Cross-View Learning

Li Zhang

Submitted in partial fulfilment of the requirement for the degree of *Doctor of Philosophy*

School of Electronic Engineering and Computer Science

Queen Mary University of London

16 May 2018

Cross-View Learning

Li Zhang

Abstract

Key to achieving more efficient machine intelligence is the capability to analysing and understanding data across different views – which can be camera views or modality views (such as *visual* and *textual*). One generic learning paradigm for automated understanding data from different views called *cross-view learning* which includes cross-view matching, cross-view fusion and cross-view generation. Specifically, this thesis investigates two of them, cross-view matching and cross-view generation, by developing new methods for addressing the following specific computer vision problems.

The first problem is *cross-view matching for person re-identification* which a person is captured by multiple non-overlapping camera views, the objective is to match him/her across views among a large number of imposters. Typically a person's appearance is represented using features of thousands of dimensions, whilst only hundreds of training samples are available due to the difficulties in collecting matched training samples. With the number of training samples much smaller than the feature dimension, the existing methods thus face the classic *small sample size* (SSS) problem and have to resort to dimensionality reduction techniques and/or matrix regularisation, which lead to loss of discriminative power for cross-view matching. To that end, this thesis proposes to overcome the SSS problem in subspace learning by matching cross-view data in a discriminative null space of the training data.

The second problem is *cross-view matching for zero-shot learning* where data are drawn from different modalities each for a different view (e.g. visual or textual), *versus* single-modal data considered in the first problem. This is inherently more challenging as the gap between different views becomes larger. Specifically, the zero-shot learning problem can be solved if the visual representation/view of the data (object) and its textual view are matched. Moreover, it requires learning a joint embedding space where different view data can be projected to for nearest neighbour search. This thesis argues that the key to make zero-shot learning models succeed is to choose the right embedding space. Different from most existing zero-shot learning models utilising a textual or an intermediate space as the embedding space for achieving cross-view matching, the proposed method uniquely explores the visual space as the embedding space. This thesis finds that in the visual space, the subsequent nearest neighbour search would suffer much less from the *hubness* problem and thus become more effective. Moreover, a natural mechanism for multiple textual modalities optimised jointly in an end-to-end manner in this model demonstrates significant advantages over existing methods.

The last problem is *cross-view generation for image captioning* which aims to automatically generate *textual sentences* from *visual images*. Most existing image captioning studies are limited to investigate variants of deep learning-based image encoders, improving the inputs for the subsequent deep sentence decoders. Existing methods have two limitations: (i) They are trained to maximise the likelihood of each ground-truth word given the previous ground-truth words and

the image, termed Teacher-Forcing. This strategy may cause a mismatch between training and testing since at test-time the model uses the previously generated words from the model distribution to predict the next word. This exposure bias can result in error accumulation in sentence generation during test time, since the model has never been exposed to its own predictions. (ii) The training supervision metric, such as the widely used cross entropy loss, is different from the evaluation metrics at test time. In other words, the model is not directly optimised towards the task expectation. This learned model is therefore suboptimal. One main underlying reason responsible is that the evaluation metrics are non-differentiable and therefore much harder to be optimised against. This thesis overcomes the problems as above by exploring the reinforcement learning idea. Specifically, a novel actor-critic based learning approach is formulated to directly maximise the reward - the actual Natural Language Processing quality metrics of interest. As compared to existing reinforcement learning based captioning models, the new method has the unique advantage of a per-token advantage and value computation is enabled leading to better model training.

Declaration

I, Li Zhang, hereby declare that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third partys copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Chapter 3

1. **L. Zhang**, T. Xiang and S. Gong. *Learning a Discriminative Null Space for Person Re-identification*. In Proc. IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, June 2016. **(CVPR)**

Chapter 4

1. **L. Zhang**, T. Xiang and S. Gong. *Learning a Deep Embedding Model for Zero-Shot Learning*. In Proc. IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii, USA, July 2017. **(CVPR)**

Chapter 5

1. **L. Zhang**, F. Sung, F. Liu, T. Xiang and S. Gong, Yongxin Yang, TM. Hospedales. *Actor-Critic Sequence Training for Image Captioning*. Neural Information Processing Systems **(NIPS)**, workshop on Visually-Grounded Interaction and Language, Long Beach, California, USA, December 2017.

Acknowledgements

There are so many people who have helped me in one way or another and to whom I owe deep gratitude for the wonderful three and half years of my PhD life.

First of all, I must express my sincere gratitude to my supervisor Tony Xiang, who through amounts of emails and weekly meetings over three years, turned me from a naive student who didn't know SVM and pooling to a researcher who has some ambitious to work in the academic. When I wrote a terrible paper draft he patiently pointed out those mistakes from character to period. When I came up some brave ideas he encouraged and push me stepping out my comfort zone. I sincerely appreciate all his support, generosity, and, most important, patience.

Second, it has been a pleasure to learn from Sean Gong, my second supervisor, for his high level supervision and witty criticisms to shape me how to think. I was also fortunate enough to work with Flood Sung in my last year PhD, who has been my main source for learning reinforcement learning and open the door of robotic vision for me. A special thank to Yongxin Yang who not only influence me with his productive research and innovative thoughts but also lend me the lovely flat to live in so that I can finish my thesis in my last stage of PhD life. The most wonderful thing about my years in Queen Mary is the friends that I was so incredibly lucky to make here. My warm appreciation also goes to them: Yi-Zhe Song, Miles Hansard, Ioannis Patras, Eddy Zhu, Patrick Shi, Hanxiao Wang, Yanwei Fu, Zhenyong Fu, Elyor Kodirov, Alex Xu, Yi Li, Feng Liu, Kiya Wang, Grace Dong, Xiaobin Chang, Hang Su, Xu Lan, Wei Li, Ying Zhang, Zhiyi Cheng, Yanbei Chen, Qian Yu, Jifei Song, Da Li, Kaiyue Pang, Conghui Hu, Umar Muhammad, Kaiyang Zhou, Tianyuan Yu, Xiangyu Kong, Shuxin Ouyang, Yaowei Wang.

Third, I want to thanks my parents for their consistent encouragement and enduring love. They trust and support me for every decision I made. They are the best parents I could have. Most importantly, I would like to thank my wife Shudi Qin. It would never have been possible to have completed this work without her love.

Last but not least, my lovely daughter Bailu who was born rightly after CVPR2017 deadline, brings lots of fun to our family. I love her and she always teaches me how to be a good Dad.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Scope of the Thesis

When we look around this world, the amount of information deliver to us with its own manifestation at any given moment is enormous. For example, visual data generated by the rapid expansion of large-scale distributed multi-camera systems, with same object captured by different cameras as *views*; Hundreds of hours of videos are uploaded to YouTube every minute, which appear in multiple modalities, namely visual, audio and text *views*; A large number of bilingual news are reported every day, with the description in each language as a *view*. Our brain manage to process these *views* to create a complete, coherent and comprehensible universe that is always rich and vivid. Moreover, human find it easy to accomplish a wide variety of tasks that involve complex visual recognition and scene understanding from different viewpoints/pose/illumination/background, tasks that involve communication in image/text/audio/video and tasks that combine translation between these different modalities. This is why customer has no problem to recognise the merchandise exhibited on Amazon webpage after viewing the photos took from different camera viewpoints/pose/background. This is also why children can better gain knowledge from ”children’s picture book” – with picture and text description on it (Figure 1.1).

In order to achieve a better and efficient machine intelligence for well understanding our rich and vivid visual world, capability to learning from data across different views is essential. This is also one reason why achieving computer vision driven machine intelligence is so difficult: the

(a) Cross camera view

(b) Cross modality

Figure 1.1: Many cross-view examples in real-world: (a) Laptop displayed under different camera views; (b) Children's picture books contain *visual* image and *textual* language.

pattern/representation of an object is not uniquely determined, it is often manifest with different *views*. Therefore, automatically and efficiently understanding data across different views poses a major research challenge. One generic learning paradigm for that called *cross-view learning*, which includes:

Cross-View Matching. The object can be captured by different views – camera views or modalities (such as *visual* and *textual*), which bring about a great challenge of matching them. This kind of problem is generally called as cross-view matching. Usually, the representation of objects from different views are significantly different from each other, and the large view discrepancy makes it quite challenging to directly compare them based on the feature representation. Substantial efforts have been dedicated to eliminate the view discrepancy by learning a mapping function or a embedding space across views.

Cross-View Generation. Given a novel instance from one view, the objective of cross-view generation is to automatically generate raw data of another view(s), by learning a mapping function across views. Data from both views are available during training, but only one view is available at test time. Usually, Generative Adversarial Network (Goodfellow et al., 2014) or a encoder-decoder network are used for such cross-view generation task. Specifically, when the view refers to camera view, images from different viewpoints are synthesised (Tran et al., 2017; Huang et al., 2017; Ji et al., 2017; Yang et al., 2015); when the view is modality, data from different modality or domain are generated (Vinyals et al., 2015; Isola et al., 2016).

Cross-View Fusion. The term *fusion* means in general an approach to extraction of information acquired in several views. The goal of cross-view fusion is to integrate complementary different view information into one new view containing information the quality of which cannot

be achieved otherwise. The term quality, its meaning and measurement depend on the particular application (e.g. detection, tracking). Specifically, when the view refers to camera view, which means the images of the same modality, taken from different view angles, the goal of cross-view fusion is to supply complementary information from different views; when the view represents modality, for example, one image with high spatial resolution, the other one with low spatial but higher spectral resolution, the goal of cross-view fusion is to get an image with high spatial and spectral resolution (Dian et al., 2017).

Views. When the *view* refers to camera view, object can be captured by different camera views, cross-views matching problem is then investigated in Chapter 3; When the *view* refers to modality, two views are consider: visual and textual. More particular, only attribute and natural language are considered as textual view in this thesis. Natural language can be continue split into: (1) the online free articles (e.g. google news, Wikipedia documents) and, (2) caption descriptions corresponding to each visual image. They are three in total for textual view in this thesis.

1. Attributes describe parts (has nose), shape (cylindrical), colour (brown), and materials (furry). They (can be learned from annotations or pre-defined by human experts) allow us to describe objects and to identify them based on textual descriptions, i.e. attribute can be seen as the unstructured textual view (Farhadi et al., 2009; Ferrari and Zisserman, 2007; Parikh and Grauman, 2011; Yan et al., 2017);
2. A skip-gram language model (Mikolov et al., 2013a,b) trained on a corpus of online free google news or Wikipedia documents which can be used to extract fixed dimension word vectors according to the name of each visual class needed;
3. Sentence descriptions/captions corresponding to each image (Lin et al., 2014; Reed et al., 2016a) have started to gain popularity recently. A neural language model (e.g. LSTM) is required to output a vector representation of the description. Both Chapter 4 and 5 consider text/captions as the textual view.

We can see that all above three learned/pre-defined representation are semantic meaningful. In particular, in Chapter 4, these vector can be termed as *semantic vector*. The semantic vector represents each visual class name is termed as a class *prototype*. Therefore, *semantic space* refers to a high dimensional vector space where visual classes are usually semantically related in. The knowledge from seen classes can be transferred to unseen classes in this semantic space. More details are in Section 4.2. While in Chapter 5, *semantic* usually describes the high level visual

representation that is semantic meaningful against to the label space (e.g. the feature representation extracted from a trained Deep Neural Networks for the classification task).

Relationship between cross-view matching and cross-view generation. Specifically, this thesis investigates two of them: cross-view matching and cross-view generation. Given a collection of data with view X and view Y (whether the data are paired or unpaired (Zhu et al., 2017a)), for both cross-view matching and cross-view generation at modelling stage, the goal is to learning a mapping function $M : X \rightarrow Y$ or finding an embedding space that both views can be projected to. Then novel instances or objects from novel classes are given for inference for both problems, the only difference is that both views are given for cross-view matching while only one view is given for inference for cross-view generation.

Moreover, the two investigated problems covering widely studied cross-view real-world applications such as person re-identification, zero-shot learning and image captioning. All these problems require strong and robust cross-view learning algorithms for building corresponding models to realise automatically cross-view data analysis at large scale.

Person Re-identification. The first application is *person re-identification* (Re-ID) which refers to the problem of visually matching already detected individual or group of people across non-overlapping cameras views distributed at diverse physical locations and times (Vezzani et al., 2013; Gong et al., 2014a). In particular, for surveillance systems performed over space and time, an individual disappearing from one view would need to be matched in one or more other views at different physical locations over a period of time, and be differentiated from numerous visually similar but different candidates in those views. For most of today's intelligent surveillance systems, re-identification has become a fundamental technology which paves the way for numerous higher level and more complex systems. For example, it contributes as a critical component for a multi-camera tracking or forensic search system, which allow government agencies to fast locate suspicious criminals, and therefore prevent terrorism threatening social infrastructure and civilian safety and security; The re-identification of a group of people collectively provides valuable intelligence for crowd movement/behaviour analysis, which facilitates public spaces like airports or shopping malls to conduct better crowd control practices or develop more profitable retail floor plans; Re-identification techniques could also be integrated into smart home automation platforms, so as to enable functionalities such as elderly/baby monitoring, intrusion detection and burglary alarming (Wang et al., 2017a). To this end, this thesis focuses on addressing this

real-world cross-view person image matching problem.

Zero-Shot Learning. A recent trend in developing visual recognition models is to scale up the number of object categories. However, most existing recognition models are based on supervised learning and require a large amount (at least 100s) of training samples to be collected and annotated for each object class to capture its intra-class appearance variations (Deng et al., 2009). This severely limits their scalability – collecting daily objects such as chair is easier, but many other categories are rare (*e.g.*, a newly identified specie of beetle on a remote pacific island). None of these models can work with few or even no training samples for a given class. In contrast, human perform visual recognition effortlessly, and instantaneously. Importantly, human are great at recognising a new object without seeing any visual samples by just knowing the description of it, leveraging similarities between the textual description of the new object and previously learned concepts. For example, a child would have no problem recognising a zebra if she has seen horses before and also read elsewhere that a zebra is a horse but with black-and-white stripes on it. Humans can easily generalise the knowledge learned in the past to recognise the classes never seen before. Very recently, researchers in machine learning and computer vision community have started to propose approaches that imitate the humans recognition ability, and this is known as zero-shot learning (ZSL). Specially, the zero-shot learning is a cross-view matching problem which can be solved if the visual representation/view of the data (object) and its textual view are matched. For example, it might allow a computer to read on the Internet that a Persian cat is ”a long-haired breed of cat characterized by its round face, short muzzle and large, striking eyes” and recognise such concept in visual data based on the description. However, zero-shot learning is inherently more challenging as the gap between different views becomes larger. To this end, this thesis studies and develops novel approaches for addressing zero-shot learning problem across visual and textual views.

Image Captioning. Another application where cross-view learning is of great use is *image captioning*. At a high-level, image captioning aims to automatically describe the visual content of an image in natural language instead of merely assigning it a category. It is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. Different from the previous two cross-view matching applications, image captioning aims to provide a means for learning a generative map from visual images to human-level textual language. Therefore, it’s an application of cross-view generation. Similar to how traditional computer vision

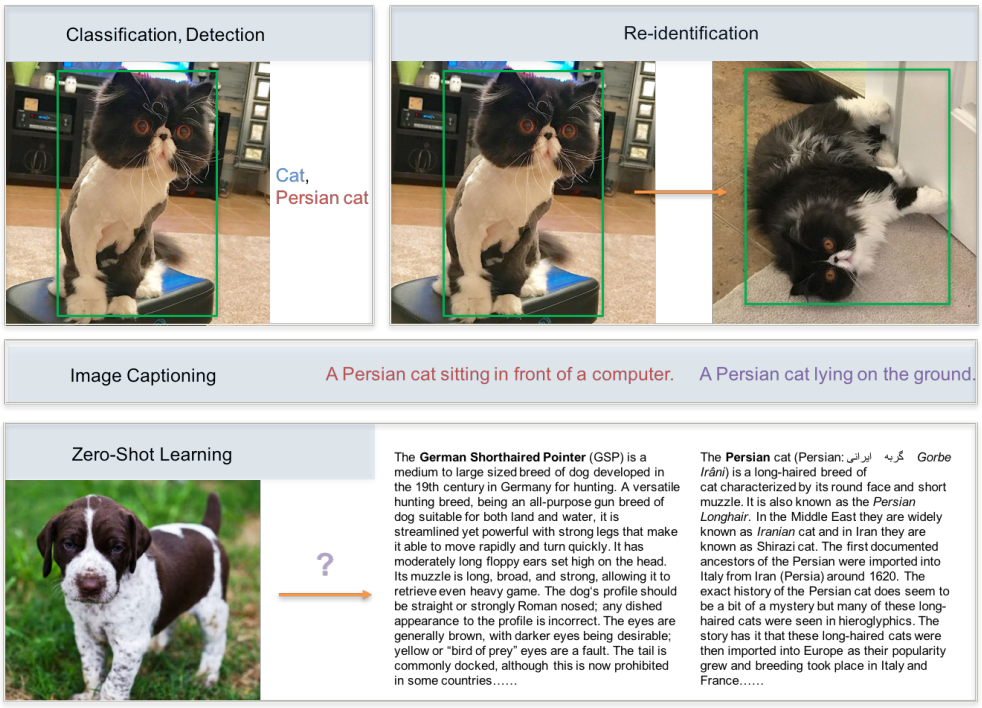


Figure 1.2: Cross-view learning problems beyond visual recognition. **Up-left:** classification and detection of visual object; **Up-right:** re-identification of object under different camera views; **Middle:** describe an image with a sentence; **Down:** recognise an unseen object.

seeks to make the world more accessible and understandable for computers, image captioning has the potential to make our world more accessible and understandable for us humans. It can serve as a tour guide, and can even serve as a visual aid for daily life of visually impaired people. To that end, a robust cross-view learning model is required to not only generate a rich visual representation but also have a strong generation capability for recovering the high-level visual representation in the form of textual language.

Long-term motivations. Given all the valuable problems and practical applications, following aspirations are motivated:

1. Cross-view learning is a step towards a long-term goal of building up intelligent machines as they require large amounts of different views data which are the channels by which the knowledge of the world can be accessed;
2. Recent successes in applying deep neural networks to computer vision and other domain such as natural language processing tasks have inspired AI researchers to explore new research opportunities at the intersection of these previously separate domains. Therefore, it is critical that we develop techniques that can relate information across different views instead of processing each independently;

3. The ultimate goal of computer vision driven artificial intelligence research is to make machines see and understand the rich and vivid visual world and endow them with the ability to communicate with us in many ways. Taking Figure 1.2 as an example, solving cross-viewing learning problems will expand the boundary of the artificial intelligence, not only can classify and detect ”Persian cat”, but also be able to recognise new breed ”German Shorthaired Pointer”, re-identify which images/videos contain same cats and describe what the cat is doing with human language.

1.2 Challenges and Limitations

Cross-view learning is hard. For making sense of the vast quantity of data generated from the rapid expansion of large-scale distributed *views*, automated cross-view learning is essential. However, it poses a number of *challenges*: (1) To a computer vision based intelligence system, image is represented as a large array of numbers indicating the brightness at any position. An ordinary image might have a few million of these pixels and an artificial intelligence agent must transform these patterns of brightness values into high-level concepts such as a ”cat”. Moreover, when the same cat (Figure 1.2) seen under different camera viewpoint and distance, or in a different pose, featuring different static and dynamic backgrounds under different lighting conditions, degrees of occlusion and other view-specific variable, its pattern of brightness values could be completely different. (2) When the *views* refer to different modalities such as *visual* and *textual*, the challenge become more severe, the model has to balance an understanding of both visual cues and textual views. For example, a natural language description – the textual view – of a ”cat” image such as ”A Persian cat sitting in front of a computer” will be represented in the computer as a sequence of integers indicating the index of each word in a vocabulary (e.g. ”A Persian cat sitting in front of a computer” might be [3, 1742, 246, 33, 20, 198, 69, 3, 498]). Therefore, the very natural task of pointing out and naming different parts of the image in fact involves a complex pattern recognition process of identifying salient subsets of a grid of a few million brightness values and annotating them with sequences of integers (Karpathy, 2016).

Encouraging progress but not enough. Despite the difficulty of these tasks, the computer vision community have recently witnessed rapid progress in the area of visual recognition. In particular, the state of the art image recognition models based on deep convolutional neural networks (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; Szegedy et al., 2015; He et al.,

2016) have become capable of distinguishing thousands of visual categories at accuracies comparable to humans, or even surpassing them in some fine-grained categories such as breeds of cats (Russakovsky et al., 2015). These advances can potentially eliminate the large view discrepancy in cross-view learning problems by simply extracting features from the learned convolutional neural networks. However, inherent challenges lie in the aforementioned cross-view learning problems still remains and innovation of more robust and advanced computer vision algorithms are motivated to propose in this thesis.

Cross-View Matching for Person Re-identification Despite the best efforts from the computer vision researchers, cross-view person image matching or person re-identification remains a largely unsolved problem. This is because that a persons appearance often undergoes dramatic changes across camera views due to changes in view angle, body pose, illumination and background clutter. Furthermore, since people are mainly distinguishable by their clothing under a surveillance setting, many passers-by can be easily confused with the target person because they wear similar clothes.

Existing approaches focus on developing discriminative feature representations that are robust against the view/pose/illumination/background changes (Gray and Tao, 2008; Yang et al., 2014; Farenzena et al., 2010; Kviatkovsky et al., 2013; Ma et al., 2012; Zhao et al., 2014; Liao et al., 2015), or learning a distance metric (Gray and Tao, 2008; Koestinger et al., 2012; Prosser et al., 2010; Zheng et al., 2013; Mignon and Jurie, 2012; Tao et al., 2013; Pedagadi et al., 2013; Li et al., 2013; Zhao et al., 2013b,a; Xiong et al., 2014; Ma et al., 2014; Lisanti et al., 2014a; Liao et al., 2015), or both jointly (Li et al., 2014; Ahmed et al., 2015). Among them, the distance metric learning methods are most popular and are the focus of this thesis. Given any feature representation and a set of training data consisting of matching image pairs across camera views, the objective is to learn the optimal distance metric that gives small values to images of the same person and large values for those of different people. Distance metric learning has been extensively studied in machine learning (Yang and Jin, 2006), and existing metric learning methods employed for Re-ID are either originated elsewhere or extensions of existing methods with modifications to address the additional challenges arising from the Re-ID task. Although they have been shown to be effective in improving the existing Re-ID benchmarks over the past five years, all these models are still limited by some of classical problems in model learning.

Small sample size problem Specifically, a key challenge for distance metric learning when applied to person re-identification is the *small sample size* (SSS) problem (Chen et al., 2000). Specifically, to capture rich person appearance whilst being robust against those condition changes mentioned above, the feature representations used by most recent Re-ID works are of high dimension – typically in the order of thousands or tens of thousands. In contrast, the number of training samples is typically small, normally in hundreds. This is because that collecting training samples of matched person pairs across views is labour intensive and tedious. As a result the sample size is much smaller (often in an order of magnitude) than the feature dimension, a problem known as the SSS problem. Metric learning methods suffer from the SSS problem because they essentially aim to minimise the intra-class (intra-person) variance (distance), whilst maximising the inter-class (inter-person) variance (distance). With a small sample size, the within-class scatter matrix becomes singular (Chen et al., 2000); to avoid it, unsupervised dimensionality reduction or regularisation are required. This in turn makes the learned distance metric sub-optimal and less discriminative (Chen et al., 2000; Zheng et al., 2005; Guo et al., 2006).

Cross-View Matching for Zero-Shot Learning Compare to the single-modal for each view in Re-ID problem, zero-shot learning (ZSL) models need to deal with the case that each *view* is associated with different modality. Specifically, ZSL models rely on learning a joint embedding space where both textual description of object classes and visual representation of object images can be projected to for cross-view matching. Usually a labelled training set of *seen classes* and the knowledge about how an *unseen class* is semantically related to the seen classes are available for model training. Seen and unseen classes are usually related in a high dimensional vector space, called *semantic space*, where the knowledge from seen classes can be transferred to unseen classes. The semantic spaces used by most early works are based on attributes (Farhadi et al., 2009; Ferrari and Zisserman, 2007; Parikh and Grauman, 2011). Given a defined attribute ontology, each class name can be represented by an attribute vector. Generally, attribute can be seen as the unstructured textual view (Yan et al., 2017). More recently, word vector space (Socher et al., 2013; Frome et al., 2013) and sentence descriptions/captions (Reed et al., 2016a) both from textual views have started to gain popularity. With the former, the class names are projected into a word vector space so that different classes can be compared, whilst with the latter, a neural language model is required to provide a vector representation of the description. The semantic vector which represents each class name is termed as a class *prototype*. With the semantic space

and a visual feature representation of image content, ZSL is typically solved in two steps: (1) A joint embedding space is learned where both the textual vectors (prototypes) and the visual feature vectors can be projected to; and (2) nearest neighbour (NN) search is performed in this embedding space to match the projection of an image feature vector against that of an unseen class prototype.

Hubness problem Despite the success of deep neural networks that learn an end-to-end model across *visual* and *textual* views in other vision problems (e.g. image captioning), very few deep ZSL model exists (Lei Ba et al., 2015; Frome et al., 2013; Socher et al., 2013; Yang and Hospedales, 2015; Reed et al., 2016a) and they show little advantage over ZSL models (Fu et al., 2014; Fu and Sigal, 2016; Akata et al., 2015; Bucher et al., 2016; Romera-Paredes and Torr, 2015; Zhang and Saligrama, 2015; Lampert et al., 2014) that utilise deep feature representations but do not learn an end-to-end embedding. The main problem prevents the deep ZSL model to succeed is the *hubness* problem (Radovanović et al., 2010). Existing models, regardless whether they are deep or non-deep, choose either the semantic space (Lampert et al., 2014; Fu and Sigal, 2016; Socher et al., 2013; Frome et al., 2013) or an intermediate embedding space (Lei Ba et al., 2015; Akata et al., 2015; Romera-Paredes and Torr, 2015; Fu et al., 2014) as the embedding space. However, since the embedding space is of high dimension and nearest neighbour search is to be performed there, the hubness problem is inevitable, that is, a few unseen class prototypes will become the nearest neighbours of many data points, i.e., hubs. Using the semantic space as the embedding space means that the visual feature vectors need to be projected into the semantic space which will shrink the variance of the projected data points and thus aggravate the hubness problem (Radovanović et al., 2010; Dinu et al., 2014).

Cross-View Generation for Image Captioning Apart from cross-view matching, this thesis discusses the problem of cross-view generation for image captioning which is introduced in Section 1.1, Image captioning model learns to capture relevant *semantic information* from visual image data by training on large numbers of visual-textual pairs. Specifically, each image will be encoded by a deep convolutional neural network into a visual vector representation. A language generating RNN, or recurrent neural network, will then decode that visual representation sequentially into a textual/natural language description.

Limitations in supervised learning based image captioning Most existing image captioning studies investigate variants of deep learning-based image encoders, improving the inputs for the sub-

sequent deep sentence decoders. Due to the supervised training strategy, they have two limitations: (i) They are trained to maximise the likelihood of each ground-truth word given the previous ground-truth words and the image, termed Teacher-Forcing. This strategy may cause a mismatch between training and testing since at test-time the model uses the previously generated words from the model distribution to predict the next word. This exposure bias can result in error accumulation in sentence generation during test time, since the model has never been exposed to its own predictions. (ii) The training supervision metric, such as the widely used cross entropy loss, is different from the evaluation metrics at test time. In other words, the model is not directly optimised towards the task expectation. This learned model is therefore suboptimal. One main underlying reason responsible is that the evaluation metrics are non-differentiable and therefore much harder to be optimised against.

1.3 Approaches

Motivated by the challenges and problems lie in cross-view learning this thesis proposes following robust cross-view learning algorithm to address them.

