

# Neural Machine Translation for Multimodal Interaction

Koel Dutta Chowdhury

B.Sc., M.Sc.

A dissertation submitted in fulfilment of the requirements for the award of

Master of Science (M.Sc.)

to the



Dublin City University  
School of Computing

Supervisors:

Dr. Yvette Graham  
Prof. Alan Smeaton

July 2019

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Masters by Research is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Koel Dutta Chowdhury

ID No.: 17210135

Date: 3 July, 2019

# Contents

|   |             |
|---|-------------|
| <b>Abstract</b>                             | <b>viii</b> |
| <b>Acknowledgements</b>                     | <b>ix</b>   |
| <b>1 Introduction</b>                       | <b>1</b>    |
| <b>2 Background</b>                         | <b>5</b>    |
| 2.1 Neural Machine Translation . . . . .    | 6           |
| 2.1.1 Encoder-Decoder Framework . . . . .   | 8           |
| 2.2 Convolutional Neural Networks . . . . . | 11          |
| 2.2.1 The VGG Network . . . . .             | 14          |
| 2.2.2 Residual Network . . . . .            | 14          |
| 2.3 Multimodal Machine Learning . . . . .   | 15          |
| 2.3.1 Representation . . . . .              | 17          |
| 2.3.2 Translation . . . . .                 | 18          |
| 2.3.3 Alignment . . . . .                   | 19          |
| 2.4 Evaluation . . . . .                    | 20          |
| <b>3 Images in Aligned Data Condition</b>   | <b>22</b>   |
| 3.1 Overview . . . . .                      | 22          |
| 3.2 Dataset . . . . .                       | 24          |
| 3.2.1 Translated Multi30k . . . . .         | 25          |
| 3.2.2 MS-COCO . . . . .                     | 25          |
| 3.3 Experimental Settings . . . . .         | 26          |

|          |  |           |
|----------|--|-----------|
| 3.3.1    | Models . . . . .                             | 26        |
| 3.3.2    | Data Pre-processing . . . . .                | 29        |
| 3.3.3    | Hyper-parameter Setups . . . . .             | 29        |
| 3.4      | Results and Discussion . . . . .             | 31        |
| 3.5      | Summary . . . . .                            | 33        |
| <b>4</b> | <b>Images in Sparse Data Conditions</b>      | <b>35</b> |
| 4.1      | Overview . . . . .                           | 35        |
| 4.2      | Experimental Settings . . . . .              | 37        |
| 4.2.1    | Synthetic Data Generation . . . . .          | 38        |
| 4.2.2    | Resources . . . . .                          | 38        |
| 4.2.3    | Extraction of Image Features . . . . .       | 39        |
| 4.2.4    | MT Models . . . . .                          | 40        |
| 4.2.5    | Data Pre-processing . . . . .                | 42        |
| 4.2.6    | Model Hyperparameters . . . . .              | 42        |
| 4.3      | Results and Discussion . . . . .             | 44        |
| 4.4      | Summary . . . . .                            | 45        |
| <b>5</b> | <b>Images in Adversarial Data Conditions</b> | <b>47</b> |
| 5.1      | Overview . . . . .                           | 47        |
| 5.2      | Experimental Settings . . . . .              | 50        |
| 5.2.1    | Dataset . . . . .                            | 50        |
| 5.2.2    | Generating textual adversaries . . . . .     | 50        |
| 5.2.3    | Generating visual adversaries . . . . .      | 52        |
| 5.2.4    | Models . . . . .                             | 52        |
| 5.3      | Results and Discussion . . . . .             | 53        |
| 5.3.1    | Explicit Alignments . . . . .                | 57        |
| 5.4      | Summary . . . . .                            | 60        |
| <b>6</b> | <b>Conclusion</b>                            | <b>64</b> |



# List of Figures

|     |  |    |
|-----|--|----|
| 2.1 | Encoder Decoder Architecture . . . . .   | 10 |
| 2.2 | Architecture of a traditional convolutional neural network. . . . .  | 13 |
| 2.3 | Illustration of the VGG16 network architecture . . . . .   | 14 |
| 2.4 | Illustration of a residual connection, from (He et al., 2015). . . . .   | 15 |
| 4.1 | Flowchart of our Hi-En MMT System . . . . .  | 40 |
| 5.1 | An evaluation example of visual adversaries. The model sees a con-<br>gruent image (left) or an incongruent image (right). . . . . | 53 |
| 5.2 | Explicit alignment between of the <i>source</i> word <i>puppy</i> and its corre-<br>sponding visual category . . . . .             | 58 |
| 5.3 | <b>MMT+explicit</b> model with explicit alignment between the source<br>token and visual category . . . . .                        | 59 |

# List of Tables

|     |  |    |
|-----|--|----|
| 3.1 | Results for the M30k <sub>T</sub> 2017 English–German and English–French test sets. All models are trained on the original M30k <sub>T</sub> training data. Our ensemble uses four multi-modal models, all independently trained: two models IMG <sub>D</sub> , one model IMG <sub>E</sub> , and one model IMG <sub>2W</sub> . . . . . | 31 |
| 3.2 | Results for the MS-COCO 2017 English–German and English–French test sets. All models are trained on the original M30k <sub>T</sub> training data. Ensemble uses four multi-modal models, all trained independently: two models IMG <sub>D</sub> , one model IMG <sub>E</sub> , and one model IMG <sub>2W</sub> . . . . .               | 31 |
| 3.3 | Results for the best model of (Calixto et al., 2017a), which is pre-trained on the English–German WMT 2015 (Bojar et al., 2015), and different combinations of multi-modal models, all trained on the original M30k <sub>T</sub> training data only, evaluated on the M30k <sub>T</sub> 2016 test set. . . . .                         | 32 |
| 3.4 | Examples of translations produced by the multimodal systems (MMT) and a text-only model (NMT). Original: the original caption. <i>IMG<sub>D</sub></i> : the type of multimodal model output, given the textual input and the global image feature. NMT:the system output, given only the textual input. . . . .                        | 33 |
| 4.1 | Examples of manually translated captions of the Flickr30k English descriptions in Hindi using PBSMT system. . . . .  | 39 |

|     |  |    |
|-----|--|----|
| 4.2 | The overall statistics of the datasets used to train the PBSMT system. The 3rd row shows the amount of additional monolingual Hindi and English text used respectively for training the language model to create synthetic Hindi and the general-domain PBSMT system. . . .  | 43 |
| 4.3 | Evaluation metrics scores Hi-En translation systems before and after applying image features on manually translated dev data. Bold numbers indicate improvements that are statistically significant compared to NMT text with $p = 0.05$ . Evaluation is performed against the English translations of the test set using standard MT evaluation metrics, with BLEU and METEOR . . . . . | 44 |
| 4.4 | Illustrative example of translations produced by the multimodal systems (MMT) and a text-only model (NMT). Manual: the original translated caption in Hindi. NMT:the system output, given only the textual input. MMT: the multimodal system output, given the textual input and the global image feature. Reference:the gold standard reference in English. . . . .                     | 45 |
| 5.1 | Examples of adversarial textual samples that we use to attack the multimodal translation models. The <u>underlined text</u> denotes the words or phrases that are perturbed to create the adversarial example. . . .   | 51 |
| 5.2 | Examples of translations produced by the <b>hierattn</b> system (MMT) and the <b>MMT+explicit</b> model for Noun adversary. Both are the outputs, given the adversarial caption and the correct image. Original: the original caption. Noun: the adversarial caption with the <u>underlined</u> replacement. . . . .   | 54 |
| 5.3 | Behaviour of text-text translation model v/s text-image translation model in adversarial conditions. . . . .   | 54 |

|     |   |    |
|-----|---|----|
| 5.4 | Corpus-level Meteor scores for the English–German Multi30K Test 2017 data. Original: performance of systems evaluated on the original text and images. Textual: evaluation on the four different textual adversaries and the correct images. Visual: evaluation on the correct text but adversarial images. . . . .                               | 55 |
| 5.5 | Corpus-level Meteor scores for the English–German Multi30K Test 2017 data. Original: performance of the <b>hierattn</b> and the proposed <b>MMT+explicit</b> system evaluated on the original text and images. Textual: evaluation on the noun textual adversary and the correct images. . . . .  | 56 |
| 5.6 | Examples of translations produced by the <b>hierattn</b> multimodal translation system. <b>Baseline</b> : the output given the <b>Original</b> image-caption pair. NUM / PREP / NOUN / NP: The adversarial caption with the <u>underlined</u> replacement. MMT: the output of the <b>hierattn</b> system, given the adversarial sentence. . . . . | 63 |

# Neural Machine Translation for Multimodal Interaction

Koel Dutta Chowdhury

## Abstract

Typically it is seen that multimodal neural machine translation (MNMT) systems trained on a combination of visual and textual inputs produce better translations than systems trained using only textual inputs. The task of such systems can be decomposed into two sub-tasks: learning visually grounded representations from images and translation of the textual counterparts using those representations. In a multi-task learning framework, translations are generated from an attention-based encoder-decoder framework and grounded representations that are learned from pre-trained convolutional neural networks (CNNs) for classifying images.

In this thesis, I study different computational techniques to translate the meaning of sentences from one language into another considering the visual modality as a naturally occurring meaning representation bridging between languages. We examine the behaviour of state-of-the-art MNMT systems from the data perspective in order to understand the role of the both textual and visual inputs in such systems. We evaluate our models on the Multi30k, a large-scale multilingual multimodal dataset publicly available for machine learning research. Our results in the optimal and sparse data settings show that the differences in translation system performance are proportional to the amount of both visual and linguistic information whereas, in the adversarial condition the effect of the visual modality is rather small or negligible. The chapters of the thesis follow a progression starting with using different state-of-the-art MMT models for incorporating images in optimal data settings to creating synthetic image data under the low-resource scenario and extending to addition of adversarial perturbations to the textual input for evaluating the real contribution of images.

## Acknowledgments

First and foremost, my sincere gratitude to Qun Liu for his support throughout. My two supervisors Yvette Graham and Alan Smeaton provided valuable comments on various drafts of this thesis, which helped immensely. A quick thanks must go to Desmond Elliott for his contributions to my thesis. The work presented in Chapter 5 resulted from discussions with him. My most sincere thank you to all of you.

I am thankful to Christian Hardmeier (University of Edinburgh/Uppsala University) and Jennifer Foster (School of Computing, DCU) for examining this thesis and for their insightful comments on my work.

I thank Science Foundation of Ireland for the research grant (Grant 13/RC/2106) in the ADAPT Centre ([www.adaptcentre.ie](http://www.adaptcentre.ie)) to support this research.

Finally, no amount of thanks will be sufficient to express my gratitude to my amazing friends Piyush Arora, Rashmi Gupta and Daria Dzendzik for abstracting all the hardware out of my life.

# Chapter 1

## Introduction

While the holy grail of full natural language understanding still remains a distant dream in artificial intelligence, progress is being made in developing machine learning algorithms to comprehend what humans are talking or writing in natural language. Although humans can easily master a language naturally, it remains one of the most intricate tasks for a computer to be able to understand the ambiguity of human language. Natural language signals – spoken or written – are in constant evolution in line with the continuous nature of the visual world, making it difficult for a machine to perform reasonable inferences from these natural language utterances.

Previous studies in human cognition have well-established the relation between human language and underlying concepts and acquisition (Pulvermüller, 2005). The precise role of perceptual phenomena and sensory-motor signals in language acquisition and representation has been empirically shown to aid the learning process of human language learners (Bornstein et al., 2004; Landau et al., 1998). Taking inspiration from such cognitive and behavioral neuroscientific studies, various NLP downstream tasks study the problem of learning word meanings from small scale or synthetic multi-modal data. For example, acoustic signals are used to identify accents, allowing MT systems to effectively translate regionalisms and to adapt to demographic and geographical language specificity. Like speech, images are also harnessed to pair linguistic contexts with perceptual reality. This Masters project

aims to explore computational techniques to integrate image signals in order to improve NLP tasks – primarily those involving language generation, such as Machine Translation (MT). The chapters of the thesis follow a progression starting with using image in aligned and sparse data setting and arriving finally to disjoint settings in order to understand the actual role of visual modalities in multimodal machine translation (MMT) systems.

Chapter 1 introduces the topic and contributions of the thesis.

Chapter 2 discuss the related work and technical background detail.

Chapter 3 focuses on incorporating visual representations in an **aligned settings** with a combination of a Convolutional Neural Network to extract visual features and a Recurrent Neural Network to learn sentence embeddings.

Chapter 4 applies the techniques described in Chapter 3 to improve machine translation performance in **sparse data settings**, where no parallel data is available.

Chapter 5 extends the investigations of Chapter 3 and Chapter 4 to the **adversarial setup**, breaking the template that a multimodal system normally relies on to learn.

Chapter 6 outlines the findings and future research directions.

For MMT which is our focus, the use of images was first addressed in the form of a shared benchmark task (Specia et al., 2016a). Since that first introduction, there has been a great deal of work on MMT where the objective is to investigate if images can potentially help the task of translating an image description into a target language, given the description in a source language and the associated image as input. For example, an additional image can help to arrive at the correct translation for the word sense of “bank” as a financial institution compared to “bank” as the side of a river or the turning of an aeroplane. However, utilising images for linguistics units involves learning visually grounded representations (Kiros

et al., 2014b; Socher et al., 2014). Chapter 2 describes the general framework and background literature of learning visually-grounded sentence representations using different feature-extraction pipelines that allows machines to learn feature representations from *raw* input, which are more or less generally applicable for various downstream multimodal tasks. Chapter 3 describes the gains of using such visual representations. Our main interest and contribution here is the re-use of general techniques to learn linguistic representations using Recurrent Neural Networks and further develop ensemble models to contrast the results obtained using off-the-shelf text-only machine translation models.

The second contribution of the thesis is the use of visual signals for machine translation in a low-resource learning scenarios. Approaches in this direction are described in Chapter 4. We show that visual modality can be used as a pivot to find possible translations for words when there is no data available. More specifically, we generate synthetic aligned set in a low-resource language and provide an empirical evidence that the performance of machine translation on lower-resource languages can be improved by jointly training together with visual modality. Such an approach can be adopted to overcome the problems of data sparsity and more importantly, have practical implications for efficiently collecting image-captions in different languages to further address the problem of lexical ambiguity.

In addition to the recent development in using linguistic-visual multimodal representations for translation tasks, we further examine the benefit of using an additional modality for MMT in a broader sense. All existing previous work in MMT has mostly shown gains in terms of system performance by employing visual context over text-only NMT however, the inner-working of these systems still needs to be determined. In general, neural network (NN) models for language processing are often criticised for being uninterpretable black boxes (Benítez et al., 1997), namely that it is difficult to feed a trained NN model with an input and examine the network to determine why such input generates a particular output (Kádár et al., 2017). Addressing this issue is a non-trivial endeavour, even though studies on post-hoc interpretability has

enabled some insight into the inner-workings of NN models by examining specific examples or by learning grounded representations (Lipton, 2016). Thus, our last contribution in Chapter 5 explores the benefit of using visually grounded representations, but in an adversarial setting. Here we consider a scenario where the image and sentence datasets are not *aligned*. We generated adversarial captions using a set of heuristics to ensure structural similarity with respect to the original captions but contradictory semantics for testing the robustness of MMT systems. Furthermore, we also introduce a method of using the explicit alignments across modalities to strengthen MMT models against textual adversaries. We find that even though this technique requires an additional external data source, it consistently improves machine translation performance.

Lastly, in our final chapter, we not only briefly summarise our main findings, but also point to some of the limitations of our work and towards future directions. To improve performance and understanding of multimodal learning, we see several avenues for future research which are outlined.

Each of the following Chapters has been previously published. They are included in this thesis with modifications of re-aligning and re-sizing a few figures and adding some more illustrative examples.

**Chapter 3** has been published as Calixto, I., Dutta Chowdhury, K, Liu, Q (2017).

DCU System Report on the WMT 2017 multimodal Machine Translation Task.  
WMT 2017, (**EMNLP 2017**), Copenhagen, Denmark.

**Chapter 4** has been published as Dutta Chowdhury, K., Hasanuzzaman, M., and Liu, Q (2018). Multimodal Neural Machine Translation for Low-resource Language Pairs using Synthetic Data. In Proceedings of the **ACL 2018** workshop on deep learning approaches for low resource natural language processing, Melbourne, Australia.

At the time of completing the thesis, part of the work in **Chapter 5** has been submitted to a conference and is *under-review*.

# Chapter 2

## Background

In this chapter, we provide some background knowledge on two key research areas, namely Machine Translation and Computer Vision, which are pertinent to our research. After having mentioned several concepts which fall under these two areas, such as the encoder-decoder framework and VGG/ResNet, a special focus will be placed on Multimodal Machine Learning that falls within the scope of this project.

The first concrete idea of using machines to perform the translation process started with the inaugural work on *mechanical dictionaries* (Hutchins, 2004). Subsequently, Georges Artsrouni designed a machine to store source language lexicons and their corresponding translations in several languages on a memory band that could be retrieved afterwards (Daumas, 1965). Petr Trojanskij proposed a multilingual translation system that take advantage of the abstract Esperanto-based symbols to encode grammatical functions between languages (Daumas, 1965). This work laid the basic foundations for *interlingual representations* or as popularly termed *–interlingua–* in translation. This sets out to use an abstract intermediary layer between two natural languages to be translated in order to make the translation process more generalized across different language combinations. Later in the 1950’s and 1960’s rule-based Machine Translation emerged as an active field of research. Some of the notable works include *direct translation*.– the task of translation between a specific source and target language using rules with only small amount of structural

analysis. King and Wieselmann (1956)’s seminal work to handle lexical ambiguities by applying word-selection-rules based on the surrounding context words is an example of this approach. Another popular approach is the *interlingua-based* approach that solves the problem of having to write new rules for every possible language pair. This is achieved by generalising the source language input to a generic interlingua representation that is language independent. This can then be translated via rules into the target language. This can also be seen as the bridge between the brute-force nature of the direct translation models and the more instructive interlingua models that handles the lexical ambiguity in the source sentence by pre-analysing its syntax and semantics structure before finally translating it into the target language (Hutchins, 2007). However, owing to the complexity of such process needed to transfer a phrase into an interlingua and back, they are highly susceptible to increasing the number of errors.

