

Multilingual and Multimodal Learning for Brazilian Portuguese

Júlia Sato*, Helena Caseli*, Lucia Specia[†]

*Federal University of São Carlos (UFSCar)
Rod. Washington Luiz, s/n, São Carlos, Brazil
juliasato@estudante.ufscar.br, helenacaseli@ufscar.br

[†]Imperial College London
Exhibition Rd, South Kensington, London SW7 2BX, United Kingdom
l.specia@ic.ac.uk

Abstract

Humans constantly deal with multimodal information, that is, data from different modalities, such as texts and images. In order for machines to process information similarly to humans, they must be able to process multimodal data and understand the joint relationship between these modalities. This paper describes the work performed on the VTLM (Visual Translation Language Modelling) framework from (Caglayan et al., 2021) to test its generalization ability for other language pairs and corpora. We use the multimodal and multilingual corpus How2 (Sanabria et al., 2018) in three parallel streams with aligned English-Portuguese-Visual information to investigate the effectiveness of the model for this new language pair and in more complex scenarios, where the sentence associated with each image is not a simple description of it. Our experiments on the Portuguese-English multimodal translation task using the How2 dataset demonstrate the efficacy of cross-lingual visual pretraining. We achieved a BLEU score of 51.8 and a METEOR score of 78.0 on the test set, outperforming the MMT baseline by about 14 BLEU and 14 METEOR. The good BLEU and METEOR values obtained for this new language pair, regarding the original English-German VTLM, establish the suitability of the model to other languages.

Keywords: Multilingual Language Model, Multimodal Machine Translation, Brazilian Portuguese, Vision and Language

1. Introduction

Understanding different modalities together is an important aspect of human comprehension. We often use sight, hearing, smell and other senses to assimilate a single concept. This multimodal aspect of learning can be very useful for machines to process multimodal information and understand the joint relationship between these modalities.

While multimodal models are trained to be able to interpret and associate data from different modalities – such as text, audio and image simultaneously – multilingual models are meant to understand multiple languages by learning cross-lingual representations. In this context, models that learn multimodal and multilingual representations have been shown to perform better in many natural language tasks (Caglayan et al., 2021).

Recently, the area of Natural Language Processing (NLP) has been experiencing a significant paradigm shift with the proposition of several neural models (deep learning) for language processing (Devlin et al., 2019; Conneau and Lample, 2019; Radford et al., 2018). These advances are based on the use of artificial neural networks and strategies such as transfer learning and attention (Calixto et al., 2016; Caglayan et al., 2016; Libovický and Helcl, 2017).

There are several examples of NLP applications for which these strategies have reached state of the art in monolingual (Lan et al., 2020; Liu et al., 2019; Rothe et al., 2020), multilingual (Devlin et al., 2019; Conneau

and Lample, 2019) and multimodal (Tan and Bansal, 2019; Lu et al., 2019; Li et al., 2019; Lin et al., 2021) processing. However, for Portuguese these advances are still very sparse (Souza et al., 2020).

Even though the initial interest was only in multimodal (Tan and Bansal, 2019; Lu et al., 2019; Li et al., 2019) or multilingual (Devlin et al., 2019; Conneau and Lample, 2019) models, recent developments have resulted in frameworks able to do both and deliver multilingual and multimodal models (Caglayan et al., 2021; Huang et al., 2021; Ni et al., 2021).

In this context, visual modality can help machines have a better understanding of textual information. This approach has been introduced in a novel multimodal Neural Machine Translation (MNMT) task (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018), which mainly focuses on enhancing text-only translation with visual features. Therefore, multimodal machine translation (MMT) improves translation quality by using context from the additional visual modality. As a result, the translation is expected to be more accurate, since the visual context helps to reduce ambiguity.

Considering the MMT task, the Visual Translation Language Modelling (VTLM), proposed by Caglayan et al. (2021), showed that multimodal and multilingual pre-training leads to considerable improvements compared to multimodal machine translation without pre-training or only with multilingual pre-training. Thus highlighting the efficiency of both multimodal and multilingual pre-training.

