Do Videos Guide Translations? Evaluation of a Video-Guided Machine Translation dataset

Zhishen Yang
Tokyo Institute of Technology
zhishen.yang@nlp.c.titech.ac.jp

Tosho Hirasawa
Tokyo Metropolitan University
hirasawa-tosho@ed.tmu.ac.jp

Mamoru Komachi
Tokyo Metropolitan University
komachi@tmu.ac.jp

Naoaki Okazaki
Tokyo Institute of Technology
okazaki@c.titech.ac.jp

Abstract

Video-guided machine translation (VMT) is a new multimodal machine translation task aimed at using videos to guide translation. Visual information gleaned from videos is expected to provide context in the translation progress. Results from the Video-guided Machine Translation Challenge 2020 suggest that multimodal models only have marginal performance improvements over their text-only counterparts. We hypothesize that this is caused by the simple and short video descriptions in VATEX, the dataset used in the challenge. In this study, we examine our hypothesis by conducting input-degradation, visual sensitivity experiments, and human evaluation of VATEX. The results indicate that textual descriptions of videos in VATEX are sufficient for translation, which prevents the visual context from videos to guide the translation.

1 Introduction

Extending text-only machine translation, the multimodal machine translation task exploits information from other modalities to improve the translation quality. Video-guided machine translation (VMT) is a new multimodal machine translation task that provides videos, as additional inputs, for a model to translate sentences from the source to target languages. Compared to image-guided machine translation, videos provide visual and acoustic modalities with rich embedded information, such as actions, objects, and temporal transitions.

VMT refers to videos as additional information to improve the translation quality. Therefore, an ideal dataset for this task should provide videos as complements rather than redundancies. The recently proposed VATEX dataset is a dataset for VMT research and the shared task (VMT Challenge). From the results of the 2020 VMT challenges, all multimodal models only had marginal performance gains compared to their text-only counterparts. We hypothesize that this was caused by the design of the VATEX dataset: simple and short video descriptions are sufficient for translations, making videos redundant information for the models. To examine our hypothesis, similar to (Caglayan et al., 2019), we conducted input degradation experiments, visual sensitivity experiments, and human evaluations. The experimental and human evaluation results showed that when textual information is sufficient, visual information from videos becomes redundant to a multimodal model. The code used in this study is publicly available.[1]

2 Related Work

VMT Challenge 2020 The top three teams in the VMT Challenge 2020 presented recurrent neural network (RNN) and Transformer-based models. The winning team, Hirasawa et al. (2020), used a doubly attentive RNN-based model (Calixto et al., 2017) with positional encoding for video features. Two other teams proposed Transformer-based VMT models with modifications to incorporate video features.

Probing Auxiliary Modalities Probing the need for auxiliary modalities is an important topic in multimodal machine translation (MMT). In image-guided machine translation, Caglayan et al. (2019) conducted analytic experiments whereby source sentences were degraded in three different ways to simulate specific conditions in which images should be beneficial. Hessel and Lee (2020) proposed a method to isolate cross-modal interactions for multimodal classification tasks and showed that cross-modal interactions have no or little contribution to the model performance. In this study, we extend Caglayan et al. (2019) to probe the need for videos as a visual modality in VMT.

[1] https://github.com/ZhishenYang/eval_on_vatex_dataset

3 Input Degradation

Inspired by Caglayan et al. (2019), we conducted four source-side input-degradation experiments: color, noun, verb, and progressive masking. Since these input-degradation experiments are used to simulate the scarce textual context condition, we hypothesize that a multimodal model can rely on the visual context obtained from videos and will perform better than a monomodal model that only relies on the textual context.

Color Deprivation We replaced English words that represent color in the source sentences with a special token $[c]$. The masked tokens consist of 0.4% in both training set and validation set.

Noun Masking We replaced each noun in the source English sentences with a special token, $[n]$. This masked 28.2% of the tokens in both the training and validation sets.