However the concept of encoding into and decoding from an *interlingua* did not receive its due until the advent of modern multilingual NMT systems in the form of multilingual representations. A major change in MT research came into effect in 1990’s when a group of researchers at IBM developed a statistical machine translation model, popularly known as the IBM model (Brown et al., 1990). Unlike rule-based MT, these models did not come attached with immense amount of hand-coded rules creation and essentially consisted of only two modules – *language model* and *translation model*. Following this, Koehn et al. (2007a) developed a phrase-based SMT (PBSMT) systems that performed phrase-based modeling.

## 2.1 Neural Machine Translation

As neural network based components were introduced to SMT, the early 2010’s saw the emergence of the first Neural Machine Translation (NMT) systems consisting of jointly-trained neural networks capable of taking a source sentence as input and translating it into a target sentence. The first modern approach to neural language

models (NLMs), was first proposed by Bengio et al. (2003) and that was followed by other notable studies such as Morin and Bengio (2005); Mnih and Hinton (2009); Mnih and Teh (2012) that laid down the foundation stones for current neural architectures. Their work presents a feed-forward multilayer perceptron with continuous word-embeddings, a single hidden layer and a softmax output layer. Given the previous fixed number of words as context over a training corpus –  $n$ -gram language model – the model is trained with stochastic gradient descent (Cauchy, 1847) through the backpropagation algorithm (Rumelhart et al., 1985) to maximize the probability of the target word. The function of the aforementioned was to combat the “curse” of dimensionality in language modeling. As a natural development, subsequent MT systems (Schwenk, 2007; Vaswani et al., 2013; Luong et al., 2015) started adopting  $n$ -gram language models alongside traditional  $n$ -gram LMs and generally obtained improvements in terms of system performance. This inspired the birth of the Neural Machine Translation (NMT) framework with the goal of redesigning the entire MT pipeline in a way such that different modules i.e, translation models, language models, and reordering models, require no separate tuning modules.

Neural machine translation is essentially a recurrent language model that conditions on the source language sentence. To put this formally, NMT aims to directly model the conditional probability  $p(y|x)$  in order to translate a sentence in a source language,  $x_1, \dots, x_n$  into a sentence in a target language,  $y_1, \dots, y_m$ . It does so using the *sequence-to-sequence* framework also known as *encoder-decoder* (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Cho et al., 2014c).

At a high level, NMT models consist of two recurrent language models, i.e. the *encoder* and the *decoder*. In the *encoding* step, the RNN simply computes a variable length representation  $\mathbf{s}$  for each source sentence without making any prediction. Subsequently, based on that source representation, the target sentence predicts the next words in the *decoding* step. The training objective for NMT is formulated as:

$$J = \sum_{(x,y) \in \mathbb{D}} -\log p(y|x) \quad (2.1)$$

where  $\mathbb{D}$  refers to the parallel training corpus of source and target sentence pairs  $(x, y)$ .

### 2.1.1 Encoder-Decoder Framework

The encoder neural network encodes a sequence of  $N$  tokens  $x_{1...n}$  in the source language with recurrent neural networks (RNN) into a fixed-length vector representation. The RNN hidden state is updated at each time step (i.e, for each element of the sequence) and, consequently, the output of the final hidden state contains information about the whole sequence. Formally, the encoder RNN implemented as a bi-directional RNN transforms the sentence to a sequence of annotations. Each token in the source language input sequence can be represented by a concatenation of the forward and backward hidden state vectors:

$$\mathbf{h}_i = [\vec{\mathbf{h}}_i; \overleftarrow{\mathbf{h}}_i] \quad (2.2)$$

where  $\mathbf{h}_i$  is the annotation of token  $x_i$ .

Based on the source annotations, the decoder recurrent neural network (Bahdanau et al., 2015) outputs a sequence of words in the target language. Tokens in the decoder are represented by a one-hot vector  $\mathbf{y}_k$ , which is mapped into an embedding  $\mathbf{e}_k$  through a learned matrix  $\mathbf{E}_k$ :

$$\mathbf{e}_k = \mathbf{y}_k \cdot \mathbf{E}_y \quad (2.3)$$

The use of fixed-length context vector is a bottleneck in the encoder-decoder framework. To improve upon this model Bahdanau et al. (2015) introduced an attention mechanism, which lets the decoder learn to focus over a specific range of the input sequence. It allows the decoder to have access to the previously predicted token  $\mathbf{y}_{k-1}$ , the previous decoder state  $\mathbf{d}_{k-1}$ , and a context vector  $\mathbf{c}_k$  calculated over the

encoder hidden states at each time-step  $t$ :

$$\mathbf{d}_{\mathbf{k}} = \text{RNN}(\mathbf{d}_{\mathbf{k}-1}, \mathbf{y}_{\mathbf{k}-1}, \mathbf{e}_{\mathbf{k}}) \quad (2.4)$$

To initialize the decoder RNN  $\mathbf{d}_1$ , a nonlinear transform of the mean of the encoder states with learned parameter  $\mathbf{W}_{init}$  is used:

$$\mathbf{d}_1 = \tanh(\mathbf{W}_{init} \cdot \frac{1}{N} \sum_i^N \mathbf{h}_i) \quad (2.5)$$

The context vector  $c_k$  is a weighted sum over the encoder hidden states, where  $N$  denotes the length of the source sentence and each vector is weighted by the attention weight  $\alpha_{ki}$ :

$$\mathbf{c}_{\mathbf{k}} = \sum_{i=1}^N \alpha_{ki} \mathbf{h}_i \quad (2.6)$$

The  $\alpha_{ki}$  is the normalised alignment matrix between each encoder hidden state vectors  $\mathbf{h}_{1...n}$  and the decoder hidden state while producing the  $k_{th}$  token in the translation. They are computed by a feed-forward neural network, where  $\mathbf{v}_{\mathbf{a}}$ ,  $\mathbf{W}_{\mathbf{a}}$  and  $\mathbf{U}_{\mathbf{a}}$  are learned parameters:

$$\alpha_{ki} = \frac{\exp(e_{ki})}{\sum_{j=1}^N \exp(e_{ji})} \quad (2.7)$$

$$e_{ki} = \mathbf{v}_{\mathbf{a}} \tanh(\mathbf{W}_{\mathbf{a}} \mathbf{d}_{\mathbf{k}-1} + \mathbf{U}_{\mathbf{a}} \mathbf{h}_i) \quad (2.8)$$

From the hidden state  $\mathbf{d}_{\mathbf{k}}$  the network predicts the conditional distribution of the next token  $y_k$ , given a target language embedding  $\mathbf{e}_{\mathbf{k}-1}$  of the previous token, the current hidden state  $\mathbf{d}_{\mathbf{k}}$ , and the calculated context vector  $\mathbf{c}_{\mathbf{k}}$ .

The final decoding is passed through a softmax layer to predict the probability of the next word in the sequence over the output vocabulary. Figure. 2.1 shows an

Encoder-Decoder Architectures <sup>1</sup>.

$$p(y_k|y_{<k}, c) = \text{softmax}(\tanh(\mathbf{e}_{\mathbf{k}-1} + \mathbf{d}_{\mathbf{k}} + \mathbf{c}_{\mathbf{k}})) \quad (2.9)$$

The translation model is trained to minimise the negative log likelihood of predicting the target language output:

$$J_{NLL}(\theta, \phi^t) = - \sum_k \log p(y_k|y_{<k}, x) \quad (2.10)$$

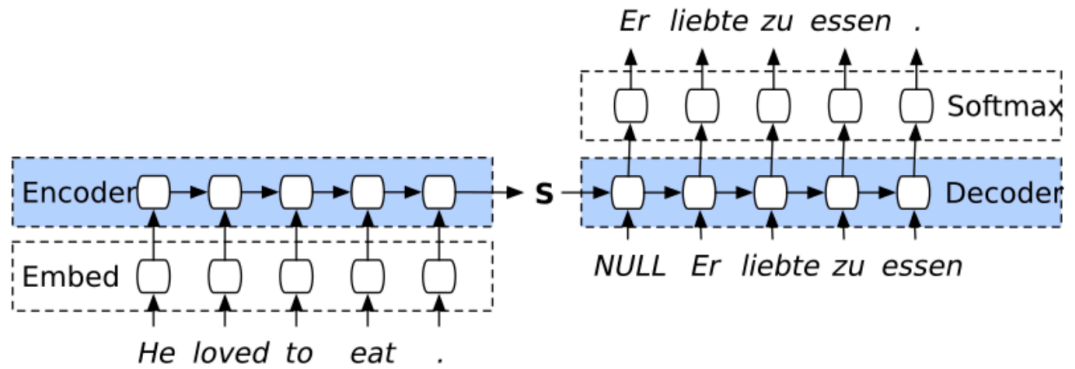


Figure 2.1: **Neural machine translation** – example of an Encoder Decoder architecture for translating a source sentence “He loved to eat” into a target sentence “Er liebte zu essen”.

As is common to most machine learning methods, considerable effort has been made to train more complex networks such as long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997; Gers et al., 1999) and gated recurrent unit networks (GRU) (Cho et al., 2014b) recurrent network variants which has shown to outperform traditional Elman networks in practice. We opted for GRUs in Chapters 3, 4 and 5.

**Gated Recurrent Neural Networks:** Gated Recurrent Unit networks introduce a particular *memory structure*, which adds an inductive bias to the network that facilitates the retention of information over multiple time steps. Intuitively GRUs

<sup>1</sup>Image source-<https://towardsdatascience.com/nlp-sequence-to-sequence-networks-part-2-seq2seq-model-encoderdecoder-model-6c22e29fd7e1>

can be seen as a sequential computer with soft-continuous read-write memory operations: the reset-gate  $\mathbf{r}_t$  decides how much of each component of the previous state is relevant to be mixed in with the current input, resulting in the current candidate memory state  $\tilde{\mathbf{h}}_t$ . The output-gate then overwrites the previous state with the current candidate.

$$\mathbf{z}_t = \sigma(\mathbf{W}_z \mathbf{w}_t + \mathbf{U}_z \mathbf{h}_{t-1} + \mathbf{b}_z) \quad (\text{update-gate})$$

$$\mathbf{r}_t = \sigma(\mathbf{W}_r \mathbf{w}_t + \mathbf{U}_r \mathbf{h}_{t-1} + \mathbf{b}_r) \quad (\text{reset-gate})$$

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}_h \mathbf{w}_t + \mathbf{r}_t \odot \mathbf{U}_h \mathbf{h}_{t-1}) \quad (\text{memory content})$$

$$\mathbf{h}_t = (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} \mathbf{z}_t \odot \tilde{\mathbf{h}}_t \quad (\text{hidden state})$$

As the attention mechanism did exceedingly well in terms of performance of RNN-based NMT systems, an alternative model came about that does away with RNNs completely by replacing them with several stacks of self-attention over the source and target sequence. This is known as *transformer*, whose architecture is best described in the original paper by Vaswani et al. (2017).

## 2.2 Convolutional Neural Networks

Convolutional Neural Network (CNNs) are multi-layered perceptrons inspired by the biological processes in the connectivity pattern neurons of an animal’s visual cortex to analyse visual imagery. These networks learn a hierarchy of blocks of image filters with learnable weights, receptive fields and pre-defined pooling operations for dimensionality reduction (LeCun et al., 1995). Contrary to the simple feed-forward network, CNNs consist of a set of distinctive layers such as *pooling* and *convolutional*. Although CNNs initially were only used to various computer vision tasks (Donahue et al., 2014; Girshick et al., 2014; Oquab et al., 2014), they are now successfully applied to many different tasks (Kalchbrenner and Blunsom, 2013)

where a convolutional neural network (CNN) is used as an encoder with an additional recurrent neural network (RNN) layer as the decoder.

We follow Gu et al. (2018) to describe the components that constitutes a CNN architecture. A convolutional layer is made up of a set of kernels or filters (these terms are used interchangeably) that are used to produce different feature maps. Each of these filters has its own set of weights that are shared, that is to say, each filter is required to apply the same operations in different parts of an image, when computing a feature map. One complete convolutional layer is made up different feature maps computed using a bank of  $n_1$  filters. Each filter detects a particular feature value at every location on the input. The output  $Y_i^l$  of layer  $l$  consists of  $n_1^l$  feature maps of size  $n_2^l \times n_3^l$ . The  $i$ -th feature map, denoted  $Y_i^l$ , is computed as in Eq.2.11:

$$\mathbf{Y}_i^l = \mathbf{B}_{i,j}^{(l)} + \sum_{j=1}^{m_1} \mathbf{K}_{i,j}^{(l)} \times \mathbf{Y}_j^{(l-1)} \quad (2.11)$$

where  $\mathbf{B}_{i,j}^{(l)}$  is the bias and  $\mathbf{K}_{i,j}^{(l)}$  denotes the filter that connects the  $j$ -th feature map in layer  $l - 1$  with  $i$ -th feature map in layer  $l$ .

In order to be able to compute non-linear features over the inputs, the feature map generated by the convolutional layer is further passed through an activation function  $g(\cdot)$ , as in Eq.2.12:

$$\mathbf{Y}_i^l = g_i f(Y_j^{(l-1)}) \quad (2.12)$$

The output is then passed to the next layer in the CNN. Previous studies suggest that rectified linear units (ReLUs) are more useful than the traditional activation functions in building CNN blocks (Nair and Hinton, 2010).

After multiple stages of convolutional and non-linearity layers, the network pro-

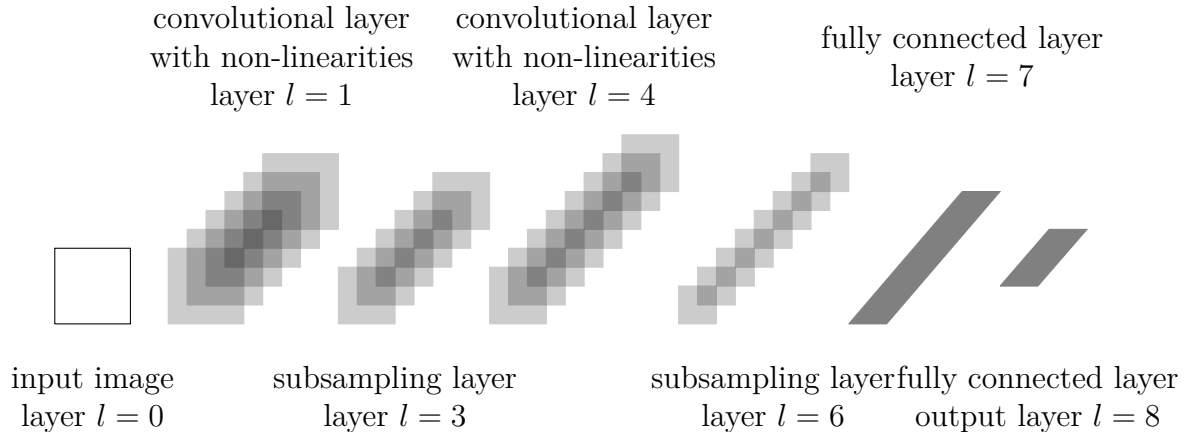


Figure 2.2: Architecture of a traditional convolutional neural network.

gresses through the downsampling layer, often termed as the *pooling layer* to reduce the spatial size of the activation maps. Typically, this layer consists of two parameters: a) the spatial extent of the filter  $F^{(l)}$  and b) the stride  $S^{(l)}$ . The primary function of the pooling layer is to provide translational invariance shifting to the network (LeCun et al., 1998). Since in most image recognition tasks, feature detection is more important than the feature’s exact location, the pooling operation aims to preserve the detected features in a low dimension by reducing the feature map size at the cost of spatial resolution. Finally, the activation maps from the combination of previous different layers are then mapped into a class probability distribution through the fully connected  $FC$  layer.  $FC$  layers are multilayer perceptrons that have full connections to all of the activations in the previous layer.

Figure 2.2 shows the original CNN architecture of LeCun et al. (1995) alternates between convolutional layers including hyperbolic tangent non-linearities and subsampling layers. In this illustration, the convolutional layers already include non-linearities and, thus, a convolutional layer actually represents two layers. The feature maps of the final subsampling layer are then fed into the actual classifier consisting of an arbitrary number of fully connected layers. The output layer usually uses softmax activation functions.

We now briefly discuss the two CNN architectures that we have used in our work namely the VGG networks (Simonyan and Zisserman, 2014) and the Residual

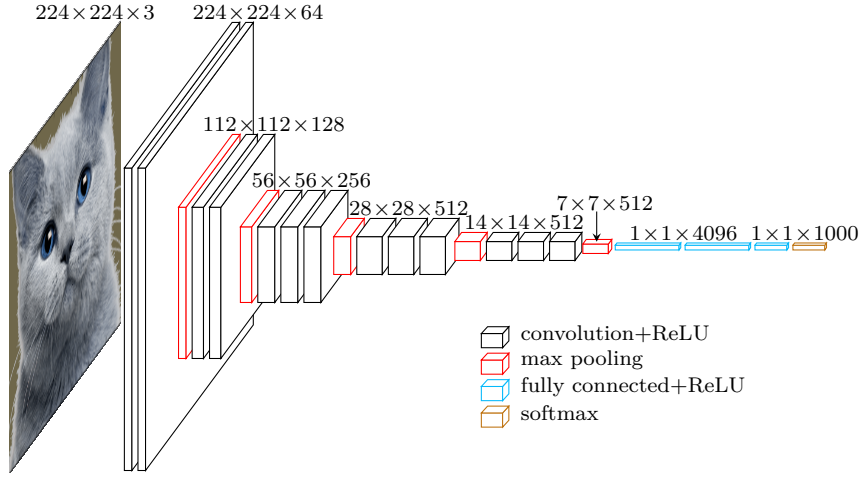


Figure 2.3: Illustration of the VGG16 network architecture

Network (He et al., 2015).

