Multimodal Representation Learning for Visual Reasoning and Text-to-Image Translation

by

Rudra

A Thesis Presented in Partial Fulfillment
of the Requirement for the Degree
Master of Science

Approved November 2018 by the
Graduate Supervisory Committee:

Yezhou Yang, Chair
Chitta Baral
Maneesh Kumar Singh

ARIZONA STATE UNIVERSITY

December 2018

ABSTRACT

Multimodal Representation Learning is a multi-disciplinary research field which aims to integrate information from multiple communicative modalities in a meaningful manner to help solve some downstream task. These modalities can be visual, acoustic, linguistic, haptic etc. The interpretation of 'meaningful integration of information from different modalities' remains modality and task dependent. The downstream task can range from understanding one modality in the presence of information from other modalities, to that of translating input from one modality to another. In this thesis the utility of multimodal representation learning for understanding one modality vis-à-vis Image Understanding for Visual Reasoning given corresponding information in other modalities, as well as translating from one modality to the other, specifically, Text to Image Translation was investigated.

Visual Reasoning has been an active area of research in computer vision. It encompasses advanced image processing and artificial intelligence techniques to locate, characterize and recognize objects, regions and their attributes in the image in order to comprehend the image itself. One way of building a visual reasoning system is to ask the system to answer questions about the image that requires attribute identification, counting, comparison, multi-step attention, and reasoning. An intelligent system is thought to have a proper grasp of the image if it can answer said questions correctly and provide a valid reasoning for the given answers. In this work how a system can be built by learning a multimodal representation between the stated image and the questions was investigated. Also, how background knowledge, specifically scene-graph information, if available, can be incorporated into existing image understanding models was demonstrated.

Multimodal learning provides an intuitive way of learning a joint representation between different modalities. Such a joint representation can be used to translate

from one modality to the other. It also gives way to learning a shared representation between these varied modalities and allows to provide meaning to what this shared representation should capture. In this work, using the surrogate task of text to image translation, neural network based architectures to learn a shared representation between these two modalities was investigated. Also, the ability that such a shared representation is capable of capturing parts of different modalities that are equivalent in some sense is proposed. Specifically, given an image and a semantic description of certain objects present in the image, a shared representation between the text and the image modality capable of capturing parts of the image being mentioned in the text was demonstrated. Such a capability was showcased on a publicly available dataset.

ACKNOWLEDGMENTS

I have come to realize the meaning behind the African proverb "it takes a village to raise a child" through the experience gained during my two years of research at ASU, and I would like to thank everyone who has helped and supported me during this time.

I would first like to thank my wonderful advisor Dr. Yezhou Yang, for his support, leadership and positive spirit. I was very fortunate that Dr. Yang joined ASU at the same time that I started here, and even more fortunate that he agreed to serve as my advisor. He not only gave me the freedom to choose my own research direction and guided me through the ups and downs of research life, he also provided me with ample cushion during turbulent times. I would like to thank Dr. Chitta Baral for his guidance during my very first reseach project. His fruitful inputs and intriguing questions helped me to improve my approach for the project. I would also like to thank Dr. Maneesh Singh for his brilliant input and guidance during the past year. Every discussion that I have had with him has helped me improve the way I approach any project.

I would also like to thank the members of Active Perception Group and Dr. Baral's lab for being part of my journey. I would like to give special thanks to Dr. Somak Aditya, who helped me understand how research should be conducted. I would like to thank Arpit Sharma, Mohammad Farhadi Bajestani, Zhiyuan Fang, Tianyi Ni, Shibin Zheng, Kausic Gunashekar, Xin Ye, Shuai Li, Divyanshu Bandil, Aman Verma, Stephen McAleer, Mo Izady for being incredible labmates and making this experience fun. I would like to give a special thanks to Trevor Richardson for all of the insightful talks we had and all of the things that he has exposed me to.

Thanks to my collaborators and collegues at Verisk Analytics, including Subbu,

Mahyar Khayatkhoei, Zheng Zhong, Han Kai Hsu, Fariba Zohrizadeh and Ya-Fang Shih. I spent a wonderful summer internship in New Jersey and look forward to beginning my industrial research career at Verisk.

Thanks to Chaynika Saikia, Amul Chugh, Nakul Chawla, Jitesh Kamble, Prakhar Khandelwal, Shiksha Patel, Gaizka Urreiztieta, Tarun Shimoga, Shuchir Inamdar, Vikas Rai, Rohan Rath. These individuals have helped, supported and inspired me in countless ways during various points of the last two years.

I would like to thank my aunts for their unconditional love and support. Finally, I would like to thank my mother. For everything.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Chapter 1

INTRODUCTION

1.1 Overview

Everyday, humans are exposed to sources of information in the real world that constitute multiple modalities at the same time. Here "modality" refers to certain type of information and/or the representation format in which the information is stored. These include, but are not limited to, textual, aural, visual, spatial or linguistic resources. For e.g., a multimedia web content on the internet is often composed of some text description accompanied by images and audio-visual content. Usually these are composed together to increase or test our reception of an idea or a concept. Humans find it very easy to assimilate these sources of information and perform very complex tasks ranging from visual reasoning and scene understanding, to that of tasks that require translation between two modalities. This capability to perform translation also lends them the ability to imagine examples in one modality given corresponding information in other modalities.

For instance, just a mere glance at the image in Figure 1.1, humans are able to extract tremendous amount of information pertaining to the visual scene. We can look at the image and immediately point out that there are "two birds sitting on a wooden branch". We can describe the various attributes of the birds, as well as detail about the activities that they are performing. Now given the question in Figure 1.1, which is an input from another modality, we promptly perform tasks starting from attribute identification based upon object mentions in the text i.e. identifying the orange, spatial attention and logical operations to identify the bird at the right of the

orange, and finally attribute identification to find the color of the bird's belly. Thus, by performing such integration of information from these two modalities along with multi-hop reasoning, we come to the conclusion that the answer should be "white".

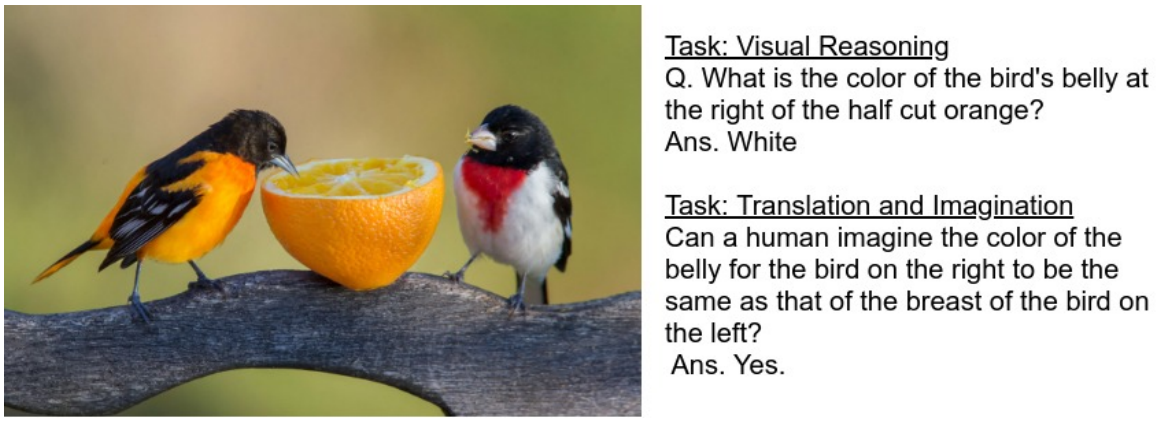

Figure 1.1: Assimilation of Information from Multiple Modalities Can Help Us Perform Multiple Tasks Such as Visual Reasoning or Multimodal Translation.

We also see no problem in combining our imaginative abilities to that of reasoning capability. For example, if someone asked us to imagine for the bird on the right to have yellow colored belly instead of white, we would be able to picture that with ease.

1.2 Challenges

Since it becomes second nature for humans to perform these tasks, we sometimes tend to forget how difficult it would be for a machine to do the same. Digesting data coming from diverse sources of information feels native to us. But for a computer, these different modalities have very different representations. For example, an image is nothing but a large array of real valued numbers specifying the intensity of various pixel values. It usually forms a very dense representation. Similarly, a text is a series of characters stored in memory as one or two bytes. But unlike images, text is usually represented in a discrete and sparse form. Thus combining such different representations into one model is not straightforward. Also, a machine has no notion

of the semantic concept of "birds" or "orange" or the color "white" or "yellow". The ability to perform scene understanding to be able to carry out visual reasoning or the capacity to envision new scenarios is not inherent to computers. One way to go around solving this is to train learning models with a large and diverse amount of data. Even though there has been a spike in the amount of data available for single modalities, there is still a dearth of data for multimodal systems. Moreover, the data currently present to train these models are noisy and often times has a lot of missing information making it difficult to make good one-to-one correspondance between modalities.

1.3 Motivation

Lending a machine the capability of comprehending information from heterogeneous sources, the ability to integrate these varied pieces of information and to be able to extract value from it has both academic motivations and practical applications. From a theoretical point of view, it's interesting to understand how this aptitude has emerged in humans over time. It gives us a playground to test out various hypothesis coming from psychology, cognitive and neurosciences that try to explain the emergence and subsequent development of this phenomenon in humans. This in turn can help us create systems that shows facsimile towards human abilities, something that has been a long standing goal of AI. It also allows us to see how advances in other fields such as mathematics, biology, physics etc. can help create computer models that assimilate diverse data while providing exploitable properties on how this data is represented internally.

From a practical standpoint, there are diverse applications where introduction of this ability can be helpful. For example, researchers on the pursuit of finding life on other planets have to constantly make sense of copious amounts of data coming from

sources such as infrared cameras, infrared spectographs, acoustic waves etc. where most of the samples are noise. A machine with this ability can help them quickly weed out undesirable examples and refocus their time on viable ones. This can also help in development of assistive technologies such as language translation models for both aural and linguistic modalities which can help break the language barrier. It can help in developing robust document understanding and web content analysis systems, better recommendation engines, automatic closed caption generations systems etc. This can also be used to better predict the occurrence of natural disasters as well as expedite medical diagnosis thus helping us to save lives.

1.4 Recent Progress

The last decade has witnessed a remarkable expansion of research in machine learning and neural networks. Deep neural nets have been around for more than 30 years, but standard training methods have serious limitations when used on architectures of more than 2 layers. With the advent of better training mechanism, higher compute power and abundance of both labeled and unlabeled data, the field has gained an unprecedented popularity. Several new areas such as meta-learning, explainability in deep learning, networks with memory, few-shot learning etc., have developed, and some previously established areas like generative models, reinforcement learning etc. have gained new momentum.

Deep learning, sometimes referred to as representation learning, has enabled us to learn higher level representations of data from a single modality using non-linear mappings. This has opened up ways of combining depictions of heterogeneous data in a more abstract sense. Furthermore, this has enabled us to train parts of our multimodal systems on single modality data and later combine their abstract representations in the form of "embeddings" in an end-to-end learning paradigm to fulfill

our goal of solving the downstream task. We have leveraged these properties in this work to solve visual reasoning and multi-modal translation tasks.

1.5 Contributions and Outline

In this thesis we mainly develop neural network models that consume and align two modalities viz. images and natural language to perform visual reasoning and text to image translation tasks.

In **Chapter 2**, we give a brief historical view of prior research on multimodal systems.

In **Chapter 3**, we present an end-to-end neural architecture that combines images and natural language to perform visual reasoning. We showcase how additional knowledge, if present, in the form of scene-graph information can be integrated with existing neural network architectures. We first convert this auxiliary information into pre-processed spatial masks using probabilistic reasoning mechanism. We then utilize the knowledge distillation paradigm to fuse this additional knowledge into existing models. We show how this multimodal fusion allows us to solve visual reasoning tasks and how the inclusion of external knowledge provides a performance boost on two publicly available datasets namely CLEVR and Sort-of-Clevr.

In **Chapter 4**, we provide an end-to-end neural network architecture that performs translation between modalities. Specifically, we showcase how a natural language sentence providing a semantic description of an image can be imagined to generate new images. For this, we train a multimodal network that learns a shared embedding space between the images and the natural language descriptions. We show the viability of our model to translate from text to corresponding images on the publicly available Caltech-UCSD Birds-200-2011 dataset.

Chapter 2

BACKGROUND

Multimodal learning has been an active area of research since the early 1970s. The field in general has been investigated by multiple communities spanning various modalities. The initial forray was made by psychologists trying to device new methods for psychotherapy and to answer how human decision making has evolved over time. They worked with multiple modalities including sound, taste, touch, appearance, aroma, attention, memories and preferences. The seminal work in this field was done by psychologist Arnold Lazarus, who originated the term behaviour therapy in psychotherapy and developed the practice of Multimodal therapy (Lazarus *et al.*, 1976). It is based on the idea that humans are biological beings that think, feel, act, sense, imagine, and interact—and that psychological treatment should address each of these modalities, both separately and together. Over time, the field has been adopted by multiple other communities. Computational approaches trying to learn and exploit representations directly from data have become a germane part of the research.

Prior research on multimodal learning can be divided into the following four eras:

- The ***behavioural*** era from 1970s until early 1980s.
- The ***computational*** era from late 1980s until 2000.
- The ***interaction*** era between 2000-2010.
- The ***deep learning*** era from 2010 until present

2.1 Behavioural Era

As stated, the behavioural era was pioneered by psychologists. Arnold Lazarus developed multimodal behaviour therapy (Lazarus *et al.*, 1976). The field later evolved into looking at integration of multi-sensory signals by humans for decision making (Mulligan and Shaw, 1980). The research also explored how humans are able to detect invariant relations between multiple modalities. Specifically, (Bahrick, 1983) looked into how infants can detect a relationship between the soundtracks and films of rigid and elastic objects in motion. Some researchers also delved into finding explanations behind various cognitive phenomenon. Most of these were related to language and gestures. One of the seminal works, now known as the McGurk effect (McGurk and MacDonald, 1976), looked into the perceptual phenomenon that demonstrates an interaction between hearing and vision in speech perception. This motivated the development of audio-visual speech recognition systems in the mid 1980s.

2.2 Computational Era

The computational era was spearheaded by the development of Audio-Visual Speech Recognition (AVSR) systems. The first AVSR system (Petajan, 1984) tried to combine lipreading from videos to enhance speech recognition capabilities of the model. It showed how the integration of acoustic and visual recognition candidates resulted in a final recognition accuracy which greatly exceeded any model trained only on acoustic recognition at that time. As computing devices started to proliferate in the mainstream market, various other works in this era were at the juncture of multimodal learning and human computer interaction. This led to the study of designing and evaluating new computer systems where human interact through multiple modalities, including both input and output modalities.

With the efficient adaptation of the backpropagation algorithm for neural networks (Werbos, 1981; Parker, 1985; LeCun, 1985) and the demonstration of emergence of useful internal representations in their hidden layers (Rumelhart and Zipser, 1986; Rumelhart *et al.*, 1986), neural networks were starting to gain momentum again. (Fels and Hinton, 1993) was one of the first works that tried to show the potential of multilayer neural networks for adaptive interfaces. More specifically, they tried to show neural networks viability for a multimodal translation task between hand-gestures and speech systems. As multimedia content started becoming the norm, it led to the need for creating a searchable library that could combine all of the componets of a multimedia document i.e. speech, image and natural language. The Informedia Digital Video Library Project (Wactlar *et al.*, 1996) was one of the first ones to inteliggently combine these modalities to create a full-content searchable digital video library.

The major algorithms used by these systems were based upon graphical models. Neural Ntworks, Hidden Markov Models and their variants became staples of these projects. Their successors are still the predominant techniques utilized for the development of such systems.

2.3 Interaction Era

The interaction era was mainly centered around the interaction between humans and machines that had access to multimodal sources of information. One of the earliest works in this was the Augmented Multi-Party Interaction project (McCowan *et al.*, 2005) that was concerned with the development of technology to support human interaction in meetings, and to provide better structure to the way meetings were run and documented. This led to the creation of 100+ hours of fully synchronized audio-visual recordings of the meetings that were transcribed and annotated.

With the improvement of speech recognition and understanding systems, there was a push to realize personalized cognitive assistants that learns from it's interaction with humans. The Cognitive Assistant that Learns and Organizes (CALO) project was among the first venture towards this direction. It attempted to integrate numerous AI technologies at that time to create a Personalized Assistant that Learns (PAL). The ability to extract information from a persons online social network (Culotta *et al.*, 2005) which included interaction with multimedia content was baked into such systems. They further pushed the frontiers in speech recognition and understanding using multimodal information (Tur *et al.*, 2008). Interestingly, Apple's SIRI was a spin-off from this project.