In this paper we report the work we performed with VTLM focusing on the Portuguese language. We used the multimodal and multilingual corpus How2 (Sanabria et al., 2018) and applied the necessary pre-processing steps to verify its performance on the VTLM model for the MMT task.

In this context, this paper is an important step towards tackling the problem of multimodal and multilingual learning involving the Portuguese language. The main contributions of this work are: (i) the generation of new linguistic-computational resources for Brazilian Portuguese, including processing scripts and the processed corpus/database generated for the experiments; (ii) the adaptation of VTLM to a new language pair and to more challenging circumstances in relation to the image-text relationship, as well as (iii) the relevant results of the experimentation.

2. Related Work

Multimodal machine translation, different from the traditional MT, considers other modalities in addition to text information in order to better translate source sentences into target ones.

Previous works propose various multimodal machine translation models and methods. Huang et al. (2016) concatenate global and regional visual features with text in order to attend to the image and the text while decoding. Calixto and Liu (2017) use global image features to initialize the encoder/decoder hidden states of RNN. Elliott and Kádár (2017) introduce a multi-task learning framework to learn visually grounded representations and learn to translate. Zhou et al. (2018) enhance the learning of a shared visual language embedding and a multimodal attention-based translator through a visual attention mechanism. Calixto et al. (2019) put forward a latent variable model to learn the interactions between visual and textual features. Ive et al. (2019) propose a translate-and-refine method based on Transformer (Vaswani et al., 2017), where images are only used by a second stage decoder. Yin et al. (2020) utilize a unified multimodal graph to capture different semantic relationships.

Unlike previous methods, Yao and Wan (2020) propose multimodal self-attention in Transformer to solve the problem of relative importance among different modalities. They show a better approach to incorporate information from other modality based on a graph perspective of Transformer and they avoid encoding irrelevant information in images by learning the representations of images based on the text. Their model is evaluated on the Multi30k dataset (Elliott et al., 2016), which contains 29,000 instances for training, 1,024 for validation and 1,000 for testing (Test2016). The model achieves a METEOR score of 55.7 and a BLEU score of 38.7 on English-German (En-De) Test2016, demonstrating the benefit of the visual modality by outperforming their text-only baseline by above 1 BLEU points.

Along this line, Liu et al. (2021) introduce a selecting method in multimodal scenarios named Gumbel-Attention. It selects the text-related parts of the image features and removes the irrelevant information using a differentiable method. They also use the Multi30k dataset and present their results on the English-German test set. The Gumbel-Attention MMT model obtains a better performance compared to Multimodal Transformer (Yao and Wan, 2020), reaching 39.2 BLEU and 57.8 METEOR on the Multi30k Test2016.

In contrast with prior studies, Long et al. (2021) introduce a machine translation method that only needs the source sentence at the inference time. They create a generative imagination-based model called ImagiT, which learns to produce visual representation from the source sentence, and then generates the target language sentence using source sentence and the “imagined representation”. Similar to previous work, the experiments are conducted on the Multi30k dataset. Their best results are on English-German (En-De) Test2017, achieving a BLEU score of 32.4 and a METEOR score of 52.5, and on English-French (En-Fr) Test2016, obtaining 59.9 BLEU and 74.3 METEOR. Their results show improvements over text-only NMT baselines, demonstrating the efficacy of their model.

Due to the scarcity of quality datasets available for multimodal translation, most current works resort only to the Multi30k dataset, which is a multilingual extension of Flickr30k (Young et al., 2014) with translations of the English image descriptions into German and French. However, following another direction Gupta et al. (2021) use the Visual Genome dataset (Krishna et al., 2016) with Hindi translations to investigate the effectiveness of their multimodal translation system. They use a pretrained multilingual sequence-to-sequence model and fine-tune it on a textual-only dataset consisting of 1,609,682 parallel sentences in English and Hindi (Kunchukuttan et al., 2018). And they bring the visual information to the textual domain by extracting object tags from the image, adding them to the source text and then fine-tuning the model on the training set with the object tags. Their model achieves state-of-the-art performance on the dataset, reaching 44.6 BLEU on the test set and 51.6 BLEU on the challenge set, which consists of sentences where there are ambiguous English words.