Verb Masking The authors of the VATEX dataset used videos from the Kinetics-600 dataset, which contains a broad range of actions. All verbs in the source sentence were replaced with a special token, $[v]$. This replaced 14.0% of the tokens in both the training and validation sets.

Progressive Masking Progressive masking aims to progressively replace the last N words in a source sentence with a special token, $[p]$. Unlike other masking experiments, progressive masking simulates a progressive low-resource scenario (Caglayan et al., 2019). We hypothesize that with the increasing number of masked tokens in the source sentences, multimodal models with access to visual information will perform better than text-only models.

We selected $N = \{2, 4, 6, 10, 20, 30\}$ in the progressive masking experiment. When $N = 30$, nearly 100% of all words were masked, and the multimodal model performed video captioning with "expected length" as the only known information.

4 Visual Sensitivity

Visual Incongruence Test Inspired by (Elliott, 2018) and (Caglayan et al., 2019), we implemented visual incongruence tests to examine the visual sensitivity of multimodal models. In this test, we fed multimodal models with incongruent visual features during the testing time. The hypothesis is that the performance of multimodal models will deteriorate when fed with visual features from irrelevant videos.

Visual Features To test whether multimodal models are sensitive to different visual features, we extracted visual features using pre-trained models for action or object classification. We conjecture that visual features correlated to subjects/objects and actions in videos will help predict nouns and verbs in video descriptions.

5 Experiments

5.1 Dataset

We used VATEX v1.1 (the latest version) [2]. The translation direction was the same as that of the VMT challenge 2020: from English to Chinese. Because the public test set is on-hold, we used a validation set to test the performance of the models. The statistics of the dataset are provided in Appendix A.

For the tokenization, we used spaCy [3] to tokenize the English sentences, and then employed byte pair encoding (Sennrich et al., 2016) to split the English tokens into subwords, where the number of merge operations was 8000. The Chinese translations were tokenized at the character level.

5.2 Visual Feature Extraction

As described in Section 4, two types of features were correlated to the nouns and verbs that we used in the experiments: ResNet-152 and I3D features. We extracted the ResNet-152 features from the per-second frames of each video. The ResNet-152 feature is the averaged convolutional features from the last convolutional layer of ResNet-152 (He et al., 2016) pretrained on ImageNet (Deng et al., 2009). The I3D features were extracted from videos using two-stream inflated 3D ConvNet (I3D) (Carreira and Zisserman, 2017) and the I3D features provided in the VMT challenge 2020 [2].

5.3 Models

For the text-only baseline models, we used an attentive RNN model (Bahdanau et al., 2015) and transformer model (Vaswani et al., 2017).

For the multimodal models, we employed two models from the VMT challenge 2020: hierarchical attentive RNN model with positional encod-

[2] https://eric-xw.github.io/vatex-website/download.html
[3] https://spacy.io/

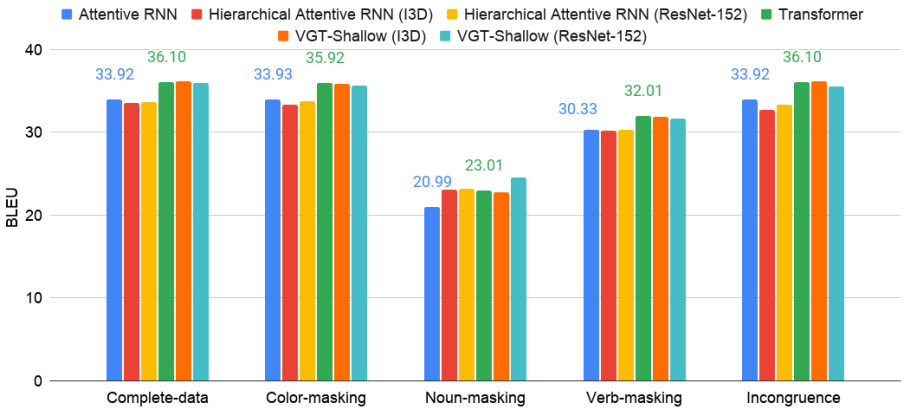

Figure 1: Corpus-level BLEU scores of the validation set. The data labels show the BLEU scores of the text-only baselines.

ing (Hirasawa et al., 2020) and VGT-Shallow[4]. See Appendix D for more details.

