Video-guided Machine Translation with Spatial Hierarchical Attention Network Encoder

Weiqi Gu Haiyue Song Chenhui Chu Sadao Kurohashi
Kyoto University, Kyoto, Japan
{gu, song, chu, kuro}@nlp.ist.i.kyoto-u.ac.jp

1 Introduction

Neural machine translation (NMT) has achieved high performance for domains where there is almost no ambiguity in data such as newspaper domain [1, 2]. However, for other domains such as spoken language or sports commentary, the ambiguity in data still remains a problem.

Multimodal machine translation (MMT) [3] is one of the key tasks focusing on incorporating multimodal content as auxiliary information sources to resolve the ambiguity, such as audio or visual data. MMT models usually take a sentence in the source language with the corresponding visual data and translate it into a sentence in the target language. Recent studies [4] assume that the spatiotemporal context information in the visual data helps to reduce the ambiguity of objects or motions in the source text data.

Previous MMT works mainly focus on Image-guided Machine Translation (IMT) task on the widely used Multi30K [5] dataset. However, video is a better information source than image because videos usually contain much more information than images. One video contains ordered sequence of frames and provides rich visual features. For each frame, it provides spatial representations for object disambiguation as an image in IMT task. Besides object disambiguation in one frame, the ordered sequences of frames can provide temporal representations for motion disambiguation.

Video-guided Machine Translation (VMT) aims to engage video data and text data for high-quality translation. Due to the lack of datasets, VMT received less attention than IMT. To cope with this problem, Wang et al. [6] collect a new large-scale and reasonable-quality multilingual video description dataset (VATEX). Each video in the dataset contains hundreds of frames and it is impractical to utilize all objects information from all frames. Existing works only used features from pretrained action detection models as temporal representations of the video to solve the motion ambiguity, thus the object ambiguity still remains a problem. As shown in Fig 1, the object 'picker' and 'bin' in English are wrongly translated into 'apple' and 'trash bin' in Chinese, which are mistranslations partially due to the object ambiguity.

Original: An apple picker takes apples from the trees and places them in a bin.
Translation: 一个苹果苹果从树上摘下苹果， 然后把它们放在一个垃圾桶里。(An apple apple takes apples from the trees and places them in a trash bin.)

Figure 1 An example of the object ambiguity problem

In this work, we propose our VMT system by using both temporal and spatial representations in a video to cope with both the motion ambiguity problem as well as the object ambiguity problem. To obtain spatial features efficiently, we propose to use a hierarchical attention network (HAN) [7] encoder to model the spatial information from object-level to video-level. The HAN framework mainly contains 2 layers for object-to-frame level and frame-to-video level abstractions, a transformer encoder layer [8] is also adopted between 2 layers to obtain contextual spatial information. Experiments on VATEX dataset show 0.2 BLEU score improvement over a strong baseline method.

2 Related Work

We introduce different kinds of auxiliary information used in MMT in this section. Pretrained image features are widely used in the initial attempts of IMT, such as using them to initializing the hidden states of the encoder and/or the decoder [9]. ResNet-50 CNN-based image classifier and information extracted from automatic object detectors shows better performance on IMT tasks [10]. Wang et

