.761290.

V. R. de Sa. Learning classification with unlabeled data. In *NIPS*, 1993.

Z. Ding, H. Zhao, and Y. Fu. *Introduction*, pages 1–6. Springer International Publishing, Cham, 2019. ISBN 978-3-030-00734-8. doi: 10.1007/978-3-030-00734-8_1. URL `https://doi.org/10.1007/978-3-030-00734-8_1`.

G. Dinu and M. Baroni. Improving zero-shot learning by mitigating the hubness problem. *ICLR Workshops*, 2015.

G. Dinu, A. Lazaridou, and M. Baroni. Improving zero-shot learning by mitigating the hubness problem. In *ICLR Workshop Paper*, 2015.

N. Djuric, M. Grbovic, and S. Vucetic. Convex kernelized sorting. In *AAAI*, 2012.

A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard. Multimodal deep learning for robust rgb-d object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 681–687, Sep. 2015. doi: 10.1109/IROS. 2015.7353446.

F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018.

M. Faruqui and C. Dyer. Improving vector space word representations using multilingual correlation. In *Proc. of EACL*, 2014.

F. Feng, X. Wang, and R. Li. Cross-modal retrieval with correspondence autoencoder. In *ACM Multimedia*, 2014.

A. Ferraresi, E. Zanchetta, M. Baroni, and S. Bernardini. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *In Proceedings of the 4th Web as Corpus Workshop (WAC-4*, 2008.

J. R. Firth. A synopsis of linguistic theory 1930-55. 1952-59:1–32, 1957.

D. A. Forsyth and J. Ponce. *Computer Vision: A Modern Approach.* Prentice Hall Professional Technical Reference, 2002. ISBN 0130851981.

A. M. Fraser and H. L. Swinney. Independent coordinates for strange attractors from mutual information. *Phys. Rev. A*, 33:1134–1140, Feb 1986. doi: 10.1103/PhysRevA.33.1134. URL `https://link.aps.org/doi/10.1103/PhysRevA.33.1134`.

C. Frogner, F. Mirzazadeh, and J. Solomon. Learning embeddings into entropic wasserstein spaces. *CoRR*, abs/1905.03329, 2019. URL `http://arxiv.org/abs/1905.03329`.

A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *Neural Information Processing Systems (NIPS)*, 2013a.

A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013b.

Y. Fu, T. Hospedales, T. Xiang, Z. Fu, and S. Gong. Transductive multi-view embedding for zero-shot recognition and annotation. In *ECCV*, 2014.

Y. Fu, T. Hospedales, T. Xiang, and S. Gong. Transductive multi-view zero-shot learning. *IEEE TPAMI*, 37(11):2332 – 2345, 2015a.

Z. Fu, T. A. Xiang, E. Kodirov, and S. Gong. Zero-shot object recognition by semantic manifold distance. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2635–2644, June 2015b. doi: 10.1109/CVPR.2015.7298879.

D. Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7 (2):155 – 170, 1983. ISSN 0364-0213. doi: https://doi.org/10.1016/S0364-0213(83)80009-3. URL `http://www.sciencedirect.com/science/article/pii/S0364021383800093`.

D. Gentner and K. D. Forbus. Computational models of analogy. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(3):266–276, 2011. doi: 10.1002/wcs.105. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/wcs.105`.

G. Glavaš, R. Litschko, S. Ruder, and I. Vulić. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 710–721, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1070. URL `https://www.aclweb.org/anthology/P19-1070`.

S. Gong, M. Cristani, S. Yan, and C. C. Loy, editors. *Person Re-Identification*. Springer, 2014.

Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 2013.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. URL `http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf`.

S. Gouws, Y. Bengio, and G. Corrado. Bilbowa: Fast bilingual distributed representations without word alignments. In *ICML*, 2015.

A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1764–II–1772. JMLR.org, 2014. URL `http://dl.acm.org/citation.cfm?id=3044805.3045089`.

A. Gretton, O. Bousquet, A. Smola, and B. Scholkopf. Measuring statistical dependence with Hilbert-schmidt norms. In *ALT*, 2005.

