Video-guided Machine Translation with Spatial Hierarchical Attention Network

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

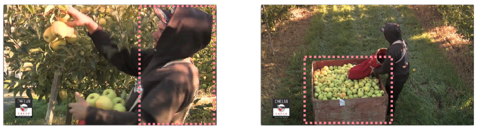
Abstract

Video-guided machine translation, as one type of multimodal machine translations, aims to engage video contents as auxiliary information to address the word sense ambiguity problem in machine translation. Previous studies only use features from pretrained action detection models as motion representations of the video to solve the verb sense ambiguity, leaving the noun sense ambiguity a problem. To address this problem, we propose a video-guided machine translation system by using both spatial and motion representations in videos. For spatial features, we propose a hierarchical attention network to model the spatial information from object-level to video-level. Experiments on the VATEX dataset show that our system achieves 35.86 BLEU-4 score, which is 0.51 score higher than the single model of the SOTA method.

1 Introduction

Neural machine translation (NMT) models relying on text data (Bahdanau et al., 2015; Wu et al., 2016) have achieved high performance for domains where there is less ambiguity in data such as the newspaper domain. For some other domains, especially real-time domains such as spoken language or sports commentary, the verb and the noun sense ambiguity largely affects the translation quality. To solve the ambiguity problem, multimodal machine translation (MMT) (Specia et al., 2016) focuses on incorporating visual data as auxiliary information, where the spatiotemporal contextual information in the visual data helps reduce the ambiguity of nouns or verbs in the source text data (Barrault et al., 2018).

Previous MMT studies mainly focus on image-guided machine translation (IMT) task (Zhao et al., 2020; Elliott et al., 2016). However, videos are better information sources than images because one video contains an ordered sequence of frames and provides much more visual features. Specifically, each frame provides spatial representations for the noun sense disambiguation as an image in IMT task. Besides the noun sense disambiguation provided by one frame, the ordered sequences of frames can provide motion representations for the verb sense disambiguation.



Source: 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 with the noun sense ambiguity problem in the VMT model by Wang et al. (2019).

The research of video-guided machine translation (VMT) starts from a large-scale video-and-language-research dataset (VATEX) (Wang et al., 2019). The authors also established a baseline using features from pretrained action detection models as motion representations of the video, which addresses the verb sense ambiguity to some extent, leaving noun sense ambiguity unsolved. Hirasawa et al. (2020) aims to solve both the verb and noun sense ambiguity problems by using frame-level action, object, and scene representations. However, without using detailed spatial information within one frame and contextual information between frames, the effect of resolving the noun ambiguity problem is limited. For example, as shown in Figure 1, the noun "bin" in English is wrongly translated into "trash bin" in Chinese, which should be translated into "box."

In this work, we propose a VMT system to address both the verb and the noun sense ambiguity

Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop, pages 87–92
August 5–6, 2021. ©2021 Association for Computational Linguistics

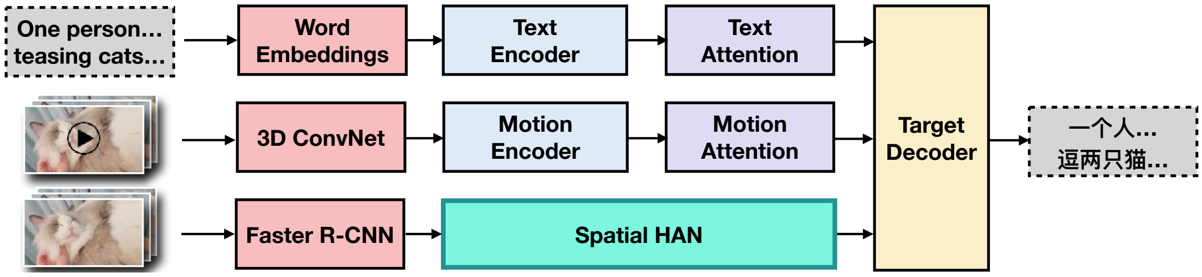


Figure 2: The proposed model with spatial HAN. The text encoder and the motion encoder are the same as those in the VMT baseline model.

problems by using both motion and spatial representations in a video. To obtain spatial representations efficiently, we propose to use a hierarchical attention network (HAN) (Werlen et al., 2018) to model the spatial information from object-level to video-level, thus we call it the spatial HAN module. Additionally, to obtain a better contextual spatial information, we add several kinds of middle layers between the object-to-frame layer and frame-to-video layer in the original HAN. Experiments on the VATEX dataset (Wang et al., 2019) show that our VMT system achieves 35.86 corpus-level BLEU-4 score on the VATEX test set, yielding a 0.51 score improvement over the single model of the SOTA method (Hirasawa et al., 2020).

