# Team RUC AI·M$^3$ Technical Report at VMT Challenge 2020: Enhancing Neural Machine Translation with Multimodal Rewards

**Yuqing Song**
Renmin University of China
syuqing@ruc.edu.cn

**Shizhe Chen**
Renmin University of China
cszhe1@ruc.edu.cn

**Qin Jin**
Renmin University of China
qjin@ruc.edu.cn

## Abstract

This notebook paper presents our model in the VATEX video-guided machine translation (VMT) challenge. We propose an *objective-side* VMT model by improving the translation system with multimodal rewards. The visual contexts are used in an additional objective to correct visually-discrepant translations and generally make the learned language embeddings more visually grounded. The proposed model significantly outperforms the state-of-the-art NMT baseline and achieves 35.28 BLEU4 score on the testing set which ranks the third place at the VATEX VMT challenge leaderboard 2020.

## 1 Introduction

The common visual world has been widely used to bridge the semantic gap in cross-lingual lexicon mappings (Chen et al., 2019; Sigurdsson et al., 2020). It is also wondered if the visual context can benefit more complex but universal sentence-level translations. Therefore, the video-guided machine translation (VMT) task (Wang et al., 2019) is proposed to explore enhancing neural machine translation (NMT) system with the video contexts.

Many endeavors (Wang et al., 2019; Calixto et al., 2017; Calixto and Liu, 2017) have been made to explore where and how to integrate the visual context into the NMT system. They take the visual information as an additional input with the source sentence and explicitly use them to help the translation of each source word. We call such models as the *input-side* VMT models. Such *input-side* VMT models can benefit to disambiguate source words and make the NMT system more robust to noise. The improvement is more obvious when the source sentences are randomly blanked out as in (Wang et al., 2019). However, the situation of words being blanked out is rather rare in realistic applications. When the source sentences are clean, the *input-side* VMT models improve little over the text-purely translation models. In this situation, the bottleneck of NMT performance is not the lack of source information, but the visual semantic learning of words. Therefore, the visual context can be utilized as an additional objective to correct visually-discrepant translations and generally make the learned language embeddings more visually grounded.

In this work, we propose an *objective-side* VMT model for the video-guided machine translation. We take into account the strengths of text-purely translation model and propose a novel *objective-side* visual assistance structure to further improve visual consistency of the translation, which employs reinforcement learning with multimodal rewarding. To be specific, we employ a video-sentence matching model to provide visual rewards for generated translations in reinforcement learning framework. In order to avoid the translation only satisfying the video content but not in accord with the source sentence, we further combine the visual guidance with textual semantic guidance to provide multimodal rewards for the translation model. Our proposed VMT model achieves the third place performance on the VMT challenge testing set.

## 2 Approach

In this section, we will describe our proposed VMT model. We first describe our NMT baseline model in Section 2.1, and then introduce the VMT model enhanced with multimodal rewards in Section 2.2. Finally, we will describe the training strategy of our model in details in Section 2.3.

### 2.1 NMT

We utilize the state-of-the-art translation model - transformer (Vaswani et al., 2017) as our NMT baseline. It consists of two modules, namely source sentence encoder and target sentence decoder. For a sentence $X = \{x_1, x_2, \cdots, x_{T_x}\}$

in the source language, the source sentence encoder encodes it into a set of distributional vectors $H^x = \{h_1^x, h_2^x, \cdots, h_{T_x}^x\}$. Then, the target sentence decoder generates the translation $Y = \{y_1, y_2, \cdots, y_{T_y}\}$ in the target language word by word conditioned on $H^x$ via attention mechanism.

**Source Sentence Encoder.** The source sentence encoder has 6 stacked encoding layers. Each encoding layer comprises a multi-head self-attention sub-layer and a fully connected feed-forward sub-layer to capture both left and right side contextual information of the source language sentence. A residual connection with layer normalization is employed to follow each of the two sub-layers. Dropout is adopted to the output of each sub-layer before it is added and normalized to prevent over-fitting.