Multimedia information retrieval also gained momentum in this era. As machine learning became ubiquitous, researchers started developing models that were capable of performing high-level feature extraction from various media and to combine them to create content retrieval systems. Annual competitions such as Digital Video Retrieval hosted at NIST also promoted the research in this field. Dynamic Bayesian Networks like asynchronous hidden markov models (Bengio, 2003a,b), conditional random fields (Lafferty *et al.*, 2001) along with other machine learning techniques became the workhorses of such systems.

2.4 Deep Learning Era

As mechanisms to realize efficient training of neural networks beyond few layers were introduced, starting with greedy layer-wise training with Restricted Boltzman Machines (RBMs) followed by fine-tuning (Bengio *et al.*, 2007), work in deep learning started showing promising results related to extraction of useful representations in single modality tasks (Hinton and Salakhutdinov, 2006; Salakhutdinov and Hinton, 2009). With the advent of large-scale multimodal datasets as well as the rise in com-

pute power of modern systems and graphic processing units (GPUs), deep learning started to become a viable option for learning representations from data. One of the pioneering work related to multimodal deep learning was done by (Ngiam *et al.*, 2011). Here, the authors looked into aligning audio-visual data to learn cross-modality features and showcased that better features for one modality (e.g., video) can be learned if multiple modalities (e.g., audio and video) are present at feature learning time. They also showed ways to learn a shared representation between modalities and evaluated it on single-modality tasks. This led to further introduction of multiple new competitions and multimodal corpora to push the research frontier. These included the Audio-Visual Emotion Challenge, Emotion Recognition in the Wild Challenge, Image and Video Captioning competitions and Visual Question Answering tasks.

As generative models saw a resurgence by the introduction of Generative Adversarial Networks (Goodfellow *et al.*, 2014), Variational Autoencoders (Kingma and Welling, 2013) and Autoregressive models (Oord *et al.*, 2016), researchers started showing impressive results in generating single modality samples in both conditional and unconditional settings. Soon, these models saw their applications in multimodal domain in tasks such as image to image translation (Isola *et al.*, 2017; Zhu *et al.*, 2017a,b; Liu *et al.*, 2017), text to image translation (Reed *et al.*, 2016b; Zhang *et al.*, 2017a), visual dialogue systems (Massiceti *et al.*, 2018; Jain *et al.*, 2018) etc.

Chapter 3

VISUAL REASONING AND MULTIMODAL REPRESENTATION

3.1 Introduction

The task of visual reasoning tests an AI system's capability to combine knowledge from multimodal domains in order to solve problems that require complex multi-step reasoning. It is usually tackled by solving the surrogate task of Visual Question Answering (VQA), which aims to combine efforts from three broad sub-fields namely image understanding, language understanding, and reasoning and is often considered as "AI complete" (Antol *et al.*, 2015). In VQA, a system is provided with an input image and a question is posed against that image. Humans usually tackle this problem by combining our ability to perform semantic concept identification, attribute identification, counting, multi-step attention, comparison and logical operations between identified concepts. For a machine, it is the AI systems job to analyze the image and the question, and reason about how to answer this question correctly. At times, additional information about the scene depicted in the image is available. It behooves the system to be able to utilize this information during training and leverage the learned representation during testing. To explicitly assess the reasoning capability of visual reasoning systems, several specialized datasets have been proposed that emphasize specifically on questions requiring complex multiple-step reasoning (CLEVR (Johnson *et al.*, 2016), Sort-of-Clevr (Santoro *et al.*, 2017)) or questions that require reasoning using external knowledge (Wang *et al.*, 2017). However, current state-of-the-art methods do not leave room for integrating such external knowledge. Several researchers (Lake *et al.*, 2016; LeCun, 2017) in their works have pointed out the

necessity of explicit modeling of such knowledge. This necessitates considering the following issues:

- *What kind of knowledge is needed?*
- *Where and how to get them?*
- *What kind of reasoning mechanism to adopt for such knowledge?*

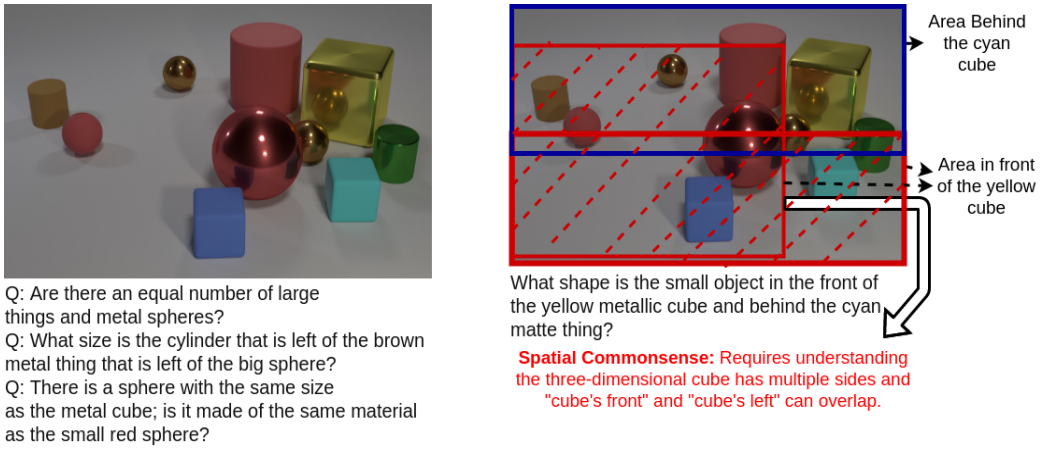

Figure 3.1: (a) An Image and a Set of Questions from the CLEVR Dataset. Questions Often Require Multiple-step Reasoning, for Example in the Second Question, One Needs to Identify the Big Sphere, Then Recognize the Reference to the Brown Metal Cube, Which Then Refers to the Root Object, That Is, the Brown Cylinder. (b) An Example of Spatial Knowledge Needed to Solve a CLEVR-type Question.

To understand the kind of external knowledge required, we investigate the CLEVR dataset proposed in (Johnson *et al.*, 2016). This dataset explicitly asks questions that require relational and multi-step reasoning. An example is provided in Fig. 3.1(a). In this dataset, the authors create synthetic images consisting of a set of objects that are placed randomly within the image. Each object is created randomly by varying its shape, color, size and texture. For each image, 10 complex questions are generated. Each question inquires about an object or a set of objects in the image. To understand which object(s) the question is referring to, one needs to decipher the clues that are provided about the property of the object or the spatial

relationships with other objects. This can be a multiple-step process, that is: first recognize object A, that refers to object B, which refers to C and so on. The failure cases of the current state-of-the-art works on this dataset often points to the lack of complex commonsense knowledge such as, *the front of cube should consist of front of all visible side of cubes.* These examples point that spatial commonsense knowledge might help answer questions such as in Fig. 3.1(b). Even though procuring such knowledge explicitly is difficult, we observe that parsing the questions and additional scene-graph information can help "disambiguate" the area of the image on which a phrase of a question focuses on.

In this chapter, we provide a framework, composed of neural network modules, built in an end-to-end manner that is not only capable of carrying out the task of visual reasoning but is also able to make use of any additional information if available. Specifically, we propose that an intuitive way of combining the information coming from these different modalities is to extract from them the spatial knowledge about the image. We showcase that in the presence of additional external information in the form of scene-graph annotations, it is possible to utilize probabilistic logical languages such as Probabilistic Soft Logic (Bach *et al.*, 2017) to encode spatial knowledge from scen-graph information. We also outline how the Knowledge Distillation (Hinton *et al.*, 2015; Vapnik and Izmailov, 2015; Hu *et al.*, 2016b) paradigm can be exploited to distill the representation learned from this additional source of information into existing visual reasoning architectures. For the case when this external information is not present, we employ in-network attention mechanism to emulate the spatial mask encoding process.

3.2 Related Works

Our work is influenced by the following thrusts of work: probabilistic logical reasoning, spatial reasoning, reasoning in neural networks, knowledge distillation; and the target application area of Visual Question Answering.

Researchers from the KR&R community, and the Probabilistic Reasoning community have come up with several robust probabilistic reasoning languages which are deemed more suitable to reason with real-world noisy data, and incomplete or noisy background knowledge. Some of the popular ones among these reasoning languages are Markov Logic Network (Richardson and Domingos, 2006), Probabilistic Soft Logic (Bach *et al.*, 2017), and ProbLog (De Raedt *et al.*, 2007). Even though these new theories are considerable large steps towards modeling uncertainty (beyond previous languages engines such as Answer Set Programming (Baral, 2003)); the benefit of using these reasoning engines has not been successfully shown on large real-world datasets. This is one of the reasons, recent advances in deep learning, especially the works of modeling knowledge distillation (Hinton *et al.*, 2015; Vapnik and Izmailov, 2015) and relational reasoning have received significant interest from the community.

Our work is also influenced by this series of works such as Region Connection Calculus etc., in the sense of what "privileged information" we expect along-with the image and the question. For the CLEVR dataset, the relations `left, right, front, behind` can be used as a closed set of spatial relations among the objects and that often suffices to answer most questions. For real images, a scene graph that encodes spatial relations among objects and regions, such as proposed in (Elliott and Keller, 2013) would be useful to integrate our methods.

Popular probabilistic reasoning mechanisms from the statistical community often define distribution with respect to Probabilistic Graphical Models. There have been

a few attempts to model such graphical models in conjunction with deep learning architectures (Zheng *et al.*, 2015). However, multi-step relational reasoning, and reasoning with external domain or commonsense knowledge [1] require the robust structured modeling of the world as adopted by KR&R languages. In its popular form, these reasoning languages often use predicates to describe the current world, such as $color(hair, red)$, $shape(object_1, sphere)$, $material(object_1, metal)$ etc; and then declare rules that the world should satisfy. Using these rules, truth values of unknown predicates are obtained, such as $ans(?x, O)$ etc. Similarly, the work in (Santoro *et al.*, 2017), defines the relational reasoning module as $RN(O) = f_\phi\Big(\sum_{i,j} g_\theta(o_i, o_j)\Big)$,where O denote all objects. In this work, the relation between a pair of objects (i.e. g_θ) and the final function over this collection of relationships i.e. f_ϕ are defined as multilayer perceptrons (MLP) and are learnt using gradient descent in an end-to-end manner. This model's simplicity and its close resemblance to traditional reasoning mechanisms motivates us to pursue further and integrate external knowledge.

Several methods have been proposed to distill knowledge from a larger model to a smaller model or from a model with access to privileged information to a model without such information. (Hinton *et al.*, 2015) first proposed a framework where a large cumbersome model is trained separately and a smaller student network learns from both groundtruth labels and the large network. Independently, (Vapnik and Izmailov, 2015) proposed an architecture where the larger (or the teacher) model has access to privileged information and the student model does not. These models together motivated many natural language processing researchers to formulate textual classification tasks as a teacher-student model, where the teacher has privileged in-

[1] An example of multi-step reasoning: if event A happens, then B will happen. The event B causes action C only if event D does not happen. For reasoning with knowledge: consider for a image with a giraffe, we need to answer "Is the species of the animal in the image and an elephant same?"

formation, such as a set of rules; and the student learns from the teacher and the ground-truth data. The imitation parameter controls how much the student *trusts* the teacher's decision. In (Hu *et al.*, 2016b), an iterative knowledge distillation is proposed where the teacher and the student learn iteratively and the convolutional network's parameters are shared between the models. In (Hu *et al.*, 2016a), the authors propose to solve sentiment classification, by encoding explicit logical rules and integrating the grounded rules with the teacher network. These applications of teacher-student network only exhibited success with classification problems with very small number of classes (less than three).

In this chapter, we show a knowledge distillation integration with privileged information which is applied to a 28-class classification, and we observe that it improves by a large margin on the baseline. In (Yu *et al.*, 2017), the authors use encoded linguistic knowledge in the form of $P(pred|obj, subj)$ to perform Visual Relationship Detection. In this work, we apply knowledge distillation in a visual question answering setting, that require both visual reasoning and question understanding.

In the absence of the scene information or in cases where such information is expensive to obtain, an attention mask over the image can be predicted inside the network based upon the posed question. Attention mechanism has been successfully applied in image captioning (Xu *et al.*, 2015; Mun *et al.*, 2017), machine translation(Bahdanau *et al.*, 2014; Vaswani *et al.*, 2017) and visual question answering (Yang *et al.*, 2016). In (Yang *et al.*, 2016), a stacked attention network was used to predict a mask over the image. They use the question vector separately to query specific image features to create the first level of attention. In contrast, we combine the question vector with the whole image features to predict a coarse attention mask.

3.3 Building Blocks

In this section, we explain the various components of our proposed framework of learning a multimodal representation for integrating additional spatial information with existing neural architectures. We start by formalizing the probabilistic reasoning mechanism which enables us to extract such spatial knowledge in the presence of scene information. Then, we describe the knowledge distillation paradigm (Hinton *et al.*, 2015) that enables us to infuse this extracted knowledge into existing networks which in our case is a relational reasoning architecture (Santoro *et al.*, 2017). We also outline the in-network computation required in the absence of the scene-graph information.

3.3.1 Probabilistic Reasoning Mechanism

In order to reason about the spatial relations among the objects in a scene and textual mentions of those objects in the question, we choose Probabilistic Soft Logic (PSL) (Bach *et al.*, 2017) as our reasoning engine. Using PSL provides us three advantages: i) (Robust Joint Modeling) from the statistical side, PSL models the joint distribution of the random variables using a Hinge-Loss Markov Random Field, ii) (interpretability) we can use clear readable declarative rules that (directly) relates to defining the clique potentials, and iii) (Convex Optimization) the optimization function of PSL is designed in a way so that the underlying function remains convex and that provides an added advantage of faster inference. We use PSL, as it has been successfully used in Vision applications (London *et al.*, 2013) in the past and it is also known to scale up better than its counterparts (Richardson and Domingos, 2006).

Hinge-Loss Markov Random Field and PSL

Hinge-Loss Markov Random Fields (HL-MRF) is a general class of continuous-valued probabilistic graphical model. An HL-MRF is defined as follows: Let $\boldsymbol{y}$ and $\boldsymbol{x}$ be two

vectors of n and n' random variables respectively, over the domain $D = [0,1]^{n+n'}$. The feasible set $\tilde{D}$ is a subset of D, which satisfies a set of inequality constraints over the random variables.

A *Hinge-Loss Markov Random Field* $\mathbb{P}$ is a probability density over D, defined as: if $(\boldsymbol{y}, \boldsymbol{x}) \notin \tilde{D}$, then $\mathbb{P}(\boldsymbol{y}|\boldsymbol{x}) = 0$; if $(\boldsymbol{y}, \boldsymbol{x}) \in \tilde{D}$, then:

$$\mathbb{P}(\boldsymbol{y}|\boldsymbol{x}) \propto exp(-f_{\boldsymbol{w}}(\boldsymbol{y}, \boldsymbol{x})). \tag{3.1}$$

PSL combines the declarative aspect of reasoning languages with conditional dependency modeling power of undirected graphical models. In PSL a set of weighted if-then rules over first-order predicates is used to specify a Hinge-Loss Markov Random field.

In general, let $\boldsymbol{C} = (C_1, ..., C_m)$ be such a collection of weighted rules where each C_j is a disjunction of literals, where each literal is a variable y_i or its negation $\neg y_i$, where $y_i \in \boldsymbol{y}$. Let I_j^+ (resp. I_j^-) be the set of indices of the variables that are not negated (resp. negated) in C_j. Each C_j can be represented as:

$$w_j : \vee_{i \in I_j^+} y_i \leftarrow \wedge_{i \in I_j^-} y_i, \tag{3.2}$$

or equivalently, $w_j : \vee_{i \in I_j^-} (\neg y_i) \bigvee \vee_{i \in I_j^+} y_i$. A rule C_j is associated with a non-negative weight w_j. PSL relaxes the boolean truth values of each ground atom a (constant term or predicate with all variables replaced by constants) to the interval [0, 1], denoted as $V(a)$. To compute soft truth values, Lukasiewicz's relaxation (Klir and Yuan, 1995) of conjunctions ($\wedge$), disjunctions ($\vee$) and negations ($\neg$) are used:

$$V(l_1 \wedge l_2) = max\{0, V(l_1) + V(l_2) - 1\}$$
$$V(l_1 \vee l_2) = min\{1, V(l_1) + V(l_2)\}$$
$$V(\neg l_1) = 1 - V(l_1).$$

In PSL, the ground atoms are considered as random variables, and the joint distribution is modeled using Hinge-Loss Markov Random Field (HL-MRF).

