DCU System Report on the WMT 2017 Multi-modal Machine Translation Task

Iacer Calixto and **Koel Dutta Chowdhury** and **Qun Liu**
{iacer.calixto,koel.chowdhury,qun.liu}@adaptcentre.ie

Abstract

We report experiments with multi-modal neural machine translation models that incorporate global visual features in different parts of the encoder and decoder, and use the VGG19 network to extract features for all images. In our experiments, we explore both different strategies to include global image features and also how ensembling different models at inference time impact translations. Our submissions ranked 3rd best for translating from English into French, always improving considerably over an neural machine translation baseline across all language pair evaluated, e.g. an increase of 7.0–9.2 METEOR points.

1 Introduction

In this paper we report on our application of three different multi-modal neural machine translation (NMT) systems to translate image descriptions. We use encoder–decoder attentive multi-modal NMT models where each training example consists of one source variable-length sequence, one image, and one target variable-length sequence, and a model is trained to translate sequences in the source language into corresponding sequences in the target language while taking the image into consideration. We use the three models introduced in Calixto et al. (2017b), which integrate *global image features* extracted using a pre-trained convolutional neural network into 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.

We are inspired by the recent success of multi-modal NMT models applied to the translation of image descriptions (Huang et al., 2016; Calixto et al., 2017a). Huang et al. (2016) incorporate global visual features into NMT with some success, and Calixto et al. (2017a) propose to use local visual features instead, achieving better results. We follow Calixto et al. (2017b) and investigate whether we can achieve better results while still using *global visual features*, which are considerably smaller and simpler to integrate when compared to local features.

We expect that, by integrating visual information when translating image descriptions, we are able to exploit valuable information from both modalities when generating the target description, effectively grounding machine translation (Glenberg and Robertson, 2000).

2 Model Description

The models used in our experiments can be viewed as expansions of the attentive NMT framework introduced by Bahdanau et al. (2015) with the addition of a visual component that incorporates visual features from images. A bi-directional recurrent neural network (RNN) with gated recurrent unit (GRU) (Cho et al., 2014) is used as the encoder. The final annotation vector for a given source position i is the concatenation of forward and backward RNN hidden states, $\boldsymbol{h}_i = [\overrightarrow{\boldsymbol{h}_i}; \overleftarrow{\boldsymbol{h}_i}]$.

We use the publicly available pre-trained convolution neural network VGG19[1] of Simonyan and Zisserman (2014) to extract global image feature vectors for all images. These features are the 4096D activations of the penultimate fully-connected layer FC7, henceforth referred to as $\boldsymbol{q}$.

We now describe the three multi-modal NMT models used in our experiments. For a detailed explanation about these models, see Calixto et al. (2017b).

[1] http://www.robots.ox.ac.uk/~vgg/research/very_deep/

Proceedings of the Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 440–444
Copenhagen, Denmark, September 711, 2017.

2.1 IMG$_{2W}$: Image as source words

In model IMG$_{2W}$, the image features are used as the first and last words of the source sentence, and the source-language attention model learns when to attend to the image representations. Specifically, given the global image feature vector $\boldsymbol{q} \in \mathbb{R}^{4096}$:

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

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 directly use $\boldsymbol{d}$ as the first and last words of the source sentence. In other words, given the word embeddings for a source sentence $X = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N)$, we concatenate the transformed image vector $\boldsymbol{d}$ to it, i.e. $X = (\boldsymbol{d}, \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N, \boldsymbol{d})$, and apply the forward and backward encoder RNN. By including images into the encoder in model IMG$_{2W}$, our intuition is that *(i)* by including the image as the *first word*, we propagate image features into the source sentence vector representations when applying the forward RNN to produce vectors $\overrightarrow{\boldsymbol{h}_i}$, 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 to produce vectors $\overleftarrow{\boldsymbol{h}_i}$.

2.2 IMG$_E$: Image for encoder initialisation

In the original attention-based NMT model of Bahdanau et al. (2015), 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 and the backward RNN, respectively.