In this paper, we chose to work with VTLM from Caglayan et al. (2021). VTLM extends the TLM framework (Conneau and Lample, 2019) with regional features and introduces a pre-training approach that combines cross-lingual and visual pre-training. It performs masked language modeling and masked region classification on a three-way parallel language and vision dataset, which is an extension of the Conceptual Captions corpus (Sharma et al., 2018) with German machine translations. VTLM achieved a BLEU score of 44.0 and a METEOR score of 61.3 on English-German (En-De) 2016 test set of Multi30k for the

MMT task, demonstrating the effectiveness of both multimodal and multilingual pre-training.

3. Method

We work with Visual Translation Language Modelling (Caglayan et al., 2021) adapted to another language pair (Brazilian Portuguese-English) and corpus (How2 (Sanabria et al., 2018)). Therefore, we first describe the VTLM objective and then present the How2 corpus.

3.1. Visual Translation Language Modelling

The VTLM objective combines multimodal and multilingual learning to generate cross-lingual and multimodal representations in order to analyze its effectiveness on the multimodal machine translation task. To accomplish this, the model joins the TLM (Translation Language Modelling), proposed by Conneau and Lample (2019), with masked region classification (MRC) (Chen et al., 2020; Su et al., 2020). VTLM defines the input x as the concatenation of m -length source language sentence $s_{1:m}^{(1)}$, n -length target language sentence $s_{1:n}^{(2)}$, and $\{v_1, \dots, v_o\}$ corresponding image features:

$$x = [s_1^{(1)}, \dots, s_m^{(1)}, s_1^{(2)}, \dots, s_n^{(2)}, v_1, \dots, v_o] \quad (1)$$

The final model combines the TLM loss with the MRC loss according to the following equation:

$$\mathcal{L} = \frac{1}{|X|} \sum_{x \in X} \log Pr(\{\hat{y}, \hat{v}\} | \tilde{x}; \theta) \quad (2)$$

where \tilde{x} is the masked input sequence, \hat{y} are the ground-truth targets for masked positions, \hat{v} are the detection labels and θ are the model parameters.

The VTLM architecture (Figure 1) extends the TLM by adding a visual modality alongside the translation pairs, and the final model processes translation pairs and projected region features in a single-stream.

In this approach, masking is random and applies to textual and visual tokens. Its proportion is 15% and it is applied separately to visual and language flows. VTLM replaces its vector of projected features by the [MASK] token, with 10% of the masking being equivalent to using region features randomly selected from all images in the batch, and the remaining 10% of the regions are left intact.

VTLM pre-training has visual and cross-lingual resources and performs masked language modeling and masked region classification on a three-way parallel language and vision dataset, which is an extension of the Conceptual Captions corpus (CC) (Sharma et al., 2018) with German machine translations.

After pre-training, the VTLM encoder is transferred to a Transformer-based (Vaswani et al., 2017) multimodal machine translation model and adjusted for the MMT task.

3.2. Corpus How2

How2 (Sanabria et al., 2018) is a multimodal and multilingual collection of approximately 80,000 instructional videos (approximately 2,000 hours) accompanied by English subtitles and around 300 hours of collected crowdsourced Portuguese translations, plus summaries of each video in English.

According to the authors, the multimodal nature of How2 improves comprehension as it helps to resolve possible ambiguities that could be found in a text-only setting. For instance, consider the example in Figure 2. The bold text shows the English subtitle for the speech, the italic text corresponds to the aligned Portuguese translation, and the text inside the rectangle is the video summary. In this example, the man is explaining how to play a golf shot and the visual context (green grass with a flagpole) or the audio context (outside with the sound of chipping a golf ball) leads to a correct interpretation of the speech, as only with the text it is not clear whether the “green” in the caption refers to the color green (“verde” in Portuguese), or to the type of surface (“green” in Portuguese).