Null Space Learning for Cross-View Matching To overcome the challenges in cross-view matching for re-id, this study argues that the SSS problem in person Re-ID distance metric learning can be best solved by learning a discriminative null space of the training data. In particular, instead of minimising the within-class variance, data points of the same classes are *collapsed*, by a transform, into a single point in a new space (see Fig. 3.1). By keeping the between-class variance non-zero, this automatically maximises the Fisher discriminative criterion and results in a discriminative subspace. The null space method, also known as the null Foley-Sammon transfer (NFST) (Guo et al., 2006) is specifically designed for the small sample case, with rigorous theoretical proof on the resulting subspace dimension. Importantly, it has a closed-form solution, no parameter to tune, requires no pre-precessing steps to reduce the feature dimension, and can be computed efficiently. Furthermore, to deal with the non-linearity of the person's appearance, a kernel version can be developed easily to further boost the matching performance within the null space. It therefore offers a perfect solution to the challenging person Re-ID problem. In addition to formulating the NSFT model as a fully supervised model to solve the person Re-ID problem, semi-supervised setting is extended to further alleviate the effects of the SSS problem by exploiting unlabelled data abundant in Re-ID applications.

Deep Embedding Learning for Cross-View Matching To address the *hubness* problem in zero-shot learning, a deep cross-view matching framework is presented, which makes it capable of dealing with the case that each view is associated with different modality. Specifically, a novel deep neural network based embedding model for ZSL is proposed which differs from existing models in that: To alleviate the hubness problem, visual space is adopted as the embedding space instead of the semantic space or an intermediate space. The resulting projection direction is from the textual view to visual view. Such a direction is opposite to the one adopted by most existing models. A theoretical analysis and some intuitive visualisations are provided to explain why this would help to counter the hubness problem.

Reinforcement Learning for Cross-View Generation To address the two limitations in cross-view generation for image captioning, the main idea is to formulate a reinforcement learning based framework to improve the quality of generated text sentence. In this way, the gradient of the expected reward can be optimised by sampling from the model during training, thus avoiding the train-test mismatch; the relevant test-time metrics such as CIDEr can be directly optimised, by treating them as reward in a reinforcement learning context. Specifically, an actor-critic model is proposed which consists of a policy network (actor) and value network (critic). The actor is trained to predict the caption as a sequential decision problem given the image, where the sequence of actions correspond to tokens. The critic predicts the value of each state (image and sequence of actions so far), which is defined as the expected task-specific reward (language metric score) that the network will receive if it outputs the current token and continues to sample outputs according to its probability distribution.

1.4 Contributions

The contributions made in this thesis are summarised below:

1. A distance metric learning model (Zhang et al., 2016) is proposed to overcome the *small sample size* challenge in Re-ID by matching people in a discriminative null space of the training data. In this null space, images of the same person are collapsed into a single point thus minimising the intra-class scatter to the extreme and maximising the relative inter-class separation simultaneously. Importantly, it has a fixed dimension, a closed-form solution and is very efficient to compute. Moreover, A novel semi-supervised learning method is developed in the null space to exploit the abundant unlabelled data to further

alleviate the effects of the SSS problem. The details are given in Chapter 3.

2. The cross-view matching framework is extended to the case of different modalities each for a different view (e.g. visual or textual), *versus* single-modal data considered in the Chapter 3. Specifically, a novel deep embedding model to match visual view/representation of the data (object) and its textual view for zero-shot visual recognition is proposed (Zhang et al., 2017b). Further, a natural mechanism for multiple textual modalities optimised jointly in an end-to-end manner in this model demonstrates significant advantages over existing methods and state-of-the-art zero-shot learning performance achieved. The model is described in Chapter 4.
3. A novel actor-critic sequence training framework for the image captioning is formulated for effectively describing image content with human-level language (Zhang et al., 2017a). In this way, the non-differentiable quality metrics of interest can be directly optimised. The proposed approach exploits the shorter episodes and ameliorates the high dimensional action space by formulating a per-token advantage and value computation strategy in this novel reinforcement learning based caption generator. The framework is discussed in Chapter 5.

1.5 Organisation of Thesis

This thesis is organised as follows, with all chapters structured as shown in Figure 1.3.

Chapter 2 presents a review on various existing cross-view learning strategies, including cross-view matching, cross-view generation, and cross-view fusion while providing further motivations for the proposed approaches of this thesis.

Chapter 3 proposes to overcome the small sample size (SSS) problem in person re-identification (Re-ID) by matching cross camera view data in a discriminative null space of the training data. Extensive experiments are conducted on five widely used person re-identification benchmarks to evaluate the advantages of such a simple approach by comparing most contemporary methods.

Chapter 4 argues the key to make deep ZSL models succeed is to choose the right embedding space and proposes a novel deep embedding model for zero-shot learning. Extensive experiments on four benchmarks show that our the proposed model beats the state-of-the-art alternatives, often by a clear margin.

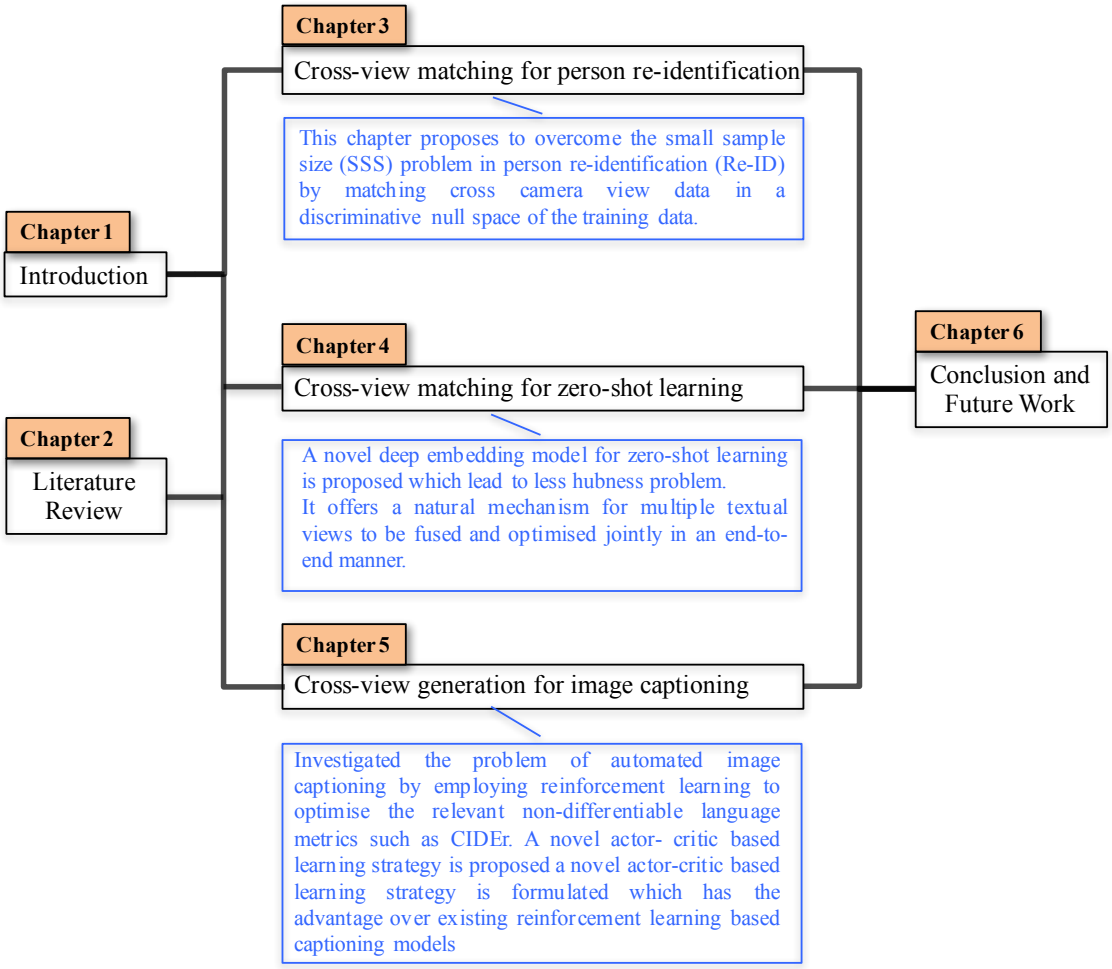


Figure 1.3: Summarisation and structure of all chapters.

Chapter 5 presents a reinforcement learning method for image captioning which aims to automatically describing image content with human-level language. Specifically, a novel actor-critic based learning approach is formulated to directly maximise the reward - the actual natural language processing quality metrics of interest. As compared to existing reinforcement learning based captioning models, the new method has the unique advantage of a per-token advantage and value computation, achieving the state of the art performance on the widely used MSCOCO benchmark.

Chapter 6 concludes the thesis, briefly introduce a number of directions to be pursued as the future work.

Chapter 2

Literature Review

This chapter presents background of several important concepts used throughout this thesis, and literature review. Specifically, Section 2.1 provides background information for machine learning and introduced existing machine learning models for cross-view learning. Section 2.2 reviews conventional cross-view matching methods briefly. Then, Section 2.3 presents cross-view generation methods. Finally, benchmark datasets are summarised in Section 2.4.

2.1 Machine Learning Tools

2.1.1 Overview on machine learning

Machine learning (Friedman et al., 2001; Bishop, 2006; Robert, 2014) is an interdisciplinary field of computer science, statistics, mathematics (esp. optimisation), and neuroscience. Formally, according to (Mitchell et al., 1997), machine learning is "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. Depending on a specific application, one can design various experiences E, tasks T, and performance measures P. Conventionally machine learning is categorised into three different settings, namely,

Supervised learning The objective is to find the mapping between input and output. There is ground truth labels for the output. Most recognition and matching problems fall into this category.

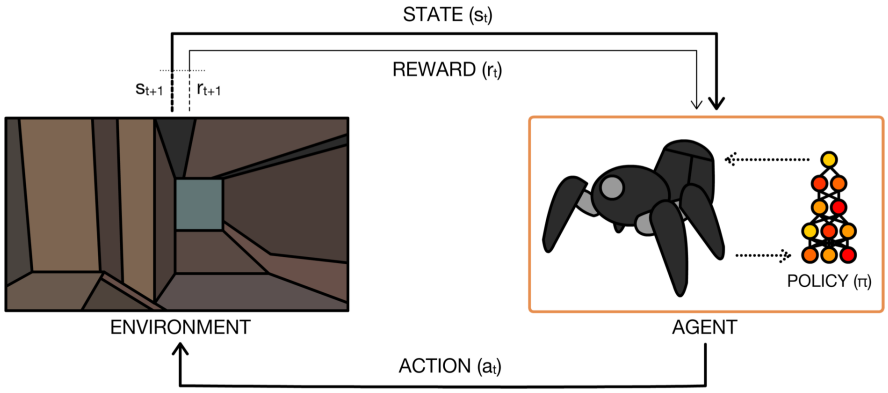


Figure 2.1: An illustration of reinforcement learning from (Arulkumaran et al., 2017). At time t, the agent receives state s_t from the environment. The agent uses its policy to choose an action a_t. Once the action is executed, the environment transitions a step, providing the next state s_{t+1} as well as feedback in the form of a reward r_{t+1}. The agent uses knowledge of state transitions, of the form $(s_t,\ a_t,\ r_{t+1},\ s_{t+1})$, in order to learn and improve its policy.

Unsupervised learning It finds the underlying structure of the given data by transforming it into another representation. There is no precise definition in this case. Dimension reduction and clustering are typical studies in unsupervised learning.

Reinforcement learning (Sutton and Barto, 1998) An agent is trying to maximise the rewards it getting in an environment by taking actions in different states. Reinforcement learning is about how to pick the best action for the agent. A more detailed illustration is shown in Figure 2.1. Two classic reinforcement learning algorithms are introduced following:

Policy gradient Gradients can provide a strong learning signal as to how to improve a parameterised policy. The REINFORCE rule can be used to compute the gradient of an expectation over a function f of a random variable X with respect to parameters θ:

$$\nabla_\theta \mathbb{E}_X[f(X;\ \theta)] = \mathbb{E}_X[f(X;\ \theta)\ \nabla_\theta \log p(X)] \tag{2.1}$$

Actor-Critic It is possible to combine value functions with an explicit representation of the policy, resulting in actor-critic methods (Barto et al., 1983). In doing so, these methods trade off variance reduction of policy gradients with bias introduction from value function methods. Actor-critic methods use the value function as a baseline for policy gradients, such that the only fundamental difference between actor-critic methods and other baseline methods are that actor-critic methods utilise a learned value function. Many state-of-the-art reinforcement learning algorithms (Mnih et al., 2016) are based on actor-critic. For instance, AlphaGo (Silver et al., 2016) utilised the actor-critic method to do self-learning in the game of Go and achieved great

success by beating human world champions. It uses Monte-Carlo rollout and the reward is only set at the end of the game with very long episode.

2.1.2 Machine learning models for cross-view learning

One typical practice of machine learning is: (1) collect the data (2) train the model, and (3) deploy it. What this study focus "cross-view learning" is thus training the model from different views data. Nowadays, a tremendous quantity of data are continually generated. It has been witnessed that many real applications involve large-scale cross-view data. For example, visual data generated by the rapid expansion of large-scale distributed multi-camera systems, with same object captured by different cameras as views; Hundreds of hours of videos are uploaded to YouTube every minute, which appear in multiple modalities, namely visual, audio and text views; A large number of bilingual news are reported every day, with the description in each language as a view. Therefore, design an appropriate cross-view learning model for different views data is essential. In this subsection, existing cross-view learning models are sequentially discussed.

Canonical Correlation Analysis One representative cross-view learning model are Canonical Correlation Analysis (CCA) (Thompson, 2005) and its variants including Kernel CCA (Hardoon et al., 2004) and multi-view CCA (Vía et al., 2007; Gong et al., 2014b). CCA is an approach to correlating linear relationships between two-view feature sets. It seeks linear transformations each for one view such that the correlation between these transformed feature sets is maximized in the common subspace while regularizing the self covariance of each transformed feature sets to be small enough. The aim of CCA is to find two projection directions w_x and w_y corresponding to each view, and maximize the following linear correlation coefficient

$$\frac{cov(w_x^T X,\ w_y^T Y)}{\sqrt{var(w_x^T X)var(w_y^T Y)}} = \frac{w_x^T C_{xy} w_y}{\sqrt{(w_x^T C_{xx} w_x)(w_y^T C_{yy} w_y)}}, \tag{2.2}$$

where X and Y indicates the data from corresponding views, the covariance matrics C_{xy}, C_{xx} and C_{yy} are calculated as $C_{xy} = \frac{1}{n} XY^T$, $C_{xx} = \frac{1}{n} XX^T$, $C_{yy} = \frac{1}{n} YY^T$. The constant $\frac{1}{n}$ can be cancelled out when calculating the correlation coefficient.

CCA has attracted a lot of researchers in past years (Rupnik and Shawe-Taylor, 2010). CCA has been extended to sparse CCA (Chen et al., 2012) and has been widely used for multi-view classification (Sun et al., 2011), clustering (Chaudhuri et al., 2009), regression (Kakade and Foster, 2007). CCA can be extended to multi-view CCA (Vía et al., 2007) by maximizing the sum

of pairwise correlations between different views. However, the main drawback of this strategy is that only correlation information between pairs of features is explored, while high-order statistics are ignored. (Luo et al., 2015) develop tensor CCA (TCCA) to generalize CCA to handle any number of views in a direct and yet natural way. In particular, TCCA can directly maximize the correlation between the canonical variables of all views, and this is achieved by analyzing the high-order covariance tensor over the data from all views (Kim et al., 2007).

Cross-view deep learning models Deep neural networks have recently demonstrated outstanding performance in a variety of tasks such as face recognition, object classification and object detection. They can significantly outperform other methods for the task of large-scale image classification. For cross-view learning, there are also some potential of improving performance through incorporating cross-view learning algorithms and deep learning methods. So far, multi-view deep representation learning has two main strategies (Wang et al., 2015). First, (Ngiam et al., 2011)] proposed Multimodal Autoencoder (MAE) which aims to achieve both within-view and across view reconstruction via a shared embedding. Second, (Andrew et al., 2013) proposed a DNN extension of CCA called deep CCA. For practical application, (Zhu et al., 2014) proposed a multi-view perceptron which is a deep model for learning face identity and view representations. (Su et al., 2015b) presented a novel CNN architecture that combines information from multiple views of a 3D shape into a single and compact shape descriptor. (Elhoseiny et al., 2016) achieved joint object categorization and pose estimation on multi-view data through employing view-invariant representation within CNNs. (Elkahky et al., 2015) presented a general recommendation framework that uses deep learning to match rich user features to item features. They also showed how to extend this framework to combine data from different domains to further improve the recommendation quality. Although these methods have realized deep learning in the cross-view learning framework, there is still a lot of room to develop cross-view deep learning in terms of methodologies and applications.

2.2 Cross-View Matching

This section mainly focuses on reviewing existing methods for cross-view matching. Particularly, recent work for two applications of cross-view learning, person re-identification and zero-shot learning which are different in perspective of *view*, are discussed in Section 2.2.1 and Section 2.2.2, respectively.

2.2.1 Person re-identification

Person re-identification (Re-ID) refers to the problem of visually matching already detected individual or group of people across non-overlapping *cameras views* distributed at diverse physical locations and times (Vezzani et al., 2013; Gong et al., 2014a). Depending on the availability of images across camera views, Re-ID can be performed in a singe-shot (i.e., only one image per camera view) or a multi-shot manner. In the following, first a number of feature learning studies are discussed. Then, contemporary distance metric learning methods for Re-ID are presented. Finally, deep neural network based methods are presented briefly. More thorough reviews can be found in (Vezzani et al., 2013; Gong et al., 2014a; Zheng et al., 2016).

Feature design The first group of methods focus on designing invariant and discriminant features (Li et al., 2014; Gray and Tao, 2008; Farenzena et al., 2010; Kviatkovsky et al., 2013; Ma et al., 2012; Zhao et al., 2014; Lisanti et al., 2014b; Yang et al., 2014; Liao et al., 2015). Designing suitable feature representation for person re-identification is a critical and challenging problem. Ideally, the extracted features should be robust to large cross-view discrepancy, different body poses, changes in illumination, background clutter, occlusion and image quality/resolution. In the context of re-id, however, it is unclear whether there exists universally important and salient features that can be applied readily to different camera views and for all individuals. The discriminative power, reliability and computability of features are largely governed by the camera-pair viewing conditions and unique appearance characteristics of different persons captured in the given views. Moreover, the difficulty in obtaining an aligned bounding box, and accurately segmenting a person from cluttered background makes extracting pure and reliable features depicting the person of interest even harder. The general trend is that the dimensions of the proposed features are getting higher. For instance the dimensions of two representations, recently proposed in (Lisanti et al., 2014b) and (Liao et al., 2015) and used in Chapter 3, are 5,138 and 26,960 respectively. However, no matter how robust the designed features are, they are unlikely to be completely invariant to the often drastic cross-view pose/illumination/background changes.

Model learning Therefore, the second group of methods focus on learning robust and discriminative distance metrics or subspaces for matching people across views, they include: KISSME (Koestinger et al., 2012), RankSVM (Prosser et al., 2010), Probabilistic Relevance Distance Comparison (Zheng et al., 2013), and (Gray and Tao, 2008; Mignon and Jurie, 2012; Tao et al.,

2013; Pedagadi et al., 2013; Li et al., 2013; Zhao et al., 2013b,a; Xiong et al., 2014; Ma et al., 2014; Lisanti et al., 2014b,a; Liao et al., 2015; Paisitkriangkrai et al., 2015).

Apart from a few exceptions (Gray and Tao, 2008; Prosser et al., 2010) based on ranking or boosting, the second groups of methods can be further divided into two major sub-groups: those on learning distance metrics (Koestinger et al., 2012; Zheng et al., 2013; Mignon and Jurie, 2012; Tao et al., 2013) and those on learning discriminative subspaces (Pedagadi et al., 2013; Lisanti et al., 2014b; Xiong et al., 2014; Liao et al., 2015). Seemingly different, these two sub-groups are closely related (Globerson and Roweis, 2005). The main idea of metric learning is to optimise the model parameters so that the cross-view inter-person distance is large whilst intra-person distance is small. Specifically, most metric learning methods focus on Mahalanobis form metrics. If the linear projection of a feature vector $\mathbf{x}_i$ in a learned discriminative subspace is denoted as $\mathbf{y}_i$, we have $\mathbf{y}_i = \mathbf{W}^T\mathbf{x}_i$. The Euclidean distance between $\mathbf{y}_i$ and $\mathbf{y}_j$ is exactly a Mahalanobis distance $||\mathbf{y}_i - \mathbf{y}_j||^2 = (\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{A}(\mathbf{x}_i - \mathbf{x}_j)$ where $\mathbf{A} = \mathbf{W}^T\mathbf{W}$ is a positive semidefinite matrix. In other words, learning a discriminative subspace followed by computing Euclidean distance is equivalent to computing a discriminative Mahalanobis distance over feature vectors in the original space. By making this connection, it is not difficult to see why both methods suffer from the same SSS problem (Section 1.2) typically associated with the subspace learning methods (Chen et al., 2000; Zheng et al., 2005; Guo et al., 2006). Most existing methods need to work with a reduced dimensionality (Pedagadi et al., 2013), achieved typically by PCA whose dimension has to be carefully tuned for each dataset. Some works additionally require introducing matrix regularisation term if the intra-class scatter matrix is used in the formulation, in order to prevent matrix singularity (Pedagadi et al., 2013; Lisanti et al., 2014b; Xiong et al., 2014; Liao et al., 2015), again with free parameters to tune. Critically, they suffer from the degenerate eigenvalue problem (i.e. several eigenvectors share the same eigenvalue), which makes the solution sub-optimal resulting in loss of discriminant ability (Zheng et al., 2005). In contrast, the proposed discriminative null space based approach (Section 3), neither dimensionality reduction before model learning nor regularisation term is required, and it has no parameters to tune.

As a solution proposed specifically to address the SSS problem, the null Foley-Sammon transfer (NFST) method has been around for a long time (Guo et al., 2006), but received very little attention apart from a recent application to the novelty detection problem (Bodesheim et al., 2013). A possible reason is that by restricting the learned discriminative projecting directions to

the null projecting directions (NPDs), on which within-class distance is always zero and between-class distance is positive, the model is *extreme*, leaving little space for further extension with clear added-value. For example, the more relaxed Fisher discriminative analysis (FDA) can be extended, gaining notable advantage, by exploit graph laplacian to preserve local data structure, known as LFDA (Sugiyama, 2006), which has been successfully applied to Re-ID (Pedagadi et al., 2013). However, a similar graph laplacian extension to NFST does not apply due to its single point per class nature. Despite the restrictions, given the acute SSS problem in Re-ID distance metric learning, the basic idea of learning a null space for overcoming this problem becomes very attractive. The general concept of collapsing same-class data points to a single point has been exploited in a Mahalanobis distance learning framework, known as maximally collapsing metric learning (MCML) (Globerson and Roweis, 2005). However, MCML does not exploit a null space. Instead, the MCML model must make approximations with plenty of free parameters to tune and no closed-form solution.

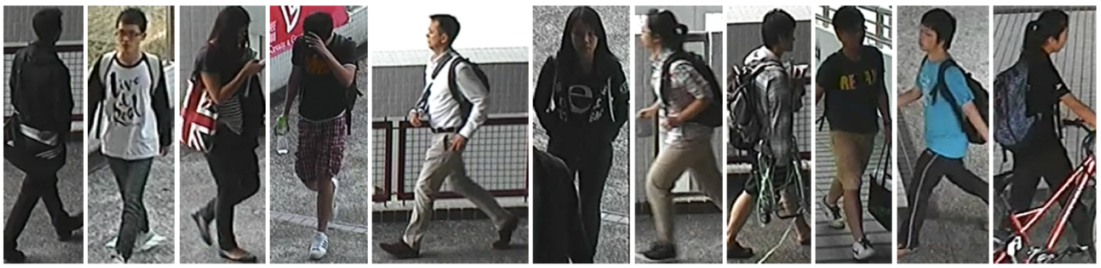
Deep learning based methods Recently, the third group of methods start to appear which are based on deep learning (Shi et al., 2016; Cheng et al., 2016; Liu et al., 2017b; Varior et al., 2016; Ustinova et al., 2015; Yi et al., 2014; Li et al., 2014; Ahmed et al., 2015; Xiao et al., 2016; Li et al., 2017; Geng et al., 2016; Chen et al., 2017) have obtained impressive performance. These approaches are largely inspired by the strong representation auto-learning capacity of deep models. They differ significantly in their network architectures, which are largely determined by the training objectives/losses. Specifically, most existing works cast the Re-ID problem as a deep metric learning problem and employ pairwise verification loss (Yi et al., 2014; Ahmed et al., 2015; Ustinova et al., 2015; Varior et al., 2016; Shi et al., 2016) or triplet ranking loss (Liu et al., 2017b; Cheng et al., 2016), or both (Wang et al., 2016a). Correspondingly the overall network architecture is a Siamese CNN network with either two or three branches for the pairwise or triplet loss respectively. (Xiao et al., 2016) uses an identity classification loss with one-branch architecture. (Geng et al., 2016) has a Siamese two-branch architecture with an identity classification loss for each branch and pairwise verification loss across the two branches. (Li et al., 2017) jointly learns the local and global branches, both subject to the same identity class supervision for maximising the complementary advantages of local and global Re-ID feature learning whilst enhancing their individual discriminative power. (Chen et al., 2017) also adopted two-branch architecture, not only learning scale-specific discriminative features by optimising multiple clas-

sification losses on the same person label information concurrently, but also maximising jointly multi-scale complementary fusion selections by multi-scale consensus regularisation in a closed-loop form. However, deep learning based Re-ID models require a large number of training data, which may be not available in many real cases. For example, deep Re-ID models are often poor in small scale due to overfitting problem, CMC rank-1 accuracy of (Xiao et al., 2016) on VIPeR lower than 40% .

Attribute/Language based methods Visual semantic attributes (Wang et al., 2017b; Liu et al., 2017d) have been exploited as a mid-level feature representation for cross-view Re-ID (Layne et al., 2014a,b, 2012; Peng et al., 2016; Su et al., 2015a, 2016). This data has been manually annotated with 15 binary attributes in (Layne et al., 2012), which include: shorts, skirt, sandals, backpack, jeans, logo, v-neck, open-outerwear, stripes, sunglasses, headphones, long-hair, short-hair, gender, carrying-object, 12 of which are appearance-based, and 3 are soft-biometrics. In many practical cases visual example is not available or tedious manual search has to be done to identify one instance that can then be used as a query. In such cases a natural language description is the only available *view*, often gathered from a number of witnesses with many variations and inconsistencies. For example, natural language descriptions of visual image can be often found in missing person sections in newspapers . (Yan et al., 2017) first extends the conventional Re-ID datasets with natural language descriptions (see Figure 2.2), which will facilitate research in person Re-ID with joint vision and language modelling. It covers the following scenarios: 1) the gallery has only the vision modality while the query has only the language modality; 2) the gallery has only vision while the query has both vision and language; and finally, 3) both the gallery and the query have vision and language. Second, (Yan et al., 2017) propose models that integrate vision and language, and demonstrate that in several Re-ID scenarios, the performance can be significantly improved by the proposed integration; Third, (Yan et al., 2017) compare natural language annotations to attribute based ones (Layne et al., 2014a, 2012), and identify their relative advantages. Compared to attribute based annotations (Layne et al., 2014a, 2012), the advantage of natural language description is its flexibility and richness.

