# OSU Multimodal Machine Translation System Report

**Mingbo Ma, Dapeng Li, Kai Zhao[†] and Liang Huang**
Department of EECS
Oregon State University
Corvallis, OR 97331, USA
`{mam, lidap, zhaok, liang.huang}@oregonstate.edu`

## Abstract

This paper describes Oregon State University's submissions to the shared WMT'17 task "multimodal translation task I". In this task, all the sentence pairs are image captions in different languages. The key difference between this task and conventional machine translation is that we have corresponding images as additional information for each sentence pair. In this paper, we introduce a simple but effective system which takes an image shared between different languages, feeding it into the both encoding and decoding side. We report our system's performance for English-French and English-German with Flickr30K (in-domain) and MSCOCO (out-of-domain) datasets. Our system achieves the best performance in TER for English-German for MSCOCO dataset.

## 1 Introduction

Natural language generation (NLG) is one of the most important tasks in natural language processing (NLP). It can be applied to a lot of interesting applications such like machine translation, image captioning, question answering. In recent years, Recurrent Neural Networks (RNNs) based approaches have shown promising performance in generating more fluent and meaningful sentences compared with conventional models such as rule-based model (Mirkovic et al., 2011), corpus-based n-gram models (Wen et al., 2015) and trainable generators (Stent et al., 2004).

More recently, attention-based encoder-decoder models (Bahdanau et al., 2014) have been proposed to provide the decoder more accurate alignments to generate more relevant words. The remarkable ability of attention mechanisms quickly update the state-of-the-art performance on variety of NLG tasks, such as machine translation (Luong et al., 2015), image captioning (Xu et al., 2015; Yang et al., 2016), and text summarization (Rush et al., 2015; Nallapati et al., 2016).

However, for multimodal translation (Elliott et al., 2015), where we translate a caption from one language into another given a corresponding image, we need to design a new model since the decoder needs to consider both language and images at the same time.

This paper describes our participation in the WMT 2017 multimodal task 1. Our model feeds the image information to both the encoder and decoder, to ground their hidden representation within the same context of image during training. In this way, during testing time, the decoder would generate more relevant words given the context of both source sentence and image.

## 2 Model Description

For the neural-based machine translation model, the encoder needs to map sequence of word embeddings from the source side into another representation of the entire sequence using recurrent networks. Then, in the second stage, decoder generates one word at a time with considering global (sentence representation) and local information (weighted context) from source side. For simplicity, our proposed model is based on the attention-based encoder-decoder framework in (Luong et al., 2015), ref-

---

[†] Current address: Google Inc., 111 8th Avenue, New York, New York, USA.

ereed as "Global attention".

On the other hand, for the early work of neural-basic caption generation models (Vinyals et al., 2015), the convolutional neural networks (CNN) generate the image features which feed into the decoder directly for generating the description.

The first stage of the above two tasks both map the temporal and spatial information into a fixed dimensional vector which makes it feasible to utilize both information at the same time.

Fig. 1 shows the basic idea of our proposed model (OSU1). The red character **I** represents the image feature that is generated from CNN. In our case, we directly use the image features that are provided by WMT, and these features are generated by residual networks (He et al., 2016).

The encoder (blue boxes) in Fig. 1 takes the image feature as initialization for generating each hidden representation. This process is very similar to neural-basic caption generation (Vinyals et al., 2015) which grounds each word's hidden representation to the context given by the image. On the decoder side (green boxes in Fig. 1), we not only let each decoded word align to source words by global attention but also feed the image feature as initialization to the decoder.
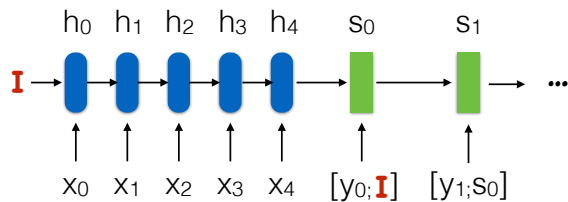


**Figure 1:** The image information is feed to both encoder and decoder for initialization. I (in red) represents the image feature that are generated by CNN.

## 3 Experiments

### 3.1 Datasets

In our experiments, we use two datasets Flickr30K (Elliott et al., 2016) and MSCOCO (Lin et al., 2014) which are provided by the WMT organization. For both datasets, there are triples that contains English as source sentence, its German and French human translations and corresponding image. The system is only trained on

Flickr30K datasets but are also tested on MSCOCO besides Flickr30K. MSCOCO datasets are considered out-of-domain (OOD) testing while Flickr30K dataset are considered in-domain testing. The datasets' statics is shown in Table 1

| Datasets | Train | Dev | Test | OOD ? |
|----------|-------|-----|------|-------|
| Flickr30K | $29,000$ | $1,014$ | $1,000$ | No |
| MSCOCO | - | - | 461 | Yes |

**Table 1:** Summary of datasets statistics.

### 3.2 Training details

For preprocessing, we convert all of the sentences to lower case, normalize the punctuation, and do the tokenization. For simplicity, our vocabulary keeps all the words that show in training set. For image representation, we use ResNet (He et al., 2016) generated image features which are provided by the WMT organization. In our experiments, we only use average pooled features.