S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *2013 IEEE International Conference on Computer Vision*, pages 2712–2719, Dec 2013. doi: 10.1109/ICCV.2013.337.

J. Guo, W. Che, D. Yarowsky, H. Wang, and T. Liu. A distributed representation-based framework for cross-lingual transfer parsing. *JAIR*, 55(1):995–1023, Jan. 2016. ISSN 1076-9757. URL `http://dl.acm.org/citation.cfm?id=3013558.3013584`.

W. Guo, J. Wang, and S. Wang. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394, 2019. ISSN 2169-3536. doi: 10.1109/ACCESS.2019.2916887.

A. Gupta, Y. Verma, and C. V. Jawahar. Choosing linguistics over vision to describe images. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI'12, pages 606–612. AAAI Press, 2012. URL `http://dl.acm.org/citation.cfm?id=2900728.2900815`.

S. Gupta, R. Girshick, P. Arbeláez, and J. Malik. Learning rich features from rgb-d images for object detection and segmentation. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 345–360, Cham, 2014. Springer International Publishing.

A. Haghighi, P. Liang, T. Berg-Kirkpatrick, and D. Klein. Learning bilingual lexicons from monolingual corpora. In *ACL*, 2008.

D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004. ISSN 0899-7667. doi: 10.1162/0899766042321814. URL `http://dx.doi.org/10.1162/0899766042321814`.

Z. S. Harris. Distributional structure. *¡i¿WORD¡/i¿*, 10(2-3):146–162, 1954. doi: 10.1080/00437956.1954.11659520. URL `https://doi.org/10.1080/00437956.1954.11659520`.

R. Harshman. Foundations of the parafac procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Papers in Phonetics*, 16, 1970.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL `http://dx.doi.org/10.1162/neco.1997.9.8.1735`.

B. Horwitz and D. Poeppel. How can eeg/meg and fmri/pet data be combined? *Human Brain Mapping*, 17(1):1–3, 2002. doi: 10.1002/hbm.10057. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.10057`.

Y. Hoshen. Non-adversarial mapping with vaes. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7528–7537. Curran Associates, Inc., 2018. URL `http://papers.nips.cc/paper/7981-non-adversarial-mapping-with-vaes.pdf`.

Y. Hoshen and L. Wolf. Nam: Non-adversarial unsupervised domain mapping. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision – ECCV 2018*, pages 455–470, Cham, 2018a. Springer International Publishing. ISBN 978-3-030-01264-9.

Y. Hoshen and L. Wolf. Non-adversarial unsupervised word translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 469–478, Brussels, Belgium, Oct.-Nov. 2018b. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/D18-1043`.

Y. Hoshen and L. Wolf. Unsupervised correlation analysis. In *CVPR*, 2018c.

T. M. Hospedales and S. Vijayakumar. Structure inference for bayesian multisensory scene understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2140–2157, Dec 2008. ISSN 0162-8828. doi: 10.1109/TPAMI.2008.25.

H. Hotelling. RELATIONS BETWEEN TWO SETS OF VARIATES*. *Biometrika*, 28(3-4):321–377, 12 1936a. ISSN 0006-3444. doi: 10.1093/biomet/28.3-4.321. URL `https://doi.org/10.1093/biomet/28.3-4.321`.

H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936b.

W. Hsieh. Nonlinear canonical correlation analysis by neural networks. *Neural Networks*, 13(10):1095 – 1105, 2000. ISSN 0893-6080. doi: https://doi.org/10.1016/S0893-6080(00)00067-8. URL `http://www.sciencedirect.com/science/article/pii/S0893608000000678`.

J. E. Hummel and K. J. Holyoak. Distributed representations of structure: A theory of analogical access and mapping. *PSYCHOLOGICAL REVIEW*, 104(3):427–466, 1997.

P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arxiv*, 2016.

P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

T. Iwata, T. Hirao, and N. Ueda. Unsupervised cluster matching via probabilistic latent variable models. In *AAAI*, 2013.