In PSL, the hinge-loss energy function $f_{\boldsymbol{w}}$ is defined as:

$$f_{\boldsymbol{w}}(\boldsymbol{y}) = \sum_{C_j \in \boldsymbol{C}} w_j \; max\Big\{1 - \sum_{i \in I_j^+} V(y_i) - \sum_{i \in I_j^-} (1 - V(y_i)), 0\Big\}. \tag{3.3}$$

The maximum-a posteriori (MAP) inference objective of PSL becomes:

$$\begin{aligned} \arg\max_{\boldsymbol{y} \in [0,1]^n} P(\boldsymbol{y}) &\equiv \arg\max_{\boldsymbol{y} \in [0,1]^n} exp(-f_{\boldsymbol{w}}(\boldsymbol{y})) \\ &\equiv \arg\min_{\boldsymbol{y} \in [0,1]^n} \sum_{C_j \in \boldsymbol{C}} w_j \; max\Big\{1 - \sum_{i \in I_j^+} V(y_i) \\ &\quad - \sum_{i \in I_j^-} (1 - V(y_i)), 0\Big\}, \end{aligned} \tag{3.4}$$

where the term $w_j \times max\big\{1 - \sum_{i \in I_j^+} V(y_i) - \sum_{i \in I_j^-} (1 - V(y_i)), 0\big\}$ measures the "distance to satisfaction" for each grounded rule C_j.

3.3.2 Knowledge Distillation Framework

While PSL provides a probabilistic knowledge representation , as shown in Figure 3.2(b), a mechanism is needed to utilize them under the deep neural networks based systems. We use the generalized knowledge distillation paradigm (Lopez-Paz *et al.*, 2015), where the teacher's network can be a larger network performing additional computation or have access to privileged information, to achieve this integration resulting in two different architectures i) **External Mask:** teacher with provided ground-truth mask, ii) **In-Network Mask:** teacher predicts the mask with additional computation. Here, we provide general formulations for both methods and give an overview of how the external mask is calculated [2] .

[2] A detailed example of how we estimate these predicates to calculate the external mask is provided in the appendix.

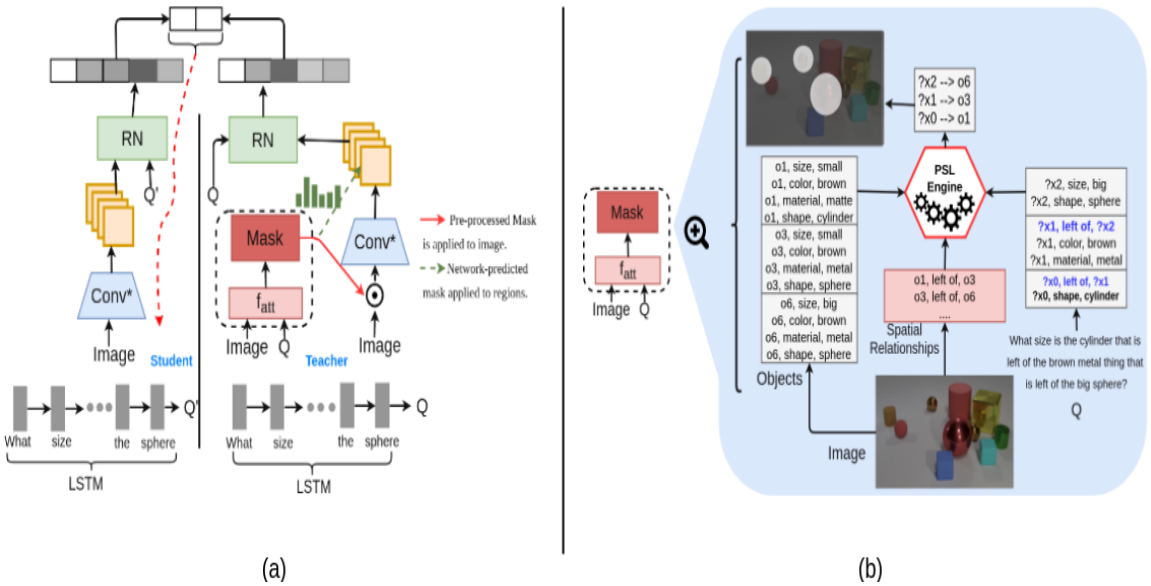

Figure 3.2: (a) The Teacher-Student Distillation Architecture: As the Base of Both Teacher and Student, We Use the Architecture Proposed by the Authors in (Santoro *et al.*, 2017). For the Experiment with Pre-processed Mask Generation, We Pass a Masked Image Through the Convolutional Network and for the Network-predicted Mask, We Use the Image and Question to Predict an Attention Mask over the Regions. (B) We Show the Internal Process of Mask Creation.

General Architecture

The general architecture for the teacher-student network is provided in Figure 3.2(a). Let us denote the teacher network as $\boldsymbol{q}_\phi$ and the student network as $\boldsymbol{p}_\theta$. In both scenarios, the student network uses the relational reasoning network (Santoro *et al.*, 2017) to predict the answer. The teacher network uses an LSTM to process the question, and a convolutional neural network to process the image. Features from the convolutional network and the final output from the LSTM is used as input to the relational reasoning module to predict an answer. Additionally in the teacher network, we predict a mask. For the External Mask setting, the mask is predicted by a reasoning engine and applied to the image, and for the attention setting, the mask is predicted using the image and text features and applied over the output from the convolution. The teacher network $\boldsymbol{q}_\phi$ is trained using softmax cross-entropy loss against the ground truth answers for each question. The student network is trained

using knowledge distillation with the following objective:

$$\theta = \arg\min_{\theta \in \Theta} \sum_{n=1}^{N} (1-\pi)\ell_1(\boldsymbol{y_n}, \sigma_\theta(\boldsymbol{x_n})) + \pi\ell_2(\boldsymbol{s_n}, \sigma_\theta(\boldsymbol{x_n})), \tag{3.5}$$

where $\boldsymbol{x_n}$ is the image-question pair, and $\boldsymbol{y_n}$ is the answer that is available during the training phase; the $\sigma_\theta(.)$ is the usual *softmax* function; $\boldsymbol{s_n}$ is the soft prediction vector of $\boldsymbol{q}_\phi$ on $\boldsymbol{x_n}$ and ℓ_i denotes the loss functions selected according to specific experiments (usually ℓ_1 is cross-entropy and ℓ_2 is euclidean norm). π is often called the imitation parameter and determines how much the student trusts the teacher's predictions.

External Mask Prediction

This experimental setting is motivated by the widely available scene graph information in large datasets starting from Sort-of-Clevr and CLEVR to Visual Genome. We use the following information about the objects and their relationships in the image: i) the list of *attribute, value* pairs for each object, ii) the spatial relationships between objects, and iii) each object's relative location in the image.

We view the problem as a special case of the bipartite matching problem, where there is one set of textual mentions (M) of the actual objects and a second set of actual objects (O). Using probabilistic reasoning we find a matching between object-mention pairs based on how the attribute-value pairs match between the objects and the corresponding mentions, and when mention-pairs are consistently related (such as *larger than, left to, next to*) as their matched object-pairs. Using the scene graph data, and by parsing the natural language question, we estimate the value of the following predicates: $attr_o(O, A, V)$, $attr_m(M, A, V)$ and $consistent(A, O, O_1, M, M_1)$. The predicate $attr_m(M, A, V)$ denotes the confidence that the value of the attribute

A of the textual mention M is V. The predicate $attr_o(O, A, V)$ is similar and denotes a similar confidence for the object O. The predicate $consistent(R, O, O_1, M, M_1)$ indicates the confidence that the textual mentions M and M_1 are consistent based on a relationship R (spatial or attribute based), if M is identified with the object O and M_1 is identified with the object O_1. Using only these two predicate values, we use the following two rules to estimate which objects relate to which textual mentions.

$$\begin{aligned} w_1 : candidate(M, O) \leftarrow & \; object(O) \wedge mention(M) \\ & \wedge attr_o(O, A, V) \wedge attr_m(M, A, V). \end{aligned} \tag{3.6}$$

$$\begin{aligned} w_2 : candidate(M, O) \leftarrow & \; object(O) \wedge mention(M) \\ & \wedge candidate(M, O) \\ & \wedge candidate(M_1, O_1) \\ & \wedge consistent(A, O, O_1, M, M_1). \end{aligned} \tag{3.7}$$

We use the grounded rules (variables replaced by constants) to define the clique potentials and use eq. 3.4 to find the confidence scores of grounded $candidate(M, O)$ predicates. Using this mention to object mapping, we use the objects that the question refers to. For each object, we use the center location, and create a heatmap that decays with distance from the center. We use a union of these heatmaps and use it as the mask. This results into a set of spherical masks over the objects mentioned in the question, as shown in Figure 3.2(b). To validate our calculated masks, we annotate the CLEVR validation set with the ground-truth objects, using the ground-truth structured program. We observe that our PSL-based method can achieve a 75% recall and 70% precision in predicting the ground-truth objects for a question.

In Figure 3.3, we provide more details of the calculated PSL predicates for the example question and image in Figure 3.2(b). We use this top collection of objects

and their relative locations to create small spherical masks over the relevant objects in the images.

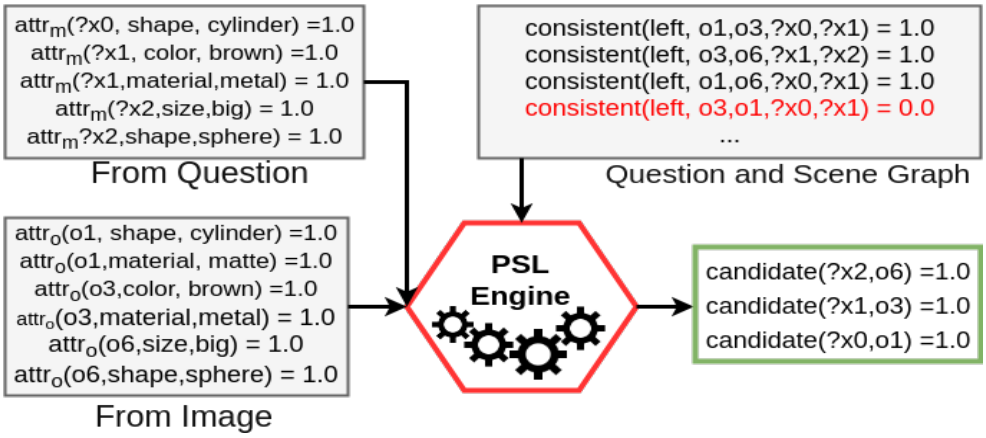

Figure 3.3: We Elaborate on the Calculated Psl Predicates for the Example Image and Question in Figure 3.2(b). The Underlying Optimization Benefits from the Negative Examples (the *consistent* Predicate with 0.0, Marked in Red). Hence, These Predicates Are Also Included in the Program.

In-Network Mask Prediction

The External Mask setting requires privileged information such as scene graph data about the image, which includes the spatial relations between objects. Such information is often expensive to obtain. Hence, in one of our experiments, we attempt to emulate the mask creation inside the network. We formulate the problem as attention mask generation over image regions using the image ($\boldsymbol{x_I} \in \mathbb{R}^{64\times64\times3}$) and the question ($\boldsymbol{x_q} \in \mathbb{R}^{w\times d}$). The calculation can be summarized by the following equations:

$$
\begin{aligned}
r_I &= conv^*(\boldsymbol{x_I}). \quad q_{emb} = LSTM(\boldsymbol{x_q}). \\
v &= tanh(W_I r_I + W_q q_{emb} + b). \\
\alpha &= \exp(v) / \sum_{r=1}^{x*y} \exp(v_r),
\end{aligned}
\tag{3.8}
$$

where r_I is $x \times y$ regions with o_c output channels, $q_{emb} \in \mathbb{R}^h$ is the final hidden state output from $LSTM$ (hidden state size is h); $W_I (\in \mathbb{R}^{xyo_c \times xy})$ and $W_q (\in \mathbb{R}^{xy\times h})$

are the weights and b is the corresponding bias. Finally, the attention α over regions is obtained by exponentiating the weights and then normalizing them. The attention α is then reshaped and element-wise multiplied with the region features extracted from the image. This is considered as a mask over the image regions conditioned on the question vector and the image features.

3.4 Experiments and Results

We propose two architectures, one where the teacher has privileged information and the other where the teacher performs additional calculation using auxiliary in-network modules. We perform experiments to validate whether the direct addition of information (external mask), or additional modules (model with attention) improves the teacher's performance over the baseline. We also perform similar experiments to validate whether this learned knowledge can be distilled to existing neural networks (student model) . Additionally, we conduct ablation studies on the probabilistic logical mechanism using which we predict a ground-truth mask from the question and the scene information.

3.4.1 Setup

As our testbed, we use the "Sort-of-Clevr" from (Santoro *et al.*, 2017) and the CLEVR dataset from (Johnson *et al.*, 2016). As the original Sort-of-Clevr dataset is not publicly available, we create the synthetic dataset as described by the authors. We use similar specification, i.e., there are 6 objects per image, where each object is either a circle or a rectangle, and we use 6 colors to identify each different object. Unlike the original dataset, we generate natural language questions along with their one-hot vector representation. In our experiments we primarily use the natural-language question. We only use the one-hot vector to replicate results of the baseline

Relational Network (RN) [3] . For our experiments, we use 9800 images for training, 200 images each for validation and testing. There are 10 question-answer pairs for each image. For Sort-of-Clevr, we use four convolutional layers with 32, 64, 128 and 256 kernels, ReLU non-linearities, and batch normalization. The questions were passed through an LSTM where the word embeddings are initialized with 50-dimensional Glove embeddings (Pennington *et al.*, 2014). The LSTM output and the convolutional features are passed through the RN network [4] . The baseline model was optimized with a cross-entropy loss function using the Adam optimizer with a learning rate of $1e^{-4}$ and mini-batches of size 64. For CLEVR, we use the Stacked Attention Network (Yang *et al.*, 2016) with the similar convolutional network and LSTM as above. We get similar results with VGG-16 as the convolutional network. Instead of the RN layer, we pass the two outputs through two levels of stacked attention, followed by a fully-connected layer. On top of this basic architecture, we define the student and teacher networks. The student network uses the same architecture as the baseline. We propose two variations of the teacher network, and we empirically show how these proposed changes improve upon the performance of the baseline network.

3.4.2 External Mask Prediction

In this setting of the experiment, the ground-truth mask, as calculated in 3.3.2, is element-wise multiplied to the image and then the image is passed through the convolutional network. We experiment with both sequential and iterative knowledge distillation. In the sequential setting, we first train the teacher network for 100 epochs

[3] We were unable to replicate the results of (Santoro *et al.*, 2017) on CLEVR dataset. Thus we use another baseline (Stacked Attention Network) and show how our method improves on that baseline. Based on our experiments, the best accuracy obtained by the baseline reasoning network is 68% with a batch-size of 640 on a single-GPU worker, after running for 600 epochs over the dataset.

[4] A four-layer MLP consisting of 2000 units per layer with ReLU non-linearities is used for g_θ; and a four-layer MLP consisting of 2000, 1000, 500, and 100 units with ReLU non-linearities used for f_ϕ.

with random embedding size of 32, batch size as 64, learning rate 0.0001. In the previous attempts to use distillation in natural language processing (Hu *et al.*, 2016a; Kim and Rush, 2016), the optimal value of π has been reported as $\min(0.9, 1 - 0.9^t)$ or 0.9^t. Intuitively, either at the early or at the latter stages, the student almost completely *trusts* the teacher. However, our experiments show different results. For the student network, we employ a hyperparamter search on the value of imitation parameter π and use two settings, where π is fixed throughout the training and in the second setting, π is varied using $\min(\pi, 1 - \pi^t)$. We vary the loss ℓ_2 among cross entropy and euclidean norm.

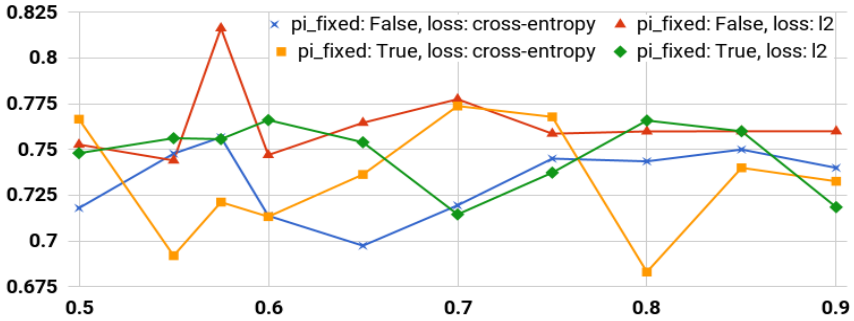

Figure 3.4: External Mask Prediction: Test Accuracy for Different Hyperparamter Combination to Obtain the Best Imitation Parameter (π) for Student for Sequential Knowledge Distillation.