2.2.2 Zero-shot learning

Zero-shot learning (ZSL) is a variant of cross-view learning which aims to matching the visual view of a unseen data (objects) and its textual view. Specifically, ZSL models rely on learning



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k)

(a) ... cross body bag which looks like from the brand Adidas ...
(b) ... that has black coloured text reading "Live Real" printed on it ...
(c) ... carrying a UK flag printed hand bag ...
(d) ... carrying a small transparent bottle with a green cap ...
(e) ... has a bandage on his right hand with a strap from his neck as he might have had an accident ...
(f) ... a black casual tee with a big "e" printed on it ...
(g) ... is seen rushing ...
(h) ... seems to be carrying an interesting metal object with green pipes dangling ...
(i) ... which has the text "RELAX" written on it in yellow ...
(j) ... is striding to the left ...
(k) ... pushes a red bicycle with both hands ...

Figure 2.2: Example images from CUHK03 with discriminative sections of their natural language descriptions. Figure credits to (Yan et al., 2017).

a joint embedding space where both textual description of seen and unseen classes and visual representation of object images can be projected to for cross-view matching.

Semantic space Seen and unseen classes are usually related in a high dimensional vector space, called semantic space, where the knowledge from seen classes can be transferred to unseen classes. Existing ZSL methods differ in what semantic spaces are used: typically either attribute (Farhadi et al., 2009; Ferrari and Zisserman, 2007; Parikh and Grauman, 2011), word vector (Socher et al., 2013; Frome et al., 2013), or text description (Reed et al., 2016a). It has been shown that an attribute space is often more effective than a word vector space (Akata et al., 2015; Zhang and Saligrama, 2015; Lampert et al., 2014; Romera-Paredes and Torr, 2015). This is hardly surprising as additional attribute annotations are required for each class. Similarly, state-of-the-art results on fine-grained recognition tasks have been achieved in (Reed et al., 2016a) using sentence descriptions/captions to construct the semantic space. Again, the good performance is obtained at the price of more manual annotation: 10 sentence descriptions need to be collected for each image, which is even more expensive than attribute annotation. This is why the word vector semantic space is still attractive: it is 'free' and is the only choice for large scale recognition with many unseen classes (Fu and Sigal, 2016). In this thesis, all three semantic spaces are considered.

Fusing multiple semantic spaces Multiple semantic spaces are often complementary to each other; fusing them thus can potentially lead to improvements in recognition performance. Score-

level fusion is perhaps the simplest strategy (Fu et al., 2015b). More sophisticated multi-view embedding models have been proposed. (Akata et al., 2015) learn a joint embedding semantic space between attribute, text and hierarchical relationship which relies heavily on hyperparameter search. Multi-view canonical correlation analysis (CCA) has also been employed (Fu et al., 2014) to explore different modalities of testing data in a transductive way. Differing from these models, our neural network based model has an embedding layer to fuse different semantic spaces and connect the fused representation with the rest of the visual-semantic embedding network for end-to-end learning. Unlike (Fu et al., 2014), it is inductive and does not require to access the whole test set at once.

Embedding model Existing methods also differ in the visual-semantic embedding model used. They can be categorised into two groups: (1) The first group learns a mapping function by regression from the visual feature space to the semantic space with pre-computed features (Lampert et al., 2014; Fu and Sigal, 2016) or deep neural network regression (Socher et al., 2013; Frome et al., 2013). For these embedding models, the semantic space is the embedding space. (2) The second group of models implicitly learn the relationship between the visual and semantic space through a common intermediate space, again either with a neural network formulation (Lei Ba et al., 2015; Yang and Hospedales, 2015) or without (Lei Ba et al., 2015; Akata et al., 2015; Romera-Paredes and Torr, 2015; Fu et al., 2014). The embedding space is thus neither the visual feature space, nor the semantic space. This study shows that using the visual feature space as the embedding space is intrinsically advantageous due to its ability to alleviate the *hubness* problem (Section 1.2).

Deep ZSL model All recent ZSL models use deep CNN features as inputs to their embedding model. However, few are deep end-to-end models. Existing deep neural network based ZSL works (Frome et al., 2013; Socher et al., 2013; Lei Ba et al., 2015; Yang and Hospedales, 2015; Reed et al., 2016a) differ in whether they use the semantic space or an intermediate space as the embedding space, as mentioned above. They also use different losses. Some of them use margin-based losses (Frome et al., 2013; Yang and Hospedales, 2015; Reed et al., 2016a). Socher *et al* (Socher et al., 2013) choose a Euclidean distance loss. (Lei Ba et al., 2015) takes a dot product between the embedded visual feature and semantic vectors and consider three training losses, including a binary cross entropy loss, hinge loss and Euclidean distance loss. The work in (Reed et al., 2016a) differs from the other models in that it integrates a neural language model into its

neural network for end-to-end learning of the embedding space as well as the language model.

The hubness problem The phenomenon of the presence of 'universal' neighbours, or hubs, in a high-dimensional space for nearest neighbour search was first studied by Radovanovic et al. (Marco et al., 2015). They show that hubness is an inherent property of data distributions in a high-dimensional vector space, and a specific aspect of the curse of dimensionality. A couple of recent studies (Dinu et al., 2014; Shigeto et al., 2015) noted that regression based zero-shot learning methods suffer from the hubness problem and proposed solutions to mitigate the hubness problem. Among them, the method in (Dinu et al., 2014) relies on the modelling of the global distribution of test unseen data ranks w.r.t. each class prototypes to ease the hubness problem. It is thus transductive. In contrast, the method in (Shigeto et al., 2015) is inductive: It argued that least square regularised projection functions make the hubness problem worse and proposed to perform reverse regression, i.e., embedding class prototypes into the visual feature space. Our model also uses the visual feature space as the embedding space but achieve so by using an end-to-end deep neural network which yields far superior performance on ZSL.

2.3 Cross-View Generation

Whilst most existing approaches to cross-view learning are devoted to tasks of matching (Section 2.2.1 and 2.2.2), extending cross-view learning frameworks to scenarios of generating raw data of novel view(s) conditioned on other view(s) is non-trivial due to the significant view discrepancy and the ability of the model to balance an understanding of both views. This section mainly focuses on reviewing existing approaches to cross-view generation. In particular, this section is separated into three subsections based on the realm of *view*: Fist, a number of image synthesis frameworks are discussed in Section 2.3.1; Then, image captioning models are introduced in Section 2.3.2; At last, group of work on machine translation are presented in Section 2.3.3.

2.3.1 Image synthesis

Generative Adversarial Network As one of the most significant improvements on the research of deep generative model, Generative Adversarial Network (GAN) (Goodfellow et al., 2014) has drawn substantial attention from computer vision society and achieved impressive results in image synthesis. We first briefly revisit GAN.

GAN consists of a generator G and a discriminator D that compete in a two-player minimax

game. The discriminator tries to distinguish a real image x from a synthetic one $G(z)$, and the generator tries to synthesize realistic-looking images that can fool discriminator. Concretely, D and G play the following game on $V(D,\ G)$:

$$\min_G \max_D V(D,\ G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{x\sim p_z(z)}[\log(1-D(G(z)))] \tag{2.3}$$

It is proved in (Goodfellow et al., 2014) that this minimax game has a global optimum when the distribution p_g of the synthetic samples and the distribution p_d of the training samples are the same. Under mild conditions (e.g., G and D have enough capacity), p_g converges to p_d. In practice, it is better for G to maximize $\log(D(G(z)))$ instead of minimizing $\log(1-D(G(z)))$.

The min-max two-player game provides a simple yet powerful way to estimate target distribution and generate novel image samples (Denton et al., 2015). With its power for distribution modelling, the GAN can encourage the generated images to move towards the true image manifold and thus generates photorealistic images with plausible high frequency details. Recently, modified GAN architectures, conditional GAN (Mirza and Osindero, 2014) in particular, have been successfully applied to vision tasks like image-to-image translation (Isola et al., 2016), super-resolution (Ledig et al., 2017), style transfer (Li and Wand, 2016), text-to-image-synthesis (Reed et al., 2016b) and camera view synthesis (Huang et al., 2017). These successful applications of GAN motivate us to develop frontal view synthesis methods based on GAN.

Camera View Synthesis When the *view* refers to camera view, the goal of view synthesis is to create unseen novel view(s) images based on a set of available existing views. From an optimisation point of view, generating novel view(s) from incompletely observed profile is an ill-posed problem.

There is one line of work use Generative Adversarial Network (GAN) (Goodfellow et al., 2014) to achieve this. (Huang et al., 2017) proposes a Two-Pathway Generative Adversarial Network (TP-GAN) for photorealistic frontal view face synthesis by simultaneously perceiving global structures and local details from a single image. Four landmark located patch networks are proposed to attend to local textures in addition to the commonly used global encoder-decoder network. The combination of adversarial loss, symmetry loss and identity preserving loss functions leverage both frontal face distribution and pre-trained discriminative deep face models to guide an identity preserving inference of frontal views from profiles. Some samples generated by TP-GAN are shown in Figure 2.3. (Tran et al., 2017) proposes Disentangled Representation

Figure 2.3: Frontal view synthesis by TP-GAN (Huang et al., 2017). The upper half shows the 90° profile image (middle) and its corresponding synthesized (left) and ground truth frontal face (right). The lower half shows the synthesized frontal view faces from profiles of 90°, 75° and 45° respectively.

learning-Generative Adversarial Network (DR-GAN) that takes one or multiple face images as the input, producing an identity representation that is both discriminative and generative, and can synthesis identity-preserving faces at any views specified by the pose code.

The other line of work of camera view synthesis is to use encoder-decoder network. (Ji et al., 2017) proposes a novel CNN architecture for view synthesis called ”Deep View Morphing” that does not suffer from lacking of texture details, shape distortions, or high computational complexity. To synthesize a middle view of two input images, a rectification network first rectifies the two input images. An encoder-decoder network then generates dense correspondences between the rectified images and blending masks to predict the visibility of pixels of the rectified images in the middle view. A view morphing network finally synthesizes the middle view using the dense correspondences and blending masks. Apart from the above 2D image synthesis work, (Yang et al., 2015) proposes a novel recurrent convolutional encoder-decoder network that is trained end-to-end on the task of rendering rotated 3D faces and chairs starting from a single image. The recurrent structure allows the proposed model to capture long-term dependencies along a sequence of transformations.

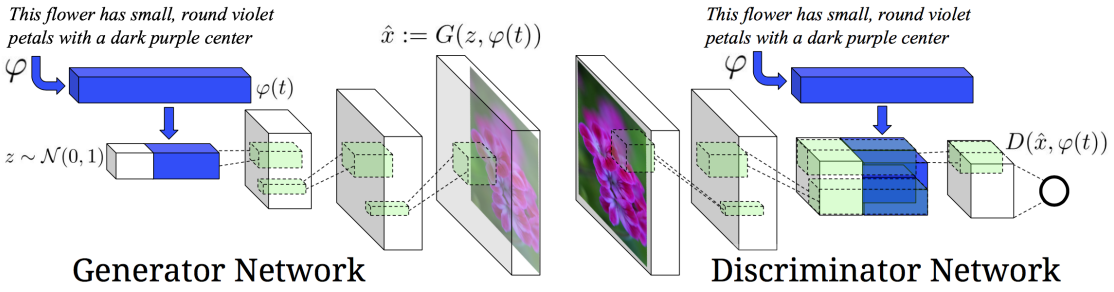


Figure 2.4: An example of text-to-image synthesis network: text-conditional convolutional GAN architecture (Reed et al., 2016b). Text encoding $\varphi(t)$ is used by both generator and discriminator. It is projected to a lower-dimensions and depth concatenated with image feature maps for further stages of convolutional processing. Figure credits to (Reed et al., 2016b).

Text-to-Image Synthesis The goal of text-to-image synthesis is automatic generating realistic images conditioned on text descriptions. (Reed et al., 2016b) develops a simple and effective end-to-end GAN architecture and training strategy that enables compelling text to image synthesis of bird and flower images from human-written descriptions. The model can synthesise many plausible visual interpretations of one given text caption. The approach is to train a deep convolutional generative adversarial network (DC-GAN) conditioned on text features encoded by a hybrid character-level convolutional recurrent neural network. Both the generator network G and the discriminator network D perform feed-forward inference conditioned on the text feature. Specifically, in the generator G, first sampling from the noise prior z and encoded the text query t using text encoder φ. The description embedding $\varphi(t)$ is first compressed using a fully-connected layer to a small dimension followed by leaky-ReLU and then concatenated to the noise vector z. Details of the network architecture is shown in Figure 2.4.

Another line of work that utilise GAN to generate image from text view and visual view is Fashion Synthesis (Zhu et al., 2017b). It extends the DeepFashion dataset (Liu et al., 2016) by collecting sentence descriptions for 79K images. Given an input image of a person and a sentence description of a new desired outfit (e.g.*a lady dressed in sleeveless white clothes*), (Zhu et al., 2017b) first generates a segmentation map $\widetilde{S}$ using the generator from the first GAN. Then they render the new image with another GAN, with the guidance from the segmentation map generated in the previous step. At test time, the final rendered image is obtained with a forward pass through the two GAN networks (see Figure 2.5).

Image-to-Image Translation When the *view* retains to the different representations of a particular image, such as an RGB image, a gradient field, an edge map or a Monet style painting,

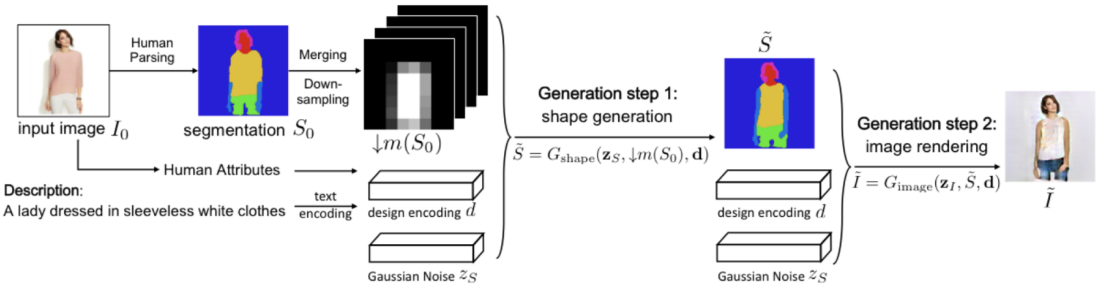


Figure 2.5: An example of Fashion synthesis network (Zhu et al., 2017b).

then automatic image-to-image translation is another variant of cross-view generation, which is defined as translating one possible representation of an image into another.

(Isola et al., 2016) training a conditional GAN to predict one view images conditioned on another paired view (e.g., generating aerial photos from maps). The discriminator D learns to classify between real and synthesised pairs. The generator learns to fool the discriminator. Unlike an unconditional GAN, both the generator and discriminator observe an input image. (Zhu et al., 2017a) learns to do the same: capturing special characteristics of one view image and figuring out how these characteristics could be translated into the other views, but it takes a further step that all in the absence of any paired training examples by introducing an additional cycle consistency loss. Neural Style Transfer (Li and Wand, 2016; Johnson et al., 2016) is another way to perform image-to-image translation, which synthesises a novel image by combining the content of one image with the style of another image (typically a painting) based on matching the Gram matrix statistics of pre-trained deep features.

2.3.2 Image captioning

If the mapping function is learned from visual view to textual view, which is rightly an inverse mapping compared with text-to-image synthesis, then we called this as image captioning. Image captioning aims to automatically describe the visual content of an image in natural language instead of merely assigning it a category. Image captioning model does indeed develop the ability to generate accurate new captions when presented with completely new scenes, indicating a deeper understanding of the objects and context in the visual images. Moreover, it learns how to express that knowledge in natural-sounding English phrases despite receiving no additional language training other than reading the human textual captions. Figure 2.7 shows the image captioning model generates a completely new caption using concepts learned from similar scenes

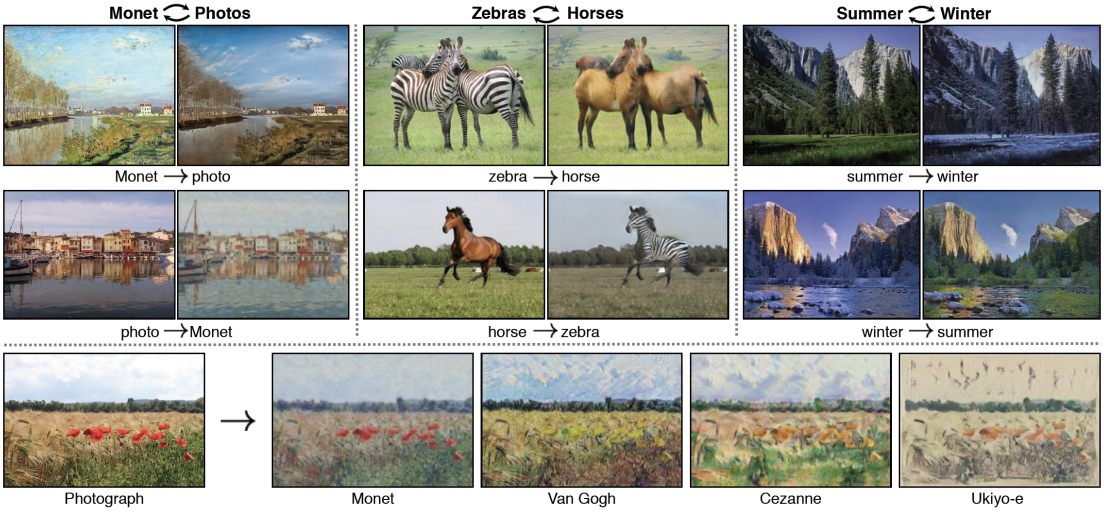


Figure 2.6: Image synthesis examples from (Zhu et al., 2017a). Given any two unordered image collections X and Y, (Zhu et al., 2017a)learns to automatically "translate" an image from one into the other and vice versa: (left) Monet paintings and landscape photos from Flickr; (center) zebras and horses from ImageNet; (right) summer and winter Yosemite photos from Flickr. Example application (bottom): using a collection of paintings of famous artists, model learns to render natural photographs into the respective styles.

in the training set.

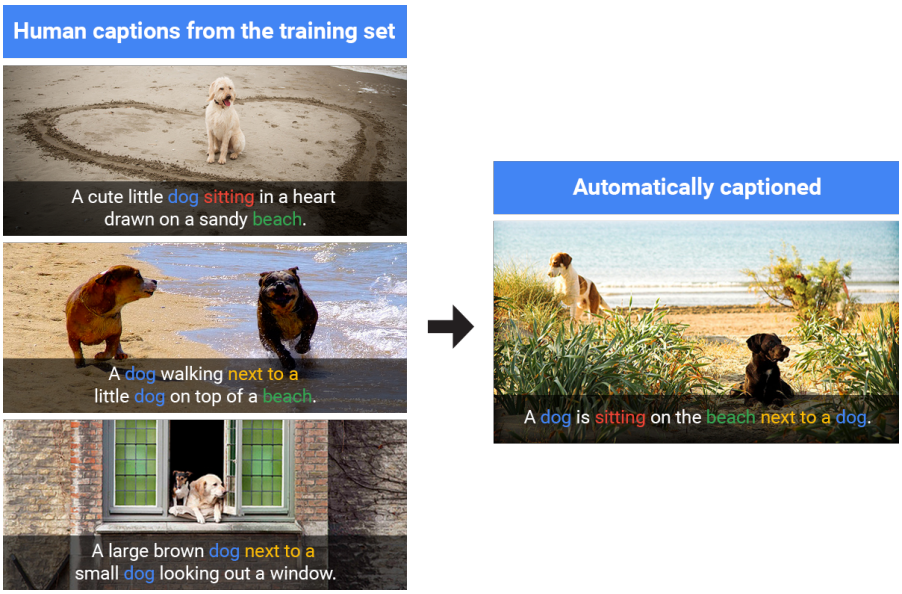


Figure 2.7: Image captioning model generates a completely new caption using concepts learned from similar scenes in the training set

Image captioning with supervised learning There is now extensive work on image captioning (Vinyals et al., 2015; Fang et al., 2015; Xu et al., 2015; Devlin et al., 2015; Mao et al., 2015; Wu et al., 2016; Vinyals et al., 2016; Liu et al., 2017a) which are based on supervised learning. The typical pipeline is based on a Convolutional Neural Network (CNN) image encoder and a re-

current neural network (RNN) based sentence decoder (Vinyals et al., 2015, 2016). Specifically, each image will be encoded by a deep convolutional neural network into a visual vector representation. A language generating RNN, or recurrent neural network, will then decode that visual representation sequentially into a textual/natural language description (Figure 2.8). The CNN image representation can be entered into the RNN in different manners. While some (Vinyals et al., 2015; Karpathy and Fei-Fei, 2015) use it only to compute the initial state of the RNN, others enter it in each RNN iteration (Mao et al., 2015; Donahue et al., 2015).

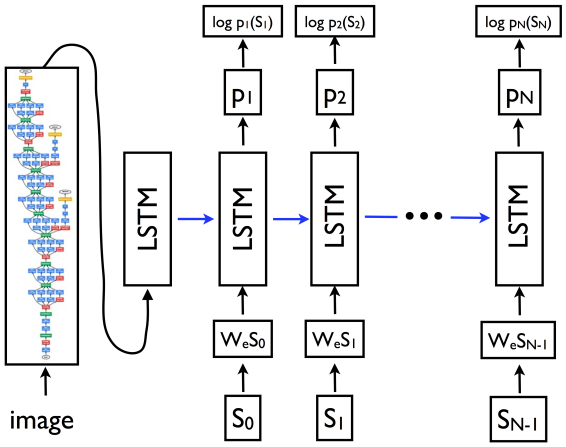


Figure 2.8: CNN-RNN model architecture for image captioning. Figure credits to (Vinyals et al., 2015, 2016).

(Xu et al., 2015) were the first to propose an attention-based approach for image captioning, in which the RNN state update includes the visual representation of an image region. Which image region is attended to is determined based on the previous state of the RNN. They propose a ”soft” variant in which a convex combination of different region descriptors is used, and a ”hard” variant in which a single region is selected. The latter is found to perform slightly better, but is more complex to train due to a non-differentiable sampling operator in the state update. In their approach the positions in the activation grid of a convolutional CNN layer is the loci of attention. Each position is described with the corresponding activation column across the layer’s channels.

Several works build upon the approach of (Xu et al., 2015). (You et al., 2016) learns a set of attribute detectors, similar to (Fang et al., 2015), for each word of their vocabulary. These detectors are applied to an image, and the strongest object detections are used as regions for an attention mechanism similar to that of (Xu et al., 2015). In their work the detectors are learned prior and independently from the language model. (Wu et al., 2016) also learn attribute

detectors but manually merge word tenses (walking, walks) and plural/singulars (dog, dogs) to reduce the set of attributes. (Wu and Cohen, 2016) improve the attention based encoder-decoder model by adding a reviewer module that improves the representation passed to the decoder. They show improved results for various tasks, including image captioning. (Yao et al., 2015) use a temporal version of the same mechanism to adaptively aggregate visual representations across video frames per word for video captioning.

Another topical issues addressed in the literature include improving visual feature representations (Liu et al., 2017a). Specifically, (Liu et al., 2017a) propose a novel CNN-RNN image annotation model which differs from the existing models in the selection of the image embedding layer and in the introduction of deeply supervised semantic regularisation to the embedding layer.

Image captioning with reinforcement learning Image captioning methods summarised above are all typically trained by maximising training caption likelihood through teacher forcing (Section 1.2). Recently a few studies proposed to use reinforcement learning to address the discrepancy between the standard training objective for image captioning (likelihood/teacher forcing) and the evaluation metrics of interest (CIDEr) (Ranzato et al., 2016; Liu et al., 2017c; Rennie et al., 2017). (Rennie et al., 2017) uses the basic REINFORCE algorithm (Williams, 1992).

REINFORCE with a Baseline. The policy gradient given by REINFORCE can be generalized to compute the reward $r(w^s)$ associated with an action value relative to a reference reward or baseline. The core idea of (Rennie et al., 2017) is to baseline the REINFORCE algorithm with the reward $r(\tilde{w})$ obtained by the current model under the inference algorithm used at test time.

$$\nabla_\theta \mathcal{L}(\theta) = -\mathbb{E}_{w^s \sim p_\theta}[(r(w^s) - r(\tilde{w}))\, \nabla_\theta \log p_\theta(w^s)]. \tag{2.4}$$

As a results, for each sampled caption, it has only one sentence level advantage which means that every token makes the same contribution towards the whole sentence – a clearly invalid assumption. (Ranzato et al., 2016; Liu et al., 2017c) add an additional FC layer on top of the RNN output to predict state value function. However, both treat *state* as the RNN output while the proposed image captioning model in this thesis treat the *state* as the RNN input (given image and the taken actions), so that one can build an independent value network rather than a shared RNN cell between actor and critic.

2.3.3 Machine translation

When the *views* represent different textual human-languages (e.g., German, Frensh and English etc). Machine translation (Sutskever et al., 2014; Bahdanau et al., 2015) is another variant of cross-view generation.

The models proposed recently for neural machine translation often belong to a family of encoderdecoders and encode a source sentence into a fixed-length vector from which a decoder generates a translation. (Sutskever et al., 2014) presents a general end-to-end approach to sequence learning that makes minimal assumptions on the sequence structure. It uses a multilayered Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) to map the input sequence to a vector of a fixed dimensionality, and then another deep LSTM to decode the target sequence from the vector. (Bahdanau et al., 2015) conjectures that the use of a fixed-length vector is a bottleneck in improving the performance of this basic encoderdecoder architecture, and propose to extend this by allowing a model to automatically (soft-)search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly.

Sequence generation with reinforcement learning Very recently there is one line of work on sequence generation (Bahdanau et al., 2017; Yu et al., 2017; Paulus et al., 2017) which are using reinforcement learning for model training. (Bahdanau et al., 2017) uses actor-critic for machine translation. Their actor and critic have the same encoder-decoder architecture while the critic outputs state-action value function for each possible actions for policy iteration. Given that the action space is huge for a sequence generation task, the predicted action-value function often have to rely on various tricks to penalising the variance of the outputs of the critic. Without the penalty the values of rare actions can be severely overestimated, introducing bias to the gradient estimates and causing convergence difficulties. (Yu et al., 2017) uses Monte Carlo rollout to sample actions and uses GAN to compute reward. (Paulus et al., 2017) applies the same method as (Rennie et al., 2017) to improve the performance of abstractive summarisation.

2.4 Benchmark Dataset

The proposed cross-view learning frameworks are evaluated on a number of benchmark datasets. They are summarised in Table 2.1, Table 2.2 and Table 2.3. Examples are shown in Figure 2.9, Figure 2.10, and Figure 2.11.