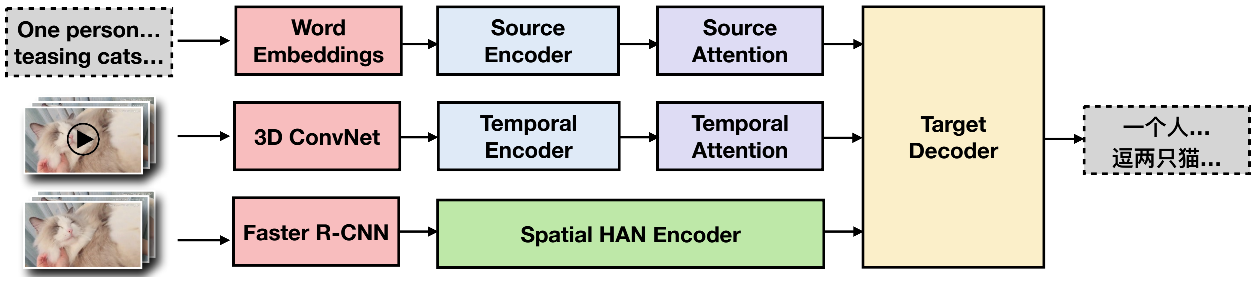

Figure 2 Proposed model with spatial HAN encoder. The source encoder and the temporal encoder are the same as in VMT baseline model, we concatenate them with our proposed spatial HAN encoder by similar target decoder.

al. [6] introduce a strong baseline model which employs the pretrained I3D model [11] for action recognition to get the motion representation while combines attention mechanism [12]. The model combining keyframes information through keyframe selection algorithm and position information of the ordered sequence of frames in a video further improves the translation quality [13].

Besides the action representation which solves the motion ambiguity, spatial information from a sequence of frames in a video could solve the problem of the object ambiguity. Therefore, we propose a novel model with a spatial HAN encoder in addition to the action detection encoder.

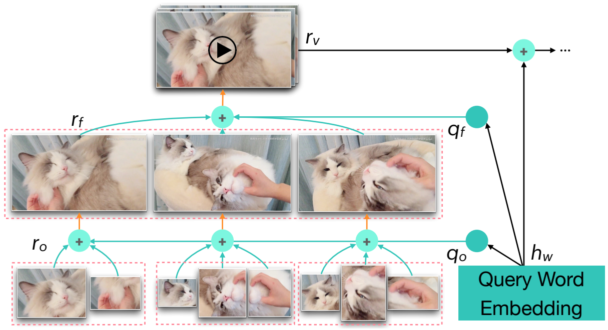

Figure 3 Structure of spatial HAN encoder. r denotes *representation* on object, frame and video levels, q denotes *query* in attention layers, h_w denotes the hidden state of the word embedding for query.

3 VMT with Spatial HAN Encoder

Figure 2 shows the overview of the proposed model. It consists of components in VMT baseline model and our proposed spatial HAN encoder. The temporal representations only provide motion disambiguation. An additional spatial encoder can provide object disambiguation. We first introduce the VMT baseline model in section 3.1. We then introduce our proposed spatial HAN encoder in section 3.2.

3.1 VMT Baseline Model

Wang et al. [6] provide a strong VMT model for the VATEX dataset related tasks. We directly use this model as our VMT baseline model. VMT baseline model mainly consists of the following three modules:

Source Encoder. Each source sentence is represented as a sequence of N word embeddings E, the Bi-LSTM [14] encoder transforms it into the sentence features $U = \{u_1, u_2, ..., u_N\}$.

Temporal Encoder. The authors use a pretrained I3D model [11] for action recognition to obtain the visual features X, then they employ a Bi-LSTM [14] temporal encoder to transform X into the motion features $M = \{m_1, m_2, ..., m_N\}$.

Target Decoder. The sentence embedding from the source language encoder and the motion embedding from the temporal encoder are concatenated and fed into the target language decoder with two attention mechanisms [15].

3.2 Spatial HAN Encoder

Besides temporal encoder and source sentence encoder in the VMT baseline model, our proposed model contains an additional spatial encoder. The intuition is that the temporal encoder only provides motion disambiguation. And an additional spatial encoder can provide object disambiguation.

After splitting one video into N frames, we extracted the object-level spatial features $S = \{s_1, s_2, ..., s_N\}$ of N frames by Faster R-CNN [16], organized them with video ID.