**Target Sentence Decoder.** The target sentence decoder is also composed of 6 decoding layers. Each decoding layer comprises of a multi-head masked self-attention sub-layer, multi-head source-attention sub-layer, and a fully connected feed-forward sub-layer. The residual connection and layer normalization are also employed to follow each sub-layer. The output of the last decoding layer is transferred to the dimensionality of the target language vocabulary size through a linear layer followed with softmax function. In our implementation, we share the same weight matrix between the target language embedding layer and the pre-softmax linear transformation.

### 2.2 NMT with Multimodal Rewarding

To ensure the generated translation satisfies semantic consistency with both the source language sentence and the visual content, we propose to further enhance the transformer baseline with multimodal rewards in the reinforcement learning framework.

we pre-train a video-sentence cross-modal matching model on parallel pairs $(V^i, Y^i)$. We divide the video into non-overlapping segments with 16 frames per segment. Then we extract segment-level features for the video from image and motion modalities. For the image modality, we utilize ResNext101 (Xie et al., 2017) pretrained on billion scale weakly-supervised data (Mahajan et al., 2018) to extract features for each segment. For the motion modality, we adopt ir-csn model (Tran et al., 2019) pretrained on Kinetics 400 to extract features. We then concatenate them to get the video embedding sequence $V^i = \{v_1^i, v_2^i, \cdots, v_T^i\}$. To capture the temporal dependencies for both the video segment sequence and sentence sequence, we adopt bi-directional GRU (Cho et al., 2014) to further encode the video and sentence sequence separately. The mean poolings of encoded hidden states for the video and sentence are then mapped to a common embedding space with fully connected embedding layer. The contrastive ranking loss with hard negative mining (Faghri et al., 2018) is utilized to train the video-sentence matching model.

After training, the video-sentence matching model is able to give higher similarity scores to translations relevant to the video than those irrelevant ones. Therefore, the visual similarity reward for the generated translation $c$ of video $v$ can be represented as follows:

$$r_{vis}(c) = s(E_y(c), E_v(v)), \tag{1}$$

where $E_y$ denotes the sentence encoder, $E_v$ denotes the video encoder, and s$(\cdot)$ denotes the cosine similarity.

To avoid the translation only satisfying the video content but not in accord with the source sentence, we further combine the visual reward with textual semantic reward $r_{text}$, which is computed by BLEU (Papineni et al., 2002) metric. Therefore, the multimodal reward for the translation $c$ can be represented as follows:

$$r(c) = \alpha r_{vis}(c) + \beta r_{text}(c), \tag{2}$$

where $\alpha$ and $\beta$ are hyper-parameters.

### 2.3 Training Strategy

We first train the NMT model with the maximum likelihood estimation of ground-truth data as follows:

$$\mathcal{L}_{XE} = -\sum_{i=1}^{N}\sum_{t=1}^{T_y} \log P(y_t^i | y_{<t}^i, X^i; \Theta) \tag{3}$$

where $\Theta$ denotes all learnable parameters in the NMT model. Then we further improve the NMT model with multimodal rewards via the "self-critical" (Rennie et al., 2017) reinforcement learning algorithm.

Firstly, we carry out Monte-Carlo sampling to sample a translation candidate $s_s^i = \{w_1^i, \cdots, w_n^i\}$ and evaluate its translation quality with the proposed multimodal rewarding function. Then we utilize the greedy search algorithm to generate a translation $s_b^i$ to provide a baseline reward for the stability of reinforcement training. Therefore,

the reinforcement learning objective for the VMT model can be expressed as follows:

$$\mathcal{L}_{RL} = -\sum_{i=1}^{N}(r(s_s^i) - r(s_b^i))\sum_{t=1}^{n}\log P(w_t^i|w_{<t}^i, X^i; \Theta) \tag{4}$$

where $r(\cdot)$ is the proposed multimodal rewarding function described in Section 2.2.