J. Jagarlamudi, S. Juarez, and H. Daumé III. Kernelized sorting for natural language processing. In *AAAI*, 2010.

T. Jebara. Kernelizing sorting, permutation and alignment for minimum volume PCA. In *Conference on Learning Theory*, 2004.

T. Jebara, R. Kondor, and A. Howard. Probability product kernels. *Journal of Machine Learning Research*, 5:819–844, 2004.

M. Kan, S. Shan, and X. Chen. Multi-view deep network for cross-view classification. In *CVPR*, 2016.

A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676, April 2017. ISSN 0162-8828. doi: 10.1109/TPAMI.2016.2598339.

A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):664–676, Apr. 2017. ISSN 0162-8828. doi: 10.1109/TPAMI.2016.2598339. URL `https://doi.org/10.1109/TPAMI.2016.2598339`.

J. R. Kettering. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 12 1971. ISSN 0006-3444. doi: 10.1093/biomet/58.3.433. URL `https://doi.org/10.1093/biomet/58.3.433`.

B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi. Multisensor data fusion: A review of the state-of-the-art. *Information Fusion*, 14(1):28 – 44, 2013. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2011.08.001. URL `http://www.sciencedirect.com/science/article/pii/S1566253511000558`.

D. Kiela and L. Bottou. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-14)*, 2014.

D.-J. Kim, J. Choi, T.-H. Oh, and I. S. Kweon. Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2012–2023, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1208. URL `https://www.aclweb.org/anthology/D19-1208`.

T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1857–1865, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL `http://proceedings.mlr.press/v70/kim17a.html`.

Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. doi: 10.3115/v1/

D14-1181. URL https://www.aclweb.org/anthology/D14-1181.

B. Klein, G. Lev, G. Sadeh, and L. Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *CVPR*, 2015.

A. Klementiev, I. Titov, and B. Bhattarai. Inducing crosslingual distributed representations of words. In *COLING*, 2012.

K. Knight, A. Nair, N. Rathod, and K. Yamada. Unsupervised analysis for decipherment problems. In *Proc. ACL-COLING*, 2006.

A. Kojima, T. Tamura, and K. Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision*, 50(2):171–184, Nov 2002. ISSN 1573-1405. doi: 10.1023/A:1020346032608. URL https://doi.org/10.1023/A:1020346032608.

R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*, 2017.

N. Krishnamoorthy, G. Malkarnenkar, R. Mooney, K. Saenko, and S. Guadarrama. Generating natural-language video descriptions using text-mined knowledge. In *AAAI*, 2013.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc. URL http://dl.acm.org/citation.cfm?id=2999134.2999257.

J. B. Kruskal. An overview of sequence comparison: Time warps, string edits, and macromolecules. *SIAM Review*, 25(2):201–237, 1983. ISSN 00361445. URL http://www.jstor.org/stable/2030214.

H. W. Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

H. W. Kuhn and B. Yaw. The hungarian method for the assignment problem. *Naval Res. Logist. Quart*, pages 83–97, 1955.

G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, Dec 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2012.162.

P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi. Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 359–368, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2390524.2390575.

D. Lahat, T. Adali, and C. Jutten. Multimodal data fusion: An overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, Sep. 2015. ISSN 0018-9219. doi: 10.1109/JPROC.2015.2460697.

P. L. Lai and C. Fyfe. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 10(05):365–377, 2000. doi: 10.1142/S012906570000034X. URL https://doi.org/10.1142/S012906570000034X. PMID: 11195936.

C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition*, 2009a.

C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009b.

C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE TPAMI*, 36(3):453–465, 2014. ISSN 0162-8828. doi: 10.1109/TPAMI.2013.140. URL http://dx.doi.org/10.1109/TPAMI.2013.140.

G. Lample, L. Denoyer, and M. Ranzato. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2017.

B. Landau, L. Smith, and S. Jones. Object perception and object naming in early development. *Trends in Cognitive Sciences*, 2(1):19 – 24, 1998. ISSN 1364-6613. doi: https://doi.org/10.1016/S1364-6613(97)01111-X. URL http://www.sciencedirect.com/science/article/pii/S136466139701111X.