### 2.2.1 The VGG Network

The VGG Network Architecture was originally proposed by Simonyan and Zisserman (2014) at Oxford University Vision group that followed a simple recipe to preserve the size of the image’s width and height through the convolution operations, i.e. only  $2 \times 2$  paddings of stride 2 and  $3 \times 3$  convolutions of stride 1 with a padding of size 1. This network is known for its linear architecture and the multiple stacked smaller size kernel layout makes it easier for the network to learn finer level properties of an image.

### 2.2.2 Residual Network

Unlike traditional sequential network architectures, such as VGG, the Residual Network relies on the “Network-In-Network Architecture” module. At first introduced by He et al. (2015) this network has accomplished state of the art results on a number of popular training image datasets such as CIFAR and MNIST. These networks address the *degradation* problem causing from naively adding more layers to a model. Contrary to the traditional convolution architectures, these networks try to fit a tiny modification of the input ( $y = x + F(x)$ ), instead of a total modification

( $y = H(x)$ ) of its input  $x$ — hence the term residual. This is illustrated in the following Figure 2.4. These networks can typically be seen as a “shortcut” established with the identity function to shorten the path that gradients traverse between a network’s output and input layers during back-propagation step.

In published work, He et al. (2015) released pre-trained versions of three networks, referred to the ResNet-50, ResNet-101 and ResNet-152 networks with different quantity of layers and residual connections in each of them. To put simply, the ResNet-101 and ResNet-152 operate over the same layer architecture as the ResNet-50, but have more building blocks in the final network architecture. In our work, we only use the ResNet-50 network (He et al., 2015) to extract image features. We specifically use the RES4FLAYER of the ResNet-50 network to extract the local features, consisting of a  $<14 \times 14 \times 1,024>$  3-tensor.

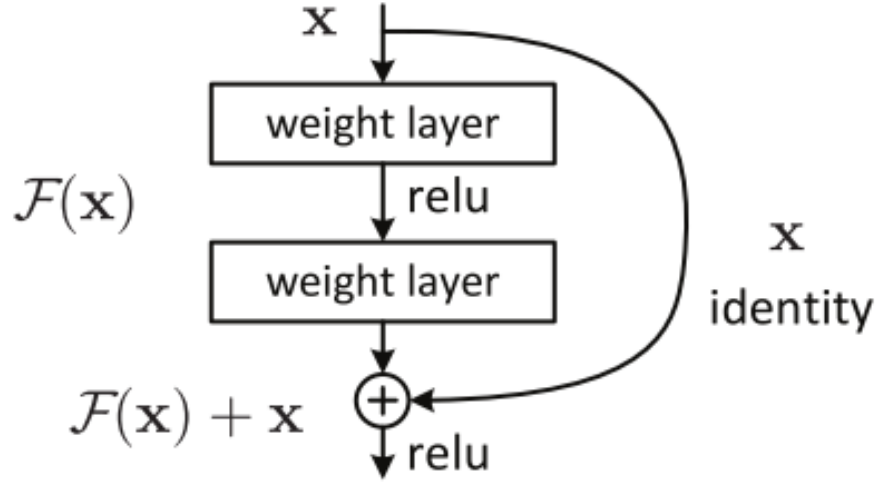


Figure 2.4: Illustration of a residual connection, from (He et al., 2015).

## 2.3 Multimodal Machine Learning

Multimodal machine learning has been receiving much attention lately, especially since the *heterogeneous modality* processing portion of such tasks have been considerably improved with the advent of new deep-learning models. Learning from different sources enhances the possibility of capturing correspondences across modalities and

gaining an in-depth understanding of natural phenomena. Following the taxonomy proposed by Baltrušaitis et al. (2019), there are five core technical factors that constitute multimodal machine learning:

1. **Representation:** combining feature vectors from cross-modalities into a joint representation in order to take advantage of their divergence, similarity and redundancy.
2. **Translation:** transforming one modality into another. For example, if there are two translation outputs of the same source then one way to know which is correct is by looking at the image.
3. **Alignment:** establishing cognitive coalition across different modalities. Identifying *concepts* from an image to their specific mention on the caption is an example.
4. **Fusion:** integrating information from two or more modalities to predict an outcome measure. For example, in multimodal sentiment analysis, visual context from facial expressions is often combined with speech signal information to predict the polarity of a given utterance.
5. **Co-learning:** transferring knowledge between tasks, modalities and their representations. This is particularly beneficial in sparse resource settings.

Each of these is a challenging problem in Natural Language Processing and demands a thorough survey of multimodal machine learning. For my thesis, I have specifically focused on the literature survey of *representation*, *translation* and *alignment*. As we aim to apply visual information to solving machine translation tasks, we need to interpret both modalities (text and image), hence *representation*, and convert them into the target language, hence *translation*. Additionally, in the last chapter of the thesis, we want to explore the textual-image alignment i.e, *alignment*.

### 2.3.1 Representation

The conceptual understanding of how best to represent modalities in a way that best captures underlying semantics as well as other contextual information is rapidly evolving. Typically, the distributional semantic models (DSMs) that compute the word representations from documents based on word co-occurrence patterns fail to capture contextualised representations both at the word-level, e.g. word-2-vec (Mikolov et al., 2013) and sentence-level, e.g. skip-thought vectors (Kiros et al., 2015). These models are not equipped to handle additional extra-linguistic modalities (Harnad, 1990; Glenberg and Robertson, 2000), thus not allowing the network to use morphological clues to form robust representations. Multi-modal distributional semantic models (DSMs) are an extension of the traditional DSMs that are able to handle this *grounding problem* (Glenberg and Robertson, 2000) by taking additional modalities into account. These are vector spaces to which one maps not only text but also different kinds of multimodal information (e.g., images, audio, speech). Typically, popular approaches of learning a joint space for multimodal interactions are based on Canonical Correlation Analysis (CCA) (Gong et al., 2014) which uses deep Convolutional Neural Networks (Donahue et al., 2014) to obtain the image features and bag-of-words features for the textual counterpart. More recently, a few works on multimodal embedding training procedures follows a max-margin objective function that is based upon a margin ranking criterion (Cohen et al., 1998) to perform a pairwise ranking between modalities. The idea is to learn a visual-semantic or multimodal embedding space of image descriptions and representations by optimising a pairwise ranking loss function (Socher et al., 2014; Kiros et al., 2014b; Karpathy and Fei-Fei, 2015; Vendrov et al., 2015; Gella et al., 2017). To that end, a pre-trained CNN is fine-tuned as an image feature extractor, followed by a learned transformation, while sentence representations are normally learned by a randomly initialized recurrent neural network.

Such joint representations aim at minimising the distance of individual unimodal representation in the coordinate space. For example: these models encourage the

representation of the word *cat* and an image of a cat to maintain a smaller distance between them than the distance between the word *cat* and *an image of a dog* (Frome et al., 2013).

### 2.3.2 Translation

Translation of one modality to another is a long-studied problem in multimodal machine learning literature. Given a candidate in one modality typically the task is to translate (transform) the same candidate in a different modality. An example of one such task is Image Captioning, where the goal is to generate a textual description for a given image. Recently, multimodal translation has gained huge attention as an active area of research for several downstream applications such as Bernardi et al. (2016); Torabi et al. (2015); Vinyals et al. (2014); Yagcioglu et al. (2015); Venugopalan et al. (2014) and also in the context of multimodal machine translation Elliott et al. (2017); Caglayan et al. (2017); Libovický and Helcl (2017) i.e., our area of interest.

Multimodal Machine Translation is a research topic that was addressed by the MT community in the form of a shared task (Specia et al., 2016b), where the goal is to translate the text with the help of an accompanying modality such as an image or video. The main idea is to use secondary information to improve the translation of ambiguous terms. There upon, various methodologies have been proposed to compare text-only and multimodal varieties of the same underlying MT framework.

Calixto et al. (2012) studied how visual information can be helpful in disambiguating machine translation outputs. Soon afterwards Hitschler et al. (2016) proposed a model that uses image features for re-ranking translations of image descriptions produced by SMT models. In the context of multimodal NMT, Huang et al. (2016) introduced a model to incorporate both local and global visual features extracted through the VGG19 network (Simonyan and Zisserman, 2014). Libovický and Helcl (2017) proposed a decoder network that learns to selectively attend to a combination of the source language and the visual data. Calixto et al. (2017a)

used a separate visual attention mechanism to incorporate spatial visual context into NMT. Other important works by Specia et al. (2016b); Caglayan et al. (2017); Elliott et al. (2017) claimed that systems trained on a combination of visual and textual inputs produce better translations than systems trained using only textual inputs. Luong et al. (2016) proposed multi-task learning settings to add neural image description as an auxiliary task to sequence-to-sequence NMT (Bahdanau et al., 2015) and reported significant improved translations in the parent translation task. However, despite all this recently reported work, the role of image in translation remains an open question at the time of writing this thesis. The work of Grönroos et al. (2018) showed that the effect of the visual modality in multimodal translation is small. In their analysis they attributed their largest gains to using additional (unconstrained) text data. More recently, Elliott (2018)’s work on evaluating systems with randomly selected visual adversaries also indicates that visual modality in the multimodal machine translation (MMT) is either unnecessary or only marginally beneficial. Similar problems in the visual question-answering domain resulted in the construction of a balanced dataset for the visual question answering domain (Zhang et al., 2016; Goyal et al., 2017).

### 2.3.3 Alignment

Multimodal alignment can be seen as the cognitive *coalition* on the level of precise concepts between two or more modalities. Alignments of cross-modalities and their relationship to model robustness and generalisation has received considerable attention in the last few years in the NLP and Computer Vision (CV) community (Karpathy and Fei-Fei, 2015; Simonyan and Zisserman, 2015; Socher et al., 2014). Proposed approaches include both supervised and unsupervised measures for aligning modalities. It is seen that latent alignment of the data during model training helps in boosting neural model performance. In the multi-modal domain, Karpathy and Fei-Fei (2015) proposed a method for aligning images with captions for cross-modal retrieval tasks. Other works include Zhu et al. (2015) who proposed a

method to aligned books with their corresponding movies-scripts by training a CNN to measure similarities between scenes and text. In a similar vein, Mao et al. (2016) used a Long Short Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997) and a CNN visual one to assess the quality between a referring expression and an image object. The intuition is that the joint semantics of both visual and textual modalities during the training procedure can provide rich supervision to learning systems (Plummer et al., 2015).

For example, in the case of multimodal machine translation systems, when there are two translation outputs of the same texts the only way to know the correct translation is by looking at the second modality e.g, an image. However, the sparse availability of annotated aligned datasets in the language and vision communities makes it difficult to extract latent alignment between modalities. Grounded representations of sentences that are learned from image-caption datasets also improves performance of sentence-level tasks when used as additional features (Kiela et al., 2017; Yoo et al., 2017; Kádár et al., 2017) to skip-thought vectors (Kiros et al., 2015). The basic idea is to keep intact the alignment of the two modalities by pulling the image-caption pairs close together and pushing the false image-caption pairs further from each other in the learned embedding space.

In addition to learning grounded representations for multimodal machine-translation, joint vision and language systems have been also applied to a wide range of tasks such as image captioning (Mao et al., 2014; Vinyals et al., 2015; Xu et al., 2015), visual question answering (Antol et al., 2015; Fukui et al., 2016; Jabri et al., 2016) and text-to-image synthesis (Reed et al., 2016).

## 2.4 Evaluation

This apart, one prime area for multimodal machine learning is that of *evaluation*. Evaluation for multimodal machine learning is particularly a subjective task which is mostly performed through the use of human assessment. Like, for speech synthesis,

naturalness and mean opinion score are considered (Zen et al., 2012; Van Den Oord et al., 2016) while realism is a measure of fit for visual data synthesis (Taylor et al., 2012).

Agreement between human assessors of translation quality is a well-established problem in the literature of MT evaluation. In addition to the standard relative ranking (RR) manual evaluation, the evaluation of the target translation are also carried out using human assessment of *adequacy* (i.e, how well the source is expressed in the target translation) and *fluency* (which determines to what extent the translation is a well-formed utterance in the target language and fluent in context). Popular methodology for crowd-sourcing human assessments of translation quality involves restructuring the task into a monolingual assessment (Graham et al., 2013, 2014).

Although human assessment are are a gold standard for evaluation, a number a automated metrics such as BLEU (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), ROUGE (Lin and Hovy, 2003), and Meteor (Denkowski and Lavie, 2014) are used as well to compute similarity scores between the generated and ground truth text.

# Chapter 3

## Images in Aligned Data Condition

### 3.1 Overview

It is assumed that Neural Machine Translation (NMT) models with additional image features are better at translating *visual terms* – words or phrases that have a direct correspondence in the image – than they are at text-only translation systems. As an illustrative example, we can consider the following two sentences:

- (1) a. I had to take out a bank loan to start my own business.
- b. By the time we reached the opposite bank, the boat was sinking fast.

Both sentences contain the word “bank”, but the meaning of the word differs entirely between them. This linguistic phenomenon, where two identical words change meaning depending on the context, is known as “polysemy” (Apresjan, 1974). Most machine translation (MT) frameworks struggle at handling polysemy because they use a single vector to represent polysemous words such as “bank”, regardless of the context in which the word is used. The problem of lexical ambiguity - remains an active area of research at the time of writing this thesis: with recent works in pre-training contextual representations (Dai and Le, 2015; Peters et al., 2018; Radford et al., 2018; Howard and Ruder, 2018; Devlin et al., 2018)— achieving state-of-the-art results in a wide variety of NLP tasks, including Natural Language Inference

(MNLI), Question Answering and others. More recently, the work of Rahman et al. (2019) demonstrates incorporating non-textual information within the input space of such networks for modeling multimodal language.

In the area of information retrieval, this problem has been recognised for several decades as it causes a mis-match when users search for “bank” meaning the financial institution but are presented with retrieved documents about river banks. Within IR, several statistical approaches to resolving or working with word senses in the indexing and retrieval tasks, have been identified (Sanderson, 1994; Kilgarrieff, 1997) though word sense disambiguation has not made it into mainstream text retrieval such as in web search, for example.

Many theories of human cognition supported by empirical evidence state that multimodal representation learning is largely driven by evidence of perceptual grounding in human concept acquisition and representation (Barsalou et al., 2003). It has been shown that visual sensory input plays a pivotal role in language acquisition by grounding meanings of words and phrases in perception. Needless to say, perceptual grounding and multimodal representations have their separate benefits, however, they can be beneficial to each other as well Kádár et al. (2017). The study of multimodal representation learning has a much briefer history than sentence-level representations and Chapter 2 situates the reader in the area.

For Multimodal Machine Translation, recent studies have shown that ambiguous and polysemous terms in principle could be disambiguated using an image as additional context (Calixto et al., 2017b; Caglayan et al., 2017; Libovický and Helcl, 2017; Elliott and Kádár, 2017) instead of using the surrounding text. In that sense, an image can act as a natural fence to deal with the relatedness in languages. One of the intriguing aspects of using *images* as a naturally occurring meaning representation is that images are also naturally universal across languages. This means that learning cross-modal representations allows a single model to transfer knowledge between modalities, thereby mitigating the low-resource problem common in cross-lingual applications and processing data (Gella et al., 2017; Kiros et al., 2014a;

Calixto et al., 2012).

In this chapter, we study how to incorporate image features to improve translation quality of an MT system when no form of data augmentation is required. To be specific, we focus on translation based on a specific data configuration, where the same images are annotated with multiple languages. To assess the impact of the existence of this *image-text* alignment has on the performance of translation, we formulate the following research question:

*Given that there is a large aligned data triple <image-source-target>, can multi-modal model offer more benefit compared to text-only MT models ?*

That is, for a given data triple (*image, source text, target text*), we examine if the model takes cues from the aligned image in order to translate the source language into the target language and to perform this translation with better quality. The approaches presented in the chapter extract global image and sentence features through separate encoders and learn to associate them. The textual portion is first encoded with a bidirectional recurrent network (Cho et al., 2014a) while the global image region features are extracted from a pre-trained convolutional neural network (Simonyan and Zisserman, 2015), presented in Section 3.3. These representations are further fused following different strategies described in Section 3.3.1 and further evaluated against one or more automatic evaluation metrics. Furthermore, we provide evidence that the performance of textual models can be improved by training jointly with additional images.

## 3.2 Dataset

To begin to outline this work, we first define the data configuration we used to address the above-mentioned research question. We need corpora that are both multilingual and multi-modal; multilingual since in training and evaluation, we need textual input in at least more than one language; multi-modal because we need

images associated with this text so as to evaluate how can we improve translation quality by exploiting them. We use the *translation* portion of the Multi30K dataset (Elliott et al., 2016, 2017), which is currently the largest collection of images paired with sentences in multiple languages used for solving multimodal machine translation tasks.

### 3.2.1 Translated Multi30k

We evaluate our model using the translation portion of the benchmark Multi30K dataset (Elliott et al., 2016). This dataset contains 31,014 images paired with an English language sentence and corresponding German, French, and Czech language translations. 29,000 instances are reserved for training, 1,014 for development, and 1,000 for evaluation.<sup>1</sup> For the experiments we only use the English-German-French portion of the dataset. The English and German sentences are pre-processed by normalising the punctuation, lowercasing and tokenising the text using the Moses toolkit. Additionally, the German text was de-compounded using the unsupervised German compound splitter (Daiber et al., 2015).

### 3.2.2 MS-COCO

We also use two out-of-domain datasets to evaluate our model namely the MS-COCO dataset of English described images (Chen et al., 2015), and the English-German WMT-2015 parallel corpus (Bojar et al., 2015). When we perform experiments with the WMT2015 corpus, we first calculate a 17,597 sub-word vocabulary using SentencePiece (Schuster and Nakajima, 2012) over the concatenation of the Multi30K and WMT15 datasets. This gives us a shared vocabulary for the external data that reduces the number of out-of-vocabulary tokens.

