DCU-UvA Multimodal MT System Report

Iacer Calixto
ADAPT Centre
School of Computing
Dublin City University
Dublin, Ireland
iacer.calixto@adaptcentre.ie

Desmond Elliott
ILLC
University of Amsterdam
Science Park 107
1098 XG Amsterdam
d.elliott@uva.nl

Stella Frank
ILLC
University of Amsterdam
Science Park 107
1098 XG Amsterdam
s.c.frank@uva.nl

Abstract

We present a doubly-attentive multimodal machine translation model. Our model learns to attend to source language and spatial-preserving $\text{CONV}_{5,4}$ visual features as separate attention mechanisms in a neural translation model. In image description translation experiments (Task 1), we find an improvement of 2.3 Meteor points compared to initialising the hidden state of the decoder with only the FC_7 features and 2.9 Meteor points compared to a text-only neural machine translation baseline, confirming the useful nature of attending to the $\text{CONV}_{5,4}$ features.

1 Introduction

Our system learns to translate image descriptions using both the source language descriptions and the images. We integrate an attention-based neural network for machine translation and image description in a unified model, in which two separate attention mechanisms operate over the language and visual modalities. We believe that this is a principled approach to learning which source words and which areas of the image to attend to when generating words in the target description.

We are inspired by recent successes in using attentive models in both neural machine translation (NMT) and neural image description. Originally, in non-attentive NMT models, the entire source sentence is encoded into a single vector which is in turn used by the decoder to generate a translation (Cho et al., 2014; Sutskever et al., 2014). In a similar vein, image description models can use a vector encoding the image as input for the description generation process (Vinyals et al., 2015; Mao et al., 2015, *inter-alia*).

Bahdanau et al. (2014) first proposed a NMT model with an attention mechanism over the source sentence. Their model is trained so that the decoder learns to attend to words in the source sentence when translating each token in the target sentence. Xu et al. (2015) introduced a similar attention-based neural image description model. In this case, the attention mechanism learns which parts of the image to attend to while generating words in the description.

When translating image descriptions, given both the source description and the source image (i.e., the setting for Task 1), we believe that both modalities can provide cues for generating the target language description. The source description provides the content for translation, but in cases where this may be ambiguous, the image features can provide contextual disambiguation. The system we propose is a first step towards integrating both modalities using attention mechanisms.

Previous work has demonstrated the plausibility of multilingual multimodal natural language processing. Elliott et al. (2015) showed how to generate descriptions of images in English and German by learning and transferring features between independent neural image description models. In comparison, our approach is a single end-to-end model over the source and target languages with attention mechanisms over both the source language and the visual features.

2 Model Description

We represent the source language with a bidirectional recurrent neural network (RNN) with a gated recurrent unit (GRU) that computes, for each word, forward and backward source annotation vectors $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$. The final source annotation vector for a word h_i is the concatenation of both $[\overrightarrow{h_i}; \overleftarrow{h_i}]$.

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 634–638,
Berlin, Germany, August 11-12, 2016.

We use the visual features released by the shared task organisers, extracted from the pre-trained VGG-19 convolutional neural network (CNN) (Simonyan and Zisserman, 2015).

The organisers release two types of visual features according to the layer they were extracted from: FC_7 features are extracted from the final fully-connected layer (FC_7), which encode information about the entire image in a 4096-dimensional feature vector; and $CONV_{5,4}$ features, extracted from the final convolutional layer ($CONV_{5,4}$), namely a 196 x 512 dimensional matrix where each row (i.e., a 512D vector) represents features from a specific spatial 'patch' and therefore encodes information about that specific 'patch' (i.e., area) of the image.

2.1 FC_7-initialised model

In this model, we use visual features extracted from the final fully-connected FC_7 layer from the pre-trained VGG-19 CNN. These features represent an abstract summary of the entire image and crucially are not spatially aware, unlike the $CONV_{5,4}$ features we use in the subsequent double-attention model. We integrate the FC_7 features into the initial state of the decoder.

We first affine-transform the 4096D FC_7 image feature vector i into the source language bidirectional RNN hidden states dimensionality, where the affine transformation parameters (W_I, b_I) are trained jointly with the model:

$$i_{\text{proj}} = i \cdot W_I + b_I. \tag{1}$$

We then simply sum these projected image features i_{proj} with the first source language context vector h_1, obtained by the encoder bidirectional RNN, and use the resulting vector as input to a feed-forward neural network f_{init} used to initialise the decoder hidden state:

$$s_0 = f_{\text{init}}(h_1 + i_{\text{proj}}) \tag{2}$$

2.2 Doubly-attentive model

The goal of the doubly-attentive model is to integrate separate attention mechanisms over the source language words and visual features in a single decoder. Similarly to the FC_7 model, we represent the source language using a bi-directional RNN with GRUs. We use visual features extracted from the $CONV_{5,4}$ layer of the VGG-19 CNN alongside the FC_7 features. The $CONV_{5,4}$ features consist of a 196 x 512 dimensional matrix, where each row represents features from a specific spatial 'patch'. Analogous to how the attention mechanism for the source language can focus on specific words or phrases in the source description, the image attention mechanism can focus on specific parts of the image (Xu et al., 2015).