The results of the hyperparameter optimization experiment is depicted in Figure 3.4. From this experiment, it can be observed that varying π over epochs gives better results than using a fixed π value for training the student. We observe a sharp increase in accuracy using the π value 0.575. This result is more consistent with the parameter value chosen by the authors in (Yu *et al.*, 2017). We also experiment by varying the word embedding (50-dimensional glove embedding and 32-dimensional word embedding) and learning rate. For sequential knowledge distillation, we get the best results with glove embedding and learning rate as $1e^{-4}$. However, we get huge improvements by using iterative knowledge distillation, where in each alternate epoch

the student learns from the teacher and the groundtruth data; and the teacher learns from its original loss function and the student's soft prediction (similar to Eqn. 3.5). Both weighted loss functions use the imitation parameter 0.9 (which remains fixed during training). We show the gradual learning of the teacher and the student till 800 epochs in Figure 3.5 and compare it with the RN baseline. We observe that: 1) the External Mask-augmented Teacher network converges faster than the baseline and 2) the Student network outperforms the baseline after 650 epochs of training.

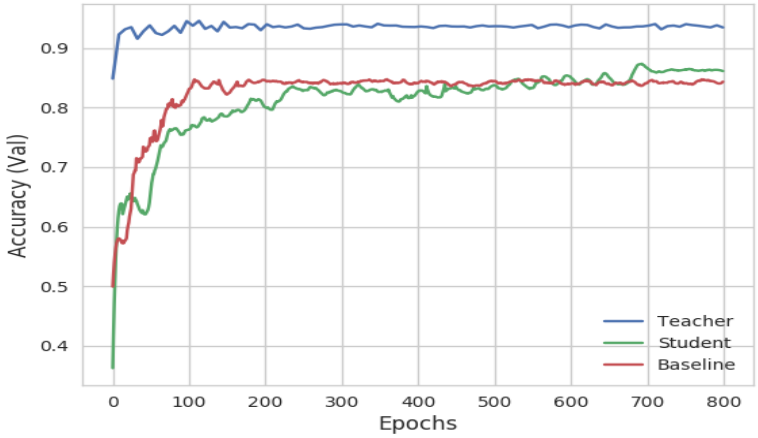

Figure 3.5: We Plot Validation Accuracy after Each Epoch for Teacher and Student Networks for Iterative Knowledge Distillation on Sort-of-clevr Dataset and Compare with the Baseline.

3.4.3 Larger Model with Attnetion

In this framework, we investigate whether the mask can be learnt inside the network with attention mechanism. We train the teacher network for 200 epochs with glove vectors of size 50, batch size as 64, learning rate as 0.0001. We have employed a hyperparamter search over learning rate, embedding type, and learning rate decay, and found that the above configuration produces best results. For the student network, we employed a similar hyperparamter search on the value of imitation parameter π and use two settings, where π is fixed throughout the training and in the

	Baseline	External Mask		In-Network Mask		Performance Boost Over Baseline (Δ)	
		Teacher	Student	Teacher	Student	Teacher	Student
Sort-of-Clevr	82% (Santoro *et al.* (2017))	**95.7**%	88.2%	87.5%	82.8%	13.7	6.2
CLEVR	53% (Yang *et al.* (2016))	**58**%	55%	-	-	5	2

Table 3.1: Test Set Accuracies of Different Architectures for the Sort-of-clevr (with Natural Language Questions) and CLEVR Dataset. For CLEVR, We Have Used the Stacked Attention Network (SAN) (Yang *et al.*, 2016) as Baseline and Only Conducted the External-mask Setting Experiment as It Already Calculates In-network Attention. Our Re-implementation of SAN Achieves 53% Accuracy on CLEVR. Accuracy Reported by (Santoro *et al.*, 2017) on SAN Is 61%. The Reported Best Accuracy for Sort-of-clevr and CLEVR Are 94% (One-hot Questions (Santoro *et al.*, 2017)) and 97.8% ((Perez *et al.*, 2017)).

second setting, π is varied using $min(\pi, 1 - \pi^t)$. We also vary the learning rate and the type of embedding (random with size 32 or glove vectors of size 50). The effect of the hyperparameter search is plotted in Fig. 3.6. We have experimented with iterative knowledge distillation and the best accuracy obtained for the teacher and the student networks are similar to that of sequential setting. The best test accuracies of the student network, the teacher with larger model and the baselines are provided in Table 3.1.

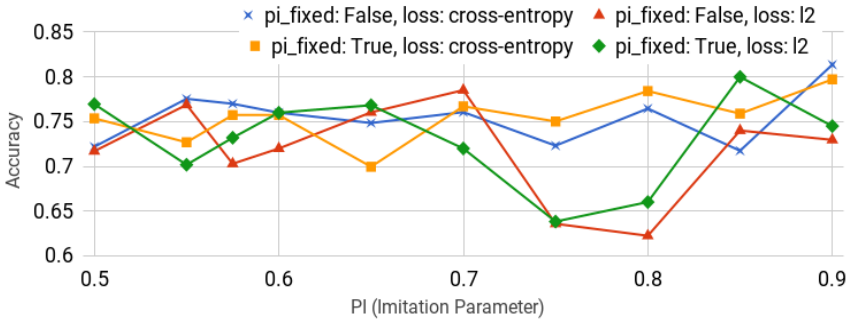

Figure 3.6: Model with Attention Mask: Test Accuracy for the Student Network for Different Hyperparamter Combination to Obtain the Best Imitation Parameter (π). We Get the Best Validation Accuracy Using the π as 0.9, ℓ_2 as Cross Entropy Loss and Varying π by over Epochs.

3.4.4 Analysis

The reported baseline accuracy on Sort-of-Clevr by (Santoro *et al.*, 2017) is 94% for both relational and non-relational questions. However, we use LSTMs to embed the natural language questions. Our implementation of the baseline achieves an overall test accuracy of 89% with one-hot question representation and 82% with LSTM embedding of the question. Addition of the pre-processed mask provides an increase in test accuracy to **95.7**%. In contrast, the teacher model with attention mask achieves **87.5**%. This is expected as the mask on the image simplifies the task by eliminating irrelevant region of image with respect to the question.

Student Learning: One may argue that adding such additional information to a model can be an unfair comparison. However, in this work, our main aim is to integrate additional knowledge (when it is available) with existing neural network architectures in a multimodal framework for the task of visual reasoning and demonstrate the benefits that such knowledge can provide. We experiment with the knowledge distillation paradigm to distill knowledge to a student. Extracted knowledge can be noisy, imperfect and often costly at test time. The distillation paradigm helps in this regard as the student network can choose to learn from the ground-truth data (putting less weight on teacher's predictions) during the training phase and doesn't require the additional knowledge during test time. For Sort-of-Clevr, we see an accuracy of **88.2**% achieved by the student network (in external mask setting), whereas for CLEVR the distillation effort increases the accuracy over the baseline method by 2%. Lastly, we show some qualitative examples of student network's output on the Sort-of-Clevr dataset (Fig. 3.7). The qualitative results indicate that our method can handle counting, spatial relationships well, but fails mostly on cases relating to shapes. This observation coupled with improvement in generalization validates that

the spatial knowledge has a significant role in our method.

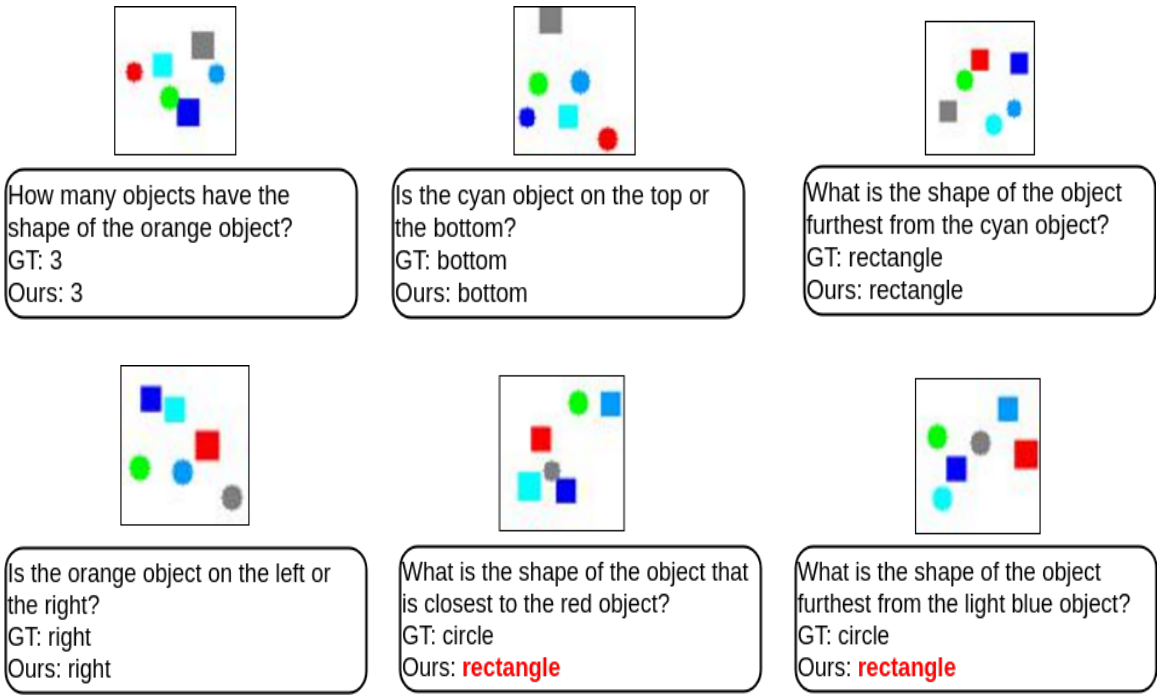

Figure 3.7: Some Example Images, Questions and Answers from the Synthetically Generated Sort-of-clevr Dataset. Red-colored Answers Indicate Failure Cases.

3.5 Conclusion

There has been a significant increase in attempts to integrate background information with state-of-the-art deep learning architectures for visual reasoning tasks that require assimilating knowledge from multiple modalities. In this chapter, we showcased our multimodal framework which attempted to integrate additional information in the form of spatial knowledge with existing neural networks to aid visual reasoning. The spatial knowledge is obtained by reasoning on the natural language question and additional scene information using the probabilistic soft logic inference mechanism. We show that such information can be encoded using a mask over the image and integrated with neural networks using knowledge distillation. Such a procedure shows significant improvements on the accuracy over the baseline network.

Chapter 4

TEXT TO IMAGE TRANSLATION

4.1 Introduction

In this chapter, we explore the general text-to-image translation problem. Text to image translation task aims at learning a conditional distribution of image given a text providing a description of said image. One way of learning this conditional distribution is by learning a joint distribution. The problem itself is ambiguous as a single input text may correspond to multiple possible output images and vice versa. Even when the text provides a semantic image description, it is unlikely to account for every pixel in the image. Thus, it is the system's task to imagine the features in the image domain corresponding to the text description. The task of learning a joint distribution from samples of marginal distribution is also imprecise as there exists an infinite number of joint distributions that can arrive the given marginal distribution (Lindvall, 2002). Thus additional assumptions needs to be placed in order to learn a joint distribution that can help us perform multimodal translation. In this work we make the assumption of a shared space representation and explore neural network based architectures for the same. We also investigate the possibility of controlling the information captured in this shared space. We propose a text-to-image translation framework based on cross-model embedding hallucination in a conditional generative modelling setting. The related ambiguity can be confined to a low-dimensional latent vector to be used for hallucination which can be sampled at test time. We also propose a single shared-latent space model which doesn't require embedding hallucination across modalities. Using the Caltech-UCSD Birds-200-2011

(Wah *et al.*, 2011b) dataset, we showcase our frameworks ability to capture the visual realization of the given text description as well as it's ability to generate diverse images.

4.1.1 Motivation

Generation of realistic images based upon text description is a challenging problem with numerous applications ranging from image editing to improving accessibility. The task requires to learn a mapping from text domain to RGB image domain. The map should be able to generate images that are realistic and capture the visual content represented in the text. Recent developments in generative modelling has spurred the synthesis of realistic images in the computer vision community. Generative Adversarial Networks (GANs) (Goodfellow *et al.*, 2014) and Variational Autoencoders (VAEs) (Kingma and Welling, 2013) as well has their conditional variants (Mirza and Osindero, 2014; Gauthier, 2014; Kingma *et al.*, 2014; Sohn *et al.*, 2015) have shown impressive performance in multiple tasks such as image super-resolution (Ledig *et al.*, 2016), image in-painting (Yeh *et al.*, 2017), attribute to image synthesis (Yan *et al.*, 2016; Vedantam *et al.*, 2017), image to image translation(Isola *et al.*, 2017; Zhu *et al.*, 2017a,b; Liu *et al.*, 2017) as well as in the task of text to image translation (Reed *et al.*, 2016b; Zhang *et al.*, 2017a; Dash *et al.*, 2017; Xu *et al.*, 2018; Zhang *et al.*, 2018).

In this chapter, we concentrate on the problem of generating realistic images given a semantic image description. An effective text to image generation model should posses the following properties: 1) **Fidelity** towards the entities and their interaction dynamics described in the text, 2) **Diversity** in the generated image for a given text by hallucinating concepts implicit or not defined in the text in order to produce a more coherent image and 3) **Controllability** in sampling to showcase the extent to

which the semantics will be captured.

Current text to image synthesis methods falter in one way or the other to capture the properties of a satisfactory text-to-image translation model. Reed et.al. (Reed *et al.*, 2016b) addressed the problem using a conditional GAN based framework. They were able to generate realistic looking images at a resolution of 64x64 but their generated images lacked details, vivid object parts as well as diversity. StackGAN (Zhang *et al.*, 2017a) utilized a two stage generation process to produce images of higher resolution (256x256) where they are able to showcase fidelity but lack diversity in the generated images. They also lack the ability to control the extent of the expression of style for the generated image. HDGAN (Zhang *et al.*, 2018) is able to generate high quality photo-realistic images using the progressive growing of GANs technique (Karras *et al.*, 2017) but they do not provide any controllability over attributes of generated samples.

In order to learn such a mapping from semantic image description to image domain while being faithful to the goal of generating diverse and perceptually realistic images with control over the sampling procedure, we combine the Variational Autoencoder paradigm with that of the Generative Adversarial Networks and explore multiple neural network architectures capable of performing the text-to-image translation task. We emphasize on the importance of learning a shared representation between these two modalities. We also propose to provide meaning to this shared latent space by capturing image regions that are mentioned in the text (also referred to as foreground in our model) using attention. We investigate stacked attention networks (Yang *et al.*, 2016) as well as compact bilinear pooling (Gao *et al.*, 2016) methods for learning said attention maps. We embed the image regions not mentioned in the semantic description (also referred to as background in our model) in a separate embedding space and learn the ability to hallucinate such an embedding space conditioned upon

the input text using co-embedding hallucination.

We show the effectiveness of our model by performing experiments using the publicly available Caltech-UCSB Birds-200-2011 (Wah *et al.*, 2011b) dataset. We perform experiments related to image generation from text descriptions for all of our architectures. For the attention based model, we generate the foreground and background separately and blend them to get the final result. We experiment with using a combination of class labels and explicit segmentation maps to bolster extraction of attention map in the shared representation learning process. We also perform quantitative evaluation of our model and report the inception score (Salimans *et al.*, 2016) metric and compare it with the current research in this field.

4.2 Related Works

Generative Modeling: Parametric modeling of the natural image distribution has been a fundamental problem in computer vision. (Hinton and Salakhutdinov, 2006) learned compressed codes for images using a stack of Restricted Boltzmann Machines(RBM). Hinton *et al.* (Hinton, 2009) used RBM and a layer-wise pretraining routine to produce probabilistic generative models. Variational Autoencoders (VAE) (Kingma and Welling, 2013) based on Helmholtz Machine (Dayan *et al.*, 1995) have been used to model the data distribution by defining an approximate density function. They model stochasticity within the network by reparameterization of latent distribution at training time. Autoregressive models (Efros and Leung, 1999; Oord *et al.*, 2016; van den Oord *et al.*, 2016) have also shown promise in capturing the natural image statistics but are slow at inference time due to their sequential nature. Recently, Generative Adversarial Networks (Goodfellow *et al.*, 2014) have shown promising results in generating high quality sample but the training instability (Salimans *et al.*, 2016) of traditional GANs often makes it hard to generate high-resolution coherent

samples. A lot of work have been proposed to stabilize GANs and improve the quality of the generated samples (Zhu *et al.*, 2017b; Salimans *et al.*, 2016; Arjovsky and Bottou, 2017; Radford *et al.*, 2015; Nguyen *et al.*, 2016; Gulrajani *et al.*, 2017) in both conditional and unconditional setting.

