Sheffield MultiMT: Using Object Posterior Predictions for Multimodal Machine Translation

Pranava Madhyastha*, Josiah Wang* and **Lucia Specia**
Department of Computer Science
University of Sheffield, UK
{p.madhyastha, j.k.wang, l.specia}@sheffield.ac.uk

Abstract

This paper describes the University of Sheffield's submission to the WMT17 Multimodal Machine Translation shared task. We participated in Task 1 to develop an MT system to translate an image description from English to German and French, given its corresponding image. Our proposed systems are based on the state-of-the-art Neural Machine Translation approach. We investigate the effect of replacing the commonly-used image embeddings with an estimated posterior probability prediction for 1,000 object categories in the images.

1 Introduction

This paper describes the University of Sheffield's submission to the second edition of the WMT17 Multimodal Machine Translation shared task. We participate in Task 1, where the challenge is to develop a Machine Translation (MT) system to automatically translate image descriptions to a target language, given an image description in a source language and its corresponding image. We submitted systems for translating from English to both German and French.

Our submission is based on the state-of-the-art attention-based Neural Machine Translation (NMT) system, which has shown better performance than conventional phrase-based statistical MT (SMT) systems in the past years. Multimodal NMT systems have been introduced (Elliott et al., 2015; Caglayan et al., 2016; Calixto et al., 2016; Huang et al., 2016) to incorporate visual information into NMT approaches, most of which condition the NMT on an image representation (typically a vector extracted from a Convolutional Neural Network (CNN) layer). However, it has not been clear thus far whether such image features actually help in the translation task and more important, if so it is not clear which aspects of the image can play a role and how.

Recent approaches to Multimodal NMT have used low level image features, including dense fully connected vectors and spatial convolutional representations from an image classification network (Elliott et al., 2015; Huang et al., 2016). They also incorporate attention mechanisms (Calixto et al., 2016). However, the effect of image features or the efficacy of the representational contribution is still an open research question.

For our submission, we propose replacing image representations used in current Multimodal NMT systems with a class-based probabilistic distribution that is estimated directly using a state-of-the-art image classification network. The core hypothesis is that such representations offer higher level semantic information and could be more beneficial to Multimodal NMT systems.

In Section 2 we discuss the motivations behind our proposed system. In Section 3 we describe our approach, which uses CNN-based image features as input (Section 3.1) to an attention based neural machine translation system (Section 3.2), resulting in a Multimodal NMT system (Section 3.3). Experimental settings are reported in Section 4, and results discussed in Section 5. A brief overview of related work are provided in Section 6.

2 Motivation

Recent work (Wu et al., 2016; You et al., 2016) exploits explicit, higher-level semantic representation of images for the tasks of image captioning and visual question answering. Instead of feeding

*P. Madhyastha and J. Wang contributed equally to this work.

a lower-level image representation directly to the model, such work explicitly explores predicting the occurrence of various concepts (objects, also referred to as attributes) in the image, and feeding such predictions to the language generation component. Our hypothesis is that such an approach, when applied to Multimodal NMT, should provide comparable, if not better results compared to systems that use image representations directly. This approach also offers the advantage of being more interpretable compared to end-to-end systems that use image representations directly. Finally, since the image classification network is trained directly to produce probabilistic class distributions, the predictions are more stable and encoded in simpler representations when compared with the fully connected, lower-level representations. This also presents an opportunity to fine tune the class distributions for the task using domain-specific data. In other words, we can tune the image network to produce better predictions on the classes that appear in the dataset of interest.

Motivated by these insights, we empirically evaluate the performance of a Multimodal NMT system with image features based on predicted class distributions. In most cases we are able to outperform the baseline system under similar settings. In the following section we describe our system in detail.

3 System description

We first describe the image features used in our system, more specifically, the probability prediction of an object category occurring in the image (Section 3.1). We then present the NMT system used (Section 3.2), and how the image features are combined to produce a Multimodal NMT system (Section 3.3) for the shared task. Figure 1 illustrates the proposed system.

3.1 Visual features

Visual features were extracted from the 152-layer version of ResNet (He et al., 2015), a Deep Convolutional Neural Network (CNN) pre-trained on 1,000 object categories (synsets) of the classification task of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2015). We extracted the final layer after applying the softmax function. This layer is a 1,000-dimensional vector providing class posterior probability estimates at image level for the 1,000 object categories, each corresponding to a distinct WordNet synset.

