Incorporating Visual Information into Neural Machine Translation

Iacer Coimbra Alves Cavalcanti Calixto

B.Sc., M.Sc.

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

Doctor of Philosophy (Ph.D.)

to the

Dublin City University
School of Computing

Supervisors:
Prof. Qun Liu
Prof. Nick Campbell (Trinity College Dublin, Ireland)

August 2017

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Ph.D. 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.

Signed:

(Candidate) ID No.: 14210086

Date: August 25, 2017

Contents

List of Figures

List of Tables

Incorporating Visual Information into Neural Machine Translation

Iacer Coimbra Alves Cavalcanti Calixto

Abstract

In this work, we study different ways to enrich Machine Translation (MT) models using information obtained from images. Specifically, we propose different models to incorporate images into MT by transferring learning from pre-trained convolutional neural networks (CNN) trained for classifying images. We use these pre-trained CNNs for image feature extraction, and use two different types of visual features: *global* visual features, that encode an entire image into one single real-valued feature vector; and *local* visual features, that encode different areas of an image into separate real-valued vectors, therefore also encoding spatial information. We first study how to train embeddings that are both multilingual and multi-modal, and use global visual features and multilingual sentences for training. Second, we propose different models to incorporate global visual features into state-of-the-art Neural Machine Translation (NMT): *(i)* as words in the source sentence, *(ii)* to initialise the encoder hidden state, and *(iii)* as additional data to initialise the decoder hidden state. Finally, we put forward one model to incorporate local visual features into NMT: *(i)* a NMT model with an independent visual attention mechanism integrated into the same decoder Recurrent Neural Network (RNN) as the source-language attention mechanism. We evaluate our models on the Multi30k, a publicly available, general domain data set, and also on a proprietary data set of product listings and images built by eBay Inc., which was made available for the purpose of this research. We report state-of-the-art results on the publicly available Multi30k data set. Our best models also significantly improve on comparable phrase-based Statistical MT (PBSMT) models trained on the same data set, according to widely adopted MT metrics.

Acknowledgments

I sincerely thank all my colleagues in the ADAPT Centre for their invaluable help. Without your support—and I mean *all of you*—this work would not have been at all possible, and I certainly would not have the knowledge I have today. I would like to specifically thank my supervisors, profs. Qun Liu and Nick Campbell, and my thesis examiners, profs. Gareth Jones and Marie-Francine Moens. My most sincere thank you to all of you.

Iacer Calixto has received funding from Science Foundation Ireland in the ADAPT Centre for Digital Content Technology (www.adaptcentre.ie) at Dublin City University funded under the SFI Research Centres Programme (Grant 13/RC/2106) co-funded under the European Regional Development Fund and the European Union Horizon 2020 research and innovation programme under grant agreement 645452 (QT21).

Chapter 1

Introduction

Machine Translation (MT) is among the most difficult tasks in Natural Language Processing (NLP) and deals with the automatic translation of text and/or speech between two natural languages. The reasons why MT is difficult are numerous. First, human languages are in constant evolution and have both an immediate dimension, e.g. how different linguistic phenomena integrate the language we use in our everyday lives, and a historic dimension, e.g. how does language evolve in time (de Saussure (1966) referred to the *synchronic* and *diachronic* dimensions of language, respectively). Second, novel words are constantly being created and integrated into a language's lexicon (i.e. neologisms), and languages are constantly borrowing words and expressions from other languages (i.e. loanwords). Third, frequently used words can often be ambiguous, e.g. *pen* as a writing device or as an enclosure in which one keeps animals. Furthermore, not only can a noun take two or more different meanings depending on the context, but one same surface form of a word can take entirely different syntactic functions (e.g. *pen* as the verb denoting the action of putting or keeping an animal in a pen). These are all examples of difficulties that MT models must address and account for if they are to successfully model all the linguistic phenomena in natural languages.

Multi-modal MT is a relatively new research topic only recently addressed by the MT community in the form of a shared task (Specia et al., 2016), where the

goal is to propose MT models that use image information to better translate image descriptions. The main idea is to improve the translations of ambiguous terms that could in principle be disambiguated by an image (e.g. an image of a jaguar could probably disambiguate whether a certain mention of *jaguar* means the car brand or the animal species). There are many different conceivable ways to extract visual information from images, as well as different MT architectures that one can incorporate visual information into.

Arguably, the two major data-driven MT paradigms are Statistical Machine Translation (SMT) and the more recent Neural Machine Translation (NMT). In SMT, a statistical model makes use of different features extracted from parallel text, such as word translation and language model (LM) probabilities. These features are typically combined in a log-linear model, which is in turn optimised to maximise some translation quality metric on held-out development data. In NMT, an Artificial Neural Network (ANN) is trained end-to-end to model the entire translation process, from reading a raw sentence in a source language to generating a translation in a target language. Similarly, a held-out development set is used for model selection. In NMT there is no need for specific features to be extracted, i.e. the model learns these features directly from the data.

In fact, the recent interest in ANNs is not at all restricted to the MT field. In general, there has been a shift of interest from models that learn specific, hand-crafted features, in favour of ANNs that model an entire problem end-to-end, as is the case of SMT versus NMT. We highlight a similar trend regarding the extraction of visual information from images. There have been many widely adopted features engineered to capture important visual information from images, such as histograms or depth, for example. These features could than be applied to the Computer Vision (CV) tasks at hand, such as performing face recognition or image classification. Nonetheless, similarly to what happened to statistical and neural MT, end-to-end neural networks, more specifically Convolutional Neural Networks (CNNs), have been successfully applied to these CV tasks, for instance to model the entire image

classification process from receiving a raw image to classifying this image into a set of possible classes, e.g. "Persian cat". Again, there is no need to compute specific image features prior to the task, since the model should learn these feature directly from the data.

State-of-the-art MT systems are almost always data-driven models, and recently NMT models have proven successful in achieving state-of-the-art results in a large number of language pairs (Bojar et al., 2016a). Similarly, image classification and object detection models have been dominated by CNNs (Russakovsky et al., 2015). One point the best-performing NMT and image classification neural networks have in common is the fact they are all *deep* neural networks, meaning that they have many layers between receiving an input and generating an output. Deep neural networks, also popularly referred to as *deep learning* models, have brought breakthroughs to many different learning tasks by structuring computational models in many subsequent layers, so that the representations learnt have multiple levels of abstraction and can model complex tasks (Lecun et al., 2015).

The research we report in this thesis is conducted within the ADAPT Centre[1], a research institute that aims to bridge the gap between the academia and the industry in research involving intelligent content. The ADAPT Centre has many important industry partners which have use-cases that can benefit from multi-modal MT. We specifically apply our research to solve an MT requirement raised by eBay Inc.[2], a large multinational company in the e-commerce area which is one of the world leaders in the field. In simple words, eBay provides a platform where users can list, buy and sell products. They have a strong requirement to make products accessible regardless of the customer's native language or country of origin, and they leverage MT to address that requirement. Moreover, most of their product listings and other product-related information are user-generated, meaning that the automatic processing of such data can suffer from many difficulties derived from these listings'

[1] `http://www.adaptcentre.ie`
[2] `http://www.ebay.com`

specialised language and grammar. In addition to that, a high percentage of user-generated content from non-business sellers is created by non-native speakers of the language themselves. eBay's placement as a world-level company with a strong use-case for multi-modal MT combined with the fact that they are an important industry partner in the ADAPT Centre enabled us to address their use-case as part of the research objectives in this work.

In this thesis, we introduce different models that incorporate visual information into MT. In the first part of our work, we investigate how to train discriminative models to distinguish between positive and negative pairs of sentences and images, and evaluate these in different tasks involving multi-modal reasoning, such as sentence-image ranking. In this model, we use global image features—these are features that encode an entire image as a whole, without differentiating between different objects or portions of the image—extracted using CNNs pre-trained for the task of image classification. We later use these multi-modal models to compute features used to re-rank n-best lists generated by MT systems.

Later on, we investigate how to incorporate image features directly into a state-of-the-art neural MT framework. We specifically study how to incorporate global and local image features—these are features that encode different areas or parts of an image separately—into the encoder–decoder NMT framework (Kalchbrenner and Blunsom, 2013; Cho et al., 2014b; Sutskever et al., 2014), in both cases using publicly available pre-trained CNNs for image feature extraction. The reason we use pre-trained CNNs and choose not to train a model from scratch is twofold: *(i)* first, because these models are known to perform well in different transfer learning scenarios (Lazaridou et al., 2015; Zhang et al., 2016); *(ii)* second, because training an image encoder from scratch with our NMT models would require prohibitively large amounts of data. In these experiments, we study how different multi-modal NMT models perform when applied to two different scenarios. In one scenario, we use a standard image description data set, where models are trained to translate image descriptions with and without the corresponding images available. In the

second scenario, these same models are used to translate real-world eBay product listings with and without including the corresponding product images as part of the training data. By assessing these two scenarios, we hope to provide a comprehensive overview of the use-cases in which a multi-modal MT model is useful.

1.1 Motivation

Previous research has already indicated that images can bring useful information to MT (Calixto et al., 2012; Hitschler et al., 2016). Additionally, MT models suffer from an obvious limitation since the meaning of a word is derived entirely from connections to other words, i.e. they do not take extra-linguistic modalities into account. In short, word forms are arbitrary symbols defined in relation to other words, and therefore lack *grounding* (Harnad, 1990; Glenberg and Robertson, 2000). We believe that incorporating images into MT is a step towards *visually grounding* translations of image descriptions. In that tone, we propose to use both global and local image features in NMT, discussed further in Section 2.2.1.1, to visually ground translations. More precisely, global features encode an entire image as a whole without differentiating between different objects or portions of the image, whereas local features encode different areas or parts of an image separately, therefore encoding spatial information. An example of global features for an image used in this work could be a vector in $\mathbb{R}^{4096}$, where all the dimensions of the vector encode information about the whole image. Likewise, an example of local features for one same image could be a 3-tensor in $\mathbb{R}^{14\times14\times1024}$, where the two first dimensions (14×14) denote a position in the image and the last dimension contains features for that portion of the image. By encoding different areas of an image in different feature vectors, local features naturally integrate spatial information.

Global image features have been successfully used in monolingual image description generation models based on the encoder–decoder framework (Elliott et al., 2015; Vinyals et al., 2015). That alone is a good indication that they might be also

well-suited for being exploited in NMT. Also, global image features are orders of magnitude smaller than local features, and are therefonre less complex and easier to integrate in a model. The "simplicity" of global image features and their relative small size compared to local image features is also one main reason to work with them in our experiments.

Local visual features are much larger and more complex than global features. As discussed above, whereas the global features we use in this work contain $4,096$ distinct real-valued features for each image, local features to encode one same image will consist of $14 \times 14 \times 1,024 = 200,704$ real-valued features. In other words, local features for one image use the same memory as global features for 49 different images. Even though bigger does not necessarily mean better, the fact that these features encode *spatial information* is worth exploring.

Neural network models are often criticised for being difficult to interpret. Particularly, to feed a trained ANN model an input and to be able to explain *why* does such input produce a particular output—to know exactly what neurons of the model contributed to generate that particular output—is often a non-trivial endeavour. However, even though fine-grained knowledge about how an ANN model works might be unpractical at times, especially in deep neural models, *post-hoc interpretability* (Lipton, 2016) is a way to try to explain how a model works by looking at particular examples produced by it or by visualising its learned representations. One highly desirable side-effect of using local features is being able to *visualise* what parts of an image contribute to the generation of specific words in a translation, which we discuss further in Chapter 7.

1.2 Research Questions

We propose to answer the following research questions in this thesis:

> **(RQ1)** *Can we use multi-modal discriminative models to improve the translation of image descriptions?*

(RQ2) *Given that there is a large number of standard text-only MT corpora, can multi-modal MT models effectively exploit this additional text-only data and provide state-of-the-art performance?*

(RQ3) *How do multi-modal MT models compare to text-only MT models when translating user-generated product listings?*

To the best of our knowledge, incorporating images into MT has just recently been addressed by the MT research community. For instance, the first shared task in multi-modal MT took place in the First Conference on Machine Translation (WMT 2016) in August, 2016. For this reason, designing an experimental setting to try to address research questions **(RQ1)**–**(RQ3)** requires defining some basic points first:

(i) what is the exact problem we wish to address;

(ii) what is the input and output data our models will require;

(iii) what existing corpora/image dataset to use in our work;

(iv) what evaluation metrics to adopt.

From now on, we shall use the term *corpora* to refer not only to text collections in its original, traditional meaning, but also to *multilingual text collections that include images* (as in *(iii)*). Unless expressly stated otherwise, the term *image description datasets* and *corpora* are used interchangeably.

(i)–(ii) For addressing *(i)* and *(ii)*, we begin to delineate the problem we wish to address by first defining the inputs and outputs available to address that problem. Having said that, from a data perspective there are two main scenarios we address in this thesis: first, an *optimal data availability* scenario and, second, a *scarse data availability* scenario. We will explain further what we mean by each of these scenarios and how we tackle each of them.

In traditional data-driven MT, training and test data consist of parallel sentence pairs. At training time, these sentence pairs are used to train a model whereas at

test time we use the sentence in the target language, i.e. the reference translation, in order to automatically evaluate the translations generated by the model, typically using one or more automatic metrics, such as BLEU (Papineni et al., 2002) or METEOR (Denkowski and Lavie, 2014). Publicly available corpora to train MT models are typically very large,[3] except for some pairs involving low-resourced languages.

First, our models must obviously be able to efficiently exploit the *optimal data availability* scenario: a situation in which there are not only parallel sentence pairs but also images available at both training and test stages. This means that multi-modal data are available at all stages and in large amounts, i.e., there are triples ⟨`image`, `source`, `target`⟩, meaning one image and one sentence pair in the source language and target languages *where these sentences describe the image*. At test time, we address the scenario where we decode a test set containing tuples with ⟨`image`, `source`, `target`⟩, i.e. each test instance contains one sentence in the source language, its reference translation and one associated image. The model should translate the source sentence into the target language *while taking the associated image into consideration* and will be evaluated against one or more automatic metrics, as already pointed out for the traditional MT scenario. In conclusion, in this scenario there is enough multi-modal and multilingual training data, and no need to resort to any form of data augmentation.

Secondly, our models must also be able to cope with the likely situation in which there is much more text-only MT training data available than multi-modal multilingual training data. In other words, the models we propose must be able to be efficiently trained on large text-only general-domain MT corpora and *fine-tuned* on typically much smaller in-domain multi-modal and multilingual data. This is a much more likely scenario and, typically, even if there are large amounts of multi-modal and multilingual training data available, we might also be interested in exploiting some very large text-only MT corpora when they exist. In general, being able to

[3]See for instance the WMT 2016 website `http://www.statmt.org/wmt16/translation-task.html`, where publicly available corpora can be downloaded to train MT models between many different language pairs.

exploit different types of data at training time, e.g. data requiring less supervision, is a desirable feature for a machine learning model.

(iii) To evaluate our work, we need corpora that are both *multilingual* and *multi-modal*: multilingual because for training and evaluating MT models we need text in at least two languages; multi-modal because we need images associated with this text so as to evaluate how we can improve translation quality by exploiting them.

Having said that, we initially shortlisted two datasets which after much consideration were deemed inadequate. One of these datasets is the Wikipedia images corpus released by Calixto et al. (2012), which contains images and *captions* in English, German and French, as well as machine translated captions. We note that image captions and image descriptions are not the same thing: the former are texts associated with an image that typically contain information that cannot be seen in the image, whereas the latter describe the actual contents of an image (Bernardi et al., 2016). The positive side of this corpus is its size—specially for the English–German and English–French language pairs—, since it contains overall 191,594 images and bilingual captions. The negative side is that bilingual captions are not necessarily translations of each other; there could be strong paraphrasing or even large chunks of the caption in one language that have no correspondence in the other language.

A second corpus we considered is the *Wikipedia ImageCLEF 2010*.[4] In general, the same negative points that apply to the Wikipedia images corpus of Calixto et al. (2012) also apply to the Wikipedia ImageCLEF 2010 image corpus. In practice, bilingual captions found in these two datasets are not good enough to be used in training and evaluating MT models.

To the best of our knowledge, there is but one publicly available dataset built for training and evaluating multi-modal MT systems, consisting of Flickr images and multi-lingual descriptions.[5] We also have access to one more dataset consisting

[4]This corpus is freely distributed for research and can be downloaded in the ImageCLEF Wikipedia Image Retrieval website: `http://imageclef.org/2010/wiki`.

[5]`https://www.flickr.com/`.

of product titles and associated product images obtained in an agreement between eBay Inc.[6] and the ADAPT Centre. This dataset is not publicly available and has been released only for the purposes of this research. We will describe both these datasets in detail in Chapter 3.

(iv) So as to address *(iv)*, one important point is how to evaluate the impact of multi-modal data on final translation quality. All models, traditional text-only MT or multi-modal MT models, can be evaluated using similar metrics since in all cases we are interested in improving *translation quality*, regardless of whether the model uses images or not. We believe that using a combination of *automatic metrics* already adopted by the MT community and manual, *qualitative evaluation* is best to address this issue. Amongst several automatic metrics adopted by the MT community, we choose to use four in our work. The first three, described below, are used because of their wide adoption and to be able to compare to existing previous work:

- **BLEU** (Papineni et al., 2002) – a widely adopted measure of weighted mean average precision;
- **Translation Edit Rate (TER)** (Snover et al., 2006) – which is an inexpensive measure that correlates fairly well with human judgements;
- **METEOR** (Lavie and Agarwal, 2007; Denkowski and Lavie, 2014) – which accounts for both precision and recall and also incorporates more complex linguistic phenomena, such as synonymy and paraphrasing.

The next metric was chosen because it is character-based, more recent, and have also shown a high correlation with human judgements according to previous years' WMT automatic metrics evaluation shared tasks (Bojar et al., 2015, 2016b):

- **chrF** (Popović, 2016) – a character-level metric that scores character n-grams with a recall bias.

[6] `http://www.ebay.com/`.

In addition to these metrics a manual, qualitative human evaluation could highlight aspects of the translations that might remain obscure by looking at automatic metrics only. For that reason, we decided to conduct an error analysis of translations generated by different models, to allow for a more in-depth analysis.

1.3 Road Map

In the remaining chapters in this thesis, we address the research questions raised in Section 1.2. We now introduce the topics discussed in each chapter.

Chapter 2 In Chapter 2, we provide the background necessary to make sense of the work presented in this thesis. Since in our work we incorporate research conducted in two different research areas, we provide a separate background for each of these main areas: Machine Translation and Computer Vision.

We start by providing a brief history of the MT field, and also provide a high-level introduction to artificial neural networks and how they can be typically applied to text processing. We then move on to introduce the two major MT paradigms discussed in this work: statistical MT and neural MT.

We continue by presenting a high-level introduction of the CV field. We discuss the main evaluation campaign in CV and briefly introduce CNNs, the technological backbone that enables virtually all the state-of-the-art methods in the area. We discuss the two specific CNNs used in our work for image feature extraction, specifically the VGG and the Residual Networks. We finish this part by discussing the research done in the area of image description generation, which is closely related to multi-modal MT.

Finally, we discuss some relevant related multi-modal MT work, which is based on the results of the first WMT multi-modal MT shared task (Specia et al., 2016).

Chapter 3 In Chapter 3, we introduce the different data sets we work with in this thesis. These emcompass some standard MT corpora, a publicly available image

description data set and a proprietary product listings data set, obtained in an agreement between the ADAPT Centre and eBay Inc. In this chapter, we also discuss some important aspects of the data sets, including a small-scale human evaluation of the eBay product listings.

Chapter 4 In Chapter 4, we introduce the notation we use throughout our work, as well as a text-only attention-based neural MT model. This is one of the baseline systems to which we compare our multi-modal models, as well as the model we use as inspiration to derive our multi-modal NMT implementations.

Chapter 5 In Chapter 5, we address research question RQ1: *Can we use multi-modal discriminative models to improve the translation of image descriptions?* We put forward a discriminative model that makes use not only of images and their English descriptions (multi-modal), but can also include descriptions in other languages whenever these are available (multilingual). We compare this model with a monolingual baseline trained on images and English descriptions only, and evaluate them in different multi-modal reasonining tasks. We show that using additional multilingual descriptions helps and that our model outperforms different baselines in the tasks of image-sentence ranking, semantic textual similarity and neural MT, where we introduce experiments to study the application of our discriminative model to re-rank n-best lists generated by different NMT models. We find that the discriminative model proposed can be efficiently used to re-rank n-best lists, and that it provides consistent gains regardless of the quality of the NMT model used to generate the n-best lists.

Chapter 6 In Chapter 6 we start to address research questions RQ2: *Given that there is a large number of standard text-only MT corpora, can multi-modal MT models effectively exploit this additional text-only data and provide state-of-the-art performance?*, and RQ3: *How do multi-modal MT models compare to text-only MT models when translating user-generated product listings?* We introduce different

multi-modal NMT models that use global image features in different ways and evaluate these models in English–German and German–English translation. We compare our models to a NMT and a PBSMT baseline, and find that using global image features leads to consistent improvements in translation quality. We also conduct ablative experiments where we measure how different models perform when pre-trained on back-translated multi-modal data, obtaining improvements with multi-modal models in all scenarios. Finally, we conduct an error analysis of the translations generated by our different systems and corroborate the findings obtained with our quantitative evaluation.

We use the same models in a series of experiments to translate eBay's product listings. Results in this scenario were unexpected, in that there was no statistically significant difference between the translations generated with multi-modal and text-only models. We finish the chapter by conjecturing some possible reasons for that outcome.

Chapter 7 In Chapter 7 we again address research questions RQ2: *Given that there is a large number of standard text-only MT corpora, can multi-modal MT models effectively exploit this additional text-only data and provide state-of-the-art performance?*, and RQ3: *How do multi-modal MT models compare to text-only MT models when translating user-generated product listings?* We introduce a multi-modal NMT model that uses local image features and evaluate it in English–German and German–English translation.

We compare this multi-modal model to a NMT and a PBSMT baseline, and find that using local image features leads to consistent improvements in translation quality for the translation of image descriptions. We again conduct ablative experiments where we measure how different models perform when pre-trained on additional text-only data, and on back-translated multi-modal data, obtaining improvements with our multi-modal model in all scenarios.

We use the same model in a series of experiments to translate product listings

from eBay. We first report experiments where we compare our multi-modal NMT model with a text-only PBSMT and NMT baseline trained on the same data, as well as use the multi-modal NMT model to re-rank n-best lists generated by a PBSMT baseline. We conduct a human evaluation of the translations generated by different systems, and find that humans prefer PBSMT output in this use-case. However, we were able to consistently improve translations by using the multi-modal NMT model to re-rank n-best lists generated with a PBSMT system.

Chapter 8 Chapter 8 concludes the thesis with general conclusions we draw from our experiments altogether. We also provide some future avenues for research.

1.4 Publications

Doubly-attentive multi-modal NMT I have co-authored one long paper where we describe the doubly-attentive model $\text{NMT}_{\text{SRC+IMG}}$ (described in Section 7.1). In this paper, we evaluate model $\text{NMT}_{\text{SRC+IMG}}$ when translating image descriptions for the Multi30k data set (discussed in Section 3):

- Calixto, I., Liu, Q., and Campbell, N. (2017b). Doubly-Attentive Decoder for Multi-modal Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924, Vancouver, Canada.

I have also co-authored one system description paper where we discussed our submission to the first multi-modal MT shared task (Specia et al., 2016),[7] where we used a preliminar version of the doubly-attentive model $\text{NMT}_{\text{SRC+IMG}}$ (described in Section 7.1) applied to translate image descriptions for the Multi30k data set (discussed in Section 3.1):

- Calixto, I., Elliott, D., and Frank, S. (2016). DCU-UvA Multimodal MT System Report. In *Proceedings of the First Conference on Machine Translation*,

[7] `http://www.statmt.org/wmt16/multimodal-task.html`

pages 634–638, Berlin, Germany.

Multi-modal NMT models with global visual features I have co-authored one long paper where we describe three models to integrate global visual features into NMT (described in Section 6.1): using an image as a word in the source sequence (IMG_W), to initialise the encoder hidden state (IMG_E), and as additional data to initialise the decoder hidden state (IMG_D).

- Calixto, I., Liu, Q., and Campbell, N. (2017c). Incorporating Global Visual Features into Attention-Based Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark.

I have also co-authored one system description paper where we report on our submission to the second multi-modal MT shared task[8] using an ensemble of different multi-modal NMT models (IMG_W, IMG_E, and IMG_D). These models were applied to translate image descriptions of the Multi30k data set (discussed in Section 3.1) and a specially curated test set based on the MSCOCO data set:

- Calixto, I., Chowdhury, K. D., and Liu, Q. (2017a). DCU System Report on the WMT 2017 Multi-modal Machine Translation Task. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark.

eBay data set I have co-authored three papers where we discussed experiments involving the eBay data set:

- Calixto, I., Stein, D., Matusov, E., Lohar, P., Castilho, S., and Way, A. (2017f). Using images to improve machine-translating e-commerce product listings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 637–643, Valencia, Spain;

[8] `http://www.statmt.org/wmt17/multimodal-task.html`

- Calixto, I., Stein, D., Matusov, E., Castilho, S., and Way, A. (2017e). Human evaluation of multi-modal neural machine translation: A case-study on e-commerce listing titles. In *Proceedings of the Sixth Workshop on Vision and Language*, pages 31–37, Valencia, Spain;

- Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., and Way, A. (2017). Is Neural Machine Translation the New State-of-the-Art? *Prague Bulletin of Mathematical Linguistics*, 10(8):109–120.

In the first paper, we discuss experiments where we use a baseline PBSMT model to generate n-best lists, and re-rank them using a multi-modal NMT model. In the second paper, we focus on the human evaluation of the translations obtained with PBSMT versus translations obtained with NMT models, both text-only and multi-modal ones. In the third paper, we study how PBSMT and NMT models compare when translating for different domains, and we include some experiments with user-generated product listings from eBay.

Multilingual Multi-Modal Embedding I have co-authored one long paper where we describe our multilingual multi-modal embedding (described in Section 5.1). We evaluate the embedding model in three different tasks, image-sentence ranking, semantic textual similarity and NMT, and find that it can be exploited with different degrees of success in the three tasks.

- Calixto, I., Liu, Q., and Campbell, N. (2017d). Multilingual Multi-modal Embeddings for Natural Language Processing. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, Varna, Bulgaria.

Other papers published and/or accepted for publication in peer-reviewed conferences Other publications I have co-authored during my PhD are:

- Ferreira, T. C., Calixto, I., Wubben, S., and Krahmer, E. (2017). Linguistic realisation as machine translation: Comparing different MT models for AMR-

to-text generation. In *Proceedings of the International Conference on Natural Language Generation*, Santiago de Compostela, Spain;

- Ganguly, D., Calixto, I., and Jones, G. F. (2016). Developing a Dataset for Evaluating Approaches for Document Expansion with Images. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portoroź, Slovenia;

- Štajner, S., Calixto, I., and Saggion, H. (2015). Automatic Text Simplification for Spanish: Comparative Evaluation of Various Simplification Strategies. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 618–626, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA;

- Hokamp, C., Calixto, I., Wagner, J., and Zhang, J. (2014). Target-Centric Features for Translation Quality Estimation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 329–334, Baltimore, Maryland, USA.

Chapter 2

Background and Related Work

In this Chapter, we provide some background introducing important and relevant work in research areas pertinent to this thesis. We first introduce two research areas central to our work, Machine Translation and Computer Vision, with a focus on Machine Translation. We provide a brief history of the field, glancing at its main events and concepts. Moreover, we provide some background on two multi-modal NLP tasks related to our work, Image Description Generation and Multi-modal Distributional Semantic Models. Finally, we discuss related work and provide an explanation of where our work lies in relation to the current state-of-the-art.

2.1 Machine Translation

The goal of Machine Translation (MT) is the automatic translation of text and/or speech between two human languages, i.e. natural languages. Historically, MT as a scientific research field can be said to have appeared around the end of the 1940s and the beginning of the 1950s; at least if we consider western universities and research groups (Hutchins, 1978; Slocum, 1985). Early developments in the field allowed for over-optimistic prognostics to be made in 1956, when it was said that a machine would be able to translate independently of humans in five years time (Lufkin, 1965). Years later, with a lack of practical results, over-optimism gave room to what many

consider as an over-pessimistic view of the field (Hutchins, 1978), when a very stern and dire report released by the Automatic Language Processing Advisory Committee (ALPAC) in 1966 advocated that "there is no immediate or predictable prospect of useful machine translation" (ALPAC, 1966). Nowadays, MT can be seen as being a mature research field, having many established research groups dedicated to its research, and a vast majority of the largest companies in the IT sector investing heavily in it.

2.1.1 Statistical Machine Translation

Statistical MT (SMT) is a data-driven approach towards MT that aims to frame translation as a statistical optimisation problem (Koehn, 2010). Statistics are learnt from a parallel training corpus and later used for translating new, previously unseen sentences. The first statistical models proposed to translate text word by word and are popularly known as the *IBM models* (Brown et al., 1993). A refinement of word-based models is the influential Phrase-Based SMT (PBSMT) model (Koehn et al., 2003), in which a model learns to translate not word by word but on the basis of contiguous sets of words, i.e. phrases, which are not necessarily linguistically motivated. Moses, an open-source PBSMT toolkit, was released by Koehn et al. (2007) and promoted the adoption of PBSMT systems in a large scale, not just in the academia but also in the industry and the public sector. Moreover, many important further developments to PBSMT also contributed to its impact. For instance, the hierarchical PBSMT model proposed by Chiang (2005) introduced the idea of hierarchical phrases, i.e. phrases composed of sub-phrases, realised by means of synchronous context-free grammars. Another important theoretical development was the log-linear model proposed by Och and Ney (2002), which incorporated different features containing information from the source and target sentences in the model, in addition to the language and translation models of the original *noisy channel* approach.

The noisy channel The noisy channel (Shannon and Weaver, 1949) is an information-theoretical model of communication where a source sends a message to a receiver through a noisy medium or channel, as illustrated in Figure 2.1.

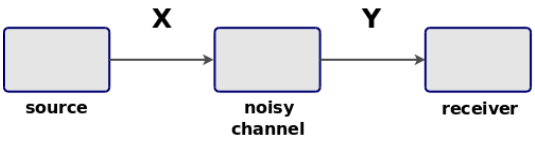


Figure 2.1: The noisy channel.

We see in Figure 2.1 that a source sends an original message $\mathbf{X}$, but the channel used to deliver the message corrupts it, i.e. thus the name noisy channel, so that the receiver observes $\mathbf{Y}$. The noisy channel in the context of MT proposes to model the transformation that a sentence in a source language undergoes when translated into a target language by a "corruption" as it is transmitted through the noisy channel.

Let us assume we wish to translate a sentence f in a foreign language, e.g. French, into a sentence e in English. In order to infer the translated sentence e^*, which has the maximal probability according to a given model, we need to traverse the model's search space as shown in Equation (2.1):

$$e^* = \arg\max_e P(e \mid f). \tag{2.1}$$

Following Bayes' theorem, the noisy channel approach for a fixed sentence f is shown in Equation (2.2):

$$P(e \mid f) = \frac{P(e)P(f \mid e)}{P(f)} \propto P(e)P(f \mid e). \tag{2.2}$$

The idea is that we observe sentences e in English, but in fact they are sentences f in French that were corrupted by a noisy channel. We have a model $P(f \mid e)$ that describes how the sentences were distorted, and also a model $P(e)$ that denote how likely an original sentence is. Here we observe two foundational features in SMT models: a *language model*, which is usually estimated on monolingual text in the target language, given by $P(e)$; and a *translation model*, which stores the probability

of a translation pair and is given by $P(f \mid e)$.

In other words, a language model computes how likely a sentence is in itself.[1] Highly likely or common sentences should be assigned a high probability, whereas rare ones should receive a low probability. A translation model will score how well a given sentence in French translates into English. Good translations should receive high probabilitites, whereas bad ones low probabilities.

In the noisy channel approach, according to Equation (2.2), we want to train a model to learn the distribution of some training data $T = \{e_i, f_i\}$, where the subscript i denote a particular sentence pair, and apply that model on to translate unseen sentences f into English.

Log-linear framework As mentioned above, an important theoretical generalisation of SMT models is the log-linear framework proposed by Och and Ney (2002). It directly models the probability $P(e \mid f)$ by summing over different features that encode information on the source, the target or both source and target languages in the model, as shown in Equation (2.3):

$$P(e \mid f) = \sum_{i=1}^{m} \exp \lambda_i \cdot h_i(e, f), \tag{2.3}$$

where there are $i = 1, \ldots, m$ different features in the model, h_i are feature functions that compute a certain relation between source- and target-language sentences—or monolingual features—, and λ_i is a weight that scales the impact of a specific feature function h_i on the overall probability $P(e \mid f)$. We note that the model described in Equation (2.3) includes the noisy channel, described in Equation (2.1), as a special case if we choose the language model feature as $h_1(e, f) = \log P(e)$ and the translation model feature $h_2(e, f) = \log P(f \mid e)$, with $\lambda_1 = \lambda_2 = 1$.

Finding the best translation for a sentence—which corresponds to the search problem introduced in Equation (2.1)—is also called *MT decoding* and is an NP-

[1] A language model can be used to estimate how likely are words, phrases, sentences or even larger chunks of text. Nevertheless, in this example we are only concerned with the probability of a sentence.

complete problem, possibly exponential in the length of the sentence to be translated (Knight, 1999). Normally, one or more heuristics are used in order to make decoding computationally feasible, such as a beam-search (Och and Ney, 2004) or cube-pruning (Chiang, 2007).

2.1.2 Neural Machine Translation

Recently, another data-driven approach very different from SMT has gained more attention: *neural MT* (NMT) models. NMT models are built using artificial neural networks (ANNs) to translate from one natural language into another. In this Section, we introduce important concepts necessary to understand ANNs, such as neurons, layers, and activation functions. We build up from simpler models, e.g. single-layer feed-forward networks, towards more complex models, leading eventually to neural MT models.

Artificial Neural Networks *Artificial Neural Networks* (ANNs) can be seen as distributed computational models inspired by the human brain, i.e. biological neural networks. They are distributed because an input signal is processed by many simple neurons or processing units, and these neurons are interconnected. Neurons in an ANN are located in layers, and a simple example of an ANN contains:

(i) one *input layer* which feeds an input signal to the next (hidden) layer;

(ii) one *hidden layer* which applies a possibly non-linear function on the input signal and feeds its output to the next (output) layer;

(iii) one *output layer* which again applies an arbitrary, possibly non-linear function to the input it receives (the output of the hidden layer) and generates an output.

We note that when applying ANNs onto NLP problems, it is very common that the input layer described in *(i)* requires input signals to be in the form of *word look-up matrices*. Word look-up matrices are simply a way to deterministically map

from discrete symbols, i.e. words, onto high-dimensional vectors, in turn used in an ANN's computations.

We now introduce an example of a simple artificial neural network with one hidden layer. Given an input vector $\boldsymbol{x} \in \mathbb{R}^{d_x}$, a hidden state $\boldsymbol{h} \in \mathbb{R}^{d_h}$ of an ANN, where d_x and d_h are the dimensionality of the input and hidden vectors, respectively, can be computed as given in Equation (2.4):

$$\boldsymbol{h} = f(\boldsymbol{W}_{ih}\boldsymbol{x}), \tag{2.4}$$

where $f(\cdot)$ is a possibly non-linear activation function (such as an element-wise sigmoid function), $\boldsymbol{W}_{ih} \in \mathbb{R}^{d_h \times d_x}$ is an input-to-hidden weight matrix and $\boldsymbol{x} \in \mathbb{R}^{d_x}$ is the input signal. Specifically, $\boldsymbol{h}$ is the output of a hidden layer as described in *(ii)*. Given the activations $\boldsymbol{h}$ of the ANN's hidden layer, the output of the network's output layer can be computed as in Equation (2.5):

$$\boldsymbol{y} = g(\boldsymbol{W}_{ho}\boldsymbol{h}), \tag{2.5}$$

where $g(\cdot)$ is a possibly non-linear activation function—such as a softmax activation function in case we wish to use this ANN in a 1–of–M classification—, $\boldsymbol{W}_{ho} \in \mathbb{R}^{M \times d_h}$ is a hidden-to-output weight matrix and $\boldsymbol{y} \in \mathbb{R}^M$ is the output vector.[2] Such a configuration could suit a problem such as the classification of an input vector $\boldsymbol{x}$ into one of M possible classes, where $\boldsymbol{y}$ is a one-hot vector representation of the output class. In other words, $\boldsymbol{y} \in \mathbb{R}^M$ is a vector for which exactly one entry $y_m, m = 1, \ldots, M$ is equal to one and all others are equal to zero. Each index m in the vector, $m = 1, \ldots, M$ encodes one class in M. We interpret $\boldsymbol{y}$ as an instance of class m if the corresponding class index $m \in M$ is active ($y_m = 1$).

In Figure 2.2, we show an example of a simple ANN architecture. In this example, the input and hidden layers have four neurons each, and the output layer has two neurons.

[2]In both Equations (2.4) and (2.5), standard bias terms are ommited for clarity.

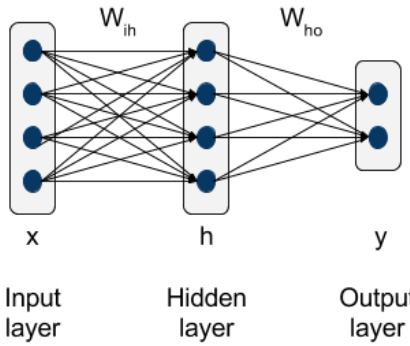


Figure 2.2: A simple artificial neural network.

In such a neural network classification model, an important decision is first how to choose $f(\cdot)$ and $g(\cdot)$ for a particular problem, and also the number of neurons d_h in the hidden layer. Assuming $f(\cdot)$ and $g(\cdot)$ as an element-wise sigmoid and softmax activation functions respectively, training the network can be seen as finding parameters W_{ih} and W_{ho} (as well as bias vectors) that best fit some training data and can be done using *backpropagation* (Rumelhart et al., 1986).

Recurrent Neural Networks *Recurrent Neural Networks* (RNNs) are a generalisation of the abovementioned ANNs for dealing with inputs $\boldsymbol{X}$ that encode a temporal relation, also referred to as sequences, e.g. an audio signal or a sentence. Now, inputs and outputs are each a *sequence of vectors* instead of only a single vector. Given an input $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$, an RNN computes a sequence of outputs $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T\}$, where T is the length of the source and target sequences, by iterating Equations (2.6) and (2.7):[3]

$$\boldsymbol{h}_t = f(\boldsymbol{W}_{ih}\boldsymbol{x}_t + \boldsymbol{W}_{hh}\boldsymbol{h}_{t-1}), \tag{2.6}$$

$$\boldsymbol{y}_t = g(\boldsymbol{W}_{ho}\boldsymbol{h}_t), \tag{2.7}$$

where $\boldsymbol{W}_{hh}$, $\boldsymbol{W}_{ih}$ and $\boldsymbol{W}_{ho}$ are weight matrices that parametrise the recurrent connection, the connection between input and hidden layers, and the connection between hidden and output layers, respectively. $\boldsymbol{h}_t$ is the hidden state or memory unit at a

[3]Standard bias terms are ommited for clarity.

time step t, $f(\cdot)$ and $g(\cdot)$ are again possibly non-linear activation functions and $\boldsymbol{y}_t$ is an output at timestep t. At each timestep t, an input vector $\boldsymbol{x}_t \in \mathbb{R}^{d_x}$, $t = 1, \ldots, T$ is read from $\boldsymbol{X}$ and one output vector $\boldsymbol{y}_t \in \mathbb{R}^{d_y}$ is generated. Also, the hidden state $\boldsymbol{h}_t$ is updated by incorporating the new input $\boldsymbol{x}_t$ and the previous hidden state $\boldsymbol{h}_{t-1}$. We denote the set of hidden states $\boldsymbol{H} = \{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_t\}$. In Figure 2.3, we illustrate the computation of one time step of an RNN, where the input and output vectors at each time step have three neurons, and the hidden state has two neurons.