A. Lazaridou, E. Bruni, and M. Baroni. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1403–1414, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P14-1132.

A. Lazaridou, G. Dinu, and M. Baroni. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 270–280, Beijing, China, July 2015a. Association for Computational Linguistics. doi: 10.3115/v1/P15-1027. URL https://www.aclweb.org/anthology/P15-1027.

A. Lazaridou, G. Dinu, and M. Baroni. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *ACL*, 2015b.

T. P. B. M. Lazaridou, Angeliki Nghia. Combining language and vision with a multimodal skip-gram model. In *ACL*, 2015.

D. L. Lee, H. Chuang, and K. Seamons. Document ranking and the vector-space model. *IEEE Softw.*, 14(2):67–75, Mar. 1997. ISSN 0740-7459. doi: 10.1109/52.582976. URL http://dx.doi.org/10.1109/52.582976.

S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, CoNLL '11, pages 220–228, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-92-3. URL http://dl.acm.org/citation.cfm?id=2018936.2018962.

Y. Li, J. Zhang, Y. Cheng, K. Huang, and T. Tan. Semantics-guided multi-level rgb-d feature fusion for indoor semantic segmentation. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1262–1266, Sep. 2017. doi: 10.1109/ICIP.2017.8296484.

J. Liao, Y. Yao, L. Yuan, G. Hua, and S. B. Kang. Visual attribute transfer through deep image analogy. *ACM Trans. Graph.*, 36(4):120:1–120:15, July 2017. ISSN 0730-0301. doi: 10.1145/3072959.3073683. URL http://doi.acm.org/10.1145/3072959.3073683.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014a. Springer International Publishing. ISBN 978-3-319-10602-1.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. *Microsoft COCO: Common Objects in Context.* 2014b.

A. Lovett, E. Tomai, K. Forbus, and J. Usher. Solving geometric analogy problems through two-stage analogical mapping. *Cognitive Science*, 33(7):1192–1231, 2009. doi: 10.1111/j.1551-6709.2009.01052.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1551-6709.2009.01052.x.

D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004. ISSN 0920-5691. doi: 10.1023/B:VISI.0000029664.99615.94. URL http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94.

M.-T. Luong, H. Pham, and C. D. Manning. Bilingual word representations with monolingual quality in mind. In *NAACL Workshop*, 2015.

V. Lyzinski, D. L. Sussman, D. E. Fishkind, H. Pao, L. Chen, J. T. Vogelstein, Y. Park, and C. E. Priebe. Spectral clustering for divide-and-conquer graph matching. *Parallel Comput.*, 47(C):70–87, Aug. 2015. ISSN 0167-8191. doi: 10.1016/j.parco.2015.03.004. URL http://dx.doi.org/10.1016/j.parco.2015.03.004.

J. Mao, W. L. Xu, Y. Yang, J. Wang, and A. L. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *CoRR*, abs/1412.6632, 2015.

J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–20, June 2016. doi: 10.1109/CVPR.2016.9.

K. Mardia, J. Kent, and J. Bibby. *Multivariate analysis.* Probability and mathematical statistics. Acad. Press, 1979. ISBN 0124712509. URL http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+02434995X&sourceid=fbw_bibsonomy.

T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In *European Conference on Computer Vision*, 2012.

T. Mikolov, G. Inc, M. View, Q. V. Le, G. Inc, I. Sutskever, and G. Inc. Exploiting similarities among languages for machine translation, 2013a.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA, 2013b. Curran Associates Inc. URL http://dl.acm.org/citation.cfm?id=2999792.2999959.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. 2013c. URL http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013d.

G. A. Miller and W. G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991. doi: 10.1080/01690969108406936. URL https://doi.org/10.1080/01690969108406936.

X. A. Miró, J. Luque, and C. Gracia. Audio-to-text alignment for speech recognition with very limited resources. In *INTERSPEECH*, 2014.

