# DeepFuse: HKU's Multimodal Machine Translation System for VMT'20

**Zhiyong Wu**
University of Hong Kong
zywu@cs.hku.hk

## Abstract

In this draft, we present our submission to the VMT'20 English to Chinese translation task. Unlike previous research that obtain visual and textual representation separately, and use attention to incorporate visual features into decoding, we propose to use a video-augmented encoder to obtain a multimodal representation for decoder. Experiments on VATEX dataset show a large improvement (6.78 in BLEU-4) over a strong baseline.

## 1 Introduction

Significant progress has been made in the field of multimodal neural machine translation in recent years, especially in image-guided machine translation (IMT). A similar task – video-guided machine translation (VMT), however, receives much less attention. The major challenge hindering progress in VMT is the relative scarcity of datasets. Recent efforts like How2 and VaTeX (Wang et al., 2019) datasets have started to alleviate this bottleneck.

Research on IMT has explored many effective methods to incorporate image information into the neural machine translation (NMT) model, which can be directly applied to VMT. However, most of these studies are based on shallow fusion: where they obtain visual and textual representation separately and fuse those representations at a certain stage. Relatively less effort, however, has been spent on learning a multimodal representation for the translation model. In this draft, we present our submission – DeepFuse, which fuses multimodal representation at multiple layers using attention mechanism. Our experiments show that such a deep fusion approach has clear gains over the previous shallow fusion approach.

## 2 Related Work

This section presents a review of recent IMT methods. We categorize those methods into two types according to the operation used, namely, simple operation-based and attention-based.

**Simple Operation-based Grounding** Early studies attempt to use simple operations to ground visual information into language, such as concatenation, weighting, or initialize first hidden state using the image representation. Elliott et al. (2015) is the first attempt for neural IMT. They propose to initialize the hidden states of the encoder and/or decoder using pre-trained image features. Later variants (Libovický and Helcl, 2017; Calixto et al., 2016; Ma et al., 2017) also demonstrate the effectiveness of this approach. In a similar fashion, Huang et al. (2016) enrich source sentence representation by appending or prepending the visual representation to the embedded source sequence. Calixto and Liu (2017) takes advantage of previous research to combine source embedding enrichment with encoder/decoder initialization. Grönroos et al. (2018) explores methods to weight output probability through a time-dependent visual decoder gate.

**Attention-based Grounding** Inspired by previous success of visual attention in image captioning (Xu et al., 2015), Caglayan et al. (2016) and Calixto et al. (2016) propose to use decoder's hidden state to select relevant visual features and generate a visual-aware representation for translation. Follow-up studies extend the decoder-based visual attention approach in different ways. Calixto et al. (2017) re-scale the visual representation before fusion. Libovický and Helcl (2017) introduces the hierarchical attention to dynamically weight and fuse representation of different modalities. More recent studies are mostly based on transformer ar-

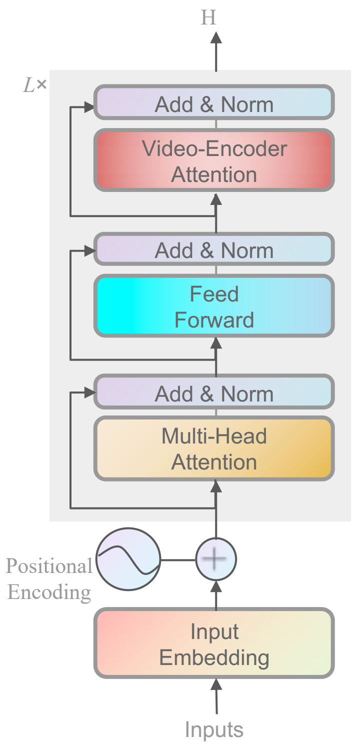


Figure 1: Overview of our visual-augmented encoder.

chitectures (Arslan et al., 2018; Libovický et al., 2018; Ive et al., 2019). However, the increasingly complex neural network architectures only yield progressively smaller gains over the previous state of the art. In contrast to the decoder-based visual attention, encoder-based approaches are relatively less explored. To this end, Delbrouck and Dupont (2017) propose to condition the mean and the variance of the batch normalisation layer on the source sentence representation for fusion.