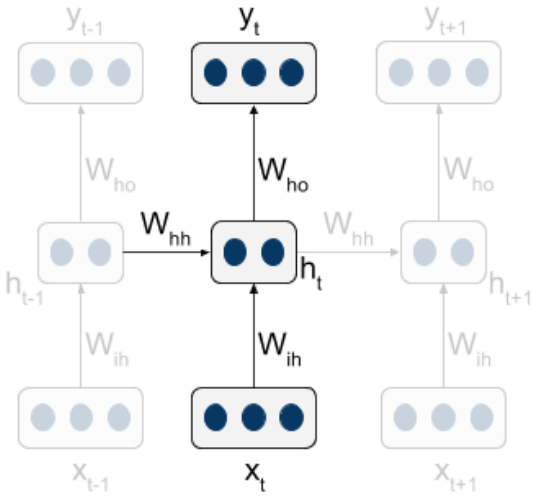


Figure 2.3: The computation of one time step of a recurrent neural network.

In order to train an RNN, we can apply a derivation of the backpropagation algorithm similarly to the ANN scenario. Nevertheless, because of the way it is structured, an RNN memory unit $\boldsymbol{h}_t$ is fully updated at every timestep t and cannot efficiently encode long-distance relations. Two improvements devised to allow for a more efficient memory unit $\boldsymbol{h}$ in RNNs are the Long Short-Term Memory (LSTM) units of Hochreiter and Schmidhuber (1997) and the more recent Gated Recurrent units (GRU) of Cho et al. (2014b).

In an LSTM unit, instead of having Equation (2.6) updating the memory vector $\boldsymbol{h}_t$ at each timestep t, there is a more complex structure that uses different *gates* to control the flow of information. In simple terms, an LSTM unit can learn not just whether a given input $\boldsymbol{x}_t$ at time t is important and should therefore be stored in memory, but also how much and what "portions" of $\boldsymbol{x}_t$ should be stored and

which should be discarded. Graves (2013) proposes an LSTM unit where the original function $f(\cdot)$ in Equation (2.6) is decomposed in different gates: *input*, *forget*, *output* and *memory*. Input and forget gates can be seen as switches the LSTM learns in order to know what to remember from the current input (timestep t, $\boldsymbol{x}_t$) and what to forget from the previous inputs (timesteps 1 to $t-1$, $\boldsymbol{h}_{t-1}$), respectively. Finally, all gates and memory taken together allow an LSTM unit to learn complex functions and long-term temporal relations that a standard RNN cannot (Sutskever et al., 2014; Graves, 2013). Similarly, GRU units are similar to LSTM units but simpler, consisting of two gates only, instead of four in the original LSTM (Cho et al., 2014b). Nevertheless, GRU and LSTM units with a comparable number of parameters have been found to be equivalent in most cases, presenting only small differences depending on the particular task at hand (Chung et al., 2014).

According to Sutskever et al. (2014), RNNs are good at mapping input sequences $\boldsymbol{X}$ to output sequences $\boldsymbol{Y}$ whenever there is a monotonic alignment between the input and output. However, in MT the relationship between the input and output sequences is complex, involves non-monotonic relations, and $\boldsymbol{X}$ and $\boldsymbol{Y}$ can have different lengths. Therefore, a different strategy is needed.

Perhaps Forcada and Ñeco (1997) are the first to attempt to apply ANNs to train models to translate between two natural languages. The authors trained very small networks compared to today's standards, e.g. with between 3 and 12 neurons in a hidden layer, and their model can be seen as a precursor of the sequence to sequence model introduced by Sutskever et al. (2014). There were one encoder and one decoder, and hidden representations were not real-valued but consisted of binary vectors instead, i.e. a sort of one-hot hidden vector representation. Training these models was difficult, and the authors did not use back-propagation but chose instead to use a non-gradient method for training. Furthermore, there are many other difficulties that must be overcome to apply ANNs end-to-end to translate sentences between two natural languages.

Some authors propose to use neural networks along with traditional SMT models,

which alleviates some of the issues which we discuss further in the next pages. Bengio et al. (2003) propose using a neural network for training language models (LMs) and report gains in perplexity in comparison to state-of-the-art n-gram LMs when evaluating their model. Their LM can be directly used as a feature in a standard log–linear SMT model. Schwenk (2007) proposes using a neural network to train an LM for large vocabulary continuous speech recognition and report improvements on different tasks and language pairs. Devlin et al. (2014) use neural networks for training an LM conditioned on both source and target sentences. They also point out that their LM scores can be used in an MT decoder and not just for reranking candidate sentences in a decoder's n-best list.

A recent resurge in neural network-based models is the application of so-called Deep Neural Networks (DNNs), which can be broadly defined as ANNs that have more than one hidden layer between its input and output layers, to model an end-to-end neural MT model. Kalchbrenner and Blunsom (2013) introduce what they name recurrent continuous translation models, which use continous vector representations for words and sentences; Cho et al. (2014b) propose to use a DNN to perform MT and use two RNNs to model the translation task; Similarly, Sutskever et al. (2014) propose using a DNN in MT by modelling the translation process as a mapping between two different sequences. One point in common in all these models is the use of the encoder–decoder framework.

Encoder–decoder The *encoder–decoder framework* tries to address the issues that appear when trying to apply neural networks to problems where inputs and outputs are sequences which have different, variable lengths, and there is a complex alignment between inputs and outputs, as well as long-range dependencies. It consists of two different neural networks that are trained together:

- an *encoder* neural network, responsible for encoding a variable–length input sequence $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{T_x}\}$, $\boldsymbol{x}_t \in \mathbb{R}^{d_x}$, into a fixed-length vector representation $\boldsymbol{v} \in \mathbb{R}^{d_h}$; and

- a *decoder* neural network, responsible for decoding a variable–length output sequence $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{T_y}\}$, $\boldsymbol{y}_t \in \mathbb{R}^{d_y}$, from this fixed-length vector representation $\boldsymbol{v}$.

We note that input and output sizes can differ ($T_x \neq T_y$). In Figure 2.4, we show an example of an encoder–decoder architecture. Common approaches for the encoder and decoder networks described above make use of convolutional neural networks (Kalchbrenner and Blunsom, 2013), RNNs with LSTM cells (Sutskever et al., 2014), or RNNs with GRU cells (Cho et al., 2014b).

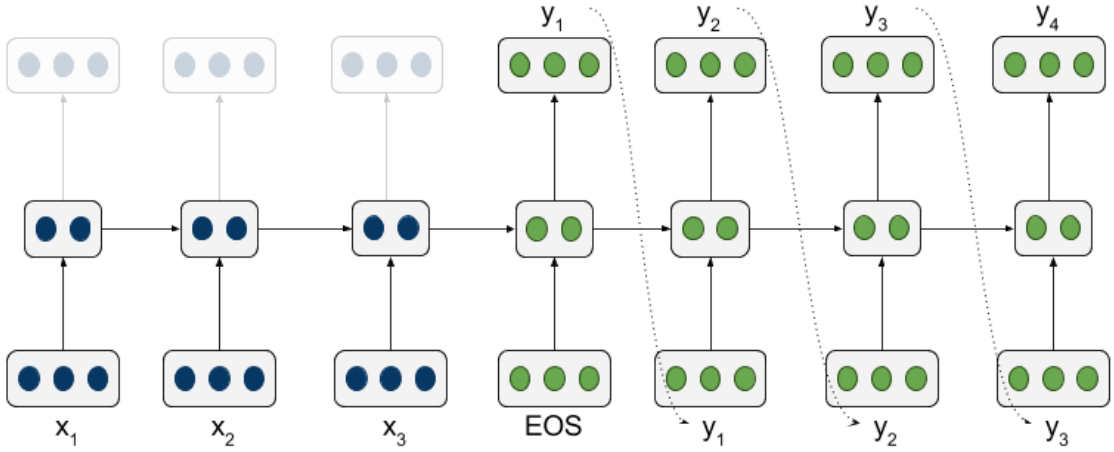


Figure 2.4: The encoder–decoder architecture with RNNs as the encoder and the decoder. In this example, the encoder is shown in blue and the decoder in green, and the input sequence is $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$, $\boldsymbol{x}_t \in \mathbb{R}^3$, and the output sequence is $\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3, \boldsymbol{y}_4\}$, $\boldsymbol{y}_t \in \mathbb{R}^3$. When the encoder RNN reads a special end-of-sentence symbol (EOS), it triggers the beginning of the decoding process. Normally, when training the decoder RNN the input to the next time step t is the output of the previous time step $t-1$, also known as the *teacher forcing algorithm* (Williams and Zipser, 1989).

Applying end-to-end neural network models to MT is far from straightforward and, in fact, successful applications of such models have only recently been reported. These models are usually computationally costly: training one neural MT model can take several days or even weeks even on very optimised hardware, i.e. state-of-the-art graphics processors (GPUs), and also using a neural MT model for exact decoding a target sentence given a source sentence can take prohibitively long times. Cho et al. (2014a) study neural network–based MT models and discuss some important

points that should be taken into consideration. The source- and target-language vocabularies must be restricted, usually to a few thousands words in order to make training and decoding feasible. For instance, Cho et al. (2014a) use $30,000$ words for source and target vocabularies each and Kalchbrenner and Blunsom (2013) use vocabularies in the same order of magnitude. Moreover, the training set size and input sentence lengths can be an issue: Cho et al. (2014a) use a training set with 348 million words to train their models and consider only sentences with up to 30 words in length; Bahdanau et al. (2015) use the same training set and consider sentences with up to 50 words; Kalchbrenner and Blunsom (2013) use around 8.5 million words and sentences with up to 80 words. Still, after these many steps to try to alleviate some of the problems in training NMT models, Kalchbrenner and Blunsom (2013) report training their model for about 15 hours, Cho et al. (2014a) train for around 4.5 days and Bahdanau et al. (2015) for about 5 days.

Attention mechanism The main bottleneck in the encoder–decoder framework is the need to encode the whole source sentence $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{T_x}\}$, $\boldsymbol{x}_t \in \mathbb{R}^{d_x}$, into a fixed-length vector representation $\boldsymbol{v} \in \mathbb{R}^{d_h}$, so that the decoder only has access to this fixed-length vector $\boldsymbol{v}$ in order to generate a translation $\boldsymbol{Y} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{T_y}\}$, $\boldsymbol{y}_t \in \mathbb{R}^{d_y}$.

Bahdanau et al. (2015) first introduced an attention mechanism into the NMT encoder–decoder framework to propose a solution to this bottleneck. The attention is a mechanism devised so that the decoder can have access to all the hidden states generated by the encoder in all time steps $\boldsymbol{H} = \{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{T_x}\}$. We denote the hidden states of the decoder as $\boldsymbol{D} = \{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_{T_y}\}$. In its original formulation, the attention mechanism is implemented as a multi-layer perceptron. Briefly, a single-layer feed-forward network is used to compute an *expected alignment* $e_{t,i}$ between each hidden vector $\boldsymbol{h}_i$ representing a source word $\boldsymbol{x}_i$, $i \in \{1, \ldots, T_x\}$, and the target word $\boldsymbol{y}_t$ at the current time step t, $t \in \{1, \ldots, T_y\}$ as showed in Equations (2.8)

and (2.9):

$$e_{t,i} = (\boldsymbol{v}_a)^T \tanh(\boldsymbol{U}_a \boldsymbol{d}_{t-1} + \boldsymbol{W}_a \boldsymbol{h}_i), \tag{2.8}$$

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{j=1}^{N} \exp(e_{t,j})}, \tag{2.9}$$

where $\alpha_{t,i}$ is the normalised alignment matrix between each source hidden vector $\boldsymbol{h}_i$ and the word $\hat{y}_t$ at time step t, and $\boldsymbol{v}_a$, $\boldsymbol{U}_a$ and $\boldsymbol{W}_a$ are trained model parameters.

Finally, a time-dependent source context vector $\boldsymbol{c}_t$ is computed as a weighted sum over the source hidden vectors, where each vector is weighted by the attention weight $\alpha_{t,i}$, as shown in Equation 2.10:

$$\boldsymbol{c}_t = \sum_{i=1}^{T_x} \alpha_{t,i} \boldsymbol{h}_i. \tag{2.10}$$

This time-dependent source context vector $\boldsymbol{c}_t$ is computed at each time step t of the decoder and replaces the fixed-length vector $\boldsymbol{v}$, used in the original encoder–decoder framework. By dynamically deriving a different context vector $\boldsymbol{c}_t$ at each time step t of the decoder, an attention-based model can learn to which source words $\boldsymbol{x}_i$ to align each target word $\boldsymbol{y}_t$. In Figure 2.5, we illustrate an attentive decoder RNN, where it has already generated the output sequence up to time step $t = 2$, and is now processing time step $t = 3$.

We note that the attention mechanism is entirely differentiable and is trained jointly with the rest of the model. It is also called a *soft-alignment*, differently from a traditional MT alignment (Och and Ney, 2003), i.e. a "hard" alignment, where a target word would be either fully aligned to a source word or not.

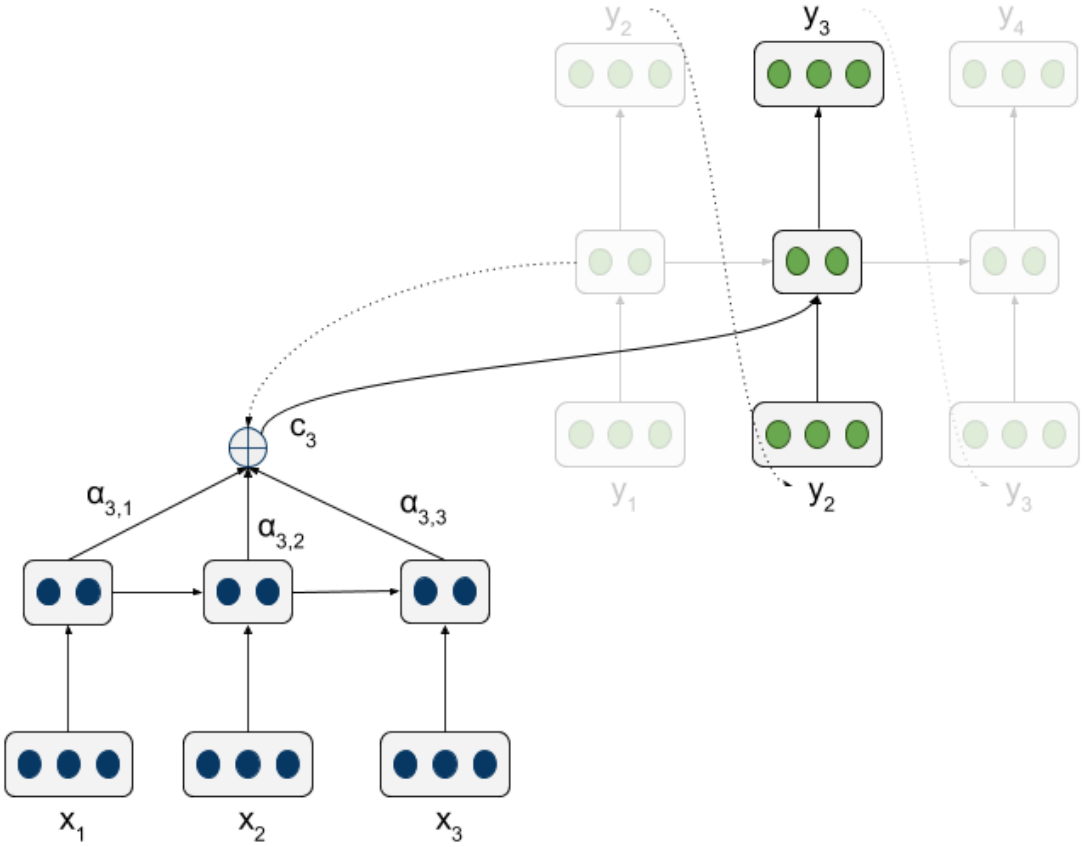


Figure 2.5: An illustration of one time step of the attention-based NMT model. In this example, the encoder is shown in blue and the decoder in green, and the input sequence is $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$, $\boldsymbol{x}_t \in \mathbb{R}^3$. The decoder RNN has already generated the output sequence up to time step $t = 2$, and is now processing time step $t = 3$. Note that the context vector $\boldsymbol{c}_t$ is time-dependent, and computed at each time step t of the decoder.

2.2 Computer Vision

Computer Vision (CV) is a research field that studies how to automatically understand images. Similarly to MT, it is also a very mature research field with many different groups in numerous countries dedicated to its research. Some of the historically challenging tasks in CV include detecting objects in images, or performing (human) pose estimation from images and videos, to name but a few. In this Section, we briefly discuss the main evaluation campaing in CV and introduce CNNs, the driving force behind virtually all current state-of-the-art image classification models. We particularly describe two CNNs in more detail, which we use in this work.

Computer Vision is a research area that directly impacts this work, and for that

reason we now discuss some relevant works and ideas in this field. Before we move on to discuss any specific work in CV, we first introduce ImageNet. ImageNet is a large image database built on top of the WordNet hierarchy (Miller, 1995), and its goal is illustrate each of WordNet's synsets with an average of 1,000 images.[4] WordNet is a lexical and semantic database structured around synsets and relations between synsets, where synsets are sets of words used in a specific context.[5] For instance, the synset for the word "jaguar" in the sentence "Jaguars are strong animals." is not the same as the one in the sentence "Jaguars and Ferrarris are very expensive." We note that the ImageNet Large Scale Visual Recognition Challenges (ILSVRCs) (Russakovsky et al., 2015) are one of the major forums for discussion and evaluation of ideas for CV. Moreover, these large scale evaluations are done by means of different *evaluation tasks*.

We now introduce the two main tasks evaluated in the ILSVRC campaigns: *image classification* and *localisation*. In image classification, the task is to assign one class to an image, where this class can be one out of $1,000$ possible classes from ImageNet. In localisation, the task is to first identify what objects appear in an image and also where every instance of each of these objects appear, i.e. their bounding boxes. Neural network-based models have since the ILSVRC 2012 campaign ranked among the best performing ones in the ImageNet Large Scale Visual Recognition Challenges (ILSVRC) for many years now (Russakovsky et al., 2015). The SuperVision model of Krizhevsky et al. (2012) represented a turning point in the field, clearly outperforming all other models in that years' campaign and leading the way in the next years. The SuperVision model uses *convolutional neural networks* to process images and to perform image classification and localisation, and now we describe how these networks work.

[4] `http://image-net.org/about-overview`

[5] More information can be found at `https://wordnet.princeton.edu/`

2.2.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are neural networks inspired by the visual cortex, and were devised to tackle problems involving image processing and understanding. They employ a combination of local receptive fields, shared weights and pooling for dimensionality reduction (Lecun et al., 1998). Although they were first proposed to address problems involving vision, they have since also been successfully applied to different tasks, e.g. NMT (Kalchbrenner and Blunsom, 2013). They can be interpreted as *deep networks* and their main, distinctive layers are *convolutional* and *pooling* layers.

One first important characteristic of CNNs is that each neuron in a layer only receives inputs from a set of neurons located in a small neighbourhood in the previous layer. That means that the activations these neurons compute are dependent on a small number of neighbouring neurons, and are therefore *local*. Local activations cause, at lower layers, neurons to extract elementary visual features such as edges, corners, or end-points and, at higher layers, they compute increasingly complex combination of these features.

We now follow Gu et al. (2015) to explain common CNNs' layers and concepts. A convolutional layer consists of a set of *kernels* used to compute different *feature maps*. Each kernel has a set of weights, which is convolved with the inputs in order to produce a feature map. These weights are shared, meaning that each kernel is constrained to apply the same operations in different parts of an image, when computing a feature map. One complete convolutional layer consists of different feature maps computed using many different kernels. Following Gu et al. (2015), the feature value at location $\langle i, j \rangle$ in the k-th feature map of the l-th layer, $z^l_{i,j,k}$, is computed as in Equation (2.11):

$$z^l_{i,j,k} = (\boldsymbol{w}^l_k)^T \boldsymbol{x}^l_{i,j} + b^l_k, \tag{2.11}$$

where $\boldsymbol{w}^l_k$ and b^l_k are the weight vector and bias of the k-th filter of the l-th layer,

respectively, and $\boldsymbol{x}^l_{i,j}$ is the input patch centered at location $\langle i, j \rangle$ of the l-th layer. The kernel is fully described by the combination of the learnt parameters $\boldsymbol{w}^l_k$ and b^l_k, and it computes the feature map $\boldsymbol{z}$ when convolved with the inputs $\boldsymbol{x}$. Note that the kernel $\langle \boldsymbol{w}^l_k, b^l_k \rangle$ that generates the feature map $\boldsymbol{z}^l_{:,:,k}$ is shared across different locations $\langle i, j \rangle$ in the input.

In order to be able to compute non-linear features over the inputs, a non-linear element-wise function is usually applied to the feature maps, as in Equation (2.12):

$$a^l_{i,j,k} = g(z^l_{i,j,k}), \tag{2.12}$$

where the $g(\cdot)$ function is normally the sigmoid, tanh or ReLU functions.

The last essential building block needed to understand CNNs is the *pooling* layer. This layer is central to bring some degree of shift invariance to the network (Lecun et al., 1998). The intuition is that, once a feature map $\boldsymbol{a}$ is computed, its exact location become less important. By sub-sampling or pooling, we reduce the dimensionality of the feature maps. A general-purpose pooling function has the format described in Equation (2.13):

$$y^l_{i,j,k} = \text{pool}(a^l_{m,n,k}), \forall (m, n) \in \mathcal{N}_{i,j}, \tag{2.13}$$

where $\mathcal{N}_{i,j}$ is the neighbourhood around location $\langle i, j \rangle$. The pooling function is normally implemented as a max- or mean-pooling, where in the former a certain number of maximum features are kept and the remaining ones are discarded, whereas in the latter the average of the features is computed.

The three types of layers are usually organised in sequential blocks: a convolutional layer followed by a non-linearity followed by a pooling layer. For instance, these are the building blocks of the seminal LeNet-5 network of Lecun et al. (1998), shown in Figure (2.6).

Since the CNN of Krizhevsky et al. (2012) was introduced in the ILSVRC campaign of 2012, many research groups have attempted to use CNNs for processing

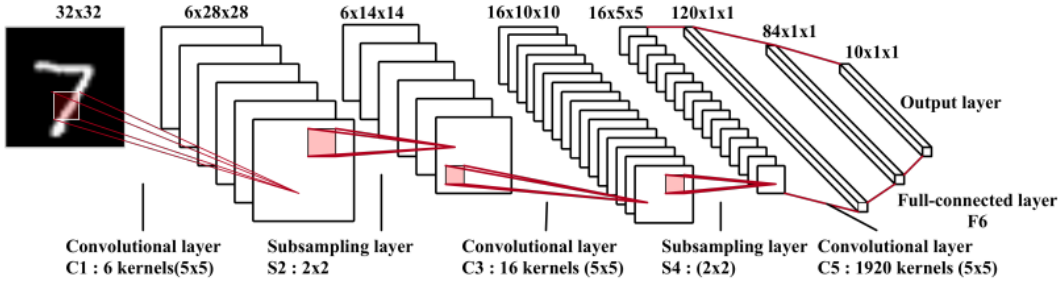


Figure 2.6: Illustration of the LeNet-5 network (Gu et al., 2015).

images, and many of them with great success (Simonyan and Zisserman, 2014; He et al., 2015; Szegedy et al., 2015). We now introduce and discuss two important research groups and their proposed CNNs, since we use their pre-trained models in our work: the Oxford University group who created the VGG networks (Simonyan and Zisserman, 2014) and the Microsoft Inc. group who put forward the Residual Networks (He et al., 2015).

2.2.1.1 VGG Networks

Simonyan and Zisserman (2014) introduced the VGG networks and released two pre-trained versions of their networks, the VGG16 and the VGG19. In Figure 2.7 we see an illustration of the VGG19 network architecture.

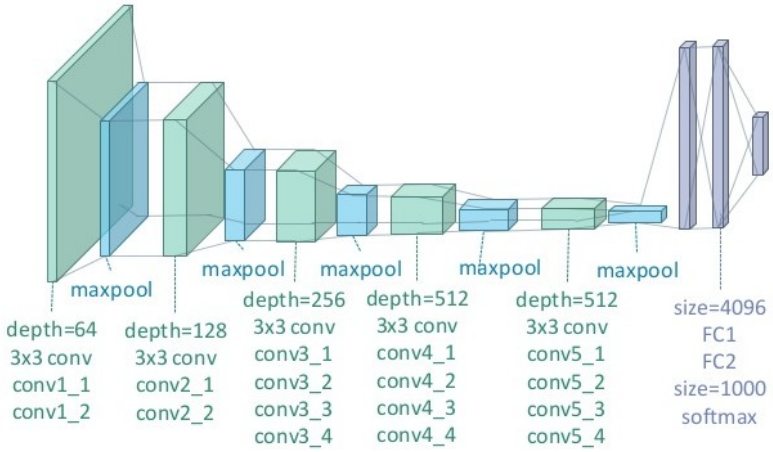


Figure 2.7: Illustration of the VGG19 network architecture.[6]

The VGG16 network has 16 layers and corresponds to the configuration C in the authors original paper (Simonyan and Zisserman, 2014), whereas the VGG19

contains 19 layers and corresponds to the network configuration E. They are actually very similar, the difference between the two being three extra convolutional layers in the VGG19 (from Figure 2.7, layers `conv3_4`, `conv4_4` and `conv5_4`).

In our work, we use the VGG19 network for image feature extraction. We specifically use two types of features: *global features* extracted from the penultimate fully-connected layer of the VGG19 network (denoted by FC2 in Figure 2.7), which consists of a $4,096$D feature vector, henceforth FC7 features; and *local features* extracted from the `conv5_4` layer of the VGG19 network, which consists of a $\langle 14 \times 14 \times 512 \rangle$ 3-tensor, henceforth CONV54. The FC7 features are global features, meaning that they encode the entire image into one $4,096$D feature vector, whereas the CONV54 features are local features, which can be seen as encoding an image in a 14×14 grid where each of the entries in the grid is represented by a 512D feature vector that only encodes information about that specific region of the image. We vectorise this 3-tensor into a 196×512 matrix $A = (a_1, a_2, \ldots, a_L)$, $a_l \in \mathbb{R}^{512}$, where each of the $L = 196$ rows consists of a 512D feature vector and each column, i.e. feature vector, represents one grid in the image. By vectorising the 3-tensor into a matrix, we can simply utilise it as a sequence (i.e., containing $L = 196$ positions) with separate features for each position (i.e., 512D feature vectors) in a similar way as in a NMT model, where we have a sequence of words in a sentence with separate features for each word (i.e. word embeddings).

2.2.1.2 Residual Networks

He et al. (2015) introduced the *residual networks*, commonly known as ResNets. One driving goal when devising these networks was experimenting with increased number of layers, under the hypothesis that by using more layers one would be able to obtain better models. Nevertheless, training very deep networks can be difficult. The authors note that naïvely adding more layers to a model, i.e. making it deeper, may cause a *degradation* problem. This problem appears when training

[6] `https://goo.gl/y0So1l`

a deep network, where the network accuracy saturates and then quickly starts to degrade (He et al., 2015; He and Sun, 2015). The authors used *residual connections* to address this problem, where they take the form illustrated in Figure 2.8. Residual connections could help training these models since they establish "shortcuts" with the identity function, effectively shortening the path that gradients have to traverse between a network's output and input layers during back-propagation.

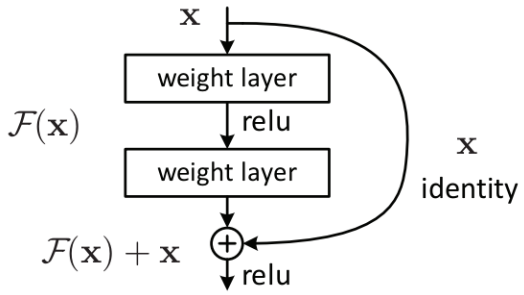


Figure 2.8: Illustration of a residual connection. (He et al., 2015)

A residual connection can be seen as a "shortcut", where an input X is fed through a set of possibly non-linear transformations $\mathcal{F}(\cdot)$, and the output of $\mathcal{F}(X)$ is added to the original input X. This makes it easier for the gradients to flow from the deeper layers back to the shallower ones when training a model using backpropagation, addressing the degradation problem.

He et al. (2015) released pre-trained versions of three networks, referred to the ResNet-50, ResNet-101 and ResNet-152 networks. As their names imply, the ResNet-50 has 50 layers, the ResNet-101 has 101 layers, and the ResNet-152 has 152 layers. The three networks structures are very similar, the main difference lying in the quantity of layers and residual connections. In other words, the ResNet-101 and ResNet-152 employ the same layer architecture as the ResNet-50, but simply has more building blocks in the final network architecture.

In our work, we use the ResNet-50 network released by He et al. (2015) for image feature extraction. We specifically use *local features* extracted from the `res4f` layer of the ResNet-50 network,[7] which consists of a $\langle 14 \times 14 \times 1,024 \rangle$ 3-tensor, henceforth

[7]Please see a visualisation for the original ResNet50 network `http://ethereon.github.io/netscope/#/gist/db945b393d40bfa26006`.

`res4f`. The `res4f` features are local features, which can be seen as encoding an image in a 14×14 grid where each of the entries in the grid is represented by a $1,024$D feature vector that only encodes information about that specific region of the image. Similarly to what we do with the local features extracted using the VGG19 network, we vectorise this 3-tensor into a $196 \times 1,024$ matrix $A = (a_1, a_2, \ldots, a_L)$, $a_l \in \mathbb{R}^{1,024}$, where each of the $L = 196$ rows consists of a $1,024$D feature vector and each column, i.e. feature vector, represents one grid in the image.

2.3 Image Description Generation

Image Description Generation (IDG) is the task where given an image, we want to generate its description. It is particularly relevant to this work, since one way to view multi-modal MT is as the task that bridges the gap between translation and image description generation. IDG, like multi-modal MT, requires two different types of *reasoning*: one of them is *visual*, where one or more models should be able to classify the image, detect objects within the image, localise these objects, and recognise interactions between objects in the image; and the other one is *linguistic*, where all of these different information generated for the image can be put into words in well-formed sentences (Bernardi et al., 2016).

IDG has been receiving much attention lately, especially since by transferring learning from pre-trained CNNs, the visual reasoning portion of this task has been considerably improved (Bernardi et al., 2016). This has been even further facilitated by the release of publicly available, pre-trained CNNs such as the VGG networks and the ResNets. To the best of our knowledge, one of the first works on IDG is that of Farhadi et al. (2010), who proposed to train a model to map images and sentences onto a common space in the form of ⟨object, action, scene⟩. The seminal work of Hodosh et al. (2013) proposed the use of Kernel Canonical Correlation Analysis (KCCA) to map images and sentences onto a multi-modal vector space. They argued that by framing IDG as a ranking task, they can better evaluate the

capacity of a model to relate sentences to images without having to deal with the difficult linguistic reasoning portion of the task.

Gong et al. (2014) propose a method based on Canonical Correlation Analysis (CCA) to the problem of automatic image description and argue that KCCA does not scale well for large training sets. They use the Flickr30k dataset as their (fully supervised) training corpus and analyse how to incorporate an additional 2 million distinct Flickr images with titles/descriptions which are not guaranteed to have similar quality, coverage or domain. They apply image features obtained using deep CNNs as reported by Donahue et al. (2014) and use bag-of-words features for the textual data. In their method, called *Stacked Auxiliary Embedding*, they map both sentences and images to one same vector space and use the 2 million (weakly) annotated images in order to improve retrieval when presenting an image to the sentence–image embedding and retrieving candidate sentences that describe it.

Socher et al. (2014) propose a method for mapping sentences and images onto a common multi-modal vector space. For generating sentence embeddings, they train an RNN that aggregates words using the output of a dependency parser. For obtaining image embeddings, they use a DNN proposed by Le et al. (2012), which was trained to classify images into one of $22,000$ ImageNet classes. Their multi-modal embedding training procedure uses a max-margin objective function that is trained to project pairs of related sentence and image vectors to be close, and pairs of unrelated (random) sentence and image vectors to be far apart. They show that their model outperforms previous methods that used KCCA approaches, for instance the method introduced by Hodosh et al. (2013).

Karpathy et al. (2014) extend the multi-modal embeddings model and propose to work on a finer granularity. Instead of mapping entire images to entire sentences as in previous work, their mappings are between objects, i.e. parts of an image, and parts of a sentence, while still using the full image–sentence mappings as well. Their model outperformed the previous state-of-the-art (Socher et al., 2014).

Vinyals et al. (2015) introduced an influential IDG model based on the encoder–

decoder approach. In their model, an image is encoded using a pre-trained CNN. Images are fed into the pre-trained CNN to obtain fully-connected features, which are in turn used to initialise the hidden state of a decoder RNN, which in turn generates the image description in natural language. Elliott et al. (2015) put forward a model to generate multilingual descriptions of images by learning and transferring features between two independent, non-attentive neural image description models.

Finally, many of the neural network-based models that presented state-of-the-art performance were based on the encoder–decoder framework and suffered from the same problems other models based on the same framework suffered: the need to squeeze all the information computed using an encoder, in this case a pre-trained CNN, into one feature vector that the decoder would in turn be conditioned upon. Xu et al. (2015) addressed this issue for IDG—similarly to how Bahdanau et al. (2015) did for NMT—by proposing an attention-based model to the task where a model learns to attend to specific parts of an image representation (the source) as it generates its description (the target) in natural language. The model was trained end-to-end and did not use fully-connected visual features, but instead used *local* features that encode spatial information. Now, instead of squeezing the entire image into one feature vector, the decoder has access to a set of local feature vectors and learns which regions of the image to attend to when generating each word in the target language. For a complete survey of models and data sets used in image description generation, please refer to Bernardi et al. (2016).

2.4 Multi-modal Distributional Semantic Models

Distributional semantic models (DSMs) compute word vector representations from text based on word co-occurrence patterns. However, these models suffer from an obvious limitation since the meaning of a word is derived entirely from connections to other words, i.e. they do not take extra-linguistic modalities into account (Harnad, 1990; Glenberg and Robertson, 2000). This is the case not only for widely adopted

word-level DSMs, e.g. word2vec (Mikolov et al., 2013), but also for sentence-level DSMs, e.g. skip-thought vectors (Kiros et al., 2015). In this work, we propose a multilingual multi-modal embedding model that incorporates images into a sentence-level DSM. For that reason, we provide a background on both text-only and multi-modal DSMs.

Multi-modal distributional semantic models try to expand DSMs and include inputs from additional modalities other than text as a means to address the *grounding* problem. At the word level, Bruni et al. (2014) propose deriving word and image vectors, where the word vector representations are based on co-occurrence counts in text corpora, and the images are represented using a bag-of-visual-words method with Scale-Invariant Feature Transform (SIFT) vectors (Lowe, 1999, 2004) extracted from a data set of tagged images. These two representations are concatenated and merged using Singular Value Decomposition. Silberer and Lapata (2014) use stacked auto-encoders to map words and images to one same shared multi-modal embedding space. Their image representation is obtained using attribute classifiers that predict visual attributes (e.g., `has wings`, `made of wood`) for given words, proposed in Farhadi et al. (2009). Lazaridou et al. (2015) expand the word2vec *skip-gram* (Mikolov et al., 2013) into a multi-modal *skip-gram* model by incorporating image features extracted from pre-trained CNNs. As we can observe, visual features obtained with pre-trained CNNs are widely used in transfer learning scenarios. More examples include visual question answering (Zhang et al., 2016), to train multi-modal word embeddings (Lazaridou et al., 2015) or in multi-modal neural machine translation (Calixto et al., 2017f).

Until this point, all these DSMs have in common that they learn models at the word-level. Nonetheless, there are many models that propose to learn sentence-level (Kiros et al., 2015; Arora et al., 2017) or even paragraph-level vector representations (Le and Mikolov, 2014). Similarly to their word-level counterparts, these models are trained based on text signals only.

At the sentence level, Kiros et al. (2014) propose a multi-modal embedding model

trained to map sentences and images into one shared multi-modal embedding space, where the sentences are encoded using RNNs. Vendrov et al. (2016) extend their model to include an asymmetric mapping function between sentences and images. In a similar vein, Socher et al. (2014) utilised Recursive Neural Networks, i.e. RNNs that operate on parse trees, as their sentence encoder. They all utilised pre-trained CNNs to extract image features and a pairwise ranking function to train their multi-modal embeddings.

2.5 Related work

In this Section, we discuss important related work in the area of multi-modal MT, comparing our models to other related models proposed in the literature as well as explaining how our models compare to the state-of-the-art.

Calixto et al. (2012) first studied how the visual context of a textual description can be helpful in the disambiguation of SMT systems. Since that introductory work much progress has been made: the introduction of the VGG and Residual networks, on the computer vision side, and of attentive neural MT networks (Bahdanau et al., 2015), on the machine translation side. Nonetheless, multi-modal MT has just recently been addressed by the MT community in a shared task (Specia et al., 2016). However, there has been a considerable amount of work on natural language generation from non-textual inputs, as discussed in Section 2.3.

In the context of NMT, Dong et al. (2015) proposed a multi-task learning approach where a model is trained to translate from one source language into multiple target languages. They used attention-based decoders where each language has one decoder RNN with a separate attention mechanism. Each translation task has a shared source-language encoder in common with all the other translation tasks. Firat et al. (2016) proposed a multi-way model trained to translate between many different source and target languages. Instead of one attention mechanism per language pair as in Dong et al. (2015), which would lead to a quadratic number of

attention mechanisms in relation to language pairs, they use a shared attention mechanism where each target language has one attention shared by all source languages. Luong et al. (2016) proposed a multi-task learning approach where they train a model using two tasks and a shared decoder: the main task is to translate from German into English and the secondary task is to generate image descriptions in English. They show improvements in the main translation task when also training for the secondary image description task. Although not an NMT model, Hitschler et al. (2016) recently used image features to re-rank translations of image descriptions generated by an SMT model and reported significant improvements.

Different research groups have proposed to include global and local visual features in re-ranking n-best lists generated by an SMT system or directly in a NMT framework with some success (Caglayan et al., 2016; Calixto et al., 2016; Huang et al., 2016; Libovický et al., 2016; Shah et al., 2016; Specia et al., 2016).

Caglayan et al. (2016) experimented with re-ranking n-best lists ($n = 1,000$) generated using a baseline Moses SMT system, using additional LM-based source and target features as well as global image features. They also trained a multi-modal NMT system, where they incorporated local image features in an attention mechanism combined with the textual attention mechanism to compute a multi-modal context vector. Calixto et al. (2016) evaluated a multi-modal NMT model where two independent attention mechanisms were used to integrate text and images. Libovický et al. (2016) combined a PBSMT and a NMT system together. They used the PBSMT system to generate translations, then fed these translations into a target-language encoder RNN. They trained a standard NMT model, as well as a multi-modal NMT model where they included global image features in the decoder initialisation. They reported results for using the PBSMT system translations only, using the multi-modal NMT model translations only, as well as a combination of these two via the target-language encoder. Shah et al. (2016) did not use NMT in their submissions and proposed to use global image features to re-rank n-best lists generated with Moses ($n = 100$). Different from others, they use the VGG16

network to extract a global image feature vector FC8, which is $1,000$D instead of the more common FC7 vector, which is $4,096$D. They report small but consistent improvements over a strong PBSMT baseline.