2 VMT with Spatial HAN

The overview of the proposed model is presented in Figure 2, which consists of components in the VMT baseline model (Hirasawa et al., 2020) and our proposed spatial HAN module.

2.1 VMT Baseline Model

Hirasawa et al. (2020) proposed a strong VMT baseline model, which consists of the following three modules.

Text Encoder. Each source sentence is represented as a sequence of N word embeddings. Then, the Bi-GRU (Schuster and Paliwal, 1997) encoder transforms them into text features $U = \{\mathbf{u_1}, \mathbf{u_2}, ..., \mathbf{u_N}\}$.

Motion Encoder. The VATEX dataset already provides motion features obtained by the pretrained I3D model (Carreira and Zisserman, 2017) for action recognition. A Bi-GRU motion encoder first transforms motion features into motion representations $M = \{\mathbf{m_1}, \mathbf{m_2}, ..., \mathbf{m_P}\}$. Then, a positional encoding (PE) layer PE (Vaswani et al., 2017) encourages the model use the order of the motion features and obtain ordered motion representations M^*, represented as:

$$M^* = \mathrm{PE}(M) \tag{1}$$

Target Decoder. The sentence embedding U from the source language encoder and the ordered motion embedding M^* from the motion encoder are processed using two attention mechanisms (Luong et al., 2015):

$$\mathbf{r_{u,t}} = \mathrm{Attention}_{u,t}(\mathbf{h_{t-1}}, U) \tag{2}$$

$$\mathbf{r_{m,t}} = \mathrm{Attention}_{m,t}(\mathbf{h_{t-1}}, M^*) \tag{3}$$

where $\mathrm{Attention}$ denotes a standard attention block, $\mathbf{h_{t-1}}$ denotes the hidden state at the previous decoding time step. Text representations $\mathbf{r_{u,t}}$ and motion representations $\mathbf{r_{m,t}}$ are allocated by another attention layer to obtain a contextual vector $\mathbf{r_{c,t}}$ at decoding time step t. The contextual vector is fed into a GRU layer for decoding:

$$\mathbf{r_{c,t}} = \mathrm{Attention}(\mathbf{h_{t-1}}, [\mathbf{r_{u,t}}, \mathbf{r_{m,t}}]) \tag{4}$$

$$\mathbf{y_t}, \mathbf{h_t} = \mathrm{f_{gru}}([\mathbf{y_{t-1}}, \mathbf{r_{c,t}}], \mathbf{h_{t-1}}) \tag{5}$$

where $\mathrm{f_{gru}}$ refers to the GRU decoding layer and $\mathbf{y}$ denotes the output target word embedding.

2.2 Spatial HAN

After splitting one video into X frames, we extract Y object-level spatial features $S_i = \{\mathbf{o_1}, \mathbf{o_2}, ..., \mathbf{o_Y}\}$ for the i-th frame. Because of the effectiveness of the PE layer (Vaswani et al., 2017) in the VMT baseline model, we also apply it to the object-level spatial features.

$$[R_o^1, R_o^2, ..., R_o^X] = \mathrm{PE}([S_1, S_2, ..., S_X]) \tag{6}$$

R_o^i denotes the object-level spatial representations of i-th frame.