Our doubly-attentive decoder is conditioned on the source sentence and the image via the two separate attention mechanisms, as well as the previous hidden state of the decoder and the previously emitted word. Therefore, in computing the decoder hidden state s_t at time step t, the decoder has access to the following information:

- i_t – the image context vector for the current time step obtained via attention over the image representation;
- c_t – the source language context vector for the current time step obtained via attention over the source sentence representation;
- s_{t-1} – the decoder's previous hidden state;
- y_{t-1} – the target word emitted by the decoder in the previous time step.

Figure 1 illustrates the computation of the decoder hidden state s_t according to our *doubly-attentive* model.

2.3 Source sequence context vector

To compute the time-dependent source sentence context vector, we follow Bahdanau et al. (2014) and use a single-layer feed-forward network f^s_{score} for computing an *expected alignment* $e^s_{t,i}$ between each source annotation vector h_i — computed as the concatenation of forward and backward source annotation vectors $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$ — and the target word to be emitted at the current time step t.

$$e^s_{t,i} = f^s_{\text{score}}(h_i, s_{t-1}, y_{t-1}), \tag{3}$$

where f^s_{score} uses all source annotation vectors $\mathbf{h}$, the decoder's previous hidden state s_{t-1} and the previously emitted word y_{t-1} in computing the expected alignments for the target word at current time step t. In Equation 4, these alignments are then normalised and converted into probabilities.

$$\alpha_{t,i} = \frac{\exp\left(e^s_{t,i}\right)}{\sum_{j=1}^{N} \exp\left(e^s_{t,j}\right)}, \tag{4}$$

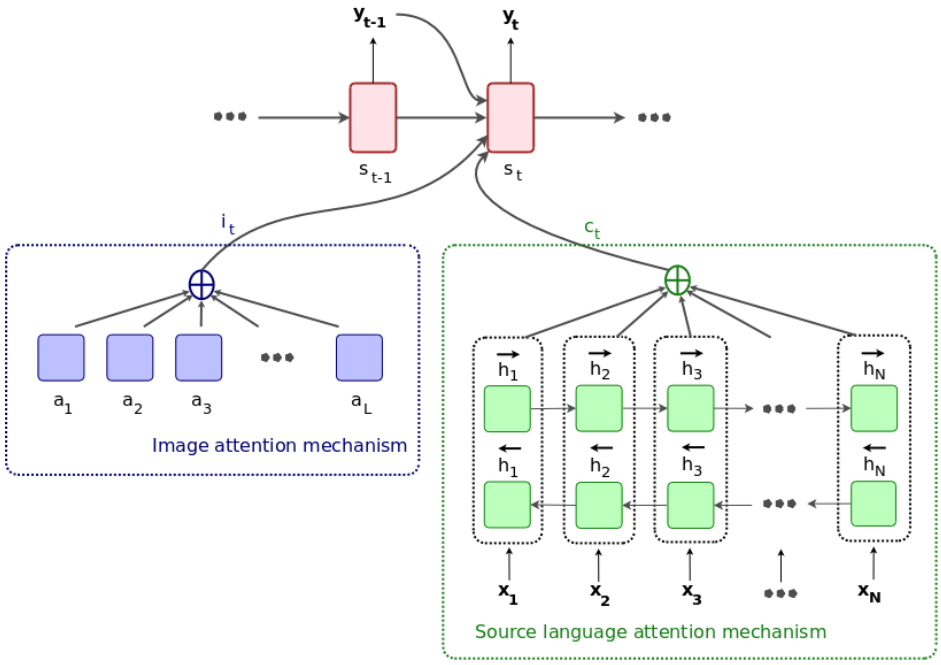

Figure 1: A doubly-attentive decoder learns to independently attend to image patches and source language words when generating translations.

where $\alpha_{t,i}$ are weights representing the attention over the source annotation vectors. The final time-dependent source context vector c_t is a weighted sum over the source annotation vectors, where each vector is weighted by the attention weight $\alpha_{t,i}$:

$$c_t = \sum_{i=1}^{N} \alpha_{t,i} h_i. \tag{5}$$

2.4 Image context vector

The time-dependent image context vectors are based on the "soft" visual attention mechanism (Xu et al., 2015). As outlined above, the image annotation vectors are the features extracted from $\text{CONV}_{5,4}$ layer, resulting in 196 vectors (each corresponding to one of the 14×14 patches in the image) of 512 dimensions each. These annotation vectors are denoted a_l (with $l = 1 \ldots 196$) and are used analogously to the hidden states h_i of the source sentence encoder.