---

<sup>1</sup>The Multi30K dataset also contains 155K independently collected descriptions in German and English. In order to make our experiments more comparable with previous work, we do not make use of this data.

### 3.3 Experimental Settings

For our experiments, we use models which can essentially be thought of as extensions of the vanilla NMT attentive model by Bahdanau et al. (2015) with additional visual context extracted from images. We follow Calixto et al. (2017b) to use a bi-directional recurrent neural network (RNN) with a gated mechanism (Cho et al., 2014a) as the *encoder*, while the concatenation of forward and backward hidden states,  $h_i = [\vec{h}_i, \overleftarrow{h}_i]$  were used as the final annotation vector for a given source position  $i$ .

For global image feature extraction, we use a publicly available pre-trained model VGG19-CNN network (Simonyan and Zisserman, 2014). This model is trained on a subset of the ImageNet database,<sup>2</sup> which is trained to classify images into one out of 1,000 Imagenet classes (Russakovsky et al., 2015). Typically, these features are the 4096D activations of the penultimate fully connected layer FC7, (henceforth referred to as  $\mathbf{q}$ ). We use three different methods to include this visual information into the NMT pipeline:

1. using an image as source words,
2. using an image to initialise the encoder hidden state and,
3. as an as additional input to initialise the decoder hidden state.

We now introduce some models used in our work to train and evaluate on the different datasets we just briefly outlined.

#### 3.3.1 Models

##### 3.3.1.1 $\text{IMG}_E$ : Image for encoder initialisation

We use two new single-layer feed-forward neural networks to compute the initial states of the forward and backward RNN, respectively instead of initialising the

---

<sup>2</sup><http://www.image-net.org>

hidden state of the encoder with the zero vector  $\vec{0}$ , as in the original attention-based NMT model of (Bahdanau et al., 2015)

We use Equation (4.1) to compute a vector  $\mathbf{d}$  from the global image feature vector  $\mathbf{q} \in \mathbb{R}^{4096}$ :

$$\mathbf{d} = \mathbf{W}_I^2.(\mathbf{W}_I^1.\mathbf{q} + \mathbf{b}_I^1) + \mathbf{b}_I^2. \quad (3.1)$$

Here  $\mathbf{W}$  and  $\mathbf{b}$  denote the projection matrix and bias vector, respectively, such that  $\mathbf{W}_I^1 \in \mathbb{R}^{4096 \times 4096}$  and  $\mathbf{b}_I^1 \in \mathbb{R}^{4096}$  while  $\mathbf{W}_I^2$  and  $\mathbf{b}_I^2$  project the image features into the same dimension as the hidden states of the source language encoder.

The encoder hidden state is initialised by the feed-forward networks computed as follows:

$$\begin{aligned} \overleftarrow{h}_{\text{init}} &= \tanh(\mathbf{W}_f \mathbf{d} + \mathbf{b}_f), \\ \overrightarrow{h}_{\text{init}} &= \tanh(\mathbf{W}_b \mathbf{d} + \mathbf{b}_b), \end{aligned} \quad (3.2)$$

where  $\mathbf{b}$  and  $\mathbf{W}$  are respectively the bias vector and the learned multi-modal projection of the image features  $\mathbf{d}$  into the encoder’s hidden state dimensionality. The suffix ‘ $f$ ’ (‘ $b$ ’) corresponds to forward (backward) states.

### 3.3.1.2 IMG<sub>D</sub>: Image for decoder initialisation

Here we use a single-layer feed-forward neural network for incorporating an image into the decoder. Generally, the decoder’s initial hidden state is computed from the encoder’s hidden states i.e. the concatenation of the last hidden states of the encoder forward RNN  $\overrightarrow{h}_N$  and backward RNN  $\overleftarrow{h}_1$ , respectively. We compute the initial hidden state  $\mathbf{s}_0$  of the decoder from using the image features as additional inputs as follows:

$$\mathbf{s}_0 = \tanh(\mathbf{W}_{di}[\overleftarrow{\mathbf{h}}_1; \overrightarrow{\mathbf{h}}_N]) + \mathbf{W}_m \mathbf{d} + \mathbf{b}_{di}, \quad (3.3)$$

where  $\mathbf{W}_{di}$  and  $\mathbf{b}_{di}$  are learned model parameters while the learned image feature  $\mathbf{d}$  is projected into the decoder hidden state dimensionality by the multi-modal projection matrix  $\mathbf{W}_m$ .

As before, given the global image vector  $\mathbf{q} \in \mathbb{R}^{4096}$ , the vector  $\mathbf{d}$  is calculated from Equation (4.1). However, in the present case, the image features are projected into the same dimensionality as the decoder’s hidden states by the parameters  $\mathbf{W}_I^2$  and  $\mathbf{b}_I^2$ .

### 3.3.1.3 IMG<sub>2W</sub> — Image as source words

In this model, we use the image features as the first and last words of the source sentence and an attention model learns when to attend to the image representations. Specifically, given the global image feature vector  $\mathbf{q} \in \mathbb{R}^{4096}$ :

$$\mathbf{d} = \mathbf{W}_I^2 \cdot (\mathbf{W}_I^1 \cdot \mathbf{q} + \mathbf{b}_I^1) + \mathbf{b}_I^2, \quad (3.4)$$

where  $\mathbf{W}_I^1 \in \mathbb{R}^{4096 \times 4096}$  and  $\mathbf{W}_I^2 \in \mathbb{R}^{4096 \times d_x}$  are image transformation matrices,  $\mathbf{b}_I^1 \in \mathbb{R}^{4096}$  and  $\mathbf{b}_I^2 \in \mathbb{R}^{d_x}$  are bias vectors, and  $d_x$  is the source words vector space dimensionality, all trained with the model. We directly use  $\mathbf{d}$  as the first and last words of the source sentence. That is, given the word embeddings for a source sentence  $X = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$ , we concatenate the transformed image vector  $\mathbf{d}$  to it, i.e.  $X = (\mathbf{d}, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N, \mathbf{d})$ , and apply the forward and backward encoder RNN passes. By including images into the encoder in model IMG<sub>2W</sub>, our intuition is that (i) by including the image as the *first word*, we propagate image features into the source sentence vector representations when applying the forward RNN to produce vectors  $\overrightarrow{\mathbf{h}}_i$ , and (ii) by including the image as the *last word*, we propagate image features into the source sentence vector representations when applying the

backward RNN to produce vectors  $\overleftarrow{\mathbf{h}}_i$ .

#### 3.3.1.4 Ensemble decoding

We use an ensemble decoding by adding in succession our best performing multi-modal models from the above-mentioned. We ensembled different models by starting with one of the best performing multi-modal models from Calixto et al. (2017b) on this data set,  $\text{IMG}_D$ , and by adding new models to the ensemble one by one, until we reach a maximum of four independent models, each of which are trained separately and on the original  $\text{M30k}_T$  training data only. In addition we also report results when pre-trained on the English-German WMT 2015 (Bojar et al., 2015) training data coupled with the local visual features extracted with the ResNet-50 network (He et al., 2015).

### 3.3.2 Data Pre-processing

The English, German and French descriptions are pre-processed by normalising the punctuation, lowercasing and token pre-processed the text using the Moses SMT Toolkit (Koehn et al., 2007b). We additionally convert space-separated tokens into subwords (Sennrich et al., 2016) by jointly training for English–German descriptions and separately for English–French descriptions. This results in final vocabulary of 74K English and 81K German subword tokens for English–German models, and 82K English and 82K French subword tokens for English–French models. We discard sentences in English, German or French if they are longer than 80 tokens.

### 3.3.3 Hyper-parameter Setups

We report our results for translating from English into German and into French. We use the 29k entries in the  $\text{M30k}_T$  training set for training our models, and the 1,014 entries in the  $\text{M30k}_T$  development set for model selection, early stopping the training procedure in case the model stops improving BLEU scores on this development set.

To build our multimodal models described in Section 3.3.1, we follow the same configuration as Calixto et al. (2017b) which implements the encoder as a bi-directional RNN with GRU, one 1024D single-layer forward RNN and one 1024D single-layer backward RNN. Throughout the experiments, the models are parameterised using 620D source and target word embeddings, and both are trained jointly with the model. All non-recurrent matrices are initialised by sampling from a Gaussian distribution ( $\mu = 0, \sigma = 0.01$ ), recurrent matrices are random orthogonal and bias vectors are all initialised to 0. Dropout is given a probability of 0.3 in source and target word embeddings, in the image features (in all MMT models), in the encoder and decoder RNNs inputs and recurrent connections, and before the read-out operation in the decoder RNN was applied. Following (Gal and Ghahramani, 2016), dropout to the encoder bidirectional RNN and decoder RNN using the same mask are also applied in all time steps. The models are trained for 25 epochs using Adam (Kingma and Ba, 2014) with learning rate of 0.002 and mini-batches of size 40, where each training instance consists of one English sentence, its corresponding image and its gold-standard translation into German and into French.

Finally, we evaluate our English–German models on three held-out test sets, the Multi30k 2016/2017 and the MS-COCO 2017 test sets (Chen et al., 2015), and our English–French models on the Multi30k 2017 and the MS-COCO 2017 test sets in terms of the following three automatic metrics:

1. **BLEU4** – a widely adopted measure of n-gram match precision (Papineni et al., 2002);
2. **METEOR** – which accounts for both precision and recall and additionally, incorporates more complex linguistic phenomena, such as synonymy and paraphrasing (Denkowski and Lavie, 2014);
3. **Translation Edit Rate (TER)** – which is an inexpensive measure that correlates fairly well with human judgements and is based on counting transformations rather than n-gram matches (Snover et al., 2006).

| Multi30k |              |                                |                               |                                  |
|----------|--------------|--------------------------------|-------------------------------|----------------------------------|
| Lang.    | Model        | BLEU4 $\uparrow$               | METEOR $\uparrow$             | TER $\downarrow$                 |
| en-de    | NMT baseline | 19.3                           | 41.9                          | 72.2                             |
| en-de    | Ensemble     | <b>29.8</b> ( $\uparrow$ 10.3) | <b>50.5</b> ( $\uparrow$ 8.6) | <b>52.3</b> ( $\downarrow$ 19.9) |
| en-fr    | NMT baseline | 44.3                           | 63.1                          | 39.6                             |
| en-fr    | Ensemble     | <b>54.1</b> ( $\uparrow$ 9.8)  | <b>70.1</b> ( $\uparrow$ 7.0) | <b>30.0</b> ( $\downarrow$ 9.6)  |

Table 3.1: Results for the M30k<sub>T</sub> 2017 English–German and English–French test sets. All models are trained on the original M30k<sub>T</sub> training data. Our ensemble uses four multi-modal models, all independently trained: two models IMG<sub>D</sub>, one model IMG<sub>E</sub>, and one model IMG<sub>2W</sub>.

| MS-COCO |              |                               |                               |                                  |
|---------|--------------|-------------------------------|-------------------------------|----------------------------------|
| Lang.   | Model        | BLEU4 $\uparrow$              | METEOR $\uparrow$             | TER $\downarrow$                 |
| en-de   | NMT baseline | 18.7                          | 37.6                          | 66.1                             |
| en-de   | Ensemble     | <b>26.4</b> ( $\uparrow$ 7.7) | <b>46.8</b> ( $\uparrow$ 9.2) | <b>54.5</b> ( $\downarrow$ 11.6) |
| en-fr   | NMT baseline | 35.1                          | 55.8                          | 45.8                             |
| en-fr   | Ensemble     | <b>44.5</b> ( $\uparrow$ 9.4) | <b>64.1</b> ( $\uparrow$ 8.3) | <b>35.2</b> ( $\downarrow$ 10.6) |

Table 3.2: Results for the MS-COCO 2017 English–German and English–French test sets. All models are trained on the original M30k<sub>T</sub> training data. Ensemble uses four multi-modal models, all trained independently: two models IMG<sub>D</sub>, one model IMG<sub>E</sub>, and one model IMG<sub>2W</sub>.

### 3.4 Results and Discussion

In Table 3.1, we show results for translating the 2017 edition of the Multi30k test sets. Table 3.2 shows the results of evaluating the models using the MS-COCO English–French test sets. Again, all models are trained on the original M30k<sub>T</sub> training data only. Additionally, we also report the results for the best model of (Calixto et al., 2017a), which is pre-trained on the English–German WMT 2015 (Bojar et al., 2015) and uses local visual features extracted with the ResNet-50 network (He et al., 2015) in Table 3.3.

It is worth noting that adding more models to the ensemble network improves the translation performance by a large margin. The gain increases to 8–10 and 7–10 points (BLEU, METEOR) for English–German and English–French Multi30k

| Multi30k (English→German)  |           |             |             |             |
|--|-----------|-------------|-------------|-------------|
|  | Ensemble? | BLEU4 ↑     | METEOR↑     | TER↓        |
| NMT <sub>SRC+IMG</sub> <sup>1</sup>  | —         | <u>39.0</u> | <u>56.8</u> | <u>40.6</u> |
| IMG <sub>D</sub>   | —         | 37.3        | 55.1        | 42.8        |
| IMG <sub>D</sub> + IMG <sub>E</sub>  | ✓         | 40.1        | 58.5        | 40.7        |
| IMG <sub>D</sub> + IMG <sub>E</sub> + IMG <sub>2W</sub>                    | ✓         | 41.0        | 58.9        | 39.7        |
| IMG <sub>D</sub> + IMG <sub>E</sub> + IMG <sub>2W</sub> + IMG <sub>D</sub> | ✓         | <b>41.3</b> | <b>59.2</b> | <b>39.5</b> |

<sup>1</sup> This model is pre-trained on the English–German WMT 2015 Bojar et al. (2015), consisting of  $\sim 4.3$ M sentence pairs.

Table 3.3: Results for the best model of (Calixto et al., 2017a), which is pre-trained on the English–German WMT 2015 (Bojar et al., 2015), and different combinations of multi-modal models, all trained on the original M30k<sub>T</sub> training data only, evaluated on the M30k<sub>T</sub> 2016 test set.

respectively using the ensemble decoding over the text-only baselines shown in Table 3.1. We further note that incorporating model IMG<sub>2W</sub> to the ensemble already consisting of models IMG<sub>E</sub> and IMG<sub>D</sub> still improves translations, according to all metrics evaluated. Similar trends were observed when tested with the MS-COCO dataset shown in Table 3.2 with an improvement of 7–9 and 8–9 points (BLEU, METEOR) respectively for the English–German and English–French pair respectively. In Table 3.3 we see the gains from using multimodal models when also trained on the additional data from (Bojar et al., 2015). We see that regardless of the type of the multimodal models, the additional image context always yields an improvement over the text-only NMT.

Our study shows that despite being trained on the same training data, there are inconsistencies in translation quality between the text-only NMT and multimodal systems, at least in terms of evaluation metrics.

In Table 5.6, we show two interesting illustrative examples for our highest scoring multimodal models, i.e, IMG<sub>D</sub> and IMG<sub>E</sub> evaluated on the M30k<sub>t</sub> 2016 datasets. In the first entry, both models IMG<sub>E</sub> and IMG<sub>D</sub> generate a perfect translation with respect to the reference whereas the NMT baseline generates an incorrect translation which is not true from observing the image. However, in the second entry, both the multimodal models extrapolate the reference+image and describe “ceremony” as



|                         |  |
|-------------------------|--|
| <b>Original:</b>        | a man is working a hotdog stand.         |
| <b>NMT:</b>             | ein Mann arbeitet ein Hotdog stehen.     |
| <b>IMG<sub>E</sub>:</b> | ein Mann arbeitet an einem Hotdog-Stand. |
| <b>IMG<sub>D</sub>:</b> | ein Mann arbeitet an einem Hotdog-Stand. |
| <b>Reference:</b>       | ein Mann arbeitet an einem Hotdog-Stand. |



|                         |   |
|-------------------------|---|
| <b>Original:</b>        | a woman with long hair is at a graduation ceremony.     |
| <b>NMT:</b>             | eine Frau mit langen Haarenist an einer StaZeremonie.   |
| <b>IMG<sub>E</sub>:</b> | eine lang haarige Frau bei einer olympischen Zeremonie. |
| <b>IMG<sub>D</sub>:</b> | eine lang haarige Frau bei einer olympischen Zeremonie. |
| <b>Reference:</b>       | eine Frau mit langen Haaren bei einer Abschluss Feier.  |

Table 3.4: Examples of translations produced by the multimodal systems (MMT) and a text-only model (NMT). Original: the original caption. *IMG<sub>D</sub>*: the type of multimodal model output, given the textual input and the global image feature. NMT: the system output, given only the textual input.

“olympischen Zeremonie” (IMG<sub>E</sub> and IMG<sub>D</sub>) with an unknown word “Olympics” (Vilar et al., 2006). We conjecture that this could be because of the fact that the training data is small and often depicts a small variation of different scenes with different forms of biases (van Miltenburg, 2016). Although the text-only NMT produces an absolute incorrect translation of the term with “StaZeremonie” that does not exist in the German language.

### 3.5 Summary

In this chapter, we evaluated multi-modal NMT models which integrate global image features into both the encoder and the decoder in aligned data settings. One of our main interests in this work was to pin-point some of the mechanisms that lead to improvements when learning multimodal as opposed to monomodal representations for solving the task of machine translation. We observe consistent improvements over a text-only NMT baseline trained on the same data, and these are typically very large i.e., 7.0–9.2 METEOR points across language pairs and test sets. Furthermore, our experiments with ensembling different multi-modal NMT models show that these

models can generate translations that compare favourably to multimodal models that use local image features. We also performed follow up experiments where we set a stronger baseline by improving our text-only model performance with training on an additional English-German parallel corpus (Bojar et al., 2015). We observed that with extra textual translation data, visual context provides improvement in performance. To sum up, our results clearly show that additional modality boosts the performance of a translation system, however the illustrative examples results from Table.5.6 makes it is unclear whether the systems successfully use images as context to aid translation.