HAN [7] framework can capture context and inter-sentence connections for translation. We propose to use a spatial encoder with HAN framework, which can extract contextual spatial information from adjacent frames within one video clip. The overview is shown in Figure 3. The object-level attention layer summarizes information from all separated objects in their respective frames.

$$q_o = l_o(h_w) \tag{1}$$

$$r_f = f_t(\text{SoftAttention}(q_o, r_o)) \tag{2}$$

where h_w denotes a hidden state of current word embedding. The function l_o is a linear layer to obtain the query q_o. We adopt a soft-dot attention [15] to transform object-level spatial features r_o into respective frame-level spatial features. Then, We obtain contextual frame-level spatial features r_f from a transformer encoder layer f_t [8].

The frame-level attention layer then summarizes representations from all ordered frames to video-level abstraction r_v:

$$q_f = l_f(h_w) \tag{3}$$

$$r_v = \text{SoftAttention}(q_f, r_f) \tag{4}$$

where l_o is a linear transformation, q_f is the query for softdot attention function.

3.3 Target Decoder with Spatial HAN Features

Because we have additional contextual spatial HAN features for the VMT task, the target decoder contains 3 kinds of inputs. We use attention mechanism [15] for both sentence embedding U from the source language encoder and the motion embedding M from the temporal encoder to obtain sentence representations r_u and motion representations r_m:

$$r_u = \text{Attention}_u(U) \tag{5}$$

$$r_m = \text{Attention}_m(M) \tag{6}$$

Sentence representations r_u, motion representations r_m and contextual spatial representations r_v are concatenated and fed into the LSTM [14] layer at each decoding step t:

$$y_t, h_t = f_{lstm}([y_{t-1}, r_{u,t}, r_{m,t}, r_{v,t}], h_{t-1}) \tag{7}$$

Where h_t is the hidden state of the target decoder at step t, $r_{u,t}$ are the sentence representations at step t, $r_{m,t}$ are the motion representations at step t and $r_{v,t}$ are the contextual spatial representations at step t. $f_{bi-lstm}$ refers to the LSTM layer.

4 Experiments

4.1 Dataset

We utilize the VATEX [6] dataset for the VMT task. VATEX is built on a subset of action classification benchmark DeepMind Kinetics-600 [17], which consists of 25,991 video clips for training, 3,000 video clips for validation and 6,000 video clips for public test. Each video clip has 5 parallel English-Chinese descriptions for the VMT task. The VATEX dataset only provides bilingual corpus and segment-level temporal motion features, doesn't provide object-level spatial features and original video clips. We recollected 23,707 video clips for training, 2,702 video clips for validation and 5,461 video clips for public test, about 10% are no longer available, which means we lack 10% spatial features in the dataset.

4.2 Settings

For the common settings in our proposed approach and in the VMT baseline model [6], we set the maximum sentence length to 40, word embedding size to 1,024, and the source encoder and temporal encoder of both 2-layer bi-LSTM with hidden dimension of 512. For our proposed spatial HAN encoder, both object-level and frame-level attention layer are soft-dot attention layer with a hidden dimension of 512. The number of layers in the transformer encoder is 6. Each layer uses multi-head attention with 8 heads and a hidden dimension of 512. The decoder is a 2-layer LSTM of hidden dimension 1,536. During training, we use Adam optimizer with a learning rate of 0.001. The vocabulary size is 10,523 for English and 2,907 for Chinese.

For the baseline model, we adopt text only score and baseline score. Here 'text only' means we only use source encoder, 'baseline' means we use both source encoder and temporal encoder in the VMT model. We retest these scores with the same experiment setting in the baseline model.

4.3 Results

We adopt corpus-level BLEU-4 score as our evaluation metric. Table 1 shows the scores of each model on the validation set and the public test set. Our proposed VMT

ex. 1: **Chinese (Source):** 一个人玩弄着一个黄色头灰色羽毛的鹦鹉。
English (Target): A man plays with a parrot with a yellow head and gray feathers.
VMT baseline: a person is playing with a <UNK> that has a yellow yellow brown brown .
Our method: a person plays with a bird that has a yellow head on it .