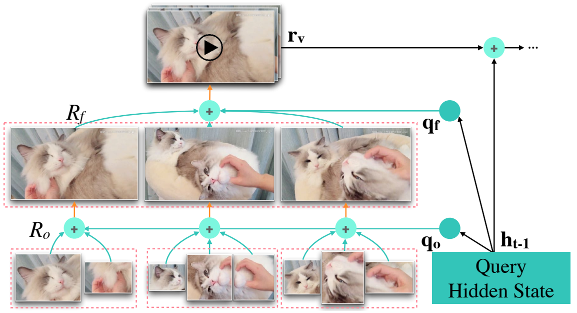


Figure 3: Structure of spatial HAN. R_o, R_f and $\mathbf{r_v}$ denote object-level, frame-level and video-level representations, $\mathbf{q}$ denotes *query* in attention layers, $\mathbf{h_{t-1}}$ denotes the hidden state at previous decoding time step.

Werlen et al. (2018) show that HAN can capture contextual and inter-sentence connections for translation. We propose to use HAN to extract contextual spatial information from adjacent frames within one video clip. With some modifications, we call the network spatial HAN.

The overview of spatial HAN is given by Figure 3. The object-level attention layer summarizes information from all separated objects in their respective frames:

$$\mathbf{q_{o,t}} = \mathrm{l_o}(\mathbf{h_{t-1}}) \tag{7}$$

$$\mathbf{r^i_{f,t}} = \mathrm{Attention_{o,t}}(\mathbf{q_{o,t}}, R^i_o) \tag{8}$$

$$R_{f,t} = \{\mathbf{r^1_{f,t}}, \mathbf{r^2_{f,t}}, ..., \mathbf{r^X_{f,t}}\} \tag{9}$$

$$R^*_{f,t} = \mathrm{f_d}(R_{f,t}) \tag{10}$$

where the function $\mathrm{l_o}$ is a linear layer to obtain the query $\mathbf{q_{o,t}}$. We adopt an attention layer to transform object-level spatial features R^i_o into respective frame-level spatial features $\mathbf{r^i_{f,t}}$. $\mathrm{f_d}$ denotes the middle encoding layer to obtain contextual frame-level spatial features $R^*_{f,t}$ at time step t.

The frame-level attention layer then summarizes representations from all ordered frames to video-level abstraction $\mathbf{r_{v,t}}$:

$$\mathbf{q_{f,t}} = \mathrm{l_f}(\mathbf{h_{t-1}}) \tag{11}$$

$$\mathbf{r_{v,t}} = \mathrm{Attention_{o,t}}(\mathbf{q_{f,t}}, R^*_{f,t}) \tag{12}$$

where $\mathrm{l_f}$ is a linear transformation, $\mathbf{q_{f,t}}$ is the query for attention function at time step t.

2.3 Target Decoder with Spatial HAN Features

The target decoder in our system contains three types of inputs: text representations $\mathbf{r_{u,t}}$, motion representations $\mathbf{r_{m,t}}$, and contextual spatial representations $\mathbf{r_{v,t}}$ from spatial HAN. The contextual vector $\mathbf{r_{c,t}}$ and the decoding GRU layer at each decoding step t become:

$$\mathbf{r_{c,t}} = \mathrm{Attention}(\mathbf{h_{t-1}}, [\mathbf{r_{u,t}}, \mathbf{r_{m,t}}, \mathbf{r_{v,t}}]) \tag{13}$$

$$\mathbf{y_t}, \mathbf{h_t} = \mathrm{f_{gru}}([\mathbf{y_{t-1}}, \mathbf{r_{c,t}}], \mathbf{h_{t-1}}) \tag{14}$$