Dataset	Cameras	Person	Instance	Chapter
VIPeR (Gray et al., 2007)	2	632	1264	3
PRID2011 (Hirzer et al., 2011)	2	749	949	3
CUHK01 (Li and Wang, 2013)	2	971	1942	3
CUHK03 (Li et al., 2014)	6	1467	14097	3
Market1501 (Zheng et al., 2015)	6	1501	32668	3

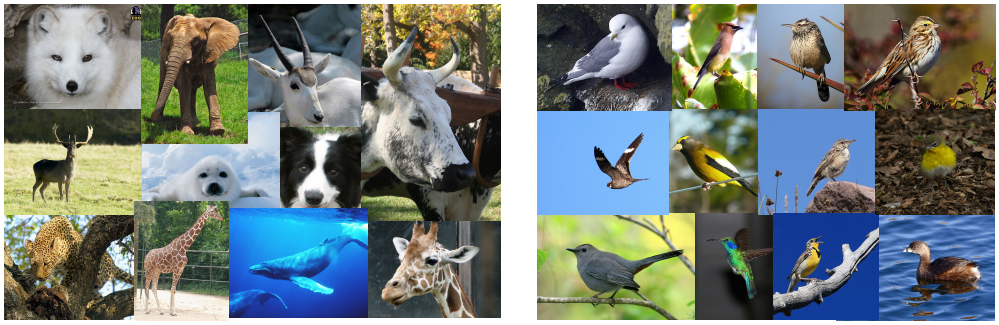
Table 2.1: Benchmark datasets for evaluation of cross-view matching for person re-identification

Dataset	Instances	SS	SS-D	Seen-Unseen	Chapter
AwA (Animals with Attributes) (Lampert et al., 2014)	30,475	A/W	85	40-10	4
CUB (CUB-200-2011) (Wah et al., 2011)	11,788	A/D	312	150-50	4
ImageNet (ILSVRC) 2010 1K (Russakovsky et al., 2015)	1.2×10^6	W	1000	800-200	4
ImageNet (ILSVRC) 2012/2010 (Russakovsky et al., 2015)	218,000	W	1000	1000-360	4

Table 2.2: Benchmark datasets for evaluation of cross-view matching for zero-shot learning. Notation: SS: semantic space; SS-D: the dimension of semantic space; A: attribute space; W: semantic word vector space; D: sentence description (only available for CUB)

Dataset	Training	Validation	Test	Caption	Chapter
MSCOCO (Lin et al., 2014)	82,783	40,504	40,775	5	5

Table 2.3: Benchmark datasets for evaluation of cross-view generation for image caption. Notation: Train, Validation and Test means the number of images in each split; Caption: number of captions annotated for each image.



(a) AwA (b) CUB

Figure 2.9: Visual examples of: (a) AwA dataset; (b) CUB dataset.

Figure 2.10: Visual examples of VIPeR dataset. Person are captured under two camera views.

a man sitting in front of a plate holding a sandwich next to a girl.
a young man is eating a hamburger while a young girl watches and laughs.
there is a man and a woman eating at the table.
a couple of people sitting at a wooden table.
a young man siting at a picnic table holding a sandwich while a young girl looks on smiling.

some people are holding a union jack umbrella.
many people stand around a bricked over square, holding their umbrellas.
a group of people on a street with umbrellas.
some people with their umbrellas and some buildings.
city street with people walking with coats and umbrellas.

a police officer patrols the streets on a motorcycle.
a police officer rides a motorcycle in front of a building.
an police officer riding on a police motorcycle.
a police officer sitting on motorbike with his police lights on.
an motorcycle cop riding his motorcycle in front of a store.

a little girl sitting at a table wiping her mouth with a napkin.
a little kid that is eating some food on the table.
an adorable little girl sitting at a black table.
little girl sitting at a table holding a napkin to her face.
a young child sitting on a table with two plates in front.

Figure 2.11: Examples of MSCOCO dataset. Five captions are given to describe for the corresponding image.

2.5 Summary

The preceding sections have given the background of important concepts used throughout this thesis and discussed important studies in the literature with respect to cross-view matching and cross-view generation methods in generic machine learning. Specifically, the main topics include cross-view matching for person re-identification and zero-shot learning, cross-view gen-

eration for image captioning. Despite the significant progress made by existing methods, there are still many limitations and many open problems to explore. In the subsequent chapters, novel approaches are presented to advance further cross-view learning problem:

1. (Chapter 3) To avoid the small sample size problem in distance metric learning Re-ID, most existing models adopted unsupervised dimensionality reduction or regularisation methods. These in turn make the learned distance metric sub-optimal and less discriminative. To overcome this problem, a distance metric learning model is proposed for Re-ID by matching people in a discriminative null space of the training data. In this null space, images of the same person are collapsed into a single point thus minimising the within-class scatter to the extreme and maximising the relative between-class separation simultaneously.
2. (Chapter 4) Despite the success of deep neural networks that learn an end-to-end model across *visual* and *textual* views in other vision problems such as image captioning, very few deep ZSL model exists and they show little advantage over ZSL models that utilise deep feature representations but do not learn an end-to-end embedding. Specifically, existing models, regardless whether they are deep or non-deep, choose either the semantic space or an intermediate embedding space as the embedding space, suffering from the *hubness* problem which is the main issue prevents the deep ZSL model to succeed. Nonetheless, how to deal with the hubness problem in ZSL is largely ignored in the literature. To alleviate the hubness problem, this study utilises the output visual feature space of a CNN subnet as the embedding space. Further, a natural mechanism for multiple textual modalities optimised jointly in an end-to-end manner in this model demonstrates significant advantages over existing methods
3. (Chapter 5): Most of existing models for image captioning rely on ”Teacher-Forcing” training strategy and optimise the supervision metric ”cross-entropy loss”. Therefore, they are likely to produce sub-optimal results. Recently proposed reinforcement learning methods for image captioning aim to overcome two limitations but suffer from invalid assumption of sentence-level poor value function design. To address those limitations, a novel actor-critic sequence training framework for the image captioning is formulated with a per-token advantage and value computation strategy.

Chapter 3

Cross-View Matching for Person Re-Identification by Learning a Discriminative Null Space

---

This chapter discusses a real-world cross-view matching problem – person re-identification (Re-ID) – with the case that *view* denotes camera view. Person re-identification (Re-ID) refers to the problem of visually matching already detected individual or group of people across non-overlapping cameras views. Most existing Re-ID methods focus on learning the optimal distance metrics across camera views. Typically a person's appearance is represented using features of thousands of dimensions, whilst only hundreds of training samples are available due to the difficulties in collecting matched training images. With the number of training samples much smaller than the feature dimension, the existing methods thus face the classic small sample size (SSS) problem and have to resort to dimensionality reduction techniques and/or matrix regularisation, which lead to loss of discriminative power. In this chapter, a distance metric learning model is proposed to overcome the *small sample size* challenge discussed in Section 1.2 by matching people in a discriminative null space of the training data. In this null space, images of the same person are collapsed into a single point thus minimising the within-class scatter to the extreme and maximising the relative between-class separation simultaneously. Importantly, it has a fixed dimension, a closed-form solution and is very efficient to compute. Extensive experiments are conducted on five widely used person re-identification benchmarks to evaluate the advantages of such a simple approach by comparing most contemporary methods

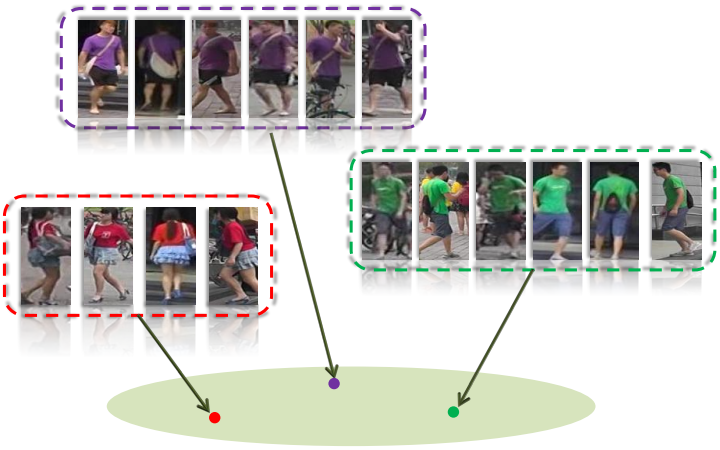


Figure 3.1: Training images of same identity are projected to a single point in a learned discriminative null space.

3.1 Background

For making sense of the vast quantity of visual data generated by the rapid expansion of large-scale distributed multi-camera systems, automated person re-identification is essential (Vezzani et al., 2013; Gong et al., 2014a). When a person is captured by multiple non-overlapping views, the objective is to match him/her across views among a large number of imposters, which is a classic cross-view matching problem.

Existing approaches for cross-view matching focus on developing discriminative feature representations that are robust against the view/pose/illumination/background changes, or learning a distance metric, dedicated to eliminate the large camera view discrepancy. Among them, the distance metric learning methods are most popular and are the focus of this chapter. Given any feature representation and a set of training data consisting of matching image pairs across camera views, the objective for Re-ID is to learn the optimal distance metric that gives small values to images of the same person and large values for those of different people. Although the distance metric learning methods have been shown to be effective in improving the existing Re-ID benchmarks over the past five years, some of them are still limited by a classical problems in model learning, small sample size (SSS) problem (Chen et al., 2000).

This chapter argues that the SSS problem in person Re-ID distance metric learning can be best solved by learning a discriminative null space of the training data. In particular, instead of minimising the within-class variance, data points of the same classes are *collapsed*, by a trans-

form, into a single point in a new space (see Fig. 3.1). By keeping the between-class variance non-zero, this automatically maximises the Fisher discriminative criterion and results in a discriminative subspace. The null space method, also known as the null Foley-Sammon transfer (NFST) (Guo et al., 2006) is specifically designed for the small sample case, with rigorous theoretical proof on the resulting subspace dimension. Importantly, it has a closed-form solution, no parameter to tune, requires no pre-precessing steps to reduce the feature dimension, and can be computed efficiently. Furthermore, to deal with the non-linearity of the person's appearance, a kernel version can be developed easily to further boost the matching performance within the null space. It therefore offers a perfect solution to the challenging person Re-ID problem. In addition to formulating the NSFT model as a fully supervised model to solve the person Re-ID problem, the semi-supervised setting is also extended to further alleviate the effects of the SSS problem by exploiting unlabelled data abundant in Re-ID applications.

3.2 Problem Definition

For the problem of cross-view matching, each object is captured by different views. Usually, the representation of objects from different views are significantly different from each other. The goal of cross-view matching is to eliminate the view discrepancy and ultimately matching the different views of same object.

Given a set of N training data denoted as $\mathbf{X} \in \mathbb{R}^{d \times N}$. Each column of the data descriptor matrix $\mathbf{X}$, $\mathbf{x}_i$ is a feature vector representing the i-th training sample. In the case of person re-id, this feature vector is extracted from a person detection box and contains appearance information about the person, and its dimension d is typically very high. This chapter assumes that each data point belongs to one of C classes, i.e. C different identities. The objective of learning a discriminative null space is to learn a projection matrix $\mathbf{W} \in \mathbb{R}^{d \times m}$ to project the original high-dimensional feature vector $\mathbf{x}_i$ into a lower-dimensional one $\mathbf{y}_i \in \mathbb{R}^m$ with $m < d$. Person Re-ID can then be performed by computing the Euclidean distance between two projected vectors in the learned discriminative null space.

3.3 Methodology

3.3.1 Foley-Sammon transform

The learned null Foley-Sammon transform (NFST) space is closely related to linear discriminant analysis (LDA), also known as Foley-Sammon transform (FST) (Foley and Sammon, 1975). So before we formulate NFST, let us first briefly revisit FST.

The objective of FST is to learn a projection matrix $\mathbf{W} \in \mathbb{R}^{d \times m}$ so that each column, denoted as $\mathbf{w}$, is an optimal discriminant direction that maximises the Fisher discriminant criterion:

$$\mathcal{J}(\mathbf{w}) = \frac{\mathbf{w}^\top \mathbf{S}_b \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_w \mathbf{w}}, \tag{3.1}$$

where $\mathbf{S}_b$ is the between-class scatter matrix and $\mathbf{S}_w$ is the within-class scatter matrix. The optimisation of Eq. (3.1) can be done by solving the following generalised eigen-problem:

$$\mathbf{S}_b \mathbf{w} = \lambda \mathbf{S}_w \mathbf{w}. \tag{3.2}$$

If $\mathbf{S}_w$ is non-singular, $C-1$ eigenvectors $\mathbf{w}^{(1)}, ..., \mathbf{w}^{(c-1)}$ can be computed corresponding to the $C-1$ largest eigenvalues of $\mathbf{S}_w^{-1}\mathbf{S}_b$. Using them as the columns, the projection matrix $\mathbf{W}$ can project the original data into a $C-1$ dimensional discriminative subspace where the C classes become maximally separable. However, in the small sample size case, we have $d > N$; as a result, $\mathbf{S}_w$ is singular. FST thus runs in numerical problems and common solutions include reducing d by PCA or adding a regularisation term to $\mathbf{S}_w$. In (Guo et al., 2006), a more principled way to overcome the SSS problem in FST is proposed, termed as Null Foley-Sammon transform (NFST).

3.3.2 Null foley-sammon transform

NFST aims to learn a discriminative subspace where the training data points of each of the C classes are collapsed to a single point, resulting in C points in the space. In order to make this subspace discriminative, these C points should not further collapse to a single point. Formally, we aim to learn the optimal projection matrix $\mathbf{W}$ so that each of its column $\mathbf{w}$ satisfies the following two conditions:

$$\mathbf{w}^\top \mathbf{S}_w \mathbf{w} = 0, \tag{3.3}$$

$$\mathbf{w}^\top \mathbf{S}_b \mathbf{w} > 0. \tag{3.4}$$

That is, it satisfies zero within-class scatter and positive between-class scatter. This guarantees the best separability of the training data in the sense of Fisher discriminant criterion. Such a linear projecting direction $\mathbf{w}$ is called Null Projecting Direction (NPD) (Guo et al., 2006).

Next, we show that a NPD must lie in the null space of $\mathbf{S}_w$. In particular, we have the following Lemma:

Lemma 1. *Let* $\mathbf{W}$ *be a projection matrix which maps a sample* $\mathbf{x}$ *into the null space of* $\mathbf{S}_w$*, where the null space is spanned by the orthonormal set of* $\mathbf{W}$*, that is,* $\mathbf{S}_w\mathbf{W} = 0$*. If all samples are mapped into the null space of* $\mathbf{S}_w$ *through* $\mathbf{W}$*, the within-class scatter matrix* $\widehat{\mathbf{S}}_w$ *of the mapped samples is a complete zero matrix.*

Proof. Let $\mathbf{x}_n^c$ be the n^{th} sample of the $c^{th} \in \{1, ...C\}$ class which has N_c samples in total. $\mathbf{y}_n^c$ denote the mapped feature vector through $\mathbf{W}$. We have:

$$\begin{aligned}
\widehat{\mathbf{S}}_w &= \sum_{c=1}^{C}\sum_{n=1}^{N_c}(\mathbf{y}_n^c - \overline{\mathbf{y}}^c)(\mathbf{y}_n^c - \overline{\mathbf{y}}^c)^\top \\
&= \sum_{c=1}^{C}\sum_{n=1}^{N_c}(\mathbf{W}^\top\mathbf{x}_n^c - \mathbf{W}^\top\mu^c)(\mathbf{W}^\top\mathbf{x}_n^c - \mathbf{W}^\top\mu^c)^\top \\
&= \mathbf{W}^\top\sum_{c=1}^{C}\sum_{n=1}^{N_c}(\mathbf{x}_n^c - \mu^c)(\mathbf{x}_n^c - \mu^c)^\top\mathbf{W} \\
&= \mathbf{W}^\top\mathbf{S}_w\mathbf{W} = \mathbf{0}
\end{aligned}$$

where $\mathbf{y}_n^c = \mathbf{W}^\top\mathbf{x}_n^c$, $\overline{\mathbf{y}}^c = \mathbf{W}^\top\mu^c$, $\mu^c = \frac{1}{N_c}\sum_{n=1}^{N_c}\mathbf{x}_n^c$, N_c is the number of samples in class c, and μ^c is the mean vector of all data belonging to the class c.

Now with Lemma 1, we know that Eq. (3.3) holds as long as $\mathbf{w}$ is from the null space of $\mathbf{S}_w$. Next we take a look the condition in the inequality (3.4). It is easy to see that when Eq. (3.3) holds, (3.4) also holds if:

$$\mathbf{w}^\top\mathbf{S}_t\mathbf{w} > 0, \tag{3.5}$$

where $\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w$ is the total scatter matrix. We now denote the null space of the $\mathbf{S}_t$ and $\mathbf{S}_w$as:

$$\mathbf{Z}_t = \left\{\mathbf{z} \in \mathbb{R}^d \mid \mathbf{S}_t\mathbf{z} = 0\right\}, \tag{3.6}$$

$$\mathbf{Z}_w = \left\{\mathbf{z} \in \mathbb{R}^d \mid \mathbf{S}_w\mathbf{z} = 0\right\}, \tag{3.7}$$

and their orthogonal complements as $\mathbf{Z}_t^\perp$ and $\mathbf{Z}_w^\perp$ respectively. Now since $\mathbf{S}_b$ is non-negative definite, we can see that in order for the NPDs to satisfy both Eqs. (3.3) and (3.4) simultaneously, they must lie in the shared space between $\mathbf{Z}_w$ and $\mathbf{Z}_t^\perp$, that is:

$$\mathbf{w} \in (\mathbf{Z}_t^\perp \cap \mathbf{Z}_w). \tag{3.8}$$

It has been proved in (Guo et al., 2006) that there are precisely $C-1$ NPDs $\mathbf{w}$ that satisfy both Eq. (3.3) and (3.4). In other words, the discriminative null space we are looking for has $m = C-1$ dimensions.

3.3.3 Learning the discriminative null space

Let $\mathbf{X}_w$ be the matrix consisting of vectors $\mathbf{x}_i^c - \mu^c$. $\mathbf{X}_t$ be the matrix consisting of vectors $\mathbf{x}_i - \mu$ with $\mu = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$. We then have,

$$\mathbf{S}_w = \frac{1}{N}\mathbf{X}_w\mathbf{X}_w^\top, \quad \mathbf{S}_t = \frac{1}{N}\mathbf{X}_t\mathbf{X}_t^\top \tag{3.9}$$

Now we know where to look for the NPDs – the shared space between $\mathbf{Z}_w$ and $\mathbf{Z}_t^\perp$. Next, we shall see how to compute them. Let us first take a look at how to compute $\mathbf{w}$ that satisfies $\mathbf{w} \in \mathbf{Z}_t^\perp$. First we notice that:

$$\begin{aligned}\mathbf{Z}_t &= \left\{\mathbf{z} \in \mathbb{R}^d \mid \mathbf{S}_t\mathbf{z} = 0\right\} = \left\{\mathbf{z} \in \mathbb{R}^d \mid \mathbf{z}^\top\mathbf{S}_t\mathbf{z} = 0\right\} \\ &= \left\{\mathbf{z} \in \mathbb{R}^d \mid (\mathbf{X}_t^\top\mathbf{z})^\top\mathbf{X}_t^\top\mathbf{z} = 0\right\} \\ &= \left\{\mathbf{z} \in \mathbb{R}^d \mid \mathbf{X}_t^\top\mathbf{z} = 0\right\}.\end{aligned}$$

Hence, $\mathbf{Z}_t^\perp$ is the subspace spanned by zero-mean data $\mathbf{x}_i - \mu$. We can obtain the orthonormal basis $\mathbf{U} = [\mathbf{u}^{(1)}, ..., \mathbf{u}^{(N-1)}]$ of the zero-mean data using Gram-Schmidt orthonormalisation, then represent each solution $\mathbf{w}$ as:

$$\mathbf{w} = \beta_1\mathbf{u}^{(1)} + ... + \beta_{N-1}\mathbf{u}^{(N-1)} = \mathbf{U}\beta, \tag{3.10}$$

Note that there are $N-1$ basis vectors because the rank of $\mathbf{S}_t$ is $N-1$.

So now after expressing $\mathbf{w}$ using Eq. (3.10), it must satisfy $\mathbf{w} \in \mathbf{Z}_t^\perp$. The next step is the make it also satisfy $\mathbf{w} \in \mathbf{Z}_w$. This can be achieved by substituting Eq. (3.10) into Eq. (3.3) and solve the following eigen-problem:

$$(\mathbf{U}^\top\mathbf{S}_w\mathbf{U})\beta = \mathbf{0}, \tag{3.11}$$

for which we know that $C-1$ solutions $\beta^{(1)}, ..., \beta^{(C-1)}$ exist, giving $C-1$ NPDs, $\mathbf{U}\beta$.

In summary, the problem of learning the discriminative null space boils down to solving an eigen-problem which has a closed-form solution and can be solved very efficiently. Importantly, the whole optimisation algorithm has no free parameter to tune.

3.3.4 Kernelisation

The NFST model is a linear model. It has been demonstrated (Xiong et al., 2014) that many distance metric learning or discriminative subspace based methods for person Re-ID benefit from kernelisation because of the non-linearity in person's appearance. In the following we describe how the discriminative null space can be kernelised.

Given a kernel function $k(\mathbf{x}_i, \mathbf{x}_i) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$, where $\Phi(\mathbf{x}_i)$ maps $\mathbf{x}_i$ to an implicit higher dimensional space, we can compute the data kernel matrix $\mathbf{K} \in \mathbb{R}^{N \times N}$ for training data $\mathbf{X}$ as $\mathbf{K} = \Phi(\mathbf{X})^\top \Phi(\mathbf{X})$. Now the within-class scatter matrix $\mathbf{S}_w$ and total-class scatter matrix $\mathbf{S}_t$ can be kernelised as:

$$\mathbf{K}_w = \mathbf{K}(\mathbf{I} - \mathbf{L})(\mathbf{I} - \mathbf{L})^\top \mathbf{K},$$
$$\mathbf{K}_t = \mathbf{K}(\mathbf{I} - \mathbf{M})(\mathbf{I} - \mathbf{M})^\top \mathbf{K},$$

where $\mathbf{I}$ is an $N \times N$ identity matrix, $\mathbf{L}$ is a block diagonal matrix with block sizes equal to the number of data points N_c for each class $c \in \{1, ...C\}$ and $\mathbf{M}$ is an $N \times N$ matrix with all entries equal to $\frac{1}{N}$.

Now to write Eq. (3.11) in its kernelised form, we need to replace $\mathbf{S}_w$ with $\mathbf{K}_w$, and compute the orthonormal basis of $\mathbf{K}_t$ to replace $\mathbf{U}$. The orthonormal basis of $\mathbf{K}_t$ can be computed using kernel PCA. First, we compute the centred kernel matrix $\widetilde{\mathbf{K}}$. Second, the eigendecomposition of $\widetilde{\mathbf{K}}$ is written as $\mathbf{K}_t = \mathbf{V}\mathbf{E}\mathbf{V}^\top$ with $\mathbf{E}$ being the diagonal matrix containing $N-1$ non-zero eigenvalues and $\mathbf{V}$ containing the corresponding eigenvectors in its columns. Now the scaled eigenvectors $\widetilde{\mathbf{V}} = \mathbf{V}\mathbf{E}^{-1/2}$ contain coefficients for the kernelised orthonormal basis used to replace $\mathbf{U}$ in Eq. (3.11). Let

$$\mathbf{H} = ((\mathbf{I} - \mathbf{M})\widetilde{\mathbf{V}})^\top \mathbf{K}(\mathbf{I} - \mathbf{L}), \tag{3.12}$$

and with Eq. (3.9), we can rewrite Eq. (3.11) as:

$$\mathbf{H}\mathbf{H}^\top \beta = \mathbf{0}. \tag{3.13}$$

By solving the eigen-problem Eq. (3.13), we obtain the final $C-1$ null projection directions (NPDs) as:

$$\mathbf{w}^{(i)} = ((\mathbf{I} - \mathbf{M})\widetilde{\mathbf{V}})^\top \beta^{(i)} \quad \forall i = 1, ..., C-1. \tag{3.14}$$

3.3.5 Semi-supervised learning

The NFST method is a fully supervised method. When applied to the problem of Re-ID, the labelled training set is used to learn the projection $\mathbf{W}$. The test data are then projected into the same subspace and matched by computing the Euclidean distance between a query sample and a set of gallery samples.

In a real-world application scenario, the labelled training data are scarce but there are often plenty of unlabelled data (person images collected from different views) that can be used to alleviate the small sample size problem. To this end, the NFST method is extended to the semi-supervised setting. More specifically, given a training set $\mathbf{X}$ contains a labelled subset $\mathbf{X}^l$ of N^l samples and an unlabelled subset $\mathbf{X}^u$ of N^u samples. Using the NFST method described above, we can first learn an initial projection matrix $\mathbf{W}^0$ using $\mathbf{X}^l$ only. Then $\mathbf{X}^u$ is projected to the lower-dimensional subspace through $\mathbf{W}^0$ and becomes $\mathbf{Y}^u_{\mathbf{W}^0}$. To utilise the unlabelled data $\mathbf{X}^u$, we use their projections $\mathbf{Y}^u_{\mathbf{W}^0}$ to build a cross-view correspondence matrix $\mathbf{A} \in \mathbb{R}^{N_u \times N_u}$ which captures the identity relationship for the unlabelled people across views. Note, since the data are unlabelled, the true cross-view correspondence relationship is unknown. We therefore use $\mathbf{A}$ to represent a soft cross-view correspondence relationship. That is, each person in one view can correspond to multiple people in another view depending on their visual similarity in the learned discriminative subspace parameterised by $\mathbf{W}^0$. To this end, we first construct a k-nearest-neighbour (k-nn) graph G across camera views with N_u vertices, where each vertex represents a unlabelled data point. $\mathbf{A}$ is then computed as the weight matrix of G using a heat kernel. With this k-nn graph, we then create pseudo-classes, each consisting one vertex from one view and its k-nearest-neighbours from the other view. Next these pseudo-classes are augmented with the labelled classes in $\mathbf{X}^l$ to create a new training set, denoted $\mathbf{P}$, on which a new project matrix $\mathbf{W}^1$ is computed using NFST. Re-learning the projection matrix runs iteratively till the average distance for the k-nearest-neighbours stop decreasing. In our experiments, we found that the algorithm converges rapidly with less iterations.

This semi-supervised learning is essentially based on self-training, a popular strategy taken by many semi-supervised learning methods (Zhu, 2005). For any self-training based methods, preventing model drift is of paramount importance. Apart from examining the average distance for the k-nearest-neighbours, another measure taken is to rank the k-nearest-neighbours and take only the top f percent with the smallest distance to create the pseudo-classes. The complete

Algorithm 1: Semi-supervised null space learning

Input: $\mathbf{X}^l$, $\mathbf{X}^u$, k, $\mathbf{P}^0 = \mathbf{0}$.

Output: The learned projection $\mathbf{W}$.

1: Estimate $\mathbf{W}^0$ using $\mathbf{X}^l$;
2: $t = 0$;
3: **while** *not converged* **do**
4: project $\mathbf{X}^u$ through $\mathbf{W}^t$ to obtain $\mathbf{Y}^u_{\mathbf{W}^t}$
5: build k-nn graph G with $\mathbf{Y}^u_{\mathbf{W}^t}$
6: take top f percent to create the pseudo-classes $\mathbf{P}^{t+1}$
7: learn $\mathbf{W}^{t+1}$ with $\mathbf{X}^l + \mathbf{P}^{t+1}$
8: $t = t + 1$
9: **end while**

semi-supervised null space learning algorithm is summarised in Algorithm. 1.

3.4 Experiments

3.4.1 Datasets

Datasets Five widely used datasets are selected for experiments, including the three largest benchmarks available (CUHK01, CUHK03, and Market1501).

VIPeR (Gray et al., 2007) contains 632 identities and each has two images captured outdoor from two views with distinct view angles. All images are scaled to 128 $\times$ 48 pixels. The 632 people's images are randomly divided into two equal halves, one for training and the other for testing. This is repeated for 10 times and the averaged performance is reported.

PRID2011 (Hirzer et al., 2011) consists of person images recorded from two cameras. Specifically, it has two camera views. View *A* captures 385 people, whilst View *B* contains 749 people. Only 200 people appear in both views. The single shot version of the dataset is used in our experiments as in (Hirzer et al., 2012): In each data split, 100 people with one image from each view are randomly chosen from the 200 present in both camera views for the training set, while the remaining 100 of View *A* are used as the probe set, and the remaining 649 of View *B* are used as gallery. Experiments are repeated over the 10 splits provided in (Hirzer et al., 2012).