All those previous attempts on IMT, except (Delbrouck and Dupont, 2017), can be categorized as shallow fusion method, where they obtain visual and textual features separately, and only fuse those representations at the decoder. Inspired by the recent success of multimodal representation learning (Lu et al., 2019), we propose to fuse different modalities at multiple encoder layers to obtain a multimodal representation for the decoder.

## 3 DeepFuse

In this section, we present the proposed DeepFuse system, which contains a visual-augmented encoder and a standard transformer decoder.

### 3.1 Visual-augmented Encoder

An illustration of the proposed visual-augmented encoder is shown in Figure 1. The encoder contains L=6 identical layers, each of which includes three sub-layers. The first sub-layer is a self-attention module, followed by a position-wise feed-forward network, and finally, a video-encoder attention module. Residual connection (He et al., 2016) and layer normalization (Ba et al., 2016) is applied between sub-layers. That is, the output of each sub-layer is LayerNorm(x+Sublayer(x)), where Sublayer(x) is the function implemented by the sub-layer itself.

Given a sentence as a sequence of embedding $\mathbf{X} = x_1, x_2, ..., x_n$ and the paired video, we first use a pretrained 3D convolutional neural network (Carreira and Zisserman, 2017) to convert the video into a sequence of segment-level features $\mathbf{E} = e_1, e_2, ..., e_m$. The encoder process the input $\mathbf{X}$ and $\mathbf{E}$ as follows:

**1.Self-attention** The input $\mathbf{X}$ is first transformed into query, key, and value vectors as $Q, K, V$, respectively. We compute the output of the self-attention module as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_h\right) W^O \quad (1)$$

where $d_k$ is the dimension of keys,

$$head_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (2)$$

and $\text{Attention}(Q, K, V)$ is the Scaled Dot-Product Attention computed as follow:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (3)$$

**2.Position-wise Feed-Forward** After the self-attention module, we apply a fully connected feed-forward network as:

$$\text{FFN}(x) = \max\left(0, xW_1 + b_1\right) W_2 + b_2 \quad (4)$$

**3.Video-encoder Attention** Given the output of the previous sub-layers as $\mathbf{H}^L$, we apply attention mechanism to select video representation that is relevant to the source sentence:

$$\overline{\mathcal{H}} = \text{Attention}\left(\mathbf{H}^L, \mathbf{K}_E, \mathbf{V}_E\right) \quad (5)$$

where $\mathbf{K}_E, \mathbf{V}_E$ are linear transformer of $\mathbf{E}$. We then compute a weighted sum of $\mathbf{H}^L$ and $\overline{\mathcal{H}}$ to obtain the multimodal representation:

$$\mathcal{H} = \mathbf{H}^L + \lambda \overline{\mathcal{H}} \tag{6}$$

where $\lambda$ is position-wise scalar weight learned with parameter $\mathbf{W}_\lambda$ and $\mathbf{U}_\lambda$:

$$\lambda = \text{sigmoid}\left(\mathbf{W}_\lambda \overline{\mathcal{H}} + \mathbf{U}_\lambda \mathbf{H}^L\right) \tag{7}$$

Finally, the output is normalized as $LayerNorm(\mathbf{H}^L + \mathcal{H})$

### 3.2 Decoder

The decoder is also composed of a stack of L=6 identical layers. We follow the design of the transformer to design our decoder.

## 4 Experiment

**The VaTeX dataset** is a bilingual collection of video descriptions, built on a subset of 41,250 video clips from the action classification benchmark DeepMind Kinetics-600 (Kay et al., 2017). VaTeX adds 10 Chinese and 10 English crowdsourced captions describing each video, half of which are independent annotations, and the other half Chinese–English parallel sentences. With low-approval samples removed, the released version of the dataset contains 206,345 translation pairs in total. We follow the official split to divide the dataset into training (26k videos), validation (3k videos), and private test splits (6k videos). All our analysis is based on the publicly available validation set.