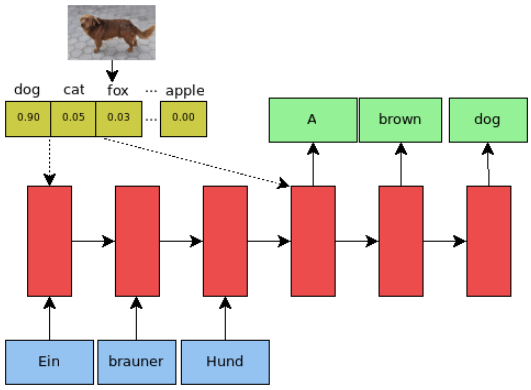


Figure 1: An illustration of our Multimodal NMT system. Departing from usual methods, we replace the lower-level image CNN representation with a vector representing the output of a 1,000-way visual classifier, where each element in the vector represents the estimated posterior probability of an object category occurring in the image. We experiment with conditioning the image representation on either the encoder or the decoder (dashed lines), and also at each source word (not shown in the Figure).

While ResNet has been reported to perform extremely well in classification tasks (3.57% top-5 error rate in the ILSVRC2015 challenge[1], where a prediction is considered correct if the gold standard category is within a system's top 5 guesses), it is worth noting that the model is built for and tuned to the 1,000 categories of ILSVRC, some of which include very fine-grained classifications like various dog species. Thus, many of these categories may not be relevant to the shared task data which is based on the Flickr30K dataset (Young et al., 2014). Conversely, many objects depicted in Flickr30K may also not be covered in the ILSVRC dataset.

3.2 Neural Machine Translation

We use a standard LSTM-based bidirectional encoder-decoder architecture with global attention (Luong et al., 2015). All our NMT models have the following architecture: the input and output vocabulary are limited to words that appear at least twice in the training data and the remaining words are replaced by the $< UNK >$ token. The hidden layer dimensionality is set to 256 and the

[1] http://image-net.org/challenges/LSVRC/2015/results

word dimensionality is set to 128, for both the encoder and decoder, as this configuration was found to lead to faster training times without sacrificing translation performance. At decoding time, we perform greedy decoding by outputing the most probable word at each time step.

3.3 Multimodal Neural Machine Translation

To add visual features, we extend the above mentioned architecture in the following ways:

1. Image features initialising the Encoder (**InitEnc**): As shown in Figure 1, we use the predicted class distribution to initialise only the encoder (i.e. images as the first token). This can be seen as conditioning the encoder on the predicted class distribution.

2. Image features initialising the Decoder (**InitDec**): As we see in Figure 1, here we initialize the decoder's first hidden state with the predicted class distribution.

3. Image features conditioning each input token (**Proj**): In this projected representation approach, we first perform an affine transformation with a weight matrix W, where $W \in \mathcal{R}^{c \times d}$ (c and d are dimensionality of the class distribution and dimensionality of the word vectors, respectively). This is followed by a non-linearity function to squash the resulting output. We add this representation to each source word representation. The weight matrix W is learned. This can be seen as composing each source token with the visual feature at each time step.

4 Experimental settings

We use our own implementation of a multimodal NMT approach and explore a number of variants of this model in order to understand the effects of using the classification layer instead of a lower level CNN layer as input to the NMT system.

4.1 Data

The shared task is based on the Multi30K (Elliott et al., 2016) dataset. Each image contains one English description taken from Flickr30K and professional translations into German and French. In this year's edition of the shared task, the source language is English (EN) and the target languages are German (DE) and French (FR). The dataset contains 29,000 training and 1,014 development instances: an image, a description in source language, and a description for each target language. There are two test sets:

1. An in-domain test set (**Flickr**) with 1,000 images.

2. An out-of-domain test set (**MSCOCO**) with 461 images whose captions were selected to contain ambiguous verbs.

4.2 Visual features

The primary visual feature explored in this paper is the class posterior probability estimates of ResNet-152 for 1,000 object categories (**Softmax**). As a comparison, we also extract the penultimate layer of ResNet-152 (**Pool5**).

The visual features are combined with the NMT model using the three configurations described in Section 3.3 (**InitEnc**, **InitDec**, **Proj**). We also compare our systems to a text-only baseline (Section 3.2).

4.3 NMT model

We implemented our NMT system (Section 3.2) in PyTorch. We use a single layer bidirectional LSTM based encoder-decoder model. We used ReLU as the projection non-linearity and used dropout with probability of 0.2. We used the *Adadelta* optimizer (Zeiler, 2012) with the default learning rate (0.01). The batch size was set to 20. We trained it for 50 epochs and selected the model that performs best on the validation set using *BLEU* as the metric.

We normalised punctuations, lowercased and tokenised the input text using the script provided in Moses (Koehn et al., 2007). Our experiments were performed with the vocabulary size of 6,000 English words, 6,500 French words and 8,000 German words after removing words that appeared only once in the training set (these words were replaced with $< UNK >$, as described in Section 3.2). At decoding time, we post-processed the output translations by replacing $< UNK >$ with an empty string.