ex. 2: **Chinese (Source):** 一名中年男子打开冰箱, 将物品取出。
English (Target): A middle aged man opens up a fridge and beings to remove items.
VMT baseline: a middle aged man opens a refrigerator and takes it out .
Our method: a middle aged man opens a refrigerator and removes the items out .

ex. 3: **Chinese (Source):** 一个人打开一扇门，显示浴缸和洗手台。
English (Target): A door is opened showing a bathtub and bathroom sink.
VMT baseline: a person opens a door and shows the bathtub and washing the phone .
Our method: a person opens a door and shows the bathtub and washes .

ex. 4: **Chinese (Source):** 一个穿红色衣服的小女孩正在打开一件大礼物。
English (Target): A little girl in a red dress is unwrapping a large present.
VMT baseline: a little girl in a red dress is opening a large present .
Our method: a little girl is opening a large present and is talking to her .

Figure 4 Four examples from Chinese to English translation. Ex. 1: VMT baseline model gives object omission error, which leads to structural errors. Ex. 2: There is object ambiguity problem in the VMT baseline method. Ex. 3: A wrong object translation in the VMT baseline model. In the above 3 examples, our method has correct object and description translations. Ex. 4 is a wrong translation in our method, where some noise information from the video affected the translation.

Table 1 Corpus-level BLEU-4 scores of English to Chinese translation.

Model	Valid	Test
Text only	29.6	29.6
VMT baseline	30.6	31.1
VMT with spatial HAN encoder	**31.2**	**31.3**

Table 2 50 translation examples from VMT baseline model and proposed method. We notice that most errors are from the object ambiguity and omission problem.

	Baseline	Our Method
Correct	31	**37**
Incorrect	19	13

system with spatial HAN encoder achieves 31.2 score on the validation set and 31.3 score on the public test set, showing 0.2 BLEU score improvement over the VMT baseline model.

Because the reference sentences in public test set are hidden, we divide the former half of the original validation set into a new validation set and the latter half into a new test set to analyze the details of translation results. We train on the newly divided dataset, and compare the results on the new test set. We analyze 50 examples randomly selected from the test set to observe whether our model can translate sentences successfully. The results are shown in Table 2, our method has 6 more correct translations than the VMT baseline model. Figure 4 shows the details of several example analyses from Chinese to English. We observed that our method can alleviate object ambiguity and omission problem in the translation, but sometimes the auxiliary information from video clips may result in wrong translations.

5 Conclusion

In this work, we propose a VMT system with spatial HAN encoder, which achieves a 0.2 BLEU score improvement over a strong VMT baseline model. The result shows the effectiveness of spatial features for object disambiguation. Our future work will focus on VMT baseline modification, especially the alignment between source, temporal and spatial representations.

Acknowledgments

This work was supported by ACT-I, JST.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv e-prints*, p. arXiv:1409.0473, September 2014.

[2] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, Vol. abs/1609.08144, , 2016.

[3] Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pp. 543–553, Berlin, Germany, August 2016. Association for Computational Linguistics.

[4] Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pp. 304–323, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[5] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pp. 70–74. Association for Computational Linguistics, 2016.

[6] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 4580–4590. IEEE, 2019.

[7] Lesly Miculicich Werlen, Dhananjay Ram, Nikolaos Pappas, and James Henderson. Document-level neural machine translation with hierarchical attention networks. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 2947–2954. Association for Computational Linguistics, 2018.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017.

[9] Desmond Elliott, Stella Frank, and Eva Hasler. Multilingual image description with neural sequence models, 2015.

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

[11] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 4724–4733. IEEE Computer Society, 2017.

[12] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[13] Tosho Hirasawa, Zhishen Yang, Mamoru Komachi, and Naoaki Okazaki. Keyframe segmentation and positional encoding for video-guided machine translation challenge 2020.

[14] Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.*, Vol. 45, No. 11, pp. 2673–2681, 1997.

[15] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[16] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and VQA. *CoRR*, Vol. abs/1707.07998, , 2017.

[17] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, Vol. abs/1705.06950, , 2017.