The expected alignments $e^i_{t,l}$ over the image features are computed by a single layer feed-forward network f^i_{score}:

$$e^i_{t,l} = f^i_{\text{score}}(a_l, s_{t-1}, y_{t-1}), \tag{6}$$

where f^i_{score} uses all image annotation vectors $\boldsymbol{a}$, the decoder previous hidden state s_{t-1} and the previously emitted word y_{t-1} in computing the expected image–target word alignments at current time step t. In Equation 7 these expected alignments are further normalised and converted into probabilities, as in the source context vector.

$$\alpha^i_{t,l} = \frac{\exp\left(e^i_{t,l}\right)}{\sum_{j=1}^{L} \exp\left(e^i_{t,j}\right)}, \tag{7}$$

where $\alpha^i_{t,i}$ are the model's *image attention weights*. A time-dependent image context vector i_t is then computed by using these attention weights.

$$i_t = \sum_{l=1}^{N} \alpha^i_{t,l} a_l. \tag{8}$$

Ideally, this image context vector i_t captures the image patches that are relevant to the current state of the decoder and for generating the next word.

3 Experiments

We report results for Task 1, which uses the translated data in the Multi30K corpus (Elliott et al., 2016). English and German descriptions in the Multi30K were normalised and tokenized, and compounds in German descriptions were further split in a pre-processing step[1].

[1]We use the scripts in the Moses SMT Toolkit to normalise, tokenize and split compounds (Koehn et al., 2007).

	Meteor
Moses	52.3
$\text{CONV}_{5,4}$-Multimodal NMT	46.4
FC_7-Multimodal NMT	44.1
Text-only Attention NMT	43.5
Elliott et al. (2015)	24.7

Table 1: Results for our models on Task 1. We find that attending over the source language and $\text{CONV}_{5,4}$ visual features is better than not using image features (text-only, attentive NMT model) and also just initialising an attention-based decoder with FC_7 features.

Throughout, we parameterise our models using 300D word embeddings, 1000D hidden states, and 1000D context vectors; the source and target languages are estimated over the entire vocabularies. Our non-recurrent matrices are initialised by sampling from a Gaussian distribution ($\mu = 0, \sigma = 0.01$), recurrent matrices are orthogonal and bias vectors are all initialised to zero. We apply dropout on the source language words (encoder) and before the readout operation (decoder) with probability of 0.3 and apply no other regularisation. We apply early stopping for model selection based on Meteor scores (Denkowski and Lavie, 2014), and if it has not increased for 20 epochs on a validation set, training is halted. The models are trained using the Adadelta optimizer (Zeiler, 2012) with an initial learning rate of 0.005.

In Table 1 we compare our models — $\text{CONV}_{5,4}$ and FC_7-Multimodal NMT — against a text-only, attention-based NMT baseline, the Moses translation baseline (Koehn et al., 2007) and the multilingual image description baseline (Elliott et al., 2015). First, it is clear that the Moses SMT baseline is very strong, given that it is only trained over the parallel text without any visual information. Our models are unable to match the performance of Moses, however, we do see a substantial increase of 20-22 Meteor points compared to the independent image description models (Elliott et al., 2015). The magnitude of the difference shows the importance of learning the source and target language representations in a single joint model.

We also observe improvements in Meteor when we compare our double-attentive $\text{CONV}_{5,4}$ model against the FC_7 initialised model (2.3 points) and against a text-only NMT model (2.9 points).

Our results indicate that incorporating image features in multimodal models helps, as compared to our text-only NMT baseline. Even though our neural models — both text-only and multimodal models — fall short of the SMT baseline performance, we believe that the use of neural architectures for this task is more principled, due to the ability to incorporate images and translations in one network that is trained end-to-end.

4 Conclusions and Future Work

We present a model which incorporates multiple multimodal attention mechanisms into a neural machine translation decoder. Source language and visual attention mechanisms have been well-studied in the recent literature, but our results indicate that multimodal attention appears to be more complex than simply combining two independent attention mechanisms. In particular, we hoped to find a greater improvement from adding visual features, relative to text-only models. However, the Multi30k dataset is relatively small, with a small vocabulary and simple syntactic structures (Elliott et al., 2016). Whereas SMT models can be trained effectively on such datasets, neural models usually perform best when a large amount of data is available. We believe that as the amount of data in multimodal translation datasets increase, neural models will become more competitive.

In future work we plan to study why the source language attention mechanism contributes more to the model than the visual attention. We believe that using the source language context vector c_t may help when computing the image context vector i_t. We also plan to investigate other attention mechanisms, for instance the "hard" attention as proposed by Xu et al. (2015). Soft attention may be too diffuse in this setting, especially over the large set of image context vectors.

Acknowledgements

This research is partially supported by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre (`www.adaptcentre.ie`) at Dublin City University. DE was supported by the NWO Vici grant nr. 277-89-002. SF was supported by European Unions Horizon 2020 research and innovation programme under grant agreement nr. 645452.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

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

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. *CoRR*, abs/1605.00459.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 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*, pages 177–180. Association for Computational Linguistics.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-RNN). In *ICLR*.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pages 3104–3112.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*.

Matthew D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701.