H. B. Mitchell. *Data Fusion Concepts and Ideas; 2nd ed.* Springer, Berlin, 2012. doi: 10.1007/978-3-642-27222-6. URL http://cds.cern.ch/record/1501761.

M. Mitchell, J. Dodge, A. Goyal, K. Yamaguchi, K. Stratos, X. Han, A. Mensch, A. Berg, T. Berg, and H. Daumé III. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756, Avignon, France, Apr. 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E12-1076.

T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195, 2008. ISSN 0036-8075. doi: 10.1126/science.1152876. URL https://science.sciencemag.org/content/320/5880/1191.

D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to Linear Regression Analysis (4th ed.)*. Wiley & Sons, 2006. ISBN 0471754951.

J. Mueller, D. Gifford, and T. Jaakkola. Sequence to better sequence: Continuous revision of combinatorial structures. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2536–2544, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/mueller17a.html.

T. Mukherjee, M. Yamada, and T. Hospedales. Learning unsupervised word translations without adversaries. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 627–632, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D18-1063.

F. Mémoli. Gromov-wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11(4):417–487, 2011. URL http://dblp.uni-trier.de/db/journals/focm/focm11.html#Memoli11.

A. Nemirovski and U. Rothblum. On complexity of matrix scaling. *Linear Algebra and its Applications*, 302-303:435 – 460, 1999. ISSN 0024-3795. doi: https://doi.org/10.1016/S0024-3795(99)00212-8. URL http://www.sciencedirect.com/science/article/pii/S0024379599002128.

J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In

ICML, 2011. URL http://www.icml-2011.org/papers/399_icmlpaper.pdf.

X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1089–1096. Curran Associates, Inc., 2008. URL http://papers.nips.cc/paper/3193-estimating-divergence-functionals-and-the-likelihood-ratio-by-penalized-convex-risk-minimizatio pdf.

M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*, 2014. URL http://arxiv.org/abs/1312.5650.

P. L. Nunez and R. B. Silberstein. On the relationship of synaptic activity to macroscopic measurements: Does co-registration of eeg with fmri make sense? *Brain Topography*, 13(2): 79–96, Dec 2000. ISSN 1573-6792. doi: 10.1023/A:1026683200895. URL https://doi.org/10.1023/A:1026683200895.

A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio, 2016. URL http://arxiv.org/abs/1609.03499. cite arxiv:1609.03499.

V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011.

S. Ouyang, T. Hospedales, Y.-Z. Song, X. Li, C. C. Loy, and X. Wang. A survey on heterogeneous face recognition: Sketch, infra-red, 3d and low-resolution. *Image and Vision Computing*, 56:28 – 48, 2016.

M. Palatucci, D. Pomerleau, G. Hinton, and T. Mitchell. Zero-shot learning with semantic output codes. In *Neural Information Processing Systems (NIPS)*, December 2009a.

M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell. Zero-shot learning with semantic output codes. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1410–1418. Curran Associates, Inc., 2009b. URL http://papers.nips.cc/paper/3650-zero-shot-learning-with-semantic-output-codes.pdf.

K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900. doi: 10.1080/14786440009463897. URL http://dx.doi.org/10.1080/14786440009463897.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://www.aclweb.org/anthology/D14-1162.

G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11 (5-6):355–602, 2019. URL https://arxiv.org/abs/1803.00567.

B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015.

B. A. Plummer, P. Kordas, M. H. Kiapour, S. Zheng, R. Piramuthu, and S. Lazebnik. Conditional image-text embedding networks. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision – ECCV 2018*, pages 258–274, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01258-8.

R. Prabhavalkar, K. Rao, T. N. Sainath, B. Li, L. Johnson, and N. Jaitly. A comparison of sequence-to-sequence models for speech recognition. In F. Lacerda, editor, *INTERSPEECH*, pages 939–943. ISCA, 2017. URL http://dblp.uni-trier.de/db/conf/interspeech/interspeech2017.html#PrabhavalkarRSL17.