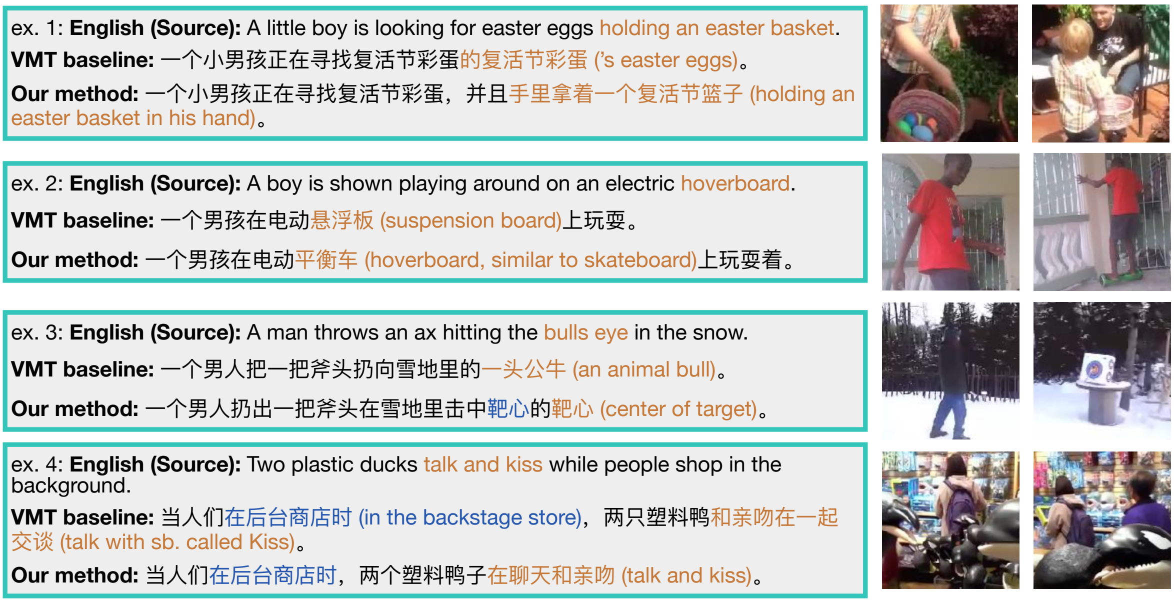
3 Experiments

3.1 Dataset

The dataset we used for the VMT task is VATEX, which is built on a subset of action classification benchmark DeepMind Kinetics-600 (Kay et al., 2017). It 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 is accompanied with 5 parallel English-Chinese descriptions for the VMT task. However, the VATEX dataset only contains parallel sentences and segment-level motion features. To extract spatial features, we recollected $23,707$ video clips for training, $2,702$ video clips for validation, and $5,461$ video clips for public test, where about 10% clips are no longer available on the Internet. Therefore, we lack 10% spatial features for the dataset, so the experiment comparison is inherently unfair for our method.

3.2 Settings

We directly used the implementation of Hirasawa et al. (2020) as our VMT baseline model. For the common settings in our proposed approach and in the VMT baseline model, we set the maximum sentence length to 40, word embedding size to $1,024$, and the text encoder and motion encoder of both 2-layer bi-GRU with hidden dimension of 512. For our proposed spatial HAN, we used Faster R-CNN (Anderson et al., 2017) to extract object-level features as the input. The hidden dimensions of both object-level and frame-level attention layers were 512. As for the middle layer $\mathrm{f_d}$ in spatial HAN, we examined GRU and LSTM with the hidden dimension of 512, and spatial HAN without the middle layer. The target decoder was a 2-layer GRU with the hidden dimension of 512. During training, we used Adam optimizer with a learning rate of 0.001 and early stop with patience to 10 times. The vocabulary contained lower-cased English and characterized Chinese tokens that occurred more than five times in the training set, which is provided by Hirasawa et al. (2020) whose sizes are $7,949$ for English and $2,655$ for Chinese.



ex. 1: **English (Source):** A little boy is looking for easter eggs holding an easter basket.

VMT baseline: 一个小男孩正在寻找复活节彩蛋的复活节彩蛋 ('s easter eggs)。

Our method: 一个小男孩正在寻找复活节彩蛋，并且手里拿着一个复活节篮子 (holding an easter basket in his hand)。

ex. 2: **English (Source):** A boy is shown playing around on an electric hoverboard.

VMT baseline: 一个男孩在电动悬浮板 (suspension board)上玩耍。

Our method: 一个男孩在电动平衡车 (hoverboard, similar to skateboard)上玩耍着。

ex. 3: **English (Source):** A man throws an ax hitting the bulls eye in the snow.

VMT baseline: 一个男人把一把斧头扔向雪地里的一头公牛 (an animal bull)。

Our method: 一个男人扔出一把斧头在雪地里击中靶心的靶心 (center of target)。

ex. 4: **English (Source):** Two plastic ducks talk and kiss while people shop in the background.