Similarly to what we do for model IMG$_{2W}$ described in Section 2.1, given a global image feature vector $\boldsymbol{q} \in \mathbb{R}^{4096}$, we compute a vector $\boldsymbol{d}$ using Equation (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 hidden states of the source language encoder.

The feed-forward networks used to initialise the encoder hidden state are computed as in (2):

$$\begin{aligned} \overleftarrow{\boldsymbol{h}}_{\text{init}} &= \tanh\left(\boldsymbol{W}_f \boldsymbol{d} + \boldsymbol{b}_f\right), \\ \overrightarrow{\boldsymbol{h}}_{\text{init}} &= \tanh\left(\boldsymbol{W}_b \boldsymbol{d} + \boldsymbol{b}_b\right), \end{aligned} \quad (2)$$

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. $\overrightarrow{\boldsymbol{h}}_{\text{init}}$ and $\overleftarrow{\boldsymbol{h}}_{\text{init}}$ are directly used as the forward and backward RNN initial hidden states, respectively.

2.3 IMG$_D$: Image for decoder initialisation

To incorporate an image into the decoder, we introduce a new single-layer feed-forward neural network. Originally, the decoder initial hidden state is computed using a summary of the encoder hidden states. This is often the concatenation of the last hidden states of the encoder forward RNN and backward RNN, respectively $\overrightarrow{\boldsymbol{h}_N}$ and $\overleftarrow{\boldsymbol{h}}_1$, or the mean of the source-language annotation vectors $\boldsymbol{h}_i$.

We propose to include the image features as additional input to initialise the decoder's hidden state, as described in (3):

$$\boldsymbol{s}_0 = \tanh\left(\boldsymbol{W}_{di}[\overleftarrow{\boldsymbol{h}}_1; \overrightarrow{\boldsymbol{h}_N}] + \boldsymbol{W}_m \boldsymbol{d} + \boldsymbol{b}_{di}\right), \quad (3)$$

where $\boldsymbol{s}_0$ is the decoder initial hidden state, $\boldsymbol{W}_m$ 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 learned model parameters.

Once again we compute $\boldsymbol{d}$ by applying Equation (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.

3 Experiments

We report results for Task 1, specifically when translating from English into German (en–de) and French (en–fr). We conducted experiments on the constrained version of the shared task, which means that the only training data we used is the data released by the shared task organisers, i.e. the *translated Multi30k* (M30k$_T$) data set (Elliott et al., 2016) with the additional French image descriptions, included for the 2017 run of the shared task.

Our encoder is a bi-directional RNN with GRU, one 1024D single-layer forward RNN and one 1024D single-layer backward RNN. Throughout, we parameterise our models using 620D source and target word embeddings, and both are trained

jointly with our model. All non-recurrent matrices are initialised by sampling from a Gaussian distribution ($\mu = 0, \sigma = 0.01$), recurrent matrices are random orthogonal and bias vectors are all initialised to $\vec{0}$. We apply dropout (Srivastava et al., 2014) with a probability of 0.3 in source and target word embeddings, 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.

The translated Multi30k training and validation sets contain 29k and 1014 images respectively, each accompanied by a sentence triple, the original English sentence and its gold-standard translations into German and into French.

We use the scripts in the Moses SMT Toolkit (Koehn et al., 2007) to normalise, lowercase and tokenize English, German and French descriptions and we also convert space-separated tokens into subwords (Sennrich et al., 2016). The subword models are trained jointly for English–German descriptions and separately for English–French descriptions using the English-German and English-French WMT 2015 data (Bojar et al., 2015). English–German models have a final vocabulary of 74K English and 81K German subword tokens, and English–French models 82K English and 82K French subword tokens. If sentences in English, German or French are longer than 80 tokens, they are discarded.