N. Quadrianto, L. Song, and A. J. Smola. Kernelized sorting. In *NIPS*, 2009.

M. Radovanovic, A. Nanopoulos, and M. Ivanovic. Nearest neighbors in high-dimensional data: The emergence and influence of hubs. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 865–872, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553485. URL https://doi.org/10.1145/1553374.1553485.

M. Radovanovic, A. Nanopoulos, and M. Ivanovic. Hubs in space: Popular nearest neighbors in high-dimensional data. *J. Mach. Learn. Res.*, 11:2487–2531, Dec. 2010. ISSN 1532-4435.

N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pages 251–260, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-933-6. doi: 10.1145/1873951.1873987. URL http://doi.acm.org/10.1145/1873951.1873987.

S. Reed, Z. Akata, H. Lee, and B. Schiele. Learning deep representations of fine-grained visual descriptions. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2016.

Z. Ren, H. Jin, Z. Lin, C. Fang, and A. Yuille. Joint image-text representation by gaussian visual semantic embedding. In *Proceeding of ACM International Conference on Multimedia (ACM MM)*, 2016.

A. Rohrbach, M. Rohrbach, and B. Schiele. The long-short story of movie description. In J. Gall, P. Gehler, and B. Leibe, editors, *Pattern Recognition*, pages 209–221, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24947-6.

B. Romera-Paredes and P. H. S. Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, 2015.

O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

S. Ruder. A survey of cross-lingual embedding models. *CoRR*, abs/1706.04902, 2017. URL http://arxiv.org/abs/1706.04902.

M. R. Rudolph, F. J. R. Ruiz, S. Mandt, and D. M. Blei. Exponential Family Embeddings, Aug. 2016. URL http://arxiv.org/abs/1608.00778.

G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, Nov. 1975. ISSN 0001-0782. doi: 10.1145/361219.361220. URL http://doi.acm.org/10.1145/361219.361220.

B. Schölkopf, A. J. Smola, and K.-R. Müller. Advances in kernel methods. chapter Kernel Principal Component Analysis, pages 327–352. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-19416-3. URL http://dl.acm.org/citation.cfm?id=299094.299113.

H. Schwenk. Continuous space language models. *Computer Speech and Language*, 21, 2007.

C. E. Shannon. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, Jan. 2001. ISSN 1559-1662. doi: 10.1145/584091.584093. URL http://doi.acm.org/10.1145/584091.584093.

T. Shen, T. Lei, R. Barzilay, and T. Jaakkola. Style transfer from non-parallel text by cross-alignment. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6830–6841. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7259-style-transfer-from-non-parallel-text-by-cross-alignment.pdf.

S. T. Shivappa, M. M. Trivedi, and B. D. Rao. Audiovisual information fusion in human–computer interfaces and intelligent environments: A survey. *Proceedings of the IEEE*, 98(10):1692–1715, Oct 2010. ISSN 0018-9219. doi: 10.1109/JPROC.2010.2057231.

C. Silberer and M. Lapata. Grounded models of semantic representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1423–1433, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2390948.2391110.

C. Silberer and M. Lapata. Learning grounded meaning representations with autoencoders. In *ACL*, 2014.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015a.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015b.

S. L. Smith, D. H. P. Turban, S. Hamblin, and N. Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *CoRR*, abs/1702.03859, 2017. URL http://arxiv.org/abs/1702.03859.

R. Socher, M. Ganjoo, C. D. Manning, and A. Y. Ng. Zero Shot Learning Through Cross-Modal Transfer. In *Advances in Neural Information Processing Systems 26*. 2013a.

R. Socher, M. Ganjoo, C. D. Manning, and A. Y. Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, 2013b.

R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014. doi: 10.1162/tacl_a_00177. URL `https://www.aclweb.org/anthology/Q14-1017`.

P. Soucy and G. W. Mineau. Beyond tfidf weighting for text categorization in the vector space model. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, pages 1130–1135, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc. URL `http://dl.acm.org/citation.cfm?id=1642293.1642474`.