To conclude, this chapter is dedicated not only to presenting our main findings, but also to pointing to some of the limitations of our work and towards future directions. A further natural extension to the work in this chapter is to take advantage of image data sets for low-resource language translation which is the main focus of Chapter 4. We also investigate the issue of disjoint settings in details in Chapter 5 and examine to what extent these visual modalities are accountable for translation. In future, we also plan to study how to generalise these models to other multi-modal natural language processing tasks, e.g. visual question answering and multimodal language generation.

# Chapter 4

## Images in Sparse Data Conditions

### 4.1 Overview

While there has been a surge of interest in tackling language translation problems using additional information such as an image or video in order to provide some context in high-resource scenarios, it still continues to be a challenging problem in the context of low-resources and out-of-domain settings (Koehn and Knowles, 2017). In general, being able to exploit a variety of data and types of data at the training regime, e.g. data requiring comparatively lesser supervision, is a desirable feature for neural MT models. However, in a non-traditional setup where the training data is scarce, there is a concern that such models will perform poorly with languages having limited resources, especially in comparison with well-resourced major languages. To tackle such situations, early approaches in MT involved the use of *pivot languages*, the language for which sufficient data is available in both sides of a low-resource language pair (Wu and Wang, 2007) as an intermediate step in the translation of the low-resource source and target pair. Along the same line for multimodal representation learning, Gella et al. (2017); Kádár et al. (2017) employ images as a pivot between multiple languages by optimizing a contrastive loss function while Rajendran et al. (2015) use the English representations as the pivot to learn shared multimodal multilingual vector spaces.

Although not directly comparable, recent work in neural multimodal machine translation has also constructed models to translate image captions by using an image as a bridge between source and target language pair (Calixto et al., 2017b; Caglayan et al., 2019, 2017; Frank et al., 2018; Libovický and Helcl, 2017) – however, such techniques have not been explored yet in the context of low-resource multimodal machine translation.

Typically, in the context of multimodal MT, a three-way parallel corpus, which contains bilingual texts and corresponding images is needed. By definition, descriptions of the additional modality, e.g. images in multiple languages, can be seen as the multiple views of the same or closely related data (Kádár et al., 2017). Conceptually, this can be also referred to as *multitask representation learning* (Johnson et al., 2017) — as the models tend to learn generalised internal representations for sentences via a shared joint grounded representation spanning modalities. Chapter 3 explores the issue of such *alignments* in a more pragmatic scenario, where parallel image–sentence corpora are available for different languages. A further natural extension to the work presented in that chapter is to take advantage of the models describe in Section.3.3 in low-resource data settings – which is the main focus of this chapter.

To that end, we formulate the following research question:

*Can a multi-modal context generate a better translation than the text-only MT model in a sparse data-scenario ?*

The additional views of the same data (modality, in our case) can help to overcome the problems of data sparsity and more importantly, have practical implications for efficiently handling the problem of lexical ambiguity. One example is the word “gram” in English (source) which can be translated into different forms in Hindi (target) based on the context (e.g. either a village or chickpea). In such contexts, an additional image or the surrounding contextual words can help with the disambiguation. Although large amounts of parallel texts are available for Hindi-English NMT

translation, there is no such dataset available for training and evaluating Hindi-English multi-modal MT systems. In exploring this, we extend research on creating data for image captioning in the dimension of language and study how to translate image descriptions in a low-resource language, i.e. *Hindi* for an unlabelled image into *English*. Through a review of the same models we observe in Chapter 3, we provide evidence that the performance on lower-resource language can be improved by additionally training with images. To the best of our knowledge, it is the first time a purely multimodal neural machine translation is applied to a dataset that includes an Indian language (Hindi).

## 4.2 Experimental Settings

Arguably, the main downside of applying MMT models in a low-resource language scenario is that there is no amount of publicly available training data, which restricts its applicability to such languages. The current off-the-shelf MMT models are trained on the *translation* portion of the Multi30k (Elliott et al., 2016) data, which is only available for high-resourced languages. We adopt a simple approach, by means of producing candidate translations to expand the current M30<sub>k</sub> dataset to include a new language, namely, *Hindi*. Based on an existing English-image parallel corpus, we develop both a synthetic training dataset as well as a manually translated the validation and test dataset for Hindi. Note that although the English (En) and Hindi(Hi) languages belong to the same family of languages (Indo-European), they differ significantly in terms of word order, syntax and morphological structure (Bharati et al., 1995). While English maintains a Subject-Verb-Object (SVO) pattern, Hindi follows a Subject-Object-Verb (SOV) convention. Moreover, compared to English, Hindi has a more complex inflection system, where nouns, verbs and adjectives can be inflected according to number, gender and case. These issues, combined with the data scarcity problem, make the Hi→En multimodal machine translation a challenging task.

### 4.2.1 Synthetic Data Generation

We now concisely describe the pipeline, illustrated later in Figure 4.1, which we developed to perform multimodal machine translation under the low-resource scenarios in the following steps:

- To create a synthetic in-domain Hindi-English parallel corpus for the image description translation task, we translated the English descriptions of images in the Flickr30k dataset, into Hindi, using a phrase-based statistical machine translation (PBSMT) system (Koehn et al., 2007a). In doing this we are inspired by (Kunchukuttan et al., 2017) to use a PBSMT system over NMT to create low-resource baselines. For the *Hindi*  $\rightarrow$  English translation, their system achieves better results with a PBSMT trained on the same corpus. In addition, we manually translated the data split from *English* source-side into *Hindi* target-side for validation and evaluation. In this we were assisted by two bi-lingual speakers of Hindi and English. One of the speakers translated the datasets into Hindi while the other verified the translation.
- We used this synthetic training data to build both the text-only baseline and multimodal system. For tuning, we used the manually translated validation split.
- Finally, we manually translated the English portion of the test split into Hindi to test our models. Some examples of manually translated descriptions are shown in Table 4.1 where the first column represents the original English captions and the second column represents the manual translation of those captions into Hindi.

### 4.2.2 Resources

Due to the unavailability of an in-domain Hindi-English parallel corpus for our caption translation task, we use an out-of-domain Hindi-English parallel corpus

| English Source Sentence                                  | Hindi Translation ( <i>Manual</i> )   |
|--|---|
| A man in an orange hat starring at something .           | एक नारंगी टोपी में एक आदमी घूर रहा है ।                                       |
| People are fixing the roof of a house.                   | लोग एक घर की छत ठीक कर रहे हैं ।  |
| Group of Asian boys wait for meat to cook over barbecue. | एशियाई लड़कों का समूह बारबेक्यू पर खाना बनाने के लिए मांस का इंतजार करता है । |
| The person in the striped shirt is mountain climbing.    | धारीदार शर्ट में व्यक्ति पहाड़ चढ़ाई कर रहा ।                                 |

Table 4.1: Examples of manually translated captions of the Flickr30k English descriptions in Hindi using PBSMT system.

which is compiled from a variety of existing sources such as OPUS (Tiedemann, 2012), HindEn (Bojar et al., 2014) and TED (Abdelali et al., 2014) as well as corpora developed at the Center for Indian Language Technology, IIT-B<sup>1</sup> over several years. The details of the IITB English-Hindi corpus can be found at (Kunchukuttan et al., 2017). We now describe the existing resources that we used to create the synthetic data. The raw corpus statistics are provided in Table 4.2.

### 4.2.3 Extraction of Image Features

For the visual component, we use the publicly available pre-trained convolutional neural network (CNN) models to extract the global image vectors as described in Simonyan and Zisserman (2014). Their network is trained and evaluated as an extensive set of deep CNN models for classifying images into one out of the 1,000 classes in ImageNet (Russakovsky et al., 2015). For all the images, the *global* image feature vectors, which are the 4096-D activations of the penultimate fully connected layer FC7, henceforth referred to as **g**, are extracted using the 19-layer VGG network (VGG19). We integrate these global image features into our model in 2 ways, namely:

- (1) a. to initialise the encoder hidden state and

---

<sup>1</sup>[www.cfilt.iitb.ac.in](http://www.cfilt.iitb.ac.in)

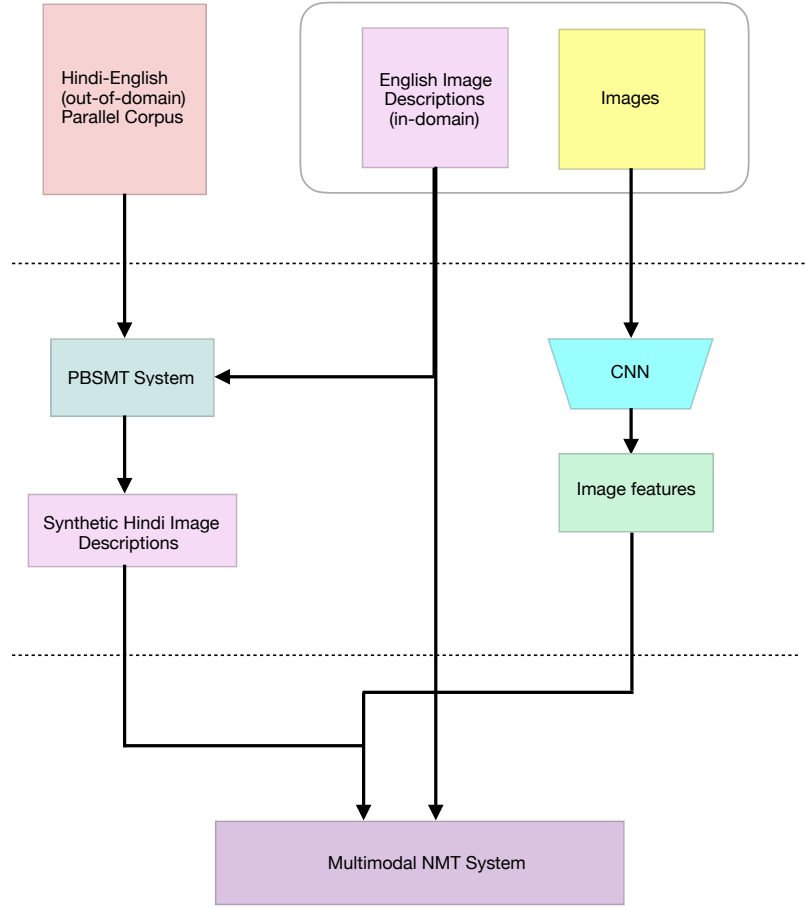


Figure 4.1: Flowchart of our Hi-En MMT System

- b. as additional input to initialise the decoder hidden state.

#### 4.2.4 MT Models

For multimodal NMT, we explore the standard multimodal attention with *global* visual features for our experiment. We use same models as described in 3.3 with encoder-decoder initialisation (INIT) (Calixto et al., 2017b) where we initialise all encoders as well the first decoder layer with a non-linear transformation of the global visual features. For the encoder, we use a bi-directional recurrent neural network (RNN) with gated recurrent unit (GRU) (Cho et al., 2014a), while the concatenation of forward and backward hidden states,  $h_i = [\vec{h}_i, \overleftarrow{h}_i]$  serves as the final annotation vector for a given source position  $i$ . Our model diagram is shown in Figure 4.1.

**IMG<sub>E</sub>:** Instead of initialising the hidden state of the encoder with the zero vector  $\vec{0}$ , as in the original attention-based NMT model of Bahdanau et al. (2015) to linearly project the concatenation of textual and visual context vectors for obtaining the final multimodal context vector, we use Equation (4.1) to compute a vector  $\mathbf{d}$  from the global image feature vector  $\mathbf{g} \in \mathbb{R}^{4096}$ :

$$\mathbf{d} = \mathbf{V}_I^2 \cdot (\mathbf{V}_I^1 \cdot \mathbf{g} + \mathbf{b}_I^1) + \mathbf{b}_I^2. \quad (4.1)$$

Here  $\mathbf{V}$  and  $\mathbf{b}$  denote the projection matrix and bias vector, respectively, such that  $\mathbf{V}_I^1 \in \mathbb{R}^{4096 \times 4096}$  and  $\mathbf{b}_I^1 \in \mathbb{R}^{4096}$  while  $\mathbf{V}_I^2$  and  $\mathbf{b}_I^2$  project the image features into the same dimensionality as the hidden states of the source language encoder.

The encoder’s hidden state is initialised by the feed-forward networks computed as follows:

$$\begin{aligned} \overleftarrow{h}_{\text{init}} &= \tanh(\mathbf{V}_f \mathbf{d} + \mathbf{b}_f), \\ \overrightarrow{h}_{\text{init}} &= \tanh(\mathbf{V}_b \mathbf{d} + \mathbf{b}_b), \end{aligned} \quad (4.2)$$

where  $\mathbf{b}$  and  $\mathbf{V}$  are respectively the bias vector and the multi-modal projection matrix for projecting the image features  $\mathbf{d}$  into the encoder’s hidden state dimensionality. The suffix ‘ $f$ ’ (‘ $b$ ’) corresponds to forward (or backward) states.

**IMG<sub>D</sub>:** A new single-layer feed-forward neural network is used for incorporating an image into the decoder. Originally, the initial hidden state of the decoder is computed from the last hidden states of the encoder’s forward RNN and backward RNN, respectively  $\overrightarrow{h}_N$  and  $\overleftarrow{h}_1$ , or from the mean of the source-language annotation vectors  $h_i$ . However, here we compute the initial hidden state  $\mathbf{s}_0$  of the decoder by

including the global image features as additional inputs as follows:

$$\mathbf{s}_0 = \tanh(\mathbf{V}_{di}[\overleftarrow{\mathbf{h}}_1; \overrightarrow{\mathbf{h}}_N]) + \mathbf{V}_m \mathbf{d} + \mathbf{b}_{di}, \quad (4.3)$$

where  $\mathbf{V}_{di}$  and  $\mathbf{b}_{di}$  are learned model parameters while the image feature  $\mathbf{d}$  is projected into the decoder hidden state dimensionality by the multi-modal projection matrix  $\mathbf{V}_m$ .

As before, given the global image vector  $\mathbf{g} \in \mathbb{R}^{4096}$ , the vector  $\mathbf{d}$  is calculated from Equation (4.1). However, in the present case, the image features are projected into the same dimensionality as the decoder’s hidden states by the parameters  $\mathbf{V}_I^2$  and  $\mathbf{b}_I^2$ .

#### 4.2.5 Data Pre-processing

The Hindi side of the out-of-domain dataset is normalised using the Indic\_NLP\_Library<sup>2</sup> to ensure a canonical Unicode representation. We used the scripts from the above library to tokenise and transcribe the Hindi sentences. For English, we used the scripts from the Moses tokeniser *tokenizer.perl*<sup>3</sup> to tokenise and to turn into lower-case the English representations for our experiments.

#### 4.2.6 Model Hyperparameters

**Out-of-domain PBSMT training:** We use similar settings to those reported in Kunchukuttan et al. (2017) to create the synthetic Hindi dataset. They used the news stories from the WMT 2014 English-Hindi shared task (Bojar et al., 2014) as their validation (dev) and test corpora which we concatenate together to create our dev set. The training and dev corpus consist of 1,492,827 and 3207 sentence segments respectively. We used the HindMono (Bojar et al., 2014) corpus which contains roughly 45 million sentences to build our language model in Hindi. The

<sup>2</sup>[https://bitbucket.org/anoopk/indic\\_nlp\\_library](https://bitbucket.org/anoopk/indic_nlp_library)

<sup>3</sup><https://github.com/moses-smt/mosesdecoder/blob/RELEASE-3.0/scripts/tokenizer/tokenizer.perl>

| Splits             | Data-type | English    | Hindi      |
|--------------------|-----------|------------|------------|
| Train <sup>1</sup> | tokens    | 20,667,259 | 22,171,543 |
| Dev <sup>1</sup>   | tokens    | 68459      | 74027      |
| Monolingual        | sentences | 20,638,520 | 45,075,279 |

<sup>1</sup> The total number of 1,492,827 training and 3207 validation sentences were used to train the PBSMT system

Table 4.2: The overall statistics of the datasets used to train the PBSMT system. The 3rd row shows the amount of additional monolingual Hindi and English text used respectively for training the language model to create synthetic Hindi and the general-domain PBSMT system.

corpus statistics are shown in Table 4.2.

For training the out-of-domain system, we use the Moses (Koehn et al., 2007a) SMT system. We used the SRILM toolkit (Stolcke, 2002) for building a 4-gram language model and GIZA++ Och and Ney (2000) with the grow-diag-final-and heuristic for extracting the phrases. The trained system is tuned using Minimum Error Rate Training (Och, 2003). For other parameters of Moses, default values are used. If the sentences in English or Hindi are longer than 80 tokens, they are discarded. Additionally, we use the News Crawl: articles from 2016 from WMT17 <sup>4</sup> for English to train the language model. This contains roughly 20 million sentences for English, with details shown in Table 4.2.

**Multimodal training:** We follow the similar settings of Calixto et al. (2017b) as described in Section.3.3.3 for training our multimodal models. All models are trained for 25 epochs using Adam (Kingma and Ba, 2014) with a learning rate of 0.002 and a batch size of 40 sentences, where each training instance consists of one English sentence, one Hindi sentence and one image. In addition, we used early-stopping with patience 10 based on validation loss and output parameters were saved after every 10,000 iterations.