**Experimental settings.** We adopt the byte pair encoding algorithm to encode the source sentence and cap the size of English to 20,000. Each training batch contains 4000 tokens. We use Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\tau = 10^{-9}$ and varied the learning rate according to:

$$\text{lrate } = d_{model}^{-0.5} \cdot \min\left(K^{-0.5}, K \cdot N^{-1.5}\right) \tag{8}$$

where K is the current number of step and N=4000 is the number of warm-up steps. During training, the value of label smoothing, the attention dropout, and residual dropout are both set to 0.1.

**Baselines.** We consider the three following baselines: (1) lstm: Text-only LSTM-based encoder-decoder NMT (2)vatex: The LSTM-based video-guided machine translation system proposed in VATEX (2) Transformer: standard text-only transformer architecture proposed by Vaswani et al. (2017).

| Model | Valid | Test |
|---|---|---|
| lstm | - | 26.85 |
| vatex | - | 29.12 |
| Transformer | 34.55 | - |
| DeepFuse | 35.77 | 35.8 |

Table 1: BLEU-4 scores of the VMT'20 English to Chinese translation.

### 4.1 Results

Results are shown in Tabel 1. Our method shows a clear improvement over strong baseline VATEX. We also observe a significant improvement over text-only transformer, suggesting that the DeepFuse can use visual information to improve the translation. Since the server room in our department in under renovation recently, we are not able to do more ablation study or analysis at the moment. We will report more experiments and results in the workshop presentation.

## 5 Conclusion

Although we observe an improvement from incorporating video information into a neural machine translation system, the improvement is modest. The largest differences in quality between the systems we experimented with can be attributed to the quality of the underlying text-only NMT system. It would be interesting future work to explore more effective ways to fuse video and text representation for translation task.

## References


Hasan Sait Arslan, Mark Fishel, and Gholamreza Anbarjafari. 2018. Doubly attentive transformer machine translation. *arXiv preprint arXiv:1807.11605*.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Ozan Caglayan, Walid Aransa, Yaxing Wang, Marc Masana, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, and Joost van de Weijer. 2016. Does multimodality help human and machine for translation and image captioning? In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 627–633, Berlin, Germany. Association for Computational Linguistics.

Iacer Calixto, Desmond Elliott, and Stella Frank. 2016. Dcu-uva multimodal mt system report. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 634–638.

Iacer Calixto and Qun Liu. 2017. Sentence-level multilingual multi-modal embedding for natural language processing. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 139–148, Varna, Bulgaria. INCOMA Ltd.

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 (Volume 1: Long Papers)*, pages 1913–1924, Vancouver, Canada. Association for Computational Linguistics.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Jean-Benoit Delbrouck and Stéphane Dupont. 2017. Modulating and attending the source image during encoding improves multimodal translation. *arXiv preprint arXiv:1712.03449*.

Desmond Elliott, Stella Frank, and Eva Hasler. 2015. Multilingual image description with neural sequence models. *arXiv preprint arXiv:1510.04709*.

Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, and Raúl Vázquez. 2018. The MeMAD submission to the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 603–611, Belgium, Brussels. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

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: Volume 2, Shared Task Papers*, pages 639–645, Berlin, Germany. Association for Computational Linguistics.

Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Distilling translations with visual awareness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6538, Florence, Italy. Association for Computational Linguistics.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202, Vancouver, Canada. Association for Computational Linguistics.

Jindřich Libovický, Jindřich Helcl, and David Mareček. 2018. Input combination strategies for multi-source transformer decoder. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 253–260, Belgium, Brussels. Association for Computational Linguistics.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

Mingbo Ma, Dapeng Li, Kai Zhao, and Liang Huang. 2017. Osu multimodal machine translation system report. *arXiv preprint arXiv:1710.02718*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, 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)*.

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. In *International conference on machine learning*, pages 2048–2057.