Ensemble Sequence Level Training for Multimodal MT: OSU-Baidu WMT18 Multimodal Machine Translation System Report

Renjie Zheng [*1] **Yilin Yang** [*1] **Mingbo Ma** [† 1,2] **Liang Huang** [† 2,1]
[1]School of EECS, Oregon State University, Corvallis, OR
[2]Baidu Research, Sunnyvale, CA
zheng@renj.me {yilinyang721, cosmmb, liang.huang.sh}@gmail.com

Abstract

This paper describes multimodal machine translation systems developed jointly by Oregon State University and Baidu Research for WMT 2018 Shared Task on multimodal translation. In this paper, we introduce a simple approach to incorporate image information by feeding image features to the decoder side. We also explore different sequence level training methods including scheduled sampling and reinforcement learning which lead to substantial improvements. Our systems ensemble several models using different architectures and training methods and achieve the best performance for three subtasks: En-De and En-Cs in task 1 and (En+De+Fr)-Cs task 1B.

1 Introduction

In recent years, neural text generation has attracted much attention due to its impressive generation accuracy and wide applicability. In addition to demonstrating compelling results for machine translation (Sutskever et al., 2014; Bahdanau et al., 2014), by simple adaptation, similar models have also proven to be successful for summarization (Rush et al., 2015; Nallapati et al., 2016), image or video captioning (Venugopalan et al., 2015; Xu et al., 2015) and multimodal machine translation (Elliott et al., 2017; Caglayan et al., 2017; Calixto and Liu, 2017; Ma et al., 2017), which aims to translate the caption from one language to another with the help of the corresponding image.

However, the conventional neural text generation models suffer from two major drawbacks. First, they are typically trained by predicting the next word given the previous ground-truth word. But at test time, the models recurrently feed their own predictions into it. This "exposure bias" (Ranzato et al., 2015) leads to error accumulation during generation at test time. Second, the models are optimized by maximizing the probability of the next ground-truth words which is different from the desired non-differentiable evaluation metrics, e.g. BLEU.

Several approaches have been proposed to tackle the previous problems. Bengio et al. (2015) propose scheduled sampling to alleviate "exposure bias" by feeding back the model's own predictions with a slowly increasing probability during training. Furthermore, reinforcement learning (Sutton et al., 1998) is proven to be helpful to directly optimize the evaluation metrics in neural text generation models training. Ranzato et al. (2015) successfully use the REINFORCE algorithm to directly optimize the evaluation metric over multiple text generation tasks. Rennie et al. (2017); Liu et al. (2017) achieve state-of-the-art on image captioning using REINFORCE with baseline to reduce training variance.

Moreover, many existing works show that neural text generation models can benefit from model ensembling by simply averaging the outputs of different models (Elliott et al., 2017; Rennie et al., 2017). Garmash and Monz (2016) claim that it is essential to introduce diverse models into the ensemble. To this end, we ensemble models with various architectures and training methods.

This paper describes our participation in the WMT 2018 multimodal tasks. Our submitted systems include a series of models which only consider text information, as well as multimodal models which also include image information to initialize the decoders. We train these models using scheduled sampling and reinforcement learning. The final outputs are decoded by ensembling those models. To the best of our knowledge, this is the first multimodal machine translation system that achieves the state-of-the-art using sequence level learning methods.

[*] Equal contribution
[†] Contributions made while at Baidu Research

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 638–642

https://doi.org/10.18653/v1/W18-6443

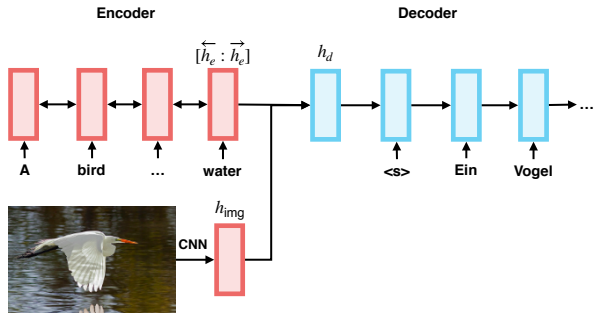

Figure 1: Multimodal Machine Translation Model

2 Methods

Our model is based on the sequence-to-sequence RNN architecture with attention (Bahdanau et al., 2014). We incorporate image features to initialize the decoder's hidden state as shown in Figure 1. Originally, this hidden state is initialized using the concatenation of last encoder's forward and backward hidden states, $\overrightarrow{h_e}$ and $\overleftarrow{h_e}$ resp. We propose to use the sum of encoder's output and image features h_{img} to initialize the decoder. Formally, we have the final initialization state h_d as:

$$h_d = \tanh(W_e[\overrightarrow{h_e}; \overleftarrow{h_e}] + W_{\text{img}} h_{\text{img}} + b). \quad (1)$$

where W_e and W_{img} project the encoder and image feature vector into the decoder hidden state dimensionality and b is the bias parameter. This approach has been previously explored by Calixto and Liu (2017).