<sup>4</sup><http://www.statmt.org/wmt17/translation-task.html>

| Multi30k (Hindi→ English)                |            |         |             |             |
|--|------------|---------|-------------|-------------|
|  | In-domain? | Images? | BLEU↑       | METEOR↑     |
| PBSMT                                    | ✓          | –       | 22.7        | 30.2        |
| PBSMT <sub>out-domain</sub> <sup>1</sup> | –          | –       | 21.6        | 29.6        |
| NMT <sub>text</sub>                      | ✓          | –       | 23.3        | 29.7        |
| IMG <sub>D</sub>                         | ✓          | ✓       | <b>24.2</b> | <b>30.7</b> |
| IMG <sub>E</sub>                         | ✓          | ✓       | 23.9        | 29.9        |

<sup>1</sup> This model is pre-trained on the the dataset as described in Table 4.2

Table 4.3: Evaluation metrics scores Hi-En translation systems before and after applying image features on manually translated dev data. Bold numbers indicate improvements that are statistically significant compared to NMT text with  $p = 0.05$ . Evaluation is performed against the English translations of the test set using standard MT evaluation metrics, with BLEU and METEOR

## 4.3 Results and Discussion

The comparative evaluation results of our systems are presented in Table 4.3.

We see from the results that the text-only NMT model outperforms the phrase based SMT model in terms of BLEU score. Our results indicate that incorporating image features in multimodal models helps, as compared to our text-only SMT and NMT baselines. This is reflected in the fact that both the image models are shown to produce better results in terms of BLEU scores with respect to both the SMT and NMT text-only counterpart. Taken together, this confirms the quantitative benefit of the visual modality in NMT models in the low-resource language scenarios.

It is worth noting that IMG<sub>E</sub> yields only a small improvement over the text-only NMT counterpart while IMG<sub>D</sub> performs consistently better in terms of both metrics (BLEU by  $\uparrow 0.9$  and METEOR by  $\uparrow 1$ ) in comparison to the strong text-only NMT and SMT baseline. We conjecture that using image features directly to initialise the encoder hidden state causes the model to overfit and prevents learning.

We now provide an illustrative example shown in Table 4.4 of our highest-scoring MMT and NMT system with respect to the original English reference and the manual source in Hindi.

In the first entry, although the NMT system without images incorrectly trans-




---

**Manual:** तदो लोग अजीब विदेशी जैसी वेशभूषा पहनने, एक नीले और एक बैंगनी, एक सड़ क मोः

**NMT:** two people dressed in exotic costumes wear a blue and one flag in a blue, are standing in a road .

**MMT:** two people wearing funny foreign attire, one blue and one *purple*, are standing in a street.

**Reference:** two people wearing odd alien-like costumes, one blue and one purple, are standing in a road.

---

Table 4.4: Illustrative example of translations produced by the multimodal systems (MMT) and a text-only model (NMT). Manual: the original translated caption in Hindi. NMT: the system output, given only the textual input. MMT: the multimodal system output, given the textual input and the global image feature. Reference: the gold standard reference in English.

lated the color ‘purple’ (as can be seen from Table 4.4, where the costumes are clearly in two different colours) the multi-modal model translated it correctly, yielding an improvement in the sentence-level BLEU ( $\uparrow$  **21.47**) score.

In terms of translations, we see that both the models extrapolate the reference and translate “alien-like costumes” into “exotic costumes” (text-only model) and as a “funny foreign attire” (multimodal model). We attribute this to the fact that the training set is small with fixed set vocabulary – it only renders a small variations in representing different scenes.

## 4.4 Summary

Within the scope of this chapter, we investigate the potential impact of using candidate Hindi descriptions of the Flickr30k dataset generated via back-translation for multimodal machine translation and provided a benchmark baseline result on this corpus. The main idea is to incorporate image features into different parts of the encoder and decoder, and to evaluate whether or not they provide any substantial gains over a vanilla NMT model under the low-resource language settings.

Our work show that a multimodal NMT system trained on synthetic training

dataset can still improve translation quality by exploiting additional visual cues. We also show that when collecting data in a different language, it is better to collect captions for the existing images because we can exploit the visually grounded word and sentence-representations. Nevertheless, our result shows that despite being trained on the same in-domain En–Hi training data, there are differences in translation quality between the SMT and NMT system, at least in terms of evaluation metrics.

These results are not necessarily surprising given that the grammatical syntax between the two languages is poorly represented in the synthetic Hindi training data. In addition to this, Hindi as a language presents many of the well-known issues that NMT currently struggles with (resource sparsity, rich morphology and complex inflection structure).

To conclude, we have proposed a strategy to use synthetic data for training an MMT system for a scenario where parallel data is not available. However, our method is still dependent on large amounts of out-of-domain comparable data, which is still far from what one can gather in a truly low-resource language scenario. In exploring both phrase-based method as well as additional back translation techniques, we hope to make our method available to low-resourced language pairs in the future.

Finally, we saw how our MMT systems trained on candidate captions achieved improvement over the baseline trained only on the textual data. It is therefore vital that MMT resources are developed for low-resource languages.

# Chapter 5

## Images in Adversarial Data Conditions

### 5.1 Overview

While recent work on Multimodal Machine Translation (MMT) systems has shown gains in terms of system performance by employing visual context over text-only NMT, there are several scientific challenges contained therein, such as adversarial data conditions. MMT systems operate over the two very different modalities – text and images which are pooled together, where the information provided by each modality may have different levels of resistance to adversarial attacks. Pooling them together as in MMT in such a way as to get maximum overall benefit, is one such challenge.

In general, the robustness of machine translation models to adversarial samples has been studied considerably in the last two years with notable works by Belinkov and Bisk (2017); Tramèr et al. (2017); Ebrahimi et al. (2018); Khayrallah and Koehn (2018). In a similar vein, Michel and Neubig (2018) proposed a benchmark dataset of noisy texts to evaluate the robustness of translation systems. However, the study of robustness for Multimodal Machine Translation systems have never been the subject of debate in the literature until recently. Elliott (2018) argued that the ad-

ditional visual modality is not necessarily used by showing that system performance did not change when the system was evaluated with randomly selected images. We conjecture that the most plausible reason for such textual predominance is that the Multi30K dataset is inevitably biased towards the source text portion, which makes it sufficient as context to perform the translation without using the visual data. Our observation about this linguistic dominance is in line with the findings of Grönroos et al. (2018) who demonstrated that the effects of visual modality are rather small in multimodal translation models and attributed their largest gains to using additional (unconstrained) text data. More recently, Caglayan et al. (2019) presented an experiment in which colours and entities were masked during the training of a multimodal translation model. They found that training the model under these conditions resulted in the system relying on the visual modality to recover the masked words during evaluation. Although, their results show that the visual modality can be used to recover the masked tokens in the source sentences, it is not clear if these systems will perform similarly when there is a mismatch between the textual and visual concepts. To that end, we construct hard negative textual adversaries with contradictory meanings to explore the robustness of systems to textual adversaries.

In this chapter, we explore the role of both modalities used in combination in multimodal translation systems to address the following research question:

*Can multi-modal models exploit the visual modality to generate better quality automatic machine translation than single-modality models, even in adversarial conditions ?*

For this, the data from the two modalities refer to the same content, so it could be an image and a textual caption describing that image. We construct textual adversarial samples to probe the contribution of the textual input in these systems and we follow the work reported in Elliott (2018) to construct the visual adversaries. Our textual adversaries are based on minimally manipulating the textual examples, e.g:

- (1)
  - a. The woman runs through a park.
  - b. \*The woman runs through a car park.

In this sense, the perturbation in adversarial caption (1b) still retains most aspects of the original caption but it depicts a completely unrelated scene to the original one. We are interested in how significantly these types of perturbations affect the performance of multimodal translation systems. We expect that small perturbations should only result in small changes to the resulting translation. If the system is sufficiently modelling the visual modality, we expect it to ignore this type of perturbation, and to produce the correct translation by leveraging the image.

Put simply, our analysis is to evaluate  $p(y|\bar{x}, c)$ , where  $\bar{x}$  is some corruption to *textual input*  $x$  and  $c$  is the *visual signal*. If the model is usefully representing both  $x$  and  $c$ , it should be able to discard small errors in  $x$ . Additionally, we also strive towards aligning these two modalities in an explicit way. The majority of previous work in MMT has focused on using visual representations (both global and spatial local preserving features) only as an additional context and great progress has been achieved in these endeavours. While a fixed set of visual concepts constitute a convenient modelling assumption for such models, they are in fact restrictive when compared to the “multiplex mentions of rich descriptions” (Karpathy and Fei-Fei, 2015) that a human can compose. Thus, we focus on building a model that is able to reason about the contents of the images and their mentions in the corresponding descriptions.

Concretely, our contributions are twofold:

1. We evaluate the robustness of three state-of-the-art multimodal translation systems in the presence of adversarial textual data. This evaluation is based on four types of textual adversaries described in Section 5.2.2. We also probe the visual awareness of these models by exposing them to randomly sampled images.
2. In the light of the above results, we infer explicit alignments between differ-

ent modalities and use them as additional information alongside the learned embeddings of the source tokens to generate translations.

## 5.2 Experimental Settings

### 5.2.1 Dataset

Each system in this analysis that we present here is trained on the 29,000 English-German-image triplets in the *translation* data which is part of the Multi30K dataset (Elliott et al., 2016). The analysis is performed on the Multi30K Test 2017 split Elliott et al. (2017). The predicted translations are evaluated against human references using Meteor 1.5 Denkowski and Lavie (2014). Note that the translations of the textual adversaries are evaluated against the gold standard, not what the model *should* predict, given the adversarial input.

### 5.2.2 Generating textual adversaries

We define *visual term* as a word or phrase that describes something clearly illustrated in the image. In our experiments, we replace a *visual term* in a sentence to create the textual adversary. In total, we experiment with four types of adversaries: numeral replacement, noun head replacement, preposition replacement, and switching the order of the noun phrases in a sentence. We follow the methodology introduced in (Young et al., 2014; Hodosh and Hockenmaier, 2016; Shi et al., 2018) to create these samples. The numeral detection, noun phrase detection, and preposition detection are performed using syntactic analyses from the SpaCy toolkit Honnibal and Johnson (2015). Table 5.1 presents an overview and examples of each type of textual attack.

**Replace Numeral (Num):** Our simplest attack is to replace the numeral in a sentence with a different quantity. To achieve this, we detect the tokens in a sentence that represent numbers (based on their part-of-speech tags) and replace them with an alternative. In addition, we treat the indefinite articles “a” and “an”


|   | Type | Original and Adversarial   |
|---|------|--|
|  | Noun | Two people walking on the <u>beach</u> .<br>Two people walking on the <i>chestnut</i> .            |
|   | Num  | <u>Two</u> people walking on the beach.<br><i>Four</i> people walking on the beach.                |
|   | NP   | <u>Two people</u> walking on <u>the beach</u> .<br><i>The beach</i> walking on <i>two people</i> . |
|   | Prep | Two people walking <u>on</u> the beach.<br>Two people walking <i>through</i> the beach.            |

Table 5.1: Examples of adversarial textual samples that we use to attack the multimodal translation models. The underlined text denotes the words or phrases that are perturbed to create the adversarial example.

as the numeral “one” because they are typically used as numerals in image captions. Furthermore, subsequent noun phrase chunks are either singularised or pluralised accordingly.

**Replace Noun head (Noun):** We extract the list of all concrete<sup>1</sup> noun-heads (Zwicky, 1985) from the COCO dataset Lin et al. (2014) and swap them with the noun heads in our data. We use WordNet Miller (1998) heuristic hypernymy rules that allows us to replace noun heads with terms that are semantically different. As an example

- (2) a. The woman runs through a field.

With consideration to the caption in (1a), in (2a), “field” is a hypernym of “park”, and so a caption does not create a good adversary for (1a). However, (1b) does create a semantically different adversarial example.

<sup>1</sup>We compute the concreteness of words following (Turney et al., 2011) and consider only those heads with concreteness measure. The degree of concreteness in a word’s context is correlated with the likelihood that the word is used in a literal sense and not metaphorically (Turney et al., 2011)  $\theta > 0.6$ .

**Switch Noun Phrases (NP):** For each caption, the extracted the noun phrases are shuffled and put back to their original position to form the samples. In the example in Table 5.1, we refer to *two people* and to *the beach* respectively as the partitive first noun phrase ( $NP_1$ ) and second noun phrase ( $NP_2$ ). The position of  $NP_1$  and  $NP_2$  are swapped in our method. As a result, the *adversarial* caption depicts a completely different scene. Such examples allow us to evaluate whether our models can identify semantically important changes in word-order even when the bag-of-words representation of two captions are same.

**Replace Preposition (Prep):** Finally, we detect the prepositions used in a sentence and randomly replace them with different prepositions. The translation system should be least sensitive to this type of preposition because it typically results in the smallest change in the meaning of the sentence, as compared to switching the noun phrases.

### 5.2.3 Generating visual adversaries

Visual concepts and their relationships with textual data is expected to provide rich supervision to multimodal translation systems. In addition to evaluating the robustness of these systems to textual adversaries, we also determine the interplay with visual adversaries. We pair each caption with a randomly sampled image from the test data to break the alignment between learned word semantics and visual concepts. We hypothesise that the performance of these systems would demonstrate inferior performances when incorporated with incongruent visual inputs.

### 5.2.4 Models

We evaluate the performance of the following three pre-trained multimodal translation systems<sup>2</sup> under the four different types of textual adversaries:

---

<sup>2</sup>We use these three systems to make our evaluation comparable to Elliott (2018)



Figure 5.1: An evaluation example of visual adversaries. The model sees a congruent image (left) or an incongruent image (right).

**decinit:** A learned transformation of the global 2048D visual data is used to initialise the decoder hidden state. (Caglayan et al., 2017). Similar to this model is  $\text{IMG}_D$  (Calixto et al., 2017b) described in Chapter. 3 where the decoder is initialized with the sum of global visual features.

**trgmul:** The target language word embeddings and global 2048-dimensional visual representations are interacted through element-wise multiplication (Caglayan et al., 2017).

**hierattn:** The decoder learns to selectively attend to a combination of the source language and a  $7 \times 7 \times 512$  volume of spatial-location preserving visual features (Libovický and Helcl, 2017).

## 5.3 Results and Discussion

**Under Adversarial Conditions** Table 5.4 shows the results of evaluating the three state-of-the-arts models using the textual and visual adversaries. We further perform a non-parametric Wilcoxon signed-rank test and rejected the null hypothesis that *randomly sampled images have no impact on the quality of multimodal translation compared to the correct images*. We observe that the visual adversaries lead to a small drop in translation performance, which supports the claim of Elliott



|   |                      |  |
|---|----------------------|--|
|  | <b>Original:</b>     | A man paddles an inflatable canoe.                 |
|   | <b>Noun:</b>         | A <u>city</u> paddles an inflatable canoe.         |
|   | <b>MMT:</b>          | Eine Stadt paddelt in einem aufblasbaren Kanu.     |
|   | <b>MMT+explicit:</b> | Ein Mann paddelt ein aufblasbares Kanu.            |
|   | <b>Reference:</b>    | Ein Mann paddelt in einem aufblasbaren Kanu .      |
|  | <b>Original:</b>     | Two white dogs cuddle their heads together.        |
|   | <b>Noun:</b>         | Two white <u>maps</u> cuddle their heads together. |
|   | <b>MMT:</b>          | Zwei weiße Karten kuscheln ihre Köpfe aneinander.  |
|   | <b>MMT+explicit:</b> | Vier weiße Hunde jagen ihre Köpfe zusammen.        |
|   | <b>Reference:</b>    | Zwei weiße Hunde stecken die Köpfe zusammen.       |

Table 5.2: Examples of translations produced by the **hierattn** system (MMT) and the **MMT+explicit** model for Noun adversary. Both are the outputs, given the adversarial caption and the correct image. Original: the original caption. Noun: the adversarial caption with the underlined replacement.

(2018) that these systems can translate without significant performance losses in the presence of the wrong images. For the textual adversaries, all three systems suffer the most from the numeral replacements, which suggests that the counting ability of the models depend on the available linguistic information. However, that the changes in concept representatives (NP/Nouns) are somewhat similar indicates that none of the models actually understands the underlying semantics of the texts i.e. the difference between “*cat drinks milk*” and “*milk drinks cat*”. To further understand the role of visual component for the purposes of translation, we also carried out as a lower-bound experiment, i.e. – how a text-only NMT systems performs with the adversarial sentences. The results of these experiments are shown in Table.5.3.

| <i>Type</i> | <b>NMT</b> | <b>MMT</b> |
|-------------|------------|------------|
| Noun        | 38.26      | 38.8       |
| NP          | 37.20      | 37.0       |
| Num         | 34.95      | 35.0       |
| Prep        | 40.73      | 42.8       |

Table 5.3: Behaviour of text-text translation model v/s text-image translation model in adversarial conditions.

|          | Original    | Textual |      |      |      | Visual |
|----------|-------------|---------|------|------|------|--------|
|          |             | Noun    | Num  | NP   | Prep |        |
| trgmul   | <b>52.1</b> | 40.9    | 37.3 | 40.9 | 46.3 | 52.3   |
| decinit  | <b>51.5</b> | 40.4    | 37.5 | 40.5 | 45.8 | 51.9   |
| hierattn | <b>48.2</b> | 38.8    | 35.0 | 37.0 | 42.8 | 46.2   |

Table 5.4: Corpus-level Meteor scores for the English–German Multi30K Test 2017 data. Original: performance of systems evaluated on the original text and images. Textual: evaluation on the four different textual adversaries and the correct images. Visual: evaluation on the correct text but adversarial images.