5 Results and discussion

We present our results on the Flickr test dataset in Table 1, for both EN–DE and EN–FR. We observe that for the **Softmax** feature, **InitDec** consistently outperformed **InitEnc** and **Proj**. It also performed better than the text-only baseline for both

Flickr	Feature	Model	Meteor	BLEU
EN–DE	-	Baseline	43.7	24.4
	Pool5	Proj	–	–
		InitEnc	43.0	23.5
		InitDec	44.3	24.6
	Softmax	Proj	43.4	24.2
		InitEnc	42.4	23.3
		InitDec	44.5	25.0
EN–FR	-	Baseline	62.2	44.2
	Pool5	Proj	–	–
		InitEnc	61.1	43.5
		InitDec	61.0	43.4
	Softmax	Proj	61.5	43.6
		InitEnc	61.0	43.3
		InitDec	62.8	45.0

Table 1: Results on the Flickr test data, for both English–German (EN–DE) and English–French (EN–FR). **Proj** was not evaluated for **Pool5** as its performance is very poor on the development set.

MSCOCO	Feature	Model	Meteor	BLEU
EN–DE	–	Baseline	39.6	20.7
	Pool5	Proj	–	–
		InitEnc	39.1	20.4
		InitDec	39.5	20.4
	Softmax	Proj	40.0	21.0
		InitEnc	37.5	18.8
		InitDec	40.7	21.4
EN–FR	-	Baseline	57.4	37.2
	Pool5	Proj	–	–
		InitEnc	56.7	36.5
		InitDec	56.7	36.9
	Softmax	Proj	57.0	36.8
		InitEnc	55.5	35.5
		InitDec	57.3	37.2

Table 2: Results on the MSCOCO test data, for both English–German (EN–DE) and English–French (EN–FR). Again, **Proj** was not evaluated for **Pool5** as its performance was very poor on the development set.

languages. In the case of **Pool5**, **InitDec** seemed to perform slightly better than **InitEnc** for German, but both yielded similar scores for French. We also observed that by using the **Pool5** feature in the **Proj** configuration, the NMT system failed to learn any useful information with extremely low *BLEU* scores on the development set, even with an increased number of epochs. Thus we do not evaluate these on the test sets.

Table 2 displays the empirical results on the MSCOCO test dataset. Similar trends are observed here for **Softmax**: **InitDec** outperformed **Proj** and **InitEnc**. For this test set, **InitDec** outperformed the baseline for EN–DE and performed comparably to the baseline for EN–FR. Interestingly, the variant with **Pool5** as a feature did not seem to perform as well, producing slightly lower scores than the baseline on this test set. Further investigation is needed to determine the reason for this phenomenon.

Overall, we observed better results for **Softmax** compared to **Pool5** with the settings used in our submission. However, more experiments need to be performed to confirm the usefulness of the posterior probabilities for the task.

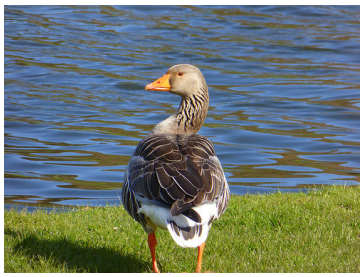
Figure 2 shows example output translations from English to German and French for the test sets, for our best performing variant **InitDec** conditioned on **Softmax** class posterior predictions. We compare the output against a text-only baseline. In the first example from the Flickr test set, **InitDec** produced an exact match against the reference for German, and an equally correct translation for French (differing only in the translation for 'bank'). In the second image from the MSCOCO test set, the German translation is much closer to the reference than the baseline. In the case of the French translation, the difference between the baseline and **InitDec** is much smaller, reflecting the quantitative results.

We conjecture that further hyperparameter search (increasing LSTM layers, dimensionality of the embeddings and hidden layers, etc.) and increasing the vocabulary size or using BPE could potentially improve the performance of our system on the task.

6 Related work

There has been interest in recent years in the task of generating image descriptions (also known as image captioning). Bernardi et al. (2016) provide a detailed discussion on various image description generation approaches that have been developed.