Finally, we use the 29K entries in the M30k$_T$ training set for training our models, and the 1,014 entries in the M30k$_T$ development set for model selection, early stopping the training procedure in case the model stops improving BLEU scores on this development set. We evaluate our English–German models on three held-out test sets, the Multi30k 2016/2017 and the MSCOCO 2017 test sets, and our English–French models on the Multi30k 2017 and the MSCOCO 2017 test sets.

We evaluate translation quality quantitatively in terms of BLEU4 (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), and TER (Snover et al., 2006).

Multi30k 2017

Lang.	Model	BLEU4 ↑	METEOR↑	TER↓
en–de	NMT baseline	19.3	41.9	72.2
en–de	Ensemble	**29.8** (↑ 10.3)	**50.5** (↑ 8.6)	**52.3** (↓ 19.9)
en–fr	NMT baseline	44.3	63.1	39.6
en–fr	Ensemble	**54.1** (↑ 9.8)	**70.1** (↑ 7.0)	**30.0** (↓ 9.6)

Table 1: Results for the M30k$_T$ 2017 English–German and English–French test sets. All models are trained on the original M30k$_T$ training data. Our ensemble uses four multi-modal models, all independently trained: two models IMG$_D$, one model IMG$_E$, and one model IMG$_{2W}$.

MSCOCO 2017

Lang.	Model	BLEU4 ↑	METEOR↑	TER↓
en–de	NMT baseline	18.7	37.6	66.1
en–de	Ensemble	**26.4** (↑ 7.7)	**46.8** (↑ 9.2)	**54.5** (↓ 11.6)
en–fr	NMT baseline	35.1	55.8	45.8
en–fr	Ensemble	**44.5** (↑ 9.4)	**64.1** (↑ 8.3)	**35.2** (↓ 10.6)

Table 2: Results for the MSCOCO 2017 English–German and English–French test sets. All models are trained on the original M30k$_T$ training data. Ensemble uses four multi-modal models, all trained independently: two models IMG$_D$, one model IMG$_E$, and one model IMG$_{2W}$.

3.1 Results

In Table 1, we show results when translating the Multi30k 2017 test sets. Models are trained on the original M30k$_T$ training data only. The NMT baseline is the attention-based NMT model of Bahdanau et al. (2015) and its results are the ones reported by the shared task organisers. When compared to other submissions of the multi-modal MT task under the constrained data regime, our models ranked sixth best when translating the English–German Multi30k 2017, and fourth best when translating the English–German MSCOCO 2017 test sets. When translating both the Multi30k 2017 and the MSCOCO 2017 English–French test sets, our models are ranked third best, scoring only 1–2 points (BLEU, METEOR) less than the best system.

In Table 2, we show results when translating the MSCOCO 2017 English–German and English–French test sets. Again, all models are trained on the original M30k$_T$ training data only. When compared to other submissions of the multi-modal MT task under the constrained data regime, our submission ranked fourth best for the English–German and third best for the English–French lan-

Multi30k 2016 (English→German)				
	Ensemble?	**BLEU4** ↑	**METEOR**↑	**TER**↓
$NMT_{SRC+IMG}$[1]	×	39.0	56.8	40.6
IMG_D	×	37.3	55.1	42.8
IMG_D + IMG_E	✓	40.1 (↑ 1.1)	58.5 (↑ 1.7)	40.7 (↑ 0.1)
IMG_D + IMG_E + IMG_{2W}	✓	41.0 (↑ 2.0)	58.9 (↑ 2.1)	39.7 (↓ 0.9)
IMG_D + IMG_E + IMG_{2W} + IMG_D	✓	**41.3** (↑ **2.3**)	**59.2** (↑ **2.4**)	**39.5** (↓ **1.1**)

[1] This model is pre-trained on the English–German WMT 2015 (Bojar et al., 2015), consisting of ∼4.3M sentence pairs.

Table 3: Results for the best model of Calixto et al. (2017a), which is pre-trained on the English–German WMT 2015 (Bojar et al., 2015), and different combinations of multi-modal models, all trained on the original $M30k_T$ training data only, evaluated on the $M30k_T$ 2016 test set.

guage pair, scoring only 1 to 1.5 points less than the best system. These are promising results, specially taking into consideration that we are using global image features, which are smaller and simpler than local features (used in Calixto et al. (2017a)).