**Analysis:** The intuition behind our effort is to assess how multi-modal models that make use of images to visually ground translations perform when translating visual term, and in principle increase translation quality by doing so. Table 5.6 shows qualitative examples of translations under textual adversarial conditions for the **hierattn** system. We also show the output of the same system given the original image–caption pair. In these examples, we see that the system produces incorrect translations with respect to either the sentence or the image. In NUM, pluralizing “A” with “Two” causes the model to generate an *unknown word*<sup>3</sup> “Japan” instead of “Halloween”. It is likely that the model has learned good representations of both “A” and “two” because these words occur frequently in the training data, it is evident that model fails to distinguish singulars against plurals, resulting in an incorrect translation. In PREP, swapping “in” for “up” causes the model to include an *incorrect lexical choice* “fische” (“fish”) instead of “waterfall”, which is not true, given the image. This example shows that a small lexical error can have a catastrophic effect on the output. This may be because - at least in Multi30K - the semantics of spatial relations are not diverse enough. For NOUN, switching “man” for “city” causes the model to generate an output containing the *mistranslated unit* “Stadt”(“city”), although a man is clearly visible in the image. This implies that additional visual signals is not always helpful in the obvious situations where we wish

<sup>3</sup>We use the error taxonomy from Vilar et al. (2006).

to translate direct visual terms. In NP, we see that the systems fail to fully capture the information contained in the image, resulting in *under-translation*. However, unlike the output under the adversarial condition which simply left an important visual concept “people” untranslated, the model with the original sentence translates “People” into “Menschen”. An inspection of the training data shows that there are sentences that describe ‘people fishing’, and so the model may be exploiting the distribution in the training data.

This analysis shows that the visual modality does not help the system to recover the correct translation, given textual adversaries. Overall, our evaluation also offers new insights on the limitations of these systems. From the experimental results we find that the information contained in the images are ignored by these models when there is an error in the corresponding caption, thereby hinting at the misalignment between them. We conjecture that the most plausible reason for the linguistic dominance is due to the representation or modelling limitations of these models.

|               | Textual     | Original    |
|---------------|-------------|-------------|
| Model         |             |             |
| MMT+explicit. | <b>52.5</b> | <b>52.9</b> |
| hierattn      | 48.2        | 38.8        |

Table 5.5: Corpus-level Meteor scores for the English–German Multi30K Test 2017 data. Original: performance of the **hierattn** and the proposed **MMT+explicit** system evaluated on the original text and images. Textual: evaluation on the noun textual adversary and the correct images.

In addition to evaluating the performance of the models mentioned in Section 5.2.4, we also conduct follow-up experiments to address the limitations of implicit alignment between different modalities in these models. We follow a simple *word-image matching* approach through the use of corresponding image region to maintain the cognitive alignment between these modalities. To do this, we treat the corresponding mentions of an image in the source sentence as weak labels and subsequently supplement them with additional visual information. Our approach is

inspired by (Frome et al., 2013) who associate words and images through a semantic embedding. More closely related is the work of Karpathy and Fei-Fei (2015), who decompose images and sentences into fragments and infers their inter-modal alignment based on grounding dependency tree relations. In contrast to their model which uses a ranking objective, we only align the *learned word embeddings* of the object category associated with the image to the corresponding source token in the sentence. Finally we use these alignment information to align and guide our Multi-modal machine translation system.

### 5.3.1 Explicit Alignments

**Visual category:** We define “*object category*” in an image to be the instances of semantic objects that provide direct information about the source word. Following prior work (Karpathy and Fei-Fei, 2015), we observe that continuous segments of sentence descriptions make meaningful references to objects in an image. One important consideration is how to identify suitable object categories for comparison so that we can best utilise the associated source token in the sentence representations. In doing so, we extracted the associated object categories using the *Oracle annotations* of the Flickr30k dataset (Plummer et al., 2015) and the OpenImage Detection Challenge (Kuznetsova et al., 2018). Our categories can be classified into two types: the smaller set of more general Oracle categories such as “*people, animals*” and the broader set of more specific predicted categories such as “*man, woman*”.

**Alignment:** Subsequently, to align the detected categories and their corresponding mentions, we calculate their similarity matrix.<sup>4</sup> Our simple approach involves flattening the sentence representation into a group of tokens (labels) and matching them with the associated category information of the image. This connection is made using standard metrics such as cosine similarity. In accordance with the revolutionary work of Frome et al. (2013) our goal is to make the model so informed enough

---

<sup>4</sup>To reduce the training costs of the visual models, we avoid the use of *pool5* features of the direct image regions instead of the pre-trained word-level embeddings of the category.

that it is able to draw reasonable conclusions about candidate labels it has never seen visually before. For example, if tested on images it has never observed before, for example a photo of a puppy, and asked whether the correct token is more likely *dog* or some other unfamiliar token (say, *apple*), our model has a fighting chance of guessing correctly because the language model of the source representation ensures that the representation of *puppy* is close to the representation of dogs the model *has* seen, while the representation of *apple* is closer to those of other fruits. Figure 5.2 is an example of the tokens in the source sentence are aligned to its matching category of the visual region.

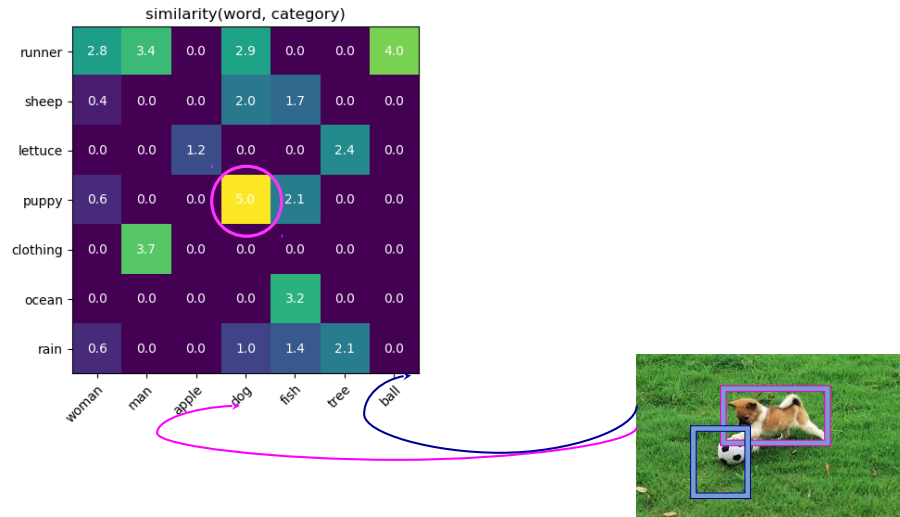


Figure 5.2: Explicit alignment between of the *source* word *puppy* and its corresponding visual category

**MMT+explicit:** Finally, for the first run of the experiment we assign object category only to the corresponding noun-head of the annotated sentence and set the categories of other tokens to empty. And then for each of the source words the learned word embedding is fused with the pre-trained word embeddings of the cat-

egory through concatenation which is then passed into our encoder. The encoder is a bi-directional RNN with GRU (Cho et al., 2014a), where a forward RNN generates a forward annotation vector at each *encoding step*. Similarly, a backward RNN generates a backward annotation vectors and subsequently, these two are concatenated to produce the final context vector which in turn is used by the decoder. Our decoder is an RNN with a conditional GRU with attention over the learned word embeddings. To reduce the training costs of the visual models, we avoid the use of *pool5* features of the direct image regions instead of the pre-trained word-level embeddings of the category. Our model **MMT+explicit** integrates the source language word embeddings and pre-trained visual categories in the encoding stage as illustrated in Figure 5.3.

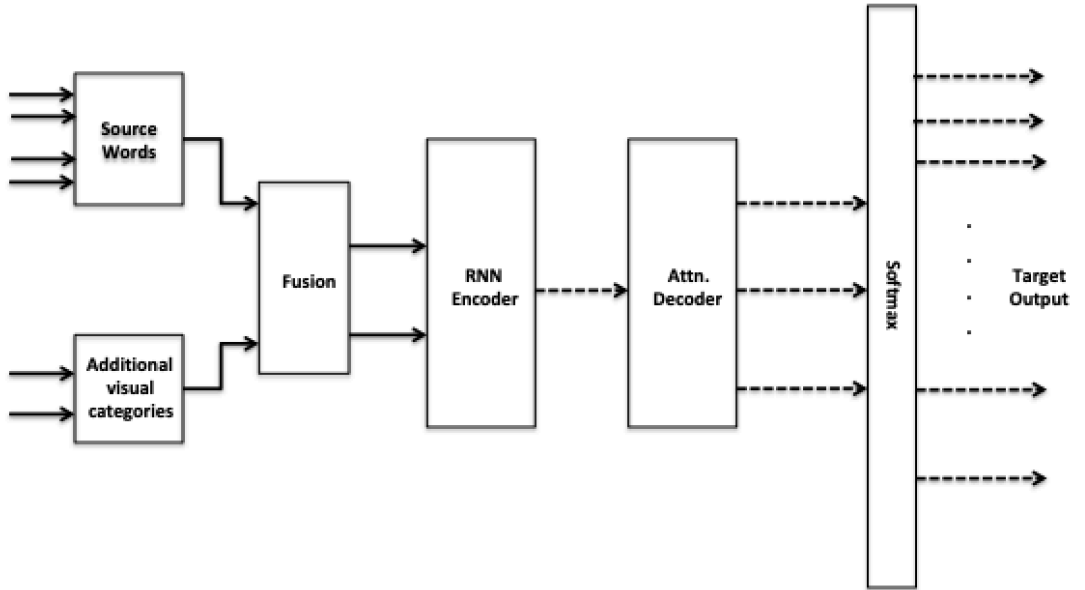


Figure 5.3: **MMT+explicit** model with explicit alignment between the source token and visual category

**With Explicit Alignments** A quick glance at the aforementioned models, we find that the way visual modality is integrated into these models impose limits on their variety. To address this issue, we conduct experiments under the same adversarial setup but instead, using our MMT+EXPLICIT model described in Section. 5.3. The result of our first experiment with noun-heads is presented in Table.5.5.

Our result clearly show that the systems perform well with respect to the evaluation metrics and is able to penalise the noun adversary. To get a sense of how the MMT+EXPLICIT model performs is presented in Figure 5.2. In both sentences, adding category embeddings improve the translation. In the first example, though the MMT says “Stadt ” (“city”), with category embeddings, the model says “Mann”, which is true, given the image. In the second example, we perturb “dogs” with “maps.” Though with category embeddings, the model gets the correct translation, “Hunde”, the model without category embeddings produces an output containing the mistranslated unit “Karten” (“cards”), although dogs are clearly visible in the image. We notice that as the amount of linguistic information in the form of learned category embedding increases, the MMT+EXPLICIT system gradually becomes less perplexed by the mismatch between the visual and textual modality or, put in another way, less sensitive to the adversarial condition.

## 5.4 Summary

**Limitations of the dataset:** Our evaluation offers new insights on the limitations of the current state-of-the-art systems. The results indicate that these systems are primarily performing text-based translations, which is supported by the observation that the visual adversaries do not harm the systems as much as their textual counterparts. We conjecture that the most plausible reason for such textual predominance is that the Multi30K dataset is inevitably biased towards the source text portion, which makes it sufficient as context to perform the translation. Our observation about this linguistic dominance is in line with the work of Grönroos et al. (2018) who demonstrated that the effects of visual modality are rather small in multimodal translation models and attributed their largest gains to using additional (unconstrained) text data.

**Limitations of alignments:** The current MMT models do not capture the intended sense of image concepts, resulting in translations that are completely unrelated to the image (refer to 5.6). The problem seems to be deep-rooted with the process of learning cross-modal representations of concrete concepts such as “objects” from the image data. To solve this, we devise our model to integrate image information as a naturally occurring meaning representation of the source sentence – such that – it acts as complementary information rather than redundant. Our initial investigation using grounded visual information suggests input noise such as errors in the source text can easily be mitigated. We conducted experiments where our model is not fed with visual category features at decoding time but also trained with them right from the scratch. The results suggest that such a model can learn to ignore the textual adversary. In fact, their gain is prominent in the adversarial condition. For future work, we would like to test our proposed model in all adversarial as well for the sparse data conditions.

To sum up, we presented a systematic analysis of the potential contribution of images for the task of multimodal machine translation. Specifically, we explored the behaviour of state-of-the-art systems on adversarial sentences that share some aspect of the correct caption to understand the impact of the textual data. Our results indicate that the systems are primarily performing text-based translations, and this is confirmed by the observation that the visual adversaries do not harm the systems as much as their textual counterparts.

To solve this problem, we introduce a novel multi-modal NMT model to incorporate visual category information visual information into NMT. We have reported improved results on the M30kT testset under adversarial conditions, improving on previous multi-modal attention based models. Our approach is able to exploit the visual category of the image regardless of the domain and aligns them to the corresponding source mention in the encoding step. As future work, we intend to devise models with an adversarial ranking loss that forces the model to pick up important differences between the true image-text pair and distractors. Furthermore, we would

also like to extend this approach to check whether the models are more sensitive to semantically incorrect input or grammatically incorrect input.



|   |                    |   |
|---|--------------------|---|
|    | <b>Original:</b>   | A group of young people dressed up for Halloween.           |
|   | <b>Baseline:</b>   | eine Gruppe junger Menschen verkleidet verkleidet .         |
|   | NUM:               | <u>Two</u> groups of young people dressed up for halloween. |
|   | MMT:               | Zwei Gruppen von jungen Menschen in Japan .                 |
|   | <b>Reference :</b> | Eine Gruppe junger Leute verkleidet sich für Halloween.     |
|    | <b>Original:</b>   | A beautiful waterfall in the middle of a forest             |
|   | <b>Baseline:</b>   | ein schöner Wasserfall in der Mitte eines Waldes .          |
|   | PREP:              | A beautiful waterfall <u>up</u> the middle of a forest      |
|   | MMT:               | Eine schöne Fische in einem Wald .                          |
|   | <b>Reference:</b>  | Ein schöner Wasserfall mitten im Wald .                     |
|  | <b>Original:</b>   | A man paddles an inflatable canoe.                          |
|   | <b>Baseline:</b>   | ein Mann paddelt in einem aufblasbaren Kanu .               |
|   | NOUN:              | A <u>city</u> paddles an inflatable canoe .                 |
|   | MMT:               | Eine Stadt paddelt in einem aufblasbaren Kanu .             |
|   | <b>Reference:</b>  | ein mann paddelt in einem aufblasbaren kanu .               |
|  | <b>Original:</b>   | People fishing off a pier.                                  |
|   | <b>Baseline:</b>   | Menschen beim Angeln.                                       |
|   | NP:                | A <u>pier</u> fishing off <u>people</u> .                   |
|   | MMT:               | Ein Pier beim Angeln .                                      |
|   | <b>Reference:</b>  | Leute fischen an einem Pier .                               |

Table 5.6: Examples of translations produced by the **hierattn** multimodal transaltion system. **Baseline:** the output given the **Original** image-caption pair. NUM / PREP / NOUN / NP: The adversarial caption with the underlined replacement. MMT: the output of the **hierattn** system, given the adversarial sentence.

# Chapter 6

## Conclusion

Traditional machine translation algorithms typically only take into consideration linguistic contexts and they achieve this by learning representations only at the word and sentence levels. The aim of this thesis was to investigate and to make advances towards learning and understanding both *visual* as well as *linguistic* information for machine translation. We considered three scenarios to perform this

1. An *aligned* setting where the data is made up of triplets of L1 sentences, and their manual translation into L2, and an associated image;
2. *Sparse* settings where for an image and its associated description in L1, there is no data available in the L2 language and finally,
3. *Disjoint* settings where either the L1-L2 parallel corpus or the *image-textual* data is not aligned.

In this chapter, we not only summarise our main findings briefly, but we also to point out some of the limitations of our work and we give some pointers towards future directions.

We started the thesis by illustrating the benefits of using the visual modality for MT systems and progressed towards re-assessing the real contribution of additional modalities in these systems. Our multimodal NMT models mainly use image features extracted from pre-trained CNNs, specifically the VGG and the Residual

Networks. We use global image features for the three models **IMG<sub>2W</sub>**, **IMG<sub>E</sub>** and **IMG<sub>D</sub>** described in Chapters 3 and 4, and local image features for our proposed model, **NMT+explicit** in Chapter 5. Revisiting the state-of-the-art and the experimental results we presented, we observed in **Chapter 3** that the multimodal variants perform significantly better than their text-only counterparts when the alignment between visual and textual concepts is maintained. All multimodal models in the *aligned* scenarios using either global image features, consistently improved the translation quality of image descriptions in comparison to the NMT-baselines. Furthermore, our experiments with ensembled multimodal NMT models introduced in Calixto et al. (2016) show that these models can generate translations that compare favourably to multimodal models that use local image features. Thus from this chapter we obtain results that gives a strong indication that training with visual signals leads to improved results in an aligned setting such as ours. However, it remains inconclusive as to how general our findings are due to the limited number of languages we considered.

In **Chapter 4**, following a review of the aforementioned MMT systems, we proposed a methodology to generate synthetic *aligned* data in a low-resource data setting. The main objective of this chapter was to improve the models’ capacity to learn from *visual signals* even when the dataset is sparse. We specifically developed a joint data generation and training system for low-resource NMT, which yields benchmark results for unsupervised scenarios where abundant comparable data is available. Our results provide evidence that visual grounding can provide a useful inductive bias to improve translation performance. However, our methodologies in *sparse* data scenarios is still dependent on large amounts of out-of-domain comparable data, which is still far from what one can gather in a truly low-resource scenario. We hope to make our method more generalizable to more low-resourced language pairs in the future by exploring popular methods that involves both phrase-based extraction mechanisms and iterative back-translation between monolingual corpora.

The role of image in translation remains an open question at the time of writing

this thesis. Recently, Caglayan et al. (2019) show that multimodal systems are insensitive to images in general, however, masking entities from the source language sentence during training can help to overcome this problem. Along this direction, Gong et al. (2014) found that adding images as extra context to translation systems only has only marginal effect on translation performance and attributed their largest gains to adding more textual training samples. Although, the first to address this issue was Elliott (2018) who investigated the actual role of visual context in translation tasks. He introduced a measure of the image awareness of multimodal translation models and showed that additional visual modality is not necessarily used by demonstrating that the performance of a system did not change when it was evaluated with randomly selected images. These findings raise new questions about how to model the visual input in multimodal translation systems.