6 Results

In this section, we discuss the results from the input-degradation and visual sensitivity experiments in Figure 1 and 2 [5]. Without any input degradation, all transformer-based models had higher BLEU scores than those of the RNN-based models. Among the multimodal models, VGT-Shallow (I3D) achieved the best BLEU scores.

Color Deprivation Since only a small fraction of tokens was masked, compared with models trained on complete data, the differences between the multimodal and monomodal models were marginal. The text-only Transformer and attentive RNN model had slightly higher BLEU scores than those of the multimodal models.

Noun Masking All models had lower BLEU scores compared to those of their complete data baselines. All multimodal models, except for the VGT-shallow (I3D), had higher BLEU scores compared to those of the text-only models. Noun-masking has a larger masking scale; therefore, compared to the verb masking and color deprivation experiments, the multimodal models can exploit the visual context to infer missing information.

Verb Masking In verb masking, although all models had deteriorated performances, the differences between the multimodal and monomodal models were minimal.

Visual Sensitivity Test Even when feeding with incongruent visual features during testing, the multimodal models suffered slight performance deterioration as compared with training and testing on a complete dataset. This indicates that visual features have a minimal influence on multimodal models when given complete text.

In our experiments, the multimodal models with different architectures had different sensitivities to visual features. The hierarchical attentive RNN using ResNet-152 had higher BLEU scores than those obtained when using I3D features. However, for VGT-Shallow, this only occurred for the noun masking experiment.

Progressive Masking Figure 2 shows the result of progressive masking. For an increasing number of masked tokens, the multimodal models started to take advantage of visual modalities, and therefore outperformed the text-only models. Moreover, the Hierarchical Attentive RNN had the best BLEU score when nearly 100% of tokens were masked.

7 Human Evaluation

The authors of VATEX used a post-editing annotation strategy to collect parallel English-Chinese translation pairs, in which automatic translation systems were employed. Based on the results of the input-degradation experiments, we hypothesize that if English-Chinese translations are sufficient, videos become redundant in post-editing. To test our hypothesis, we conducted a human evaluation task based on a post-editing annotation strategy.

We randomly selected 500 videos from VATEX's validation set to construct the human evaluation set,

[4] https://www.youtube.com/watch?v=zHwXPmIQajA&t=517s

[5] See Appendix B for more details.

EN:	A person is showing three little kids making faces and being historically.
VX:	一个人正在展示三个小孩做鬼脸和历史。 (A person is showing three kids making funny faces and history.)
ZH:	一个人正在展示历史上的三个做鬼脸的孩子。 (A person is showing three grimacing children in history.)
PE:	三个孩子在电脑前，一个男孩的椅子翻了。 (Three children were in front of the computer, and one boy's chair turned over.)

Table 1: Example of human evaluation. "VX" shows the translation from the VATEX dataset; "ZH" is the human translation; and "PE" is the post-edited Chinese translation. The parenthesized sentences are obtained using Google Translation on Chinese sentences.

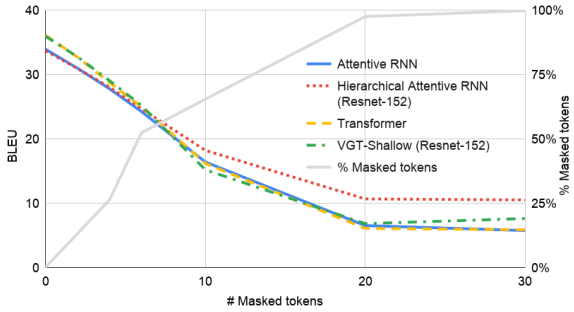

Figure 2: Progressive masking: the multimodal models outperform the monomodal models with increasing percentages of masked tokens.

and selected two English video descriptions from each video's parallel translation pairs in random order. Therefore, a total of 1000 instances exist in the human evaluation set; each instance has an English video description and a video URL.