As discussed previously, translation systems are traditionally trained using cross entropy loss. To overcome the discrepancy between training and inference distributions, we train our models using scheduled sampling (Bengio et al., 2015) which mixes the ground truth with model predictions, further adopting the REINFORCE algorithm with baseline to directly optimize translation metrics.

2.1 Scheduled Sampling

When predicting a token $\hat{y}_t$, scheduled sampling uses the previous model prediction $\hat{y}_{t-1}$ with probability ϵ or the previous ground truth prediction y_{t-1} with probability $1 - \epsilon$. The model prediction $\hat{y}_{t-1}$ is obtained by sampling a token according to the probability distribution by $P(y_{t-1}|h_{t-1})$. At the beginning of training, the sampled token can be very random. Thus, the probability ϵ is set very low initially and increased over time.

One major limitation of scheduled sampling is that at each time step, the target sequences can be incorrect since they are randomly selected from the ground truth data or model predictions, regardless of how input was chosen (Ranzato et al., 2015). Thus, we use reinforcement learning techniques to further optimize models on translation metrics directly.

2.2 Reinforcement Learning

Following Ranzato et al. (2015) and Rennie et al. (2017), we use REINFORCE with baseline to directly optimize the evaluation metric.

According to the reinforcement learning literature (Sutton et al., 1998), the neural network, θ, defines a policy p_θ, that results in an "action" that is the prediction of next word. After generating the end-of-sequence term (EOS), the model will get a reward r, which can be the evaluation metric, e.g. BLEU score, between the golden and generated sequence. The goal of training is to minimize the negative expected reward.

$$L(\theta) = -\mathbb{E}_{w^s \sim p_\theta}\left[r(w^s)\right]. \quad (2)$$

where sentence $w^s = (w_1^s, ..., w_T^s)$.

In order to compute the gradient $\nabla_\theta L(\theta)$, we use the REINFORCE algorithm, which is based on the observation that the expected gradient of a non-differentiable reward function can be computed as follows:

$$\nabla_\theta L(\theta) = -\mathbb{E}_{w^s \sim p_\theta}\left[r(w^s)\nabla_\theta \log p_\theta(w^s)\right]. \quad (3)$$

The policy gradient can be generalized to compute the reward associated with an action value *relative* to a reference reward or *baseline* b:

$$\nabla_\theta L(\theta) = -\mathbb{E}_{w^s \sim p_\theta}\left[(r(w^s) - b)\nabla_\theta \log p_\theta(w^s)\right]. \quad (4)$$

The baseline does not change the expected gradient, but importantly, it can reduce the variance of the gradient estimate. We use the baseline introduced in Rennie et al. (2017) which is obtained by the current model with greedy decoding at test time.

$$b = r(\hat{w^s}) \quad (5)$$

where $\hat{w^s}$ is generated by greedy decoding.

For each training case, we approximate the expected gradient with a single sample $w^s \sim p_\theta$:

$$\nabla_\theta L(\theta) \approx -(r(w^s) - b)\nabla_\theta \log p_\theta(w^s). \quad (6)$$

	Train	Dev.	Vocab.	Vocab. after BPE
En	2,900	1,014	10,212	7,633
De	2,900	1,014	18,726	5,942
Fr	2,900	1,014	11,223	6,457
Cs	2,900	1,014	22,400	8,459

Table 1: Statistics of Flickr30K Dataset

2.3 Ensembling

In our experiments with relatively small training dataset, the translation qualities of models with different initializations can vary notably. To make the performance much more stable and improve the translation quality, we ensemble different models during decoding to achieve better translation.

To ensemble, we take the average of all model outputs:

$$\hat{y_t} = \sum_{i=1}^{N} \frac{\hat{y_t^i}}{N} \quad (7)$$

where $\hat{y_t^i}$ denotes the output distribution of ith model at position t. Similar to Zhou et al. (2017), we can ensemble models trained with different architectures and training algorithms.