Our implementation is adapted from on Pytorch-based OpenNMT (Klein et al., 2017). We use two layered bi-LSTM (Sutskever et al., 2014) on the source side as encoder. Our batch size is 64, with SGD optimization and a learning rate at 1. For English to German, the dropout rate is 0.6, and for English to French, the dropout rate is 0.4. These two parameters are selected by observing the performance on development set. Our word embeddings are randomly initialized with 500 dimensions. The source side vocabulary is 10,214 and the target side vocabulary is 18,726 for German and 11,222 for French.

### 3.3 Beam search with length reward

During test time, beam search is widely used to improve the output text quality by giving the decoder more options to generate the next possible word. However, different from traditional beam search in phrase-based MT where all hypotheses know the number of steps to finish the generation, while in neural-based generation, there is no information about what is the most ideal number of steps to finish the decoding. The above issue also leads to another problem that the beam search in neural-based MT prefers shorter sequences due to

probability-based scores for evaluating different candidates. In this paper, we use Optimal Beam Search (Huang et al., 2017) (OBS) during decoding time. OBS uses bounded length reward mechanism which allows a modified version of our beam search algorithm to remain optimal.

Figure 2 and Figure 3 show the BLEU score and length ratio with different rewards for different beam size. We choose beam size equals to 5 and reward equals to 0.1 during decoding.

## 3.4 Results

WMT organization provides three different evaluating metrics: BLEU (Papineni et al., 2002), METEOR (Lavie and Denkowski, 2009) and TER (Snover et al., 2006).

Table 2 to Table 5 summarize the performance with their corresponding rank among all other systems. We only show a few top performing systems in the tables to make a comparison. OSU1 is our proposed model and OSU2 is our baseline system without any image information. For MSCOCO dataset, the translation from English to German (Table 3), which is the hardest tasks compared with others since it is from English to German on OOD dataset, we achieve best TER score across all other systems.

| System | Rank | TER | METEOR | BLEU |
|---|---|---|---|---|
| UvA-TiCC | 1 | **47.5** | 53.5 | **33.3** |
| NICT | 2 | 48.1 | **53.9** | 31.9 |
| LIUMCVC | 3 & 4 | 48.2 | 53.8 | 33.2 |
| CUNI | 5 | 50.7 | 51 | 31.1 |
| OSU2$^\dagger$ | 6 | 50.7 | 50.6 | 31 |
| OSU1$^\dagger$ | 8 | 51.6 | 48.9 | 29.7 |

**Table 2:** Experiments on Flickr30K dataset for translation from English to German. 16 systems in total. † represents our system.

| System | Rank | TER | METEOR | BLEU |
|---|---|---|---|---|
| OSU1$^\dagger$ | 1 | **52.3** | 46.5 | 27.4 |
| UvA-TiCC | 2 | 52.4 | 48.1 | 28 |
| LIUMCVC | 3 | 52.5 | **48.9** | **28.7** |
| OSU2$^\dagger$ | 8 | 55.9 | 45.7 | 26.1 |

**Table 3:** Experiments on MSCOCO dataset for translation from English to German. 15 systems in total. † represents our system.

As describe in section 2, OSU1 is the model with image information for both encoder and

| System | Rank | TER | METEOR | BLEU |
|---|---|---|---|---|
| LIUMCVC | 1 | **28.4** | **72.1** | **55.9** |
| NICT | 2 | 28.4 | 72 | 55.3 |
| DCU | 3 | 30 | 70.1 | 54.1 |
| OSU2$^\dagger$ | 5 | 32.7 | 68.3 | 51.9 |
| OSU1$^\dagger$ | 6 | 33.6 | 67.2 | 51 |

**Table 4:** Experiments on Flickr30K dataset for translation from English to French. 11 systems in total. † represents our system.

| System | Rank | TER | METEOR | BLEU |
|---|---|---|---|---|
| LIUMCVC | 1 | **34.2** | **65.9** | **45.9** |
| NICT | 2 | 34.7 | 65.6 | 45.1 |
| DCU | 3 | 35.2 | 64.1 | 44.5 |
| OSU2$^\dagger$ | 4 | 36.7 | 63.8 | 44.1 |
| OSU1$^\dagger$ | 6 | 37.8 | 61.6 | 41.2 |

**Table 5:** Experiments on MSCOCO dataset for translation from English to French. 11 systems in total.

decoder, and OSU2 is only the neural machine translation baseline without any image information. From the above results table we found that image information would hurt the performance in some cases. In order to have more detailed analysis, we show some test examples for the translation from English to German on MSCOCO dataset.

Fig 4 shows two examples that NMT baseline model performances better than OSU1 model. In the first example, OSU1 generates several unseen objects from given image, such like knife. The image feature might not represent the image accurately. For the second example, OSU1 model ignores the object "box" in the image.

Fig 5 shows two examples that image feature helps the OSU1 to generate better results. In the first example, image feature successfully detects the object "drink" while the baseline completely neglects this. In the second example, the image feature even help the model figure out the action of the cat is "sleeping".