Given an instance from the human evaluation set, a human translator is first asked to perform an English-Chinese translation, then watch the video, and post-edit the translation only if it does not properly describe the video. We also asked the translator to provide reasons for the post-editing, as well as any remarks. Four professional translators were recruited, two for the translation and two for post-editing, to guarantee the translation quality.

The average editing distance between original Chinese translations from VATEX and human translations is 13.3, which indicates that the annotation strategy of VATEX cannot generate as high-quality translation as a human translator. The translator post-edited 104 Chinese translations, 10.4% of the total number of instances, with an average editing distance of 5.1, compared with human translations, and 14.5 compared with original translations from the VATEX dataset. We found that 98% of post-edits are categorized as "Source English description is inaccurate" which means the translators actually used information gleaned from videos to correct the wrong description in English source sentences.

Based on the above results, we found that in most cases, source English sentences provide sufficient information for the human translator to perform translations; videos help to correct incorrect descriptions in the source sentences rather than to disambiguate translations. These findings also indicate that the short and simple sentences in the VATEX dataset are sufficient for translation purposes, and their videos only provide rather redundant information, which is consistent with the automatic metrics evaluations.

Table 1 shows an example of post-editing correction. The edit distance between the Chinese video description in VATEX and our human post-editing is 15. In the source English descriptions, "being historically" is an ambiguous and incorrect phrase, and we also cannot align it with any parts in the video. Therefore, the translator changed most parts of the sentence to align with the video content.

8 Conclusion

In this study, we probe textual modality dominance and the contributions of visual modality in VMT tasks by analyzing a large-scale VMT dataset: VATEX. Results from input-degradation and visual sensitivity experiments indicate that multimodal models tend to ignore the visual modality when textual modality has sufficient information to perform translation. We ascribe the experimental results to the simple and short video descriptions in VATEX, providing sufficient information to accomplish translation, eventually preventing the visual context from videos to engage in the translation process.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 19H01118. The research results have been achieved by "Research and Development of Deep Learning Technology for Advanced Multilingual Speech Translation", the Commissioned Research of National Institute of Information and Communications Technology (NICT), Japan.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations*.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4159–4170, Minneapolis, Minnesota. 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 (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.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978, Brussels, Belgium. 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.

Jack Hessel and Lillian Lee. 2020. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 861–877, Online. Association for Computational Linguistics.

Tosho Hirasawa, Zhishen Yang, Mamoru Komachi, and Naoaki Okazaki. 2020. Keyframe segmentation and positional encoding for video-guided machine translation challenge 2020. *arXiv preprint arXiv:2006.12799*.

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)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

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*, pages 5998–6008.

A Dataset Statistics

Table 2 shows the statistics of dataset that we obtained. Due to some of the YouTube URLs are invalid, we exclude these data from

B Experiment Results

Table 3 shows the corpus-level BLEU scores of each model in Figure 1.

C Hyper-parameters of baseline models

C.1 Attentive NMT

The attentive RNN model has a 2-layer bidirectional GRU encoder with 512 hidden dimensions, a 2-layer GRU decoder with 512 hidden dimensions, and an embedding layer of 1024 dimensions. During training, we used the Adam optimizer with a learning rate of 0.0004, and batch size of 64 sentences.

C.2 Transformer

The Transformer model in our experiment is the same as the Transformer base model in Vaswani et al. (2017). We employed the same training strategy as Vaswani et al. (2017) to train our Transformer baseline.

D Video-guided NMT

Video-guided Transformer (VGT) models were proposed at ALVR 2020. One of VGT models, VGT-Shallow, achieved the best performance as a single model system. Although the author provided a strong baseline, the paper[6] is not publicly available. However, the video about their models is available[7], and we implemented their model based on the video.