## 3 Experiments

### 3.1 Dataset

We utilize the VATEX (Wang et al., 2019) dataset for the VMT task, which consists of 34,402 videos with 5 English-Chinese parallel pairs per video. Besides, there are also additional 5 English and Chinese non-parallel captions annotated for each video. We follow the official split with 25,467 videos for training, 2,935 videos for validation and 6,000 videos for testing in the experiments. To avoid the noise, we use the giza-pp [1] tool to filter mis-aligned pairs, which results in 100,000 parallel pairs for training. In the final submission, we enlarge the training set with back-translation. We pre-train a Chinese-to-English translation model on VATEX dataset, and use it to generate English translations for all the Chinese annotations in the dataset. The generated pseudo pairs are merged to the annotated parallel pairs for the final training. We adopt the byte pair encoding (bpe) algorithm (Sennrich et al., 2016) to get both the source and target language vocabulary, which contains 9,594 and 14,305 words in English and Chinese respectively.

### 3.2 Implementation Details

For the transformer, we follow (Vaswani et al., 2017) to set $d_{model} = 512$, $d_{ff} = 2048$ and 8 heads for multi-head attention. During training, each batch contains a set of sentence pairs with approximately 2048 source tokens and 2048 target tokens. We utilize label-smoothing with value set as 0.1. The learning rate is varied under a warm-up strategy with 8,000 steps. At inference phase, we use beam search with a beam size of 5 and length penalty of 0.7. The hyper-parameter $\alpha$ and $\beta$ in multimodal rewards are set to 1 and 2 respectively.

### 3.3 Experimental Results

Table 1 presents the performances of our NMT and VMT models. The VMT model with visual reward alone achieves 1 point BLEU4 improvement over the NMT model on the validation and testing set consistently. Training with textual reward achieves competitive performance. Combining the visual and textual reward for VMT model achieves additional gains with 34.76 BLEU4 score. The last three rows show the performances with training set enlarged. Enlarging the training set with back translation improves both the NMT and VMT models significantly. We finally ensemble VMT models with multiple runs and achieve 35.28 BLEU4 score on the testing set.

Table 1: Performances of different models for EN-ZH VMT on VATEX dataset.

| Methods | $r_{vis}$ | $r_{text}$ | Val. | Test |
|---|---|---|---|---|
| NMT | | | 33.47 | 33.10 |
| VMT | √ | | 34.48 | 34.00 |
| VMT | | √ | 34.44 | - |
| VMT | √ | √ | 34.76 | - |
| Ensemble | √ | √ | 35.26 | 34.54 |
| NMT | | | 34.54 | - |
| VMT | √ | √ | 35.82 | - |
| **Ensemble** | √ | √ | **36.13** | **35.28** |

## 4 Conclusion

In this work, we propose an *objective-side* VMT model by enhancing neural machine translation with multimodal rewards. We adopt the state-of-the-art NMT model and further improve it with both visual and textual rewards in reinforcement learning framework. Our whole system achieves the third place performance on the video-guided machine translation challenge 2020.

## References


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, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 992–1003. Association for Computational Linguistics.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1913–1924. Association for Computational Linguistics.


[1] https://github.com/moses-smt/giza-pp

Shizhe Chen, Qin Jin, and Alexander G. Hauptmann. 2019. Unsupervised bilingual lexicon induction from mono-lingual multimodal data. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 8207–8214. AAAI Press.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Empirical Methods in Natural Language Processing*, pages 1724–1734.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference*, page 12.

Dhruv Mahajan, Ross B. Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. 2018. Exploring the limits of weakly supervised pretraining. In *Computer Vision - ECCV*, volume 11206 of *Lecture Notes in Computer Science*, pages 185–201. Springer.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Conference on Computer Vision and Pattern Recognition*, pages 1179–1195.

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*. The Association for Computer Linguistics.

Gunnar A. Sigurdsson, Jean-Baptiste Alayrac, Aida Nematzadeh, Lucas Smaira, Mateusz Malinowski, João Carreira, Phil Blunsom, and Andrew Zisserman. 2020. Visual grounding in video for unsupervised word translation. *CoRR*, abs/2003.05078.

Du Tran, Heng Wang, Matt Feiszli, and Lorenzo Torresani. 2019. Video classification with channel-separated convolutional networks. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 5551–5560. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *The IEEE International Conference on Computer Vision (ICCV)*.

Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition*.