To the best of our knowledge, according to the 2016 multi-modal machine translation shared task, the best published results of a purely multi-modal NMT model are those of Huang et al. (2016), who proposed to use global visual features, obtained with the VGG19 network, extracted for an entire image and also for regions of the image obtained using the RCNN of Girshick et al. (2014). Their best model improves over a strong text-only attention-based NMT baseline and is comparable to results obtained with an SMT model trained on the same data. For this reason, we use their models as baselines in our experiments whenever appropriate.

Our work differs from previous work in that, first, we propose attention-based multi-modal NMT models. This is an important difference since the use of attention in NMT has become standard and is the current state-of-the-art (Luong et al., 2015; Jean et al., 2015; Firat et al., 2016; Sennrich et al., 2016b). Second, we study different forms to integrate both global and local image features into attention-based NMT. Third, in one branch of our work we propose a *doubly-attentive decoder* where we effectively fuse two mono-modal attention mechanisms into one multi-modal decoder (this is directly related to our research question **(RQ2)** *Given that there is a large number of standard text-only MT corpora, can multi-modal MT models effectively exploit that additional text-only data and provide state-of-the-art performance?*). We train the entire model jointly and end-to-end but still preserve the two independent attention mechanisms, differently from Caglayan et al. (2016). We argue that maintaining the attention mechanisms independent is key to been able to effectively pre-train our models on large text-only MT corpora. Finally, we are interested in how to merge textual and visual representations into multi-modal representations when generating words in the target language, which differs substantially from text-only translation tasks even when these translate from many source language into many target languages (Dong et al., 2015; Firat et al., 2016).

In the next two chapters, we introduce the data sets used in this work (Chapter 3) and put forward the mathematical notation and baseline NMT model used to derive our multi-modal NMT models (Chapter 4). In Chapter 5 we present our Multilingual Multi-modal Embedding model, which can be interpreted as a multi-modal sentence-level DSM. Further on, in Chapters 6 and 7 we introduce multi-modal NMT models that incorporate global and local image features, respectively.

Chapter 3

Data sets

In this chapter, we introduce and discuss important characteristics of the data sets used in this work. These are the Multi30k (Section 3.1), the WMT 2015 English-German data (Section 3.2) and the eBay data sets (Section 3.3). We also provide a quick discussion (Section 3.3.3) of the eBay data sets, as well as a small qualitative evaluation of its contents.

In our work, we need different types of data according to the different models we propose, and we propose and evaluate two types of models in this work:

- discriminative multilingual and multi-modal neural ranking models; and
- multi-modal NMT models.

Each of these two types of models use similar but ultimately different data:

1. The discriminative ranking models need multilingual sentences accompanied by an image. These multilingual sentences need not be parallel, i.e. the sentences need not be translation pairs, as long as they all describe the same image.
2. The multi-modal NMT models need parallel bilingual sentences and an image, and typically will also need a test set with parallel sentences and an image for model evaluation.

The main difference between the discriminative neural ranking and multi-modal NMT models is that the former needs only multilingual sentences with images, whereas the latter also needs the multilingual sentences to be parallel, i.e. translation pairs. We now introduce some corpora and/or data sets we use in our work to train and evaluate the different families of models we have just briefly described.

3.1 Multi30k

The original Flickr30k data set contains ∼30K images and 5 English sentence descriptions for each image (Young et al., 2014). Images were collected from Flickr and their descriptions were obtained by asking humans to describe the contents of image. Recently, Elliott et al. (2016) released two multilingual expansions of the original Flickr30k, which we call the translated and the comparable Multi30k datasets, henceforth referred to as M30k_T and M30k_C, respectively.

We note that both the M30k_T and the M30k_C data sets were just recently released as the official data to support the first shared task on multi-modal machine translation (Specia et al., 2016), as part of the WMT 2016.[1]

3.1.1 Translated Multi30k (M30k_T)

For each of the images in the Flickr30k, the M30k_T has one of its English descriptions manually translated into German by a professional translator. Training, validation and test sets contain $29,000$, $1,014$ and $1,000$ images, respectively, each accompanied by one sentence pair (the original English sentence and its German translation).

Since this data set contains images and bilingual parallel sentence descriptions, it is used for training our multi-modal NMT models. In Table 3.1, we show some statistics for the M30k_T training data set as well as its coverage compared to the Multi30k development and test sets. We highlight that words with a frequency of

[1]Both data sets were released in January, 2016.

one account for a maximum of 2% of the corpus, as well as words with a frequency less or equal to five account for a maximum of 5% of the corpus.

	English		**German**	
Training set				
# words (total)	377,501		375,048	
Vocabulary size	10,748		14,863	
Average word frequency	35.1		25.2	
# words with frequency=1	4,744	(1.2%)	7,577	(2.0%)
# words with frequency≤3	9,934	(2.6%)	14,526	(3.8%)
# words with frequency≤5	13,978	(3.7%)	19,120	(5.0%)
Development set				
# words (total)	13,308		13,331	
Vocabulary size	1,976		2,278	
Training set coverage	91.0%		87.6%	
Test set				
# words (total)	12,968		12,604	
Vocabulary size	1,913		2,139	
Training set coverage	91.7%		80.3%	

Table 3.1: Translated Multi30k training, development and test data sets statistics.

3.1.2 Comparable Multi30k (M30k_C)

For each of the 30K images in the Flickr30k, the M30k_C has five descriptions in German collected independently of the English descriptions. Training, validation and test sets contain 29,000, 1,014 and 1,000 images, respectively, each accompanied by five sentences in English and five sentences in German. One important difference between this data set and the M30k_T is that the latter contains parallel sentences describing images, whereas sentence descriptions in the M30k_C are comparable sentences, not translations.

In Table 3.2, we show some statistics for the M30k_C training data set. In this scenario, singletons, i.e. words with a frequency of one in the corpus, account for less than 1% of the text, and for comparison words with a frequency less or equal to five account for a maximum of 2.3% of the text. This practically halves the relative

	English		German	
# words (total)	1,943,430		1,624,398	
Vocabulary size	21,722		30,153	
Average word frequency	89.4		53.8	
# words with frequency=1	8,905	(0.4%)	15,022	(0.9%)
# words with frequency≤3	18,955	(0.9%)	28,501	(1.7%)
# words with frequency≤5	26,232	(1.3%)	37,458	(2.3%)

Table 3.2: Comparable Multi30k training data set statistics.

number of singletons and words with frequency ≤ 5 compared to the translated Multi30k. Overall, the Multi30k data sets do not contain much ambiguity, have a relatively small vocabulary and sentences with simple syntactic structures (Elliott et al., 2016). We highlight these characteristics to contrast with the eBay data, described in Section 3.3. The relative simplicity of the Multi30k data sets implies that translating it should be easier than translating the eBay data sets.

3.2 WMT 2015 English–German corpora

We also use the parallel English–German corpora released for the WMT 2015 translation task (Bojar et al., 2015) in some experiments when pre-training model $\text{NMT}_{\text{SRC+IMG}}$ (described in Section 7.2). These consist of three corpora: *the Europarl corpus* (Koehn, 2005), consisting of transcriptions of speeches from the European Parliament; *the Common Crawl corpus*, consisting of a large amount of text crawled from the Web; and *the News Commentary corpus*, consisting of news articles (Bojar et al., 2015). We remove any empty entries in these corpora and concatenate them together.

The final concatenated corpus contains 4,310,018 parallel sentences. In Table 3.3, we show some statistics for the concatenation of these corpora.

	English		German	
# words (total)	103,082,259		102,843,424	
Vocabulary size	936,706		1,750,853	
Average word frequency	110.0		58.7	
# words with frequency=1	523,309	(0.5%)	1,083,200	(1.0%)
# words with frequency≤3	957,790	(0.9%)	1,851,826	(1.8%)
# words with frequency≤5	1,216,372	(1.1%)	2,271,536	(2.2%)

Table 3.3: Some statistics for the concatenation of the Europarl, the Common Crawl and the News Commentary corpora.

3.3 eBay data sets

We now describe the data sets of product listings and images obtained in an agreement between eBay Inc. and the ADAPT Centre. These datasets are not publicly available and have been released only for the purposes of this research. In general, the eBay data sets consist of user-generated data and are very noisy in comparison to both the Multi30k and the WMT 2015 English-German data, making it harder to work with.

3.3.1 eBay24k

The eBay24k data set was curated based on product images and listings, both of them created by eBay users. Originally, 7,280 product listings in English and their accompanying images, both user-generated, were collected. The English listings were machine-translated into German using an in-house SMT model, and the MT output, i.e. translated German product listings, was post-edited by humans. Also, 17,526 product listings in German and their accompanying images, both user-generated, were collected. The German listings were machine-translated into English using another in-house SMT model, and the output of this model, i.e. translated English product listings, was post-edited by humans.

The final eBay24k is the concatenation of both 7,280 ⟨English, post-edited German, image⟩ and 17,526 ⟨German, post-edited English, image⟩ entries. After cleaning the data, the eBay24k training set consists of 23,697 tuples of product listings

and images. The development and test sets contain respectively 480 and 444 randomly selected entries from the original curated data. Each entry in the training, development and test sets consists of: *(i)* a listing in English, *(ii)* a listing in German and *(iii)* a product image.

We refer to the 23,697 training set triples as the *eBay24k* data set. In Table 3.4, we show statistics for the eBay24k training data set. Note that these statistics are computed on *lowercased* text, whereas for the Multi30k ($M30k_T$ and $M30k_C$) and the WMT 2015 data, the same statistics are computed on *truecased* text.

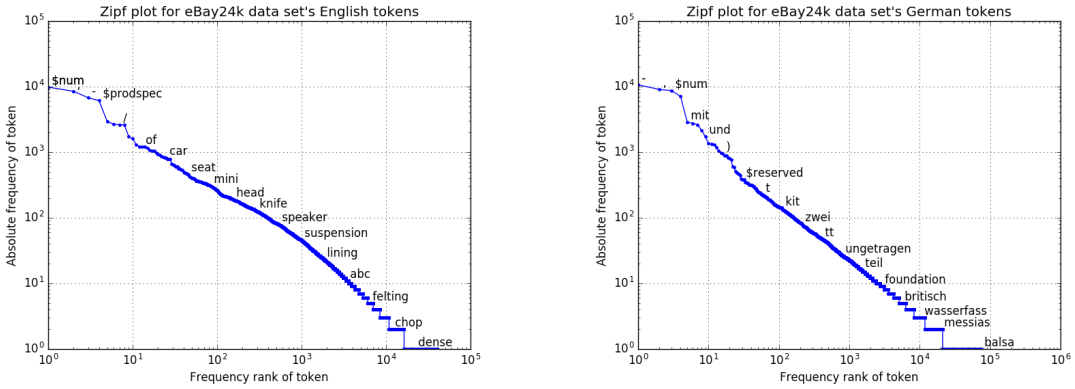
	English		**German**	
Training set				
# words (total)	299,903		258,638	
Vocabulary size	40,174		73,225	
Average word frequency	7.4		3.5	
# words with frequency=1	23,831	(7.9%)	52,078	(20.1%)
# words with frequency≤3	42,019	(14.0%)	81,360	(31.5%)
# words with frequency≤5	52,591	(17.5%)	95,401	(36.9%)
Development set				
# words (total)	6,396		6,518	
Vocabulary size	2,715		2,982	
eBay24k coverage	81.5%		71.3%	
eBay24k + eBay80k coverage	84.1%		79.8%	
Test set				
# words (total)	6,001		6,084	
Vocabulary size	2,632		2,849	
eBay24k coverage	80.8%		71.4%	
eBay24k + eBay80k coverage	84.3%		79.9%	

Table 3.4: eBay24k training, development and test data sets statistics.

In Figure 3.1, we present a graph showing the Zipf distribution (Powers, 1998) of the tokens in the eBay24k data set.

3.3.2 eBay80k

The curation of parallel product listings with an accompanying product image is costly and time-consuming, since it involves humans in the loop for post-editing



(a) English word frequencies.

(b) German word frequencies.

Figure 3.1: Word frequencies for the eBay24k data set.

the machine-translated listings, thus the relatively small number of entries in the eBay24k. More easily accessible are monolingual German product listings accompanied by the product image, since this type of data can be retrieved directly from eBay's data bases without need for additional post-editing effort.

As additional data, eBay released 83,832 tuples of German product listings and images, again all used-generated, henceforth the eBay80k data set. We follow Sennrich et al. (2016a) and back-translate the German listings into English using the text-only NMT baseline described in Chapter 4 trained on the textual part of the eBay24k data set, i.e. ⟨German, English⟩ listings. We then simply add the back-translated sentences to the original data set. Sennrich et al. (2016a) studied the application of back-translation to text-only NMT models and found that it can be useful when one needs to incorporate target language data into NMT models in the form of ⟨synthetic source sentence, original target sentence⟩. A back-translation model is trained to translate from the target into the source language, and used to translate the original target sentences into the source language. The synthetic source is then just added together with the original target sentence to the original training data and used as is. We refer to these 83,832 triples ⟨original image, synthetic English listing, original German listing⟩ as the *back-translated eBay80k* data set.

Finally, since we do not use the eBay80k by itself as a training set, but instead use

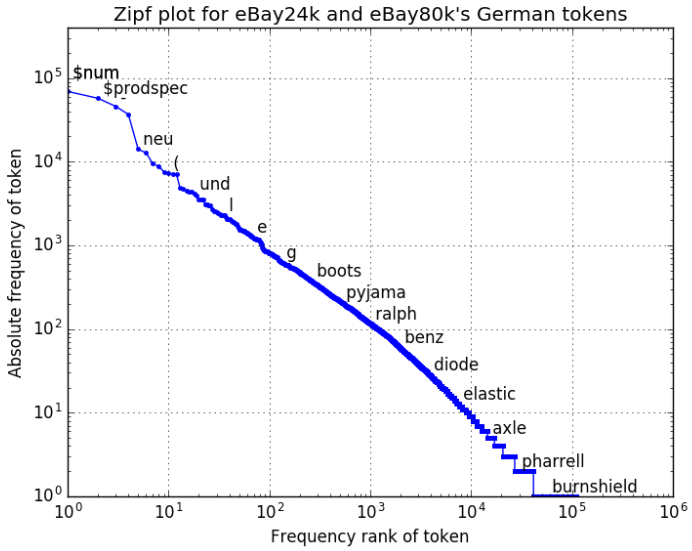


Figure 3.2: German word frequencies for the concatenation of the eBay24k and the eBay80k data sets.

the concatenation of the eBay24k and the eBay80k, in Table 3.5 we show statistics for this concatenated data set. Note that once again these statistics are computed on *lowercased* text. Similarly to the eBay24k, in Figure 3.2 we show a graph showing the Zipf distribution (Powers, 1998) of the German tokens in the concatenation of the eBay24k and the eBay80k data sets.

	English		**German**	
# words (total)	1,127,955		1,017,371	
Vocabulary size	71,448		106,751	
Average word frequency	15.7		9.5	
# words with frequency=1	45,458	(4.0%)	66,763	(6.6%)
# words with frequency≤3	74,216	(6.6%)	113,381	(11.1%)
# words with frequency≤5	89,596	(7.9%)	139,540	(13.7%)

Table 3.5: Concatenation of the eBay24k and eBay80k data sets statistics.

3.3.3 Discussion

Translating user-generated product titles has particular challenges; they are often ungrammatical and can be difficult to interpret in isolation even by a native speaker of the language, as can be seen in Table 3.6.

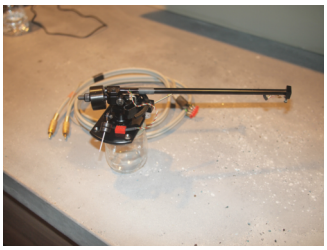
Image	Language	Product Description
	(en)	mary kay cheek color mineral pick citrus bloom shy blush bold berry + more
	(de)	mary kay mineral cheek colour farbauswahl citrus bloom shy blush bold berry + mehr
	(en)	just rewired original mission 774 fluid damped low mass tonearm , very good cond .
	(de)	vor kurzem neu verkabelter flüssigkeitsgedämpfter leichter original - mission 774 - tonarm , sehr guter zustand

Table 3.6: Example of two product listings and their corresponding image.

We note that the average word frequencies for German are very low (3.5 for the eBay24k and 9.5 for the concatenation of the eBay24k and eBay80k), especially when compared to the $M30k_T$ (25.2), the $M30k_C$ (53.8), or the WMT 2015 data (58.7). Also, the number of low frequency words in the eBay24k data set is very high: 36.9% of the German words appear a maximum of 5 times in the whole data set. Concatenating it with the eBay80k helps to lower this quantity to 13.7%, but that is still very high compared to the $M30k_T$ (5%), the $M30k_C$ (2.3%), or the WMT 2015 data (2.2%).

To further demonstrate this issue, in Table 3.7 we show perplexity scores obtained with LMs trained on three sets of different German corpora: the $M30k_C$ (Section 3.1.2), eBay's in-domain data (the concatenation of the German listings in the eBay24k and eBay80k data sets) and a concatenation of the German sentences in the WMT 2015 English–German parallel corpora (Section 3.2). These are 5-gram LMs trained with KenLM (Heafield et al., 2013) using modified Kneser-Ney smoothing (Kneser and Ney, 1995) on tokenized, lowercased data. We see that different LM perplexities on the eBay24k test set are high even for an LM trained on the eBay's in-domain data. These perplexity scores indicate that *fluency* might not be a good metric to use in this part of our study, i.e. we should not expect a fluent

machine-translated output of a model trained on poorly fluent training data.

LM training corpus	**#sentences** (×1000)	**Perplexity** (×1000)	
		eBay24k	**Multi30k**
WMT'15	4310.0	60.1	0.5
$M30k_C$	29.0	25.2	0.05
eBay24k	99.0	**1.8**	4.2

Table 3.7: Perplexity of eBay24k and Multi30k's test sets using LMs trained on different corpora. WMT'15 is the concatenation of the Europarl, Common Crawl and News Commentary corpora (the German side of the parallel English–German corpora).

Clearly, user-generated product listings are not very fluent in terms of grammar or even predictable word order. To better understand whether this has an impact on semantic intelligibility, we also wanted to assess how challenging to understand they are for a human reader. Accordingly, we asked humans how they perceive product listings with and without having the associated images available, under the hypothesis that images bring additional understanding to their corresponding listings.

Native speakers are presented with an English (German) product listing. Half of them are also shown the product image, whereas the other half is not. For the first group, we ask two questions: *(i)* in the context of the product image, how easy is to understand the English (German) product listing (i.e. *difficulty*) and *(ii)* how well does the English (German) product listing describe the product image (i.e. *adequacy*). For the second group, we just ask *(i)* how easy is to understand the English (German) product listing. In all cases humans must select from a five-level Likert scale where in *(i)* answers range from *1–Very easy* to *5–Very difficult* and in *(ii)* from *1–Very well* to *5–Very poorly*.

Table 3.8 suggests that the intelligibility of both the English and German product listings are perceived to be somewhere between "easy" and "neutral" when images are also available. It is notable that, in the case of German listings, there is a statistically significant difference between the group that had access to the image

Language	N	Difficulty		Adequacy
		listing only	listing+image	listing+image
English	20	2.50 ± 0.84	2.40 ± 0.84	2.45 ± 0.49
German	15	2.83 ± 0.75	**2.00** ± 0.50	2.39 ± 0.78

Table 3.8: Difficulty in understanding product titles with and without images and adequacy of product titles and images. N is the number of raters.

and the product listing (M=2.00, SD=.50) and the group that only viewed the listings (M=2.83, ST=.30), where $F(1,13) = 6.72$, $p < 0.05$. Furthermore, humans find that product listings describe the associated image somewhere between "well" and "neutral" with no statistically significant differences between the adequacy of product listings and images in different languages.

Altogether, we have a strong indication that images can indeed help an MT model translate product listings, especially for translations into German.

We now move on to Chapter 4, where we introduce the mathematical notation adopted in this work, as well as the baseline NMT model used as a basis to derive our multi-modal NMT models.

Chapter 4

Notation and Baseline NMT

In this chapter, we formalise the notation used in our multi-modal NMT models throughout our work. We follow the notation introduced by Bahdanau et al. (2015) and Firat et al. (2016), and now describe the text-only NMT model used as a baseline in most of our experiments and also as a basis to derive our multi-modal models in Chapters 6 and 7.

4.1 Text-only Neural Machine Translation

Bahdanau et al. (2015) first introduced an attention mechanism into the NMT encoder–decoder framework. Given a source sequence $X = (x_1, x_2, \cdots, x_N)$ and its translation $Y = (y_1, y_2, \cdots, y_M)$, an NMT model aims at building a single neural network that translates X into Y by directly learning to model $p(Y \mid X)$. Each x_i is a row index in a source lookup matrix $\boldsymbol{W}_x \in \mathbb{R}^{|V_x| \times d_x}$, the *source word embeddings matrix*, and each y_j is an index in a target lookup matrix $\boldsymbol{W}_y \in \mathbb{R}^{|V_y| \times d_y}$, the *target word embeddings matrix*. V_x and V_y are source and target vocabularies and d_x and d_y are source and target word embeddings dimensionalities, respectively.

A bidirectional RNN with GRU is used as the encoder. A forward RNN $\overrightarrow{\Phi}_{\text{enc}}$ reads X word by word, from left to right, and generates a sequence of *forward annotation vectors* $(\overrightarrow{\boldsymbol{h}}_1, \overrightarrow{\boldsymbol{h}}_2, \cdots, \overrightarrow{\boldsymbol{h}}_N)$ at each encoder time step $i \in [1, N]$. Similarly,

a backward RNN $\overleftarrow{\Phi}_{\text{enc}}$ reads X from right to left, word by word, and generates a sequence of *backward annotation vectors* $(\overleftarrow{\boldsymbol{h}}_1, \overleftarrow{\boldsymbol{h}}_2, \cdots, \overleftarrow{\boldsymbol{h}}_N)$, as in (4.1):

$$\begin{aligned}\overrightarrow{\boldsymbol{h}_i} &= \overrightarrow{\Phi}_{\text{enc}}\big(\boldsymbol{W}_x[x_i], \overrightarrow{\boldsymbol{h}}_{i-1}\big),\\ \overleftarrow{\boldsymbol{h}_i} &= \overleftarrow{\Phi}_{\text{enc}}\big(\boldsymbol{W}_x[x_i], \overleftarrow{\boldsymbol{h}}_{i+1}\big).\end{aligned} \tag{4.1}$$

The final annotation vector for a given time step i is the concatenation of forward and backward vectors, as shown in (4.2):

$$\boldsymbol{h}_i = \big[\overrightarrow{\boldsymbol{h}_i}; \overleftarrow{\boldsymbol{h}_i}\big]. \tag{4.2}$$

In other words, each source sequence X is encoded into a sequence of annotation vectors $C = (\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_N)$, which are in turn used by the decoder: essentially a neural language model (LM) (Bengio et al., 2003) conditioned on the previously emitted words and the source sentence via an attention mechanism.

A multilayer perceptron (MLP) is used to initialise the decoder's hidden state $\boldsymbol{s}_0$ at time step $t = 0$, where the input to this network is the concatenation of the last forward and backward vectors $\big[\overrightarrow{\boldsymbol{h}_N}; \overleftarrow{\boldsymbol{h}_1}\big]$ computed in Equation (4.2). The MLP is described in Equation (4.3):

$$\boldsymbol{s}_0 = \tanh\big(\boldsymbol{W}_{di}[\overrightarrow{\boldsymbol{h}_N}; \overleftarrow{\boldsymbol{h}_1}] + \boldsymbol{b}_{di}\big), \tag{4.3}$$

where $\boldsymbol{W}_{di}$ and $\boldsymbol{b}_{di}$ are model parameters. Since RNNs normally better store information about recent inputs in comparison to more distant ones (Hochreiter and Schmidhuber, 1997; Bahdanau et al., 2015), by using $\overleftarrow{\boldsymbol{h}_1}$ and $\overrightarrow{\boldsymbol{h}_N}$ we expect to initialise the decoder's hidden state with a strong source sentence representation, i.e. a representation with a strong focus on both the first and the last tokens in the source sentence.

At each time step t of the decoder, a *time-dependent* source context vector $\boldsymbol{c}_t$ is computed based on the annotation vectors C and the decoder previous hidden

state $\boldsymbol{s}_{t-1}$. This is part of the formulation of the *conditional GRU* and is described further in Equation (4.2). In other words, the encoder is a bi-directional RNN with GRU and the decoder is an RNN with a conditional GRU.

Given a hidden state $\boldsymbol{s}_t$, the probabilities for the next target word are computed using one projection layer followed by a softmax, as illustrated in Equation (4.4):

$$p(y_t = k \mid \boldsymbol{y_{<t}}, X) \propto \exp(\boldsymbol{L_o} \tanh(\boldsymbol{L_s s}_t + \boldsymbol{L_w E_y}[\hat{y}_{t-1}] + \boldsymbol{L_c c_t})), \tag{4.4}$$

where matrices $\boldsymbol{L}_o$, $\boldsymbol{L}_s$, $\boldsymbol{L}_w$ and $\boldsymbol{L}_c$ are transformation matrices and $\boldsymbol{c}_t$ is a time-dependent source context vector generated by the conditional GRU.

4.2 Conditional Gated Recurrent Unit (GRU)

The conditional GRU,[1] illustrated in Figure 4.1, has three main components computed at each time step t of the decoder:

- REC_1 computes a hidden state proposal $\boldsymbol{s}'_t$ based on the previous hidden state $\boldsymbol{s}_{t-1}$ and the previously emitted word $\hat{y}_{t-1}$;
- ATT_{src}[2] is an attention mechanism over the hidden states of the source-language RNN and computes $\boldsymbol{c}_t$ using all source annotation vectors C and the hidden state proposal $\boldsymbol{s}'_t$;
- REC_2 computes the final hidden state $\boldsymbol{s}_t$ using the hidden state proposal $\boldsymbol{s}'_t$ and the time-dependent source context vector $\boldsymbol{c}_t$.

We use the conditional GRU in our text-only attention-based NMT model. First, a single-layer feed-forward network is used to compute an *expected alignment* $e^{\text{src}}_{t,i}$ between each source annotation vector $\boldsymbol{h}_i$ and the target word $\hat{y}_t$ to be emitted at

[1] `https://github.com/nyu-dl/dl4mt-tutorial/blob/master/docs/cgru.pdf`.

[2] ATT_{src} is named ATT in the original technical report.

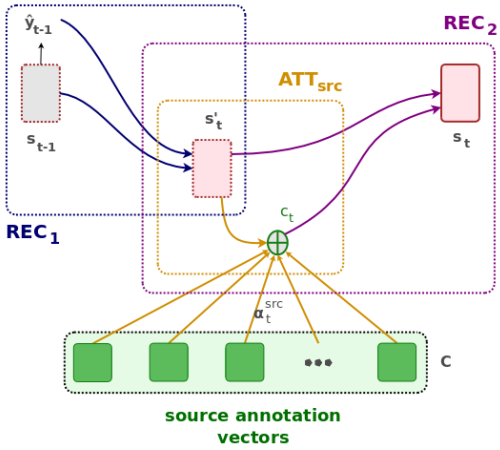


Figure 4.1: An illustration of the conditional GRU: the steps taken to compute the current hidden state $\boldsymbol{s}_t$ from the previous state $\boldsymbol{s}_{t-1}$, the previously emitted word $\hat{y}_{t-1}$, and the source annotation vectors C, including the candidate hidden state $\boldsymbol{s}'_t$ and the source-language attention vector $\boldsymbol{c}_t$.

the current time step t, as showed in Equations (4.5) and (4.6):

$$e^{\text{src}}_{t,i} = (\boldsymbol{v}^{\text{src}}_a)^T \tanh(\boldsymbol{U}^{\text{src}}_a \boldsymbol{s}'_t + \boldsymbol{W}^{\text{src}}_a \boldsymbol{h}_i), \tag{4.5}$$

$$\alpha^{\text{src}}_{t,i} = \frac{\exp\left(e^{\text{src}}_{t,i}\right)}{\sum_{j=1}^{N} \exp\left(e^{\text{src}}_{t,j}\right)}, \tag{4.6}$$

where $\alpha^{\text{src}}_{t,i}$ is the normalised alignment matrix between each source annotation vector $\boldsymbol{h}_i$ and the word $\hat{y}_t$ to be emitted at time step t, and $\boldsymbol{v}^{\text{src}}_a$, $\boldsymbol{U}^{\text{src}}_a$ and $\boldsymbol{W}^{\text{src}}_a$ are model parameters.

A time-dependent source context vector $\boldsymbol{c}_t$ is computed as a weighted sum over the source annotation vectors, where each vector is weighted by the attention weight $\alpha_{t,i}$, as in Equation 4.7:

$$\boldsymbol{c}_t = \sum_{i=1}^{N} \alpha_{t,i} \boldsymbol{h}_i. \tag{4.7}$$

Finally, we use the time-dependent source context vector $\boldsymbol{c}_t$ an input to REC_2, which computes the final hidden state $\boldsymbol{s}_t$ using the hidden state proposal $\boldsymbol{s}'_t$, and

the time-dependent source context vector $\boldsymbol{c}_t$, as in Equation (4.8):

$$\begin{aligned}
\boldsymbol{r}_t &= \sigma(\boldsymbol{W}_r^{\text{src}}\boldsymbol{c}_t + \boldsymbol{U}_r\boldsymbol{s}'_t), \\
\boldsymbol{z}_t &= \sigma(\boldsymbol{W}_z^{\text{src}}\boldsymbol{c}_t + \boldsymbol{U}_z\boldsymbol{s}'_t), \\
\underline{\boldsymbol{s}}_t &= \tanh(\boldsymbol{W}^{\text{src}}\boldsymbol{c}_t + \boldsymbol{r}_t \odot (\boldsymbol{U}\boldsymbol{s}'_t)), \\
\boldsymbol{s}_t &= (1 - \boldsymbol{z}_t) \odot \underline{\boldsymbol{s}}_t + \boldsymbol{z}_t \odot \boldsymbol{s}'_t,
\end{aligned} \tag{4.8}$$

where the parameters $\boldsymbol{W}_r^{\text{src}}$, $\boldsymbol{W}_z^{\text{src}}$, $\boldsymbol{U}_r$, $\boldsymbol{U}_z$, $\boldsymbol{W}^{\text{src}}$ and $\boldsymbol{U}$ are trained with the model.

Chapter 5

Multilingual Multi-modal Embedding

In this chapter, we introduce a model to train embeddings that are both multilingual and multi-modal. This model is a first step towards integrating multilingual linguistic content and visual content in a fully-fledged NMT framework. It directly addresses our research question **(RQ1)** *Can we use multi-modal discriminative models to improve the translation of image descriptions?* We train one model and report experiments on applying it on three downstream tasks: image-sentence ranking, semantic textual similarity and NMT. We find that it can effectively be used not only to improve the translation of image descriptions in a n-best list re-ranking scenario, but also in other NLP tasks.

Distributional semantic models (DSMs) compute word vector representations from text based on word co-occurrence patterns. However, these models suffer from an obvious limitation since the meaning of a word is derived entirely from connections to other words, i.e. they do not take extra-linguistic modalities into account and thus lack *grounding* (Glenberg and Robertson, 2000). This is the case not only for widely adopted word-level DSMs, e.g. word2vec (Mikolov et al., 2013), but also for sentence-level DSMs, e.g. skip-thought vectors (Kiros et al., 2015), order embeddings (Vendrov et al., 2016).

In this chapter, we address this issue and expand on the idea of training sentence-level multi-modal embeddings (Kiros et al., 2014; Socher et al., 2014), introducing a model that can be trained not only on images and their monolingual descriptions but also on additional multilingual image descriptions when these are available, henceforth the Multilingual Multi-modal Embedding (MLMME) model. We believe that having multiple descriptions of one image, regardless of its language, is likely to increase the coverage and variability of ideas described in the image, which may lead to a better generalisation of the depicted scene semantics. Moreover, a similar description expressed in different languages may differ in subtle but meaningful ways. By applying the proposed model, we expect that the embedding obtained from an image and the embeddings of its multilingual descriptions be *close* to one another.

To that end, we introduce a novel training objective function that uses pairwise ranking (Cohen et al., 1999) adapted to the case of three or more input sources, i.e. an image and multilingual sentences (Section 5.1). Our objective function links images and multiple sentences in an arbitrary number of languages, and we validate our idea in experiments where we use the Multi30k data set (Chapter 3).

We evaluate our embeddings in three different tasks (Section 5.3):

- an image–sentence ranking (ISR) task, in both directions, where we find that multilingual signals improve ISR to a large extent, i.e. the median ranks for English are improved from 8 to 5 and for German from 11 to 6, although the impact on ranking sentences given images is less conclusive (Section 5.3.1);

- two sentence textual similarity (STS) tasks, finding consistent improvements over a comparable monolingual baseline and outperforming the best published SemEval results (Section 5.3.2);

- a neural machine translation (NMT) task, where we use our model to re-rank n-best lists generated by different NMT models and report consistent improvements (Section 5.3.3).

5.1 Model description

Our MLMME model is composed of two main components: one *textual* and one *visual*. In the textual component, we have K different languages L_k, $k \in [1, K]$, and for each language we use a recurrent neural network (RNN) with gated recurrent units (GRU) (Cho et al., 2014b) as a sentence encoder. Let $S^k = \{w_1^k, \ldots, w_{N_k}^k\}$ denote sentences composed of word indices in a language L_k, and $X^k = (\boldsymbol{x}_1^k, \boldsymbol{x}_2^k, \cdots, \boldsymbol{x}_{N_k}^k)$ the corresponding word embeddings for these sentences, where N_k is the sentence length. An RNN Φ_{enc}^k reads X^k word by word, from left to right, and generates a sequence of annotation vectors $(\boldsymbol{h}_1^k, \boldsymbol{h}_2^k, \cdots, \boldsymbol{h}_{N_k}^k)$ for each embedding $\boldsymbol{x}_i^k$, $i \in [1, N_k]$. For any given input sentence, we use the corresponding encoder RNN last annotation vector $\boldsymbol{h}_{N_k}^k$ for that language L_k as the sentence representation, henceforth $\boldsymbol{v}^k$.

In our visual component we use publicly available pre-trained models for image feature extraction. In this model we use the 19-layer VGG network (VGG19) (Simonyan and Zisserman, 2014) to extract feature vectors for all images in our dataset. More specifically, we use global image features extracted from the penultimate fully-connected layer of the VGG19 network, which consists of a $4,096$D feature vector, henceforth FC7.

Each training example consists of a tuple *(i)* sentences S^k in L_k, $\forall k \in [1, K]$, and *(ii)* the associated image these sentences describe. Given a training instance, we retrieve the embeddings $X^k = \{\boldsymbol{x}_1^k, \ldots, \boldsymbol{x}_{N_k}^k\}$ for each sentence S^k using one separate word embedding matrix for each language k. A sentence embedding representation $\boldsymbol{v}^k$ is then obtained by applying the encoder Φ_{enc}^k onto each embedding $\boldsymbol{x}_{1:N_k}^k$ and using the last annotation vector $\boldsymbol{h}_{N_k}^k$ of each RNN, after it has consumed the last token in each sentence. Note that our encoder RNNs for different languages share no parameters. An image feature vector $\boldsymbol{q} \in \mathbb{R}^{4096}$ is extracted using a pre-trained CNN (i.e., this corresponds to the abovementioned FC7 feature vector) so that $\boldsymbol{d} = W_I \cdot \boldsymbol{q}$ is an image embedding and W_I is an image transformation matrix trained with the

model. Also, image embeddings $\boldsymbol{d}$ and sentence embeddings $\boldsymbol{v}^k$, $\forall k \in [1, K]$, are normalised to unit norm and have the same dimensionality. Finally, $s_i(\boldsymbol{d}, \boldsymbol{v}^k) = \boldsymbol{d} \cdot \boldsymbol{v}^k$, $\forall k \in [1, K]$ is a function that computes the similarity between images and sentences in all languages, and $s_s(\boldsymbol{v}^k, \boldsymbol{v}^l) = \boldsymbol{v}^k \cdot \boldsymbol{v}^l$, $\forall k, l \in [1, K], k \neq l$ computes the similarity between sentences in two different languages.

We now describe two *pairwise ranking* functions used in our objective, one that scores sentences and images, and another one that scores sentences in two different languages. Our model takes into consideration not only the relation between sentences in a given language and images—computed by the $s_i(\cdot, \cdot)$ function—, but also sentences in different languages in relation to each other, computed by the $s_s(\cdot, \cdot)$ function. Our sentence–image (i.e., *multi-modal*) ranking function is given in Equation (5.1):

$$\begin{aligned} \mathrm{R_{MM}} = & \sum_{k=1}^{K} \sum_{j} \sum_{r} \max\{0, \alpha - s_i(\boldsymbol{d}_j, \boldsymbol{v}_j^k) + s_i(\boldsymbol{d}_j, \boldsymbol{v}_r^k)\} + \\ & \sum_{k=1}^{K} \sum_{j} \sum_{r} \max\{0, \alpha - s_i(\boldsymbol{d}_j, \boldsymbol{v}_j^k) + s_i(\boldsymbol{d}_r, \boldsymbol{v}_j^k)\}, \end{aligned} \tag{5.1}$$

where ($\boldsymbol{v}_j^k$,$\boldsymbol{d}_j$) is a positive image–sentence pair, $\boldsymbol{v}_r^k$ (subscript r for *random*) is a contrastive or non-descriptive sentence embedding in language L_k for image embedding $\boldsymbol{d}_j$ and vice-versa, and α is a model parameter, i.e. the *margin*. $\mathrm{R_{MM}}$ learns to rank a sentence embedding $\boldsymbol{v}_j^k$ in any language L_k, $k \in [1, K]$, against an image embedding $\boldsymbol{d}_j$, and vice-versa.

Our sentence–sentence (*multilingual*) ranking function is given in Equation (5.2):

$$\begin{aligned} \mathrm{R_{ML}} = & \sum_{k,l=1}^{K} \sum_{j} \sum_{r} \max\{0, \alpha - s_s(\boldsymbol{v}_j^k, \boldsymbol{v}_j^l) + s_s(\boldsymbol{v}_j^k, \boldsymbol{v}_r^l)\} + \\ & \sum_{k,l=1}^{K} \sum_{j} \sum_{r} \max\{0, \alpha - s_s(\boldsymbol{v}_j^l, \boldsymbol{v}_j^k) + s_s(\boldsymbol{v}_j^l, \boldsymbol{v}_r^k)\}, \\ & k \neq l, \end{aligned} \tag{5.2}$$

where $(\boldsymbol{v}_j^k, \boldsymbol{v}_j^l)$ is a positive sentence pair in languages L_k and L_l, respectively, $\boldsymbol{v}_r^k$ is a contrastive or non-descriptive sentence embedding in language L_k for sentence $\boldsymbol{v}_j^l$ in language L_l and vice-versa. In both $\mathrm{R_{MM}}$ and $\mathrm{R_{ML}}$, contrastive terms are chosen randomly from the training set and resampled at every epoch.