Ensemble decoding We now report on how can ensemble decoding be used to improve multi-modal MT. In Table 3, we show results when translating the Multi30k 2016's test set. We ensembled different models by starting with one of Calixto et al. (2017b)'s best performing multi-modal models on this data set, IMG_D, and by adding new models to the ensemble one by one, until we reach a maximum of four independent models, all of which are trained separately and on the original $M30k_T$ training data only. We also report results for the best model of Calixto et al. (2017a), which is pre-trained on the English–German WMT 2015 (Bojar et al., 2015) and uses local visual features extracted with the ResNet-50 network (He et al., 2015).

We first note that adding more models to the ensemble seems to always improve translations by a large margin (∼ 3 BLEU/METEOR points). Adding model IMG_{2W} to the ensemble already consisting of models IMG_E and IMG_D still improves translations, according to all metrics evaluated. This is an interesting result, since compared to these other two multi-modal models, model IMG_{2W} performs the worst according to BLEU, METEOR and chrF3 (see Calixto et al. (2017b)). Our best results are obtained with an ensemble of four different multi-modal models.

By using an ensemble of four different multi-modal NMT models trained on the translated Multi30k training data, we were able to obtain translations comparable to or even better than those obtained with the strong multi-modal NMT model of Calixto et al. (2017a), which is pre-trained on large amounts of WMT data and uses local image features.

4 Conclusions and Future work

In this work, we evaluated multi-modal NMT models which integrate *global image features* into both the encoder and the decoder. We experimented with ensembling different multi-modal NMT models introduced in Calixto et al. (2017b), and results show that these models can generate translations that compare favourably to multi-modal models that use local image features. We observe consistent improvements over a text-only NMT baseline trained on the same data, and these are typically very large (e.g., 7.0–9.2 METEOR points across language pairs and test sets). In future work we plan to study how to generalise these models to other multi-modal natural language processing tasks, e.g. visual question answering.

Acknowledgments

This project 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).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations, ICLR 2015*. San Diego, California.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal, pages 1–46. http://aclweb.org/anthology/W15-3001.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017a. Doubly-Attentive Decoder for Multi-modal Neural Machine Translation. In *Proceedings of the 55th Conference of the Association for Computational Linguistics: Volume 1, Long Papers*. Vancouver, Canada (Paper Accepted).

Iacer Calixto, Qun Liu, and Nick Campbell. 2017b. 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 (Paper Accepted).

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. 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)*. Doha, Qatar, pages 1724–1734. http://www.aclweb.org/anthology/D14-1179.

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

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German Image Descriptions. In *Proceedings of the 5th Workshop on Vision and Language, VL@ACL 2016*. Berlin, Germany. http://aclweb.org/anthology/W/W16/W16-3210.pdf.

Yarin Gal and Zoubin Ghahramani. 2016. A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. In *Advances in Neural Information Processing Systems, NIPS*. Barcelona, Spain, pages 1019–1027. http://papers.nips.cc/paper/6241-a-theoretically-grounded-application-of-dropout-in-recurrent-neural-networks.pdf.

A. Glenberg and D. Robertson. 2000. Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language* http://psych.wisc.edu/glenberg/.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385* .

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based Multimodal Neural Machine Translation. In *Proceedings of the First Conference on Machine Translation*. Berlin, Germany, pages 639–645. http://www.aclweb.org/anthology/W/W16/W16-2360.

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 *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics, Prague, Czech Republic, ACL '07, pages 177–180. http://dl.acm.org/citation.cfm?id=1557769.1557821.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Philadelphia, Pennsylvania, ACL '02, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany, pages 1715–1725. http://www.aclweb.org/anthology/P16-1162.

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

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas, AMTA*. Cambridge, MA, USA, pages 223–231.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15:1929–1958. http://jmlr.org/papers/v15/srivastava14a.html.