Conditional Image Generation: Using the above stated methods, conditional image generation has also been studied. The conditioning parameter can range from class labels or attributes (Chen *et al.*, 2016; Odena *et al.*, 2016; Yan *et al.*, 2016; Vedantam *et al.*, 2017) to natural language image descriptions (Reed *et al.*, 2016b,c; Zhang *et al.*, 2017a) or on images themselves (Taigman *et al.*, 2016; Isola *et al.*, 2017; Zhu *et al.*, 2017a; Liu *et al.*, 2017). Both VAE (Sohn *et al.*, 2015; Walker *et al.*, 2016) and autoregressive models (van den Oord *et al.*, 2016) have shown promising results. Conditional GAN based frameworks have resulted in substantial boost in the quality of the results (Reed *et al.*, 2016b; Zhang *et al.*, 2017a; Xu *et al.*, 2018; Zhang *et al.*, 2018).

Fusion of Variational and Adversarial Learning: VAEs and GANs have been combined before in (Larsen *et al.*, 2015; Berthelot *et al.*, 2017; Zhu *et al.*, 2017b). In BEGAN (Berthelot *et al.*, 2017) the discriminator of a GAN is replaced by an autoencoder, whereas, in (Larsen *et al.*, 2015) the decoder of a VAE is the same as the generator of a GAN. In (Zhu *et al.*, 2017b), the embedding produced by a VAE is being used as the noise code for a conditional image-to-image translation GAN. In (Rosca *et al.*, 2017), the authors explored various ways in which variational and adversarial objectives can be fused together. In their work, they learn both the posterior distribution and the likelihood distribution of the variational objective using adversarial training. Unlike their work, we use the adversarial training in the embedding space to learn a conditional co-embedding distribution. We also explore an experimental setup where we have adversarial object on the likelihood space. We

also investigate if the Joint-VAE framework described in (Vedantam *et al.*, 2017) can be extended to natural language sentences and formulated as a Joint-VAE-GAN framework. Inspired by the (Liu *et al.*, 2017) work on image to image translation, we also explore if such a model can be adapted to combine the visual and the text modality.

Text-to-Image Generation: Several methods have been proposed to generate images from unstructured text. Mansimov et al (Mansimov *et al.*, 2015) built an AlignDRAW model by learning to estimate alignment between text and the generating canvas. Nguyen et al (Nguyen *et al.*, 2016) used an approximate Langevin sampling approach to generate images conditioned on text. However, their sampling approach requires an inefficient iterative optimization process. Reed et.al (Reed *et al.*, 2016b) proposed a conditional GAN architecture called GAN-INT to generate plausible images conditioned upon semantic image description. Even though their method was able to produce images that looked plausible with respect to the input text condition, the synthesized images lacked details and diversity. In their follow up work called GAWWN (Reed *et al.*, 2016c), they utilized additional supervision in the form of bounding boxes and key-points to generate more realistic images at a higher resolution (128x128). In (Zhang *et al.*, 2017a), the authors proposed stacking two GANs together to provide more information to the second GAN in the sequence. This helped them in generating images of higher resolution (256x256) and provide more details to the synthesized images. In HDGAN (Zhang *et al.*, 2018), the authors leveraged a hierarchically nested architecture based upon (Karras *et al.*, 2017) to generate photo-realistic images of even higher resolution (512x512). Even though their generated images look very realistic, they lack diversity and control in generating a sample. They also suffer implicit mode collapse as their background and orientation of foreground objects do not show much variability. Motivated by the fact that variational

autoencoder based GANs tend to discourage such implicit mode collapse (Rosca *et al.*, 2017), we utilize this paradigm as part of our framework.

By using a VAE, we are grounding the generator of our GAN using a perceptual similarity metric, the reconstruction loss in our case, thus helping in mitigating some of issues related to mode-collapse. Also, since our co-embedding hallucination is being done on a lower dimensional embedding space, making it less prone to the complications related to unstability in GAN training. For the model with attention, by generating the foreground object separately from the background, we are allowing for a greater variability to be present in the foreground. At the same time, since background generation is conditioned upon foreground embedding space, we are allowing for the model to generate coherent images. We have trained our model in an end-to-end framework.

4.3 Building Blocks

Contemprorary generative models viz. Generative Adversarial Netwroks (Goodfellow *et al.*, 2014) and Variational Autoencoder (Kingma and Welling, 2013) form the fundamental units of our framework. We briefly discuss them below:

4.3.1 Variatinal Autoencoders

Variational Autoencoders are latent variable models that describe a stochastic process by which modeled data is assumed to be generated. Thus, it provides us a process by which synthetic data can be simulated from model distribution. Let x be the observed data points and z be the latent variables. Let $p(x, z)$ define the parametric model distribution, and $p(x|z)$ be the *generative mdoel* defined over the latent variables. Given a dataset $\mathcal{X} = \{x^1, x^2, ..., x^N\}$, we wish to perform the maximum

likelihood learning of the parameters of the generative model:

$$\log p(X) = \sum_{i=1}^{N} \log p(x^i) \tag{4.1}$$

In general, the marginal likelihood is intractable to compute for generative models that have high-dimensional latent variables and flexible priors and likelihoods. A solution is to introduce $q(z|x)$, an approximate parameteric inference model defined over the latent variables, and optimize the *variational lower bound* on the marginal log-likelihood of each observation x:

$$\log p(X) \geq \mathbb{E}_{q(z|x)}[\log p(x, z) - \log q(z|x)] = \mathcal{L}(x; \theta) \tag{4.2}$$

where θ represents the parameters aka weights of the p and q models.

There are various ways of optimizing the lower bound $\mathcal{L}(x; \theta)$; for continous z, it can done efficiently by using the re-parameterization trick introduced in (Kingma and Welling, 2013). This way of optimizing the variational lower bound with a parametric inference network and reparameterization of continuous latent variables is usually called Variational Autoencoder (VAE) and draws its inspiration from Helmholtz machines (Dayan *et al.*, 1995). The "autoencoding" terminology comes from the fact that the lower bound $\mathcal{L}(x; \theta)$ can be re-arranged:

$$\mathcal{L}(x; \theta) = \mathbb{E}_{q(z|x)}[\log p(x|z)] - D_{KL}(q(z|x)||p(z)) \tag{4.3}$$

where the first term can be seen as the expectation of negative reconstruction error and the KL divergence term can be seen as a regularizer, which as a whole could be seen as a regularized autoencoder loss with $q(z|x)$ being the encoder and $p(x|z)$ being the decoder. In the context of 2D images modeling, the decoding distribution $p(x|z)$ is usually chosen to be a simple factorized distribution, i.e. $p(x|z) = \prod_i p(x_i|z)$, and

this setup often yields a sharp decoding distribution $p(x|z)$ that tends to reconstruct original datapoint x exactly.

4.3.2 Generative Adversarial Networks

Generative Adversarial Networks are implicit latent variable models that overcome the intractable marginal likelihood computation performed in VAEs by never actually computing them. Instead they rely on implicit signals to learn the parameters of their generative model. In their vanilla form, it consists of two adversarial model, a generator G and a discriminator D, playing a mini-max game. G is a generative model or neural network that tries to capture the data distribution whereas D is a discriminative model that tries to estimate the probability whether the sample came from training data $\mathcal{X}$ or was generated by G. Both G and D are designed to learn a non-linear mapping using neural network architecture.

To learn the generative distribution p_g over data $\mathcal{X}$, the generator builds a mapping function from a prior noise distribution $p_z(z)$ to data space as $G(z;\theta_g)$. And the discriminator, $D(x;\theta_d)$, outputs a single scalar representing the probability that sample x came form training data $\mathcal{X}$ rather than p_g. Here, θ_g and θ_d are parameters of the Generator G and discriminator D respectively.

In the vanilla form, G and D are both trained simultaneously: we adjust parameters for G to minimize $\log(1 - D(G(z))$ and adjust parameters for D to minimize $\log D(X)$, as if they are following the two-player min-max game with value function $V(G, D)$:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{x\sim p_x(z)}[\log(1 - D(G(z)))] \tag{4.4}$$

Around their inception, these models have been criticized with the issues related

to unstable training arising from saddle point optimization problem as well as the generator only partially covering the actual data distribution, termed as the mode collapse problem. Several improvements have been made on top of this vanilla formulation (Arjovsky and Bottou, 2017; Arjovsky *et al.*, 2017; Gulrajani *et al.*, 2017; Miyato *et al.*, 2018) as well as empirical results (Radford *et al.*, 2015; Salimans *et al.*, 2016; Rosca *et al.*, 2017; Brock *et al.*, 2018) have resulted in various tricks that tend to help stabilize the training of the two networks and mitigate the issue of mode-collapse.

4.4 Using Cross Modal Hallucination

In this section we discuss a direct cross modal embedding hallucination formulation based upon the Joint VAE-GAN framework inspired from (Vedantam *et al.*, 2017). We learn three separate embedding spaces, two of them capturing unique information contained in the two input modalities and a shared embedding space capturing the common information between the two modalities. The overall architecture for training and test phase can be summed up in Figure 4.1.

Instead of using an attribute vector as in (Vedantam *et al.*, 2017), we train our model on unstructured text. Using the Joint-VAE formulation, we learn a separate latent space for both the input text and image. We also learn a shared latent space in order to capture the space of common concepts present in the image and the corresponding text. We define the joint distribution as:

$$p_\theta(x, y, z) = p_\theta(z)p_\theta(x|z)p_\theta(y|z) \tag{4.5}$$

Here, $p_\theta(x|z)$ and $p_\theta(y|z)$ are the decoders for image and the text respectively. The objective function similar to (Vedantam *et al.*, 2017) can be written as :

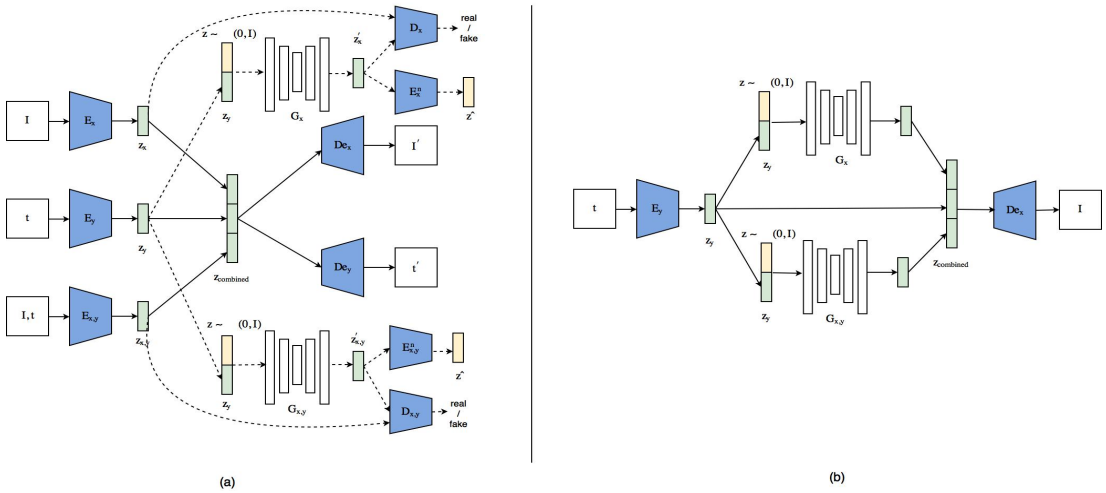

Figure 4.1: Joint-VAE-GAN Network Architecture to Perform Text-to-Image Translation Task by Hallucinating Image and Shared Embedding from Text Embeddings. Part (a) Refers to the Network Being Used During the Training Phase Whereas Part (b) Refers to the Network Being Used During Inference. The Image (I), Text (t) and Shared (I, t) Encoders Are Denoted by E, the Decoders by De. Image and Shared Space Hallucinators Are Shown by G and Their Discriminators and Encoders by D and E^n Respectively.

$$
\begin{aligned}
\mathcal{L}_{Joint-VAE}(x,y) = & \mathop{\mathbb{E}}_{z\sim q(z|x,y)}[\lambda_x^{xy} log(p(x|z)) + \lambda_y^{xy} log(p(y|z))] \\
& - KL(q(z|x,y), p(z)) + \mathop{\mathbb{E}}_{z\sim q(z|x)}[\lambda_x^{x} log(p(x|z))] \\
& - KL(q(z|x), p(z)) + \mathop{\mathbb{E}}_{z\sim q(z|y)}[\lambda_y^{y} log(p(y|z))] \\
& - KL(q(z|y), p(z)) \qquad (4.6)
\end{aligned}
$$

Here, x and y for the input image and text pair. $p(z)$ is the prior on the latent space, where, $p(z) \sim N(0, I)$ and modeling the joint distribution of $p(z, x, y)$ helps us learn a joint posterior space. In order to have access to the image and shared embedding spaces during inference, we train two conditional GANs as side arms of the Join-VAE framework as can be seen in Figure 4.1. GANs train a generator G and discriminator D by formulating their objective as an adversarial game. The discriminator attempts to differentiate between real target from the dataset and fake samples produced by the

generator. In our setting, our real as well as the fake supports are moving objectives converging over time when the setup is being trained in an end-to-end fashion. Here, as the Joint-VAE learns a better embedding space, the discriminator gets better input as real targets and thus over time provides better loss to the generator. This helps us to prevent the case where the discriminator gets too strong too quickly. Thus, it allows the generator to learn to map from the text to the image and shared embedding space gradually.

We use a conditional version of LSGAN (Mao *et al.*, 2017) and can write the objective functions for D and G for the image embedding hallucinator as follows:

$$\begin{aligned}\min_D V_{LSGAN}(D) = 1/2 \mathop{\mathbb{E}}_{e_x \sim p(z|x), e_y \sim p(z|y)} [(D(e_x, e_y) - 1)^2] \\ + 1/2 \mathop{\mathbb{E}}_{z \sim p_z(z), e_y \sim p(z|y)} [(D(G(z, e_y), e_y))^2]\end{aligned} \tag{4.7}$$

$$\min_G V_{LSGAN}(G) = 1/2 \mathop{\mathbb{E}}_{z \sim p_z(z), e_y \sim p(z|y)} [(D(G(z, e_y), e_y) - 1)^2] \tag{4.8}$$

Here, e_y is the text embedding learned by the Joint-VAE-GAN. Similar to (Mirza and Osindero, 2014), we concatenate the real or the generated image e_x or shared space $e_{x,y}$ embedding with the conditioning text embedding e_y before passing it to D. In order to encourage the generator to utilize the latent code z, we use the technique described as bijectivity constraint in (Zhu *et al.*, 2017b) and encode the output of the generator using an encoder E. The input latent code z to the generator is sampled from $\mathcal{N}(0, I)$ whereas the encoder is trying to obtain it's point estimate $\hat{z} = E(G(e_y, z))$. The overall loss for our hallucinators can thus be written as follows where $\mathcal{L}_1^{latent}$ is the ℓ_1 loss between the input latent code and the encoded generator output.

$$\mathcal{L}_{hal} = arg \min_{G,E} \max_D \mathcal{L}_{LSGAN}(G, D) + \lambda_{latent} \mathcal{L}_1^{latent}(G, E) \tag{4.9}$$

During inference, the input for our model is the unstructured text providing a sematic description for the image. This is first embedded using the text encoder of the Joint-VAE-GAN framework to obtain e_y. This embedding is further used to generate the image e_x and the shared embedding space $e_{x,y}$ using their respective conditional GANs. The combined embedding can then be passed through the image decoder of the Joint-VAE-GAN framework.

4.4.1 Implementation Details

We implemented our image encoder with convolution and batch-normalization layers and leaky relu as the activation function. This results in a 512 dimensional mean and stddev vectors. Using the reparameterization trick (Kingma and Welling, 2013), we obtain our image latent space vector. The text encoder is a simple 2 layer mlp. It's input is the char-cnn-rnn embeddings provided by (Reed *et al.*, 2016a). The output is a 512 dimensional conditionally augmented or re-parameterized (Kingma and Welling, 2013; Zhang *et al.*, 2017a) embedding vector. For the shared encoder, the image is first convolved and down-sampled, using the same fundamental units as used in the image encoder, to a 4x4x512 tensor. The char-cnn-rnn embeddings are encoded using an mlp to a 256 dimensional vector. This vector is then spatially replicated before performing a depth-wise concatenation with the image feature map. A 1x1 convolution is performed on this feature map to reduce it's depth dimension to 512. It is further convolved, flattened and reduced to a 512 dimensional mean and stddev vector. Re-parameterization trick (Kingma and Welling, 2013) is applied on the mean and the stddev vectors to produce the shared space embedding.