The original work provides two variants of VGT models: VGT-Shallow and VGT-Deep. The vanilla Transformer encodes an n-tokens source sentence $x = (x_1, \cdots, x_n)$, into the hidden state $\mathbf{h} = (\mathbf{h}_1, \cdots, \mathbf{h}_n)$, before decoding the m-tokens target sentence $y = (y_i, \cdots, y_m)$ from $\mathbf{h}$. In the VGT models, the model employs an auxiliary modality $\mathbf{z} = (\mathbf{z}_1, \cdots, \mathbf{z}_k)$ of k-elements, which interact with the language modality in their encoder. Note that both VGT-Shallow and VGT-Deep have a standard Transformer decoder.

[6]"Depper Is Not Always Better: Strong Baselines for Video-guided Neural Machine Translation"

[7]https://www.youtube.com/watch?v=zHwXPmIQajA&t=517s

Split	Language	Video	Sent.	Token
Train	English	24K	121K	1,986K
	Chinese			2,891K
Valid	English	2.8K	15K	228K
	Chinese			331K

Table 2: The statistics of VATEX dataset that we used in the experiments.

In the preliminary experiments, we found that VGT-Shallow outperformed VGT-Deep, which was also reported by the the original author. Therefore, we only adopted VGT-Shallow in our experiments.

D.1 VGT-Shallow

The VGT-Shallow first encodes an input sentence using a standard Transformer encoder to retrieve the hidden state $\mathbf{h}$. Subsequently, the model employs a single fusion layer that has one visual reconstruction module, one cross-modal multi-head attention module, and one element-wise weighted sum module. Note that we exploit normalization and residual connection between modules. Specifically, the visual reconstruction module uses multi-head attention to reconstruct the auxiliary features.

$$\mathbf{h}_{\mathrm{r}}^{'} = \mathrm{multihead}_{\mathrm{r}}(\mathbf{z}, \mathbf{h}, \mathbf{h}) \quad (1)$$

where $\mathrm{multihead}_{\mathrm{r}}$ is a multi-head attention module.

The obtained reconstructed feature $\mathbf{h}_{\mathrm{r}}^{'}$ is then fed into the cross-modal attention module:

$$\mathbf{h}_{\mathrm{x}}^{'} = \mathrm{multihead}_{\mathrm{x}}(\mathbf{h}, \mathbf{h}_{\mathrm{r}}^{'}, \mathbf{h}_{\mathrm{r}}^{'}) \quad (2)$$

where $\mathrm{multihead}_{\mathrm{x}}$ is a multi-head attention module.

Finally, the fusion layer computes the element-wise sum over $\mathbf{h}$ using $\mathbf{h}_{\mathrm{x}}^{'}$ as the weight to obtain the final multimodal representation $\mathbf{h}^{'}$:

$$\mathbf{h}^{'} = \mathbf{h}_{\mathrm{x}}^{'} \odot \mathbf{h} \quad (3)$$

The model decodes the target sentence using $\mathbf{h}^{'}$ instead of $\mathbf{h}$.

Models	T	$\mathrm{T}_{\mathrm{Incongruence}}$	$\mathrm{T}_{\mathrm{Color}}$	$\mathrm{T}_{\mathrm{Noun}}$	$\mathrm{T}_{\mathrm{Verb}}$
Attentive RNN	33.92	33.92	33.93	20.99	30.33
Hierarchical Attentive RNN (I3D)	33.49*	32.75*	33.34*	23.03*	30.21
Hierarchical Attentive RNN (ResNet-152)	33.68*	33.37*	33.71*	23.18*	30.26
Transformer	36.10	36.10	**35.92**	23.01	**32.01**
VGT-Shallow (I3D)	**36.18**	**36.17**	35.83	22.80*	31.84*
VGT-Shallow (ResNet-152)	35.93*	35.48*	35.59*	**24.50***	31.67*

Table 3: Corpus-level BLEU scores on validation set. * indicates that a model is significantly different from its text-only counterpart with P-value $<= 0.05$. **bold** marks the model with the best BLEU score.