N. Srivastava and R. R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012.

T. Stathaki. *Image Fusion: Algorithms and Applications.* Academic Press, Inc., Orlando, FL, USA, 2008. ISBN 0123725291, 9780123725295.

M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, Dec 2008. ISSN 1572-9052. doi: 10.1007/s10463-008-0197-x. URL `https://doi.org/10.1007/s10463-008-0197-x`.

M. Sugiyama, T. Suzuki, and T. Kanamori. *Density Ratio Estimation in Machine Learning.* Cambridge University Press, New York, NY, USA, 1st edition, 2012. ISBN 0521190177, 9780521190176.

T. Suzuki and M. Sugiyama. Sufficient dimension reduction via squared-loss mutual information estimation. In *AISTATS*, 2010.

T. Suzuki, M. Sugiyama, J. Sese, and T. Kanamori. Approximating mutual information by maximum likelihood density ratio estimation. In Y. Saeys, H. Liu, I. Inza, L. Wehenkel, and Y. V. de Pee, editors, *Proceedings of the Workshop on New Challenges for Feature Selection in Data Mining and Knowledge Discovery at ECML/PKDD 2008*, volume 4 of *Proceedings of Machine Learning Research*, pages 5–20, Antwerp, Belgium, 15 Sep 2008. PMLR. URL `http://proceedings.mlr.press/v4/suzuki08a.html`.

R. Szeliski. *Computer Vision: Algorithms and Applications.* Springer-Verlag, Berlin, Heidelberg, 1st edition, 2010. ISBN 1848829345, 9781848829343.

M. Tapaswi, M. Bäuml, and R. Stiefelhagen. Story-based video retrieval in tv series using plot synopses. In *Proceedings of International Conference on Multimedia Retrieval*, ICMR '14, pages 137:137–137:144, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2782-4. doi: 10.1145/2578726.2578727. URL `http://doi.acm.org/10.1145/2578726.2578727`.

Y. H. Tsai, L. Huang, and R. Salakhutdinov. Learning robust visual-semantic embeddings. In *ICCV*, 2017.

L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3): 279–311, Sep 1966. ISSN 1860-0980. doi: 10.1007/BF02289464. URL `https://doi.org/10.1007/BF02289464`.

M. Turk. Multimodal interaction: A review. *Pattern Recognition Letters*, 36:189 – 195,

2014. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2013.07.003. URL http://www.sciencedirect.com/science/article/pii/S0167865513002584.

P. D. Turney. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073153. URL https://doi.org/10.3115/1073083.1073153.

P. D. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141–188, Jan. 2010. ISSN 1076-9757. URL http://dl.acm.org/citation.cfm?id=1861751.1861756.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. Translating videos to natural language using deep recurrent neural networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1494–1504, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1173. URL https://www.aclweb.org/anthology/N15-1173.

L. Vilnis and A. McCallum. Word representations via gaussian embedding. In *ICLR*, 2015.

O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3156–3164. IEEE, 2015.

G. Vivone, L. Alparone, J. Chanussot, M. Dalla Mura, A. Garzelli, G. A. Licciardi, R. Restaino, and L. Wald. A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5):2565–2586, May 2015. ISSN 0196-2892. doi: 10.1109/TGRS.2014.2361734.

I. Vulic and A. Korhonen. On the role of seed lexicons in learning bilingual word embeddings. In *ACL*, 2016.

I. Vulic and M. Moens. Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In *HLT-NAACL*, 2013.

L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016.

L. Wang, Y. Li, and S. Lazebnik. Learning two-branch neural networks for image-text matching tasks. *CoRR*, abs/1704.03470, 2017.

W. Wang, R. Arora, K. Livescu, and J. A. Bilmes. Unsupervised learning of acoustic features via deep canonical correlation analysis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4590–4594, April 2015. doi: 10.1109/ICASSP.2015.7178840.

D. Weiss, C. Alberti, M. Collins, and S. Petrov. Structured training for neural network transition-based parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 323–333, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1032. URL `https://www.aclweb.org/anthology/P15-1032`.