Our image decoder follows the DCGAN's (Radford *et al.*, 2015) discriminator's architecture. The text decoder is a mlp mapping back to the char-cnn-rnn embedding's dimension. The image, text and shared space embeddings are concatenated

before passing them through the decoders.

We use a U-net (Ronneberger *et al.*, 2015) based architecture for the conditional embedding space hallucinators. The architecture has been shown to produce strong results in the unimodal image prediction setting. We also experimented with multiple other networks for the hallucinators but found u-net based network to produce better results.

4.4.2 Result and Failure Analysis

To quantitatively evaluate the performance of our model, we calculated the inception score (Salimans *et al.*, 2016) metric for this version of our formulation. Inception score tries to formalize the concept of realism for a generated set of images by breaking the concept into two criteria:

- Every realistic image should be recognizable, which means that the score distribution for it must be, ideally, dominated by one class.
- Class distribution over the whole sample should be as close to uniform as possible, in other words, a good generator is a diverse generator.

$$I = \exp \mathbb{E}_x[D_{KL}(p(y|x)||p(y))] \tag{4.10}$$

It does so by computing the average of the KL-divergences between the conditional label distributions of samples and marginal distributions obtained from all the samples as shown in eq. 4.10. As suggested in (Salimans *et al.*, 2016) and similar to (Zhang *et al.*, 2017a), we evaluate this metric on 30k samples by passing them on a pre-trained inception model.The results are given in Table 4.1.

Some qualitative results can be shown in Figure 4.2. We perform additional post-processing step to sharpen the output images similar to (Mansimov *et al.*, 2015).

Metric	GAN-INT-CLS (Reed *et al.*, 2016b)	StackGAN-I (Zhang *et al.*, 2017a)	Ours (Joint-VAE-GAN)	HDGAN (Zhang *et al.*, 2018)
Inception Score	2.88 ± 0.04	2.95 ± 0.02	2.89 ± 0.06	3.53 ± 0.03

Table 4.1: Inception Score for Joint-VAE-GAN Formulation for 64x64 Images.

Subsequent analysis of the architecture and the results brought forth the following issues with this formulation:

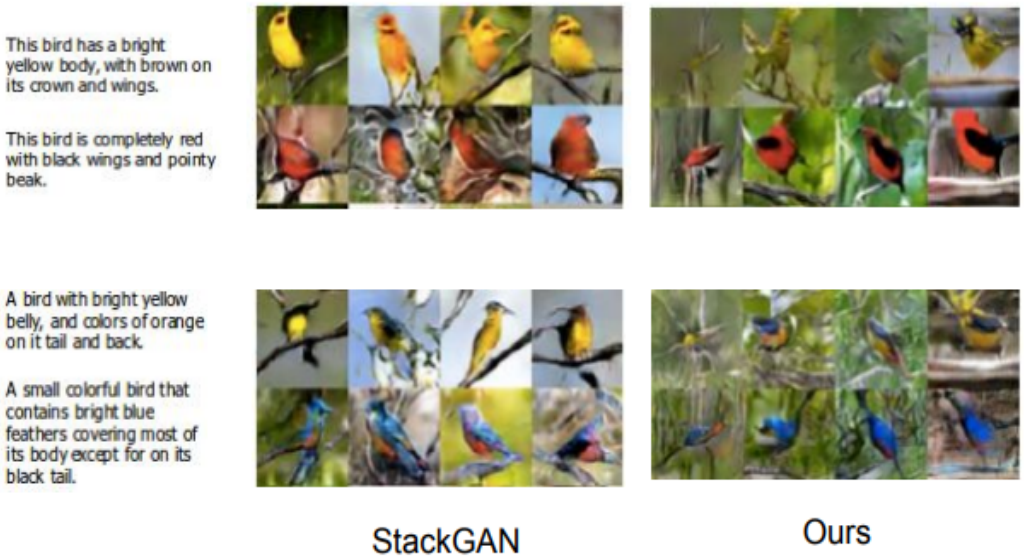

Figure 4.2: Qualitative Results of Our Joint-VAE-GAN Formulation and It's Comparison to StackGan. In Ours, the Rightmost Image Is Generated at the Caculated Latent Space During Inference for the given Input Text. The Images to the Left Are Generated by Interpolating the Hallucinated Shared Latent Space While Keeping the Other Latent Spaces as Constant.

- **Collapse between image and shared latent space:** As there are no explicit constraints on what unique information the image and the shared latent space should capture, the two can collapse to represent the same information. This is specially true for the chosen dataset for experiments as in the Caltech-UCSD Birds dataset, the images contain all of the information that can be gathered from the text providing the semantic description.

  This issue can be mitigated by learning only two representational spaces.

One for the shared information between the image and the semantic text description and the other for the rest of the information present in the image.

- **Variance over-estimation in latent space:** Empirically as shown in Figure 4.2, while interpolating the latent space, the image quickly transitions from a viable image for the given text to noise. This was seen to be true even when very small steps were taken during interpolation. This can be attributed to the variance over-estimation problem often seen in the latent space of VAE (Bowman *et al.*, 2015; Zhao *et al.*, 2017) in that it tends to overfit data, and in the mean time, learn a $q_\phi(z)$ that has variance tending to infinity. This matters in practice when dataset is small compared to the difficulty of the task which is true in our case.

  One of the most straight-forward way of solving this issue is to anneal the KL-divergence loss in the evidence lowerbound formulation.

4.5 Using a Single Shared Embedding Space

In this section we will discuss a single latent space formulation inspired by the image-to-image translation design of (Liu *et al.*, 2017). Here, we make a shared-latent space assumption which says that the corresponding text and image pairs can be mapped to the same shared-latent space. Both the text and the image modalities are modeled using the VAE-GAN framework (Larsen *et al.*, 2015). We also enforce a weight-sharing constraint similar to (Liu *et al.*, 2017) which interacts with the adversarial training objectives to enforce the shared latent space to generate corresponding images and text across domain. The VAEs relate the translated images and text with input images and text in their respective domains. The shared space assumption can be depicted in Figure 4.3

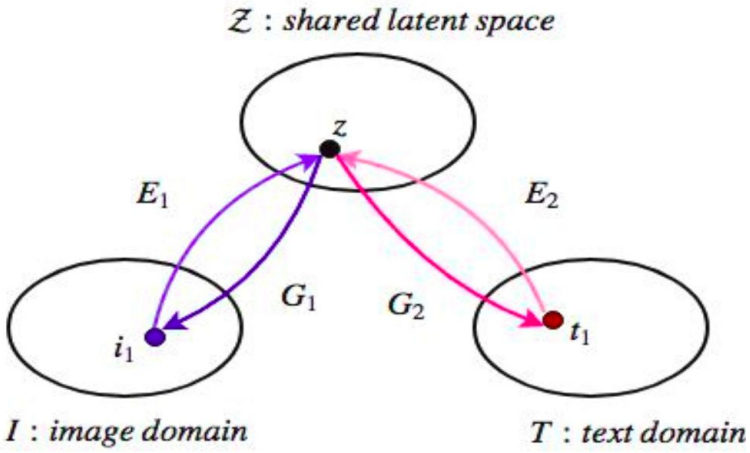

Figure 4.3: The Shared-Latent Space Assumption: We Assume a Pair of Corresponding Images and Text (i_1, t_1) from the Image and the Text Domains Can Be Mapped to the Same Latent Embedding z in the Shared Latent Space $\mathcal{Z}$. Here, E_1 and E_2 Are Encoders Mapping Images and Text to Their Latent Codes Respectively. G_1 and G_2 Are Generators Mapping from the Latent Code to Their Respective Domains.

The overall architecture with the weight-sharing constraint can be depicted in Figure 4.4. The encoder-generator pair $\{E_1, G_1\}$ forms the VAE for the image modality, termed VAE_1. It maps the input image i to a code in the latent space $\mathcal{Z}$ via the encoder E_1 and produces both mean μ and variance σ for the latent space unlike (Liu *et al.*, 2017). This is re-parameterized to obtain the latent code vector and is used to reconstruct the input image via the generator G_1. We assume the components in the latent space $\mathcal{Z}$ are conditionally independent and gaussian with unit variance. Similarly, $\{E_2, G_2\}$ constitues the VAE for the text domain, termed VAE_2. The generator-discriminator pair $\{G_1, D_1\}$ forms the GAN for the image domain and the network combination $\{E_1, G_1, D_1\}$ forms the VAE-GAN (Larsen *et al.*, 2015) for the image domain.

The task of translation from input text t to output generated image $i^{t \to i}$ during inference can be achieved by first encoding t using E_2 to the shared space $\mathcal{Z}$ to obtain the shared latent code z. Then using the image domain generator G_1, this latent code

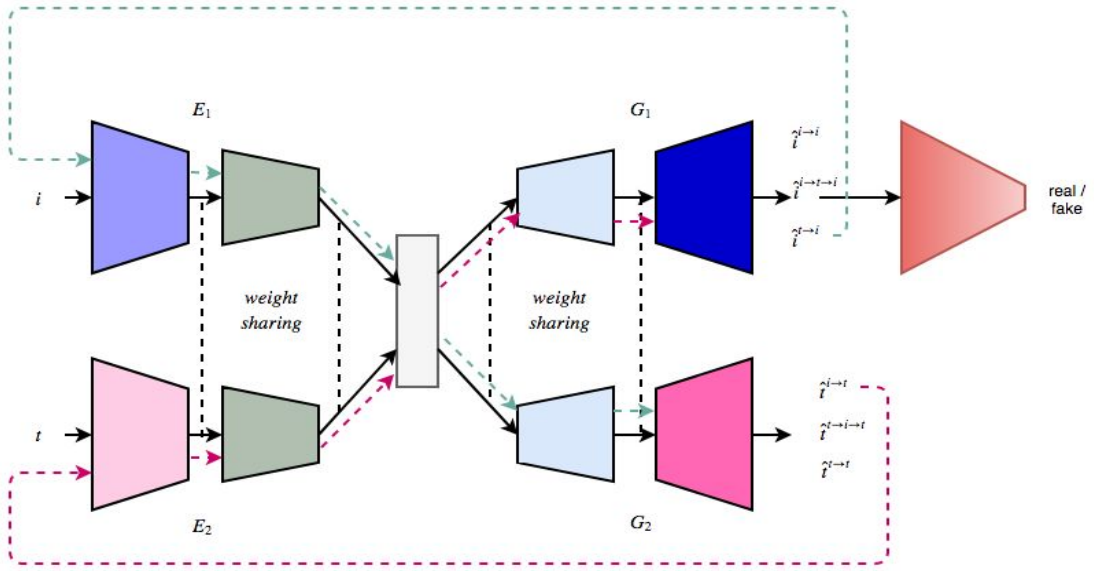

Figure 4.4: Single Shared-Latent Space VAE-GAN Architecture: Here, E_1 and E_2 Are Encoders Mapping Images and Text to Their Latent Codes Respectively. G_1 and G_2 Are Generators Mapping from the Latent Code to Their Respective Domains. The Weight Sharing Constraint Is Implemented by Tying the Weights of the Last Few Layers of E_1, E_2 and G_1, G_2 Respectively (as Shown by the Dashed Black Lines). $i^{i \to i}$ and $t^{t \to t}$ Are Self-reconstructed Images and Text Respectively. $i^{t \to i}$ and $t^{i \to t}$ Are Cross-domain Generated Images and Text Respectively. D_1 Is the Discriminator for the Image Domain. $\hat{i}^{i \to t \to i}$ Shows the Cyclically Reconstructed Image (Dashed Pink Lines) and $\hat{t}^{t \to i \to t}$ Is the Cyclically Reconstructed Text (Dashed Cyan Lines).

can be translated to the corresponding sample in the image domain.

As the shared-latent space assumption implies the cyclic-consistency constraint (Liu *et al.*, 2017; Zhu *et al.*, 2017a), the overall learning task boils down to jointly solving VAE_1 for the image domain, VAE_2 for the text domain, adversarial loss using GAN for both the reconstructed image and the translated image (unlike (Liu *et al.*, 2017), we pass both the translated and the reconstructed images from G_1 to the discriminator D_1. This helps in the learning process as pointed out also in (Rosca *et al.*, 2017)) and the cyclic-reconstruction losses imposing the cyclic-consistency constraints. The overall loss function can be stated in eq. 4.11

$$\min_{E_1,E_2,G_1,G_2} \max_{D_1} \mathcal{L}_{VAE_1}(E_1,G_1) + \mathcal{L}_{GAN_{recon}}(E_1,G_1,D_1) + \mathcal{L}_{GAN_{trans}}(E_2,G_1,D_1) + \mathcal{L}_{CC_1}(E_1,G_1,E_2,G_2) + \mathcal{L}_{VAE_2}(E_2,G_2) + \mathcal{L}_{CC_2}(E_2,G_2,E_1,G_1) \tag{4.11}$$

The VAE aims for minimizing the variational upperbound as stated in eq. 4.3. The VAE objectives for the image and the text domains can be written in eq. 4.12 and eq. 4.13 respectively.

$$\mathcal{L}_{VAE_1}(E_1,G_1) = \lambda_1 KL(q_{E_1}(z|i)||p_z(z)) - \lambda_2 \mathop{\mathbb{E}}_{z\sim q_{E_1}(z|i)}[log(p_{G_1}(\hat{i}^{i\to i}|z))] \tag{4.12}$$

$$\mathcal{L}_{VAE_2}(E_2,G_2) = \lambda_1 KL(q_{E_2}(z|t)||p_z(z)) - \lambda_2 \mathop{\mathbb{E}}_{z\sim q_{E_2}(z|t)}[log(p_{G_2}(\hat{t}^{t\to t}|z))] \tag{4.13}$$

Here, $\hat{i}^{i\to i}$ and $\hat{t}^{t\to t}$ referes to the reconstructed image and the text from their respective domains. We have an image discriminator D_1 in our formulation. Unlike (Liu *et al.*, 2017), we use both the reconstructed and the translated images for the adversarial loses as this helps in further stabilizing the optimization problem. The translated and the reconstructed adversarial losses can be summed up in eq. 4.14 and eq. 4.15 respectively.

$$\mathcal{L}_{GAN_{trans}}(E_2,G_1,D_1) = \lambda_0 \mathop{\mathbb{E}}_{i\sim P_I}[log(D_1(i))] + \lambda_0 \mathop{\mathbb{E}}_{z\sim q_{E_2}(z|t)}[log(1-D_1(G_1(z))] \tag{4.14}$$

$$\mathcal{L}_{GAN_{recon}}(E_1,G_1,D_1) = \lambda_0 \mathop{\mathbb{E}}_{i\sim P_I}[log(D_1(i))] + \lambda_0 \mathop{\mathbb{E}}_{z\sim q_{E_1}(z|i)}[log(1-D_1(G_1(z))] \tag{4.15}$$

As previously stated, the shared-latent space assumption implies the cyclic-consistency constraint. We have explicitly provided this constraint as part of our optimization problem as the task of learning multimodal translation using joint distribution is ill posed and having more constraints is beneficial. The cyclic-consistency constraints are outlined in eq. 4.16 and eq. 4.17. Here, $t^{i\to t}$ is the text translated from image and $i^{t\to i}$ is the image translated from the corresponding text.

$$\mathcal{L}_{CC_1}(E_1, G_1, E_2, G_2) = \lambda_3 KL(q_i(z|i)||p_z(z)) + \lambda_3 KL(q_t(z|t^{i\to t})||p_z(z)) - \lambda_4 \mathop{\mathbb{E}}_{z\sim q_t(z|t^{i\to t})}[log(p_{G_1}(i|z))] \tag{4.16}$$

$$\mathcal{L}_{CC_2}(E_2, G_2, E_1, G_1) = \lambda_3 KL(q_t(z|t)||p_z(z)) + \lambda_3 KL(q_i(z|i^{t\to i})||p_z(z)) - \lambda_4 \mathop{\mathbb{E}}_{z\sim q_i(z|i^{t\to i})}[log(p_{G_2}(t|z))] \tag{4.17}$$

We have modeled p_{G_1} and p_{G_2} as Gaussian distributions and utilized binary cross-entropy losses for our reconstruction terms unlike (Liu *et al.*, 2017) which treats them as Laplacian distributions. The prior distribution $p_z(z)$ is a zero mean Gaussian $p_z(z) \sim \mathcal{N}(z|0, I)$. The hyperparameters λ_1 and λ_2 control the contribution of the KL and the reconstruction terms of VAE_1 and VAE_2 respectively. The hyperparameter λ_0 controls the impact of the GAN objective functions. The hyperparmaeters λ_3 and λ_4 control the weights of the KL-divergence terms and the cyclic reconstruction terms in the cyclic consistency objective functions.