VMT baseline: 当人们在后台商店时 (in the backstage store)，两只塑料鸭和亲吻在一起交谈 (talk with sb. called Kiss)。

Our method: 当人们在后台商店时，两个塑料鸭子在聊天和亲吻 (talk and kiss)。

Figure 4: Four English to Chinese translation examples. Phrases in orange imply corresponding information and phrases in blue imply other translation errors. Ex. 1, 2 and 3 display noun sense ambiguity errors generated by the VMT baseline that make the translation unreasonable, whereas our model correctly translates these noun phrases. Ex. 4 shows a sentence structure error in the VMT baseline output, where the model wrongly recognizes the verb as the noun.

We adopt corpus-level BLEU-4 as the evaluation metric. We reported the score of the VMT baseline model denoted as "VMT baseline: Text+Motion," naming that it uses both the text and motion encoders. Besides the experiments with text, motion and spatial features obtained by our methods, denoted as "Ours: Text+Motion+Spatial," we also conducted the experiments with only text and spatial features denoted as "Ours: Text+Spatial."

3.3 Results

Model	**Valid**	**Test**
Wang et al. (2019)	-	31.10
Hirasawa et al. (2020)	35.42	35.35
VMT baseline: Text+Motion	35.55	35.59
Ours: Text+Motion+Spatial	35.71	35.82
Ours: Text+Spatial	**35.75**	**35.86**

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

Table 1 shows the results of baseline and proposed models on the validation and public test sets. Our proposed system achieves 35.75 score on the validation set and 35.86 score on the test set, showing 4.76 BLEU score improvement over the VATEX's baseline model (Wang et al., 2019), and 0.51 BLEU score improvement over the best single model with the text corpus and action features. Because of some different settings in hyperparameters, our VMT baseline has 0.24 BLEU score improvement over the best single model.

Model	**Mid Layer**	**Valid**
Text+Motion+Spatial	None	**35.71**
	LSTM	35.50
	GRU	35.54
Text+Spatial	None	**35.75**
	LSTM	35.37
	GRU	35.27

Table 2: BLEU-4 scores of our models with different settings and middle layer choice.

Table 2 shows the ablation study on different settings of middle layer choice. Without the middle layer, both the two models achieved the best validation score. The reason may be that the PE layer for object-level spatial features already provides the contextual information, thus the middle layer in spatial HAN is dispensable. We notice that our models achieve comparable BLEU score results with and without motion features. We assume that it may come from the misalignment between motion, spatial and text features, where nouns and verbs in the sentences are not aligned to spatial features and motion features strictly. Also, the amount of nouns in sentences are much more than the amount of verbs in sentences, e.g., the ratios of nouns and verbs in source training corpus are 0.29 and 0.17, thus spatial features will take on more roles.

We further conducted a pairwise human evaluation to investigate how our proposed method improves the translation. Results on 50 random samples show that our model has 12 better translations than the VMT baseline model mainly on the noun sense disambiguation, where the VMT baseline model has 6 better translations mainly on the verb sense disambiguation and syntax. This suggests that our model can alleviate the noun sense ambiguity problem. The analysis details of several examples are given by Figure 4.

4 Conclusion

In this work, we proposed a VMT system with spatial HAN, which achieved 0.51 BLEU score improvement over the single model of the SOTA method. The result also showed the effectiveness of spatial features for the noun sense disambiguation. Our future work will focus on the alignment between text, motion and spatial representations.

Acknowledgments

This work was supported by ACT-I, JST.

References

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

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, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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

João Carreira and Andrew Zisserman. 2017. 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*, pages 4724–4733. IEEE Computer Society.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74. 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. *CoRR*, abs/2006.12799.

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. 2017. The kinetics human action video dataset. *CoRR*, abs/1705.06950.

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

Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.*, 45(11):2673–2681.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. 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*, pages 543–553, 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, 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 *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4580–4590. IEEE.

Lesly Miculicich Werlen, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2947–2954. Association for Computational Linguistics.

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. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2020. Double attention-based multimodal neural machine translation with semantic image regions. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 105–114, Lisboa, Portugal. European Association for Machine Translation.