L. Wittgenstein. *Philosophical Investigations.* Wiley-Blackwell, 1953.

C. Xing, D. Wang, C. Liu, and Y. Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1104. URL `https://www.aclweb.org/anthology/N15-1104`.

K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 2048–2057. JMLR.org, 2015. URL `http://dl.acm.org/citation.cfm?id=3045118.3045336`.

R. Xu, Y. Yang, N. Otani, and Y. Wu. Unsupervised cross-lingual transfer of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2465–2474, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1268. URL `https://www.aclweb.org/anthology/D18-1268`.

M. Yamada and M. Sugiyama. Cross-domain object matching with model selection. In *AISTATS*, 2011.

M. Yamada, L. Sigal, M. Raptis, M. Toyoda, Y. Chang, and M. Sugiyama. Cross-domain matching with squared-loss mutual information. *IEEE TPAMI*, 37(9):1764–1776, 2015.

F. Yan and K. Mikolajczyk. Deep correlation for matching images and text. In *CVPR*, 2015.

Y. Yang, C. L. Teo, H. Daumé, III, and Y. Aloimonos. Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 444–454, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-937284-11-4. URL `http://dl.acm.org/citation.cfm?id=2145432.2145484`.

L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4507–4515, Dec 2015. doi: 10.1109/ICCV.2015.512.

Z. Yi, H. Zhang, P. Tan, and M. Gong. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. In *Proc. of ICCV*, 2017.

N. Yokoya, T. Yairi, and A. Iwasaki. Hyperspectral, multispectral, and panchromatic data fusion based on coupled non-negative matrix factorization. In *2011 3rd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pages 1–4, June 2011. doi: 10.1109/WHISPERS.2011.6080924.

P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014. ISSN 2307-387X. URL `https://transacl.org/ojs/index.php/tacl/article/view/229`.

S. Yuan, K. Bai, L. Chen, Y. Zhang, C. Tao, C. Li, G. Wang, R. Henao, and L. Carin. Weakly supervised cross-domain alignment with optimal transport, 2020.

A. L. Yuille and A. Rangarajan. The concave-convex procedure (cccp). In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 1033–1040. MIT Press, 2002. URL `http://papers.nips.cc/paper/2125-the-concave-convex-procedure-cccp.pdf`.

D. Zeng, Y. Yu, and K. Oyama. Audio-visual embedding for cross-modal music video retrieval through supervised deep cca. In *2018 IEEE International Symposium on Multimedia (ISM)*, pages 143–150, Dec 2018. doi: 10.1109/ISM.2018.00-21.

M. Zhang, Y. Liu, H. Luan, and M. Sun. Adversarial training for unsupervised bilingual lexicon induction. In *ACL*, 2017a.

M. Zhang, Y. Liu, H. Luan, and M. Sun. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *EMNLP*, 2017b.

M. Zhang, Y. Liu, H. Luan, and M. Sun. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1970, Vancouver, Canada, July 2017c. Association for Computational Linguistics. doi: 10.18653/v1/P17-1179. URL `https://www.aclweb.org/anthology/P17-1179`.

M. Zhang, Y. Liu, H. Luan, and M. Sun. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, Copenhagen, Denmark, Sept. 2017d. Association for Computational Linguistics. doi: 10.18653/v1/D17-1207. URL `https://www.aclweb.org/anthology/D17-1207`.

S. Zhao, J. Song, and S. Ermon. Infovae: Information maximizing variational autoencoders. *CoRR*, abs/1706.02262, 2017. URL `http://arxiv.org/abs/1706.02262`.

J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, Oct 2017. doi: 10.1109/ICCV.2017.244.

J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 19–27, Washington, DC, USA, 2015. IEEE Computer Society. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.11. URL `http://dx.doi.org/10.1109/ICCV.2015.11`.

W. Y. Zou, R. Socher, D. Cer, and C. D. Manning. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398, Seattle, Washington, USA, Oct. 2013. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/D13-1141`.