Finally, our optimisation function in Equation (5.3) minimises the linearly weighted combination of $\mathrm{R_{MM}}$ and $\mathrm{R_{ML}}$:

$$\min_{\theta_k, W_I} \beta \mathrm{R_{MM}} + (1-\beta)\mathrm{R_{ML}}, \forall k \in [1, K],$$
$$0 \geq \beta \geq 1, \tag{5.3}$$

where θ_k includes all the text encoder RNNs parameters for language L_k, and W_I is the image transformation matrix. β is a model hyperparameter that controls how much influence a particular similarity (*multi-modal* or *multilingual*) has in the overall cost. We illustrate the model in Figure 5.1.

The two extreme scenarios are $\beta = 0$, in which case only the multilingual similarity is used, and $\beta = 1$, in which case only the multi-modal similarity is used. If the number of languages $K = 1$ and $\beta = 1$, our model computes the monolingual Visual Semantic Embedding (VSE) of Kiros et al. (2014).

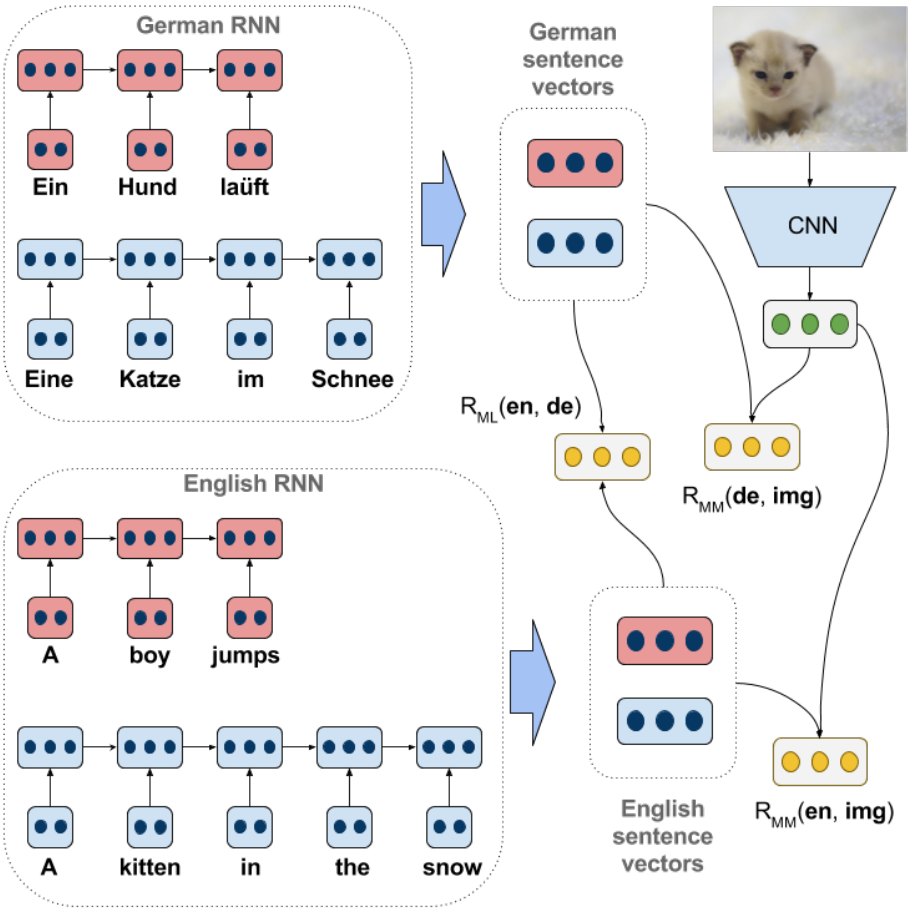


Figure 5.1: Multilingual multi-modal embedding trained with images and their English and German descriptions. The sentences in red denote contrastive examples, whereas the sentences in blue are descriptive of the image.

5.2 Experimental setup

For each language we train a separate encoder RNN with GRU with a 1024D hidden layer. Word embeddings are 620-dimensional and trained jointly with the model. All non-recurrent matrices are initialised by sampling from a Gaussian ($\mu = 0, \sigma = 0.01$), recurrent matrices are random orthogonal and bias vectors are all initialised to $\vec{0}$. We apply dropout (Srivastava et al., 2014) with a probability of 0.5 in both text and image representations, which are in turn mapped onto a 2048D multi-modal embedding space. We set the margin $\alpha = 0.2$ and the number of randomly sampled instances is $r = 127$. Our models are trained using stochastic gradient descent with Adam (Kingma and Ba, 2014) and minibatches of 128 instances.

	English							German				
	Skip-T.	VSE		Ours				VSE	Ours			
		paper	current	β=1	β=.75	$\beta = 0.5$	$\beta = 0.25$	current	β=1	β=.75	$\beta = 0.5$	$\beta = 0.25$
Sentence to image												
r@1	18.2	16.8	16.5	23.0 (+6.2)	**24.9** (+8.1)	22.3 (+5.5)	21.3 (+4.5)	13.5	**21.6** (+8.1)	20.3 (+6.8)	20.3 (+6.8)	19.5 (+6.0)
r@5	41.9	42.0	41.9	49.3 (+7.3)	**52.3** (+10.3)	48.3 (+6.3)	45.5 (+3.5)	36.6	**48.8** (+12.2)	45.0 (+8.4)	43.7 (+7.1)	43.0 (+6.4)
r@10	53.5	56.5	54.4	61.1 (+4.6)	**63.6** (+7.1)	58.4 (+1.9)	56.7 (+0.2)	49.0	**59.5** (+10.5)	56.6 (+7.6)	55.4 (+6.4)	54.4 (+5.4)
mrank	9	8	9	6	**5**	6	7	11	**6**	7	8	8
Image to sentence												
r@1	26.8	23.0	30.7	**33.1** (+2.4)	30.7 (+0.0)	27.4 (−3.3)	26.7 (−4.0)	30.5	**32.3** (+1.7)	24.9 (−5.6)	23.0 (−7.5)	21.8 (−8.7)
r@5	54.9	50.7	**57.8**	57.2 (−0.6)	55.4 (−2.4)	54.5 (−3.3)	51.4 (−6.4)	56.0	**58.6** (+2.6)	52.3 (−3.7)	48.4 (−7.6)	49.8 (−6.2)
r@10	67.5	62.9	**70.6**	68.7 (−1.9)	65.6 (−5.0)	64.0 (−6.6)	61.9 (−8.7)	**68.9**	68.1 (−0.8)	63.6 (−5.3)	62.8 (−6.1)	61.3 (−7.6)
mrank	5	5	**4**	**4**	**4**	**4**	5	**4**	**4**	5	6	6

Table 5.1: Two monolingual baselines, one is the Skip-thought vectors (Skip-T.) of Kiros et al. (2015), the other is the VSE model of Kiros et al. (2014), and our MLMME model on the M30k$_C$ test set. Best monolingual results are underlined and best overall results appear in bold. We show improvements over the best monolingual baseline in parenthesis. Best viewed in colour.

5.3 Results and Analysis

As our main baseline, we retrain Kiros et al. (2014) monolingual models separately on the M30k$_C$'s English and German sentences (+images), whereas model MLMME is trained on the entire M30k$_C$.

When processing English sentences and images, we additionally use the pre-trained Skip-Thought vectors (Kiros et al., 2015), more specifically the 4800D vectors (*combine-skip*) as a second baseline. We follow the authors description on how to do that:[1] *(i)* we use their pre-trained encoders to compute the English sentence representations, i.e. a 4800D vector; *(ii)* we train (i.e. fine-tune) their image-sentence ranking model on the M30k$_C$ training set; *(iii)* we select the model with the best performance of the M30k$_C$ validation set and use it to compute results in the test set.

5.3.1 Image–Sentence Ranking

In image-sentence ranking, the task is to rank a set of images given a sentence so that the image best matching the sentence is ranked as high as possible (and vice-versa, for sentences given images). In Table 5.1, we show results for the monolingual English Skip-thought vectors of Kiros et al. (2015), the monolingual VSE English and German models of Kiros et al. (2014) and our MLMME models on the M30k$_C$ data

[1] `https://github.com/ryankiros/skip-thoughts#image-sentence-ranking`

set and evaluated on images and bilingual sentences. Recall-at-k ($r@k$) measures the mean number of times the correct result appear in the top-k retrieved entries and *mrank* is the median rank.

First, we note that multilingual models show consistent improvements in ranking images given sentences. All our models, regardless of the value of the hyperparameter β ($= .25, .5, .75, 1$), show strong improvements in $r@k$ (up to +12.2) and median rank (in English, the mrank is reduced from 8 to 5 and in German from 11 to 6 in comparison to the best model by Kiros et al. (2014)). Nevertheless, when ranking sentences given images, results are less conclusive. The best results achieved by our multilingual models, for both languages, are observed when $\beta = 1$, with the $r@k$ slightly deteriorating as we include more multilingual similarity, i.e. $\beta = .75, .5, .25$, and the median rank also slightly increasing for English (from 4 to 5) and German (from 4 to 6). In short, model MLMME consistently improves over all baselines when ranking images given sentences, and applying model MLMME with $\beta = 1$ to rank sentences given images performs comparably to the monolingual VSE baseline and clearly improves over using the Skip-Thought model on the same task.

Using image features is crucial in *grounding* the sentence vector representations. We note that using β=0 in Equation (5.3) is equivalent to using only multilingual similarity scores (eq. (5.2)), and no multi-modal similarities (eq. (5.1)). However, in the training data there are multiple sentences describing one same image, and by not using the multi-modal similarity the model loses the ability to generalise and project semantically similar sentences, i.e. sentences that describe one same image, close together. In other words, the model has no way of mapping the comparable sentences that describe one same image together, since the link between these sentences are the image they describe.

In practice, we note that using $\beta = 0$ leads to a model that cannot learn to rank sentences given images and vice-versa. This happens because the optimisation function $\min_{\theta_k, W_I} \beta \mathrm{R_{MM}} + (1 - \beta)\mathrm{R_{ML}}$, degenerates into $\min_{\theta_k, W_I} \mathrm{R_{ML}}$, $\forall k \in [1, K]$, i.e. it does not take any multi-modal similarity into consideration. For this reason,

we do not include $\beta = 0$ in our hyperparameter search in our experiments.

5.3.2 Semantic Textual Similarity

In the semantic textual similarity (STS) task,[2] we use our model to compute the distance between a pair of sentences (distances are equivalent to cosine similarities and therefore lie in the $[0, 1]$ real interval). Gold standard scores for all test sets are given in the $[0, 5]$ interval, where 0 means complete dissimilarity and 5 complete similarity. We simply use the cosine similarity distance computed by our model and scale it by 5, directly comparing it to the gold standard scores.

We note that in our STS experiments, we use the same models trained on the M30k_C applied onto the image–sentence ranking task (Section 5.3.1). We report results for all semantic similarity tasks for which test sets are publicly available (Agirre et al., 2012, 2013, 2014, 2015, 2016). These test sets include excerpts from the news domain, machine translation evaluation, forum answers, video descriptions, among others. Some of these test sets are highly out-of-domain when compared to the images and their descriptions used to train our MLMME models. Moreover, since there is no SemEval data set including the German language, we only use the English SemEval test sets. As an illustration, in Table 5.2 we show examples of entries from the different test sets.

Specifically, we embed both sentences in each of the test sets with our English encoder, trained as part of the MLMME, and also the VSE English encoder as our main baseline. We note that the vocabulary of the MLMME and VSE models are derived from the M30k_C training data, and in case there are any out-of-vocabulary words in the test sets, they are replaced by a special UNK symbol.

Amongst all test sets, there are two in-domain similarity tasks—image description similarity for years 2014 and 2015—and all the other tasks can be considered general- or out-of-domain.

In Table 5.3, the entries corresponding to the corresponding year's best SemEval

[2] `http://alt.qcri.org/semeval2017/task1/`

SemEval 2012 (Agirre et al., 2012) – **MSRpar**	
Sent. 1	The problem likely will mean corrective changes before the shuttle fleet starts flying again .
Sent. 2	He said the problem needs to be corrected before the space shuttle fleet is cleared to fly again .
Score	4.4
SemEval 2013 (Agirre et al., 2013) – **OnWN**	
Sent. 1	measure the depth of a body of water
Sent. 2	any large deep body of water .
Score	0.8
SemEval 2014 (Agirre et al., 2014) – **Tweet-news**	
Sent. 1	Hollywood Accepts Chinese Censorship (Will Movies Get Any Better ?)
Sent. 2	In Hollywood Movies for China , Bureaucrats Want a Say
Score	2.4
SemEval 2014 (Agirre et al., 2014) – **Image descriptions**	
Sent. 1	A cat standing on tree branches .
Sent. 2	A black and white cat is high up on tree branches .
Score	3.6
SemEval 2015 (Agirre et al., 2015) – **Headlines**	
Sent. 1	The foundations of South Africa are built on Nelson Mandela 's memory
Sent. 2	Australian politicians lament over Nelson Mandela 's death
Score	1.3
SemEval 2015 (Agirre et al., 2015) – **Image descriptions**	
Sent. 1	The couple is sitting near the water in lawn chairs .
Sent. 2	The boy hops from one picnic table to the other in the park .
Score	0.0
SemEval 2016 (Agirre et al., 2016) – **Plagiarism**	
Sent. 1	There are two main approaches for dynamic programming .
Sent. 2	There are four steps in Dynamic Programming : 1 .
Score	1.0

Table 5.2: Example entries for different SemEval test sets (Agirre et al., 2012, 2013, 2014, 2015, 2016).

model are the ones reported by the official shared task at the time the official results were released.[3] In Table 5.3, we note that our multilingual model consistently improves on the monolingual baseline of Kiros et al. (2014) in the two in-domain similarity tasks, staying competitive even compared to the best performing model in the SemEval shared task (entries marked with a † in Table 5.3). In fact, the only time model MLMME outperforms the best comparable SemEval model is in the image description similarity tasks (in 2014, our best model achieves 0.826 Pearson rank correlation, whereas the best results in SemEval 2014 is 0.821; in 2015, our best model achieves 0.886 Pearson rank correlation, versus 0.864 for the best SemEval

[3]These best SemEval models are the ones which ranked first overall considering all test sets in each year, not necessarily the model that ranked first in the specific image description test set.

Test set	Kiros	Our model				SemEval
		β=1	β=.75	β=.5	β=.25	best model
SemEval 2012 (Agirre et al., 2012)						
MSRpar	**.083**	.017	.043	.031	.013	.630
MSRvid	.799†	.780†	.792†	**.809**†	.805†	.873
SMT Europarl	.420	.414	.426	**.446**†	.401	.528
OnWN	**.539**	.462	.473	.519	.496	.664
SMT news	**.376**	.346	.337	.340	.333	.493
SemEval 2013 (Agirre et al., 2013)						
FNWN	**.092**	.036	.014	.033	.079	.581
headlines	**.442**	.409	.391	.407	.388	.764
OnWN	.389	.544	.575	**.585**	.571	.752
SemEval 2014 (Agirre et al., 2014)						
deft-forum	**.339**	.239	.188	.230	.244	.482
deft-news	**.524**	.351	.401	.347	.390	.765
headlines	**.442**	.349	.350	.379	.391	.764
images	.791†	.797†	.819†	**.826**†	.817†	.821
OnWN	.520	.560	.556	.579	**.624**	.858
Tweet-news	.402	.345	.344	**.404**	.376	.763
SemEval 2015 (Agirre et al., 2015)						
answers–forums	.248	.231	.234	**.284**	.244	.739
answers–students	**.584**	.424	.444	.425	.459	.772
belief	**.488**	.460	.439	.455	.479	.749
headlines	.424	.409	.407	**.447**	.442	.825
images	.834†	.880†	.882†	.885†	**.886**†	.864
SemEval 2016 (Agirre et al., 2016)						
answer–answer	**.399**	.212	.253	.288	.362	.692
headlines	.314	**.316**	.282	.309	.303	.827
plagiarism	**.573**	.473	.502	.534	.515	.841
postediting	**.710**	.701	.685	.699	.680	.835
question–question	.336	**.353**	.332	.212	.252	.687

Table 5.3: Pearson rank correlation scores for semantic textual similarities in different SemEval test sets (Agirre et al., 2012, 2013, 2014, 2015, 2016). Best overall scores (ours vs. baseline) in bold. We underline a score in case it improves on the monolingual baseline of Kiros et al. (2014) and mark it with † in case its difference from the best SemEval result is less than 10%.

2015's results).

One interesting point to note is that the only two evaluation sets where the β parameter is monotonically aligned to the correlations with the human judgements are the two in-domain tasks (image description similarity in 2014 and 2015). In these two tasks, the monolingual baseline of Kiros et al. (2014) is the worst perfoming model, and the correlations with human judgements monotonically increase as we increase β from 1.0 to 0.25. In all other tasks, there is no monotonic relation between the value of β and the human judgements.

In general, results on general domain similarity tasks are mixed, e.g. answers or headlines, and both MME and MLMME show weak correlation with human judgements. It is noteworthy that all models, baseline and multilingual, perform far worse than the best corresponding SemEval model in virtually all general-domain tasks (see entries marked with † in Table 5.3). Only once one configuration of one of our models remained competitive according to the state-of-the-art, and that was our multilingual model with $\beta = 0.5$ in the Europarl SMT task (differences $< 10\%$ compared to the best performing model). When we consider only the general-domain similarity tasks, the monolingual baseline of Kiros et al. (2014) has a higher Pearson rank correlation about 54% of the time, i.e. our model performs better about 46% of the time.

5.3.3 Neural Machine Translation (NMT)

In this section, we study how to incorporate image features to re-rank n-best lists generated with text-only NMT models. Arguably, the main advantage of using discriminative models, e.g. MLMME, to re-rank n-best lists instead of directly training multi-modal NMT models (Chapters 6 and 7) is the time it takes for training. Whereas training the discriminative MLMME model on the Flickr30k data set takes $\sim$ 6 hours, training a multi-modal NMT model on the same data set usually takes many days. The discriminative MLMME model is trained for ranking, which is considerably faster than training an NMT model. A multi-modal NMT

model must, for each target word, compute an expensive `softmax` operation which involves a normalisation over the entire target vocabulary, which can cause training to take considerably more time.

With these experiments, we wish to address two main questions:

(i) how does the quality of the baseline MT model used to generate the n-best list affect the final results?

(ii) how does model MLMME, trained on images and their descriptions, perform when applied to re-rank n-best lists from an in-domain test set (i.e. image description data), but generated by an MT model trained on a different domain?

In *(i)*, our intuition is that an MT model that is too weak will not produce n-best lists good enough to lead to good results with re-ranking. On the other hand, if the model is already highly optimised (e.g., by running a grid search over many hyperparameters), it might be difficult to improve translations further with n-best re-ranking. In *(ii)*, we want to study whether using off-the-shelf (i.e., general-domain) MT models to generate the n-best lists could work as well as training an in-domain MT model.

5.3.3.1 Experimental setup

In order to answer the two questions above, we train different models on different sets of English–German data. In order to train baseline models that perform better or worse, we use different hyper-parameter settings to train the baseline text-only NMT model described in Chapter 4 on the M30k$_{\mathrm{T}}$ training set to translate from English into German. We use this model to generate n-best lists ($n = 20$) for each entry in the M30k$_{\mathrm{T}}$ validation and test sets. We use the monolingual VSE model of Kiros et al. (2014) trained on German sentences and images to compute the distance between them, and our MLMME models trained with $\beta \in \{0.25, 0.5, 0.75, 1.0\}$ to compute the distance between German and English sentences with $s_s(\cdot, \cdot)$, and between a German sentence and an image using $s_i(\cdot, \cdot)$, for all entries in the M30k$_{\mathrm{T}}$

validation and test sets. We then train an n-best list re-ranker on the M30k$_{\mathrm{T}}$ validation set's 20-best lists with k-best MIRA (Crammer and Singer, 2003; Cherry, 2012), and use the new distances as additional features to the original MT log-likelihood $p(Y \mid X)$. We finally apply the optimised weights to re-rank the test set's 20-best lists.

How does `n`-best re-ranking perform when `n`-best lists are generated by MT models of different quality? In order to address this question, we apply our discrimininative MLMME model trained on the comparable Multi30k data set to re-rank n-best lists generated by three different models.

We train one *weak* model, one *regular* model and one *optimised* model on the translated Multi30k training data set (without images) to translate from English into German (Equations 4.1–4.8). In order to train these three different models, we search for the best dropout and L2 regularisation weight combination by observing model performance on the validation set. The search space for the dropout hyper-parameter is the set $\{0.0, 0.1, 0.2, \ldots, 0.9, 1.0\}$, and for the L2 regularisation weight is the set $\{0.0, 1\mathrm{e}{-1}, 1\mathrm{e}{-2}, \ldots, 1\mathrm{e}{-9}, 1\mathrm{e}{-10}\}$.

The configuration which performs the worst is the one with no regularisation, i.e. dropout probability 0.0 and L2 regularisation weight 0.0; the configuration which performs the best on the translated Multi30k validation set also uses no L2 regularisation (i.e., L2 weight is 0.0), but dropout with probability 0.2.

Weak model Our *weak* model is the model described in Section 4.1 trained with no regularisation, i.e. L2 regularisation weight is 0.0 and dropout probability is 0.0. It corresponds to the model with the worst performance on the translated Multi30k validation set.

Regular model Our *regular* model is the model described in Section 4.1 with some medium-performance regularisation. Specifically, from the hyper-parameter search on the translated Multi30k validation set, we use a weight of $1\mathrm{e}{-8}$ to scale

the L2 regularisation term and a dropout of 0.5.

Optimised model Our *optimised* model is the model described in Section 4.1 that has the best performance on the translated Multi30k validation set, according to our dropout and L2 regularisation hyper-parameter search. This corresponds to the model with no L2 regularisation (i.e., L2 weight is 0.0), and dropout with probability 0.2.

How does model MLMME, trained on images and their descriptions, perform when applied to re-rank n-best lists of a test set of the same domain, but generated by an MT model trained on a different domain?

In order to address this question, we apply our discriminative model MLMME trained on the comparable Multi30k data set to re-rank n-best lists generated by a baseline text-only NMT model trained on data from a different domain.

We therefore train the baseline text-only NMT model described in Chapter 4 on data from the news domain. Specifically, we use the WMT 2015 English–German corpora described in Section 3.2 to train a text-only NMT baseline, and use this baseline to generate n-best lists.

5.3.3.2 Results

How does n-best re-ranking perform when n-best lists are generated by MT models of different quality? Following the abovementioned experimental setup, we first use multiple models to generate n-best lists ($n \in \{20, 50\}$) for each entry in the M30k_T validation and test sets. In this set of experiments we use three different models to generate the n-best lists (described in Section 5.3.3.1 above): the *weak model*, the *regular model*, and the *optimised model*. Second, we use the monolingual VSE model of Kiros et al. (2014) trained on German sentences and images to compute the distance between translations into German and images, for all entries in the M30k_T validation and test sets. We also use our MLMME models trained with $\beta \in \{0.25, 0.5, 0.75, 1.0\}$ to compute the distance between German and

English sentences with $s_s(\cdot, \cdot)$, and between a German sentence and an image using $s_i(\cdot, \cdot)$, for all entries in the M30k$_{\mathrm{T}}$ validation and test sets. We then train an n-best list re-ranker on the M30k$_{\mathrm{T}}$ validation set's 20-best (50-best) lists with k-best MIRA (Crammer and Singer, 2003; Cherry, 2012), and use the new distances as additional features with the original MT log-likelihood $p(Y \mid X)$. We finally apply the optimised weights to re-rank the test set's 20-best (50-best) lists.

In Table 5.4, we show results obtained with these different NMT systems: the *weak model*, the *regular model*, and the *optimised model*. We first note that the difference between 1-best translations obtained with the optimised and the weak model (best and the worst systems respectively), according to automatic MT metrics, is considerable: 9.6 BLEU, 9.2 METEOR, and 11.2 TER.

In order to measure the quality of the n-best lists generated by the different models, we also compute their oracle scores. The difference between the oracle scores for the n-best lists generated by the weak and the regular model is considerable: 8.8/10.4 BLEU, 7.9/8.0 METEOR, and 5.3/9.3 TER, for the 20-best and 50-best lists respectively. Nevertheless, the difference between the oracle scores for the n-best lists generated by the regular and the optimised model is not nearly as strong: 1.2/−0.3 BLEU, 0.0/0.1 METEOR, and 3.4/0.8 TER, again for the 20-best and 50-best lists respectively. However, when we analyse the metrics scores obtained by re-ranked models, we see a considerable difference between the improvements brought by VSE and MLMME features to the regular and optimised models.

Weak model First of all, when we use VSE features to re-rank n-best lists generated by the weak model, translations do not change much. MLMME features have a strong impact on METEOR scores, suggesting that they are making translations more adequate by improving their word-level recall. Using MLMME features to re-rank significantly improves METEOR in relation to the baseline and to the translations obtained with the VSE-features re-ranked model, for all values of β and for all n-best list sizes.

	Discriminative re-ranking?	N	BLEU		METEOR		TER	
NMT (weak model)	baseline	1	25.7		43.1		56.1	
	+ VSE	20	25.8	(+0.1)	43.2	(+0.1)	56.1	(-0.0)
	+ MLMME, $\beta = 1$	20	26.1	(+0.4)	**44.4**†‡	**(+1.3)**	55.5	(-0.6)
	+ MLMME, $\beta = 0.75$	20	26.1	(+0.4)	44.3†‡	(+1.2)	55.9	(-0.2)
	+ MLMME, $\beta = 0.5$	20	26.0	(+0.3)	43.9†‡	(+0.8)	55.9	(-0.2)
	+ MLMME, $\beta = 0.25$	20	**26.3**†‡	**(+0.6)**	44.3†‡	(+1.2)	**55.2**†‡	**(-0.9)**
	oracle	20	33.1		51.4		46.5	
	+ VSE	50	25.8	(+0.1)	43.5†	(+0.4)	56.1	(-0.0)
	+ MLMME, $\beta = 1$	50	26.2	(+0.5)	**44.6**†‡	**(+1.5)**	55.4	(-0.7)
	+ MLMME, $\beta = 0.75$	50	**26.4**†	**(+0.7)**	44.5†‡	(+1.4)	55.6	(-0.5)
	+ MLMME, $\beta = 0.5$	50	25.9	(+0.2)	43.9†	(+0.8)	55.9	(-0.0)
	+ MLMME, $\beta = 0.25$	50	**26.4**†‡	(+0.7)	44.5†‡	(+1.4)	**55.0**†‡	**(-1.1)**
	oracle	50	36.2		53.8		43.4	
NMT (regular model)	baseline	1	32.4		50.7		51.9	
	+ VSE	20	32.2	(-0.2)	50.7	(+0.0)	52.6	(+0.7)
	+ MLMME, $\beta = 1$	20	**33.8**†‡	**(+1.4)**	**51.4**†‡	**(+0.7)**	49.0‡	(-2.9)
	+ MLMME, $\beta = 0.75$	20	33.5‡	(+1.1)	51.3†‡	(+0.6)	49.0‡	(-2.9)
	+ MLMME, $\beta = 0.5$	20	**33.8**†‡	**(+1.4)**	**51.4**†‡	**(+0.7)**	**48.6**†‡	**(-3.3)**
	+ MLMME, $\beta = 0.25$	20	33.7‡	(+1.3)	**51.4**†‡	**(+0.7)**	49.4‡	(-2.5)
	oracle	20	41.9		59.3		41.2	
	+ VSE	50	32.7	(-0.3)	50.8	(+0.1)	51.4	(-0.5)
	+ MLMME, $\beta = 1$	50	**34.2**†‡	**(+1.8)**	**51.6**†‡	**(+0.9)**	48.3‡	(-3.6)
	+ MLMME, $\beta = 0.75$	50	34.1†‡	(+1.7)	**51.6**†‡	**(+0.9)**	47.6†‡	(-4.3)
	+ MLMME, $\beta = 0.5$	50	34.0†‡	(+1.6)	51.4†‡	(+0.7)	**47.3**†‡	**(-4.6)**
	+ MLMME, $\beta = 0.25$	50	34.1†‡	(+1.7)	**51.6**†‡	**(+0.9)**	48.5‡	(-3.4)
	oracle	50	46.6		61.8		34.1	
NMT (optimised model)	baseline	1	35.3		52.3		44.9	
	+ VSE	20	32.3	(-3.0)	49.8	(-2.5)	46.5	(+1.6)
	+ MLMME, $\beta = 1$	20	35.3‡	(+0.0)	**52.7**†‡	(+0.4)	**44.5**‡	(-0.4)
	+ MLMME, $\beta = 0.75$	20	35.2‡	(-0.1)	52.6‡	(+0.3)	44.6‡	(-0.3)
	+ MLMME, $\beta = 0.5$	20	35.1‡	(-0.2)	52.3‡	(+0.0)	44.9‡	(-0.0)
	+ MLMME, $\beta = 0.25$	20	**35.7**‡	**(+0.4)**	**52.7**‡	**(+0.4)**	**44.5**‡	**(-0.4)**
	oracle	20	43.2		59.7		37.8	
	+ VSE	50	30.7	(-4.6)	47.9	(-4.4)	48.6	(+3.7)
	+ MLMME, $\beta = 1$	50	35.4‡	(+0.1)	**52.7**†‡	**(+0.4)**	**44.4**†‡	(-0.5)
	+ MLMME, $\beta = 0.75$	50	35.2‡	(-0.1)	52.5‡	(+0.2)	44.7‡	(-0.2)
	+ MLMME, $\beta = 0.5$	50	35.1‡	(-0.2)	52.3‡	(+0.0)	44.7‡	(-0.2)
	+ MLMME, $\beta = 0.25$	50	**35.6**‡	**(+0.3)**	52.6‡	(+0.3)	**44.4**†‡	(-0.5)
	oracle	50	46.3		61.9		34.9	

Table 5.4: MT evaluation metrics computed for 1-best translations generated with three baseline NMT models and for 20- and 50-best lists generated by the same models, re-ranked using VSE and MLMME as discriminative features. Results improve significantly over the corresponding 1-best baseline (†) or over the translations obtained with the VSE re-ranker (‡) with $p = 0.05$.

The model re-ranked with MLMME features with $\beta = 0.25$ is clearly the best performing one in this scenario. It is the only model that significantly improves on the three automatic metrics over both the 1-best baseline and the VSE-features re-ranked model, for all n-best lists sizes ($p = 0.05$).

Regular model Again, when we use VSE features to re-rank n-best lists generated by the regular model, translations do not change much. Nevertheless, VSE-features re-ranked models are the only ones to show some small deterioration in relation to the baseline, even though these differences are not statistically significant.

Models re-ranked with MLMME features are consistently better than the baseline, for all values of β and $n \in \{20, 50\}$. They also show strong improvements on METEOR scores—similarly to when MLMME features are applied to re-rank n-best lists generated with the weak model—in relation to both the baseline and to the translations obtained with the VSE-features re-ranked model, suggesting that they are still making translations more adequate by improving their word-level recall.

When applied to re-rank 50-best lists, MLMME features also significantly improve BLEU scores in relation to the baseline and to the translations obtained with the VSE-features re-ranked model, in spite of the values of β.

Optimised model Improving on the baseline using VSE or MLMME features becomes harder when applied to n-best lists generated by the optimised model. When the baseline model used to generate the n-best lists already provides very strong results, it becomes harder for any additional features, VSE or MLMME, to lead to improved translations. From looking at the results, perhaps the most apparent outcome is the poor results obtained when using VSE features in this scenario. Using the additional VSE features to re-rank consistently and significantly degrades translations, for all n-best list sizes ($n = \{20, 50\}$).

The same does not happen when using MLMME features to re-rank n-best lists. MLMME features lead to translations that consistenly improve over those obtained with the VSE-features re-ranked model, for all different configurations of MLMME models ($\beta = \{0.25, 0.5, 0.75, 1.0\}$). However, the absolute improvement is the smallest among the three models, which is to be expected since a strong baseline should be more difficult to improve on.

Model MLMME with $\beta = 0.25$ or $\beta = 1.0$ seems to achieve the best results

regardless of the n-best list sizes. These are the only two models that also significantly improve on the corresponding 1-best baseline according to at least one of the metrics.

How does model MLMME, trained on images and their descriptions, perform when applied to re-rank n-best lists of a test set of the same domain, but generated by an MT model trained on a different domain? Following the abovementioned experimental setup, we first use a text-only baseline model to generate n-best lists ($n \in \{20, 50\}$) for each entry in the M30k$_{\mathrm{T}}$ validation and test sets. However, in this set of experiments we use an *out-of-domain* (OOD) baseline model to generate the n-best lists. This OOD model is the text-only baseline introduced in Chapter 4 trained on the English–German WMT 2015 corpora, described in Section 3.2. Another important difference between this set of experiments and the ones reported for question *(i) How does* ***n****-best re-ranking perform when* ***n****-best lists are generated by MT models of different quality?* is that we do not use the M30k$_{\mathrm{T}}$ validation set for model selection, but instead use a held out set of 1K sentences from the WMT 2015 corpora as our validation set.

We use the monolingual VSE model of Kiros et al. (2014) trained on German sentences and images to compute the distance between images and target sentences in the n-best lists, and our MLMME models trained with $\beta \in \{.25, .5, .75, 1\}$ to compute the distance between German and English sentences with $s_s(\cdot, \cdot)$, and between a German sentence and an image using $s_i(\cdot, \cdot)$, for all entries in the M30k$_{\mathrm{T}}$ validation and test sets. We similarly train an n-best list re-ranker on the M30k$_{\mathrm{T}}$ validation set's 20-best (50-best) lists with k-best MIRA (Crammer and Singer, 2003; Cherry, 2012), and use the new distances as additional features to the original MT log-likelihood $p(Y \mid X)$. We finally apply the optimised weights to re-rank the test set's 20-best (50-best) lists. In Table 5.5, we show results obtained for the set of experiments where we evaluate how VSE and MLMME models perform when applied to re-rank n-best lists generated by an NMT baseline model trained on out-of-domain

data.

	Discriminative re-ranking?	N	BLEU		METEOR		TER	
Out-of-domain NMT baseline	—	1	21.0		41.6		57.7	
	+ VSE	20	21.2	(+0.2)	41.8	(+0.2)	57.2†	(-0.5)
	+ MLMME, $\beta = 1$	20	21.6†‡	**(+0.6)**	42.4†‡	**(+0.8)**	56.4†‡	**(-1.3)**
	+ MLMME, $\beta = 0.75$	20	21.7†‡	**(+0.7)**	42.4†‡	**(+0.8)**	56.6†‡	**(-1.1)**
	+ MLMME, $\beta = 0.5$	20	21.2	(+0.2)	42.0†	(+0.4)	56.8†	(-0.9)
	+ MLMME, $\beta = 0.25$	20	**22.1**†‡	**(+1.1)**	**42.7**†‡	**(+1.1)**	**56.2**†‡	**(-1.5)**
	oracle	20	29.2		49.1		46.2	
	+ VSE	50	21.0	(+0.0)	41.8	(+0.2)	59.9	(+2.2)
	+ MLMME, $\beta = 1$	50	21.3	(+0.3)	42.5†‡	**(+0.9)**	59.2‡	(+1.5)
	+ MLMME, $\beta = 0.75$	50	**21.9**†‡	**(+0.9)**	42.5†‡	**(+0.9)**	57.8‡	(-0.1)
	+ MLMME, $\beta = 0.5$	50	21.7†	(+0.7)	42.4†‡	**(+0.8)**	**57.6**‡	(-0.1)
	+ MLMME, $\beta = 0.25$	50	21.6‡	(+0.6)	**42.7**†‡	**(+1.1)**	59.2‡	(+1.5)
	oracle	50	32.2		51.4		43.4	

Table 5.5: MT evaluation metrics computed for translations for the M30k$_{\mathrm{T}}$ test set. We show results for 1-best translations generated with an out-of-domain baseline NMT model and for 20- and 50-best lists generated by the same model, re-ranked using VSE and MLMME as discriminative features. Results improve significantly over the corresponding 1-best baseline (†) or over the translations obtained with the VSE re-ranker (‡) with $p = 0.05$.

When re-ranking 20-best lists, VSE features improve results marginally in comparison to the baseline, i.e. statistically significant improvements are only observed in TER, with no significant difference according to BLEU or METEOR. By contrast, MLMME features improve results in almost all configurations. In fact, the only situation in which re-ranking with these features did not significantly improve over both the baseline and the VSE re-ranked translations, is for $\beta = 0.5$. When $\beta \in \{1.0, 0.75, 0.25\}$, translations are significantly better than the baseline and re-ranked translations obtained with the VSE features, according to all metrics evaluated.

When re-ranking 50-best lists, one of the first things to notice is the apparent deterioration in TER scores. Nevertheless, none of these differences in TER scores are statistically significant compared to the baseline. Using VSE features degrades translations' TER the most (+2.5 points), while scores are practically unaltered when using MLMME features with $\beta \in \{0.5, 0.75\}$ ($\pm$0.1 points).

We note that BLEU and METEOR scores do not suffer the negative impact observed for TER scores. According to these two metrics, re-ranking with VSE

features does not impact translations, similarly to the results when re-ranking 20-best lists. From a different perspective, MLMME features consistently improve METEOR scores, which is a result similar to that obtained when re-ranking n-best lists obtained with NMT baselines trained on in-domain data (results for question *(i) How does* $\boldsymbol{n}$*-best re-ranking perform when* $\boldsymbol{n}$*-best lists are generated by MT models of different quality?* and in Table 5.4). This suggests that they are making translations more adequate by improving their word-level recall, which is a very good result since NMT is known to suffer from adequacy issues (Tu et al., 2016).

5.3.4 Analysis

We propose a novel model that incorporates both multilingual and multi-modal similarities and introduce a modified pairwise ranking function to optimise our model (Equation 5.3), which shows gains in three different tasks: ISR, STS and NMT. Results obtained with the Multi30k data set demonstrate that our model can learn meaningful multimodal embeddings, effectively making use of multilingual signals and leading to consistently better results in comparison to a comparable monolingual model.

We demonstrate through results on the ISR task that incorporating multilingual sentences leads to consistent improvements over a comparable monolingual model for ranking images given sentences. Nevertheless, results are not so clear for ranking sentences given images. MLMME models that perform best in this task are the ones trained using only multi-modal similarities, i.e. no multilingual similarity is taken into account into the cost function. Results when ranking English sentences given images are slightly better than those for ranking German sentences. When ranking sentences given images, the recall@k metrics (r@1, r@5 and r@10) decrease as we use less multi-modal and more multilingual similarities (i.e., $\beta \rightarrow 0$), regardless of the language. The same happens with the median rank (mrank), which also tends to deteriorate as β moves from 1 towards 0. When ranking images given English sentences, both the recall@k (r@1, r@5 and r@10) and the median rank (mrank)

metrics first improve as we move from $\beta = 0$ toward $\beta = 0.75$, and then deteriorate as we use less multi-modal and more multilingual similarities (i.e., $\beta \to 0$). When ranking images given German sentences, all the metrics evaluated deteriorate as we use less multi-modal and more multilingual similarities (i.e., $\beta \to 0$).

Using VSE and MLMME models in the STS task also leads to good results for the in-domain tasks. MLMME models outperform a comparable VSE model and also compare favourably in the two in-domain image description similarity tasks: 0.826 (ours) vs. 0.821 (best overall submitted SemEval model) for image description similarity (2014) and 0.886 (ours) vs. 0.864 (best overall submitted SemEval model) for image description similarity (2015). Overall, using models VSE and MLMME to measure textual similarity for out-of-domain tasks does not show good correlation with human judgements.