4.5.1 Implementation Details

We implemented both the image and text encoder and generator in fully convolutional manner. The image encoder and generator resemble DCGANs discriminator and generator architectures respectively. The basic building blocks are convolutional and batch-normalization layers followed by leaky-relu as the activation function. The text encoder and generator follow the encoder and decoder architectures provided in (Zhang *et al.*, 2017b). The output of the text generator for each word in the output sentence is a probability distribution over all of the unique words in the text vocabulary. The reconstruction loss is taken between the input list of vocabulary index and the indices with the highest probaility value for each of the words in the output sentence. The image discriminator is a multi-scale discriminator (Karras *et al.*, 2017).

The implementation was validated on the Caltech-UCSD CUB-Birds dataset (Wah *et al.*, 2011b).The shared-latent space code is a vector of dimension 512. We tried multiple values for our hyperparameters and settled upon the following dictionary of values: $\{\lambda_0 : 10.0, \lambda_1 : 0.1, \lambda_2 : 50.0, \lambda_3 : 0.1, \lambda_4 : 100.0\}$. Unlike (Liu *et al.*, 2017), we also perform logistic annealing of our KL-divergence term (Bowman *et al.*, 2015). We also experimented with the trade-off between the number of iterations of the VAE-GAN generator and the discriminator and found 2 discriminator iterations per image and text VAE-GAN iteration combined to be most useful. We used Adam optimizer for training with a learning rate of 0.0001 for both the VAE-GAN generator and image discriminator. The momentums for the optimizer were set to 0.5 and 0.999. We trained with a batch size of 64. During training, for each of the input image, one text description out of the 5 provided description was chosen at random. We used 300 dimensional glove embeddings to embeded our input sentences. The 80/20 train and test split was achieved on the entire dataset. The experiments were done with both cropped and un-cropped images using the bounding box information provided in the dataset.

4.5.2 Result and Failure Analysis

Language Model Evaluation

As an aside, we also examined the performance of the fully convolutional language model using the Hotel Reviews dataset. We follow the architecture outlined in (Zhang *et al.*, 2017b) for the reconstruction task and have a 4-layer convolutional encoder followed by a 4-layer convolutional decoder. The last layer of the encoder and the first layer of the decoder of this architecture have their weights tied with the image encoder and generator respectively when used as part of the overall architecture.

Unlike (Zhang *et al.*, 2017b), which implements an autoencoder, we utilize the VAE paradigm thus lending our model the ability to generate novel sentences. The loss function of our model is equivalent to eq. 4.13. For the Hotel Reviews dataset, we closely followed the training criterion and hyperparameter values outlined in (Zhang *et al.*, 2017b). We also performed stochastic annealing of the KL-divergence term in eq. 4.13. For quantitative evaluation, we calculated BLEU score (Papineni *et al.*, 2002) between our reconstructed and input senteces similar to (Zhang *et al.*, 2017b). The results are provided in Table 4.2.

Model	**BLEU**
CNN-DCNN (Zhang *et al.*, 2017b)	94.2
Ours	90.8

Table 4.2: BLEU (Papineni *et al.*, 2002) Score Comparison Between Our Language Model and (Zhang *et al.*, 2017b).

Qualitative results for the hotel reviews dataset compairing our output with (Zhang *et al.*, 2017b) is given Table 4.3 as well as some reconstructed and generated samples from the CUB Birds dataset (Wah *et al.*, 2011a) is given in Table 4.4

Ground-truth:	on every visit to nyc , the hotel beacon is the place we love to stay . so conveniently located to central park , lincoln center and great local restaurants . the rooms are lovely . beds so comfortable , a great little kitchen and new wizz bang coffee maker . the staff are so accommodating and just love walking across the street to the fairway supermarket with every imaginable goodies to eat .
CNN-DCNN	on every visit to nyc , the hotel beacon is the place we love to stay . so closely located to central park , lincoln center and great local restaurants . biggest rooms are lovely . beds so comfortable , a great little kitchen and new UNK suggestion coffee maker . the staff turned so accommodating and just love walking across the street to former fairway supermarket with every food taxes to eat .
Ours	on every visit to nyc , the hotel beacon is the place we to to stay . so conveniently located to central park , lincoln center and great lovely restaurants . to rooms are lovely . beds so clean , a great little kitchen and new water bang coffee maker . the staff are so accommodating and just lovely walking across the street to the fairway supermarket with every imagine good to eat .

Table 4.3: Reconstructed Paragraph of the Hotel Reviews Example Used in (Zhang *et al.*, 2017b)

Ground-truth	Ours (reconstructed)	Ours (generated)
this bird has a large head , a black bill , and a white breast .	this bird has a large head , a black bill , and a white breast .	low tourquoise lakc posterior long distinguished beak .
a small bird with a bright red breast region and white in the wingbars region having a white belly and black crown .	a small bird with a yellow red breast belt and white in the wingbars and very a white belly and black crown .	than smaller wings birds black wings blue .
the small bird has long tarsus , short wings , and a medium sized bill .	this small bird has long tarsus with short wings , and a very belly bill .	this purple yellow bird orange blue plumage white under breast .

Table 4.4: Reconstructed and Generated Sentences from the CUB Birds Dataset (Wah *et al.*, 2011a).

Translation Model Evaluation

Similar to previous section, we calculated the inception score (Salimans *et al.*, 2016) metric for our single shared-latent space formulation. The results are provided in Table 4.5.

Metric	GAN-INT-CLS (Reed *et al.*, 2016b)	StackGAN-I (Zhang *et al.*, 2017a)	Ours (un-cropped) (Single Shared-Latent Space)	HDGAN (Zhang *et al.*, 2018)
Inception Score	2.88 ± 0.04	2.95 ± 0.02	2.97 ± 0.03	3.53 ± 0.03

Table 4.5: Inception Score for Single Shared-Latent Space Formulation for 64x64 Images.

Some qualitative results are provided in Figure 4.5. The generated images capture the attributes present in the text as is also evident by the inception score, but the results are blurry. This can be attributed to the VAE-GAN architecture iteself which does generate sharper images compared to vanilla VAE paradigm but they are not as sharp as vanilla GAN outputs because of the reconstruction terms in the loss function. A post-processing step used in the previous section can help in reducing the blurriness of the results. This might also help in increasing the evaluation metric *i.e.* inception score.

As we have tied each of the outputs of the generative manifold with that of input samples, the generative manifold is not as unbounded as using only GAN based

Figure 4.5: Generated Images for Corresponding Text from Single Share-latent Space Text to Image Translation Model.

architecture viz (Zhang *et al.*, 2018). Though this helps in mitigating the issue of mode collapse to some extent, it puts a trade-off on how diverse and novel the generated samples can be and thus hinders achieving higher inception score. Some results of interpolation in the latent space for a given text input is provided in Figure 4.6. From the results we see that interpolation in the latent space using different z vector results in a change in the pose of the generated bird image as well as a change in the background. Unlike previous results for the Joint-VAE-GAN model, there is no apparent issue related to variance over-estimation. To further improve this model, future directions can be into improving the language model and adding a text discriminator D_2 on top of the text generator G_2 to learn a more stable manifold. Incorporating progressive growing strategy similar to (Karras *et al.*, 2017; Zhang *et al.*, 2018) is also going to help in improving the results.

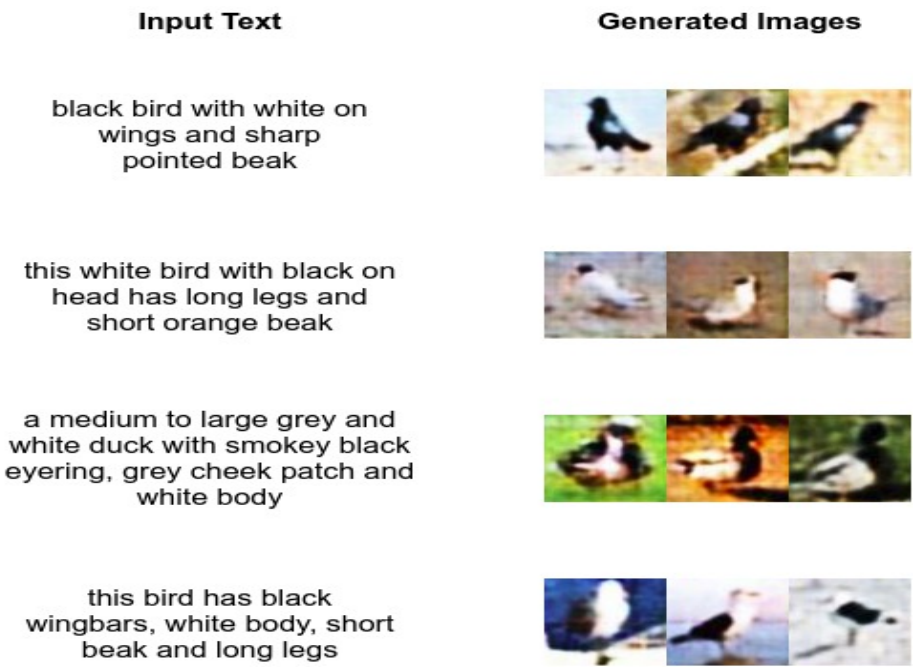

Figure 4.6: Diverse Generated Images for a given Input Text by Interpolating in the Latent Space.

4.6 Conclusion

Multimodal translation tasks are ambiguous. Learning a joint-representation between multiple modalities along-with imposing additional constraints on the said representation can help us learn realize a translation framework. In this chapter, we showcased our text-to-image translation framework using a combination of contemprorary generative modeling techniques. We utilized a variational autoencoder in conjunction with a generative adversarial network to learn a shared representation between the text and the image modalities. We demonstrated the outcomes of such a formulation and provided both qualitative and quantitative results, using the inception score, on publicly available dataset. We also analyzed failure cases and provided possible ways to overcome these issues and future directions for the work.

Chapter 5

CONCLUSIONS

The presented work shows the effectiveness of multimodal learning for tackling visual reasoning and multimodal translation tasks. In chapter 3, we showcased how a multimodal framework can be adopted to distill the representation learned from external knowledge into existing neural network architectures. We answered the fundamental questions related to the whereabouts of such knowledge for the task of visual reasoning. Specifically we showed how additional information present in the form of scene-graph information can be integrated with existing architectures by leveraging probabilistic reasoning mechanisms. The probabilistic soft logic engine was used identify the object mentions in the questions and their corresponding location in the text to generate a spatial mask over the image depicting regions of interest. This knowledge was distilled into existing visual reasoning architectures in a generalized knowledge distillation framework. We also demonstrated how such a representation can be emulated inside the model using attention. The efficacy of the framework was demonstrated on two publicly available datasets i.e. CLEVR and Sort-of-Clevr.

In chapter 4, we tackled the task of multimodal translation and proposed a gneeralized framework for text-to-image translation. Using the publicly available Caltech-UCSD Birds-200-2011 dataset, we demonstrated how semantic description of images can be consumed in our framework to generate novel image samples that retain the properties depicted in the text. Using the inception score as the quantitative criteria, we provided how our framework fares against other research in this task. We also analyzed the failure cases of our framework and proposed ways to overcome them which we will be incorporating in our future work.

REFERENCES

Antol, S., A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick and D. Parikh, "Vqa: Visual question answering", in "Proceedings of the IEEE International Conference on Computer Vision", pp. 2425–2433 (2015).

Arjovsky, M. and L. Bottou, "Towards principled methods for training generative adversarial networks", arXiv preprint arXiv:1701.04862 (2017).

Arjovsky, M., S. Chintala and L. Bottou, "Wasserstein gan", arXiv preprint arXiv:1701.07875 (2017).

Bach, S. H., M. Broecheler, B. Huang and L. Getoor, "Hinge-loss markov random fields and probabilistic soft logic", Journal of Machine Learning Research (JMLR) To appear (2017).

Bahdanau, D., K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate", arXiv preprint arXiv:1409.0473 (2014).

Bahrick, L. E., "Infants' perception of substance and temporal synchrony in multimodal events", Infant Behavior and Development **6**, 4, 429–451 (1983).

Baral, C., *Knowledge representation, reasoning and declarative problem solving* (Cambridge university press, 2003).

Bengio, S., "An asynchronous hidden markov model for audio-visual speech recognition", in "Advances in Neural Information Processing Systems", pp. 1237–1244 (2003a).

Bengio, S., "Multimodal authentication using asynchronous hmms", in "International Conference on Audio-and Video-Based Biometric Person Authentication", pp. 770–777 (Springer, 2003b).

Bengio, Y., P. Lamblin, D. Popovici and H. Larochelle, "Greedy layer-wise training of deep networks", in "Advances in neural information processing systems", pp. 153–160 (2007).

Berthelot, D., T. Schumm and L. Metz, "Began: Boundary equilibrium generative adversarial networks", arXiv preprint arXiv:1703.10717 (2017).

Bowman, S. R., L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz and S. Bengio, "Generating sentences from a continuous space", arXiv preprint arXiv:1511.06349 (2015).

Brock, A., J. Donahue and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis", arXiv preprint arXiv:1809.11096 (2018).

Chen, X., Y. Duan, R. Houthooft, J. Schulman, I. Sutskever and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets", in "Advances in Neural Information Processing Systems", pp. 2172–2180 (2016).

Culotta, A., R. Bekkerman and A. McCallum, "Extracting social networks and contact information from email and the web", Tech. rep., MASSACHUSETTS UNIV AMHERST DEPT OF COMPUTER SCIENCE (2005).

Dash, A., J. C. B. Gamboa, S. Ahmed, M. Liwicki and M. Z. Afzal, "Tac-gan-text conditioned auxiliary classifier generative adversarial network", arXiv preprint arXiv:1703.06412 (2017).

Dayan, P., G. E. Hinton, R. M. Neal and R. S. Zemel, "The helmholtz machine", Neural computation **7**, 5, 889–904 (1995).

De Marneffe, M.-C., B. MacCartney, C. D. Manning *et al.*, "Generating typed dependency parses from phrase structure parses", in "Proceedings of LREC", vol. 6 (2006).

De Raedt, L., A. Kimmig and H. Toivonen, "Problog: A probabilistic prolog and its application in link discovery", in "Proceedings of the 20th International Joint Conference on Artifical Intelligence", IJCAI'07, pp. 2468–2473 (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2007).

Efros, A. A. and T. K. Leung, "Texture synthesis by non-parametric sampling", in "Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on", vol. 2, pp. 1033–1038 (IEEE, 1999).

Elliott, D. and F. Keller, "Image description using visual dependency representations", in "Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing", pp. 1292–1302 (2013).

Fels, S. S. and G. E. Hinton, "Glove-talk: A neural network interface between a data-glove and a speech synthesizer", IEEE transactions on Neural Networks **4**, 1, 2–8 (1993).

Gao, Y., O. Beijbom, N. Zhang and T. Darrell, "Compact bilinear pooling", in "Proceedings of the IEEE conference on computer vision and pattern recognition", pp. 317–326 (2016).

Gauthier, J., "Conditional generative adversarial nets for convolutional face generation", Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester **2014**, 5, 2 (2014).

Goodfellow, I., 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", pp. 2672–2680 (2014).

Gulrajani, I., F. Ahmed, M. Arjovsky, V. Dumoulin and A. C. Courville, "Improved training of wasserstein gans", in "Advances in Neural Information Processing Systems", pp. 5769–5779 (2017).

Hinton, G., O. Vinyals and J. Dean, "Distilling the Knowledge in a Neural Network", http://arxiv.org/pdf/1503.02531v1.pdf (2015).

Hinton, G. E., "Deep belief networks", Scholarpedia **4**, 5, 5947 (2009).

Hinton, G. E. and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks", science **313**, 5786, 504–507 (2006).

Hu, Z., X. Ma, Z. Liu, E. Hovy and E. Xing, "Harnessing deep neural networks with logic rules", in "Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)", pp. 2410–2420 (Association for Computational Linguistics, Berlin, Germany, 2016a), URL `http://www.aclweb.org/anthology/P16-1228`.

Hu, Z., Z. Yang, R. Salakhutdinov and E. Xing, "Deep neural networks with massive learned knowledge", in "Proceedings of the 2016 Conference on EMNLP", pp. 1670–1679 (ACL, Austin, Texas, 2016b), URL `https://aclweb.org/anthology/D16-1173`.