CUHK01 (Li and Wang, 2013) contains 971 identities with each person having two images in each camera view. All the images are normalised to 160 $\times$ 60 pixels. Following the standard set-

ting, images from camera A are used as probe and those from camera B as gallery. We randomly partition the dataset into 485 people for training and 486 for testing (multi-shot) following (Liao et al., 2015; Zhao et al., 2014), again over 10 trials.

CUHK03 (Li et al., 2014) contains 13,164 images of 1,360 identities, captured by six surveillance cameras with each person only appearing in two views. It provides both manually labelled pedestrian bounding boxes and bounding boxes automatically detected by the deformable-part-model (DPM) detector (Felzenszwalb et al., 2010). A real-world Re-ID system has to rely on a person detector; the latter version of the data is thus ideal for testing performance given detector errors. We report results on both of the manually labelled and detected person images. The 20 training/test splits provided in (Li et al., 2014) is used under and the single-shot setting as in (Liao et al., 2015) – two images are randomly chosen for testing; one is for probe and the other for gallery.

Market1501 (Zheng et al., 2015) is the biggest Re-ID benchmark dataset to date, containing 32,668 detected person bounding boxes of 1,501 identities. Each identity is captured by six cameras at most, and two cameras at least. During testing, for each identity, one query image in each camera is selected, therefore multiple queries are used for each identity. Note that, the selected 3,368 queries in (Zheng et al., 2015) are hand-drawn, instead of DPM-detected as in the gallery. Each identity may have multiple images under each camera. We use the provided fixed training and test set, under both the single-query and multi-query evaluation settings.

3.4.2 Settings

Feature Representations By default the recently proposed Local Maximal Occurrence (LOMO) features (Liao et al., 2015) are used for person representation. The descriptor has 26,960 dimensions. To test our method's ability to fuse different representations, we also consider another histogram-based image descriptor proposed in (Lisanti et al., 2014b). These include colour histogram, HOG and LBP which are concatenated resulting in 5138 dimensions. In addition, deep learning feature (Li et al., 2017) is also adopted for comparing with deep learning based method.

Evaluation metrics Cumulated Matching Characteristics (CMC) curve is adopted to evaluate the performance of person re-identification methods for all datasets in this chapter. Note that for the Market1501 dataset, since there are on average 14.8 cross-camera ground truth matches for each query, we additionally use mean average precision(mAP) as in (Zheng et al., 2015) to evaluate the performance.

Parameter setting There is no free parameter to tune for our model. However, with the kernelisation, kernel selection is necessary. Unless stated otherwise, RBF kernel is used with the kernel width determined automatically using the mean pairwise distance of samples. For other compared methods, different model specific parameters have to be tuned carefully to report the highest results. Note that under the semi-supervised null space learning algorithm, there are free parameters: the value of k in the k-nn graph is fixed to 3 for all experiments. The percentage of neighbours f kept for creating pseudo classes are fixed at 40%. We found that the results are not sensitive to the values of these parameters.

3.4.3 Fully supervised learning results

For the fully supervised setting, all the labels of the training data are used for model learning. For different datasets, we select different most representative and competitive alternative methods for comparison.

Results on VIPeR We first evaluate our method against the state-of-the-art on VIPeR. We compare with 17 existing methods. Among them, the distance metric learning based methods are RPLM (Hirzer et al., 2012), MtMCML (Ma et al., 2014), Mid-level Filter (Zhao et al., 2014), SCNCD (Yang et al., 2014), Similarity Learning (Chen et al., 2015), LADF (Li et al., 2013), ITML (Davis et al., 2007), LMNN (Weinberger et al., 2005), KISSME (Koestinger et al., 2012), and MCML (Globerson and Roweis, 2005), whilst the others are discriminative subspace learning based methods including kCCA (Lisanti et al., 2014b), MFA (Xiong et al., 2014), kLFDA (Xiong et al., 2014), and XQDA (Liao et al., 2015). Note that XQDA can be considered as hybrid between metric learning and subspace learning. In addition, deep learning based model is also compared (Ahmed et al., 2015). For fair comparison, whenever possible (i.e. code is available and features can be replaced), we compare with these methods using the same LOMO features. Otherwise, the reported results are presented.

From the results shown in Table 3.1, we can have the following observations: (1) Our method achieves the highest performance when a single type of features are used (Rank 1 of 42.28% compared to the closest competitor XQDA (Liao et al., 2015) which gives 40.00%). (2) For fair comparison against methods which fuse more than one types of features (Paisitkriangkrai et al., 2015) or more than one models (Zhao et al., 2014), we also present our result obtained by a simple score-level fusion using the two types of features described earlier. Our method (Ours (Fusion)) beats the nearest rival (Paisitkriangkrai et al., 2015) by over 5% on Rank 1. (3) The discriminative

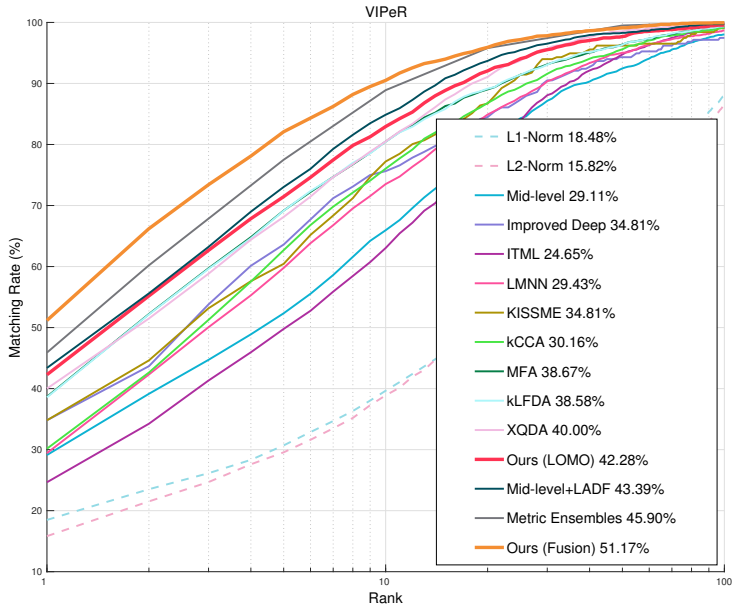


Figure 3.2: CMC curve comparison between the proposed method and the existing state-of-the-art models on VIPeR.

subspace learning based methods seem to be more competitive compared with the distance metric learning based methods. Note that all of them have been kernelised and we observe a significant drop in performance without kernelisation. This confirms the conclusion drawn in (Xiong et al., 2014) that kernelisation is critical for addressing the non-linearity problem in re-id. (4) The most related methods MCML (Globerson and Roweis, 2005) and MtMCML (Ma et al., 2014) yeild much poorer results[1], indicating that the principle of collapsing same-class samples is better realised in a subspace learning framework which provides an exact and closed-form solution. (5) The deep learning based method (Ahmed et al., 2015) does not work well on this small dataset despite the fact that the model has been pre-trained on the far-larger CUHK01+CUHK03 datasets. This suggests that the model learned from other datasets are not transferable by the simple model fine-tuning strategy and small sample size remains a bottle-neck for applying deep learning to Re-ID. Fig. 3.2 shows the CMC curve comparison between proposed method and existing state-of-the-art models on VIPeR.

Results on PRID2011 We compare the state-of-the-art (Hirzer et al., 2012; Paisitkriangkrai et al., 2015) results reported on PRID2011 in Table 3.2. With access to the implementation codes, we also compare with the methods in (Liao et al., 2015; Xiong et al., 2014; Lisanti et al., 2014b) using the same LOMO features. The results show clearly with a single feature type, our method

[1]The result of MCML is from (Ma et al., 2014) using different features. We did have access to the code of MCML. However, no matter how hard we try, it would not converge to a meaningful solution using the higher-dimensional LOMO features.

Rank	**1**	**5**	**10**	**20**
RPLM (Hirzer et al., 2012)	27.00	55.30	69.00	83.00
MtMCML (Ma et al., 2014)	28.83	59.34	75.82	88.51
MCML (Globerson and Roweis, 2005)	20.19	47.31	63.96	77.69
Mid-level (Zhao et al., 2014)	29.11	52.34	65.95	79.87
SCNCD (Yang et al., 2014)	37.80	68.50	81.20	90.40
LADF (Li et al., 2013)	30.22	64.70	78.92	90.44
Improved Deep (Ahmed et al., 2015)	34.81	63.61	75.63	84.49
Similarity Learning (Chen et al., 2015)	36.80	70.40	**83.70**	91.70
ITML (LOMO) (Davis et al., 2007)	24.65	49.78	63.04	78.39
LMNN (LOMO) (Weinberger et al., 2005)	29.43	59.78	73.51	84.91
KISSME (LOMO) (Koestinger et al., 2012)	34.81	60.44	77.22	86.71
kCCA (LOMO) (Lisanti et al., 2014b)	30.16	62.69	76.04	86.80
MFA (LOMO) (Xiong et al., 2014)	38.67	69.18	80.47	89.02
kLFDA (LOMO) (Xiong et al., 2014)	38.58	69.15	80.44	89.15
XQDA (LOMO) (Liao et al., 2015)	40.00	68.13	80.51	91.08
Ours (LOMO)	**42.28**	**71.46**	82.94	**92.06**
Mid-level+LADF (Zhao et al., 2014)	43.39	73.04	84.87	93.70
Metric Ensembles (Paisitkriangkrai et al., 2015)	45.90	77.50	88.90	95.80
Ours (Fusion)	**51.17**	**82.09**	**90.51**	**95.92**

Table 3.1: Fully supervised results on VIPeR

is the state-of-the-art; when fusing two types of features, the result is improved dramatically (over 10% increase on both Rank 1 and 5), and significantly higher than the reported results of the feature fusion method in (Paisitkriangkrai et al., 2015), which fuses four different types of features including the deep convolutional neural network (CNN) features.

Rank	**1**	**5**	**10**	**20**
L1-Norm (LOMO)	7.20	17.20	24.00	27.50
L2-Norm (LOMO)	16.30	30.0	37.90	47.90
RPLM (Hirzer et al., 2012)	15.00	32.00	42.00	54.00
kCCA (LOMO) (Lisanti et al., 2014b)	14.30	37.40	47.60	62.50
MFA (LOMO) (Xiong et al., 2014)	22.30	45.60	57.20	68.20
kLFDA (LOMO) (Xiong et al., 2014)	22.40	46.50	58.10	68.60
XQDA (LOMO) (Liao et al., 2015)	26.70	49.90	61.90	73.80
Ours (LOMO)	**29.80**	**52.90**	**66.00**	**76.50**
Metric Ensembles (Paisitkriangkrai et al., 2015)	17.90	39.00	50.00	62.00
Ours (Fusion)	**40.90**	**64.70**	**73.20**	**81.00**

Table 3.2: Fully supervised results on PRID2011

Results on CUHK01 & CUHK03 Compared with VIPeR and PRID2011, these two datasets are much bigger with thousands of training samples. However, the sample size is still much smaller than the feature dimension, i.e. the SSS problem still exists. Table 3.3 shows that on CUHK01, our method beats all compared existing methods at low ranks and when two types of features are fused, the margin is significant. As for CUHK03, there are two versions: the one with manually cropped person images, and the one with bounding boxes produced by a detector. The latter obviously is harder as reflected by the decrease of matching accuracy for all compared methods. But it is also a better indicator of real-world performance. It can be seen from Table 3.4 that, as expected, on this much larger dataset, the deep learning based model (Ahmed et al., 2015) with its millions of parameters becomes much more competitive – with manually cropped images, our result with single feature type is higher on Rank 1 but lower on other ranks. However, with the detector boxes, our method is less affected and outperforms the deep model in (Ahmed et al., 2015) by a big margin. In addition, our performance is further boosted by fusing two types of features. Fig. 3.3 shows the CMC curve comparison between proposed method and existing

state-of-the-art models on CUHK01.

Rank	**1**	**5**	**10**	**20**
SalMatch (Zhao et al., 2013a)	28.45	45.85	55.67	67.95
Mid-level Filter (Zhao et al., 2014)	34.30	55.06	64.96	74.94
Improved Deep (Ahmed et al., 2015)	47.53	71.60	80.25	87.45
kCCA (LOMO) (Lisanti et al., 2014b)	56.30	80.66	87.94	93.00
MFA (LOMO) (Xiong et al., 2014)	54.79	80.08	87.26	92.72
kFLDA (LOMO) (Xiong et al., 2014)	54.63	80.45	86.87	92.02
XQDA (LOMO) (Liao et al., 2015)	63.21	83.89	**90.04**	94.16
Ours (LOMO)	**64.98**	**84.96**	89.92	**94.36**
Metric Ensembles (Paisitkriangkrai et al., 2015)	53.40	76.40	84.40	90.50
Ours (Fusion)	**69.09**	**86.87**	**91.77**	**95.39**

Table 3.3: Fully supervised results on CUHK01

Dataset	CUHK03 (Manual)				CUHK03 (Detected)			
Rank	**1**	**5**	**10**	**20**	**1**	**5**	**10**	**20**
DeepReID (Li et al., 2014)	20.65	51.50	66.50	80.00	19.89	50.00	64.00	78.50
Improved Deep (Ahmed et al., 2015)	54.74	**86.50**	**93.88**	**98.10**	44.96	76.01	83.47	93.15
XQDA (LOMO) (Liao et al., 2015)	52.20	82.23	92.14	96.25	46.25	78.90	88.55	94.25
Ours (LOMO)	**58.90**	85.60	92.45	96.30	**53.70**	**83.05**	**93.00**	**94.80**
Metric Ensembles (Paisitkriangkrai et al., 2015)	62.10	89.10	94.30	97.80	-	-	-	-
Ours (Fusion)	**62.55**	**90.05**	**94.80**	**98.10**	**54.70**	**84.75**	**94.80**	**95.20**

Table 3.4: Fully supervised results on CUHK03. '-' means that no reported results is available.

Results on Market1501 This dataset is the largest and most realistic dataset with natural detector errors abundant in the provided data as they were collected in front of a busy supermarket. The baseline presented in (Zheng et al., 2015) is not competitive because it is based on a weaker BoW features and L2-Norm distance. We compare our method with four alternatives with the same LOMO features. The results in Table 3.5 again show that our method significantly outperforms the alternatives, under both the single query and multi-query settings and with both evaluation metrics. This is despite the fact that with 12,936 training samples, the SSS problem is

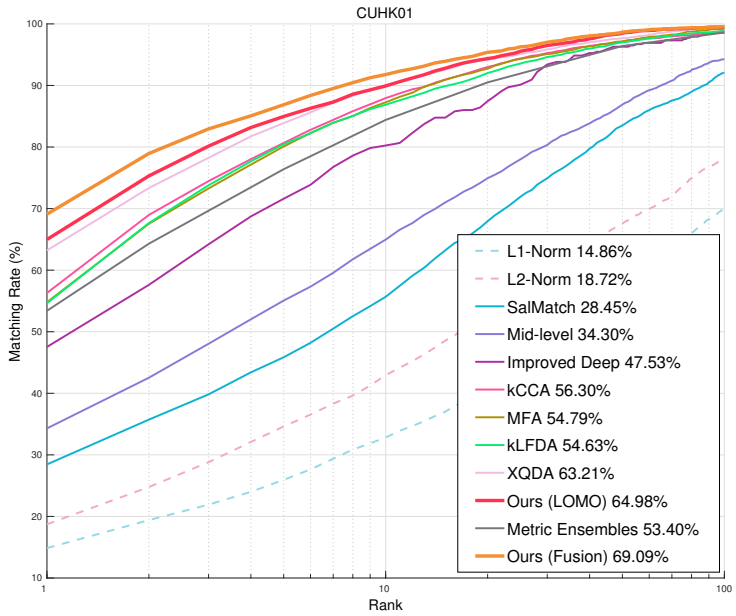


Figure 3.3: CMC curve comparison between the proposed method and the existing state-of-the-art models on CUHK01.

the least severe in this dataset.

Results with deep learning features As reviewed in Section 2.2.1, deep learning based methods have obtained impressive performance, especially on large-scale dataset Market1501. To evaluate the proposed method with deep learning features, JLML (Li et al., 2017) is selected for comparison. JLML is pre-trained on ImageNet (ILSVRC2012) for classification task. Subsequently, Market1501 is used for model fine-tuning. Particularly, $1,024$-D joint representation is obtained by concatenating the local (512D) and global (512D) feature vectors extract from end-to-end trained model JLML (Li et al., 2017). The proposed model is adopted on the extracted deep feature while JLML only use L2 distance for testing. Table 3.6 shows that our method boost almost 20 points on rank-1 single query comparing with our results on Table 3.5. However, JLML still outperforms the proposed method with a clear margin. This is due to the JLML can better discriminate identity/people under different camera views in this end-to-end trained $1,024$D dimension space. One possible extension of current model (see Section 6.2) could be integrating the proposed model into learning an end-to-end neural networks, as NFST is an extreme case of Linear Discriminative Analysis (Dorfer et al., 2015).

3.4.4 Semi-supervised learning results

For semi-supervised setting, we use the VIPeR and PRID2011 datasets. The same data splits are used as in the fully-supervised setting. The difference is that only one third of the training data

Query	*singleQ*		*multiQ*	
Evaluation metrics	**Rank-1**	**mAP**	**Rank-1**	**mAP**
Baseline (Zheng et al., 2015)	34.38	14.10	42.64	19.47
Baseline (+HS) (Zheng et al., 2015)	-	-	47.25	21.88
KISSME (LOMO) (Koestinger et al., 2012)	40.50	19.02	-	-
MFA (LOMO) (Xiong et al., 2014)	45.67	18.24	-	-
kLFDA (LOMO) (Xiong et al., 2014)	51.37	24.43	52.67	27.36
XQDA (LOMO) (Liao et al., 2015)	43.79	22.22	54.13	28.41
Ours (LOMO)	**55.43**	**29.87**	**67.96**	**41.89**
Ours (Fusion)	**61.02**	**35.68**	**71.56**	**46.03**

Table 3.5: Fully supervised results on Market1501

Query	*singleQ*		*multiQ*	
Evaluation metrics	**Rank-1**	**mAP**	**Rank-1**	**mAP**
JLML (Li et al., 2017)	**85.1**	**65.5**	**89.7**	**74.5**
Ours (Deep)	79.5	59.4	84.9	68.4

Table 3.6: Fully supervised results comparing with deep learning based method (Li et al., 2017) on Market1501

are labelled following the setting in (Liu et al., 2014; Kodirov et al., 2015). For comparison, apart from the state-of-the-art methods in (Liu et al., 2014; Kodirov et al., 2015), we also choose three subspace learning based methods trained on the labelled data only.

The results in Table 3.7 show that the performance of our method is clearly superior to that of the compared alternatives. The advantage is more significant on PRID2011. This dataset has only 100 pairs or 200 training samples; with only one third of them labelled, the SSS problem becomes the most acute than any experiment we conducted before in Section 3.4.3. Comparing Table 3.7 with Table 3.2, it is apparent that the performance of all three compared subspace learning methods, kCCA, kLFDA, and XQDA degrades drastically. In contrast, the performance of our method decreases much more gracefully from 29.80% to 24.70% on Rank 1. This is partly because our self-training based method can exploit the unlabelled data. It also shows that it can better cope with the SSS problem in its extreme.

Dataset	VIPeR				PRID2011			
Rank	**1**	**5**	**10**	**20**	**1**	**5**	**10**	**20**
SSCDL (Liu et al., 2014)	25.60	53.70	68.20	83.60	-	-	-	-
kCCA (LOMO) (Lisanti et al., 2014b)	13.64	37.97	53.77	69.94	5.80	16.00	24.70	36.00
kLFDA (LOMO) (Xiong et al., 2014)	25.47	53.25	66.49	80.13	12.00	27.10	37.80	50.30
XQDA (LOMO) (Liao et al., 2015)	28.04	56.30	69.65	81.74	12.60	29.40	40.20	53.00
IterativeLap (LOMO) (Kodirov et al., 2015)	29.43	49.05	59.18	69.62	18.70	34.60	43.50	52.30
Ours (LOMO)	**31.68**	**59.40**	**72.78**	**84.91**	**24.70**	**46.80**	**58.20**	**68.20**

Table 3.7: Semi-supervised Re-ID results on VIPeR and PRID2011

3.4.5 Running cost

We compare the run time of our method with XQDA, kLFDA and MFA on Market1501. We calculate the overall training time over 12,936 samples and test time over 3,368 queries. All algorithms are implemented in Matlab and run on a server with 2.6GHz CPU cores and 384GB memory. Table 3.8 shows that for training, our method is the most efficiently, whilst on testing it is much slower than XQDA, but faster than kLFDA and MFA. Considering the test time is over 3,368 queries, it is more than adequate for real-time applications.

Method	Ours	XQDA (Liao et al., 2015)	kLFDA (Xiong et al., 2014)	MFA (Xiong et al., 2014)
Training	393.1	3233.8	995.2	437.8
Testing	31.3	1.6	43.4	43.2

Table 3.8: Run time comparison on Market1501 (in seconds)

3.5 Summary

This chapter proposed to solve the person Re-ID problem by learning a discriminative null space of the training samples. Compared with existing Re-ID models, the employed NFST model is much simpler, with a closed-form solution and no parameters to tune. Yet, it is very effective in dealing with the SSS problem faced by the Re-ID methods. Extensive experiments on five widely used benchmarks show that our method achieves the state-of-the-art performance on all of them under both fully supervised and semi-supervised settings.

On the other hand, although the mostly recent state-of-the-art results are obtained by deep Re-ID models, the contribution of this chapter in the era of deep learning still can not be ignored. Deep learning based model often overfit on the small datasets due to the SSS problem. A possible way is to combine these two: a deep model trained using large dataset for extracting features and the discriminative null space model to adapt to the target small dataset.

Chapter 4

Cross-View Matching for Zero-Shot Learning by Deep Embedding Learning

---

In Chapter 3, a distance metric learning framework was formulated for cross-view matching problem, but one key assumption is that both of the *views* pertain to same modality. In this chapter, a novel deep embedding model for zero-shot learning (ZSL) problem is proposed, which makes it capable of dealing with the case that each *view* is associated with different modality. Specifically, the proposed deep neural network based embedding model differs from the existing models in that: To alleviate the hubness problem discussed in Section 1.2, visual space is adopted as the embedding space instead of the semantic space or an intermediate space. The resulting projection direction is from the *textual* view to *visual* view. Such a direction is opposite to the one adopted by most existing models. Moreover, a theoretical analysis and some intuitive visualisations are provided to explain why this would help to counter the hubness problem. Further, this framework design also provides a natural mechanism for multiple textual views (e.g., attributes and sentence descriptions) to be fused and optimised jointly in an end-to-end manner. Extensive experiments on four benchmarks show that the proposed method beats all the state-of-the-art models presented to date, often by a clear margin.

4.1 Background

ZSL models rely on learning a joint embedding space where both textual/semantic description of object classes and visual representation of object images can be projected to for cross-view

matching. Specifically, the zero-shot learning problem can be solved if the visual view of the data (object) and its textual view are matched. Despite the success of deep neural networks that learn an end-to-end model across *visual* and *textual* views in other vision problems (e.g. image captioning), very few deep ZSL model exists and they show little advantage over ZSL models that utilise deep feature representations but do not learn an end-to-end embedding.

End-to-end learning of a deep neural network embedding based ZSL model, which is the focus of this chapter, offers a number of advantages . First, end-to-end optimisation can potentially lead to learning a better embedding space. For example, if sentence descriptions are used as the input to a neural language model such as recurrent neural networks (RNNs) for computing a semantic space, both the neural language model and the CNN visual feature representation learning model can be jointly optimised in an end-to-end fashion. Second, a neural network based joint embedding model offers the flexibility for addressing various transfer learning problems such as multi-task learning and multi-domain learning (Yang and Hospedales, 2015). Third, when multiple semantic spaces are available, this model can provide a natural mechanism for fusing the multiple modalities. However, despite all these intrinsic advantages, in practice, the few existing end-to-end deep models for ZSL in the literature (Lei Ba et al., 2015; Frome et al., 2013; Socher et al., 2013; Yang and Hospedales, 2015; Reed et al., 2016a) fail to demonstrate these advantages and yield only weaker or merely comparable performances on benchmarks when compared to non-deep learning alternatives.

In this chapter, a novel Deep neural network based Embedding Model (DEM) for ZSL is proposed, which differs from existing models in that: (1) This chapter argues that the key to make deep ZSL models succeed is to choose the right embedding space. Instead of embedding into a semantic space or an intermediate space, the output visual feature space of a CNN subnet is adopted as the embedding space. This is because that in this space, the subsequent nearest neighbour search would suffer much less from the hubness problem and thus become more effective. A theoretical analysis and some intuitive visualisations are provided to explain why this would help us counter the hubness problem. (2) A simple yet effective multi-modality fusion method is developed in our neural network model which is flexible and importantly enables end-to-end learning of the semantic space representation.

4.2 Problem definition

Assume a labelled training set of N training samples is given as $\mathcal{D}_{tr} = \{(\mathbf{I}_i, \mathbf{y}_i^u, t_i^u), i = 1, \ldots, N\}$, with associated class label set $\mathcal{T}_{tr}$, where $\mathbf{I}_i$ is the i-th training image, $\mathbf{y}_i^u \in \mathbb{R}^{L\times 1}$ is its corresponding L-dimensional semantic representation vector, $t_i^u \in \mathcal{T}_{tr}$ is the u-th training class label for the i-th training image. Given a new test image $\mathbf{I}_j$, the goal of ZSL is to predict a class label $t_j^v \in \mathcal{T}_{te}$, where t_j^v is the v-th test class label for the j-th test instance. We have $\mathcal{T}_{tr} \cap \mathcal{T}_{te} = \varnothing$, i.e., the training (seen) classes and test (unseen) classes are disjoint. Note that each class label t^u or t^v is associated with a pre-defined semantic space representation $\mathbf{y}^u$ or $\mathbf{y}^v$ (e.g. attribute vector), referred to as semantic class prototypes. For the training set, $\mathbf{y}_i^u$ is given because each training image $\mathbf{I}_i$ is labelled by a semantic representation vector representing its corresponding class label t_j^u.

As introduced in Section 1.1, zero-shot learning is also a cross-view matching problem. The textual view in the context of ZSL usually relates to the class label representation in the textual space, which means the textual *view* for same class objects are identical. Specifically, the ZSL problem can be solved if the visual view of the data (object) and its textual view are matched.

Semantic representation All three kind of semantic representation from textual view (Section 1.1) are considered in this chapter. More details about semantic representation are described in Section 4.4.

1. Attribute: Attribute vector $\mathbf{y}_i^u$ for each visual class are given, each dimension of the attribute vector is semantic meaningful which describe the object class, e.g. *black, white, blue, brown, patches, spots, stripes, furry, tail, horns, claws, tusks, smelly, flies, ...*
2. Word vector: A skip-gram language model (Mikolov et al., 2013a,b) trained on a corpus of 4.6M Wikipedia documents is used to extract fixed dimension word vectors $\mathbf{y}_i^u$ to represent each object class;
3. Sentence descriptions/captions: A neural language model (e.g. LSTM) is required to output a vector representation $\mathbf{y}_i^u$ from sentence descriptions/captions corresponding to each image (Reed et al., 2016a). More details are given in Section 4.3.3.