Moreover, we have also evaluated how well do the Visual Semantic Embedding (VSE) model of Kiros et al. (2014) and our Multilingual Multi-Modal Embedding (MLMME) model perform when used to compute features to re-rank n-best lists generated by models trained on in-domain and out-of-domain data. We found that VSE features perform well on less optimised NMT models trained on in-domain data, but they become less attractive as the baseline NMT models used to generate n-best lists gets better, getting to the point of significantly harming BLEU, METEOR and TER in the case of a highly optimised model. When applied to n-best lists obtained with models trained on out-of-domain data, they are also not very attractive and either do not affect translations, e.g. BLEU or METEOR, or affect translations negatively, e.g. TER.

Overall, MLMME features outperform VSE features in this set of experiments in all scenarios. When applied to the in-domain scenario, MLMME features seem to have a stronger impact on re-ranking n-best lists generated with the regular model compared to the weak and optimised models. Nonetheless, when applied to translations generated with the optimised model, MLMMEs with $\beta = 0.25$ or $\beta = 1.0$ seem to achieve the best results. These models significantly improve on

their corresponding 1-best baseline according to one or more of the metrics evaluated. When applied to the out-of-domain scenario, MLMME features also have an overall positive impact. They consistently and significantly increase METEOR scores, for all n-best lists sizes ($n \in \{20, 50\}$), which is a strong finding since NMT models are known to suffer from adequacy issues (Tu et al., 2016).

Finally, MLMME models take considerably less time to train compared to a fully fledged multi-modal NMT model: training MLMME models take ∼3–6 hours, whereas training a text-only attention-based NMT model should take ∼3–4 days.[4] Likewise, using MLMME models to compute features at inference time is fast: it takes the time to encode the source sentence with the source–language RNN, the target sentence with the target language RNN, the image with a pre-trained CNN, and then performing three dot products, i.e. source·target, target·image, and source·image. Arguably, our results indicate MLMME models to be attractive candidates for inclusion in an NLP pipeline for image descriptions.

In the next chapter, we introduce three different multi-modal NMT models that directly utilise global visual features. Even though these take more time to train, they directly optimise the probability of a translation given both the source sentence and the image, potentially leading to better overall results.

[4]This is the case of training an English–German translation model on the Multi30k data set.

Chapter 6

Incorporating Global Visual Features into NMT

In this chapter, we introduce three multi-modal NMT models that directly incorporate global visual features in different parts of the encoder and the decoder. All these multi-modal NMT models are trained end-to-end and directly optimise the probability of a translation given both the source sentence and the image, therefore differing from using images in an external ranking model for re-ranking at decoding time, proposed in Chapter 5.

The models described in this chapter can be seen as expansions of the text-only attention-based NMT framework described in Chapter 4 with the addition of a *visual component* to incorporate global image features.

Simonyan and Zisserman (2014) trained and evaluated an extensive set of deep convolutional neural network (CNN) models for classifying images into one out of the $1,000$ classes in ImageNet (Russakovsky et al., 2015). We use their 19-layer VGG network (VGG19) to extract image feature vectors for all images in our dataset. We feed an image to the pre-trained VGG19 network and use the 4096D activations of the penultimate fully-connected layer FC7[1] as our *image feature vector*, henceforth referred to as $\boldsymbol{q}$.

[1]We use the activations of the FC7 layer, which encode information about the entire image, of the VGG19 network (configuration E) in Simonyan and Zisserman (2014)'s paper.

We put forward three different methods to incorporate global image features into NMT. The main idea is to integrate image features in different parts of the encoder and the decoder, and evaluate whether they provide different gains depending on where and how they are integrated in the NMT model. We propose to incorporate global image features into the attentive NMT framework:

- using an image as words in the source sentence (Section 6.1.1),
- using an image to initialise the source language encoder (Section 6.1.2), and
- the target language decoder (Section 6.1.3).

We also evaluated a fourth mechanism to incorporate images into NMT, namely to use an image as one of the different contexts available to the decoder at each time step of the decoding process. We add the image features directly as an additional context, in addition to $\boldsymbol{W}_y[\tilde{y}_{t-1}]$, $\boldsymbol{s}_{t-1}$ and $\boldsymbol{c}_t$, to compute the hidden state $\boldsymbol{s}_t$ of the decoder at a given time step t, as illustrated in Figure 6.1. We corroborate previous findings by Vinyals et al. (2015) in that adding the image features as such causes the model to overfit, ultimately preventing learning.[2]

[2]For comparison, a model trained to translate from English into German on the translated Multi30k training set and evaluated on the translated Multi30k test set (described in Section 3.1.1) achieve just 3.8 BLEU, 15.5 METEOR and 93.0 TER.

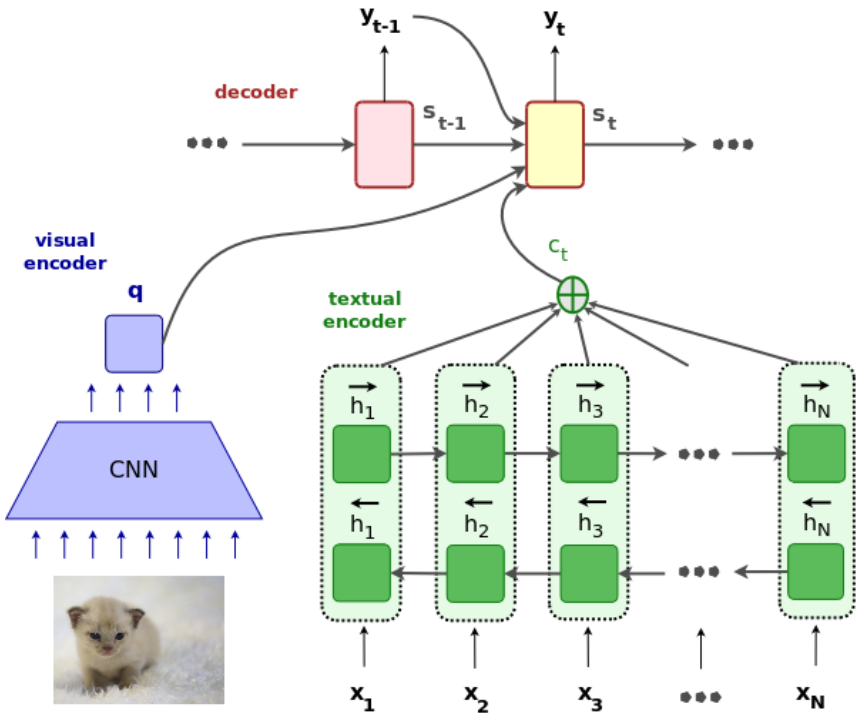


Figure 6.1: Using image features as an additional context at each time step t of the decoder.

6.1 Models

6.1.1 IMG_W: Image as Words in the Source Sentence

One way we propose to incorporate images into the encoder is to project an image feature vector into the space of the words of the source sentence. We use the projected image as the first and/or last word of the source sentence and let the attention model learn when to attend to the image representation. Model $\text{IMG}_{1\text{W}}$ uses the image features as the first word only, and model $\text{IMG}_{2\text{W}}$ uses the image features as the first and last words of the source sentence. By including images into the encoder in models $\text{IMG}_{1\text{W}}$ and $\text{IMG}_{2\text{W}}$, 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 $\overrightarrow{\Phi}_{\text{enc}}$ (vectors $\overrightarrow{\boldsymbol{h}_i}$), 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 $\overleftarrow{\Phi}_{\text{enc}}$ (vectors $\overleftarrow{\boldsymbol{h}_i}$).

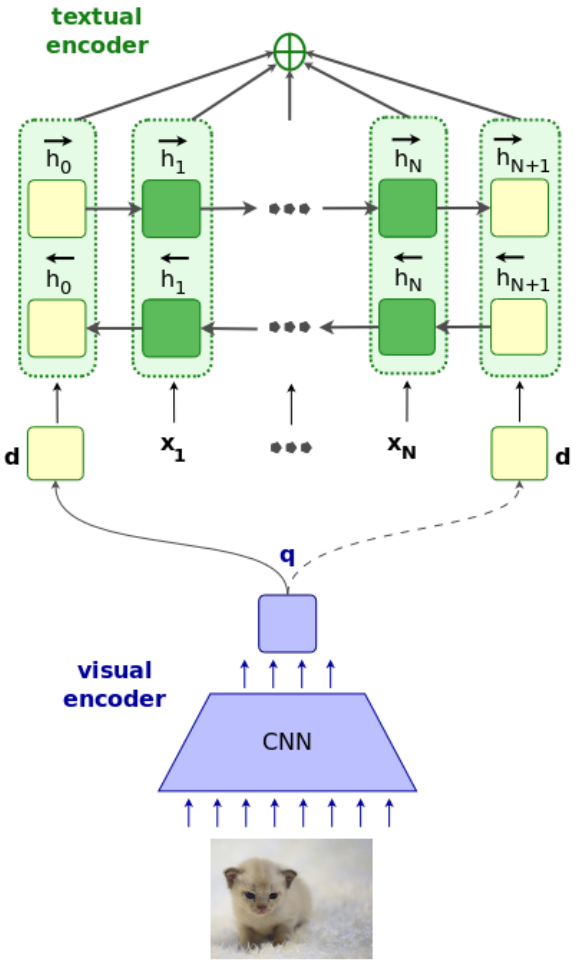


Figure 6.2: An encoder bidirectional RNN that uses image features as words in the source sequence.

Specifically, given the global image feature vector $\boldsymbol{q} \in \mathbb{R}^{4096}$, we compute (6.1):

$$\boldsymbol{d} = \boldsymbol{W}_I^2 \cdot (\boldsymbol{W}_I^1 \cdot \boldsymbol{q} + \boldsymbol{b}_I^1) + \boldsymbol{b}_I^2, \tag{6.1}$$

where $\boldsymbol{W}_I^1 \in \mathbb{R}^{4096 \times 4096}$ and $\boldsymbol{W}_I^2 \in \mathbb{R}^{4096 \times d_x}$ are image transformation matrices, $\boldsymbol{b}_I^1 \in \mathbb{R}^{4096}$ and $\boldsymbol{b}_I^2 \in \mathbb{R}^{d_x}$ are bias vectors, and d_x is the source words vector space dimensionality, all trained with the model. We then directly use $\boldsymbol{d}$ as words in the source words vector space: as the first word only (model $\text{IMG}_{1\text{W}}$), and as the first and last words of the source sentence (model $\text{IMG}_{2\text{W}}$).

An illustration of this idea is given in Figure 6.2, where a source sentence that originally contained N tokens, after including the image as a source word will contain $N+1$ tokens (model $\text{IMG}_{1\text{W}}$) or $N+2$ tokens (model $\text{IMG}_{2\text{W}}$). In model $\text{IMG}_{1\text{W}}$,

the image is projected as the first source word only (solid line in Figure 6.2); in model IMG_{2W}, it is projected into the source words space as both first and last words (both solid and dashed lines in Figure 6.2).

Given a source sequence $X = (x_1, x_2, \cdots, x_N)$, we concatenate the transformed image vector $\boldsymbol{d}$ to $\boldsymbol{W}_x[X]$ and apply the forward and backward encoder RNN passes, generating hidden vectors as in Figure 6.2. When computing the context vector $\boldsymbol{c}_t$ (Equations (4.5) and (4.6)), we effectively make use of the transformed image vector, i.e. the $\alpha_{t,i}$ attention weight parameters will use this information to attend or not to the image features.

6.1.2 IMG_{E}: Image for Encoder Initialisation

In the original text-only attention-based NMT model described in Chapter 4, the hidden state of the encoder is initialised with the zero vector $\overrightarrow{0}$. Instead, we propose to use two new single-layer feed-forward neural networks to compute the initial states of the forward RNN $\overrightarrow{\Phi}_{\text{enc}}$ and the backward RNN $\overleftarrow{\Phi}_{\text{enc}}$, respectively, as illustrated in Figure 6.3.

Similarly to the processing of the image features in the models described in Section 6.1.1, given a global image feature vector $\boldsymbol{q} \in \mathbb{R}^{4096}$, we compute a vector $\boldsymbol{d}$ using Equation (6.1), only this time the parameters $\boldsymbol{W}_I^2$ and $\boldsymbol{b}_I^2$ project the image features into the same dimensionality as the textual encoder's hidden states.

The feed-forward networks used to initialise the encoder hidden state are computed as in Equation (6.2):

$$\begin{aligned} \overleftarrow{\boldsymbol{h}}_{\text{init}} &= \tanh\left(\boldsymbol{W}_f \boldsymbol{d} + \boldsymbol{b}_f\right), \\ \overrightarrow{\boldsymbol{h}}_{\text{init}} &= \tanh\left(\boldsymbol{W}_b \boldsymbol{d} + \boldsymbol{b}_b\right), \end{aligned} \tag{6.2}$$

where $\boldsymbol{W}_f$ and $\boldsymbol{W}_b$ are multi-modal projection matrices that project the image features $\boldsymbol{d}$ into the encoder forward and backward hidden states dimensionality, respectively, and $\boldsymbol{b}_f$ and $\boldsymbol{b}_b$ are bias vectors.

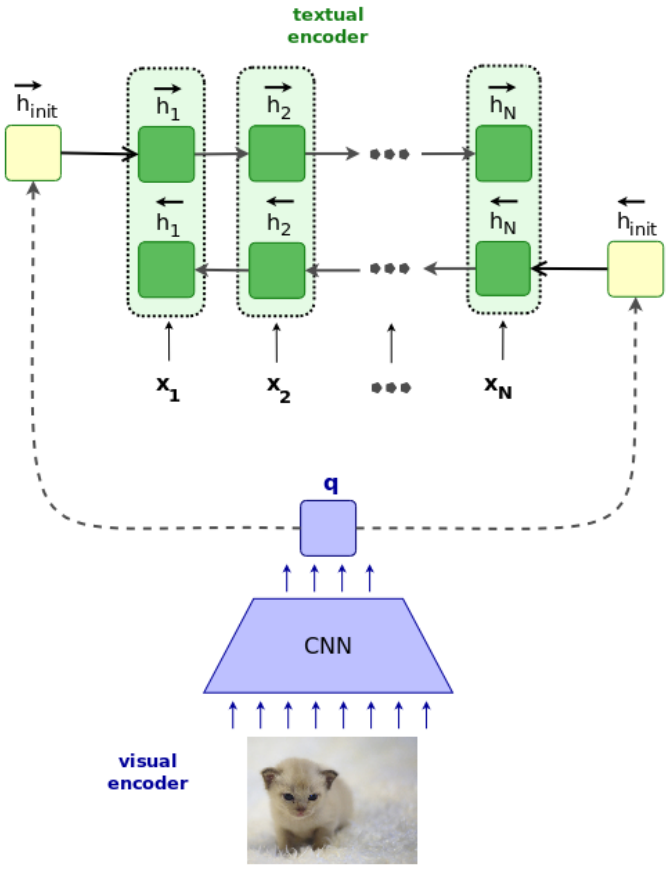


Figure 6.3: Using an image to initialise the encoder hidden states.

6.1.3 IMG$_\mathbf{D}$: Image for Decoder Initialisation

To incorporate an image into the decoder, we introduce a new single-layer feed-forward neural network to be used instead of the one described in Equation (4.3). Originally, the decoder initial hidden state was computed using the concatenation of the last hidden states of the encoder forward RNN ($\overrightarrow{\Phi}_{\text{enc}}$) and backward RNN ($\overleftarrow{\Phi}_{\text{enc}}$), respectively $\overrightarrow{\boldsymbol{h}_N}$ and $\overleftarrow{\boldsymbol{h}}_1$.

Our proposal is that we include the image features as additional input to initialise the decoder hidden state at time step $t = 0$, as described in Equation (6.3):

$$\boldsymbol{s}_0 = \tanh\left(\boldsymbol{W}_{di}[\overleftarrow{\boldsymbol{h}}_1; \overrightarrow{\boldsymbol{h}_N}] + \boldsymbol{W}_m\boldsymbol{d} + \boldsymbol{b}_{di}\right), \tag{6.3}$$

where $\boldsymbol{W}_m$ is a multi-modal projection matrix that projects the image features $\boldsymbol{d}$ into the decoder hidden state dimensionality and $\boldsymbol{W}_{di}$ and $\boldsymbol{b}_{di}$ are the same as in

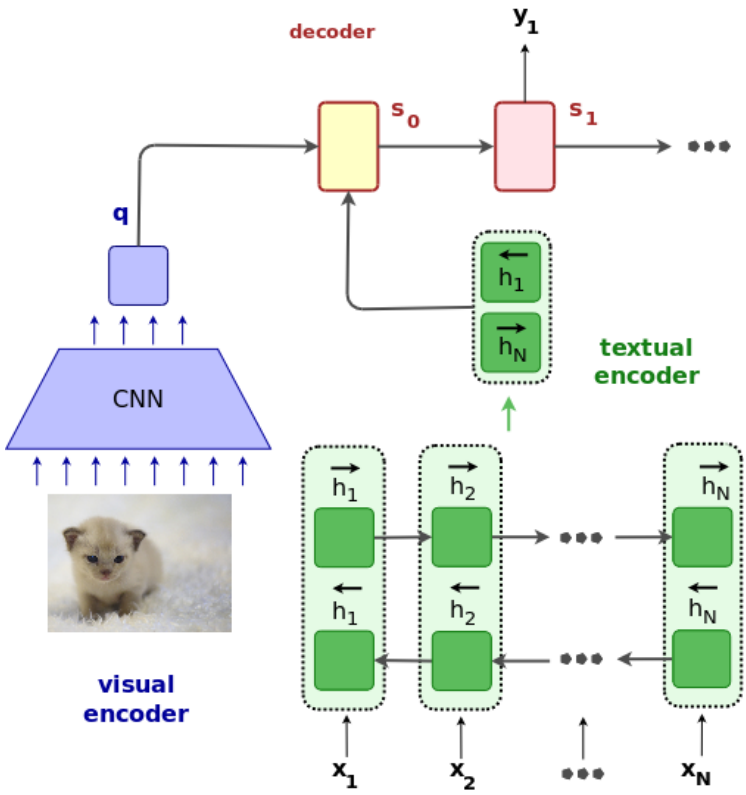


Figure 6.4: Image as additional data to initialise the decoder hidden state s_0.

Equation (4.3).

Once again we compute $\boldsymbol{d}$ by applying Equation (6.1) onto a global image feature vector $\boldsymbol{q} \in \mathbb{R}^{4096}$, only this time the parameters $\boldsymbol{W}_I^2$ and $\boldsymbol{b}_I^2$ project the image features into the same dimensionality as the decoder hidden states. We illustrate this idea in Figure 6.4.

6.2 Experimental setup

Our encoder is a bidirectional RNN with GRU (one 1024D single-layer forward RNN and one 1024D single-layer backward RNN). Source and target word embeddings are 620D each 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 as $\vec{0}$. Our decoder RNN also uses GRU and is a neural LM (Bengio et al., 2003) conditioned on its previous emissions and the source sentence by means of the source

attention mechanism.

Image features are obtained by feeding images to the pre-trained VGG19 network of Simonyan and Zisserman (2014) and using the activations of the penultimate fully-connected layer FC7. We apply dropout with a probability of 0.2 in both source and target word embeddings and with a probability of 0.5 in the image features, in the encoder and decoder RNNs inputs and recurrent connections, and before the readout operation in the decoder RNN. We follow Gal and Ghahramani (2016) and apply dropout to the encoder bidirectional RNN and decoder RNN using the same mask in all time steps.

Our models are trained using stochastic gradient descent with Adadelta (Zeiler, 2012) and minibatches of size 40 for improved generalisation (Keskar et al., 2017), where each training instance consists of one English sentence, one German sentence and one image. We apply early stopping for model selection based on BLEU scores, so that if a model does not improve BLEU scores on the validation set for more than 20 epochs, training is halted.

We evaluate translation quality quantitatively in terms of BLEU4 (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), TER (Snover et al., 2006), and chrF3 scores[3] (Popović, 2015) and we report statistical significance for the three first metrics using approximate randomisation computed with MultEval (Clark et al., 2011).

We use the scripts in the Moses SMT Toolkit (Koehn et al., 2007) to normalise, truecase and tokenize English and German descriptions and we also convert space-separated tokens into subwords (Sennrich et al., 2016b). All models trained on the the Multi30k data sets use a common vocabulary of $83,093$ English and $91,141$ German subword tokens, and those trained on the eBay data sets use a common vocabulary of $32,025$ English and $32,488$ German subword tokens. If sentences in English or German are longer than 80 tokens, they are discarded.

[3]We specifically compute character 6-gram F3 scores.

Experiments on the Multi30k data sets In a first set of experiments, we train models on the Multi30k data set to translate from English into German and from German into English. By doing this, even though the data sets only contain one language pair, we expect to shed some light on the possible differences that might arise when translating from a morphologically-rich into a morphologically-poor language, and vice-versa. Additionally, translations into English are arguably more accessible to the research community. We hope that they can also make communicating our main findings easier and make our results available to a broader audience.

English→German As our main baseline we train the text-only attention-based NMT model described in Chapter 4, in which only the textual part of M30k_T (English–German) is used for training. We also train a PBSMT model built with Moses on the same data. The LM is a 5–gram LM with modified Kneser-Ney smoothing (Kneser and Ney, 1995) trained on the German side of the M30k_T (lowercased or truecased) dataset. We use minimum error rate training (Och, 2003) for tuning the model parameters for BLEU scores. Our third baseline is the best comparable multi-modal model by Huang et al. (2016), and also their best model with additional object detections: respectively models `m1` (image at head) and `m3` in the authors' paper.

German→English Again, as our main baseline we train the text-only attention-based NMT model described in Chapter 4, in which only the textual part of M30k_T (German–English) is used for training. We also train a PBSMT model built with Moses on the same data. The LM is a 5–gram LM with modified Kneser-Ney smoothing (Kneser and Ney, 1995) trained on the English side of the M30k_T dataset. We use minimum error rate training (Och, 2003) for tuning the model parameters for BLEU scores. To the best of our knowledge, there are no published results for multi-modal NMT models trained to translate from German into English.

Experiments on the eBay data sets In a second set of experiments, we train models on the eBay data sets to translate from English into German. We specifically use the eBay24k (Section 3.3.1) and the M30k$_{\mathrm{T}}$ (Section 3.1.1) data sets to train our baselines. In order to measure the impact caused by the size of the training data, we also include the back-translated eBay80k data set in our experiments (Section 3.3.2). With these additional experiments, we expect to be able to assess how different models perform when applied to an arguably more difficult translation scenario, i.e. user-generated, noisy data.

English→German We train a text-only PBSMT and an attention-based NMT model for comparison. The PBSMT model is built using the Moses SMT Toolkit (Koehn et al., 2007), and is trained on the concatenation of the in-domain parallel product listings of the eBay24k and the 29K general-domain parallel English–German descriptions of the M30k$_{\mathrm{T}}$. The language model (LM) is a 5–gram LM with modified Kneser-Ney smoothing (Kneser and Ney, 1995). We use minimum error rate training (Och, 2003) for tuning the model parameters for BLEU scores. The text-only NMT baseline is again the one described in Chapter 4 and is trained on the M30k$_{\mathrm{T}}$'s English–German descriptions.

Additionally, in order to measure the impact that additional back-translated data have on multi-modal models IMG$_{\mathrm{2W}}$, IMG$_{\mathrm{E}}$, and IMG$_{\mathrm{D}}$ in the e-commerce translation scenario, we also train the same baselines but including the back-translated eBay80k data set in our experiments (Section 3.3.2). In order to be able to use the same amount of training data to train the PBSMT baseline, instead of using the additional back-translated data directly we use the ∼80K gold-standard German product descriptions from the eBay80k in order to estimate the LM used in the PBSMT baseline. The reason we did not directly used the back-translated data is because it degraded the PBSMT results in preliminary experiments with this data set. In the case of the NMT models, the additional back-translated data is simply concatenated to the original training data.

6.3 Experiments on the Multi30k data sets

The Multi30K dataset contains images and bilingual descriptions. Overall, it is a small dataset with a small vocabulary whose sentences have simple syntactic structures and not much ambiguity (Elliott et al., 2016). This is reflected in the fact that even the simplest baselines perform fairly well on it: the smallest BLEU score for translating into German is 32.9 (PBSMT), which is still good; for translating into English, the smallest BLEU is 32.8 (PBSMT), which can still be considered reasonably high for a baseline.

6.3.1 English→German

Translations for the translated Multi30k (English→German)								
	BLEU4↑		METEOR↑		TER↓		chrF3↑	
PBSMT	32.9		<u>54.1</u>		<u>45.1</u>		<u>67.4</u>	
NMT	<u>33.7</u>		52.3		46.7		64.5	
Huang	35.1		52.2		—		—	
+ RCNN	36.5		54.1		—		—	
IMG_{1W}	$37.1^{\dagger\ddagger}$	**(↑ 3.4)**	$54.5^{\ddagger}$	**(↑ 0.4)**	$42.7^{\dagger\ddagger}$	**(↓ 2.4)**	66.9	**(↓ 0.5)**
IMG_{2W}	$36.9^{\dagger\ddagger}$	**(↑ 3.2)**	$54.3^{\ddagger}$	**(↑ 0.2)**	$\mathbf{41.9}^{\dagger\ddagger}$	**(↓ 3.2)**	66.8	**(↓ 0.6)**
IMG_{E}	$37.1^{\dagger\ddagger}$	**(↑ 3.4)**	$55.0^{\dagger\ddagger}$	**(↑ 0.9)**	$43.1^{\dagger\ddagger}$	**(↓ 2.0)**	67.6	**(↑ 0.2)**
IMG_{D}	$\mathbf{37.3}^{\dagger\ddagger}$	**(↑ 3.6)**	$\mathbf{55.1}^{\dagger\ddagger}$	**(↑ 1.0)**	$42.8^{\dagger\ddagger}$	**(↓ 2.3)**	**67.7**	**(↑ 0.3)**
IMG_{2W+D}	$35.7^{\dagger\ddagger}$	**(↑ 2.0)**	$53.6^{\ddagger}$	**(↓ 0.5)**	$43.3^{\dagger\ddagger}$	**(↓ 1.8)**	66.2	**(↓ 1.2)**
IMG_{E+D}	$37.0^{\dagger\ddagger}$	**(↑ 3.3)**	$54.7^{\ddagger}$	**(↑ 0.6)**	$42.6^{\dagger\ddagger}$	**(↓ 2.5)**	67.2	**(↓ 0.2)**

Table 6.1: BLEU4, METEOR, chrF3 (higher is better) and TER scores (lower is better) on the $M30k_T$ test set for the two text-only baselines PBSMT and NMT, the two multi-modal NMT models by Huang et al. (2016) and our MNMT models that: *(i)* use images as words in the source sentence (IMG_{1W}, IMG_{2W}), *(ii)* use images to initialise the encoder (IMG_E), and *(iii)* use images as additional data to initialise the decoder (IMG_D). Best text-only baselines are underscored and best overall results appear in bold. We highlight in parentheses the improvements brought by our models compared to the best corresponding text-only baseline score. Results differ significantly from PBSMT baseline (†) or NMT baseline (‡) with $p = 0.05$.

From Table 6.1 we see that our multi-modal models perform well, with models IMG_E and IMG_D improving on both baselines according to all metrics analysed. We also note that all models but IMG_{2W+D} perform consistently better than the strong multi-modal NMT baseline of Huang et al. (2016), even when this model has access

Translations for the translated Multi30k (English→German)								
	BLEU4↑		METEOR↑		TER↓		chrF3↑	
original training data								
IMG_{2W}	36.9		54.3		41.9		66.8	
IMG_{E}	37.1		55.0		43.1		67.6	
IMG_{D}	37.3		55.1		42.8		67.7	
+ back-translated training data								
PBSMT	34.0		55.0		44.7		68.0	
NMT	35.5		53.4		43.3		65.3	
IMG_{2W}	36.7†‡	(↑ **1.2**)	54.6 ‡	(↓ **0.4**)	42.0†‡	(↓ **1.3**)	66.8	(↓ **1.2**)
IMG_{E}	**38.5**†‡	(↑ **3.0**)	55.7†‡	(↑ **0.9**)	**41.4**†‡	(↓ **1.9**)	68.3	(↑ **0.3**)
IMG_{D}	**38.5**†‡	(↑ **3.0**)	**55.9**†‡	(↑ **1.1**)	41.6†‡	(↓ **1.7**)	**68.4**	(↑ **0.4**)
Improvements (original vs. + back-translated)								
IMG_{2W}	↓ 0.2		↑ 0.1		↑ 0.1		↑ 0.0	
IMG_{E}	↑ **1.4**		↑ 0.7		↓ **1.8**		↑ **0.7**	
IMG_{D}	↑ 1.2		↑ **0.8**		↓ 1.2		↑ **0.7**	

Table 6.2: BLEU4, METEOR, TER and chrF3 scores on the $M30k_T$ test set for models trained on original and additional back-translated data. Best text-only baselines are underscored and best overall results in bold. We highlight in parentheses the improvements brought by our models compared to the best baseline score. Results differ significantly from PBSMT baseline (†) or NMT baseline (‡) with $p = 0.05$. We also show the improvements each model yield in each metric when only trained on the original $M30k_T$ training set vs. also including additional back-translated data.

to more data (+RCNN features).[4] Combining image features in the encoder and the decoder at the same time (last two entries in Table 6.1) does not seem to improve results compared to using the image features in only the encoder or the decoder. To the best of our knowledge, it is the first time a purely neural model significantly improves over a PBSMT model in all metrics on this data set.

Arguably, the main downside of applying multi-modal NMT in a real-world scenario is the small amount of publicly available training data (∼30k training instances), which restricts its applicability. For that reason, we back-translated the German sentences in the $M30k_C$ and created additional 145k synthetic triples (synthetic English sentence, original German sentence and image).

In Table 6.2, we present results for some of the models evaluated in Table 6.1 but

[4]In fact, model IMG_{2W+D} still improves on the multi-modal baseline of Huang et al. (2016) when trained on the same data.

when also trained on the additional data. In order to add more data to the PBSMT baseline, we simply added the German sentences in the M30k_{C} as additional data to train the LM, since adding the synthetic sentence pairs to train the baseline PBSMT model, as we did with all neural MT models, degraded the results. Both our models IMG_{E} and IMG_{D} that use global image features to initialise the encoder and the decoder, respectively, improve significantly according to BLEU, METEOR and TER with the additional back-translated data, and also achieved better chrF3 scores. Model IMG_{2W}, that uses images as words in the source sentence, does not significantly differ in BLEU, METEOR or TER ($p = 0.05$), but achieves a lower chrF3 score than the comparable PBSMT model. Although model IMG_{2W} trained on only the original data has the best TER score (= 41.9), both models IMG_{E} and IMG_{D} perform comparably with the additional back-translated data (= 41.4 and 41.6, respectively), though the difference between the latter and the former is still not statistically significant ($p = 0.05$).

We see in Tables 6.1 and 6.2 that our models which use images directly to initialise either the encoder or the decoder are the only ones to consistently outperform the PBSMT baseline according to the chrF3 metric, a character-based metric that includes both precision and recall, and has a recall bias. That is also a noteworthy finding, since chrF3 has been shown to have a high correlation with human judgements (Stanojević et al., 2015).

In Table 6.3 we see translations for two entries in the test M30k set. In the first entry, all but the SMT and the IMG_{2W} models generated a translation that perfectly matched the reference. In the second entry, we have an interesting case where although the reference translation available is not entirely correct—there is one dog with brown and black fur in the image, whereas the German description mentions a brown only dog ("Ein brauner Hund")—, the multi-modal models translated it correctly.

In Table 6.4 we show two more translations for two arguably more complicated examples in the test M30k set. In the first entry, the last three multi-modal models

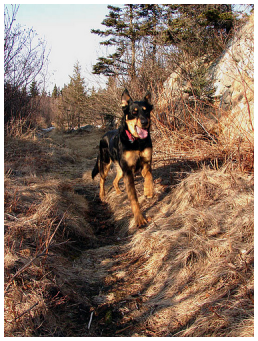
	**ref.**	ein Mann arbeitet an einem Hotdog-Stand .
	PBSMT	ein Mann arbeitet ein Hotdog **stehen** .
	NMT	ein Mann arbeitet an einem Hotdog-Stand .
	IMG_{1W}	ein Mann arbeitet an einem Hotdog-Stand .
	IMG_{2W}	ein Mann arbeitet **einen Hot Dog** .
	IMG_{E}	ein Mann arbeitet an einem Hotdog-Stand .
	IMG_{D}	ein Mann arbeitet an einem Hotdog-Stand .
	**ref.**	Ein brauner Hund läuft über den Sand Strand .
	PBSMT	**ein braun und schwarzer** Hund läuft auf einem Pfad im Wald .
	NMT	ein **brauner Hund steht** an einem **Sand Strand** .
	IMG_{1W}	ein braun-schwarzer Hund läuft auf einem Pfad im Wald .
	IMG_{2W}	ein braun-schwarzer Hund läuft im Wald auf einem Pfad .
	IMG_{E}	ein braun-schwarzer Hund läuft im Wald auf einem Pfad .
	IMG_{D}	ein braun-schwarzer Hund läuft im Wald auf einem Pfad .

Table 6.3: Some translations into German for the M30k test set.

extrapolate the reference+image and describe "ceremony" as a "wedding ceremony" (IMG_{2W}) and as an "Olympics ceremony" (IMG_{E} and IMG_{D}). This could be due to the fact that the training set is small, depicts a small variation of different scenes and contains different forms of biases (van Miltenburg, 2015). In the second entry, we have a longer reference with 16 tokens. Both models IMG_{E} and IMG_{D} only mistranslate the compound "Straße Ecke". Model IMG_{1W} and the NMT baseline mistranslate this compound and also another one, "Kopf Schmuck". IMG_{2W} describes the men as "sitting", which is not true from observing the image. Finally, the SMT model outputs a sentence with some grammar errors as well as uses a verb that does not exist in the German language ("convene").

	**ref.**	eine Frau mit langen Haaren bei einer Abschluss Feier .
	PBSMT	eine Frau mit langen Haaren steht **an einem Abschluss**
	NMT	eine Frau mit langen Haaren **ist an einer StaZeremonie** .
	IMG$_{1W}$	eine Frau mit langen Haaren **ist an einer warmen Zeremonie teil** .
	IMG$_{2W}$	eine Frau mit langen Haaren steht bei einer **Hochzeit Feier** .
	IMG$_{E}$	eine **lang haarige** Frau bei einer **olympischen** Zeremonie .
	IMG$_{D}$	eine **lang haarige** Frau bei einer **olympischen** Zeremonie .
	**ref.**	mehrere ältere Männer, einige mit traditionellen Kopfbedeckungen, treffen sich an einer **Straßen Ecke**.
	PBSMT	mehrere ältere Männer, **einer in** traditioneller **Kopf Schmuck convene** an einer **Straße Ecke**.
	NMT	mehrere ältere Männer, manche mit traditionellen **Kopf Schmuck**, unterhalten sich an einer **Straße Ecke**.
	IMG$_{1W}$	mehrere ältere Männer, einige mit traditionellen **Kopf Schmuck**, stehen an einer **Straße Ecke**.
	IMG$_{2W}$	mehrere ältere Männer mit traditioneller Kopfbedeckung **sitzen** an einer **Straße Ecke**.
	IMG$_{E}$	mehrere ältere Männer, einige davon mit traditioneller Kopfbedeckung, unterhalten sich an einer **Straße Ecke**.
	IMG$_{D}$	mehrere ältere Männer, einige davon mit traditioneller Kopfbedeckung, unterhalten sich an einer **Straße Ecke**.

Table 6.4: More translations into German for the M30k test set.

6.3.2 German→English

Translations for the translated Multi30k (German→English)								
	BLEU4↑		METEOR↑		TER↓		chrF3↑	
PBSMT	32.8		34.8		43.9		61.8	
NMT	<u>38.2</u>		<u>35.8</u>		<u>40.2</u>		<u>62.8</u>	
IMG$_{2W}$	39.5 ‡	**(↑ 1.3)**	37.1†‡	**(↑ 1.3)**	37.1†‡	**(↓ 3.1)**	63.8	**(↑ 1.0)**
IMG$_{E}$	41.1†‡	**(↑ 2.9)**	37.7†‡	**(↑ 1.9)**	37.9†‡	**(↓ 2.3)**	65.7	**(↑ 2.9)**
IMG$_{D}$	41.3†‡	**(↑ 3.1)**	37.8†‡	**(↑ 2.0)**	37.9†‡	**(↓ 2.3)**	65.7	**(↑ 2.9)**
IMG$_{2W+D}$	39.9†‡	**(↑ 1.7)**	37.2†‡	**(↑ 1.4)**	**37.0**†‡	**(↓ 3.2)**	64.4	**(↑ 1.6)**
IMG$_{E+D}$	**41.9**†‡	**(↑ 3.7)**	**37.9**†‡	**(↑ 2.1)**	37.1†‡	**(↓ 3.1)**	**66.0**	**(↑ 3.2)**

Table 6.5: BLEU4, METEOR, chrF3 (higher is better) and TER scores (lower is better) on the M30k$_T$ test set for the two text-only baselines PBSMT and NMT, and our MNMT models that: *(i)* use images as words in the source sentence (IMG$_{1W}$, IMG$_{2W}$), *(ii)* use images to initialise the encoder (IMG$_E$), and *(iii)* use images as additional data to initialise the decoder (IMG$_D$). Best text-only baselines are underscored and best overall results appear in bold. We highlight in parentheses the improvements brought by our models compared to the best corresponding text-only baseline score. Results differ significantly from NMT baseline (†) or PBSMT baseline (‡) with $p = 0.01$.

In Table 6.5, we show results obtained when training our models to translate from German into English. We note that the scores obtained by the PBSMT baseline are very low compared to neural models (especially BLEU4 scores). In order to investigate it further, we also trained an additional PBSMT baseline on the lowercased translated Multi30k data set, for which we show results in Table 6.6. We hypothesised that using truecased data could have led to higher data sparcity, making it harder for PBSMT models to learn properly. However, we cannot draw such conclusions from the results in Table 6.6, since all the differences in all of the four metrics are inconsistent and not statistically significant. Finally, training the PBSMT baseline on either the lowercased or the truecased corpus led to results which are clearly worse than those obtained with all other neural models according to all four automatic metrics evaluated.

BLEU4↑	METEOR↑	TER↓	chrF3↑
PBSMT on truecased translated Multi30k			
32.8	34.8	43.9	61.8
PBSMT on lowercased translated Multi30k			
32.8 (↑ 0.0)	34.9 (↑ 0.1)	44.1 (↓ 0.2)	62.1 (↑ 0.3)

Table 6.6: BLEU4, METEOR, chrF3 (higher is better) and TER scores (lower is better) on the $M30k_T$ test set for the PBSMT baseline when trained and evaluated on truecased vs. lowercased data. We highlight in parentheses the improvements brought by using lowercased instead of truecased data.

Overall, all multi-modal models perform well on the translated Multi30k test set. Model IMG_{E+D} performed best in three out of four metrics (BLEU4, METEOR and chrF3), whereas model IMG_{2W+D} performed best according to TER scores. We again observe that the model that uses images as words in the source sentence (IMG_{2W}) is the worst-performing among the multi-modal models. It scores the worst among these models according to BLEU4, METEOR and chrF3, but again we observe that it has one of the best overall TER scores (= 37.1), which is the same trend observed for translations from English into German, discussed in Section 6.3.1.