A natural extension to these works was to conduct experiments with both visual and textual adversaries in order to understand the role of both modalities to such systems. To the best of our knowledge, until this, no extensive studies have been done to understand the role of each modalities for MMT in a systematic manner. Thus in **Chapter 5** we studied the potential contribution of both modalities for the machine translation task in a *disjoint* data-setting. We introduced hard-negative adversarial samples in the text domain and studied the performance of existing frameworks. Our hard-negative adversarial sentences retained most aspects of the original sentence while depicting a completely unrelated scene. We also probed the visual awareness of these models by exposing them to random sample images.

Our results indicated that the systems are primarily performing text-based translations, which is supported by the observation that the visual adversaries did not harm the system performance as much as their textual counterparts. We also obtained new insights on the limitations of the current dataset (Elliott et al., 2016) as well as the state-of-the-art frameworks (Libovický and Helcl, 2017; Caglayan et al., 2017, 2016). Concurrent to our work is work published in (Caglayan et al., 2019) who show that the visual modality can be used to recover the masked tokens

in the source sentences. However, entity omissions does not harm the alignments across modalities as much as the hard-negative adversaries and that explains why the models were not particularly successful in translating adversarial noises.

We proposed a model that integrates explicit linguistic knowledge to help the model make reasonable inferences if tested on images it has never seen before. The result obtained from the first run of the experiment with noun-heads highlighted that explicit alignment can help to immune MMT models against any kind of input errors in the source. For future work, we want to introduce more knowledge bases and human priors to maintain the semantic consistency in MMT pipelines. We will put more emphasis on the specific visual term in the image, aligning them with corresponding mention in the source data. Furthermore, we would also like to extend this approach to check whether the models are more sensitive to semantically incorrect input or to grammatically incorrect input. Another interesting direction for future work would be to devise models with a max-margin ranking loss that forces the model to distinguish important differences between the true image-text pair and the hard negatives (Huang et al., 2018).

The visual grounding approaches presented in the thesis involved extracting global visual features through separate image encoders and further to associate them with sentences. Global visual features have also been proven to perform strongly in various transfer learning scenarios for different downstream NLP applications such as visual question answering (Zhang et al., 2016) or visual semantic word embeddings (Gella et al., 2017; Kádár et al., 2017). Another approach for using image is described in Chapter 5 that involves learning latent alignments between sentence fragments – usually words – and image regions using similar framework presented in (Karpathy and Fei-Fei, 2015). Here the sentence and the local region features of images are separately encoded with a bidirectional RNN and from a pre-trained CNN respectively. Furthermore, a dot product between these two computes the region-word interactions. Local visual features or attention over specific image regions has also been used in a transfer learning image description generation task

(Xu et al., 2015), which is closely related to the task of multimodal machine translation. We believe that an interesting avenue for future work involves the application of local image descriptors, using more advanced attention mechanisms to compute region-word interactions, in a multi-step way (Nam et al., 2017; Huang et al., 2017). Progress towards learning better visually-grounded representations can be also be extended in multi-task learning strategies (Ruder, 2017).

A major limitation, however, of most of our experiments is that they are based on specific data configurations, where the same images are annotated with multiple languages (Elliott et al., 2016). The current version of the Multi30K data set probably does not contain many training samples where the models need to take the visual modality into account for translation (Elliott, 2018). Thus, generation of more *balanced* datasets (Goyal et al., 2017) that captures the joint semantics of both modalities— we think is an exciting future avenue for visually grounded and multilingual translation tasks.

# Bibliography

- Abdelali, A., Guzman, F., Sajjad, H., and Vogel, S. (2014). The amara corpus: Building parallel language resources for the educational domain. In *LREC*, volume 14, pages 1044–1054.
- Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., and Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.
- Apresjan, J. D. (1974). Regular polysemy. *Linguistics*, 12(142):5–32.
- Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations, ICLR 2015*, San Diego, California.
- Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.
- Barsalou, L. W., Simmons, W. K., Barbey, A. K., and Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in cognitive sciences*, 7(2):84–91.
- Belinkov, Y. and Bisk, Y. (2017). Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173*.
- Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.
- Benítez, J. M., Castro, J. L., and Requena, I. (1997). Are artificial neural networks black boxes? *IEEE Transactions on neural networks*, 8(5):1156–1164.
- Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., and Plank, B. (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442.

- Bharati, A., Chaitanya, V., Sangal, R., and Ramakrishnamacharyulu, K. (1995). *Natural language processing: a Paninian perspective*. Prentice-Hall of India New Delhi.
- Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., and Turchi, M. (2015). Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.
- Bojar, O., Diatka, V., Rychlý, P., Straňák, P., Tamchyna, A., and Zeman, D. (2014). Hindi-English and Hindi-only Corpus for Machine Translation. In *Proceedings of the Ninth International Language Resources and Evaluation Conference (LREC’14)*, Reykjavik, Iceland. ELRA, European Language Resources Association. in prep.
- Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S.-Y., Pascual, L., Pêcheux, M.-G., Ruel, J., Venuti, P., and Vyt, A. (2004). Cross-linguistic analysis of vocabulary in young children: Spanish, dutch, french, hebrew, italian, korean, and american english. *Child Development*, 75(4):1115–1139.
- Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. *Computational linguistics*, 16(2).
- Caglayan, O., Aransa, W., Bardet, A., García-Martínez, M., Bougares, F., Barrault, L., Masana, M., Herranz, L., and Van de Weijer, J. (2017). Lium-cvc submissions for wmt17 multimodal translation task. *arXiv preprint arXiv:1707.04481*.
- Caglayan, O., Aransa, W., Wang, Y., Masana, M., García-Martínez, M., Bougares, F., Barrault, L., and van de Weijer, J. (2016). Does multimodality help human and machine for translation and image captioning? In *Proceedings of the First Conference on Machine Translation*, pages 627–633.
- Caglayan, O., Madhyastha, P., Specia, L., and Barrault, L. (2019). Probing the need for visual context in multimodal machine translation. *arXiv preprint arXiv:1903.08678*.
- Calixto, I., de Campos, T., and Specia, L. (2012). Images as context in statistical machine translation. In *The 2nd Annual Meeting of the EPSRC Network on Vision & Language (VL’12)*, Sheffield, UK. EPSRC Vision and Language Network.

- Calixto, I., Elliott, D., and Frank, S. (2016). DCU-UvA Multimodal MT System Report. In *Proceedings of the First Conference on Machine Translation*, pages 634–638.
- Calixto, I., Liu, Q., and Campbell, N. (2017a). Doubly-Attentive Decoder for Multimodal Neural Machine Translation. In *Proceedings of the 55th Conference of the Association for Computational Linguistics: Volume 1, Long Papers*, Vancouver, Canada (Paper Accepted).
- Calixto, I., Liu, Q., and Campbell, N. (2017b). Incorporating global visual features into attention-based neural machine translation. *CoRR*, abs/1701.06521.
- Cauchy, A. (1847). Méthode générale pour la résolution des systemes d’équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538.
- Chen, X., Fang, H., Lin, T., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. (2015). Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325.
- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014a). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar.
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014b). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- Cho, K., van Merriënboer, B., Gulcehre, C., Bougares, F., Schwenk, H., and Bengio, Y. (2014c). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*.
- Cohen, W. W., Schapire, R. E., and Singer, Y. (1998). Learning to order things. In *Advances in Neural Information Processing Systems*, pages 451–457.
- Dai, A. M. and Le, Q. V. (2015). Semi-supervised sequence learning. In *Advances in neural information processing systems*, pages 3079–3087.
- Daiber, J., Quiroz, L., Wechsler, R., and Frank, S. (2015). Splitting compounds by semantic analogy. In Haji269, J. and Branco, A., editors, *Proceedings of the 1st Deep Machine Translation Workshop*, pages 20–28. Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics.

- Daumas, M. (1965). Les machines à traduire de georges artsrouni. *Revue d'histoire des sciences et de leurs applications*, 18(3):283–302.
- Denkowski, M. and Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. (2014). Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655.
- Ebrahimi, J., Lowd, D., and Dou, D. (2018). On adversarial examples for character-level neural machine translation. *arXiv preprint arXiv:1806.09030*.
- Elliott, D. (2018). Adversarial evaluation of multimodal machine translation. In *Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium.
- Elliott, D., Frank, S., Barrault, L., Bougares, F., and Specia, L. (2017). Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 215–233, Copenhagen, Denmark.
- Elliott, D., Frank, S., Sima'an, K., and Specia, L. (2016). Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*.
- Elliott, D. and Kádár, A. (2017). Imagination improves multimodal translation. *arXiv preprint arXiv:1705.04350*.
- Frank, S., Elliott, D., and Specia, L. (2018). Assessing multilingual multimodal image description: Studies of native speaker preferences and translator choices. *Natural Language Engineering*, 24(3):393–413.
- Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al. (2013). Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129.
- Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., and Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.

- Gal, Y. and Ghahramani, Z. (2016). A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. In *Advances in Neural Information Processing Systems, NIPS*, pages 1019–1027, Barcelona, Spain.
- Gella, S., Sennrich, R., Keller, F., and Lapata, M. (2017). Image pivoting for learning multilingual multimodal representations. *arXiv preprint arXiv:1707.07601*.
- Gers, F. A., Schmidhuber, J., and Cummins, F. (1999). Learning to forget: Continual prediction with lstm.
- Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587.
- Glenberg, A. M. and Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of memory and language*, 43(3):379–401.
- Gong, Y., Wang, L., Hodosh, M., Hockenmaier, J., and Lazebnik, S. (2014). Improving image-sentence embeddings using large weakly annotated photo collections. In *European Conference on Computer Vision*, pages 529–545. Springer.
- Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. (2017). Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.
- Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2013). Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41.
- Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2014). Is machine translation getting better over time? In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 443–451.
- Grönroos, S.-A., Huet, B., Kurimo, M., Laaksonen, J., Merialdo, B., Pham, P., Sjöberg, M., Sulubacak, U., Tiedemann, J., Troncy, R., et al. (2018). The memad submission to the wmt18 multimodal translation task. *arXiv preprint arXiv:1808.10802*.
- Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et al. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354–377.

- Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.
- He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385*.
- Hitschler, J., Schamoni, S., and Riezler, S. (2016). Multimodal pivots for image caption translation. *arXiv preprint arXiv:1601.03916*.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Hodosh, M. and Hockenmaier, J. (2016). Focused evaluation for image description with binary forced-choice tasks. In *Proceedings of the 5th Workshop on Vision and Language*, pages 19–28.
- Honnibal, M. and Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378.
- Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.
- Huang, J., Li, Y., Ping, W., and Huang, L. (2018). Large margin neural language model. *arXiv preprint arXiv:1808.08987*.
- Huang, P.-Y., Liu, F., Shiang, S.-R., Oh, J., and Dyer, C. (2016). Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 639–645.
- Huang, Y., Wang, W., and Wang, L. (2017). Instance-aware image and sentence matching with selective multimodal lstm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2310–2318.
- Hutchins, J. (2004). Two precursors of machine translation: Artsrouni and trojanskij. *International Journal of Translation*, 16(1):11–31.
- Hutchins, J. (2007). Machine translation: A concise history. *Computer aided translation: Theory and practice*, 13:29–70.
- Jabri, A., Joulin, A., and van der Maaten, L. (2016). Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739.

- Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. (2017). Google’s multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Kádár, A., Chrupała, G., and Alishahi, A. (2017). Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43(4):761–780.
- Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709.
- Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.
- Khayrallah, H. and Koehn, P. (2018). On the impact of various types of noise on neural machine translation. *arXiv preprint arXiv:1805.12282*.
- Kiela, D., Conneau, A., Jabri, A., and Nickel, M. (2017). Learning visually grounded sentence representations. *arXiv preprint arXiv:1707.06320*.
- Kilgariff, A. (1997). I don’t believe in word senses. *Computers and the Humanities*, 31(2):91–113.
- King, G. W. and Wieselmann, I. L. (1956). Stochastic methods of machine translation. *International Telemeter Corporation*.
- Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kiros, R., Salakhutdinov, R., and Zemel, R. (2014a). Multimodal neural language models. In *International Conference on Machine Learning*, pages 595–603.
- Kiros, R., Salakhutdinov, R., and Zemel, R. S. (2014b). Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.
- Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Torralba, A., Urtasun, R., and Fidler, S. (2015). Skip-thought vectors. corr abs/1506.06726.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007a). Moses: Open source

- toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007b). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual meeting of Association for Computational Linguistics*, pages 177–180.
- Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.
- Kunchukuttan, A., Mehta, P., and Bhattacharyya, P. (2017). The iit bombay english-hindi parallel corpus. *arXiv preprint arXiv:1710.02855*.
- Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Mallocci, M., Duerig, T., et al. (2018). The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*.
- Landau, B., Smith, L., and Jones, S. (1998). Object perception and object naming in early development. *Trends in cognitive sciences*, 2(1):19–24.
- LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995.
- LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.
- Libovický, J. and Helcl, J. (2017). Attention strategies for multi-source sequence-to-sequence learning. *arXiv preprint arXiv:1704.06567*.
- Lin, C.-Y. and Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

- Luong, M., Le, Q. V., Sutskever, I., Vinyals, O., and Kaiser, L. (2016). Multi-task sequence to sequence learning. In *ICLR*.
- Luong, M.-T., Kayser, M., and Manning, C. D. (2015). Deep neural language models for machine translation. In *CoNLL*.
- Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., and Murphy, K. (2016). Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.
- Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., and Yuille, A. (2014). Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*.
- Michel, P. and Neubig, G. (2018). Mtnt: A testbed for machine translation of noisy text. *arXiv preprint arXiv:1809.00388*.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- Miller, G. (1998). *WordNet: An electronic lexical database*. MIT press.
- Mnih, A. and Hinton, G. (2009). A scalable hierarchical distributed language model. In *NIPS*.
- Mnih, A. and Teh, Y. W. (2012). A fast and simple algorithm for training neural probabilistic language models. In *ICML*.
- Morin, F. and Bengio, Y. (2005). Hierarchical probabilistic neural network language model. In *AISTATS*.
- Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.
- Nam, H., Ha, J.-W., and Kim, J. (2017). Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 299–307.
- Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Association for Computational Linguistics.
- Och, F. J. and Ney, H. (2000). Giza++: Training of statistical translation models.

- Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL ’02, pages 311–318, Philadelphia, Pennsylvania.
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.
- Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 2641–2649. IEEE.
- Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature reviews neuroscience*, 6(7):576.
- Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. URL <https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/languageunderstandingpaper.pdf>.
- Rahman, W., Hasan, M. K., Zadeh, A., Morency, L.-P., and Hoque, M. E. (2019). M-bert: Injecting multimodal information in the bert structure. *arXiv preprint arXiv:1908.05787*.
- Rajendran, J., Khapra, M. M., Chandar, S., and Ravindran, B. (2015). Bridge correlational neural networks for multilingual multimodal representation learning. *arXiv preprint arXiv:1510.03519*.
- Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. (2016). Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*.
- Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.
- Sanderson, M. (1994). Word sense disambiguation and information retrieval. In *SIGIR'94*, pages 142–151. Springer.
- Schuster, M. and Nakajima, K. (2012). Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.
- Schwenk, H. (2007). Continuous space language models. *Computer Speech and Languages*, 21(3):492–518.
- Sennrich, R., Haddow, B., and Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany.
- Shi, H., Mao, J., Xiao, T., Jiang, Y., and Sun, J. (2018). Learning visually-grounded semantics from contrastive adversarial samples. *arXiv preprint arXiv:1806.10348*.
- Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*.
- Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas, AMTA*, pages 223–231, Cambridge, MA, USA.
- Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., and Ng, A. Y. (2014). Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association of Computational Linguistics*, 2(1):207–218.
- Specia, L., Frank, S., Sima'an, K., and Elliott, D. (2016a). A shared task on multi-modal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.

- Specia, L., Frank, S., Sima'an, K., and Elliott, D. (2016b). A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 543–553.
- Stolcke, A. (2002). Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.
- Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *NIPS*.
- Taylor, S. L., Mahler, M., Theobald, B.-J., and Matthews, I. (2012). Dynamic units of visual speech. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 275–284. Eurographics Association.
- Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *LREC*, volume 2012, pages 2214–2218.
- Torabi, A., Pal, C., Larochelle, H., and Courville, A. (2015). Using Descriptive Video Services to Create a Large Data Source for Video Annotation Research. *ArXiv e-prints*.
- Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. (2017). Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*.
- Turney, P. D., Neuman, Y., Assaf, D., and Cohen, Y. (2011). Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 680–690. Association for Computational Linguistics.
- Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *SSW*, 125.
- van Miltenburg, E. (2016). Stereotyping and bias in the flickr30k dataset. *arXiv preprint arXiv:1605.06083*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Vaswani, A., Zhao, Y., Fossum, V., and Chiang, D. (2013). Decoding with large-scale neural language models improves translation. In *EMNLP*.

- Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.
- Vendrov, I., Kiros, R., Fidler, S., and Urtasun, R. (2015). Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*.
- Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., and Saenko, K. (2014). Translating Videos to Natural Language Using Deep Recurrent Neural Networks. *ArXiv e-prints*.
- Vilar, D., Xu, J., Luis Fernando, D., and Ney, H. (2006). Error analysis of statistical machine translation output. In *LREC*, pages 697–702.
- Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2014). Show and Tell: A Neural Image Caption Generator. *ArXiv e-prints*.
- Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3156–3164. IEEE.
- Wu, H. and Wang, H. (2007). Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21(3):165–181.
- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057.
- Yagcioglu, S., Erdem, E., Erdem, A., and Cakici, R. (2015). A distributed representation based query expansion approach for image captioning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 106–111. Association for Computational Linguistics.
- Yoo, K. M., Shin, Y., and Lee, S.-g. (2017). Improving visually grounded sentence representations with self-attention. *arXiv preprint arXiv:1712.00609*.
- Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

- Zen, H., Braunschweiler, N., Buchholz, S., Gales, M. J., Knill, K., Krstulovic, S., and Latorre, J. (2012). Statistical parametric speech synthesis based on speaker and language factorization. *IEEE transactions on audio, speech, and language processing*, 20(6):1713–1724.
- Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D., and Parikh, D. (2016). Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5014–5022.
- Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.
- Zwicky, A. M. (1985). Heads. *Journal of linguistics*, 21(1):1–29.