## 4 Conclusion

We describe our system submission to the shared WMT'17 task "multimodal translation task I". The results for English-German and English-French on Flickr30K and MSCOCO datasets are reported in this paper. Our proposed model is simple but effective and we achieve the best performance in TER for
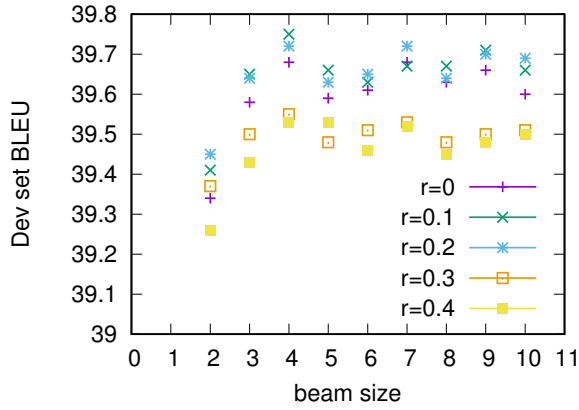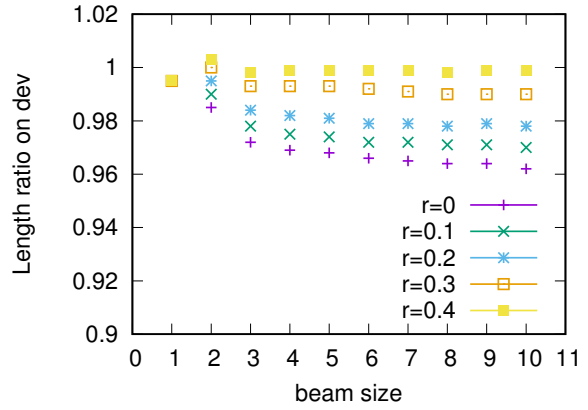
**Figure 2:** BLEU vs. beam size



**Figure 3:** length ratio vs. beam size



| input | a finger pointing at a hotdog with cheese , sauerkraut and ketchup . |
|---|---|
| OSU1 | ein finger zeigt auf einen hot dog mit einem messer , wischmobs und napa . |
| OSU2 | ein finger zeigt auf einen hotdog mit hammer und italien . |
| Reference | ein finger zeigt auf einen hotdog mit käse , sauerkraut und ketchup . |



| input | a man reaching down for something in a box |
|---|---|
| OSU1 | ein mann greift nach unten , um etwas zu irgendeinem . |
| OSU2 | ein mann greift nach etwas in einer kiste . |
| Reference | ein mann bückt sich nach etwas in einer schachtel . |

**Figure 4:** Two testing examples that image information confuses the NMT model.



| input | there are two foods and one drink set on the clear table . |
|---|---|
| OSU1 | da sind zwei speisen und ein getränk am klaren tisch . |
| OSU2 | zwei erwachsene und ein erwachsener befinden sich auf dem rechteckigen tisch . |
| Reference | auf dem transparenten tisch stehen zwei speisen und ein getränk . |



| input | a camera set up in front of a sleeping cat . |
|---|---|
| OSU1 | eine kameracrew vor einer schlafenden katze . |
| OSU2 | eine kamera vor einer blonden katze . |
| Reference | eine kamera , die vor einer schlafenden katze aufgebaut ist |

**Figure 5:** Two testing examples that image information helps the NMT model.

English-German for MSCOCO dataset.

# 5 Acknowledgment

# References

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

D. Elliott, S. Frank, K. Sima'an, and L. Specia. 2016. Multi30k: Multilingual english-german image descriptions. *Proceedings of the 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* .

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *Conference on Computer Vision and Pattern Recognition CVPR* .

Liang Huang, Kai Zhao, and Mingbo Ma. 2017. Optimal beam search for neural text generation (modulo beam size). In *EMNLP 2017* .

G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *ArXiv e-prints* .

Alon Lavie and Michael J. Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine Translation* .

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context .

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

Danilo Mirkovic, Lawrence Cavedon, Matthew Purver, Florin Ratiu, Tobias Scheideck, Fuliang Weng, Qi Zhang, and Kui Xu. 2011. Dialogue management using scripts and combined confidence scores. *US Patent* pages 7,904,297.

Ramesh Nallapati, Bowen Zhou, and Mingbo Ma. 2016. Classify or select: Neural architectures for extractive document summarization. *CoRR* .

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* .

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization .

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* .

Amanda Stent, Rashmi Prasad, and Marilyn Walker. 2004. Trainable sentence planning for complex information presentation in spoken dialog systems. *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics* .

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *Proceedings of the 27th International Conference on Neural Information Processing Systems* .

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. *IEEE Conference on Computer Vision and Pattern Recognition* pages 3156–3164.

Tsung-Hsien Wen, Milica Gasic, Dongho Kim, Nikola Mrksic, Pei-hao Su, David Vandyke, and Steve J. Young. 2015. Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. *CoRR* .

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)* .

Zhilin Yang, Ye Yuan, Yuexin Wu, William W. Cohen, and Ruslan Salakhutdinov. 2016. Review networks for caption generation. *Advances in Neural Information Processing Systems* .