Both models that use the image to initialise the encoder or decoder (IMG_E

and IMG_D) fare well when translating from German into English. Their BLEU, METEOR and chrF3 scores are very close to the best-performing model $\text{IMG}_\text{E+D}$ (differences are at most 0.8, 0.2 and 0.3 in each metric, respectively). Finally, model $\text{IMG}_\text{E+D}$ that uses the global image features in order to initialise both the encoder and the decoder is the best one in this test set and language direction.

In Table 6.7 we show some samples of translations for the M30k_T's test set obtained with different models. The examples were selected based on the difference between a model's output and the NMT baseline according to METEOR. We selected one example for each model, meaning that each model has at least one example where it considerably improves over the text-only NMT baseline. In the first example, we see that the translation obtained with the PBSMT system has a grammar mistake (highlighted in bold). All other models, text-only and multi-modal, output correct English and translate the German sentence well. In the second example, the PBSMT did not translate the German word "grob" (rough). The baseline NMT model incorrectly translated "spielen grob miteinander" (play roughly with each other) as "play tug-of-war". All multi-modal models generate fluent English and do not omit any important parts of the source sentence, except for model IMG_E that emits the "grob" (roughly) from the translation. In the third example, the baseline NMT model is clearly the worse among all models. It fails to mention that the people in the field are football players, and wrongly translates "in Aktion" (in action) as "in progress", which damages the translation. All other models, PBSMT and multi-modal NMT, generate fluent English and translate the source reasonably well. The only model that does not explicitly mention that the football players in the field are playing (i.e., "in action") is model IMG_2W. In the fourth example, results are more mixed. Models PBSMT, IMG_E, IMG_D, and $\text{IMG}_\text{E+D}$ all translate the sentence well and perfectly reproduce the reference translation. The NMT baseline's output says the woman is "taking a picture", which is close in meaning to "operating the camera" used in the reference translation. Model $\text{IMG}_\text{2W+D}$'s output says that the woman is "serving a camera", which is incorrect. In the fifth and

last example, both baselines do not translate the source well. The PBSMT model's translation says the dog is jumping "indoors", while the NMT baseline's translation is wrong ("jumps over a jump outside"). All other multi-modal models perfectly replicate the reference translation.

	**src.**	ein junges Paar sitzt auf dem Gehsteig und entspannt gemeinsam .
	ref.	a young couple sits on the sidewalk and relaxes together .
	PBSMT	a young couple sits on the sidewalk **and relaxing together** .
	NMT	a young couple sits on the sidewalk relaxing .
	IMG_{2W}	a young couple sitting on the sidewalk relaxing .
	IMG_{E}	a young couple sitting on the sidewalk relaxing together .
	IMG_{D}	a young couple sitting on the sidewalk relaxing together .
	IMG_{2W+D}	a young couple sits on the sidewalk relaxing together .
	IMG_{E+D}	a young couple sitting on the sidewalk relaxing together .
	**src.**	zwei braune Hunde spielen grob miteinander .
	ref.	two brown dogs playing in a rough manner .
	PBSMT	two brown dogs play **grob** .
	NMT	two brown dogs play **tug-of-war** .
	IMG_{2W}	two brown dogs play in a rough manner .
	IMG_{E}	two brown dogs play with each other .
	IMG_{D}	two brown dogs playing rough with each other
	IMG_{2W+D}	two brown dogs are playing in a rough manner .
	IMG_{E+D}	two brown dogs playing in a rough manner .
	**src.**	mehrere Fußballer auf einem Feld in Aktion .
	ref.	several football players on a field in action .
	PBSMT	several soccer players in action on a field .
	NMT	several people running in a field in progress .
	IMG_{2W}	several soccer players in a field .
	IMG_{E}	several soccer players on a field in action .
	IMG_{D}	several soccer players on a field in action .
	IMG_{2W+D}	several soccer players on a field in action .
	IMG_{E+D}	several footballers in a field in action .
	**src.**	die blau gekleidete Frau bedient eine Kamera vor zwei anderen Frauen .
	ref.	the woman in blue is operating a camera in front of two other women .
	PBSMT	a woman in blue is operating a camera in front of two other women .
	NMT	the woman in blue is taking a picture with two other women .
	IMG_{2W}	the woman in blue is serving a camera to two other women .
	IMG_{E}	the woman in blue is operating a camera in front of two other women .
	IMG_{D}	the woman in blue is operating a camera in front of two other women .
	IMG_{2W+D}	the woman in blue is serving a camera to two other women .
	IMG_{E+D}	the woman in blue is operating a camera in front of two other women .
	**src.**	ein Hund springt im Freien über ein Hindernis .
	ref.	a dog jumps over an obstacle outside .
	PBSMT	a dog jumps over a hurdle indoors .
	NMT	a dog jumps over a jump outside .
	IMG_{2W}	a dog jumps over an obstacle outside .
	IMG_{E}	a dog jumps over an obstacle outside .
	IMG_{D}	a dog jumps over an obstacle outside .
	IMG_{2W+D}	a dog jumps over an obstacle outside .
	IMG_{E+D}	a dog jumps over an obstacle outside .

Table 6.7: Some translations into English for the M30k test set.

Similarly to the experimental setup adopted to evaluate the English–German models, in Table 6.8 we report results for models trained on additional synthetic data where the additional data was again obtained similarly to the English–German scenario. We back-translated the English sentences in the $M30k_C$ and created additional 145k synthetic triples (synthetic German sentence, original English sentence and image). In order to add more data to the PBSMT baseline, we simply added the English sentences in the $M30k_C$ as additional data to train the LM, since adding the synthetic sentence pairs to train the baseline PBSMT model, as we did with all neural MT models, deteriorated the results.

We first note that in spite of the considerable improvements the PBSMT baseline obtains with the additional data (↑ 4.0 BLEU4, ↑ 1.6 METEOR, ↓ 3.1 TER and ↑ 2.7 chrF3), it is still consistently outperformed by the NMT baseline according to all four automatic metrics, and by a large margin. In fact, the PBSMT model improves roughly up to the quality level that the baseline NMT model achieves without any additional back-translated training data. Furthermore, more advanced data selection techniques could be applied to extract further benefits from PBSMT models, but that is outside the scope of this work.

In general, once again the multi-modal NMT models trained using the additional synthetic data fare well when translating into English. All models show nominal improvements over the NMT baseline according to all metrics, the only exception being model IMG_{2W} decreasing BLEU4 scores by 0.2 (but still either mantaining or improving performance according to the other metrics in comparison to the NMT baseline). Model IMG_E is the best performing model in this scenario, being the only model to significantly improve on both the PBSMT and the NMT baselines trained with additional data according to BLEU, METEOR and TER. It also achieves the highest chrF3 scores among all models. Even though model IMG_D still performs as the second best model, it is slightly worse than IMG_E (0.4–0.7 difference according to different metrics).

Once again, results are impressive when we look at the improvements that the

additional data brings to multi-modal models. Improvements range between 1.5–3.8 points according to different metrics, and the average improvement across different metrics is 2.51 points, which is typically considered a considerable improvement regardless of the particular metric.[5] The smallest per-metric average improvement was found for METEOR (= 1.8), and the largest one was found for chrF3 (= 2.93). This is interesting and also intriguing since both metrics have a recall bias, the important difference being that METEOR is word-level and chrF3 is character-level.

Finally, improvements on recall-based metrics are a welcome finding for neural translation models. The attention mechanism in Bahdanau et al. (2015), used in our work, does not explicitly take attention weights from previous time steps into account. This means that when deciding which source words align to a target word, the model does not have access to the previous attention weights explicitly, i.e. it only has *implicit* information from the previous attention weights via the previous hidden state of the decoder. This causes the model to eventually suffer from *under-translation* and *over-translation*: in first case, a model "forgets" to translate parts of the source sentence into the target language; in the second case, a model translates the same words or phrases in the source sentence into the target language multiple times. Both these phenomena have a direct impact in adequacy and by consequence in recall-oriented metrics (Tu et al., 2016).

[5]Although a difference of x points in one metric does not necessarily mean an equal variation in another metric, we can interpret the four metrics ranges to lie between 0%–100%.

Translations for the translated Multi30k (German→English)								
	BLEU4↑		METEOR↑		TER↓		chrF3↑	
original training data								
IMG_{2W}	39.5		37.1		37.1		63.8	
IMG_{E}	41.1		37.7		37.9		65.7	
IMG_{D}	41.3		37.8		37.9		65.7	
+ back-translated training data								
PBSMT	36.8		36.4		40.8		64.5	
NMT	<u>42.6</u>		<u>38.9</u>		<u>36.1</u>		<u>67.6</u>	
IMG_{2W}	42.4 ‡	(↓ 0.2)	39.0 ‡	(↑ 0.1)	**34.7**†‡	**(↓ 1.4)**	67.6	(↑ 0.0)
IMG_{E}	**43.9**†‡	**(↑ 1.3)**	**39.7**†‡	**(↑ 0.8)**	34.8†‡	**(↓ 1.3)**	**68.6**	**(↑ 1.0)**
IMG_{D}	43.4 ‡	(↑ 0.8)	39.3 ‡	(↑ 0.4)	35.2 ‡	(↓ 0.9)	67.8	(↑ 0.2)
Improvements (original vs. + back-translated)								
IMG_{2W}	**↑ 2.9**		**↑ 1.9**		**↓ 2.4**		**↑ 3.8**	
IMG_{E}	**↑ 2.8**		**↑ 2.0**		**↓ 3.1**		**↑ 2.9**	
IMG_{D}	**↑ 2.1**		**↑ 1.5**		**↓ 2.7**		**↑ 2.1**	

Table 6.8: BLEU4, METEOR, TER and chrF3 scores on the $M30k_T$ test set for models trained on original and additional back-translated data. Best text-only baselines are underscored and best overall results in bold. We highlight in parentheses the improvements brought by our models compared to the best baseline score. Results differ significantly from NMT baseline (†) or PBSMT baseline (‡) with $p = 0.05$. We also show the improvements each model yields in each metric when only trained on the original $M30k_T$ training set vs. also including additional back-translated data.

6.3.3 Error Analysis

We believe it is both interesting and important to know what specific types of errors the different models we propose make. Moreover, an error analysis of translations generated by different models can help shed light on the reasons why certain models perform better than others, and if so, in which particular scenarios. It is also important to effectively verify whether there are systematic mistakes that different models make, so we can conjecture the reasons for these mistakes and consider how they might be addressed.

For instance, one intuitive assumption we make regarding the quality of translations obtained with our multi-modal models is that they are better at translating *visual terms*, or terms in a sentence that have a strong visual component to their meaning. These would typically be nouns or verbs, and we define *visual terms* as

either a single word or a phrase that describes something clearly illustrated in the image. Some examples include the colour of an object, a mention to an object, or mentions to animals and people in the image, for instance.

In our investigation, we randomly select 50 sentences from the translated Multi30k test set (described in Section 3.1.1) and compare the translations generated by models trained on two sets of data: the original translated Multi30k data set (M30k_T) and the M30k_T training set plus additional back-translated data. The baselines we evaluate are a PBSMT and a text-only NMT model, and the multi-modal models we evaluate are IMG_2W, IMG_E, IMG_D, $\text{IMG}_\text{2W+D}$ and $\text{IMG}_\text{E+D}$, as explained in Section 6.3.2. In short, we back-translated the English sentences in the M30k_C and created additional 145k synthetic triples (synthetic German sentence, original English sentence and image). In order to add more data to the PBSMT baseline, we simply added the English sentences in the M30k_C as additional data to train the LM, since adding the synthetic sentence pairs to train the baseline PBSMT model, as we did with all neural MT models, deteriorated the results. These are all models trained to translate from German into English, and the reason we perform our error analysis on the translations into English is to make it more useful to a broader number of people in the research community. These models are thoroughly explained in Section 6.3.2, and their results can be found in Tables 6.5 and 6.8.

Error taxonomy We follow previous work and adopt an error taxonomy that is simple to understand and address our needs. Our error taxonomy was adapted from the one introduced in Vilar et al. (2006), with a few differences. These differences are mostly due to the fact that we want to measure how our models translate terms that describe concepts that have a direct correspondence in the image, which we refer to as *visual terms*. Additionally, some of the fine-grained distinctions in the taxonomy proposed in Vilar et al. (2006) are not necessary in our work; in these cases, we just kept the high-level error without differentiating specific sub-errors further.

Finally, in our work the possible categories to select from are:

- *missing words* – there are words missing in the translation, which can be:
  - *content words*, which are central to convey the meaning of the sentence;
  - *filler words*, which are only necessary to make the sentence grammatical;
- *word order* – translations have wrong word order;
- *incorrect words* – words were incorrectly translated:
  - *wrong sense* – includes cases where there is a wrong disambiguation, lexical choice and also spurious translations;
  - *incorrect form* – there are spelling mistakes or mistakes in the inflected word, although the base form is correct;
  - *extra words* – some of the source words are translated more than once, i.e. over-translation;
  - *style* – the translation makes sense, i.e. the main sentence meaning is conveyed, but it does not read fluently.
- *unknown word* – there are parts of the source sentence that were left untranslated.
- *punctuation* – there is a wrong punctuation mark.

In order to measure how well our models translate visual terms, we also mark whenever a model translates a visual term correctly and incorrectly. Additionally, we are also interested in the cases where a model can generate novel information from the image, i.e. the textual description generated by the model does not have an obvious corresponding mention in the source sentence to align to, but can be inferred, at least in principle, from the image. Finally, one last case we investigate is when a model translates visual terms incorrectly, but there is something interesting about the mistake made by the model. Interestingness is clearly a subjective quality, and

we typically select examples where the translation is wrong but there is a reasonable (visual) examplanation for the mistake, e.g. the model translates "Elephant trunk" as "Elephant hose".

Finally, we add one more category, *visual*, to the original categories proposed in Vilar et al. (2006), that has four subcategories:

- *correct* – a visual term is correctly translated;
- *spurious* – a visual term is incorrectly translated;
- *incorrect but interesting* – a visual term is incorrectly translated but there is something interesting about the mistake;
- *novel* – a visual term is generated without a corresponding mention in the source sentence to align to, meaning that the visual term *could have* been inferred from the image.

Moreover, in order to reduce the ambiguity of what a *visual term* is, we propose that a single word or a phrase should be considered a visual term if it describes something clearly illustrated in the image. Since the Multi30k data set consists of images and their descriptions, there will likely be many terms that fall under the *visual term* category.

Error Analysis Tool We use the BLAST tool (Stymne, 2011) in our error analysis. BLAST is a simple open-source tool implemented in Java designed to aid humans in performing Machine Translation error analysis. It allows its user to select with which error taxonomies to work with, and is simple to install and to use.

6.3.3.1 Results

In Table 6.9, we present the error analysis of translations generated by models trained on the M30k_T training set for 50 randomly selected sentences from the M30k_T test set.

	NMT	PBSMT	IMG_{2W}	IMG_{E}	IMG_{D}	IMG_{2W+D}	IMG_{E+D}
missing words							
content	5	5	10	**0**	2	1	2
filler	1	2	**0**	**0**	**0**	**0**	**0**
incorrect words							
wrong sense	32	37	29	**21**	28	**21**	22
incorrect form	7	11	**1**	5	2	4	4
extra words	7	**1**	3	12	5	2	5
style	1	1	1	1	**0**	2	2
visual terms							
correct	24	9	26	**33**	29	32	30
spurious	31	46	26	**22**	25	**21**	**21**
incorrect but interesting	0	0	3	**0**	1	**2**	**4**
novel	0	0	**6**	**6**	3	**6**	5
others							
word order	**0**	3	**0**	1	**0**	1	**0**
unknown word	**0**	31	1	**0**	**0**	1	**0**
punctuation	1	2	**0**	**0**	**0**	**0**	**0**

Table 6.9: Results of the error analysis of translations obtained for 50 randomly selected sentences from the $M30k_T$ test set. Models are all trained on the $M30k_T$ training set. We show the quantity of different errors by each model and error type.

Missing words In general, multi-modal models clearly outperform the two text-only baselines. The behaviour of model IMG_{2W} is unexpected, since it generates translations with the most content words missing, even more than the baselines. This indicates that this model is the one that suffers the most from the undertranslation problem. A possible reason for that could be that by adding the image features as words in the source sentence, it becomes more difficult for a model to properly infer the word alignments since the image features are propagated into the source sentence via the recurrent connections of the bidirectional encoder RNN. However, models IMG_{E} and IMG_{D} do not suffer from this problem and show less undertranslation problems than both baselines. One interesting finding is that when combining models IMG_{2W} and IMG_{D} into model IMG_{2W+D}, the undertranslation problem is partially solved and the number of missing content words is reduced to one out of 50 sentences analysed.

Incorrect words We note that, in most cases, multi-modal models outperform both baselines. The PBSMT baseline is the one that produces more incorrect translations units, including *wrong sense* and *incorrect form* error types. However, it does not suffer from the overtranslation problem, also discussed by Tu et al. (2016). In fact, model IMG_E and the NMT baseline are the ones which produce translations with more repetitive, over-translated content. IMG_2W, IMG_D and the other multi-modal model combinations suffer less from that problem, but still suffer from it to a certain extent (2 to 5 errors in 50 sentences). Introducing some form of attention memory—making the model aware of its previous attention weights for the words generated in the previous time steps—is likely to improve these type of errors, as discussed by Mi et al. (2016) and Tu et al. (2016). Moreover, all models, baselines and multi-modal, do not suffer much from style issues.

The PBSMT baseline clearly has pronounced out-of-vocabulary issues, derived from its lack of ability to extrapolate from a fixed set vocabulary. None of the neural models, baseline or multi-modal, suffer from this issue. This is a very important characteristic of these models, and partially derives from the fact that all the data fed to these models is preprocessed into subwords (Sennrich et al., 2016b). We do not use sentences preprocessed into subwords with our PBSMT models since in preliminary experiments when we tried to do so, results were consistently deteriorated.

Translation of visual terms Regarding the translation of visual terms, the PBSMT baseline is clearly the worst performing one, with 46 spurious translations and only 9 correct ones. The NMT baseline performs considerably better but is still the second worst with 31 spurious translations, although with 24 correct ones; this is nonetheless a strong improvement over the PBSMT baseline. Model IMG_E is the one with the least number of spurious and incorrect but interesting translations of visual terms, with 22 translated units in these categories, and also with the highest number of correct visual terms translations, a total of 33 correct translations. It is closely followed by models $\text{IMG}_\text{2W+D}$, $\text{IMG}_\text{E+D}$ and IMG_D, with 23, 25, and 26

Model	Translation
source	zwei Männer fahren in Kajaks , eines orange und das andere blau , einen Fluss hinab .
reference	two guys are kayaking , one orange one blue , down a creek .
NMT	two men are kayaking in orange , one in orange and the other in blue , going down a river .
PBSMT	two men are kayaks in an orange , blue , and the other in a river .
IMG_{2W}	two men are kayaking down a river in yellow kayak , orange and the other blue .
IMG_{E}	two men are kayaking through kayaks , one in orange and the other blue , are **paddling down** a river .
IMG_{D}	two men are kayaking in kayaks , one in orange and the other blue is going down a river .
IMG_{2W+D}	two men are kayaking , orange and the other blue , **paddling down** a river .
IMG_{E+D}	two men are kayaking in kayaks , one in orange and the other blue , is **paddling down** a river .

Table 6.10: Examples of translations for the example 219 in the M30k test set, where some translations involve novel visual terms (highlighted in bold-faced text).

translated units in these categories, respectively.

In general, all multi-modal models analysed generate a small number of novel visual terms, although these are not very frequent (maximum of 6 examples in the 50 sentences analysed). Their interestingness and uniqueness is due to novel visual terms being likely to have been generated mostly from the image, since these are terms in the target sentence without any obvious words in the source sentence from which they could have been translated.

Examples of novel visual terms We now introduce and discuss some translations of novel visual terms generated by some multi-modal models.

In Table 6.10, we show an example where neither the source German sentence nor the English reference translation contained the translated unit "paddling down", although paddles are clearly visible in the image. Looking into the $M30k_T$ training sentences, there are few examples where a sentence describing people in kayaks or kayaking also include the words "paddle" or "paddling", so the models have not necessarily taken that information strictly from the image, although the image seems to have helped since neither the PBSMT nor the NMT baselines included

Model	Translation
source	eine Gruppe junger Menschen trinkt Shots in einem Mexikanischen Setting .
reference	a group of young people take shots in a Mexican setting .
NMT	a group of young people are **having fun** in an auditorium .
PBSMT	a group of young people drinking at a Shots Mexikanischen Setting .
IMG_{2W}	a group of young people having drinks in a **Mexican restaurant** .
IMG_{E}	a group of young people drinking apples in a **Mexican restaurant** .
IMG_{D}	a group of young people drinking food in a **Mexican restaurant** .
IMG_{2W+D}	a group of young people **having fun** in a Mexican room .
IMG_{E+D}	a group of young people drinking dishes in a **Mexican restaurant** .
	

Table 6.11: Examples of translations for the example 300 in the M30k test set, where some translations involve novel visual terms (highlighted in bold-faced text).

these terms.

In Table 6.11, we draw attention to the example containing two novel visual terms of interest. In the two cases, neither the source German sentence nor the English reference translation contained the translated units "having fun" or "Mexican restaurant", although both could have been inferred at least partially from the image. In this example, one of the novel visual terms, "having fun", is also generated by the baseline NMT model, making it clear that at times what seems like a translation extracted exclusively from the image may have been learnt from the textual part of the training data. However, none of the two text-only NMT or PBSMT baselines translated "Mexikanischen Setting" into "Mexican restaurant", but four out of the five multi-modal models did. The multi-modal models also had problems translating the German "trinkt Shots" (drinking shots). We observe translations such as "having drinks" (IMG_{2W}), which although not a novel translation is still a correct one, but also "drinking apples" (IMG_{E}), "drinking food" (IMG_{D}), and

Model	Translation
source	zwei Nonnen posieren für ein Foto .
reference	two nuns are posing for a picture .
NMT	two men pose for a picture .
PBSMT	two Nonnen posing for a picture .
IMG_{2W}	two girls posing for a picture .
IMG_{E}	two **women** pose for a picture .
IMG_{D}	two **women** pose for a picture .
IMG_{2W+D}	two **women** pose for a picture .
IMG_{E+D}	two **women** pose for a picture .
	

Table 6.12: Examples of translations for the example 720 in the M30k test set, where some translations involve novel visual terms (highlighted in bold-faced text).

"drinking dishes" (IMG_{E+D}), which are clearly incorrect.

In Table 6.12, we bring a simpler example that still demonstrates the strengths multi-modal models bring when translating visual terms. In this example, four out of five multi-modal models translate "Nonnen" (nuns) as "women", whereas the other one translates it as "girl", which is incorrect but still arguably better than the two baselines; the NMT model translates it as "men", and the PBSMT baseline copied the source word "Nonnen" as is, i.e. it is an out-of-vocabulary word. This example showcases that the PBSMT baseline can still leave words untranslated, i.e. out-of-vocabulary, and a strong text-only NMT baseline can still make basic mistakes, even when translating simple sentences like this.

Model	Translation
source	ein Mann verwe ndet elektro nische Geräte .
reference	a man is using electronic equipment .
NMT	a man is working with a pair of equipment .
PBSMT	a man verwe ndet elektro nische equipment .
IMG_{2W}	a man is working on some equipment .
IMG_{E}	a man is playing a **DJ equipment** .
IMG_{D}	a man is working on welding equipment .
IMG_{2W+D}	a man is working on some equipment .
IMG_{E+D}	a man is playing a piece of equipment .

Table 6.13: Examples of translations for the example 339 in the M30k test set, where some translations involve novel visual terms (highlighted in bold-faced text).

In Table 6.13, we discuss an interesting example. Here, the German source sentence is incorrect; it looks like it was probably incorrectly tokenized. One of the ways to fix sentence "ein Mann verwe ndet elektro nische Geräte ." is writing it as "ein Mann **verwendet elektronische** Geräte ." This is the German sentence for which the reference translation "a man is using electronic equipment ." is correct.

We note that the PBSMT model is unable to cope with these errors in the source sentence. Its translation basically left "verwe ndet elektro nische" untranslated, which has a clear negative impact in the quality of the output. Nonetheless, all NMT models (including the baseline) have managed to translate "verwe ndet elektro nische" more or less accurately. The translation generated by the baseline NMT

model mentions "a pair of equipment", which is again wrong but conveys some of the meaning in the source. Most of the translations generated by the multi-modal models are better, with one translation in special. Model IMG_E translates "verwe ndet elektro nische Geräte" (is using electronic equipment) as "is playing a DJ equipment", which is surprisingly correct and accurate, although this information is clearly not in the source nor in the reference translation. We again looked into the M30k_T training sentences, and there are a few examples where sentences describe "DJ" or "DJ equipment". That means that the models have not necessarily taken that information strictly from the image since they have seen these in these training sentences, although the image seems to have helped since neither the PBSMT nor the NMT baselines included these terms.

Results including back-translated training sentences In Table 6.14, we show results for models trained on additional multi-modal back-translated data, as described in Section 6.3.3. The error analysis follows the same protocol as the one where the models are trained on the translated Multi30k, described in Table 6.9.

In general, we observe the same trends in models trained using additional back-translated data (Table 6.14) and models trained on only the M30k_T original training data (Table 6.9).

One important change with the additional back-translated data is that the PBSMT decreases its error involving general terms (visual or non-visual) to levels close to the neural models. However, it still suffers considerably from out-of-vocabulary words (28 unknown words in 50 sentences analysed) and also from word order issues, and in both cases these issues are practically inexistent in neural models.

There is a clear trend towards multi-modal models translating visual terms better than both baselines, which is to be expected. The best overall model, meaning the model that makes the least number of errors, is model IMG_D, which has 19 visual terms translation errors. This is followed by models IMG_E, IMG_2W and the NMT baseline, with 21, 22 and 28 visual terms translation errors, respectively.

	NMT	PBSMT	IMG_{2W}	IMG_E	IMG_D
			missing words		
content	**2** (↓**3**)	6 (↑**1**)	3 (↓**7**)	**2** (↑**2**)	**2** (↓**0**)
filler	– (↓**1**)	1 (↓**1**)	– (↓**0**)	– (↓**0**)	1 (↑**1**)
			incorrect words		
wrong sense	30 (↓**2**)	22 (↓**15**)	**19** (↓**10**)	22 (↑**1**)	21 (↓**1**)
incorrect form	**1** (↓**6**)	5 (↓**6**)	5 (↑**4**)	4 (↓**1**)	3 (↓**1**)
extra words	2 (↓**5**)	4 (↑**3**)	2 (↓**10**)	2 (↓**3**)	**1** (↓**1**)
style	5 (↑**4**)	4 (↑**3**)	– (↓**0**)	**2** (↓**0**)	3 (↑**1**)
			visual terms		
correct	27 (↑**3**)	16 (↑**7**)	33 (↑**0**)	34 (↑**2**)	**36** (↑**6**)
spurious	27 (↓**4**)	39 (↓**7**)	20 (↓**6**)	19 (↓**3**)	**16** (↓**9**)
incorrect but interesting	1 (↑**1**)	– (↓**0**)	2 (↓**1**)	2 (↑**2**)	3 (↑**2**)
novel	1	–	1	2	3
			others		
word order	–	**7**	–	1	–
unknown word	–	**28**	–	–	–
punctuation	–	–	–	–	–

Table 6.14: Results of the error analysis of translations obtained for 50 randomly selected sentences from the $M30k_T$ test set. Models are all trained on the $M30k_T$ plus the back-translated $M30k_C$ training set. We show the quantity of different errors by each model and error type, and also, in parentheses, the difference between the current number of errors vs. the number of errors for the same model trained on only the $M30k_T$ training data.

Final remarks In general, the additional back-translated training data reduced errors in all models, baselines and multi-modal. We note that the most damaging type of error we evaluate to the perceived quality of a translation is the *wrong sense* in the *incorrect words* category and the *spurious* translations in the *visual* category, and that these errors were, on average, drastically reduced by the addition of back-translated data. When analysing the impact of the additional back-translated data in the translation of visual terms, both the number of spurious and incorrect errors taken together consistently decreased, as well as the number of correct translations consistently increased. Overall, the number of errors in the *incorrect words* categories is reduced from 205 by 45, a total of 22%. Additionally, the number of errors in the *visual term* categories is reduced from 154 by 25, a total of 16%.

These are all strong findings that support our initial intuition that multi-modal models are not only quantitatively but also qualitatively better than text-only ones when translating image descriptions. We demonstrate that multi-modal models reduce not only errors related to the translation of visual terms, but also considerably reduce more general errors, e.g. *incorrect words* category. This is in itself an interesting finding, since it implies that adding multi-modal, visual signals is helpful not only in the obvious situations where we wish to translate visual terms. By contrast, our error analysis indicates that improvements are distributed across visual and non-visual portions of the text, which is a surprising collateral impact. Finally, adding back-translated multi-modal data helps multi-modal models improve and has a general positive impact on the final translation quality, again improving both in the translation of visual and non-visual terms.

6.4 Experiments on the eBay data set

We now report experiments where we apply our models onto the eBay data sets. In Table 6.15 we show results for two text-only baselines, PBSMT and NMT, and three multi-modal models that use image global features, IMG_{2W}, IMG_{E}, and IMG_{D}. We compute four automatic MT metrics, BLEU, METEOR, TER and chrF3, as well as character-level precision and recall. Although not MT metrics *per se*[6], we include precision and recall because these two dimensions help us better understand the translations obtained with different models.

We first note that the PBSMT model shows ambiguous trends with the additional back-translated data; although BLEU and METEOR scores are slightly improved, TER and chrF3 scores are slightly deteriorated. We note that the character-level precision is slightly increased, while the character-level recall is slighly decreased. In general, the differences brought by the additional back-translated data to PBSMT

[6]Character-level precision and recall were not devised to correlate well with human judgements of translation quality. Nonetheless, they are components used to compute the chrF3 metric (Popović, 2015).

Model	Training data	BLEU	METEOR	TER	chrF3	character precision	character recall
PBSMT	eBay24k + M30k$_T$	25.9	44.9	**56.0**	**61.6**	63.9	**61.3**
	+ back-translated eBay80k	**26.5** ↑ 0.6	**45.3** ↑ 0.4	56.2 ↑ 0.2	61.1 ↓ 0.5	**64.3** ↑ 0.4	60.8 ↓ 0.5
Text-only NMT	eBay24k + M30k$_T$	19.4	37.7	61.2	54.9	59.5	54.5
	+ back-translated eBay80k	23.7 ↑ 4.3	**42.3** ↑ 4.6	56.5 ↓ 4.7	59.3 ↑ 4.4	64.3 ↑ 4.8	58.8 ↑ 4.3
IMG$_{2W}$	eBay24k + M30k$_T$	19.6	37.7	60.1	55.3	60.9	54.7
	+ back-translated eBay80k	**24.2** ↑ 4.6	41.8 ↑ 4.1	**55.7** ↓ 4.4	58.9 ↑ 3.6	**64.8** ↑ 3.9	58.3 ↑ 3.6
IMG$_E$	eBay24k + M30k$_T$	19.0	37.0	61.7	54.8	59.6	54.3
	+ back-translated eBay80k	23.7 ↑ 4.7	42.1 ↑ 5.1	57.0 ↓ 4.7	58.9 ↑ 4.1	63.9 ↑ 4.3	58.4 ↑ 4.1
IMG$_D$	eBay24k + M30k$_T$	19.6	37.6	60.7	55.0	60.1	54.5
	+ back-translated eBay80k	**24.2** ↑ 4.6	**42.3** ↑ 4.7	56.5 ↓ 4.2	**59.4** ↑ 4.4	64.2 ↑ 4.1	**58.9** ↑ 4.4
Improvements							
IMG$_{2W}$ vs. Text-only NMT		↑ 0.5	↓ 0.5	↓ 0.8	↓ 0.4	↑ 0.5	↓ 0.5
IMG$_{2W}$ vs. PBSMT		↓ 2.3	↓ 3.5	↓ 0.3	↓ 2.7	↑ 0.5	↓ 3.0
IMG$_E$ vs. Text-only NMT		↑ 0.0	↓ 0.2	↑ 0.5	↓ 0.4	↓ 0.4	↓ 0.4
IMG$_E$ vs. PBSMT		↓ 2.8	↓ 3.2	↑ 1.0	↓ 2.7	↓ 0.4	↓ 2.9
IMG$_D$ vs. Text-only NMT		↑ 0.5	↑ 0.0	↑ 0.0	↑ 0.1	↓ 0.1	↑ 0.1
IMG$_D$ vs. PBSMT		↓ 2.3	↓ 3.0	↑ 0.5	↓ 2.2	↓ 0.1	↓ 2.4

Table 6.15: Comparative results with PBSMT, text-only NMT and multi-modal models IMG$_{2W}$, IMG$_E$ and IMG$_D$. Best overall PBSMT and neural MT results in bold. We show improvements brought by the additional back-translated data and also the relative differences between different models (best viewed in colour).

are not significant ($p = 0.05$).

Overall, the impact that the additional back-translated data brings into neural models is consistently positive, across all neural models and evaluation metrics. The gains the neural models show range between 3.6 chrF3 (model IMG$_{2W}$) and 5.1 METEOR (model IMG$_E$), with an average improvement of 4.45 points over the four MT evaluation metrics.

Even though the impact brought by the additional back-translated data is uneven between PBSMT and neural MT models, PBSMT models are still the best ones according to these experiments. Looking carefully at character-level precision and recall, we note that PBSMT translations are better precisely according to recall-oriented metrics, since character-level precision scores achieved by neural models with access to additional back-translated data are comparable to the ones obtained by the PBSMT models (the best PBSMT model has 64.3 character precision, while the best and worst multi-modal NMT models with additional back-translated data have 64.8 and 63.9, respectively).

We expected that the multi-modal models IMG$_{2W}$, IMG$_E$, and IMG$_D$ would

improve more over the NMT baseline, as was the case with the experiments with the $M30k_T$ in Section 6.3. One point to note is that we did not find statistically significant differences, in any of the metrics evaluated, between multi-modal models and the text-only NMT baseline.

6.5 Final Remarks

In this chapter, we have introduced different ideas to incorporate global image features into state-of-the-art attention-based NMT, by using images as words in the source sentence, to initialise the encoder hidden state and as additional data in the initialisation of the decoder hidden state. We corroborate previous findings in that using image features directly at each time step of the decoder causes the model to overfit and prevents learning. The intuition behind our effort is to use global image feature vectors to visually *ground* translations and consequently increase translation quality. Extensive experiments show that adding global image features into attention-based NMT is useful and improves in the translation of image descriptions over both NMT and PBSMT baselines, as well as a strong multi-modal NMT baseline in the English–German translation scenario, according to all metrics evaluated.

When applied to the Multi30k data set, we note that the idea of using images as words in the source sentence, also entertained by Huang et al. (2016), does not perform as well as directly using the images in the encoder or decoder initialisation regardless of the target language. Morever, the fact that multi-modal NMT models can benefit from back-translated data regardless of the translation direction is an interesting finding.

Model IMG_{2W} consistently achieves the lowest TER scores, but from looking at the other metrics computed as well as from a manual evaluation of the translations, it is not the best multi-modal NMT model evaluated. As a general conclusion, models IMG_E, IMG_D, and IMG_{E+D} are the best performing ones in both directions, with

small variations between them according to the target language. Whereas models IMG_E and IMG_D perform best when translating into German, model $\text{IMG}_\text{E+D}$ is the best performing one when translating into English.

PBSMT can still be considered as a viable approach to translate into German when a small training set is available. Nevertheless, when more training data is available, unless one has the time/budget to apply more advanced data selection techniques, it does not scale well according to our experiments on the Multi30k data. NMT models, both text-only and multi-modal, can directly use the concatenation of the original data plus synthetic examples obtained from a different source, and still significantly improve their translations. This finding is also independent of the language direction, i.e. it holds for both English–German as well as German–English translation.

Moreover, we conducted an extensive error analysis of translations for a random set of 50 sentences of the M30k_T test set, where we specifically investigated errors in the translation of visual and non-visual terms. We found that the additional back-translated data consistently improves translations, and that the multi-modal NMT models IMG_2W, IMG_E, and IMG_D are qualitatively better than both text-only baselines when translating not only visual terms but also non-visual terms. These are strong findings that corroborate the good quantitative results obtained in other experiments carried out with the Multi30k data set.

The results we obtained when applying translation models that include global image features to translate user-generated product listings were not as good as those obtained with the M30k_T. Adding back-translated data to all neural models, text-only and multi-modal, led to consistently better translations according to all metrics evaluated, which did not happend when adding the additional data to a PBSMT model. Nonetheless, we found no statistically significant difference between a baseline NMT model and any of the multi-modal models IMG_2W, IMG_E, or IMG_D, all of them trained using additional back-translated data.

We did not expect these results, and have a few hypotheses to explain why they

happened. First, the M30k$_{\mathrm{T}}$ and eBay data sets are very different. The eBay data is user-generated, noisy and arguably much more difficult to work with. That is not to mention the product images in the eBay data, which also present additional difficulties to our multi-modal models to cope with, as per our discussion in Section 3.3.3. The pre-trained CNN models we use for visual feature extraction were trained to classify images from ImageNet, and using them to extract features for product images can be an additional difficulty we did not anticipate.

In order to address these concerns, we randomly selected two dozen entries from the eBay test set, and fed the corresponding entries' product image into a pre-trained ResNet-50, described in Section 2.2.1.2. We then manually evaluated the outputs of the classification, which are one out of $1,000$ possible classes in ImageNet. The results of the image classification were not perfect but are nonetheless surprisingly good, an indication that the image features we extract for the product images are in fact reasonably good, therefore unlikely to be the main culprit in our experimental setup.

One main difference between the experiments with back-translated data for the M30k$_{\mathrm{T}}$ and the eBay data sets is the ratio of sentences available for each image. In the original M30k$_{\mathrm{T}}$, there are one parallel sentence pair for each image and therefore this ratio is 1-to-1. With the additional back-translated sentences, this ratio is increased from 1-to-1 to 6-to-1, since all the back-translated sentences are obtained from the M30k$_{\mathrm{C}}$, which share the same images. In the eBay data set, the sentence-to-image ratio before and after adding more back-translated data is always 1-to-1. This is because the additional back-translated product listings were obtained for different entries, and not different product listings sharing one same product image. We conjecture that that could have played a role in the results obtained with the multi-modal NMT models. Finally, the eBay data sets have a very high percentage of low-frequency words, and especially if compared to the Multi30k data, e.g. 36.9% of the eBay24k German tokens have frequency lower than or equal to 5, whereas only 5% of the Multi30k German tokens fall under the same category. All these

taken together make the eBay data set much more difficult to work with.

In the next chapter, we propose a multi-modal NMT model that integrates local image features. We expect that by integrating local features in a separate visual attention mechanism, the model can learn when to focus on them and when not to. In principle, such a model can therefore be more resilient to noisy data and more likely to deliver better translations under these scenarios, e.g. when translating the eBay user-generated data sets.

Chapter 7

Incorporating Local Visual Features into NMT

In this chapter, we introduce a multi-modal NMT model that incorporates local visual features in a separate visual attention mechanism. Local features are much larger than global ones, and therefore can possibly encode more fine-grained information about an image and its objects. We demonstrate that this multi-modal NMT model makes efficient use of training data, since we can effectively pre-train it using both multi-modal data as well as monolingual parallel MT data.

Similarly to Chapter 6, once again our models can be seen as expansions of the attention-based NMT framework described in Chapter 4 with the addition of a *visual component* to incorporate image features. However, differently from the models introduced in Chapter 6, we are now interested in exploiting *local image features*, i.e. image features that encode spatial information.