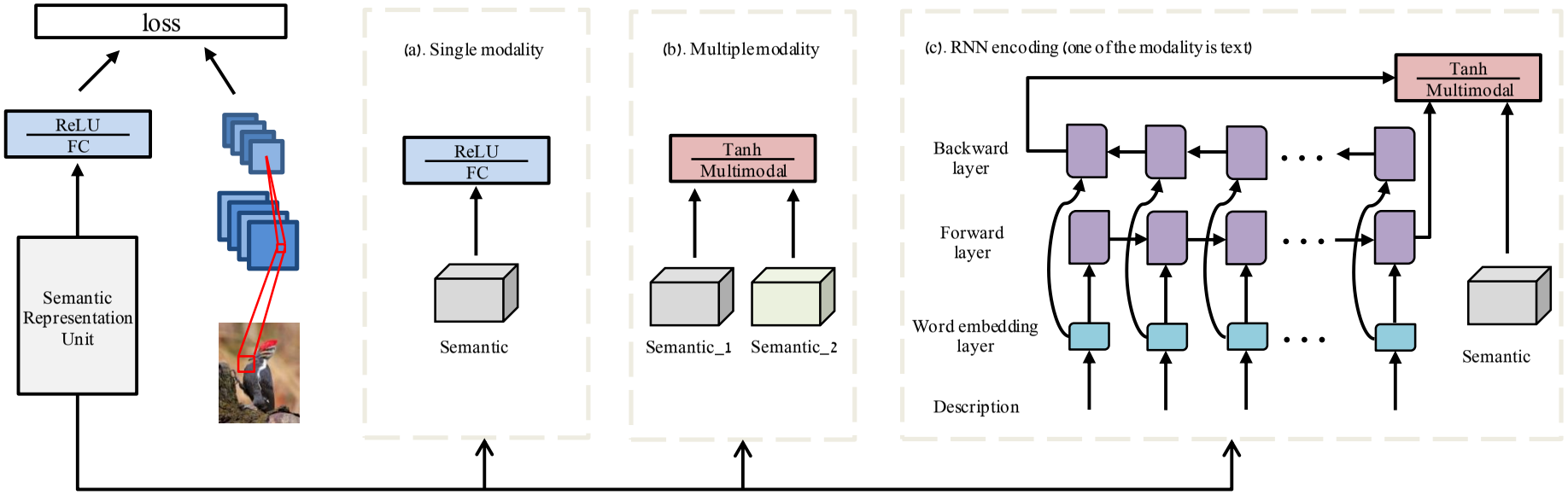


Figure 4.1: Illustration of the network architecture of our deep embedding model. The detailed architecture of the semantic representation unit in the left branch (semantic encoding subnet) is given in (a), (b) and (c) which correspond to the single modality (semantic space) case, the multiple (two) modality case, and the case where one of the modalities is text description. For the case in (c), the semantic representation itself is a neural network (RNN) which is learned end-to-end with the rest of the network.

4.3 Methodology

4.3.1 Model architecture

The architecture of the proposed model is shown in Fig. 4.1. It has two branches. One branch is the visual encoding branch, which consists of a CNN subnet that takes an image $\mathbf{I}_i$ as input and outputs a D-dimensional feature vector $\phi(\mathbf{I}_i) \in \mathbb{R}^{D\times 1}$. This D-dimensional visual feature space will be used as the embedding space where both the image content and the semantic representation of the class that the image belongs to will be embedded. The semantic embedding is achieved by the other branch which is a semantic encoding subnet. Specifically, it takes an L-dimensional semantic representation vector of the corresponding class $\mathbf{y}_i^u$ as input, and after going through two fully connected (FC) linear + Rectified Linear Unit (ReLU) layers outputs a D-dimensional semantic embedding vector. Each of the FC layer has an l_2 parameter regularisation loss. The two branches are linked together by a least square embedding loss which aims to minimise the discrepancy between the visual feature $\phi(\mathbf{I}_i)$ and its class representation embedding vector in the visual feature space. With the three losses, our objective function is as follows:

$$\mathcal{L}(\mathbf{W}_1,\mathbf{W}_2) = \frac{1}{N}\sum_{i=1}^{N} ||\phi(\mathbf{I}_i) - f_1(\mathbf{W}_2 f_1(\mathbf{W}_1\mathbf{y}_i^u))||^2 + \lambda(||\mathbf{W}_1||^2 + ||\mathbf{W}_2||^2) \tag{4.1}$$

where $\mathbf{W}_1 \in \mathbb{R}^{L\times M}$ are the weights to be learned in the first FC layer and $\mathbf{W}_2 \in \mathbb{R}^{M\times D}$ for the second FC layer. λ is the hyperparameter weighting the strengths of the two parameter regularisation losses against the embedding loss. We set $f_1(\mathbf{.})$ to be the Rectified Linear Unit (ReLU) which introduces nonlinearity in the encoding subnet (Krizhevsky et al., 2012).

After that, the classification of the test image $\mathbf{I}_j$ in the visual feature space can be achieved by simply calculating its distance to the embed prototypes:

$$v = \arg\min_{v} \mathcal{D}(\phi(\mathbf{I}_j), f_1(\mathbf{W}_2 f_1(\mathbf{W}_1\mathbf{y}^v))) \tag{4.2}$$

where $\mathcal{D}$ is a distance function, and $\mathbf{y}^v$ is the semantic space vector of the v-th test class prototype.

4.3.2 Multiple semantic space fusion

As shown in Fig. 4.1, we can consider the semantic representation and the first FC and ReLU layer together as a semantic representation unit. When there is only one semantic space considered, it is illustrated in Fig. 4.1(a). However, when more than one semantic space is used,

e.g., we want to fuse attribute vector with word vector for semantic representation of classes, the structure of the semantic representation unit is changed slightly, as shown in Fig. 4.1(b).

More specifically, we map different semantic representation vectors to a multi-modal fusion layer/space where they are added. The output of the semantic representation unit thus becomes:

$$f_2(\mathbf{W}_1^{(1)} \cdot \mathbf{y}_i^{u_1} + \mathbf{W}_1^{(2)} \cdot \mathbf{y}_i^{u_2}), \tag{4.3}$$

where $\mathbf{y}_i^{u_1} \in \mathbb{R}^{L_1 \times 1}$ and $\mathbf{y}_i^{u_2} \in \mathbb{R}^{L_2 \times 1}$ denote two different semantic space representations (e.g., attribute and word vector), "+" denotes element-wise sum, $\mathbf{W}_1^{(1)} \in \mathbb{R}^{L_1 \times M}$ and $\mathbf{W}_1^{(2)} \in \mathbb{R}^{L_2 \times M}$ are the weights which will be learned. $f_2(\cdot)$ is the element-wise scaled hyperbolic tangent function (LeCun et al., 2012):

$$f_2(x) = 1.7159 \cdot \tanh(\frac{2}{3}x). \tag{4.4}$$

This activation function forces the gradient into the most non-linear value range and leads to a faster training process than the basic hyperbolic tangent function.

4.3.3 Bidirectional LSTM encoder for description

The structure of the semantic representation unit needs to be changed again, when text description is avalialbe for each training image (see Fig. 4.1(c)). In this work, we use a recurrent neural network (RNN) to encode the content of a text description (a variable length sentence) into a fixed-length semantic vector. Specifically, given a text description of T words, $x = (x_1, \ldots, x_T)$ we use a Bidirectional RNN model (Schuster and Paliwal, 1997) to encode them. For the RNN cell, the Long-Shot Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) units are used as the recurrent units. The LSTM is a special kind of RNN, which introduces the concept of gating to control the message passing between different time steps. In this way, it could potentially model long term dependencies. Following (Graves et al., 2013b), the model has two types of states to keep track of the historical records: a cell state $\mathbf{c}$ and a hidden state $\mathbf{h}$. For a particular time step t, they are computed by integrating the current inputs x_t and the previous state $(\mathbf{c}_{t-1}, \mathbf{h}_{t-1})$. During the integrating, three types of gates are used to control the messaging passing: an input gate $\mathbf{i}_t$, a forget gate $\mathbf{f}_t$ and an output gate $\mathbf{o}_t$.

We omit the formulation of the bidirectional LSTM here and refer the readers to (Graves et al., 2013b,a) for details. With the bidirectional LSTM model, we use the final output as our encoded semantic feature vector to represent the text description:

$$f(\mathbf{W}_{\overrightarrow{\mathbf{h}}} \cdot \overrightarrow{\mathbf{h}} + \mathbf{W}_{\overleftarrow{\mathbf{h}}} \cdot \overleftarrow{\mathbf{h}}), \tag{4.5}$$

where $\overrightarrow{\mathbf{h}}$ denote the forward final hidden state, $\overleftarrow{\mathbf{h}}$ denote the backward final hidden state. $f(.) = f_1(.)$ if text description is used only for semantic space unit, and $f(.) = f_2(.)$ if other semantic space need to be fused (Sec. 4.3.2). $\mathbf{W}_{\overrightarrow{\mathbf{h}}}$ and $\mathbf{W}_{\overleftarrow{\mathbf{h}}}$ are the weights which will be learned.

In the testing stage, we first extract text encoding from test descriptions and then average them per-class to form the test prototypes as in (Reed et al., 2016a). Note that since our ZSL model is a neural network, it is possible now to learn the RNN encoding subnet using the training data together with the rest of the network in an end-to-end fashion.

4.3.4 The hubness problem

How does our model deal with the hubness problem? First we show that our objective function is closely related to that of the ridge regression formulation. In particular, if we use the matrix form and write the outputs of the semantic representation unit as $\mathbf{A}$ and the outputs of the CNN visual feature encoder as $\mathbf{B}$, and ignore the ReLU unit for now, our training objective becomes

$$\mathcal{L}(\mathbf{W}) = ||\mathbf{B} - \mathbf{W}\mathbf{A}||_F^2 + \lambda ||\mathbf{W}||_F^2, \tag{4.6}$$

which is basically ridge regression. It is well known that ridge regression has a closed-form solution $\mathbf{W} = \mathbf{B}\mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top + \lambda\mathbf{I})^{-1}$. Thus we have:

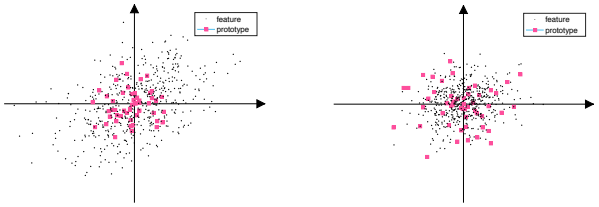
$$\begin{aligned} ||\mathbf{W}\mathbf{A}||_2 &= ||\mathbf{B}\mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top + \lambda\mathbf{I})^{-1}\mathbf{A}||_2 \\ &\leq ||\mathbf{B}||_2 ||\mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top + \lambda\mathbf{I})^{-1}\mathbf{A}||_2 \end{aligned} \tag{4.7}$$

It can be further shown that

$$||\mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top + \lambda\mathbf{I})^{-1}\mathbf{A}||_2 = \frac{\sigma^2}{\sigma^2 + \lambda} \leq 1. \tag{4.8}$$

where σ is the largest singular value of $\mathbf{A}$. So we have $||\mathbf{W}\mathbf{A}||_2 \leq ||\mathbf{B}||_2$. This means the mapped source data $||\mathbf{W}\mathbf{A}||_2$ are likely to be closer to the origin of the space than the target data $||\mathbf{B}||_2$, with a smaller variance.

Why does this matter in the context of ZSL? Figure 4.2 gives an intuitive explanation. Specifically, assuming the feature distribution is uniform in the visual feature space, Fig. 4.2(a) shows that if the projected class prototypes are slightly shrunk towards the origin, it would not change how hubness problem arises – in other words, it at least does not make the hubness issue worse.



(a) S → V (b) V → S

Figure 4.2: Illustration of the effects of different embedding directions on the hubness problem. S: semantic space, and V: visual feature space. Better viewed in colour.

However, if the mapping direction were to be reversed, that is, we use the semantic vector space as the embedding space and project the visual feature vectors $\phi(\mathbf{I})$ into the space, the training objective is still ridge regression-like, so the projected visual feature representation vectors will be shrunk towards the origin as shown in Fig. 4.2(b). Then there is an adverse effect: the semantic vectors which are closer to the origin are more likely to become hubs, i.e. nearest neighbours to many projected visual feature representation vectors. This is confirmed by our experiments (see Sec. 4.4) which show that using which space as the embedding space makes a big difference in terms of the degree/seriousness of the resultant hubness problem and therefore the ZSL performance.

Measure of hubness To measure the degree of hubness in a nearest neighbour search problem, the *skewness* of the (empirical) N_k distribution is used, following (Radovanović et al., 2010; Shigeto et al., 2015). The N_k distribution is the distribution of the number $N_k(i)$ of times each prototype i is found in the top k of the ranking for test samples (i.e. their k-nearest neighbour), and its skewness is defined as follows:

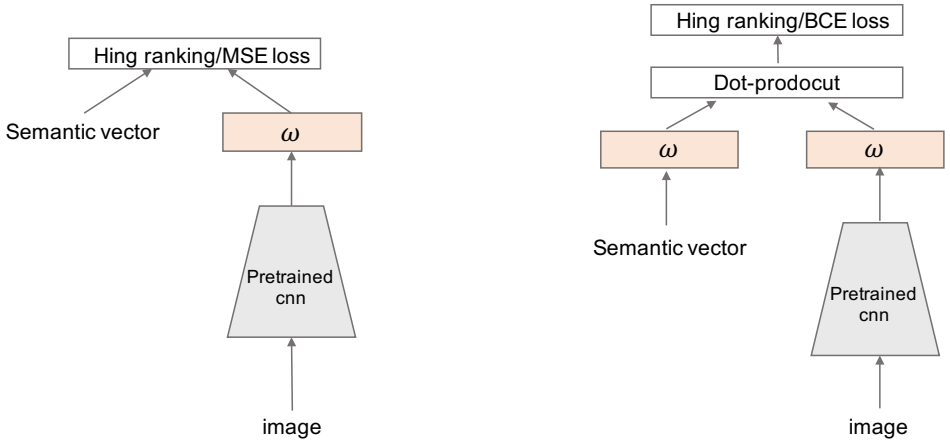
$$(N_k skewness) = \frac{\sum_{i=1}^{l}(N_k(i) - E[N_k])^3/l}{Var[N_k]^{\frac{3}{2}}}, \tag{4.9}$$

where l is the total number of test prototypes. A large *skewness* value indicates the emergence of more hubs.

4.3.5 Relationship to other deep ZSL models

We compare the proposed model with the related end-to-end neural network based models: DeViSE (Frome et al., 2013), Socher *et al.* (Socher et al., 2013), MTMDL (Yang and Hospedales, 2015), and Ba *et al.* (Lei Ba et al., 2015). Their model structures fall into two groups. In the first group (see Fig. 4.3(a)), DeViSE (Frome et al., 2013) and Socher *et al.* (Socher et al., 2013) map

the CNN visual feature vector to a semantic space by a hinge ranking loss or least square loss. In contrast, MTMDL (Yang and Hospedales, 2015) and Ba *et al.* (Lei Ba et al., 2015) fuse visual space and semantic space to a common intermediate space and then use a hinge ranking loss or a binary cross entropy loss (see Fig. 4.3(b)). For both groups, the learned embedding model will make the variance of **WA** to be smaller than that of **B**, which would thus make the hubness problem worse. In summary, the hubness will persist regardless what embedding model is adopted, as long as NN search is conducted in a high dimensional space. Our model does not worsen it, whist other deep models do, which leads to the performance difference as demonstrated in our experiments.



(a) (Frome et al., 2013; Socher et al., 2013) (b) (Yang and Hospedales, 2015; Lei Ba et al., 2015)

Figure 4.3: The architectures of existing deep ZSL models fall into two groups: (a) learning projection function ω from visual feature space to semantic space; (b) learning an intermediate space as embedding space.

4.4 Experiments

4.4.1 Dataset and settings

We follow two ZSL settings: the *old* setting and the new *GBU* setting provided by (Xian et al., 2017) for training/test splits. Under the *old* setting, adopted by most existing ZSL works before (Xian et al., 2017), some of the test classes also appear in the ImageNet 1K classes, which have been used to pretrain the image embedding network, thus violating the zero-shot assumption. In contrast, the new *GBU* setting ensures that none of the test classes of the datasets appear in the ImageNet 1K classes. Under both settings, the test set can comprise only the unseen class

samples (conventional test set setting) or a mixture of seen and unseen class samples. The latter, termed generalised zero-shot learning (GZSL), is more realistic in practice.

Datasets Four benchmarks are selected for the *old* setting: **AwA** (Animals with Attributes) (Lampert et al., 2014) consists of 30,745 images of 50 classes. It has a fixed split for evaluation with 40 training classes and 10 test classes. **CUB** (CUB-200-2011) (Wah et al., 2011) contains 11,788 images of 200 bird species. We use the same split as in (Akata et al., 2015) with 150 classes for training and 50 disjoint classes for testing. **ImageNet (ILSVRC) 2010 1K** (Russakovsky et al., 2015) consists of 1,000 categories and more than 1.2 million images. We use the same training/test split as (Mensink et al., 2012; Frome et al., 2013) which gives 800 classes for training and 200 classes for testing. **ImageNet (ILSVRC) 2012/2010**: for this dataset, we use the same setting as (Fu and Sigal, 2016), that is, ILSVRC 2012 1K is used as the training seen classes, while 360 classes in ILSVRC 2010 which do not appear in ILSVRC 2012 are used as the test unseen classes. Three datasets (Xian et al., 2017) are selected for *GBU* setting: **AwA1**, **AwA2** and **CUB**. The newly released AwA2 (Xian et al., 2017) consists of 37,322 images of 50 classes which is an extension of AwA while AwA1 is same as AwA but under the *GBU* setting.

Semantic space For **AwA**, we use the continuous 85-dimension class-level attributes provided in (Lampert et al., 2014), which have been used by all recent works. For the word vector space, we use the 1,000 dimension word vectors provided in (Fu et al., 2014, 2015a). For **CUB**, continuous 312-dimension class-level attributes and 10 descriptions per image provided in (Reed et al., 2016a) are used. For **ILSVRC 2010** and **ILSVRC 2012**, we trained a skip-gram language model (Mikolov et al., 2013a,b) on a corpus of 4.6M Wikipedia documents to extract $1,000$D word vectors for each class.

Model setting and training Unless otherwise specified, We use the Inception-V2 (Szegedy et al., 2015; Ioffe and Szegedy, 2015) as the CNN subnet in the old and conventional setting, and ResNet101 (He et al., 2016) for the GBU and generalised setting, taking the top pooling units as image embedding with dimension $D = 1024$ and 2048 respectively. The CNN subnet is pre-trained on ILSVRC 2012 1K classification without fine-tuning, the same as the recent deep ZSL works (Lei Ba et al., 2015; Reed et al., 2016a). For fair comparison with DeViSE (Frome et al., 2013), ConSE (Norouzi et al., 2014) and AMP (Fu et al., 2015b) on ILSVRC 2010, we also use the Alexnet (Krizhevsky et al., 2012) architecture and pretrain it from scratch using the 800 training classes. All input images are resized to 224×224. Fully connected layers of our

model are initialised with random weights for all of our experiments. Adam (Kingma and Ba, 2015) is used to optimise our model with a learning rate of 0.0001 and a minibatch size of 64. The model is implemented based on *Tensorflow*.

Parameter setting In the semantic encoding branch of our network, the output size of the first FC layer M is set to 300 and 700 for AwA and CUB respectively when a single semantic space is used (see Fig. 4.1(a)). Specifically, we use one FC layer for ImageNet in our experiments. For multiple semantic space fusion, the multi-modal fusion layer output size is set to 900 (see Fig. 4.1(b)). When the semantic representation was encoded from descriptions for the CUB dataset, a bidirectional LSTM encoding subnet is employed (see Fig. 4.1(c)). We use the `BasicLSTMCell` in *Tensorflow* as our RNN cell and employ ReLU as activation function. We set the input sequence length to 30; longer text inputs are cut off at this point and shorter ones are zero-padded. The word embedding size and the number of LSTM unit are both 512. Note that with this LSTM subnet, RMSprop is used in the place of Adam to optimise the whole network with a learning rate of 0.0001, a minibatch size of 64 and gradient clipped at 5. The loss weighting factor λ in Eq. (4.1) is searched by five-fold cross-validation. Specifically, 20% of the seen classes in the training set are used to form a validation set.

4.4.2 Experiments on small scale datasets

Competitors Numerous existing works reported results on AwA and CUB these two relatively small-scale datasets under old setting. Among them, only the most competitive ones are selected for comparison due to space constraint. The selected 13 can be categorised into the non-deep model group and the deep model group. All the non-deep models use ImageNet pretrained CNN to extract visual features. They differ in which CNN model is used: F_O indicates that overfeat (Sermanet et al., 2013) is used; F_G for GoogLeNet (Szegedy et al., 2015); and F_V for VGG net (Simonyan and Zisserman, 2014). The second group are all neural network based with a CNN subnet. For fair comparison, we implement the models in (Frome et al., 2013; Socher et al., 2013; Yang and Hospedales, 2015; Lei Ba et al., 2015) on AwA and CUB with Inception-V2 as the CNN subnet as in our model and (Reed et al., 2016a). The compared methods also differ in the semantic spaces used. Attributes (A) are used by all methods; some also use word vector (W) either as an alternative to attributes, or in conjunction with attributes (A+W). For CUB, recently the instance-level sentence descriptions (D) are used (Reed et al., 2016a). Note that only inductive methods are considered. Some recent methods (Zhang and Saligrama, 2016b; Fu et al.,

2014, 2015a) are tranductive in that they use all test data at once for model training, which gives them a big unfair advantage.

Comparative results on AwA under old setting From Table 4.1 we can make the following observations: (1) Our model DEM achieves the best results either with attribute or word vector. When both semantic spaces are used, our result is further improved to 88.1%, which is 7.6% higher than the best result reported so far (Zhang and Saligrama, 2016a). (2) The performance gap between our model to the existing neural network based models are particularly striking. In fact, the four models (Frome et al., 2013; Socher et al., 2013; Yang and Hospedales, 2015; Lei Ba et al., 2015) achieve weaker results than most of the compared non-deep models that use deep features only and do not perform end-to-end training. This verify our claim that selecting the appropriate visual-semantic embedding space is critical for the deep embedding models to work. (3) As expected, the word vector space is less informative than the attribute space (86.7% vs. 78.8%) even though our word vector space alone result already beats all published results except for one (Zhang and Saligrama, 2016a). Nevertheless, fusing the two spaces still brings some improvement (1.4%).

Comparative results on CUB under old setting Table 4.1 shows that on the fine-grained dataset CUB, our model also achieves the best result. In particular, with attribute only, our result of 58.3% is 3.8% higher than the strongest competitor (Changpinyo et al., 2016). The best result reported so far, however, was obtained by the neural network based DS-SJE (Reed et al., 2016a) at 56.8% using sentence descriptions. It is worth pointing out that this result was obtained using a word-CNN-RNN neural language model, whilst our model uses a bidirectional LSTM subnet, which is easier to train end-to-end with the rest of the network. When the same LSTM based neural language model is used, DS-SJE reports a lower accuracy of 53.0%. Further more, with attribute only, the result of DS-SJE (50.4%) is much lower than ours. This is significant because annotating attributes for fine-grained classes is probably just about manageable; but annotating 10 descriptions for each images is unlikely to scale to large number of classes. It is also evident that fusing attribute with descriptions leads to further improvement.

Comparative results under the GBU setting We follow the evaluation setting of (Xian et al., 2017). We compare our model with 12 alternative ZSL models in Table 4.2. We can see that on AwA1, AwA2 and aPY, the proposed model DEM is particularly strong under the more realistic GZSL setting measured using the harmonic mean (H) metric. In particular, DEM achieves state-

Model	**F**	**SS**	**AwA**	**CUB**
AMP (Fu et al., 2015b)	F_O	A+W	66.0	-
SJE (Akata et al., 2015)	F_G	A	66.7	50.1
SJE (Akata et al., 2015)	F_G	A+W	73.9	51.7
ESZSL (Romera-Paredes and Torr, 2015)	F_G	A	76.3	47.2
SSE-ReLU (Zhang and Saligrama, 2015)	F_V	A	76.3	30.4
JLSE (Zhang and Saligrama, 2016a)	F_V	A	80.5	42.1
SS-Voc (Fu and Sigal, 2016)	F_O	A/W	78.3/68.9	-
SynC-struct (Changpinyo et al., 2016)	F_G	A	72.9	54.5
SEC-ML (Bucher et al., 2016)	F_V	A	77.3	43.3
DeViSE (Frome et al., 2013)	N_G	A/W	56.7/50.4	33.5
Socher *et al.* (Socher et al., 2013)	N_G	A/W	60.8/50.3	39.6
MTMDL (Yang and Hospedales, 2015)	N_G	A/W	63.7/55.3	32.3
Ba *et al.* (Lei Ba et al., 2015)	N_G	A/W	69.3/58.7	34.0
DS-SJE (Reed et al., 2016a)	N_G	A/D	-	50.4/**56.8**
DEM	N_G	A/W(D)	**86.7/78.8**	**58.3**/53.5
DEM	N_G	A+W(D)	**88.1**	**59.0**

Table 4.1: Zero-shot classification accuracy (%) comparison on AwA and CUB (hit@1 accuracy over all samples) under the old and conventional setting. SS: semantic space; A: attribute space; W: semantic word vector space; D: sentence description (only available for CUB). F: how the visual feature space is computed; For non-deep models: F_O if overfeat (Sermanet et al., 2013) is used; F_G for GoogLeNet (Szegedy et al., 2015); and F_V for VGG net (Simonyan and Zisserman, 2014). For neural network based methods, all use Inception-V2 (GoogLeNet with batch normalisation) (Szegedy et al., 2015; Ioffe and Szegedy, 2015) as the CNN subnet, indicated as N_G.

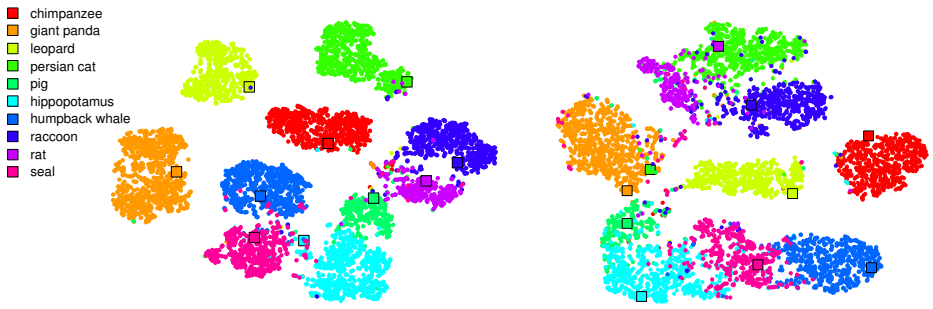
of-the-art performance on AwA1, AwA2 and SUN under conventional setting with 68.4%, 67.1% and 61.9%, outperforming alternatives by big margins.

4.4.3 Experiments on ImageNet

Comparative results on ILSVRC 2010 Compared to AwA and CUB, far fewer works report results on the large-scale ImageNet ZSL tasks. We compare our model against 8 alternatives on ILSVRC 2010 in Table 4.3, where we use hit@5 rather than hit@1 accuracy as in the small dataset experiments. Note that existing works follow two settings. Some of them (Mukherjee and Hospedales, 2016; Huang et al., 2016) use existing CNN model (e.g. VGG/GoogLeNet) pretrained from ILSVRC 2012 1K classes to initialise their model or extract deep visual features. Comparing to these two methods under the same setting, our model gives 60.7%, which beats the nearest rival PDDM (Huang et al., 2016) by over 12%. For comparing with the other 6 methods, we follow their settings and pretrain our CNN subnet from scratch with Alexnet (Krizhevsky et al., 2012) architecture using the 800 training classes for fair comparison. The results show that again, significant improvement has been obtained with our model.

Comparative results on ILSVRC 2012/2010 Even fewer published results on this dataset are available. Table 4.4 shows that our model clearly outperform the state-of-the-art alternatives by a large margin.