Currently, the two largest image description datasets are Flickr30K (Young et al., 2014) and MS COCO (Lin et al., 2014). These datasets are constructed in English and are aimed at advancing research on the generation of image descriptions in English. Recent attempts have been made to incorporate multilinguality into both these large-scale datasets, with the datasets being extended to



EN	A duck on the bank of a river
DE (**Baseline**)	eine ente an der kste eines flusses .
DE (**InitDec**)	eine ente am ufer eines flusses
DE (Reference)	eine ente am ufer eines flusses
FR (**Baseline**)	un canard sur l' eau , dans une rivière
FR (**InitDec**)	un canard sur la rive d' une rivière
FR (Reference)	un canard sur la berge d' une rivière

EN	A tennis player is moving to the side and is gripping his racquet with both hands.
DE (**Baseline**)	ein tennisspieler fhrt zur seite und greift nach seinem schlger .
DE (**InitDec**)	ein tennisspieler bewegt sich zur seite , whrend sein schlger mit beiden hnden .
DE (Reference)	ein tennisspieler bewegt sich zur seite und hlt den schlger mit beiden hnden .
FR (**Baseline**)	un joueur de tennis se déplaçant de côte et sa raquette avec les deux mains .
FR (**InitDec**)	un joueur de tennis se déplaçant côté et se met sa raquette avec les deux mains .
FR (Reference)	un joueur de tennis se déplace sur le cté et tient sa raquette avec ses deux mains .

Figure 2: Example output translations from English to German (DE) and French (FR), for the Flickr test set (top) and the MSCOCO test set (bottom). We show the results of **InitDec** using **Softmax** as the visual feature.

other languages such as German and Japanese (Elliott et al., 2016; Hitschler et al., 2016; Miyazaki and Shimizu, 2016; Yoshikawa et al., 2017).

The first known attempt at using NMT for machine translation of image descriptions is by Elliott et al. (2015), who conditioned an NMT system with a CNN image embedding (the penultimate layer of VGG-16 (Simonyan and Zisserman, 2014)) at the beginning of either the encoder or the decoder. The WMT16 shared task on Multimodal Machine Translation (Specia et al., 2016) has further encouraged research in this area. At the time, phrase-based SMT systems (Shah et al., 2016; Libovický et al., 2016; Hitschler et al., 2016) performed better than NMT systems (Calixto et al., 2016; Huang et al., 2016; Caglayan et al., 2016). Participants used either the penultimate fully connected layer or a convolutional layer of a CNN as image representation, with the exception of Shah et al. (2016) who used the classification output of VGG-16 as features to a phrase-based SMT system. In all cases, image information were found to provide only marginal improvements.

7 Conclusions and future work

We presented our approach that uses predicted class distribution as image features for the task of multimodal machine translation. We described three configurations for incorporating the visual representation and observed that the three methods perform differently. For our submission with the settings described in the paper, using ResNet-152's class posterior probability distribution seems to result in better scores than using

the same network's pool5 features. Future experiments will aim at dissecting the type of information the image features are adding to the NMT and understand deeply the contribution of predicted class based representations.

Acknowledgments

This work was supported by the MultiMT project (EU H2020 ERC Starting Grant No. 678017).

References

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research* 55:409–442.

Ozan Caglayan, Walid Aransa, Yaxing Wang, Marc Masana, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, and Joost van de Weijer. 2016. Does multimodality help human and machine for translation and image captioning? In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 627–633.

Iacer Calixto, Desmond Elliott, and Stella Frank. 2016. DCU-UvA Multimodal MT system report. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 634–638.

D. Elliott, S. Frank, K. Sima'an, and L. Specia. 2016. Multi30k: Multilingual English-German image descriptions. In *5th Workshop on Vision and Language*. pages 70–74.

Desmond Elliott, Stella Frank, and Eva Hasler. 2015. Multi-language image description with neural sequence models. *CoRR* abs/1510.04709.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR* abs/1512.03385. http://arxiv.org/abs/1512.03385.

Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. 2016. Multimodal pivots for image caption translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 2399–2409.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *First Conference on Machine Translation*. pages 639–645.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL System Demonstration Session*. pages 177–180.

Jindřich Libovický, Jindřich Helcl, Marek Tlustý, Ondřej Bojar, and Pavel Pecina. 2016. Cuni system for wmt16 automatic post-editing and multimodal translation tasks. In *First Conference on Machine Translation*. pages 646–654.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*. Springer, pages 740–755.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* .

Takashi Miyazaki and Nobuyuki Shimizu. 2016. Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 1780–1790.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115(3):211–252. https://doi.org/10.1007/s11263-015-0816-y.

Kashif Shah, Josiah Wang, and Lucia Specia. 2016. SHEF-Multimodal: Grounding machine translation on images. In *First Conference on Machine Translation*. pages 660–665.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR* abs/1409.1556.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *First Conference on Machine Translation*. pages 543–553.

Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton van den Hengel. 2016. What Value Do Explicit High Level Concepts Have in Vision to Language Problems? In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. 2017. STAIR captions: Constructing a

large-scale japanese image caption dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image Captioning With Semantic Attention. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions 2:67–78.

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701* .