We use publicly available pre-trained CNNs for image feature extraction. Specifically, we extract local image features for all images in our dataset using the 50-layer Residual network (ResNet-50) of He et al. (2015). These local features are the activations of the `res4f` layer, which can be seen as encoding an image in a 14×14 grid where each of the entries in the grid is represented by a 1024D feature vector that only encodes information about that specific region of the image. We vectorise

this 3-tensor into a 196×1024 matrix $A = (\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_L), \boldsymbol{a}_l \in \mathbb{R}^{1024}$ where each of the L = 196 rows consists of a 1024D feature vector and each column, i.e. feature vector, represents one grid in the image.

In this chapter, we introduce one model that incorporates local image features into the attention-based NMT framework by integrating these features in an additional visual attention mechanism to be incorporated in a multi-modal decoder RNN. We call this model $\text{NMT}_{\text{SRC+IMG}}$ and introduce and discuss it in Section 7.1.

7.1 $\text{NMT}_{\text{SRC+IMG}}$ — Doubly-Attentive Decoder

Model $\text{NMT}_{\text{SRC+IMG}}$ integrates two separate attention mechanisms, over the source-language words and visual features, in a single decoder RNN. The novel *doubly-attentive* decoder RNN is conditioned on the previous hidden state of the decoder and the previously emitted word, as well as the source sentence and the image via two independent attention mechanisms, as illustrated in Figure 7.1. In other words, in model $\text{NMT}_{\text{SRC+IMG}}$ a *doubly-attentive decoder* naturally incorporates *local* visual features extracted from images. Our decoder learns to attend to source-language words and parts of an image independently by means of two *separate attention mechanisms* as it generates words in the target language.

We implement this idea expanding the conditional GRU described in Section 4.2 onto a *doubly-conditional* GRU. To this end, in addition to the source-language attention, we introduce a new attention mechanism ATT_{img} to the original conditional GRU proposal. This visual attention computes a *time-dependent* image context vector $\boldsymbol{i}_t$ given a hidden state proposal $\boldsymbol{s}'_t$ and the image annotation vectors $A = (\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_L)$ using the "soft" attention mechanism (Xu et al., 2015).

This attention mechanism is very similar to the source-language attention with the addition of a *gating scalar*, explained further below. First, a single-layer feed-forward network is used to compute an *expected alignment* $e^{\text{img}}_{t,l}$ between each image annotation vector $\boldsymbol{a}_l$ and the target word to be emitted at the current time step t,

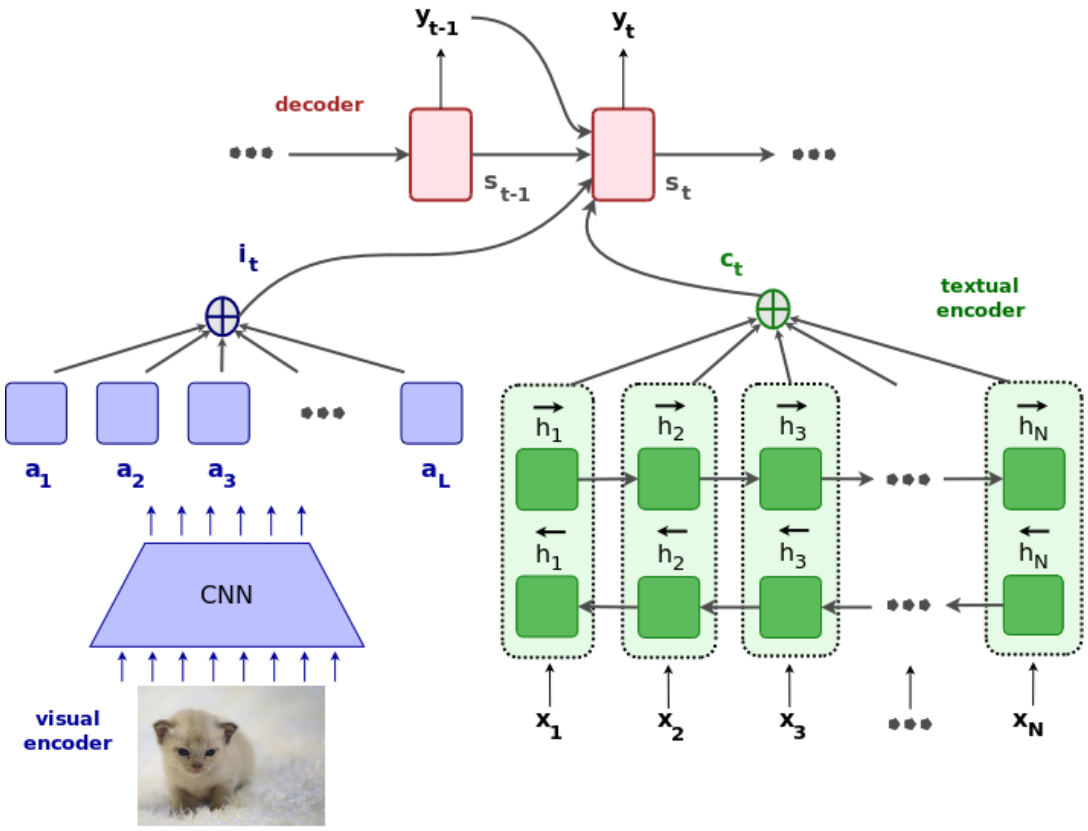


Figure 7.1: A doubly-attentive decoder learns to attend to image patches and source-language words independently when generating translations.

as in Equations (7.1) and (7.2).

$$e_{t,l}^{\text{img}} = (\boldsymbol{v}_a^{\text{img}})^T \tanh(\boldsymbol{U}_a^{\text{img}} \boldsymbol{s}'_t + \boldsymbol{W}_a^{\text{img}} \boldsymbol{a}_l), \tag{7.1}$$

$$\alpha_{t,l}^{\text{img}} = \frac{\exp\left(e_{t,l}^{\text{img}}\right)}{\sum_{j=1}^{L} \exp\left(e_{t,j}^{\text{img}}\right)}, \tag{7.2}$$

where $\alpha_{t,l}^{\text{img}}$ is the normalised alignment matrix between all the image patches $\boldsymbol{a}_l$ and the target word to be emitted at time step t, and $\boldsymbol{v}_a^{\text{img}}$, $\boldsymbol{U}_a^{\text{img}}$ and $\boldsymbol{W}_a^{\text{img}}$ are model parameters. Note that Equations (4.5) and (4.6), that compute the expected source alignment $e_{t,i}^{\text{src}}$ and the weight matrices $\alpha_{t,i}^{\text{src}}$, and Equations (7.1) and (7.2), that compute the expected image alignment $e_{t,l}^{\text{img}}$ and the weight matrices $\alpha_{t,l}^{\text{img}}$, both compute similar statistics over the source and image annotations, respectively.

In Equation (7.3) we compute $\beta_t \in [0, 1]$, a gating scalar used to weight the importance of the image context vector in relation to the target word to be emitted

at time step t.

$$\beta_t = \sigma(\boldsymbol{W}_\beta \boldsymbol{h}_{t-1} + \boldsymbol{b}_\beta), \tag{7.3}$$

where $\boldsymbol{W}_\beta$, $\boldsymbol{b}_\beta$ are model parameters. This is in turn used to compute the time-dependent image context vector $\boldsymbol{i}_t$ for the current decoder time step t, as in Equation (7.4).

$$\boldsymbol{i}_t = \beta_t \sum_{l=1}^{L} \alpha_{t,l}^{i} \boldsymbol{a}_l. \tag{7.4}$$

The only difference between Equation (4.7) (source context vector) and Equation (7.4) (image context vector) is that the latter uses a gating scalar, whereas the former does not. We use β following Xu et al. (2015) who empirically found it to improve the variability of the image descriptions generated with their model.

Finally, we use the time-dependent image context vector $\boldsymbol{i}_t$ as an additional input to a modified version of REC_2 (Section 4.2), which now computes the final hidden state $\boldsymbol{s}_t$ using the hidden state proposal $\boldsymbol{s}'_t$, and the source and image time-dependent context vectors $\boldsymbol{c}_t$ and $\boldsymbol{i}_t$, as in Equation (7.5):

$$\begin{aligned}
\boldsymbol{r}_t &= \sigma(\boldsymbol{W}_r^{\text{src}} \boldsymbol{c}_t + \boldsymbol{W}_r^{\text{img}} \boldsymbol{i}_t + \boldsymbol{U}_r \boldsymbol{s}'_j), \\
\boldsymbol{z}_t &= \sigma(\boldsymbol{W}_z^{\text{src}} \boldsymbol{c}_t + \boldsymbol{W}_z^{\text{img}} \boldsymbol{i}_t + \boldsymbol{U}_z \boldsymbol{s}'_j), \\
\underline{\boldsymbol{s}}_t &= \tanh(\boldsymbol{W}^{\text{src}} \boldsymbol{c}_t + \boldsymbol{W}^{\text{img}} \boldsymbol{i}_t + \boldsymbol{r}_t \odot (\boldsymbol{U} \boldsymbol{s}'_t)), \\
\boldsymbol{s}_t &= (1 - \boldsymbol{z}_t) \odot \underline{\boldsymbol{s}}_t + \boldsymbol{z}_t \odot \boldsymbol{s}'_t,
\end{aligned} \tag{7.5}$$

where again all matrices $\boldsymbol{W}$ and $\boldsymbol{U}$ are trained jointly with the model. Finally, in Equation (7.6) the probabilities for the next target word are computed using the new multi-modal hidden state $\boldsymbol{s}_t$, the previously emitted word $\hat{y}_{t-1}$, and the two context vectors $\boldsymbol{c}_t$ and $\boldsymbol{i}_t$:

$$p(y_t = k \mid \boldsymbol{y}_{<t}, X, A) \propto \exp(\boldsymbol{L_o} \tanh(\boldsymbol{L_s} \boldsymbol{s}_t + \boldsymbol{L_w} \boldsymbol{W_y} [\hat{y}_{t-1}] + \boldsymbol{L_{cs}} \boldsymbol{c}_t + \boldsymbol{L_{ci}} \boldsymbol{i}_t)), \tag{7.6}$$

where $\boldsymbol{L}_o$, $\boldsymbol{L}_s$, $\boldsymbol{L}_w$, $\boldsymbol{L}_{cs}$ and $\boldsymbol{L}_{ci}$ are projection matrices trained with the model.

7.2 Experiments on the Multi30k data sets

In this set of experiments, we use model $\text{NMT}_{\text{SRC+IMG}}$ to translate the Multi30k data sets, which consist of relatively clean data (Elliott et al., 2016) (more details on Section 3.1). With these experiments, we wish to observe how does model $\text{NMT}_{\text{SRC+IMG}}$ fare when translating this type of data.

We use the entire M30k_{T}'s training set for training our models, its validation set for model selection with BLEU, and its test set for evaluation. Also, since the amount of training data available is small, we build a back-translation model using the text-only NMT model described in Chapter 4 trained on the $\text{Multi30k}_{\text{T}}$ data set (German→English and English→German), without images. We use this model to back-translate the 145k German (English) descriptions in the $\text{Multi30k}_{\text{C}}$ into English (German) and include the triples (synthetic English description, German description, image) when translating into German, and the triples (synthetic German description, English description, image) when translating into English, as additional training data (Sennrich et al., 2016a). We also use the WMT 2015 text-only parallel corpora available for the English–German language pair, consisting of about 4.3M sentence pairs (Bojar et al., 2015).

We use the scripts in the Moses SMT Toolkit (Koehn et al., 2007) to normalise and tokenize English and German descriptions, and we also convert space-separated tokens into subwords (Sennrich et al., 2016b). All models use a common vocabulary of $83,093$ English and $91,141$ German subword tokens. If sentences in English or German are longer than 80 tokens, they are discarded. We train models to translate from English into German and from German into English, and report evaluation of cased, tokenized sentences with punctuation.

Our encoder is a bidirectional RNN with GRU, one 1024D single-layer forward and one 1024D single-layer backward RNN. Source and target word embeddings are 620D each and trained jointly with the model. Word embeddings and other non-recurrent matrices are initialised by sampling from a Gaussian $\mathcal{N}(0, 0.01^2)$, recurrent

matrices are random orthogonal and bias vectors are all initialised to $\vec{0}$.

Visual features are obtained by feeding images to the pre-trained ResNet-50 and using the activations of the `res4f` layer (He et al., 2015). We apply dropout with a probability of 0.5 in the encoder bidirectional RNN, the image features, the decoder RNN and before emitting a target word. We follow Gal and Ghahramani (2016) and apply dropout to the encoder bidirectional and the decoder RNN using one same mask in all time steps.

All models are trained using stochastic gradient descent with ADADELTA (Zeiler, 2012) with minibatches of size 80 (text-only NMT) or 40 ($\text{NMT}_{\text{SRC+IMG}}$), where each training instance consists of one English sentence, one German sentence and one image ($\text{NMT}_{\text{SRC+IMG}}$). We apply early stopping for model selection based on BLEU4, so that if a model does not improve on BLEU4 in the validation set for more than 20 epochs, training is halted.

Our models' translation quality are evaluated quantitatively in terms of BLEU4 (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), TER (Snover et al., 2006), and chrF3 (Popović, 2015).[1] We report statistical significance with approximate randomisation for the first three metrics using the MultEval tool (Clark et al., 2011).

7.2.1 Baselines

We wish to compare our model to two main baselines. The first one is a text-only NMT model which makes no use of any image features, but that has the same overall configuration and hyperparameters as our multi-modal NMT model. This allows us to measure how do the additional image features help neural MT. The second one is a PBSMT model that is trained on the same amount of data as our multi-modal NMT model, but again without making use of the image features. This allows us to measure how the two architectures, phrase-based versus neural, compare to each

[1] We specifically compute character 6-gram F3, and additionally character precision and recall for comparison.

other.

We train one text-only PBSMT and one text-only NMT model for each translation direction for comparison (English→German and German→English). Our PBSMT baseline is built with Moses and uses a 5–gram LM with modified Kneser-Ney smoothing. It is trained on the English→German (German→English) descriptions of the M30k$_{\mathrm{T}}$, whereas its LM is trained on the German (English) descriptions only. We use minimum error rate training to tune the model (Och, 2003) with BLEU. The text-only NMT baseline is the one described in Chapter 4 and is trained on the M30k$_{\mathrm{T}}$'s English–German descriptions, again in both translation directions.

When translating into German, we also compare our model against two publicly available results obtained with multi-modal attention-based NMT models. The first model is Huang et al. (2016)'s best model trained on the same data as our models, and the second is their best model using additional object detections, respectively models `m1` (image at head) and `m3` in the authors' paper.

We note that the text-only PBSMT and NMT baselines are comparable to the ones trained for the experiments in Chapter 6. However, these two baselines were *trained independently for both set of experiments*, i.e. the random initialisation of their parameters is different for each experiment, which explains the small variation in the results in Tables 6.1 and 7.1.

7.2.2 Results and Analysis

In Table 7.1, we show results for our text-only baselines NMT and PBSMT, the multi-modal models of Huang et al. (2016) and our MNMT models trained on the M30k$_{\mathrm{T}}$, and pre-trained on the in-domain back-translated M30k$_{\mathrm{C}}$ and the general-domain text-only English-German MT corpora from WMT 2015.

Training on M30k$_{\mathbf{T}}$ One main finding is that our model consistently outperforms the comparable model of Huang et al. (2016) when translating into German, with improvements of +1.4 BLEU and +2.7 METEOR. In fact, even when their model has

English→German						
Model	**Training data**	**BLEU4**↑	**METEOR**↑	**TER**↓	**chrF3**↑ (prec. / recall)	
NMT	$M30k_T$	33.7	52.3	46.7	65.2	(67.7 / 65.0)
PBSMT	$M30k_T$	32.9	54.3†	45.1†	**67.4**	(66.5 / 67.5)
Huang et al. (2016)	$M30k_T$	35.1 (↑ 1.4)	52.2 (↓ 2.1)	—	—	—
	+ RCNN	**36.5** (↑ 2.8)	54.1 (↓ 0.2)	—	—	—
$NMT_{SRC+IMG}$	$M30k_T$	**36.5**†‡	**55.0**†	**43.7**†‡	67.3	(66.8 / 67.4)
Improvements						
$NMT_{SRC+IMG}$ vs. NMT		↑ **2.8**	↑ **2.7**	↓ **3.0**	↑ 2.1	↓ 0.9 / ↑ 2.4
$NMT_{SRC+IMG}$ vs. PBSMT		↑ **3.6**	↑ 0.7	↓ **1.4**	↓ 0.1	↑ 0.3 / ↓ 0.1
$NMT_{SRC+IMG}$ vs. Huang		↑ 1.4	↑ 2.8	—	—	—
$NMT_{SRC+IMG}$ vs. Huang (+RCNN)		↑ 0.0	↑ 0.9	—	—	—
Pre-training data set: back-translated $M30k_C$ (in-domain)						
PBSMT (LM)	$M30k_T$	34.0	**55.0**†	44.7	**68.0**	(66.8 / 68.1)
NMT	$M30k_T$	35.5‡	53.4	43.3‡	65.2	(67.7 / 65.0)
$NMT_{SRC+IMG}$	$M30k_T$	**37.1**†‡	54.5†	**42.8**†‡	66.6	(67.2 / 66.5)
$NMT_{SRC+IMG}$ vs. best PBSMT		↑ **3.1**	↓ 0.5	↓ **1.9**	↓ 1.4	↑ 0.4 / ↓ 1.6
$NMT_{SRC+IMG}$ vs. NMT		↑ **1.6**	↑ **1.1**	↓ **0.5**	↑ 1.4	↓ 0.5 / ↑ 1.5
Pre-training data set: WMT'15 English-German corpora (general domain)						
PBSMT (concat)	$M30k_T$	32.6	53.9	46.1	67.3	(66.3 / 67.4)
PBSMT (LM)	$M30k_T$	32.5	54.1	46.0	67.3	(66.0 / 67.4)
NMT	$M30k_T$	37.8	56.7	41.0	69.2	(69.7 / 69.1)
$NMT_{SRC+IMG}$	$M30k_T$	**39.0**†‡	**56.8** ‡	**40.6** ‡	**69.6**	(69.6 / 69.6)
$NMT_{SRC+IMG}$ vs. best PBSMT		↑ **6.4**	↑ **2.7**	↓ **5.4**	↑ 2.3	↑ 3.3 / ↑ 2.2
$NMT_{SRC+IMG}$ vs. NMT		↑ **1.2**	↑ 0.1	↓ 0.4	↑ 0.4	↓ 0.1 / ↑ 0.5

Table 7.1: BLEU4, METEOR, chrF3, character-level precision and recall (higher is better) and TER scores (lower is better) on the translated Multi30k ($M30k_T$) test set. Best text-only baselines results are underlined and best overall results appear in bold. We show Huang et al. (2016)'s improvements over the best text-only baseline in parentheses. Results are significantly better than the NMT baseline (†) and the SMT baseline (‡) with $p < 0.01$ (no pre-training) or $p < 0.05$ (when pre-training either on the back-translated $M30k_C$ or WMT'15 corpora).

access to more data our model still improves by +0.9 METEOR, while maintaining the same BLEU4 scores.

Moreover, we can also conclude from Table 7.1 that PBSMT performs better at recall-oriented metrics, i.e. METEOR and chrF3, whereas NMT at precision-oriented ones, i.e. BLEU4. This is somehow to be expected, since the attention mechanism used in the NMT model by Bahdanau et al. (2015), adopted in our work, does not explicitly take attention weights from previous time steps into account, therefore lacking the notion of source coverage as in SMT (Koehn et al., 2003; Tu et al., 2016). We note that these ideas are complimentary and that incorporating

German→English						
Model	**Training data**	**BLEU4↑**	**METEOR↑**	**TER↓**	**chrF3↑** (prec. / recall)	
PBSMT	$M30k_T$	32.8	34.8	43.9	61.8	(63.4 / 61.6)
NMT	$M30k_T$	38.2	35.8	40.2	62.8	(65.5 / 62.5)
$NMT_{SRC+IMG}$	$M30k_T$	**40.6**†‡	**37.5**†‡	**37.7**†‡	**65.2**	(68.1 / 64.9)
Improvements						
$NMT_{SRC+IMG}$ vs. NMT		↑ **2.4**	↑ **1.7**	↓ **2.5**	↑ **2.4**	↑ 2.6 / ↑ 2.4
$NMT_{SRC+IMG}$ vs. PBSMT		↑ **7.8**	↑ **2.7**	↓ **1.4**	↑ **6.2**	↑ 4.7 / ↑ 3.3
Pre-training data set: back-translated $M30k_C$ (in-domain)						
PBSMT	$M30k_T$	36.8	36.4	40.8	64.5	(65.7 / 64.4)
NMT	$M30k_T$	42.6	38.9	36.1	67.6	(69.3 / 67.5)
$NMT_{SRC+IMG}$	$M30k_T$	**43.2**‡	**39.0**‡	**35.5**‡	**67.7**	(70.1 / 67.5)
Improvements						
$NMT_{SRC+IMG}$ vs. PBSMT		↑ **6.4**	↑ **2.6**	↓ **5.3**	↑ **3.2**	↑ 4.4 / ↑ 3.1
$NMT_{SRC+IMG}$ vs. NMT		↑ 0.6	↑ 0.1	↓ 0.6	↑ 0.1	↑ 0.8 / ↑ 0.0

Table 7.2: BLEU4, METEOR, chrF3, character-level precision and recall (higher is better) and TER scores (lower is better) on the translated Multi30k ($M30k_T$) test set. Best text-only baselines results are underlined and best overall results appear in bold. We show Huang et al. (2016)'s improvements over the best text-only baseline in parentheses. Results are significantly better than the NMT baseline (†) and the SMT baseline (‡) with $p < 0.01$.

coverage into model $NMT_{SRC+IMG}$ could lead to even more improvements, especially in recall-oriented metrics as METEOR and chrF3. Nonetheless, our doubly-attentive model shows consistent gains in both precision- and recall-oriented metrics in comparison to the text-only NMT baseline, i.e. it is significantly better according to BLEU4, METEOR and TER ($p < 0.01$), and it also improves chrF3 by +2.1. In comparison to the PBSMT baseline, our proposed model still significantly improves according to both BLEU4 and TER ($p < 0.01$), also increasing METEOR by +0.7 but with an associated p-value of $p = 0.071$, therefore not significant for $p < 0.05$. Although chrF3 is the only metric in which the PBSMT model scores best, the difference between our model and the latter is only 0.1, i.e. they are practically equivalent. We note that model $NMT_{SRC+IMG}$ consistently increases character recall in comparison to the text-only NMT baseline. Although it can happen at the expense of character precision, gains in recall are always much higher than any

eventual loss in precision, leading to consistent improvements in chrF3.

In Table 7.2, we observe that when translating into English and training on the original M30k$_{\mathrm{T}}$, model NMT$_{\mathrm{SRC+IMG}}$ outperforms both baselines by a large margin, according to all four metrics evaluated. We also note that both model NMT$_{\mathrm{SRC+IMG}}$'s character-level precision and recall are higher than those of the two baselines, in contrast to results obtained when translating from English into German. This suggests that model NMT$_{\mathrm{SRC+IMG}}$ might better integrate the image features when translating into an "easier" language with less morphology, although experiments involving more language pairs are necessary to definitively confirm whether this is indeed the case.

Pre-training We now discuss results for models pre-trained using different data sets. We first pre-trained the two text-only baselines PBSMT and NMT and our model NMT$_{\mathrm{SRC+IMG}}$ on the back-translated M30k$_{\mathrm{C}}$, a medium-sized in-domain image description data set (145k training instances), in both directions. We also pre-trained the same models on the English–German parallel sentences of much larger MT data sets, i.e. the concatenation of the Europarl, Common Crawl and News Commentary corpora, used in WMT 2015 ($\sim$4.3M parallel sentences). Model PBSMT (concat.) used the concatenation of the pre-training and training data for training, and model PBSMT (LM) only used the general-domain German sentences as additional data to train the LM. From Tables 7.1 and 7.2, it is clear that model NMT$_{\mathrm{SRC+IMG}}$ can learn from both in-domain, multi-modal pre-training data sets as well as text-only, general domain ones.

Pre-training on M30k$_{\mathbf{C}}$ When pre-training on the back-translated M30k$_{\mathrm{C}}$ and translating into German, our model's chrF3 is lower compared to the PBSMT baseline, which is mostly due to character recall; nonetheless, our model still improved by the same margin on the text-only NMT baseline. Moreover, our model outperforms the PBSMT baseline according to BLEU4 and TER, and the text-only NMT baseline according to all metrics ($p < 0.05$).

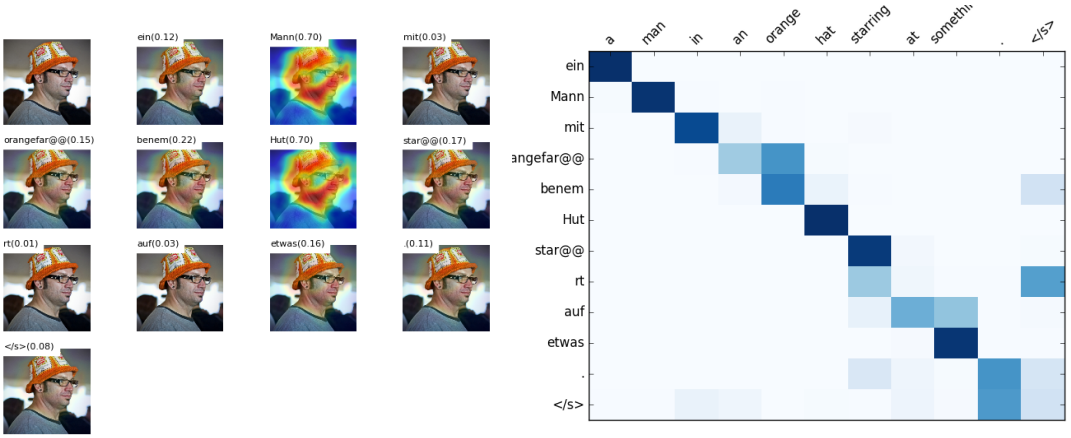
When translating into English, model $\text{NMT}_{\text{SRC+IMG}}$ still consistently scores better according to all metrics evaluated, although the differences between its translations and those obtained with the NMT baseline are no longer statistically significant ($p < 0.01$).

Pre-training on WMT 2015 corpora We also pre-trained our English–German models on the WMT 2015 corpora, which took 10 days, i.e. ∼6–7 epochs. Results show that model $\text{NMT}_{\text{SRC+IMG}}$ improves significantly over the NMT baseline according to BLEU4, and is consistently better than the PBSMT baseline according to all four metrics. In order for PBSMT models to remain competitive, we believe more advanced data selection techniques are needed. However, this line of inquiry is out of the scope of this work.

Overall, we found a strong indication that model $\text{NMT}_{\text{SRC+IMG}}$ can exploit the additional pre-training data efficiently, both general- and in-domain. While the PBSMT model is still competitive when using additional in-domain data—according to METEOR and chrF3—, the same cannot be said when using general-domain pre-training corpora. From our experiments, NMT models in general, and especially model $\text{NMT}_{\text{SRC+IMG}}$, thrive when training and test domains are mixed, which is a very common real-world scenario.

Textual and visual attention In Figure 7.2, we visualise the visual and textual attention weights for an entry of the M30k_{T} test set. In the visual attention, the β gate (written in parentheses after each word) caused the image features to be used mostly to generate the words *Mann* (man) and *Hut* (hat), two highly *visual terms* in the sentence. We observe that in general visually grounded terms, e.g. *Mann* (man) and *Hut* (hat), usually have a high associated β value, whereas other less visual terms like *mit* (with) or *auf* (at) do not. That causes the model to use the image features when it is describing a visual concept in the sentence, which is an interesting feature of our model. Interestingly, our model is very selective when choosing to use image features: it only assigned $\beta > 0.5$ for 20% of the decoded target words for the

test set, and $\beta > 0.8$ to only 8%. A manual inspection of translations shows that these words are mostly concrete nouns with a strong visual appeal.



(a) Image–target word alignments.

(b) Source–target word alignments.

Figure 7.2: Visualisation of image– and source–target word alignments for the M30k_{T} test set.

Lastly, using two independent attention mechanisms is a good compromise between model compactness and flexibility. While the attention-based NMT model baseline has ∼200M parameters, model $\text{NMT}_{\text{SRC+IMG}}$ has ∼213M, therefore using just ∼6.6% more parameters than the latter.

7.3 Experiments on the eBay data set

We now describe experiments where we apply model $\text{NMT}_{\text{SRC+IMG}}$ to translate the eBay datasets. For training our models we use eBay24k (Section 3.3.1) and the M30k_{T} (Section 3.1.1) data sets. In order to measure the impact caused by the size of the training data, we also include the back-translated eBay80k data set in our experiments (Section 3.3.2).

We use the scripts in the Moses SMT Toolkit (Koehn et al., 2007) to normalise, truecase and tokenize English and German descriptions and we also convert space-separated tokens into subwords (Sennrich et al., 2016b). All models use a common vocabulary of $32,025$ English and $32,488$ German subword tokens. If sentences in English or German are longer than 80 tokens, they are discarded.

We evaluate our models quantitatively using BLEU4 (Papineni et al., 2002) and TER (Snover et al., 2006) and report statistical significance computed using approximate randomisation with the Multeval toolkit (Clark et al., 2011).

7.3.1 Baselines

We again wish to compare our model to two main baselines. The first one is a text-only NMT model which makes no use of any image features, but that has the same overall configuration and hyperparameters as our multi-modal NMT model. This allows us to measure how do the additional image features help neural MT models translate noisy data. The second one is a PBSMT model that is trained on the same amount of data as our multi-modal NMT model, but again without making use of the image features. This allows us to measure how the two architectures, phrase-based versus neural, compare to each other.

We train a text-only PBSMT and a text-only NMT model for comparison. The PBSMT model we use in our n-best re-ranking experiments is trained on 120k in-domain parallel sentences and is built using the Moses SMT Toolkit (Koehn et al., 2007). The language model (LM) is a 5–gram LM with modified Kneser-Ney smoothing (Kneser and Ney, 1995). We use minimum error rate training (Och, 2003) for tuning the model parameters for BLEU scores. The text-only NMT baseline is the one described in Chapter 4 and is trained on the M30k$_{\mathrm{T}}$'s English–German descriptions.

In order to measure how well do text-only and multi-modal NMT models perform in re-ranking n-best lists, we train these NMT models using the eBay24k and the M30k$_{\mathrm{T}}$ data sets only. We do not include the back-translated data set when training NMT models for re-ranking n-best lists to be able to evaluate these two scenarios independently.

Model	Training data	BLEU	TER
PBSMT	eBay24k + $M30k_T$	26.1	**54.9**
	+ backtranslated eBay80k	**27.4** ↑ 1.3	55.4 ↑ 0.5
Text-only NMT	eBay24k + $M30k_T$	21.1	60.0
	+ backtranslated eBay80k	22.5 ↑ 1.4	58.0 ↓ 2.0
$NMT_{SRC+IMG}$	eBay24k + $M30k_T$	17.8	62.2
	+ backtranslated eBay80k	**25.1** ↑ 7.3	**55.5** ↓ 6.7
Improvements			
$NMT_{SRC+IMG}$ vs. Text-only NMT		↑ 2.3	↓ 2.5
$NMT_{SRC+IMG}$ vs. PBSMT		↓ 2.3	↑ 0.6

Table 7.3: Comparative results with PBSMT, text-only NMT and multi-modal models $NMT_{SRC+IMG}$. Best PBSMT and NMT results in bold.

7.3.2 Results and Analysis

In Table 7.3 we present quantitative results obtained with the two text-only PBSMT and NMT baselines and one multi-modal model $NMT_{SRC+IMG}$.

It is clear that the gains from adding more data are much more apparent to the multi-modal model $NMT_{SRC+IMG}$ than to the two text-only ones. This can be attributed to the fact that this model effectively has access to more data, i.e. image features, and consequently can learn better representations derived from them. The PBSMT model's improvements are inconsistent; its TER score even deteriorates by 0.5 with the additional data. The same does not happen with the neural MT models, which both (text-only and multi-modal) benefit from the additional data. Model $NMT_{SRC+IMG}$'s gains are more than 3× larger than that of text-only NMT and PBSMT baselines, indicating that they can properly exploit the additional data. Nevertheless, even with the added back-translated data, model $NMT_{SRC+IMG}$ still falls behind the PBSMT model both in terms of BLEU and TER. However, the difference between the two models, PBSMT and $NMT_{SRC+IMG}$, seems to be getting increasingly smaller with the increase in the data size.

In Table 7.4, we show results for re-ranking an 10-, 100- and 1,000-best lists generated by a PBSMT system trained using 120k in-domain parallel sentences. When $n = 10$, both models text-only NMT and $NMT_{SRC+IMG}$ significantly improve

Re-ranking model	Training data	n-best size	BLEU	METEOR	TER	chrF3
baseline	eBay120k		29.0	48.4	53.0	—
Text-only NMT	eBay100k	10	29.3 ↑ 0.3	48.5 ↑ 0.1	52.4 † ↓ 0.6	—
$\text{NMT}_{\text{SRC+IMG}}$	eBay24k+M30k$_{\text{T}}$	10	**29.4** ↑ 0.4	48.5 ↑ 0.1	52.1 † ↓ 0.9	—
Text-only NMT	eBay100k	100	28.9 ↓ 0.1	**48.6** ↑ 0.2	53.6 ↑ 0.6	—
$\text{NMT}_{\text{SRC+IMG}}$	eBay24k+M30k$_{\text{T}}$	100	28.9 ↓ 0.1	48.4	52.4 † ↓ 0.6	—
Text-only NMT	eBay100k	1,000	28.9 ↓ 0.1	48.1 ↓ 0.3	**51.7** † ↓ 1.3	—
$\text{NMT}_{\text{SRC+IMG}}$	eBay24k+M30k$_{\text{T}}$	1,000	28.6 ↓ 0.4	48.2 ↓ 0.2	52.1 † ↓ 0.9	—

Table 7.4: Results for re-ranking n-best lists with text-only and multi-modal NMT models. † Difference is statistically significant ($p \leq 0.05$). Best individual results are underscored, best overall results in bold.

on TER, with model $\text{NMT}_{\text{SRC+IMG}}$ performing slightly better. Both models show small but not significant increases in BLEU and METEOR ($p \leq 0.05$). As lists grow larger ($n = 100$ and $1,000$), it seems that both neural models increasingly have more difficulty to re-rank. In this scenario, the text-only NMT's differences in BLEU or TER are not statistically significant, whereas $\text{NMT}_{\text{SRC+IMG}}$'s improvements in TER are. Nonetheless, the text-only NMT slightly deteriorates BLEU (-0.1) and TER ($+0.6$), while model $\text{NMT}_{\text{SRC+IMG}}$'s TER still improved by -0.6 whereas its BLEU decreased by -0.1. In the case where $n = 1,000$, both models also show the same trend and slightly deteriorate BLEU and METEOR, but again with no statistical significance ($p \leq 0.05$). Nonetheless, both multi-modal and text-only NMT models still improve significantly according to TER. We note that model $\text{NMT}_{\text{SRC+IMG}}$'s improvements in TER are consistent across different n-best list sizes; the baseline NMT model's improvements are not.

The best nominal BLEU ($= 29.4$) and TER ($= 51.7$) scores were achieved by model $\text{NMT}_{\text{SRC+IMG}}$ when applied to re-rank 10-best lists and the text-only NMT re-ranking $1,000$-best lists, respectively. Model $\text{NMT}_{\text{SRC+IMG}}$ significantly improves on TER when $n = 10$, 100 and $1,000$, and does not significantly deteriorates translations according to any other metric, suggesting that it can efficiently exploit the additional multi-modal signals.

Remarks We investigate the potential impact of multi-modal NMT in the context of e-commerce product listings. With only a limited number of multi-modal and multilingual training data available, both text-only and multi-modal NMT models still fail to outperform a productive SMT system, contrary to recent findings (Bentivogli et al., 2016). However, the introduction of back-translated data leads to substantial improvements, especially to a multi-modal NMT model. We also found that NMT models trained on small in-domain data can still be successfully used to rescore a standard PBSMT system with significant gains in TER. Since we know from our experiments with LM perplexities that these are very high for e-commerce data, i.e. fluency is quite low, it seems fitting that BLEU scores do not improve as much.

7.3.3 Human Evaluation

Quantitative MT metrics—such as BLEU, METEOR, TER and chrF, used in this work—are very helpful and often provide a good indication of how good translations are. However, in order to have a complete picture of the translations generated by different MT models, we additionally conduct a qualitative human evaluation. In this qualitative human evaluation, we ask bilingual native German speakers:

1. to assess the *multi-modal adequacy* of translations (number of participants $\mathtt{N} = 18$);

2. to *rank* translations generated by different models from best to worst (number of participants $\mathtt{N} = 18$).

Adequacy Humans are presented with the English product listing, the product image and a translation generated by one of the models (without knowing which model). They are then asked how much of the meaning in the source is also expressed in the translation, while taking the product image into consideration. They must then select from a four-level Likert scale where the answers range from *1 – All of it*

Model	BLEU4↑	TER↓	Adequacy↓
Text-only NMT	22.5	58.0	2.71 ± .48
$\text{NMT}_{\text{SRC+IMG}}$	25.1†	**55.5**†	**2.36** ± .47
SMT	**27.4**†‡	**55.4**†	**2.36** ± .47

Table 7.5: Adequacy of translations and two automatic metrics on the eBay24k test set. Automatic metrics were computed with the MultEval tool (Clark et al., 2011) and results are significantly better than those of the text-only NMT (indicated by †) or $\text{NMT}_{\text{SRC+IMG}}$ (indicated by ‡) with $p < 0.01$.

to *4 – None of it*.

Ranking We present humans with a product image and three translations obtained from different models for a particular English product listing (without identifying the models) and ask them to rank translations from best to worst.

7.3.3.1 Results and Analysis

In Table 7.5, we contrast the human assessments of the adequacy of translations obtained with two text-only baselines PNSMT and NMT and one multi-modal model $\text{NMT}_{\text{SRC+IMG}}$ with scores obtained computing the two automatic MT metrics BLEU Papineni et al. (2002) and TER Snover et al. (2006).

Both $\text{NMT}_{\text{SRC+IMG}}$ and the PBSMT baseline improve on the text-only NMT according to both automatic metrics ($p < 0.01$) (Clark et al., 2011). Although a one-way anova did not show any statistically significant differences in adequacy between $\text{NMT}_{\text{SRC+IMG}}$ and the text-only NMT baseline ($\text{F}(2, 18) = 1.29$, $p > 0.05$), human evaluators ranked $\text{NMT}_{\text{SRC+IMG}}$ as better than the former over 88% of the time, a strong indication that images do help neural MT and bring important information that the multi-modal model $\text{NMT}_{\text{SRC+IMG}}$ can efficiently exploit.