4.4.4 Further analysis



(a) S → V (b) V → S

Figure 4.4: Visualisation of the distribution of the 10 unseen class images in the two embedding spaces on AwA using t-SNE (Maaten and Hinton, 2008). Different classes as well as their corresponding class prototypes (in squares) are shown in different colours. Better viewed in colour.

Importance of embedding space selection We argue that the key for an effective deep embedding model is the use of the CNN output visual feature space rather than the semantic space

	AwA1				AwA2				CUB				aPY				SUN			
	ZSL	GZSL			ZSL	GZSL			ZSL	GZSL			ZSL	GZSL			ZSL	GZSL		
Model	T1	u	s	H	T1	u	s	H	T1	u	s	H	T1	u	s	H	T1	u	s	H
DAP (Lampert et al., 2014)	44.1	0.0	**88.7**	0.0	46.1	0.0	84.7	0.0	40.0	1.7	67.9	3.3	33.8	4.8	78.3	9.0	39.9	4.2	25.1	7.2
IAP (Lampert et al., 2014)	35.9	2.1	78.2	4.1	35.9	0.9	87.6	1.8	24.0	0.2	**72.8**	0.4	36.6	5.7	65.6	10.4	19.4	1.0	37.8	1.8
ConSE (Norouzi et al., 2014)	45.6	0.4	88.6	0.8	44.5	0.5	**90.6**	1.0	34.3	1.6	72.2	3.1	26.9	0.0	**91.2**	0.0	38.8	6.8	39.9	11.6
CMT (Socher et al., 2013)	39.5	8.4	86.9	15.3	37.9	8.7	89.0	15.9	34.6	4.7	60.1	8.7	28.0	10.9	74.2	19.0	39.9	8.7	28.0	13.3
SSE (Zhang and Saligrama, 2015)	60.1	7.0	80.5	12.9	61.0	8.1	82.5	14.8	43.9	8.5	46.9	14.4	34.0	0.2	78.9	0.4	51.5	2.1	36.4	4.0
DeViSE (Frome et al., 2013)	54.2	13.4	68.7	22.4	59.7	17.1	74.7	27.8	52.0	**23.8**	53.0	32.8	**39.8**	4.9	76.9	9.2	56.5	16.9	27.4	20.9
SJE (Akata et al., 2015)	65.6	11.3	74.6	19.6	61.9	8.0	73.9	14.4	53.9	23.5	59.2	33.6	32.9	3.7	55.7	6.9	53.7	14.7	30.5	19.8
LATEM (Xian et al., 2016)	55.1	7.3	71.7	13.3	55.8	11.5	77.3	20.0	49.3	15.2	57.3	24.0	35.2	0.1	73.0	0.2	55.3	14.7	28.8	19.5
ESZSL (Romera-Paredes and Torr, 2015)	58.2	6.6	75.6	12.1	58.6	5.9	77.8	11.0	53.9	12.6	63.8	21.0	38.3	2.4	70.1	4.6	54.5	11.0	27.9	15.8
ALE (Akata et al., 2016)	59.9	16.8	76.1	27.5	62.5	14.0	81.8	23.9	54.9	23.7	62.8	**34.4**	39.7	4.6	73.7	8.7	58.1	**21.8**	33.1	**26.3**
SYNC (Changpinyo et al., 2016)	54.0	8.9	87.3	16.2	46.6	10.0	90.5	18.0	**55.6**	11.5	70.9	19.8	23.9	7.4	66.3	13.3	56.3	7.9	**43.3**	13.4
SAE (Kodirov et al., 2017)	53.0	1.8	77.1	3.5	54.1	1.1	82.2	2.2	33.3	7.8	54.0	13.6	8.3	0.4	80.9	0.9	40.3	8.8	18.0	11.8
DEM	**68.4**	**32.8**	84.7	**47.3**	**67.1**	**30.5**	86.4	**45.1**	51.7	19.6	57.9	29.2	35.0	**11.1**	75.1	**19.4**	**61.9**	20.5	34.3	25.6

Table 4.2: Comparative results on four datasets. Under that ZSL setting, the performance is evaluated using per-class average Top-1 (**T1**) accuracy (%), and under GZSL, it is measured using **u** = **T1** on unseen classes, **s** = **T1** on seen classes, and **H** = harmonic mean.

Model	**hit@5**
ConSE (Norouzi et al., 2014)	28.5
DeViSE (Frome et al., 2013)	31.8
Mensink *et al.* (Mensink et al., 2012)	35.7
Rohrbach (Rohrbach et al., 2011)	34.8
PST (Rohrbach et al., 2013)	34.0
AMP (Fu et al., 2015b)	41.0
DEM	**46.7**
Gaussian Embedding (Mukherjee and Hospedales, 2016)	45.7
PDDM (Huang et al., 2016)	48.2
DEM	**60.7**

Table 4.3: Comparative results (%) on ILSVRC 2010 (hit@1 accuracy over all samples) under the old and conventional setting.

Model	**hit@1**	**hit@5**
ConSE (Norouzi et al., 2014)	7.8	15.5
DeViSE (Frome et al., 2013)	5.2	12.8
AMP (Fu et al., 2015b)	6.1	13.1
SS-Voc (Fu and Sigal, 2016)	9.5	16.8
DEM	**11.0**	**25.7**

Table 4.4: Comparative results (%) on ILSVRC 2012/2010 (hit@1 accuracy over all samples) under the old and conventional setting.

as the embedding space. In this experiment, we modify our model in Fig. 4.1 by moving the two FC layers from the semantic embedding branch to the CNN feature extraction branch so that the embedding space now becomes the semantic space (attributes are used). Table 4.5 shows that by mapping the visual features to the semantic embedding space, the performance on AwA drops by 26.1% on AwA, highlighting the importance of selecting the right embedding space. We also hypothesis that using the CNN visual feature space as the embedding layer would lead to less hubness problem. To verify that we measure the hubness using the skewness score (see Sec. 4.3.4). Table 4.6 shows clearly that the hubness problem is much more severe when the wrong embedding space is selected. We also plot the data distribution of the 10 unseen classes of AwA together with the prototypes. Figure 4.4 suggests that with the visual feature space as the embedding space, the 10 classes form compact clusters and are near to their corresponding prototypes, whilst in the semantic space, the data distributions of different classes are much less separated and a few prototypes are clearly hubs causing miss-classification.

Loss	**Visual → Semantic**	**Semantic → Visual**
Least square loss	60.6	**86.7**
Hinge loss	57.7	72.8

Table 4.5: Effects of selecting different embedding space and different loss functions on zero-shot classification accuracy (%) on AwA.

N_1 **skewness**	**AwA**	**CUB**
Visual → Semantic	0.4162	8.2697
Semantic → Visual	**-0.4834**	**2.2594**

Table 4.6: N_1 skewness score on AwA and CUB with different embedding space.

Neural network formulation Can we apply the idea of using visual feature space as embedding space to other models? To answer this, we consider a very simple model based on linear ridge regression which maps from the CNN feature space to the attribute semantic space or vice versa. In Table 4.7, we can see that even for such a simple model, very impressive results are obtained with the right choice of embedding space. The results also show that with our neural network based model, much better performance can be obtained due to the introduced nonlinear-

ity and its ability to learn end-to-end.

Model	**AwA**	**CUB**
Linear regression (V $\rightarrow$ S)	54.0	40.7
Linear regression (S $\rightarrow$ V)	74.8	45.7
DEM	**86.7**	**58.3**

Table 4.7: Zero-shot classification accuracy (%) comparison with linear regression on AwA and CUB.

Choices of the loss function As reviewed in Sec. 2, most existing ZSL models use either margin based losses or binary cross entropy loss to learn the embedding model. In this work, least square loss is used. Table 4.5 shows that when the semantic space is used as the embedding space, a slightly inferior result is obtained using a hinge ranking loss in place of least square loss in our model. However, least square loss is clearly better when the visual feature space is the embedding space.

4.5 Summary

This chapter has proposed a novel deep embedding model for zero-shot learning. The model differs from existing ZSL model in that it uses the CNN output feature space as the embedding space. This chapter hypothesises that this embedding space would lead to less hubness problem compared to the alternative selections of embedding space. Further more, the proposed model offers the flexibility of utilising multiple semantic spaces and is capable of end-to-end learning when the semantic space itself is computed using a neural network. Extensive experiments show that our model achieves state-of-the-art performance on a number of benchmark datasets and validate the hypothesis that selecting the correct embedding space is the key for achieving the excellent performance.

Chapter 5

Cross-View Generation for Image Captioning by Actor-Critic Sequence Training

---

Chapter 3 and Chapter 4 presented frameworks for matching data across views. This chapter takes a further step to automatically generate the textual language description of a visual image. This is an important capability for a robot or other visual-intelligence driven AI agent that may need to communicate with human users about what it sees. Such image captioning methods are typically trained by maximising the likelihood of ground-truth annotated caption given the image. While simple and easy to implement, these approaches do not directly maximise the language quality metrics we care about such as CIDEr. This chapter investigate training image captioning methods based on actor-critic reinforcement learning in order to directly optimise non-differentiable quality metrics of interest for effectively describing image content with human-level language. By formulating a per-token advantage and value computation strategy in this novel reinforcement learning based captioning model, it is shown that it is possible to achieve the state of the art performance on the widely used image captioning benchmark.

5.1 Background

As the classic task of automatic object category recognition is beginning to approach a solved problem (Szegedy et al., 2016), interest is growing in solving a more 'end-to-end' task of generating richer descriptions of images in terms of natural language, suitable for communication to human users (Vinyals et al., 2015, 2016; You et al., 2016; Liu et al., 2017a). This task is ex-

tremely topical recently, benefiting from public benchmarks such as MSCOCO (Lin et al., 2014). Despite extensive research in recent years, leading performance on the benchmarks has not increased dramatically. It is hypothesised that this is mainly due to research focus being on the image understanding aspects of captioning, rather than the language generation aspects. In this chapter, reinforcement learning based are investigated methods for training effective language generation in captioning.

Most existing captioning studies investigate variants of deep learning-based image encoders, that feed into deep sentence decoders. They have two main issues: (i) They are trained by maximising the likelihood of each ground-truth word given the previous ground-truth words and the image using back-propagation (Ranzato et al., 2016), termed 'Teacher-Forcing' (Bengio et al., 2015). This creates a mismatch between training and testing, since at test-time the model uses the previously generated words from the model distribution to predict the next word. This exposure bias (Ranzato et al., 2016), results in error accumulation during generation at test time, since the model has never been exposed to its own predictions. (ii) While sequence models are usually trained using the cross entropy loss, the actual NLP quality metrics of interest – with which we evaluate them at test time – are non-differentiable metrics such as CIDEr (Vedantam et al., 2015). Ideally sequence models for image captioning should be trained to avoid exposure bias and directly optimise metrics for the task at hand.

To address this two identified issues in image captioning, the main idea is to formulate a reinforcement learning based work to improve the quality of generated textual description. In this way, the gradient of the expected reward can be optimised by sampling from the model during training, thus avoiding the train-test mismatch; and the relevant test-time metrics such as CIDEr can be directly optimised, by treating them as reward in a reinforcement learning context.

Specifically, an actor-critic model is proposed for image captioning. It consists of a policy network (actor) and value network (critic). The actor is trained to predict the caption as a sequential decision problem given the image, where the sequence of actions correspond to tokens. The critic predicts the value of each state (image and sequence of actions so far), which we define as the expected task-specific reward (language metric score) that the network will receive if it outputs the current token and continues to sample outputs according to its probability distribution. The value predicted by the critic can be used to train the actor (captioning policy network). Under the assumption that the critic produces the exact values, the actor is trained based on an unbi-

ased estimate of the gradient of the caption score in terms of relevant language quality metrics. Compared to most reinforcement learning applications (Mnih et al., 2013), image captioning has a much higher dimensional action (e.g., 10,000+ token/word actions) space but shorter episodes. The proposed actor-critic approach exploits the shorter episodes and ameliorates the high dimensional action space.

5.2 Problem formulation

Image captioning model aims to generate caption sequence $Y = \{y_1, \ldots, y_T\}, y_t \in \mathcal{D}$ given an image I, where $\mathcal{D}$ is the dictionary. To simplify the formulas we always use T to denote the length of an output sequence, ignoring the fact that the generated caption sequences may have different lengths. Two sets of input-output pairs (I, Y) are assumed to be available for both training and testing. The trained sequence generative model is evaluated by computing the task-specific score $R(\hat{Y}, Y)$ (e.g., BLEU, CIDEr) on the test set, where $\hat{Y}$ is the predicted caption sequence.

In this chapter, off-the-shelf conventional encoder-decoder architecture (see Figure 5.1) for image captioning is adopted which consists of a Convolutional Neural Network (CNN) as the encoder and a Recurrent Neural Network (RNN) as the decoder. In order to transform the image captioning problem into a reinforcement learning task, we consider the image caption generation process as a finite Markov decision process (MDP) $\{S, A, P, R, \gamma\}$. In the MDP setting, the state S is composed of the image feature I_e encoded by the CNN from image I and the tokens/actions $\{a_0, a_1, \ldots, a_t\}$ that are generated so far. With the definition of the state, the state transition function P is $s_{t+1} = \{s_t,\ a_{t+1}\}$, where the action a_{t+1} becomes a part of the next state s_{t+1} and the reward r_{t+1} is received. $\gamma \in [0, 1]$ is the discount factor. Under the MDP interpretation of the image captioning problem, we can apply standard reinforcement learning algorithms to maximise the cumulated reward.

5.2.1 Model

Actor-critic (Barto et al., 1983) reinforcement learning method is adopted to train the proposed model which contains a policy network (actor) and a value network (critic). In particular, the model use Inception-V3 (Szegedy et al., 2016) as the CNN subnet and Long Short-Term Memory (Hochreiter and Schmidhuber, 1997) as the RNN subnet. Both the policy network and value network are based on LSTM for sequential action or value generation.

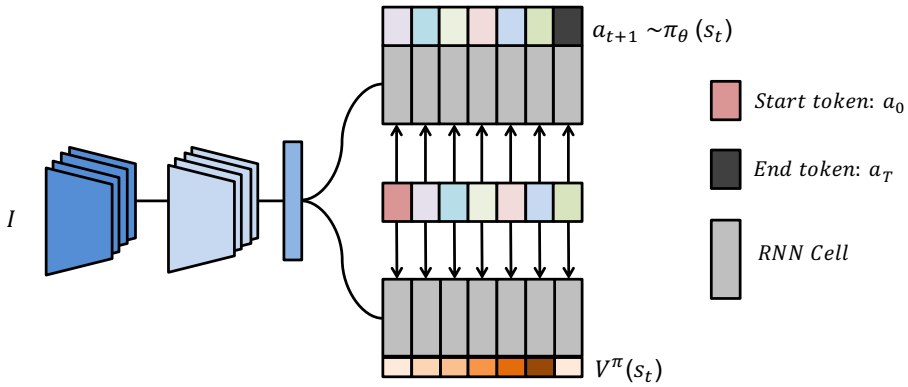


Figure 5.1: Schematic illustration of our actor-critic based captioning model (with word embedding layer omitted).

Policy network The policy network π is parametrised by θ and at time t it receives a state s_t and generates the categorical distribution over $|\mathcal{D}|$ actions (tokens), i.e. $a_{t+1} \sim \pi_\theta(s_t)$. We encode the given image I to I_e by CNN and treat I_e and the start token a_0 as the initial state s_0:

$$s_0 = \{\mathrm{I}_e,\ a_0\}. \tag{5.1}$$

With state transfer function mentioned above, we have:

$$s_t = \{\mathrm{I}_e, a_0, a_1, \ldots, a_t\}. \tag{5.2}$$

We feed state s_t into the LSTM and obtain the LSTM hidden state h_{t+1} (I_e was set as h_0). In order to build a probabilistic model for caption generation with an LSTM, we add a stochastic output layer f (typically with the softmax activation for discrete outputs) that generates outputs $a_{t+1} \in \mathcal{D}$:

$$\mathrm{h}_{t+1} = \mathrm{LSTM}(s_t),$$
$$a_{t+1} \sim f(\mathrm{h}_{t+1}).$$

Thus, the policy network defines a probability distribution $p(a_{t+1}|s_t)$ of the action a_{t+1} given current state s_t. The architecture of the policy network is the same as the standard supervised learning. Therefore, by given a target ground truth sequence $\{y_0, y_1, \ldots, y_T\}$, the supervised learning approach would be to train this network by minimising the cross entropy loss (XE).

$$\mathcal{L}_{\mathrm{XE}}(\theta) = -\sum_{t=1}^{T} \log(\pi_\theta(y_t \mid y_0, \ldots, y_{t-1})). \tag{5.3}$$

This corresponds to imitation learning of a perfect teacher in an RL context, and we use the pre-trained model as the initial policy network.

Policy gradient training Policy gradient methods maximise the expected cumulated reward by repeatedly estimating the gradient $g := \nabla_\theta \mathbb{E}[\sum_{t=1}^{T} r_t]$, where the environment issues the reward r_t according to the efficacy of the produced actions, rather than the teacher demonstrating the ideal actions directly as in Eq 5.3. For policy gradient, it is typically better to train an expression of the form:

$$g = \mathbb{E}[\sum_{t=0}^{T-1} A^\pi(s_t,\ a_{t+1}) \nabla_\theta \log \pi_\theta(a_{t+1} \mid s_t)], \tag{5.4}$$

where $A^\pi(s_t,\ a_{t+1})$ is advantage function yields almost the lowest possible variance, though in practice, the advantage function is not known and must be estimated. This statement can be intuitively justified by the following interpretation of the policy gradient: that a step in the policy gradient direction should increase the probability of better-than-average actions and decrease the probability of worse-than-average actions. The advantage function, by its definition $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$, measures whether or not the action is better or worse than the policy's default behaviour. So that the gradient term $A^\pi(s_t,\ a_{t+1}) \nabla_\theta \log \pi_\theta(a_{t+1} \mid s_t)$ points in the direction of increasing $\pi_\theta(a_{t+1} \mid s_t)$ if and only if $A^\pi(s_t,\ a_{t+1}) > 0$.

Value network Given the policy π, sampled actions and reward function, the value represents the expected future return as a function of the observed state s_t. We use V be an approximate state-value function.

$$V^\pi(s_t) = \mathbb{E}[\sum_{l=0}^{T-t-1} \gamma^l r_{t+l+1} \mid a_{t+1}, \ldots, a_T \sim \pi,\ I], \tag{5.5}$$

where parameter γ allows us to reduce variance by down-weighting rewards, at the cost of introducing bias. This parameter corresponds to the discount factor used in discounted formulations of MDPs.

The value network can be seen as an encoder. We propose to use a separate LSTM parametrised by ϕ with shared CNN. The RNN consumes state $s_t = \{\mathrm{I}_e, a_0, a_1, \ldots, a_t\}$ and produces a single value output to predict the TD target (to be defined later in Sec 5.2.2).

5.2.2 Advantage function estimation

Temporal-difference (TD) learning is utilised for advantage function estimation. Specifically, we define $Q^{\pi}(s_t, a_{t+1})$ in forward-view TD(λ) setting:

$$Q^{\pi}(s_t, a_{t+1}) = (1-\lambda)\sum_{n=1}^{\infty} G_t^n, \tag{5.6}$$

where G_t^n is the n-step expected return:

$$G_t^n = r_{t+1} + \gamma r_{t+2} + ... + \gamma^{n-1} r_{t+n} + \gamma^n V^{\pi}(s_{t+n}). \tag{5.7}$$

Therefore we have:

$$A^{\pi}(s_t, a_{t+1}) = Q^{\pi}(s_t, a_{t+1}) - V^{\pi}(s_t) = (1-\lambda)\sum_{n=1}^{\infty} G_t^n - V^{\pi}(s_t), \tag{5.8}$$

which is the same definition as Generalised Advantage Estimation (GAE) (Schulman et al., 2016) but in a forward view. Then the gradient of policy network has the form:

$$g = \mathbb{E}[\sum_{t=0}^{T-1} \big((1-\lambda)\sum_{n=1}^{\infty} G_t^n - V^{\pi}(s_t)\big)\nabla_{\theta} \log \pi_{\theta}(a_{t+1}|s_t)]. \tag{5.9}$$

5.2.3 Value function estimation

When using a nonlinear function approximator to represent the value function, the simplest approach is to solve a nonlinear regression problem:

$$\min_{\phi} ||Q^{\pi}(s_t, a_{t+1}) - V_{\phi}^{\pi}(s_t)||^2 \tag{5.10}$$

where $Q^{\pi}(s_t,\ a_{t+1}) = (1-\lambda)\sum_{n=1}^{\infty} G_t^n$.

5.2.4 λ setting for image captioning

λ setting plays an important role for the whole algorithm. If $\lambda = 0$, the advantage and value function estimations become one-step TD, whereas if $\lambda = 1$, the estimations turn out to be Monte Carlo approach. Since the episode length of image captioning is relatively shorter than popular contemporary RL problems (e.g. Atari and Mujoco games), and we have to sample the whole sequence of captions for rewarding, we set $\lambda = 1$ for our image captioning problem. Under this setting, the estimator for both advantage and value function is unbiased and the limited length of episode restricts the variance of estimation to a limited range. Concretely, with $\lambda = 1$, we have:

$$Q^{\pi}(s_t,\ a_{t+1}) = \sum_{l=0}^{T-t-1} \gamma^l r_{t+l+1} \tag{5.11}$$

5.2.5 Reward

For image captioning we can only obtain an evaluation score (e.g. CIDEr) when the caption generation process is finished. Therefore, we define the reward as follows:

$$r_t = \begin{cases} 0 & t < T \\ \text{score} & t = T \end{cases} \tag{5.12}$$

under such reward setting, we have

$$Q^{\pi}(s_t,\ a_{t+1}) = \gamma^{T-t-1} r_T. \tag{5.13}$$

Then the gradient of policy network has a sample form:

$$g = \mathbb{E}[\sum_{t=0}^{T-1} (\gamma^{T-t-1} r_T - V(s_t)) \nabla_\theta \log \pi_\theta(a_{t+1} \mid s_t)]. \tag{5.14}$$

5.3 Experiments

5.3.1 Implementation details

The Inception-v3 (Szegedy et al., 2016) is adopted as the CNN subnet, and an LSTM network is used as the RNN subnet. The number of LSTM cells is 512, equal to the dimension of the word embedding. The output vocabulary size for sentence generation is 12,000. Note that all these are exactly the same as the NICv2 (Vinyals et al., 2015) model ensuring a fair comparison.

For the CNN feature we used, semantic concept (Liu et al., 2017a) feature $\mathrm{I} \in \mathbb{R}^{1000}$ is used. These 1,000 semantic concepts are mined from the most frequent words in a set of image captions. A concept classifier is learned to predict Is as classification scores for the concepts.

Algorithm 2 describes the proposed method in detail. Our preliminary experiments show that training actor-critic from scratch can lead to an early determinisation of the policy and vanishing gradients, because neither the actor nor the critic would provide adequate training signals for one another. The actor would sample completely random tokens that receive very low reward, thus providing a very weak training signal for the critic. A random critic would be similarly useless for training the actor. To overcome these problems, staged pre-trainings are carried out. More specifically, we first pre-train the actor using standard cross entropy loss (XE) (see Eq. 5.3). After that, we pre-train the critic network by feeding it with sampled actions from the fixed pre-trained actor. The critic network is pre-trained for 2,000 iterations using Adam with a learning

rate of 5e-5. For the final stage of joint training of actor-critic, we weight critic loss by 0.5. We use Adam with an initial learning rate of 5e-5 and decrease it to 5e-6 after 1 million iterations, with minibatch size 16. The complete training procedure including pre-training is described by Algorithm 3.

Algorithm 2: Actor-Critic Training for Image Captioning

Require: Actor $\pi(a_{t+1}|s_t)$ and critic $V(s_t)$ with weights θ and ϕ respectively;
for *iteration = 1 to max iteration* **do**
  Receive a random example (I, Y) and sample sequence of actions $\{a_1, \ldots, a_T\}$ according to current policy π_θ;
  Compute TD target $Q^\pi(s_t,\ a_{t+1}) = \gamma^{T-t-1} r_T$ for $V(s_t)$;
  Update critic weights ϕ by minimising Eq. 5.10;
  Update actor weights θ using the gradient in Eq. 5.14;

Algorithm 3: Complete Actor-Critic Algorithm for Image Captioning

Initialise actor $\pi(a_{t+1}|s_t)$ and critic $V(s_t)$ with random weights θ and ϕ respectively;
Pre-train the actor to predict ground truth y_t given $\{y_1, \ldots, y_{t-1}\}$ by minimise Eq. 5.3;
Pre-train the critic to estimate $V(s_t)$ by running Algorithm 2 with fixed actor;
Run Algorithm 2

5.3.2 Datasets and setting

We evaluate the proposed method on the most widely used MSCOCO (Lin et al., 2014) dataset. The dataset contains 82,783 training images and 40,504 validation images. Each image is manually annotated with about 5 captions. The comparison against the state-of-the-art is conducted using the actual MS COCO test set comprising 40,775 images. Note that the annotation of the test set is not publicly available, so the results are obtained from the COCO evaluation server. We also follow the setting of (Vinyals et al., 2015, 2016) by using a held-out set of 4,051 images from the COCO validation set as the development set. The widely used BLEU, CIDEr, METEOR, and ROUGE scores are employed to measure the quality of generated captions.

5.3.3 Experimental results

Competitors Several state-of-the-art models are selected for comparison: MSRCap: The Microsoft Captivator (Devlin et al., 2015) combines the bottom-up based word generation model (Fang et al., 2015) with a gated recurrent neural network (Cho et al., 2014) (GRNN) for image captioning. mRNN: The multimodal recurrent neural network (Mao et al., 2015) uses a multimodal layer to combine the CNN and RNN. NICv2: The NICv2 (Vinyals et al., 2016) is an improved version of the Neural Image Caption generator (Vinyals et al., 2015). It uses a better image encoder, i.e., Inception-v3. In addition, scheduled sampling (Bengio et al., 2015) and an ensemble of 15 models are used; both improved the accuracy of captioning. V2L: The V2L model (Wu et al., 2016) uses a CNN based attribute detector to firstly generate 256 attributes, and then feed as initial input to an LSTM model to generate captions. ATT: The semantic attention model (You et al., 2016) uses both image features and visual attributes, and introduces an attention mechanism to reweight the attribute context to improve captioning accuracy. Semantic (Liu et al., 2017a) is our base model which uses a semantically regularised embedding layer as the interface between the CNN and RNN.

In addition to the traditional supervised learning method, we compare our method with three reinforcement learning based model. PG (Liu et al., 2017c) and MIXER (Ranzato et al., 2016) use policy gradient method with an additional FC layer on top of the RNN as state value network to reduce the high variance. Different from (Ranzato et al., 2016), (Liu et al., 2017c) uses the same method as (Yu et al., 2017) with k-times Monte Carlo rollout to estimate the state value target. Self-critical (Rennie et al., 2017) uses the basic REINFORCE algorithm with a reward obtained by the current model under the inference algorithm as the baseline. This simple method achieved very high performance which ranked 2nd currently on COCO captioning challenge. However, it use a multiple model ensemble. In contrast, only a single model is used for our method.