Isola, P., J.-Y. Zhu, T. Zhou and A. A. Efros, "Image-to-image translation with conditional adversarial networks", arXiv preprint (2017).

Jain, U., S. Lazebnik and A. G. Schwing, "Two can play this game: visual dialog with discriminative question generation and answering", in "The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)", (2018).

Johnson, J., B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning", arXiv preprint arXiv:1612.06890 (2016).

Karras, T., T. Aila, S. Laine and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation", arXiv preprint arXiv:1710.10196 (2017).

Kim, Y. and A. M. Rush, "Sequence-level knowledge distillation", arXiv preprint arXiv:1606.07947 (2016).

Kingma, D. P., S. Mohamed, D. J. Rezende and M. Welling, "Semi-supervised learning with deep generative models", in "Advances in Neural Information Processing Systems", pp. 3581–3589 (2014).

Kingma, D. P. and M. Welling, "Auto-encoding variational bayes", arXiv preprint arXiv:1312.6114 (2013).

Klir, G. and B. Yuan, "Fuzzy sets and fuzzy logic: theory and applications", (1995).

Lafferty, J., A. McCallum and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", (2001).

Lake, B. M., T. D. Ullman, J. B. Tenenbaum and S. J. Gershman, "Building machines that learn and think like people", Behavioral and Brain Sciences pp. 1–101 (2016).

Larsen, A. B. L., S. K. Sønderby, H. Larochelle and O. Winther, "Autoencoding beyond pixels using a learned similarity metric", arXiv preprint arXiv:1512.09300 (2015).

Lazarus, A. A. *et al.*, "Multimodal behavior therapy: I.", (1976).

LeCun, Y., "Une procedure d'apprentissage ponr reseau a seuil asymetrique", proceedings of Cognitiva 85 pp. 599–604 (1985).

LeCun, Y., "A path to ai", https://futureoflife.org/wp-content/uploads/2017/01/Yann-LeCun.pdf (2017).

Ledig, C., 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", arXiv preprint (2016).

Li, Y., D. McLean, Z. A. Bandar, J. D. O'shea and K. Crockett, "Sentence similarity based on semantic nets and corpus statistics", IEEE transactions on knowledge and data engineering **18**, 8, 1138–1150 (2006).

Lindvall, T., *Lectures on the coupling method* (Courier Corporation, 2002).

Liu, M.-Y., T. Breuel and J. Kautz, "Unsupervised image-to-image translation networks", in "Advances in Neural Information Processing Systems", pp. 700–708 (2017).

London, B., S. Khamis, S. Bach, B. Huang, L. Getoor and L. Davis, "Collective activity detection using hinge-loss markov random fields", in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops", pp. 566–571 (2013).

Lopez-Paz, D., L. Bottou, B. Schölkopf and V. Vapnik, "Unifying distillation and privileged information", URL `http://arxiv.org/abs/1511.03643`, cite arxiv:1511.03643 (2015).

Mansimov, E., E. Parisotto, J. L. Ba and R. Salakhutdinov, "Generating images from captions with attention", arXiv preprint arXiv:1511.02793 (2015).

Mao, X., Q. Li, H. Xie, R. Y. Lau, Z. Wang and S. P. Smolley, "Least squares generative adversarial networks", in "2017 IEEE International Conference on Computer Vision (ICCV)", pp. 2813–2821 (IEEE, 2017).

Massiceti, D., N. Siddharth, P. K. Dokania and P. H. Torr, "Flipdial: A generative model for two-way visual dialogue", in "The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)", (2018).

McCowan, I., J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos *et al.*, "The ami meeting corpus", in "Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research", vol. 88, p. 100 (2005).

McGurk, H. and J. MacDonald, "Hearing lips and seeing voices", Nature **264**, 5588, 746 (1976).

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

Miyato, T., T. Kataoka, M. Koyama and Y. Yoshida, "Spectral normalization for generative adversarial networks", arXiv preprint arXiv:1802.05957 (2018).

Mulligan, R. M. and M. L. Shaw, "Multimodal signal detection: Independent decisions vs. integration", Perception & Psychophysics **28**, 5, 471–478 (1980).

Mun, J., M. Cho and B. Han, "Text-guided attention model for image captioning.", in "AAAI", pp. 4233–4239 (2017).

Ngiam, J., A. Khosla, M. Kim, J. Nam, H. Lee and A. Y. Ng, "Multimodal deep learning", in "Proceedings of the 28th international conference on machine learning (ICML-11)", pp. 689–696 (2011).

Nguyen, A., J. Yosinski, Y. Bengio, A. Dosovitskiy and J. Clune, "Plug & play generative networks: Conditional iterative generation of images in latent space", arXiv preprint arXiv:1612.00005 (2016).

Odena, A., C. Olah and J. Shlens, "Conditional image synthesis with auxiliary classifier gans", arXiv preprint arXiv:1610.09585 (2016).

Oord, A. v. d., N. Kalchbrenner and K. Kavukcuoglu, "Pixel recurrent neural networks", arXiv preprint arXiv:1601.06759 (2016).

Papineni, K., S. Roukos, T. Ward and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation", in "Proceedings of the 40th annual meeting on association for computational linguistics", pp. 311–318 (Association for Computational Linguistics, 2002).

Parker, D. B., "Learning-logic", Tech. Rep. TR-47, Center for Comp. Research in Economics and Management Sci., MIT (1985).

Pennington, J., R. Socher and C. D. Manning, "Glove: Global vectors for word representation", in "Empirical Methods in Natural Language Processing (EMNLP)", pp. 1532–1543 (2014), URL `http://www.aclweb.org/anthology/D14-1162`.

Perez, E., F. Strub, H. De Vries, V. Dumoulin and A. Courville, "Film: Visual reasoning with a general conditioning layer", arXiv preprint arXiv:1709.07871 (2017).

Petajan, E. D., "Automatic lipreading to enhance speech recognition (speech reading)", (1984).

Radford, A., L. Metz and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks", arXiv preprint arXiv:1511.06434 (2015).

Reed, S., Z. Akata, H. Lee and B. Schiele, "Learning deep representations of fine-grained visual descriptions", in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition", pp. 49–58 (2016a).

Reed, S., Z. Akata, X. Yan, L. Logeswaran, B. Schiele and H. Lee, "Generative adversarial text to image synthesis", arXiv preprint arXiv:1605.05396 (2016b).

Reed, S. E., Z. Akata, S. Mohan, S. Tenka, B. Schiele and H. Lee, "Learning what and where to draw", in "Advances in Neural Information Processing Systems", pp. 217–225 (2016c).

Richardson, M. and P. Domingos, "Markov logic networks", Mach. Learn. **62**, 1-2, 107–136, URL `http://dx.doi.org/10.1007/s10994-006-5833-1` (2006).

Ronneberger, O., P. Fischer and T. Brox, "U-net: Convolutional networks for biomedical image segmentation", in "International Conference on Medical image computing and computer-assisted intervention", pp. 234–241 (Springer, 2015).

Rosca, M., B. Lakshminarayanan, D. Warde-Farley and S. Mohamed, "Variational approaches for auto-encoding generative adversarial networks", arXiv preprint arXiv:1706.04987 (2017).

Rumelhart, D. E., G. E. Hinton and R. J. Williams, "Learning internal representations by error propagation", in "Parallel Distributed Processing", edited by D. E. Rumelhart and J. L. McClelland, vol. 1, pp. 318–362 (MIT Press, 1986).

Rumelhart, D. E. and D. Zipser, "Feature discovery by competitive learning", in "Parallel Distributed Processing", pp. 151–193 (MIT Press, 1986).

Salakhutdinov, R. and G. Hinton, "Semantic hashing", International Journal of Approximate Reasoning **50**, 7, 969–978 (2009).

Salimans, T., I. Goodfellow, W. Zaremba, V. Cheung, A. Radford and X. Chen, "Improved techniques for training gans", in "Advances in Neural Information Processing Systems", pp. 2234–2242 (2016).

Santoro, A., D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia and T. Lillicrap, "A simple neural network module for relational reasoning", arXiv preprint arXiv:1706.01427 (2017).

Sohn, K., H. Lee and X. Yan, "Learning structured output representation using deep conditional generative models", in "Advances in Neural Information Processing Systems", pp. 3483–3491 (2015).

Taigman, Y., A. Polyak and L. Wolf, "Unsupervised cross-domain image generation", arXiv preprint arXiv:1611.02200 (2016).

Tur, G., A. Stolcke, L. Voss, J. Dowding, B. Favre, R. Fernández, M. Frampton, M. Frandsen, C. Frederickson, M. Graciarena *et al.*, "The calo meeting speech recognition and understanding system", in "Spoken Language Technology Workshop, 2008. SLT 2008. IEEE", pp. 69–72 (IEEE, 2008).

van den Oord, A., N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves *et al.*, "Conditional image generation with pixelcnn decoders", in "Advances in Neural Information Processing Systems", pp. 4790–4798 (2016).

Vapnik, V. and R. Izmailov, "Learning using privileged information: similarity control and knowledge transfer.", Journal of machine learning research **16**, 55 (2015).

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is all you need", arXiv preprint arXiv:1706.03762 (2017).

Vedantam, R., I. Fischer, J. Huang and K. Murphy, "Generative models of visually grounded imagination", CoRR **abs/1705.10762**, URL `http://arxiv.org/abs/1705.10762` (2017).

Wactlar, H. D., T. Kanade, M. A. Smith and S. M. Stevens, "Intelligent access to digital video: Informedia project", Computer **29**, 5, 46–52 (1996).

Wah, C., S. Branson, P. Welinder, P. Perona and S. Belongie, "The caltech-ucsd birds-200-2011 dataset", (2011a).

Wah, C., S. Branson, P. Welinder, P. Perona and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset", Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011b).

Walker, J., C. Doersch, A. Gupta and M. Hebert, "An uncertain future: Forecasting from static images using variational autoencoders", in "European Conference on Computer Vision", pp. 835–851 (Springer, 2016).

Wang, P., Q. Wu, C. Shen, A. Dick and A. van den Hengel, "Fvqa: fact-based visual question answering", IEEE Transactions on Pattern Analysis and Machine Intelligence (2017).

Werbos, P. J., "Applications of advances in nonlinear sensitivity analysis", in "Proceedings of the 10th IFIP Conference, 31.8 - 4.9, NYC", pp. 762–770 (1981).

Xu, K., J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention", in "ICML", pp. 2048–2057 (2015).

Xu, T., P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang and X. He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks", arXiv preprint (2018).

Yan, X., J. Yang, K. Sohn and H. Lee, "Attribute2image: Conditional image generation from visual attributes", in "European Conference on Computer Vision", pp. 776–791 (Springer, 2016).

Yang, Z., X. He, J. Gao, L. Deng and A. Smola, "Stacked attention networks for image question answering", in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition", pp. 21–29 (2016).

Yeh, R. A., C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson and M. N. Do, "Semantic image inpainting with deep generative models", in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition", pp. 5485–5493 (2017).

Yu, R., A. Li, V. I. Morariu and L. S. Davis, "Visual relationship detection with internal and external linguistic knowledge distillation.", IEEE International Conference on Computer Vision (ICCV) (2017).

Zhang, H., T. Xu, H. Li, S. Zhang, X. Huang, X. Wang and D. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks", in "IEEE Int. Conf. Comput. Vision (ICCV)", pp. 5907–5915 (2017a).

Zhang, Y., D. Shen, G. Wang, Z. Gan, R. Henao and L. Carin, "Deconvolutional paragraph representation learning", in "Advances in Neural Information Processing Systems", pp. 4169–4179 (2017b).

Zhang, Z., Y. Xie and L. Yang, "Photographic text-to-image synthesis with a hierarchically-nested adversarial network", (2018).

Zhao, S., J. Song and S. Ermon, "Infovae: Information maximizing variational autoencoders", arXiv preprint arXiv:1706.02262 (2017).

Zheng, S., S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang and P. H. S. Torr, "Conditional random fields as recurrent neural networks", in "Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)", ICCV '15, pp. 1529–1537 (IEEE Computer Society, Washington, DC, USA, 2015), URL `http://dx.doi.org/10.1109/ICCV.2015.179`.

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

Zhu, J.-Y., R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang and E. Shechtman, "Toward multimodal image-to-image translation", in "Advances in Neural Information Processing Systems", pp. 465–476 (2017b).

APPENDIX A

VISUAL REASONING AND MULTIMODAL REPRESENTATION

A.1 External Mask Prediction Example

We first describe how we obtain the predicate confidence scores for both CLEVR (Johnson *et al.*, 2016) and Sort-of-Clevr (Santoro *et al.*, 2017) datasets. We use the image and the question from Figure 3.2 as the running example.

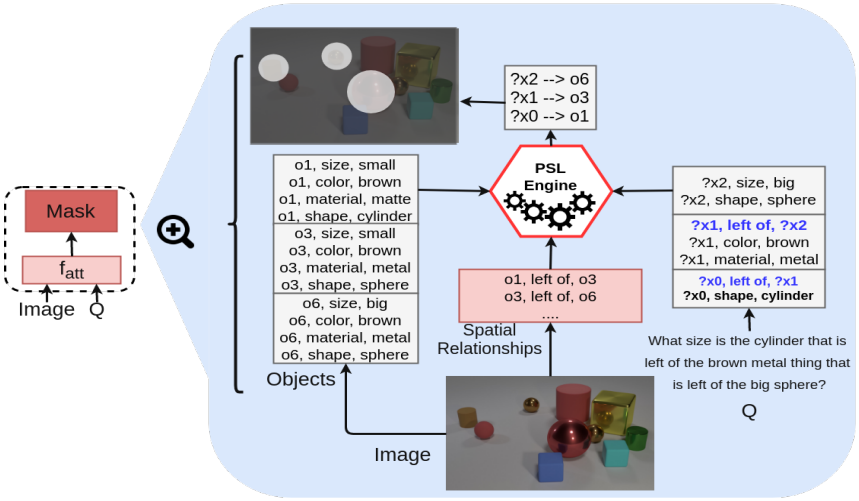

Figure A.1: Internal Process of Mask Creation.

From chapter 3, the two equations required to estimate which object mentions are related to which textual mentions are as follows:

$$w_1 : candidate(M, O) \leftarrow object(O) \wedge mention(M) \\ \wedge attr_o(O, A, V) \wedge attr_m(M, A, V). \tag{A.1}$$

$$w_2 : candidate(M, O) \leftarrow object(O) \wedge mention(M) \\ \wedge candidate(M, O) \\ \wedge candidate(M_1, O_1) \\ \wedge consistent(A, O, O_1, M, M_1). \tag{A.2}$$

$attr_o(O, A, V)$ was directly obtained by leveraging the synthetic data generation process, which is similar to CLEVR dataset generation (Johnson *et al.*, 2016). For example $attr_o(o_1, size, small) = 1.0, attr_o(o_1, material, matte) = 1.0$ for the leftmost brown cylinder for the image I. To obtain confidence scores for $attr_m(M, A, V)$, we parse the natural language question using the Stanford syntactic dependency parser (De Marneffe *et al.*, 2006) to obtain all nouns. For all the nouns, we extract the qualifying adjectives and each qualifying adjective is assigned to an attribute (shape, size, color, material) using a similarity measure (average similarity based on Word2vec and

WordNet [1]). For the example question, we obtain $attr_m(?x0, shape, cylinder) = 1.0$, $attr_m(?x1, color, brown) = 1.0$. Then, for each textual mention M, we maintain a list of objects, where an object is only filtered out if the object and mention have a conflicting property-value pair. To obtain the $consistent(R, O, O_1, M, M_1)$ values, we perform the following steps: 1) for each mention-pair (M, M_1), we choose a corresponding candidate object-pair (O, O_1), 2) for the mention-pair we extract the shortest-path from the syntactic dependency tree and match with the type of attribute (*size, shape, left, right, beside*) using the highest word-similarity measure, 3) if the attribute is a property (such as shape, size, color), then the mentioned relation is found (*same, as large as, larger than, greater than*) and the property values of objects O and O_1 are used to check their consistency. If they are consistent we use 1.0 or else we use 0.0 as the score; and 4) if the attribute is spatial (such as *left to, right to, beside, next*) then we check the spatial relationship and use the confidence of 1.0 if the object-pair O, O_1 is consistent, otherwise we use 0.0; for example $consistent(\text{left}, o3, o6, ?x1, ?x2) = 1.0$ in the example image. Using the above predicate values, we use the PSL engine to infer the candidate objects and calculate the ground-truth mask.

[1] WordNet-based word pair similarities is calculated as a product of `length` (of the shortest path between sysnsets of the words), and `depth` (the depth of the subsumer in the hierarchical semantic net) (Li *et al.*, 2006).