If we compare $\text{NMT}_{\text{SRC+IMG}}$ and the PBSMT system, the latter outperforms the former according to BLEU, but are virtually no different according to TER (Table 7.5). In fact, the adequacy scores for both these models are, on average, the same according to scores computed over $\texttt{N} = 18$ different human assessments. Nonetheless, even though both models $\text{NMT}_{\text{SRC+IMG}}$ and PBSMT are found to

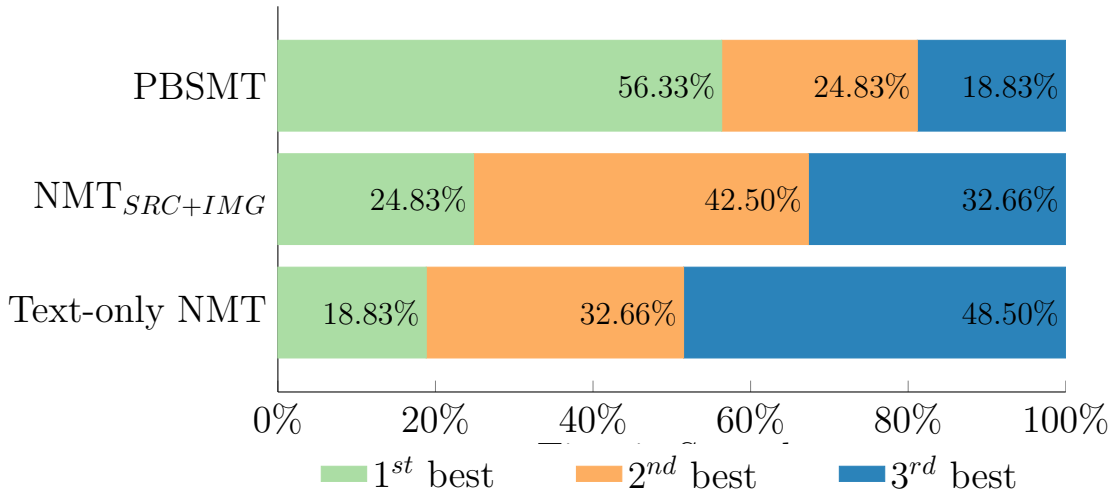


Figure 7.3: Models PBSMT, text-only NMT and NMT$_{\text{SRC+IMG}}$ ranked by humans from best to worst.

produce equally adequate output, translations obtained with the PBSMT model are ranked best by humans over 56.3% of the time, while translations obtained with the multi-modal model NMT$_{\text{SRC+IMG}}$ are ranked best 24.8% of the time, as can be seen in Figure 7.3.

Translations generated by model NMT$_{\text{SRC+IMG}}$ contain many neologisms, possibly due to training these models using sub-word tokens rather than just words (Sennrich et al., 2016a). Some examples are: "sammlerset", "garagenskateboard", "kampfaltschlocker", "schneidsattel" and "oberreceiver". Since the PBSMT model was trained directly on words and consequently does not present these issues, we argue that this generative quality of the NMT models and the data sets evaluated in this work could have made translations more confusing for native German speakers to understand, at least partially explaining the preference for the PBSMT translations.

Images bring important information to NMT models in this context; in fact, translations obtained with a multi-modal NMT model are preferred to ones obtained with a text-only model over 88% of the time. Nevertheless, humans still prefer phrase-based SMT over NMT output. We attribute this to the nature of the task: listing titles have little syntactic structure yet many rare words, which especially for subword units produce many confusing neologisms. The core neural MT models still have to be improved significantly to address these challenges. However, in

contrast to SMT, they already provide an effective way of improving MT quality with information contained in images.

7.4 Final Remarks

In this chapter, we have proposed and evaluated model $\text{NMT}_{\text{SRC+IMG}}$, which incorporates local image features into attention-based NMT. Similarly to the experiments with global image features in Chapter 6, the main goal in this chapter is to use local image features to visually ground translations and increase translation quality.

We showed through extensive experiments that using local image features in model $\text{NMT}_{\text{SRC+IMG}}$ to translate image descriptions of the Multi30k data set leads to improvements over both text-only PBSMT and NMT baselines, as well as over the multi-modal NMT model of Huang et al. (2016), the best performing purely neural model in the first shared multi-modal MT shared task (Specia et al., 2016). Moreover, additional experiments demonstrated the feasibility of exploiting larger text-only MT corpora when using model $\text{NMT}_{\text{SRC+IMG}}$. This performs well when pre-trained on both in-domain as well as out-of-domain corpora, which indicates its suitability to be used in real-world use-cases.

Finally, we also conducted experiments where we analysed the translation of eBay's product listings data set, arguably a much more difficult scenario for MT. Even though model $\text{NMT}_{\text{SRC+IMG}}$ consistently improves over a corresponding text-only NMT baseline, it still does not outperform a PBSMT baseline according to our experiments. Additionally, in our human experiments we found that humans normally prefer translations of product listings generated by a PBSMT system to translations generated with model $\text{NMT}_{\text{SRC+IMG}}$. Nonetheless, humans also found translations of a PBSMT system and model $\text{NMT}_{\text{SRC+IMG}}$ to describe the accompanying product image equally adequately.

Finally, despite these inconsistent results when translating product listings with model $\text{NMT}_{\text{SRC+IMG}}$, we successfully used it to re-rank n-best lists generated with

a PBSMT system and report consistent improvements in TER, which is in itself a promising finding.

Chapter 8

Conclusions and Future Work

In this thesis we have introduced and discussed different multi-modal Neural Machine Translation (NMT) models. Our multi-modal NMT models all make use of state-of-the-art pre-trained CNNs for image feature extraction, specifically the VGG and the Residual Networks. We use *global* image features in three of our models (IMG_{2W}, IMG_{E}, and IMG_{D}), and *local* image features in a fourth model, $\text{NMT}_{\text{SRC+IMG}}$. Global visual features are widely used in transfer learning scenarios, such as in visual question answering (Zhang et al., 2016), to train multi-modal word embeddings (Lazaridou et al., 2015) or in re-ranking n-best lists generated by a PBSMT model (Hitschler et al., 2016). Local visual features have also been proven to perform strongly in a transfer learning image description generation scenario (Xu et al., 2015), which is closely related to multi-modal machine translation. All these taken altogether, we have had strong indication that including image features can indeed help MT.

We have included many different important dimensions when evaluating our different multi-modal NMT models: *(i)* we have studied the application of our models to translate in two different domains, general-purpose image descriptions and eBay's product listings; *(ii)* we have applied our models to translate into two different language settings, English→German and German→English; *(iii)* we included ablative experiments where we added more training data to our models, both in the form

of back-translated synthetic data and text-only Machine Translation data, in both cases trying to simulate a real-world scenario where these sub-optimal data might be available; *(iv)* we have conducted an error analysis of translations obtained for the general-purpose image descriptions data set Multi30k, and carried out a human evaluation of translations obtained for eBay's product listings translations.

All multi-modal models, using either global or local features, consistently improved the translations of general-purpose image descriptions in comparison to the baselines. This holds true when translating both into English and into German. Among the models using global image features, IMG_{2W} performed the worst, which is the most similar to the models introduced by Huang et al. (2016). Model $\text{NMT}_{\text{SRC+IMG}}$, which uses local image features, shows consistent improvements and also can efficiently exploit additional data, either in the form of back-translated data added to the original training data, or in the form of text-only MT corpora, incorporated into the model in a pre-training stage.

Applying our multi-modal NMT models to eBay's product listings data set proved to be much more difficult than to translate the general-purpose image description Multi30k data set. Models IMG_{2W}, IMG_{E}, and IMG_{D} that use global image features were not significantly different from the text-only NMT baseline in this use-case, which was an unexpected finding. We conjecture that the low average word frequencies for German (3.5 for the eBay24k and 9.5 for the concatenation of the eBay24k and eBay80k), as well as the high number of low frequency words in both of these data sets (36.9% of the words in the eBay24k and 13.7% of the words in the concatenation of the eBay24k and eBay80k) could be the culprit in preventing neural MT models from outperforming a comparable PBSMT baseline system.

However, model $\text{NMT}_{\text{SRC+IMG}}$ improves over a comparable text-only NMT baseline in all scenarios, which is a positive finding that indicates that it is flexible and applicable to different use-cases. Still, it does not outperform a PBSMT baseline when applied to translate user-generated product listings, according to our experiments. Moreover, in our human evaluation we found that humans normally

prefer translations of product listings generated by a PBSMT system to translations generated with either text-only NMT and $\text{NMT}_{\text{SRC+IMG}}$, both neural models. Nonetheless, when asked about the multi-modal adequacy of translations, humans found translations of a PBSMT system and model $\text{NMT}_{\text{SRC+IMG}}$ to describe the accompanying product image equally adequately. Finally, despite these inconsistent results when translating product listings with model $\text{NMT}_{\text{SRC+IMG}}$, we successfully used it to re-rank n-best lists generated with a PBSMT system and report consistent improvements in TER, which is in itself a promising finding.

At an initial stage of this thesis, we started experimenting with incorporating both global and local image features into NMT. These preliminary efforts resulted in a system description paper submitted to the multi-modal MT shared task in WMT 2016 (Calixto et al., 2016). In this submission, we compared our first implementations of the multi-modal models IMG_{D} (described in Chapter 6) and a *combination* of IMG_{D} and $\text{NMT}_{\text{SRC+IMG}}$ (described in Chapter 7), both applied to translate image descriptions for the Multi30k data set (discussed in Section 3). This combined model effectively used global image features for decoder initialisation, as in IMG_{D}, and local image features in an independent attention mechanism, as in $\text{NMT}_{\text{SRC+IMG}}$. At that time, we obtained promising results since using the additional local features improved translations in comparison to using only global features, but even our best multi-modal NMT models still lagged behind results obtained with a phrase-based SMT baseline system. We hypothesise that the main point in this submission that caused this difference was the relatively small size of the network we trained in comparison to state-of-the-art networks we report in this thesis. For comparison, word embedding matrices were 300D, whereas in our current experiments in this thesis they are 620D, and (source and target) word embeddings strongly contribute to the final capacity of a model. In further experiments, we increased the size of the networks and did not find significant improvements from using additional local features in comparison to only global ones. For this reason, a short time after our submission to the multi-modal MT shared task in WMT 2016 we decided to explore

both global and local features in separate models. Nevertheless, we believe that an in-depth study of how multi-modal NMT models can effectively exploit both types of image features in a single architecture can be an interesting avenue for future work.

In Chapter 5 we started by devising a discriminative model to incorporate images and text applicable to different Natural Language Processing tasks. To this end, we introduced model MLMME, which exploits multilingual and multi-modal data to train sentence-level embeddings. We evaluated our model on the three NLP tasks of image-sentence ranking, sentence textual similarity, and neural machine translation, and showed that it improves over a comparable text-only baseline in all the three tasks, also outperforming the best submission in the comparable SemEval at the time in the in-domain sentence textual similarity task. In Section 5.3.3, we further explored how effective are MLMME, and under what conditions MLMME improves re-ranking n-best lists generated by text-only NMT models. We found that using MLMME in re-ranking is very robust to the overall quality of the baseline NMT model used to generate the n-best lists, improving translations even when a very weak or very strong baseline is used. The same did not hold when we used the VSE model to re-rank, which deteriorated translations specially when used to re-rank the stronger baselines.

In Chapter 6 we introduced experiments in which we study the use of global image features extracted with the pre-trained VGG19 network in neural MT. We propose three different models, IMG_{2W}, IMG_{E}, and IMG_{D}, each of them incorporating global features in a different way. We evaluate these models on two main scenarios: when translating general-purpose image descriptions from Flickr, i.e. the Multi30k data set, and when translating eBay product listings. When translating the Multi30k, all models show strong improvements over the the corresponding text-only baselines, even though model IMG_{2W} does not perform as well as the others. Results are positive both when translating into German and into English, showing that the three models can integrate visual information effectively for both target languages.

To the best of our knowledge, this is the first time that a purely neural translation model has been found to outperform a PBSMT baseline on this data set, according to all metrics evaluated.

However, NMT models which incorporate global image features' results on the eBay data set are not as positive. Neither of the three multi-modal models significantly improved over the corresponding text-only NMT baseline, and in general the PBSMT had the best translations in this use-case. We emphasise that none of the multi-modal NMT models deteriorated the translations in comparison to the NMT baseline as well. In order to able to pinpoint the reason for these negative results, we first ran experiments where we evaluated the quality of the image features obtained for the eBay product images: we fed a random sample of a few product images from the eBay data sets into the pre-trained VGG19 CNN and inspected the results, specifically one out of $1,000$ possible classes from ImageNet (Russakovsky et al., 2015). The results were surprisingly accurate, meaning that the classification was correct for a vast majority of the cases investigated. This suggested that the culprit in our multi-modal NMT models was not the visual component, but had to do with the textual part of the data. When we look deeper into the eBay product listings, we find that it has very low average word frequencies for both English and German words, and a very high number of singletons, i.e. words that appear only once in the text, compared to both the Multi30k and the WMT 2015 descriptions. Moreover, perplexities computed with different in-domain and out-of-domain LMs on the eBay test set suggest that it is indeed a very difficult test set for automatic processing (Section 3.3.3). One last possible reason we conjecture for the negative results has to do with the vocabulary size of the networks we trained. Vocabulary sizes in the eBay experiments were much smaller than those in the Multi30k experiments, i.e. whereas the dictionary sizes in the eBay experiments were ∼30K, those of the Multi30k experiments were ∼80–90K. More exhaustive experiments comparing performance on both datasets is an interesting potential avenue for future work.

In Chapter 7, we reported experiments on using model $\text{NMT}_{\text{SRC+IMG}}$, which

incorporates local image features, to translate image descriptions from Flickr, i.e. the Multi30k data set, and also eBay product listings. Similarly to our multi-modal NMT models that use global image features, results when translating image descriptions in the Multi30k agreed with our expectations, i.e. translations obtained with our multi-modal model $\text{NMT}_{\text{SRC+IMG}}$ were significantly better than those obtained with any of the text-only baselines, NMT or PBSMT. These results are consistent regardless of the language direction, i.e. the multi-modal model improves when translating into English and also when translating into German. Finally, an important result we observed is that this model can be efficiently pre-trained with both medium-sized back-translated multi-modal data sets and with large text-only MT corpora. Additionally, results are consistent when translating both into a morphologically poor and a morphologically rich language, i.e. German→English and English→German scenarios, respectively.

Differently from Chapter 6, when applying model $\text{NMT}_{\text{SRC+IMG}}$ onto the eBay data set, we found that it consistently outperformed a comparable text-only NMT baseline. However, when we compared it to a PBSMT baseline, results were mixed: the PBSMT baseline was better according to some metrics, e.g. BLEU, but according to other metrics there was no difference between the two models, e.g. TER. In order to better understand these quantitative results and put them in context, we conducted a qualitative evaluation where we asked humans to assess the multi-modal adequacy of product listings and their translations, i.e. whether listings and their translations actually described the contents of the product image and vice-versa—a necessary condition for the image features extracted from CNNs pre-trained for image classification and object detection to be useful in transfer learning scenarios such as neural machine translation of product listings—, and also to rank translations generated by different models from best to worst. We found that humans preferred translations generated by a PBSMT model; however, if they could only choose between translations obtained with a text-only NMT model and the multi-modal model $\text{NMT}_{\text{SRC+IMG}}$, they chose the latter about 88% of the time. This shows

unequivocally that including product images in model $\text{NMT}_{\text{SRC+IMG}}$ also helps neural translation models better translate product listings, although not yet to the point of outperforming the strong PBSMT baseline. Finally, in order to leverage the good translations generated with PBSMT models in the product listings scenario, we used a PBSMT baseline to generate n-best lists and used a text-only NMT model and the multi-modal model $\text{NMT}_{\text{SRC+IMG}}$ to re-rank them. We found that using the multi-modal model $\text{NMT}_{\text{SRC+IMG}}$ to re-rank consistently improves TER scores independently of the n-best list sizes, whereas using the baseline text-only model NMT does not have the same effect.

We now revisit the research questions introduced in Chapter 1, and elaborate on how our experiments and findings in this work address each of them.

> **(RQ1)** *Can we use multi-modal discriminative models to improve the translation of image descriptions?*

In order to address **(RQ1)**, we introduced the discriminative model MLMME that integrates multilingual image descriptions in an arbitrary number of languages and images by means of their global features obtained with a pre-trained CNN. We applied this model to improve the translation of image descriptions from the Multi30k data set, but also in other NLP tasks such as image–sentence ranking and sentence textual similarity. In n-best list re-ranking experiments, we trained NMT models on different data sets, in-domain and general-domain, and using different regularisation hyperparameters, dropout probability and L2 weight. With these experiments, we validated that model MLMME can effectively be used to improve translations in a domain adaptation scenario, especially for smaller n-best list sizes. For larger n-best list sizes, we observed that some of our models could degrade TER scores, although consistently improving METEOR scores and either improving or mantaining BLEU scores. Nonetheless, we emphasise the consistent improvements on the recall-oriented metrics, e.g. METEOR, since recall is known to be a weak spot of NMT models (Mi et al., 2016; Tu et al., 2016). We also validated that model

MLMME can be used to re-rank and improve translations regardless of the quality of the baseline NMT model used to generate the n-best lists, by showing consistent improvements on translations obtained with weak, medium and strong baselines, regardless of the n-best list sizes. Thus, these experiments allowed us to provide an answer to research question **(RQ1)**.

> **(RQ2)** *Given that there is a large number of standard text-only MT corpora, can multi-modal MT models effectively exploit this additional text-only data and provide state-of-the-art performance?*

Our experiments in Chapters 6 and 7 address research question **(RQ2)**. In these two chapters, we evaluated our multi-modal models in three different scenarios: *(i)* in a "normal" training data regime, where the only data used for training is the M30k_T training data; *(ii)* with additional synthetic in-domain training data, where we back-translated entries in the M30k_C training data and added the new synthetic triples to the original M30k_T training set; *(iii)* with additional general-domain text-only training data, where we used publicly-available MT training corpora for pre-training our models, and further fine-tuned these models on the original M30k_T training data.

Our multi-modal models that use global image features and model $\text{NMT}_{\text{SRC+IMG}}$, that uses local image features, can efficiently exploit additional data, either in the form of back-translated in-domain data (as in *(ii)*) or general-domain text-only MT corpora (as in *(iii)*). We systematically evaluated model $\text{NMT}_{\text{SRC+IMG}}$ by pre-training using general-domain text-only MT corpora and in-domain multi-modal back-translated data, and we observed consistent improvements in all metrics both when translating into English and into German, in all cases.

We have therefore positively answered the research question **(RQ2)**.

> **(RQ3)** *How do multi-modal MT models compare to text-only MT models when translating user-generated product listings?*

In order to answer research question **(RQ3)**, we also refer to the experiments we reported in Chapters 6 and 7. From experiments on the eBay data set, none of our multi-modal NMT models that incorporates global image features improved translations compared to a text-only NMT baseline; we also note that none of these models degraded translations. We ran experiments where we evaluated the quality of the image features obtained for the eBay product images when used to classify these product images into one out of the $1,000$ classes from ImageNet, and found that the classification was correct for a vast majority of the cases investigated, i.e. the culprit was the textual part of the data, not the visual. Additionally, in general the PBSMT had the best translations as corroborated by a qualitative evaluation where we asked humans to rank translations from best to worst.

However, when applying model $\text{NMT}_{\text{SRC+IMG}}$ onto the eBay data set, we found that it consistently outperformed a comparable text-only NMT baseline, but when compared to a PBSMT baseline, results were mixed. We found that humans still preferred translations generated by a PBSMT model; however, if they could only choose between translations obtained with a text-only NMT model and the multi-modal model $\text{NMT}_{\text{SRC+IMG}}$, they chose the latter about 88% of the time. This shows unequivocally that including product images in model $\text{NMT}_{\text{SRC+IMG}}$ also helps neural translation models to better translate product listings, although not yet to the point of outperforming the strong PBSMT baseline. Finally, we used a PBSMT baseline to generate n-best lists and used a text-only NMT model and the multi-modal model $\text{NMT}_{\text{SRC+IMG}}$ to re-rank them. We found that using the multi-modal model $\text{NMT}_{\text{SRC+IMG}}$ to re-rank consistently improves TER scores independently of the n-best list sizes, whereas using the baseline text-only model NMT does not have the same effect.

These findings answer the research question **(RQ3)**. Models that use global image features do not fare well in translating product listings, both compared to a text-only NMT model, in which case there are no significant differences between the two, or a PBSMT system trained on the same data, in which case the multimodal

models are significantly worse. Although not a completely positive answer, we found that the multi-modal model $\text{NMT}_{\text{SRC+IMG}}$ improves over a comparable text-only NMT baseline and can still be used to improve translations generated with a baseline PBSMT model in an n-best re-ranking scenario.

8.1 Contributions

We have introduced different models to incorporate images into NMT. We have proposed models to incorporate image features in a re-ranking stage, by training a discriminative model on multilingual and multi-modal data, and also directly into a NMT model, in which case we proposed different models to integrate global and local image features. Our contributions are:

- We have proposed a novel discriminative model that is trained on images and their multilingual descriptions in an arbitrary number of languages. We have evaluated this model in three different tasks, image–sentence ranking, semantic textual similarity and neural machine translation. Our investigations show that: it consistently outperforms all baselines in sentence→image ranking; it shows mixed results in image→sentence ranking; it outperforms the previously published state-of-the-art in two in-domain semantic textual similarity tasks; and it can be used to consistently improve translations obtained with NMT models in n-best list re-ranking experiments.

- We have found that our discriminative model MLMME can be effectively used to improve translations regardless of the quality of the NMT model used to generate the n-best lists, and also in a domain adaptation scenario where the NMT model was trained on OOD data.

- We have proposed three different NMT models that incorporate global image features, and successfully applied them to translate image descriptions and product listings. These models can be pre-trained on back-translated in-

domain data and consistently improve translations of image descriptions from the Multi30k compared to two text-only baselines. Our models also improve on the best published results by multi-modal neural models in the literature.

- NMT models that incorporate global image features are not significantly different than a comparable text-only NMT baseline when applied to translate eBay product listings using the product images. In this scenario, i.e. the translation of eBay user-generated product listings, a PBSMT system delivers the best translations.

- We have proposed a NMT model that integrates local image features, and successfully applied it to translate image descriptions and product listings. Our model can be efficiently pre-trained on general-domain text-only MT corpora as well as in-domain back-translated multi-modal data. It consistently improved translations of image descriptions compared to both a text-only NMT and PBSMT baselines, and consistently improved the translation of product listings compared to a text-only NMT baseline.

- We have shown that although our multi-modal model $\text{NMT}_{\text{SRC+IMG}}$ does not yet outperform a PBSMT baseline on product listings translation, it can still be used to re-rank n-best lists generated by a PBSMT system and consistently improve TER scores of translations.

8.2 Future Work

In our work, we have introduced different multi-modal NMT models and applied them to translate data from two different domains: image descriptions from Flickr, in the experiments where we used the Multi30k data sets, and eBay product listings, in the experiments using the eBay data sets. The former is a more ideal scenario where the data is clean and simple, whereas the latter is a real-world industrial application with all the pitfalls that come with it. In the course of our experimentations, we

have identified many possible avenues for future research which we believe could lead to interesting and improved results.

In Chapter 5, we propose a discriminative model trained to distinguish between positive and negative sentence–sentence and/or sentence–image pairs, i.e. positive pairs are descriptive of one another, whereas negative pairs are not (Section 5.1). This model uses cosine similarity to score the similarity between two sentences in different languages, or one sentence and one image. Luong et al. (2015) introduce different types of attention mechanisms for NMT, and we believe we could adapt his `score` functions and study the effects of applying the `general` and the `concat` functions (Luong et al., 2015) instead of cosine similarity scores in Equations 5.1 and 5.2. Additionally, we would like to conduct additional experiments with data sets that include image descriptions in more than two languages, as is the case of the Multi30k.

In Chapters 6 and 7, we propose different multi-modal NMT models. Recently, there have been many different proposals for improvements aiming to expand and ammeliorate the encoder–decoder framework. We would like to include the notion of attention coverage in our models, i.e. inform the attention mechanism about the attention decisions chosen for the previously emitted target words. This could likely help the NMT model with problems of under- and over-translation, issues from which NMT models are known to suffer (Mi et al., 2016; Tu et al., 2016; Yang et al., 2017).

Additionally, we believe that an interesting avenue for future work involves the in-depth investigation of different forms of integrating both global and local features into one multi-modal NMT model, whether it can boost translation quality even more and, if so, under what circumstances.

We would also like to try to build different NMT models that use only local image features. Instead of merging the mono-modal states, i.e. one that encode the source sentence and another one that encode the image, at the level of the decoder recurrent hidden state, as in model $\text{NMT}_{\text{SRC+IMG}}$, we could try out different ideas. One idea

is to first train two separate networks, one for image description generation and another one for machine translation,[1] making use of all the specialised training data available to these two tasks; then, either merge these two pre-trained models using (a likely small) in-domain multilingual and multi-modal data set, or ensembling the two different models at inference time.

[1]One condition is that the two networks have the same target language, i.e. the IDG target language and the NMT target language must be the same.

Bibliography

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., Rigau, G., Uria, L., and Wiebe, J. (2015). SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Mihalcea, R., Rigau, G., and Wiebe, J. (2014). SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., and Wiebe, J. (2016). SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., and Guo, W. (2013). *SEM 2013 shared task: Semantic Textual Similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA.

Agirre, E., Diab, M., Cer, D., and Gonzalez-Agirre, A. (2012). SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 385–393, Montréal, Canada.

ALPAC (1966). *Language and Machines: Computers in Translation and Linguistics*. The National Academies Press, Washington, DC.

Arora, S., Liang, Y., and Ma, T. (2017). A Simple but Tough-to-Beat Baseline for Sentence Embeddings. In *International Conference on Learning Representations, ICLR 2017*, Toulon, France.

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.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A Neural Probabilistic Language Model. *J. Mach. Learn. Res.*, 3:1137–1155.

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. (2016). Neural versus Phrase-Based Machine Translation Quality: a Case Study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 257–267, Austin, Texas.

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(1):409–442.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. (2016a). Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany.

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., Graham, Y., Kamran, A., and Stanojević, M. (2016b). Results of the WMT16 Metrics Shared Task. In *Proceedings of the First Conference on Machine Translation*, pages 199–231, Berlin, Germany. Association for Computational Linguistics.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Comput. Linguist.*, 19(2):263–311.

Bruni, E., Tran, N. K., and Baroni, M. (2014). Multimodal distributional semantics. *J. Artif. Int. Res.*, 49(1):1–47.

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, Berlin, Germany.

Calixto, I., Chowdhury, K. D., and Liu, Q. (2017a). DCU System Report on the WMT 2017 Multi-modal Machine Translation Task. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark.

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, Berlin, Germany.

Calixto, I., Liu, Q., and Campbell, N. (2017b). Doubly-Attentive Decoder for Multi-modal Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924, Vancouver, Canada.

Calixto, I., Liu, Q., and Campbell, N. (2017c). Incorporating Global Visual Features into Attention-Based Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark.

Calixto, I., Liu, Q., and Campbell, N. (2017d). Multilingual Multi-modal Embeddings for Natural Language Processing. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, Varna, Bulgaria.

Calixto, I., Stein, D., Matusov, E., Castilho, S., and Way, A. (2017e). Human evaluation of multi-modal neural machine translation: A case-study on e-commerce listing titles. In *Proceedings of the Sixth Workshop on Vision and Language*, pages 31–37, Valencia, Spain.

Calixto, I., Stein, D., Matusov, E., Lohar, P., Castilho, S., and Way, A. (2017f). Using images to improve machine-translating e-commerce product listings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 637–643, Valencia, Spain.

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., and Way, A. (2017). Is Neural Machine Translation the New State-of-the-Art? *Prague Bulletin of Mathematical Linguistics*, 10(8):109–120.

Cherry, Colinand Foster, G. (2012). Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436, Montrèal, Canada.

Chiang, D. (2005). A Hierarchical Phrase-based Model for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 263–270, Ann Arbor, Michigan. Association for Computational Linguistics.

Chiang, D. (2007). Hierarchical phrase-based translation. *Comput. Linguist.*, 33(2):201–228.

Cho, K., van Merrienboer, B., Bahdanau, D., and Bengio, Y. (2014a). On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014b). 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.

Chung, J., Gülçehre, Ç., Cho, K., and Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. Technical Report Arxiv report 1412.3555, Université de Montréal. Presented at the Deep Learning workshop at NIPS2014.

Clark, J. H., Dyer, C., Lavie, A., and Smith, N. A. (2011). Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability.

In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 176–181, Portland, Oregon.

Cohen, W. W., Schapire, R. E., and Singer, Y. (1999). Learning to order things. *Journal of Artificial Intelligence Research*, 10:243–270.

Crammer, K. and Singer, Y. (2003). Ultraconservative Online Algorithms for Multiclass Problems. *J. Mach. Learn. Res.*, 3:951–991.

de Saussure, F. (1966). *Course in general linguistics*. New York: McGraw-Hill.

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*, Gothenburg, Sweden. The Association for Computer Linguistics.

Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R., and Makhoul, J. (2014). Fast and Robust Neural Network Joint Models for Statistical Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1370–1380, Baltimore, Maryland. Association for Computational Linguistics.

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 *Proceedings of The 31st International Conference on Machine Learning (ICML)*, page 647–655, Beijing, China.

Dong, D., Wu, H., He, W., Yu, D., and Wang, H. (2015). Multi-Task Learning for Multiple Language Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China.

Elliott, D., Frank, S., and Hasler, E. (2015). Multi-Language Image Description with Neural Sequence Models. *CoRR*, abs/1510.04709.

Elliott, D., Frank, S., Sima'an, K., and Specia, L. (2016). Multi30K: Multilingual English-German Image Descriptions. In *Proceedings of the 5th Workshop on Vision and Language, VL@ACL 2016*, Berlin, Germany.

Farhadi, A., Endres, I., Hoiem, D., and Forsyth, D. (2009). Describing objects by their attributes. In *Proceedings of the IEEE Computer Society Conference*

on Computer Vision and Pattern Recognition (CVPR, pages 1778–1785, Miami, Florida, USA.

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., and Forsyth, D. (2010). Every Picture Tells a Story: Generating Sentences from Images. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, pages 15–29, Berlin, Heidelberg. Springer-Verlag.

Ferreira, T. C., Calixto, I., Wubben, S., and Krahmer, E. (2017). Linguistic realisation as machine translation: Comparing different MT models for AMR-to-text generation. In *Proceedings of the International Conference on Natural Language Generation*, Santiago de Compostela, Spain.

Firat, O., Cho, K., and Bengio, Y. (2016). Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Forcada, M. L. and Ñeco, R. P. (1997). Recursive Hetero-associative Memories for Translation. In *Proceedings of the International Work-Conference on Artificial and Natural Neural Networks: Biological and Artificial Computation: From Neuroscience to Technology*, IWANN '97, pages 453–462, London, UK, UK. Springer-Verlag.

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.

Ganguly, D., Calixto, I., and Jones, G. F. (2016). Developing a Dataset for Evaluating Approaches for Document Expansion with Images. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portoroź, Slovenia.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 580–587, Washington, DC, USA.

Glenberg, A. and Robertson, D. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language*, 43:379–401.

Gong, Y., Wang, L., Hodosh, M., Hockenmaier, J., and Lazebnik, S. (2014). Improving Image-Sentence Embeddings Using Large Weakly Annotated Photo Collections. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision – ECCV 2014*, volume 8692 of *Lecture Notes in Computer Science*, pages 529–545. Springer International Publishing.

Graves, A. (2013). Generating Sequences With Recurrent Neural Networks. *CoRR*, abs/1308.0850.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., and Wang, G. (2015). Recent advances in convolutional neural networks. *arXiv preprint arXiv:1512.07108*.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42:335–346.

He, K. and Sun, J. (2015). Convolutional neural networks at constrained time cost. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 5353–5360, Boston, MA, USA.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385*.

Heafield, K., Pouzyrevsky, I., Clark, J. H., and Philipp (2013). Scalable Modified Kneser-Ney Language Model Estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria.

Hitschler, J., Schamoni, S., and Riezler, S. (2016). Multimodal Pivots for Image Caption Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2399–2409, Berlin, Germany.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780.

Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing Image Description As a Ranking Task: Data, Models and Evaluation Metrics. *J. Artif. Int. Res.*, 47(1):853–899.

Hokamp, C., Calixto, I., Wagner, J., and Zhang, J. (2014). Target-Centric Features for Translation Quality Estimation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 329–334, Baltimore, Maryland, USA.

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*, pages 639–645, Berlin, Germany.

Hutchins, W. J. (1978). Machine translation and machine-aided translation. *Journal of Documentation*, 34:119–159.

Jean, S., Cho, K., Memisevic, R., and Bengio, Y. (2015). On Using Very Large Target Vocabulary for Neural Machine Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Beijing, China. The Association for Computational Linguistics.

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent Continuous Translation Models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*, pages 1700–1709, Seattle, USA.

Karpathy, A., Joulin, A., and Fei-Fei, L. (2014). Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, pages 1889–1897, Cambridge, MA, USA. MIT Press.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. (2017). On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In *International Conference on Learning Representations, ICLR 2017*, Toulon, France.

Kingma, D. P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6980.

Kiros, R., Salakhutdinov, R., and Zemel, R. S. (2014). Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *CoRR*, abs/1411.2539.

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Torralba, A., Urtasun, R., and Fidler, S. (2015). Skip-thought Vectors. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pages 3294–3302, Cambridge, MA, USA. MIT Press.

Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. In *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 181–184, Detroit, Michigan.

Knight, K. (1999). Decoding Complexity in Word-replacement Translation Models. *Comput. Linguist.*, 25(4):607–615.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical Phrase-based Translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Edmonton, Canada. Association for Computational Linguistics.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

Lavie, A. and Agarwal, A. (2007). Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Lazaridou, A., Pham, N. T., and Baroni, M. (2015). Combining Language and Vision with a Multimodal Skip-gram Model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163, Denver, Colorado.

Le, Q., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G., Dean, J., and Ng, A. (2012). Building high-level features using large scale unsupervised learning. In Langford, J. and Pineau, J., editors, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, ICML '12, pages 81–88, Edinburgh, Scotland, GB. Omnipress.

Le, Q. V. and Mikolov, T. (2014). Distributed Representations of Sentences and Documents. In *Proceedings of The 31st International Conference on Machine Learning (ICML)*, page 1188–1196, Beijing, China.

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Libovický, J., Helcl, J., Tlustý, M., Bojar, O., and Pecina, P. (2016). CUNI System for WMT16 Automatic Post-Editing and Multimodal Translation Tasks. In *Proceedings of the First Conference on Machine Translation*, pages 646–654, Berlin, Germany.

Lipton, Z. C. (2016). The Mythos of Model Interpretability. In *ICML 2016 Workshop on Human Interpretability in Machine Learning (WHI 2016)*, New York, NY.

Lowe, D. G. (1999). Object Recognition from Local Scale-Invariant Features. In *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2*, ICCV '99, pages 1150–, Washington, DC, USA. IEEE Computer Society.

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vision*, 60(2):91–110.

Lufkin, J. M. (1965). Human vs machine translation of foreign languages. *Engineering Writing and Speech, IEEE Transactions on*, 8(1):8–14.

Luong, M.-T., Le, Q. V., Sutskever, I., Vinyals, O., and Kaiser, L. (2016). Multi-Task Sequence to Sequence Learning. In *Proceedings of the International Conference on Learning Representations (ICLR), 2016*, San Juan, Puerto Rico.

Luong, T., Pham, H., and Manning, C. D. (2015). Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1412–1421, Lisbon, Portugal.

Mi, H., Sankaran, B., Wang, Z., and Ittycheriah, A. (2016). Coverage Embedding Models for Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 955–960, Austin, Texas.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Commun. ACM*, 38(11):39–41.

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*, ACL '03, pages 160–167, Sapporo, Japan.

Och, F. J. and Ney, H. (2002). Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 295–302, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Och, F. J. and Ney, H. (2004). The Alignment Template Approach to Statistical Machine Translation. *Comput. Linguist.*, 30(4):417–449.

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.

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal.

Popović, M. (2016). chrF deconstructed: beta parameters and n-gram weights. In *Proceedings of the First Conference on Machine Translation*, pages 499–504, Berlin, Germany. Association for Computational Linguistics.

Powers, D. M. W. (1998). Applications and Explanations of Zipf's Law. In *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*, NeMLaP3/CoNLL '98, pages 151–160.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations.

In Rumelhart, D. E., McClelland, J. L., and PDP Research Group, C., editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1*, chapter Learning Internal Representations by Error Propagation, pages 318–362. MIT Press, Cambridge, MA, USA.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Schwenk, H. (2007). Continuous Space Language Models. *Comput. Speech Lang.*, 21(3):492–518.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany.

Sennrich, R., Haddow, B., and Birch, A. (2016b). 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.

Shah, K., Wang, J., and Specia, L. (2016). SHEF-Multimodal: Grounding Machine Translation on Images. In *Proceedings of the First Conference on Machine Translation*, pages 660–665, Berlin, Germany.

Shannon, C. E. and Weaver, W. (1949). *The Mathematical Theory of Communication.*

Silberer, C. and Lapata, M. (2014). Learning Grounded Meaning Representations with Autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 721–732, Baltimore, Maryland.

Simonyan, K. and Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR*, abs/1409.1556.

Slocum, J. (1985). A Survey of Machine Translation: Its History, Current Status, and Future Prospects. *Comput. Linguist.*, 11(1):1–17.

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., Manning, C., and Ng, A. (2014). Grounded Compositional Semantics for Finding and Describing Images with Sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218.

Specia, L., Frank, S., Sima'an, K., and Elliott, D. (2016). A Shared Task on Multimodal Machine Translation and Crosslingual Image Description. In *Proceedings of the First Conference on Machine Translation, WMT 2016*, pages 543–553, Berlin, Germany.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Stanojević, M., Kamran, A., Koehn, P., and Bojar, O. (2015). Results of the WMT15 Metrics Shared Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, Lisbon, Portugal.

Stymne, S. (2011). Blast: A Tool for Error Analysis of Machine Translation Output. In *Proceedings of the ACL-HLT 2011 System Demonstrations*, pages 56–61, Portland, Oregon. Association for Computational Linguistics.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems, NIPS*, NIPS'14, pages 3104–3112, Cambridge, MA, USA. MIT Press.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9.

Tu, Z., Lu, Z., Liu, Y., Liu, X., and Li, H. (2016). Modeling Coverage for Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85, Berlin, Germany.

van Miltenburg, E. (2015). Stereotyping and Bias in the Flickr30K Dataset. In *Proceedings of the Workshop on Multimodal Corpora, MMC-2016*, pages 1–4, Portorož, Slovenia.

Vendrov, I., Kiros, R., Fidler, S., and Urtasun, R. (2016). Order-Embeddings of Images and Language. In *International Conference on Learning Representations, ICLR 2016*, San Juan, Puerto Rico.

Vilar, D., Xu, J., D'haro, L., and Ney, H. (2006). Error Analysis of Statistical Machine Translation Output. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, Genoa, Italy. European Language Resources Association (ELRA). ACL Anthology Identifier: L06-1244.

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

Štajner, S., Calixto, I., and Saggion, H. (2015). Automatic Text Simplification for Spanish: Comparative Evaluation of Various Simplification Strategies. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 618–626, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Williams, R. J. and Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural Comput.*, 1(2):270–280.

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 *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2048–2057, Lille, France. JMLR Workshop and Conference Proceedings.

Yang, Z., Hu, Z., Deng, Y., Dyer, C., and Smola, A. (2017). Neural machine translation with recurrent attention modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 383–387, Valencia, Spain.

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.

Zeiler, M. D. (2012). ADADELTA: An Adaptive Learning Rate Method. *CoRR*, abs/1212.5701.

Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D., and Parikh, D. (2016). Yin and Yang: Balancing and answering binary visual questions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, Nevada, USA.