Results The results on development set are summarised in Table 5.1. We report a significant improvement from 1.007 to 1.162 on CIDEr over the log-likelihood baseline when single model greedy search is used for decoding. We can also see that our method is better than attention (Xu et al., 2015) and memory cell (Graves et al., 2016) which are added on top of the LSTM cell. For fair comparison with the current state-of-the-art method (Rennie et al., 2017), we implement it with same semantic CNN input (Liu et al., 2017a) on development set with single model for

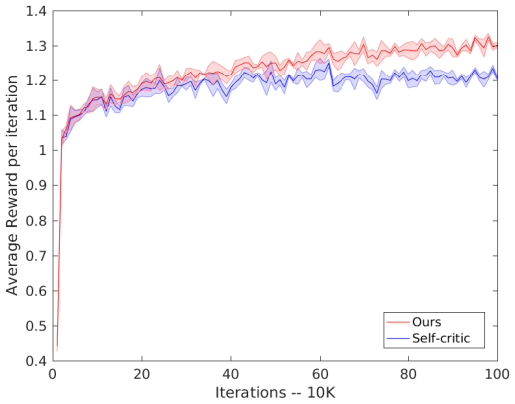


Figure 5.2: Average training reward curve for ours and (Rennie et al., 2017). We recorded the reward for 1 million iteration and plot every 10k iteration.

evaluation. The training reward curves of our method and (Rennie et al., 2017) are shown in Figure 5.2. Our method is clearly superior to that of (Rennie et al., 2017) due to the per-token advantage and value computation strategy adopted in our actor-critic based reinforcement learning framework.

We also submitted our single model results to the official evaluation server to compare with the eight baselines mentioned above. The evaluation is done with both 5 and 40 reference captions (C5 and C40). Our model is ranked the 3rd on the MSCOCO image captioning challenge leaderboard. Ours is the highest ranked single models, and is only surpassed by multi-model ensembles. Table 5.2 shows that, compared to the supervised learning based methods, our method significantly outperforms all of them in all metrics despite using only a single model rather than model ensemble. Comparing to the other two reinforcement learning based methods (Liu et al., 2017c; Ranzato et al., 2016), our method still achieved better performance except ROUGE-L c40.

Figure 5.3 shows some qualitative examples of our models captioning compared against using the same encoder-decoder architecture, but with standard cross-entropy (XE) training.

Computational Cost We compare the training time of our method with several alternatives. All algorithms are implemented in *Tensorflow* and run on an NVIDIA P100 card, with a mini-batch size of 16. Table 5.3 shows that for training, our method is the most efficient one. This is mainly due to the fact that our model does not have attention cell. Furthermore, for the Self-critical (Rennie et al., 2017) model, it needs to sample twice (random sampling + greedy decoding) for each iteration which is expensive.

Metric	CIDEr-D	BLEU-4	METEOR	ROUGE-L
NIC (Vinyals et al., 2015)	0.855	0.277	0.237	-
NICv2 (Vinyals et al., 2016)	0.998	0.321	0.257	-
Semantic (Liu et al., 2017a)	1.007	0.302	0.256	0.539
Semantic (Liu et al., 2017a)+**Attention** (Xu et al., 2015)	1.042	0.311	0.263	0.543
Semantic (Liu et al., 2017a)+**Attention** (Xu et al., 2015)+**Memory** (Graves et al., 2016)	1.057	0.318	0.266	0.547
Semantic (Liu et al., 2017a)+**Self-critical** (Rennie et al., 2017)	1.140	0.323	0.266	0.554
Ours	**1.162**	**0.344**	**0.267**	**0.558**

Table 5.1: Single model greedy search scores on the MSCOCO development set

Metric	B-1		B-2		B-3		B-4		METEOR		ROUGE-L		CIDEr-D	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
MSRCap (Devlin et al., 2015)	0.715	0.907	0.543	0.819	0.407	0.710	0.308	0.601	0.248	0.339	0.526	0.680	0.931	0.937
mRNN (Mao et al., 2015)	0.716	0.890	0.545	0.798	0.404	0.687	0.299	0.575	0.242	0.325	0.521	0.666	0.917	0.935
V2L (Wu et al., 2016)	0.725	0.892	0.556	0.803	0.414	0.694	0.306	0.582	0.246	0.329	0.528	0.672	0.911	0.924
NICv2 (Vinyals et al., 2016)	0.713	0.895	0.542	0.802	0.407	0.694	0.309	0.587	0.254	0.346	0.530	0.682	0.943	0.946
ATT (You et al., 2016)	0.731	0.900	0.565	0.815	0.424	0.709	0.316	0.599	0.250	0.335	0.535	0.682	0.943	0.958
Semantic (Liu et al., 2017a)	0.743	0.917	0.578	0.840	0.434	0.735	0.323	0.621	0.255	0.343	0.540	0.691	0.986	1.002
PG (Liu et al., 2017c)	0.754	0.918	0.591	0.841	0.445	0.738	0.332	0.624	0.257	0.340	0.550	**0.695**	1.013	1.032
MIXER (Ranzato et al., 2016)	0.747	-	0.579	-	0.431	-	0.317	-	0.258	-	0.545	-	0.991	-
Ours	**0.778**	**0.929**	**0.612**	**0.855**	**0.459**	**0.745**	**0.337**	**0.625**	**0.264**	**0.344**	**0.554**	0.691	**1.102**	**1.121**

Table 5.2: Results from the official MS-COCO image captioning challenge leaderboard (`https://www.codalab.org/competitions/3221#results`)

Model	Time
Semantic (Liu et al., 2017a)+Attention (Xu et al., 2015)	0.10
Semantic (Liu et al., 2017a)+Self-critical (Rennie et al., 2017)	0.13
Semantic (Liu et al., 2017a)+Self-critical (Rennie et al., 2017)+Attention (Xu et al., 2015)	0.18
Ours	**0.07**

Table 5.3: Training time for one minibatch on COCO dataset (in seconds)

5.4 Summary

This chapter have investigated the problem of automated image captioning by employing reinforcement learning to optimise the relevant non-differentiable language metrics such as CIDEr. A novel actor-critic based learning strategy is formulated which has the advantage over existing reinforcement learning based captioning models in that a per-token advantage and value computation is enabled leading to better model training. State-of-the-art performance is achieved using our computational efficient model on the MSCOCO benchmark.

Human : A man doing a trick on this skateboard..

XE : a man riding a skateboard on a cement ledge.

Ours : a man doing a trick on a skateboard.

Human : A motorcycle carrying many wheels is parked.

XE : a motorcycle parked next to a yellow wall.

Ours : a yellow motorcycle parked in front of a street.

Human : Large black motorcycle sitting next to a white building.

XE : a motorcycle parked next to a building.

Ours : a black motorcycle parked next to a building.

Human : A small personal pizza sits in a pizza box.

XE : a pizza is sitting on a box with a box of pizza.

Ours : a person holding a box of pizza.

Human : A group of skiers in the mountains reach a sign.

XE : a group of people standing on top of slope.

Ours : a group of people skiing on a snow covered slope.

Human : Old and new trains navigating a rail yard.

XE : a train is on the tracks in a city.

Ours : a black and white photo of trains on a train yard.

Figure 5.3: Qualitative results of image captioning on the MS COCO dataset.

Chapter 6

Conclusion and Future Work

This chapter summarise the achievements of the work presented in this thesis, and also discuss possible directions for future research.

6.1 Conclusion

This thesis has presented a collection of cross-view learning methods for automated analysis and understanding cross-view data. In particular, cross-view matching models for person re-identification and zero-shot learning have been investigated and explored. Besides, cross-view generation model is also studied for image captioning, aiming to automatically generate textual descriptions conditioned on visual images. These problems are inherently challenging due to the significant appearance and modality variations across views, or intrinsic learning strategy desiderata. Specifically,

1. In Chapter 3, a null space learning method for person re-identification problem is presented. This allows to obtain a more discriminative subspace for cross-view person image matching. Following are the observations in this chapter: (1) This chapter first identified the small sample size (SSS) problem suffered by all existing metric learning based re-id methods and argued that their solutions to this problem is suboptimal. (2) A null space learning method is then presented to overcome the SSS problem in cross-view person Re-ID. In this null space, images of the same person are collapsed into a single point thus minimising the within-class scatter to the extreme and maximising the relative between-

class separation simultaneously. Compared with existing Re-ID models, the employed NFST model is much simpler, with a closed-form solution and no parameters to tune. Yet, it is very effective in dealing with the SSS problem faced by the Re-ID methods. (3) A novel semi-supervised learning method is developed in the null space to exploit the abundant unlabelled data to further alleviate the effects of the SSS problem. Extensive experiments carried out on five person re-identification benchmarks show that such a simple and computationally very efficient approach beats all state-of-the-art methods often by a large margin.

2. In Chapter 4, a novel deep embedding model for zero-shot learning is proposed. Specifically, the proposed deep neural network based embedding model differs from existing models in that: To alleviate the hubness problem, visual space is adopted as the embedding space instead of the semantic space or an intermediate space. The resulting projection direction is from the textual view to visual view. Such a direction is opposite to the one adopted by most existing models. Moreover, a theoretical analysis and some intuitive visualisations are provided to explain why this would help to counter the hubness problem. Further, this framework design also provides a natural mechanism for multiple textual views (e.g., attributes and sentence descriptions) to be fused and optimised jointly in an end-to-end manner. Extensive experiments on four benchmarks show show that the proposed method beats all the state-of-the-art models presented to date, often by a clear margin.
3. In Chapter 5, first, two main issues in most existing captioning methods were identified. (i) They are trained by maximising the likelihood of each ground-truth word given the previous ground-truth words and the image using back-propagation (Ranzato et al., 2016), termed 'Teacher-Forcing' (Bengio et al., 2015). This creates a mismatch between training and testing, since at test-time the model uses the previously generated words from the model distribution to predict the next word. This exposure bias (Ranzato et al., 2016), results in error accumulation during generation at test time, since the model has never been exposed to its own predictions. (ii) While sequence models are usually trained using the cross entropy loss, the actual NLP quality metrics of interest – with which we evaluate them at test time – are non-differentiable metrics such as CIDEr (Vedantam et al., 2015). Ideally sequence models for image captioning should be trained to avoid exposure bias

and directly optimise metrics for the task at hand. Second, to address the identified two issues in image captioning, the main idea is to formulate a reinforcement learning based work to improve the quality of generated textual description. In this way, the gradient of the expected reward can be optimised by sampling from the model during training, thus avoiding the train-test mismatch; and the relevant test-time metrics such as CIDEr can be directly optimised, by treating them as reward in a reinforcement learning context. Specifically, a novel actor-critic based learning strategy is formulated which has the advantage over existing reinforcement learning based captioning models in that a per-token advantage and value computation is enabled leading to better model training. State-of-the-art performance is achieved using our computation efficient model on the MSCOCO benchmark.

Although the newly proposed methods have explored several applications and challenges in cross-view learning, other directions and dimensions are also possibly promising to investigate and explore, and a few of them are discussed below.

6.2 Future Work

The potential research directions for future work beyond the proposed methods are summarised as follows to end this thesis.

1. **Deep null space learning for person re-identification**: The initial effort of exploiting null space of the training samples for person image matching across views or person re-identification has shown effective and encouraging. However, the features are still hand-crafted. Recently, deep neural networks have been becoming a dominate solution for the appearance representation, they have also been shown to be effective for re-identification (Xiao et al., 2016). One possible extension could be integrating the current model into learning an end-to-end neural networks, as NFST is an extreme case of Linear Discriminative Analysis (Dorfer et al., 2015). Algorithmically, incorporating appropriate losses is important (Li et al., 2017; Geng et al., 2016).
2. **Deep embedding learning for sentence retrieval**: Aside from zero-shot learning, one could use the proposed cross-view matching framework of learning an embedding model for a range of other visual-textual cross-view problems. For example, sentence retrieval, to find amidst a set of sentences the one best describing the content of a given image or video, would be an area that could be improved upon. The embedding model transform the

text embedding from sentence vectorisation into a higher-dimensional visual feature space, which is capable of alleviating the *hubness* problem (Section 1.2), potentially allowing the training of robust systems for sentence retrieval.

3. **Reinforcement learning for multi-label image classification**: Reinforcement learning method has demonstrated favourable capabilities of generating text from image with CNN-RNN model. This indicates its promising potentials to deal with other visual-textual cross-view tasks with CNN-RNN architecture. Current CNN-RNN model for multi-label image classification (Wang et al., 2016b) optimise the likelihood of the generated text labels which is not the evaluation metric of interest. This may lead the learned model suboptimal. Therefore, the proposed actor-critic sequence training model for multi-label image classification is interesting to investigate and explore.

Bibliography

E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:1425–1438, 2016.

G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *International Conference on Machine learning*, pages 1247–1255, 2013.

K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath. A brief survey of deep reinforcement learning. *arXiv preprint arXiv:1708.05866*, 2017.

D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *International Conference on Learning Representations*, 2015.

D. Bahdanau, P. Brakel, K. Xu, A. Goyal, R. Lowe, J. Pineau, A. Courville, and Y. Bengio. An actor-critic algorithm for sequence prediction. In *International Conference on Learning Representations*, 2017.

A. G. Barto, R. S. Sutton, and C. W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, 1983.

S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, 2015.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

P. Bodesheim, A. Freytag, E. Rodner, M. Kemmler, and J. Denzler. Kernel null space methods for novelty detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

M. Bucher, S. Herbin, and F. Jurie. Improving semantic embedding consistency by metric learning for zero-shot classiffication. In *European Conference on Computer Vision*, 2016.

S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha. Synthesized classifiers for zero-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan. Multi-view clustering via canonical correlation analysis. In *International Conference on Machine learning*, pages 129–136. ACM, 2009.

D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang. Similarity learning on an explicit polynomial kernel feature map for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

L.-F. Chen, H.-Y. M. Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu. A new lda-based face recognition system which can solve the small sample size problem. *Pattern Recognition*, 2000.

X. Chen, L. Han, and J. Carbonell. Structured sparse canonical correlation analysis. In *Artificial Intelligence and Statistics*, pages 199–207, 2012.

Y. Chen, X. Zhu, and S. Gong. Person re-identification by deep learning multi-scale representations. In *IEEE International Conference on Computer Vision Workshop*, 2017.

D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *International Conference on Machine learning*, 2007.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

E. L. Denton, S. Chintala, R. Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in Neural Information Processing Systems*, pages 1486–1494, 2015.

J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, and M. Mitchell. Language models for image captioning: The quirks and what works. In *Annual Meeting of the Association for Computational Linguistics*, 2015.

R. Dian, L. Fang, and S. Li. Hyperspectral image super-resolution via non-local sparse tensor factorization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5344–5353, 2017.

G. Dinu, A. Lazaridou, and M. Baroni. Improving zero-shot learning by mitigating the hubness problem. In *ICLR workshop*, 2014.

J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2625–2634, 2015.

M. Dorfer, R. Kelz, and G. Widmer. Deep linear discriminant analysis. *arXiv preprint arXiv:1511.04707*, 2015.

M. Elhoseiny, T. El-Gaaly, A. Bakry, and A. Elgammal. A comparative analysis and study of multiview cnn models for joint object categorization and pose estimation. In *International Conference on Machine learning*, pages 888–897, 2016.

A. M. Elkahky, Y. Song, and X. He. A multi-view deep learning approach for cross domain user modeling in recommendation systems. In *Proceedings of the 24th International Conference on World Wide Web*, pages 278–288. International World Wide Web Conferences Steering Committee, 2015.

H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. From captions to visual concepts and back. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

M. Farenzena, L. Bazzani, A. Perina, M. Cristani, and V. Murino. Person re-identification by symmetry-driven accumulation of local features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

V. Ferrari and A. Zisserman. Learning visual attributes. In *Advances in Neural Information Processing Systems*, 2007.

D. Foley and J. Sammon. An optimal set of discriminant vectors. *IEEE Transactions on Computers*, 1975.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, 2013.

Y. Fu and L. Sigal. Semi-supervised vocabulary-informed learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Y. Fu, T. M. Hospedales, T. Xiang, Z. Fu, and S. Gong. Transductive multi-view embedding for zero-shot recognition and annotation. In *European Conference on Computer Vision*, 2014.

Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Transductive multi-view zero-shot learning. *PAMI*, 2015a.

Z. Fu, T. Xiang, E. Kodirov, and S. Gong. Zero-shot object recognition by semantic manifold distance. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015b.

M. Geng, Y. Wang, T. Xiang, and Y. Tian. Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244*, 2016.

A. Globerson and S. T. Roweis. Metric learning by collapsing classes. In *Advances in Neural Information Processing Systems*, 2005.

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

Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision*, 106(2):210–233, 2014b.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

A. Graves, N. Jaitly, and A.-r. Mohamed. Hybrid speech recognition with deep bidirectional lstm. In *ASRU*, 2013a.

A. Graves, A. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *ICASSP*, 2013b.

A. Graves, G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. Grabska-Barwińska, S. G. Colmenarejo, E. Grefenstette, T. Ramalho, J. Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 2016.

D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European Conference on Computer Vision*, 2008.

D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, volume 3, pages 1–7, 2007.

Y.-F. Guo, L. Wu, H. Lu, Z. Feng, and X. Xue. Null foley–sammon transform. *Pattern Recognition*, 2006.

D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer, 2011.

M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *European Conference on Computer Vision*, 2012.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neurocomputing*, 1997.

C. Huang, C. C. Loy, and X. Tang. Local similarity-aware deep feature embedding. In *Advances in Neural Information Processing Systems*, 2016.

R. Huang, S. Zhang, T. Li, and R. He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. *arXiv preprint arXiv:1704.04086*, 2017.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine learning*, 2015.

P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016.

D. Ji, J. Kwon, M. McFarland, and S. Savarese. Deep view morphing. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016.

S. M. Kakade and D. P. Foster. Multi-view regression via canonical correlation analysis. In *International Conference on Computational Learning Theory*, pages 82–96. Springer, 2007.

A. Karpathy. *Connecting Images and Natural Language*. PhD thesis, Stanford University, 2016.

A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

T.-K. Kim, S.-F. Wong, and R. Cipolla. Tensor canonical correlation analysis for action classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

E. Kodirov, T. Xiang, and S. Gong. Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification. In *British Machine Vision Conference*, 2015.

E. Kodirov, T. Xiang, and S. Gong. Semantic autoencoder for zero-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

I. Kviatkovsky, A. Adam, and E. Rivlin. Color invariants for person reidentification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *PAMI*, 2014.

R. Layne, T. M. Hospedales, S. Gong, and Q. Mary. Person re-identification by attributes. In *British Machine Vision Conference*, volume 2, page 8, 2012.

R. Layne, T. M. Hospedales, and S. Gong. Attributes-based re-identification. In *Person Re-Identification*, pages 93–117. Springer, 2014a.

R. Layne, T. M. Hospedales, and S. Gong. Re-id: Hunting attributes in the wild. In *British Machine Vision Conference*, 2014b.

Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, 2012.

C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

J. Lei Ba, K. Swersky, S. Fidler, and R. Salakhutdinov. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *IEEE International Conference on Computer Vision*, 2015.

C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pages 702–716. Springer, 2016.

W. Li and X. Wang. Locally aligned feature transforms across views. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

W. Li, X. Zhu, and S. Gong. Person re-identification by deep joint learning of multi-loss classification. In *International Joint Conference of Artificial Intelligence*, 2017.

Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

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 *European Conference on Computer Vision*, 2014.

G. Lisanti, I. Masi, A. Bagdanov, and A. Del Bimbo. Person re-identification by iterative re-weighted sparse ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014a.

G. Lisanti, I. Masi, and A. Del Bimbo. Matching people across camera views using kernel canonical correlation analysis. In *Proceedings of the International Conference on Distributed Smart Cameras*. ACM, 2014b.

F. Liu, T. Xiang, T. M. Hospedales, W. Yang, and C. Sun. Semantic regularisation for recurrent image annotation. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017a.

H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 2017b.

S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy. Improved image captioning via policy gradient optimization of spider. *IEEE International Conference on Computer Vision*, 2017c.

X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu. Semi-supervised coupled dictionary learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. *IEEE International Conference on Computer Vision*, 2017d.

Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1096–1104, 2016.

Y. Luo, D. Tao, K. Ramamohanarao, C. Xu, and Y. Wen. Tensor canonical correlation analysis for multi-view dimension reduction. *IEEE transactions on Knowledge and Data Engineering*, 27(11):3111–3124, 2015.

B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *European Conference on Computer Vision Workshop*, 2012.

L. Ma, X. Yang, and D. Tao. Person re-identification over camera networks using multi-task distance metric learning. *IEEE Transactions on Image Processing*, 2014.

L. v. d. Maaten and G. Hinton. Visualizing data using t-SNE. *JMLR*, 2008.

J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). In *International Conference on Learning Representations*, 2015.

B. Marco, L. Angeliki, and D. Georgiana. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *ACL*, 2015.

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.

A. Mignon and F. Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *arXiv preprint arXiv:1301.3781*, 2013a.

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*, 2013b.

M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

T. M. Mitchell et al. Machine learning, 1997.

V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning. In *Advances in Neural Information Processing Systems*, 2013.

V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine learning*, 2016.

T. Mukherjee and T. Hospedales. Gaussian visual-linguistic embedding for zero-shot recognition. In *EMNLP*, 2016.

J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *International Conference on Machine learning*, pages 689–696, 2011.

M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. In *International Conference on Learning Representations*, 2014.

S. Paisitkriangkrai, C. Shen, and A. ven den Hengel. Learning to rank in person re-identification with metric ensembles. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

D. Parikh and K. Grauman. Relative attributes. In *IEEE International Conference on Computer Vision*, 2011.

R. Paulus, C. Xiong, and R. Socher. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*, 2017.

S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1306–1315, 2016.

B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by support vector ranking. In *British Machine Vision Conference*, 2010.

M. Radovanović, A. Nanopoulos, and M. Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *JMLR*, 2010.

M. Ranzato, S. Chopra, M. Auli, and W. Zaremba. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations*, 2016.

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

S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *International Conference on Machine learning*, 2016b.

S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

C. Robert. *Machine Learning: a Probabilistic Perspective*. MIT Press, 2014.

M. Rohrbach, M. Stark, and B. Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

M. Rohrbach, S. Ebert, and B. Schiele. Transfer learning in a transductive setting. In *Advances in Neural Information Processing Systems*, 2013.

B. Romera-Paredes and P. Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine learning*, 2015.

J. Rupnik and J. Shawe-Taylor. Multi-view canonical correlation analysis. In *Conference on Data Mining and Data Warehouses*, pages 1–4, 2010.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 2015.

J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016.

M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 1997.

P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li. Embedding deep metric for person re-identification: A study against large variations. In *European Conference on Computer Vision*, 2016.

Y. Shigeto, I. Suzuki, K. Hara, M. Shimbo, and Y. Matsumoto. Ridge regression, hubness, and zero-shot learning. In *ECML/PKDD*, 2015.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 2016.

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

R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems*, 2013.

C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao. Multi-task learning with low rank attribute embedding for person re-identification. In *IEEE International Conference on Computer Vision*, pages 3739–3747, 2015a.

C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian. Deep attributes driven multi-camera person re-identification. In *European Conference on Computer Vision*, pages 475–491. Springer, 2016.

H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *IEEE International Conference on Computer Vision*, pages 945–953, 2015b.

M. Sugiyama. Local fisher discriminant analysis for supervised dimensionality reduction. In *International Conference on Machine learning*, 2006.

L. Sun, S. Ji, and J. Ye. Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):194–200, 2011.

I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112, 2014.

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

D. Tao, L. Jin, Y. Wang, Y. Yuan, and X. Li. Person re-identification by regularized smoothing kiss metric learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2013.

B. Thompson. Canonical correlation analysis. *Encyclopedia of statistics in behavioral science*, 2005.

L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

E. Ustinova, Y. Ganin, and V. Lempitsky. Multiregion bilinear convolutional neural networks for person re-identification. *arXiv preprint arXiv:1512.05300*, 2015.

R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *European Conference on Computer Vision*, 2016.

R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

R. Vezzani, D. Baltieri, and R. Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys*, 2013.

J. Vía, I. Santamaría, and J. Pérez. A learning algorithm for adaptive canonical correlation analysis of several data sets. *Neural Networks*, 20(1):139–152, 2007.

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

O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *PAMI*, 2016.

C. Wah, S. Branson, P. Perona, and S. Belongie. Multiclass recognition and part localization with humans in the loop. In *IEEE International Conference on Computer Vision*, 2011.

F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016a.

H. Wang et al. *Minimising Human Annotation for Scalable Person Re-Identification*. PhD thesis, Queen Mary University of London, 2017a.

J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu. Cnn-rnn: A unified framework for multi-label image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016b.

J. Wang, X. Zhu, S. Gong, and W. Li. Attribute recognition by joint recurrent learning of context and correlation. In *IEEE International Conference on Computer Vision*, volume 2, 2017b.

W. Wang, R. Arora, K. Livescu, and J. Bilmes. On deep multi-view representation learning. In *International Conference on Machine learning*, pages 1083–1092, 2015.

K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems*, 2005.

R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 1992.

Q. Wu, C. Shen, L. Liu, A. Dick, and A. van den Hengel. What value do explicit high level concepts have in vision to language problems? In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Z. Y. Y. Y. Y. Wu and R. S. W. W. Cohen. Encode, review, and decode: Reviewer module for caption generation. *Advances in Neural Information Processing Systems*, 2016.

Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 69–77, 2016.

Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *arXiv preprint arXiv:1707.00600*, 2017.

T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *European Conference on Computer Vision*, 2014.

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 *International Conference on Machine learning*, 2015.

F. Yan, K. Mikolajczyk, and J. Kittler. Person re-identification with vision and language. *arXiv preprint arXiv:1710.01202*, 2017.

J. Yang, S. E. Reed, M.-H. Yang, and H. Lee. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In *Advances in Neural Information Processing Systems*, pages 1099–1107, 2015.

L. Yang and R. Jin. Distance metric learning: A comprehensive survey. *Michigan State Universiy*, 2006.

Y. Yang and T. M. Hospedales. A unified perspective on multi-domain and multi-task learning. In *International Conference on Learning Representations*, 2015.

Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li. Salient color names for person re-identification. In *European Conference on Computer Vision*, 2014.

L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In *IEEE International Conference on Computer Vision*, pages 4507–4515, 2015.

D. Yi, Z. Lei, S. Liao, and S. Z. Li. Deep metric learning for person re-identification. In *IEEE International Conference on Pattern Recognition*, 2014.

Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

L. Yu, W. Zhang, J. Wang, and Y. Yu. Seqgan: sequence generative adversarial nets with policy gradient. In *AAAI*, 2017.

L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

L. Zhang, F. Sung, F. Liu, T. Xiang, S. Gong, Y. Yang, and T. M. Hospedales. Actor-critic sequence training for image captioning. In *NIPS Workshop on Visually-Grounded Interaction and Language*, 2017a.

L. Zhang, T. Xiang, and S. Gong. Learning a deep embedding model for zero-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017b.

Z. Zhang and V. Saligrama. Zero-shot learning via semantic similarity embedding. In *IEEE International Conference on Computer Vision*, 2015.

Z. Zhang and V. Saligrama. Zero-shot learning via joint latent similarity embedding. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016a.

Z. Zhang and V. Saligrama. Zero-shot recognition via structured prediction. In *European Conference on Computer Vision*, 2016b.

R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *IEEE International Conference on Computer Vision*, 2013a.

R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013b.

R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision*, 2015.

L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

W. Zheng, L. Zhao, and C. Zou. Foley-sammon optimal discriminant vectors using kernel approach. *IEEE TNN*, 2005.

W.-S. Zheng, S. Gong, and T. Xiang. Re-identification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3), 2013.

J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017a.

S. Zhu, S. Fidler, R. Urtasun, D. Lin, and C. C. Loy. Be your own prada: Fashion synthesis with structural coherence. In *IEEE International Conference on Computer Vision*, 2017b.

X. Zhu. Semi-supervised learning literature survey. *Citeseer*, 2005.

Z. Zhu, P. Luo, X. Wang, and X. Tang. Multi-view perceptron: a deep model for learning face identity and view representations. In *Advances in Neural Information Processing Systems*, pages 217–225, 2014.