3 Experiments

3.1 Datasets

We perform experiments using Flickr30K (Elliott et al., 2016) which are provided by the WMT organization. Task 1 (Multimodal Machine Translation) consists of translating an image with an English caption into German, French and Czech. Task 1b (Multisource Multimodal Machine Translation) involves translating parallel English, German and French sentences with accompanying image into Czech.

As shown in Table 1, both tasks have 2900 training and 1014 validation examples. For preprocessing, we convert all of the sentences to lower case, normalize the punctuation, and tokenize. We employ byte-pair encoding (BPE) (Sennrich et al., 2015) on the whole training data including the four languages and reduce the source and target language vocabulary sizes to 20k in total.

3.2 Training details

The image feature is extracted using ResNet-101 (He et al., 2016) convolutional neural network trained on the ImageNet dataset. Our implementation is adapted from Pytorch-based OpenNMT (Klein et al., 2017). We use two layered bi-LSTM (Sutskever et al., 2014) as the encoder and share the vocabulary between the encoder and the decoder. We adopt length reward (Huang et al., 2017) on En-Cs task to find the optimal sentence length. We use a batch size of 50, SGD optimization, dropout rate as 0.1 and learning rate as 1.0. Our word embeddings are randomly initialized of dimension 500.

To train the model with scheduled sampling, we first set probability ϵ as 0, and then gradually increase it 0.05 every 5 epochs until it's 0.25. The reinforcement learning models are trained based on those models pre-trained by scheduled sampling.

	En-De	En-Fr	En-Cs
NMT	39.64	58.36	31.27
NMT+SS	40.19	58.67	31.38
NMT+SS+RL	40.60	58.80	31.73
MNMT	39.27	57.92	30.84
MNMT+SS	39.87	58.80	31.21
MNMT+SS+RL	40.39	58.78	31.36
NMT Ensemble	**42.54**	61.43	**33.15**
MIX Ensemble	42.45	**61.45**	33.11

Table 2: BLEU scores of different approaches on the validation set. Details of the ensemble models are described in Table 9.

3.3 Results for task 1

To study the performance of different approaches, we conduct an ablation study. Table 2 shows the BLEU scores on validation set with different models and training methods. Generally, models with scheduled sampling perform better than baseline models, and reinforcement learning further improves the performance. Ensemble models lead to substantial improvements over the best single model by about +2 to +3 BLEU scores. However, by including image information, MNMT per-

Task	System	NMT+SS	NMT+SS+RL	MNMT+SS	MNMT+SS+RL
En-De	NMT	7	6	0	0
	MIX	7	6	5	4
En-Fr	NMT	9	5	0	0
	MIX	9	0	3	0
En-Cs	NMT	7	6	0	0
	MIX	7	6	5	4

Table 3: Number of different models used for ensembling.

	Rank	BLEU	METEOR	TER
OSU-BD-NMT	1	**32.3**	50.9	49.9
OSU-BD-MIX	2	32.1	50.7	**49.6**
LIUMCVC-MNMT-E	3	31.4	51.4	52.1
UMONS-DeepGru	4	31.1	**51.6**	53.4
LIUMCVC-NMT-E	5	31.1	51.5	52.6
SHEF1-ENMT	6	30.9	50.7	52.4
Baseline	-	27.6	47.4	55.2

Table 4: En-De results on test set. 17 systems in total. (Only including constrained models).

	Rank	BLEU	METEOR	TER
LIUMCVC-MNMT-E	1	**39.5**	59.9	41.7
UMONS	2	39.2	**60**	41.8
LIUMCVC-NMT-E	3	39.1	59.8	41.9
OSU-BD-NMT	4	39.0	59.5	**41.2**
SHEF-MLT	5	38.9	59.8	41.5
OSU-BD-MIX	9	38.6	59.3	41.5
Baseline	-	28.6	52.2	58.8

Table 5: En-Fr results on test set. 14 systems in total. (Only including constrained models).

	Rank	BLEU	METEOR	TER
OSU-BD-NMT	1	**30.2**	29.5	**50.7**
OSU-BD-MIX	2	30.1	**29.7**	51.2
SHEF1-ENMT	3	29.0	29.4	51.1
SHEF-LT	4	28.3	29.1	51.7
SHEF-MLT	5	28.2	29.1	51.7
SHEF1-MFS	6	27.8	29.2	52.4
Baseline	-	26.5	27.7	54.4

Table 6: En-Cs results on test set. 8 systems in total. (Only including constrained models).

	En-Cs	Fr-Cs	De-Cs	(En+Fr+De)-Cs
NMT	**31.27**	28.48	26.96	29.47
MNMT	30.84	27.02	25.99	29.23

Table 7: BLEU scores on validation set for task 1B

forms better than NMT only on the En-Fr task with scheduled sampling.

Table 4, 5 and 6 show the test set performance of our models on En-De, En-Fr and En-Cs subtasks with other top performance models. We rank those models according to BLEU. Our submitted systems rank first in BLEU and TER on En-De and En-Cs subtasks.

3.4 Results for task 1B

Table 7 shows the results on validation set without sequence training. En-Cs, Fr-Cs, De-Cs are models trained from one language to another. (En+Fr+De)-Cs models are trained using multiple source data. Similar to the Shuffle method discussed in multi-reference training (Zheng et al., 2018), we randomly shuffle the source data in all languages and train using a traditional attention based-neural machine translation model in every epoch. Since we do BPE on the whole training data, we can share the vocabulary of different languages during training. The results show that models trained using single English to Czech data perform much better than the rest.

	Rank	BLEU	METEOR	TER
OSU-BD-NMT	1	**26.4**	28.0	**52.1**
OSU-BD-MIX	1	**26.4**	**28.2**	52.7
SHEF1-ARNN	3	25.2	27.5	53.9
SHEF-CON	4	24.7	27.6	**52.1**
SHEF-MLTC	5	24.5	27.5	52.5
SHEF1-ARF	6	24.1	27.1	54.6
Baseline	-	23.6	26.8	54.2

Table 8: Task 1B multi-source translation results on test set. 6 systems in total.

Task	System	Model Rank				Team Rank			
		Num †	BLEU	MET.	TER	Num ‡	BLEU	MET.	TER
En-De	NMT	11	1	4	2	5	1	3	1
	MIX	11	2	5	1				
En-Fr	NMT	11	4	9	1	6	3	5	1
	MIX	11	9	10	3				
En-Cs	NMT	6	1	1	1	3	1	1	1
	MIX	6	2	2	3				
En-Cs (1B)	NMT	6	1	2	1	3	1	1	1
	MIX	6	1	1	5				

Table 9: Rank of our models. † represents the total number of models. ‡ represents the total number of teams.

Table 8 shows results on test set. The submitted systems are the same as those used in En-Cs task of task 1. Although we only consider the English source during training, our proposed systems still rank first among all the submissions.

4 Conclusions

We describe our systems submitted to the shared WMT 2018 multimodal translation tasks. We use sequence training methods which lead to substantial improvements over strong baselines. Our ensembled models achieve the best performance in BLEU score for three subtasks: En-De, En-Cs of task 1 and (En+De+Fr)-Cs task 1B.

Acknowledgments

This work was supported in part by DARPA grant N66001-17-2-4030, and NSF grants IIS-1817231 and IIS-1656051. We thank the anonymous reviewers for suggestions and Juneki Hong for proofreading.

References

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

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1171–1179.

Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost van de Weijer. 2017. Lium-cvc submissions for wmt17 multimodal translation task. In *Proceedings of the Second Conference on Machine Translation*, pages 432–439.

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 992–1003.

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, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233.

Ekaterina Garmash and Christof Monz. 2016. Ensemble learning for multi-source neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1409–1418.

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. When to finish? 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*.

Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. 2017. Improved image captioning via policy gradient optimization of spider. In *Proc. IEEE Int. Conf. Comp. Vis*, volume 3, page 3.

Mingbo Ma, Dapeng Li, Kai Zhao, and Liang Huang. 2017. Osu multimodal machine translation system report. In *Proceedings of the Second Conference on Machine Translation*, pages 465–469.

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

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *CVPR*, volume 1, page 3.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

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*.

Richard S Sutton, Andrew G Barto, et al. 1998. *Reinforcement learning: An introduction*. MIT press.

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond J. Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence - video to text. In *ICCV*.

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)*.

Renjie Zheng, Mingbo Ma, and Huang Liang. 2018. Multi-reference training with pseudo-references for neural translation and text generation. *arXiv preprint arXiv:1808.09564*.

Long Zhou, Wenpeng Hu, Jiajun Zhang, and Chengqing Zong. 2017. Neural system combination for machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 378–384.