Therefore, How2 is an important resource for multimodal tasks such as multimodal machine translation.

3.3. VTLM for Video Subtitles

The multimodal and multilingual How2 corpus is used in all stages of experimentation. In addition to language differences, using this corpus for the MMT task brings additional challenges compared to using the Multi30k corpus, as in most previous works (Yao and Wan, 2020; Liu et al., 2021; Long et al., 2021; Caglayan et al., 2021). The fundamental reason is that How2 is a collection of videos – which means that the text associated with each image (video frame) is not a simple description of it, but a subtitle that may not be related to its corresponding frame due to the constant motion of a video – while Multi30k is a collection of static images, that is, each sentence has a single image that is semantically aligned with the sentence (the sentence is a description of the image). As a consequence, multimodal machine translation using How2 is considerably more difficult compared to multimodal machine translation using Multi30k.

Further, another pertinent challenge is to make the VTLM model work under these circumstances. As VTLM has as input the translation pairs and projected region features, it requires a three-way parallel multimodal corpus. Therefore, we use the How2 corpus with aligned English-Portuguese-Visual information.

In this context, it was necessary to have object features and English and Portuguese texts. The video frames from How2 and their corresponding bilingual subtitles were made available by Sanabria et al. (2018), so the next step was the feature extraction. This process is illustrated in Figure 3.

As there were between two to fifteen frames corresponding to a single caption, we had to select only one

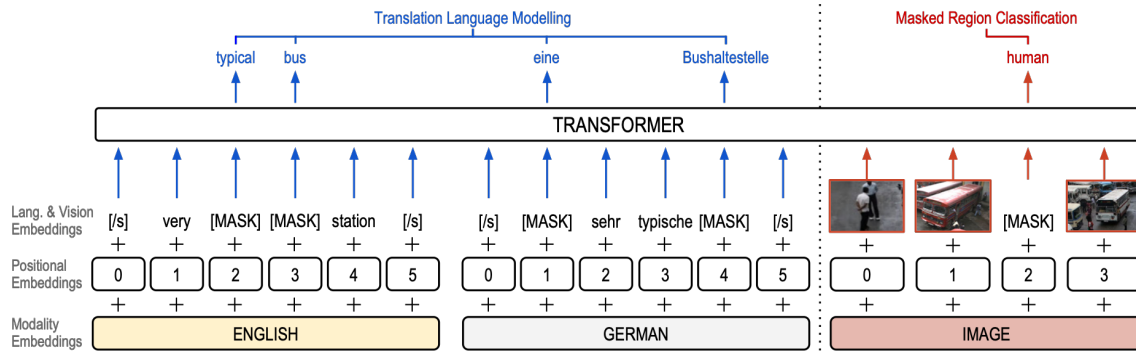


Figure 1: VTLM architecture from (Caglayan et al., 2021).

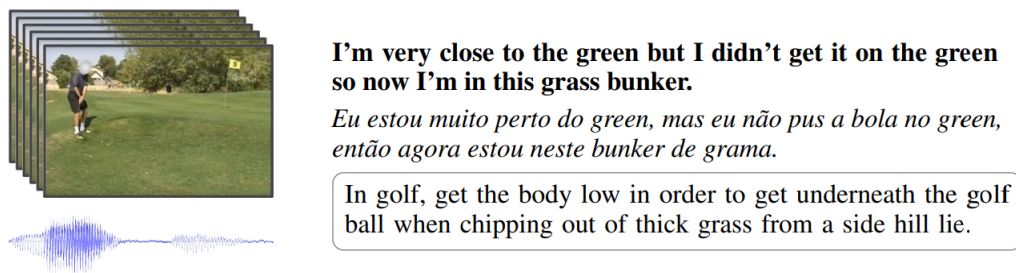


Figure 2: How2 example from (Sanabria et al., 2018).

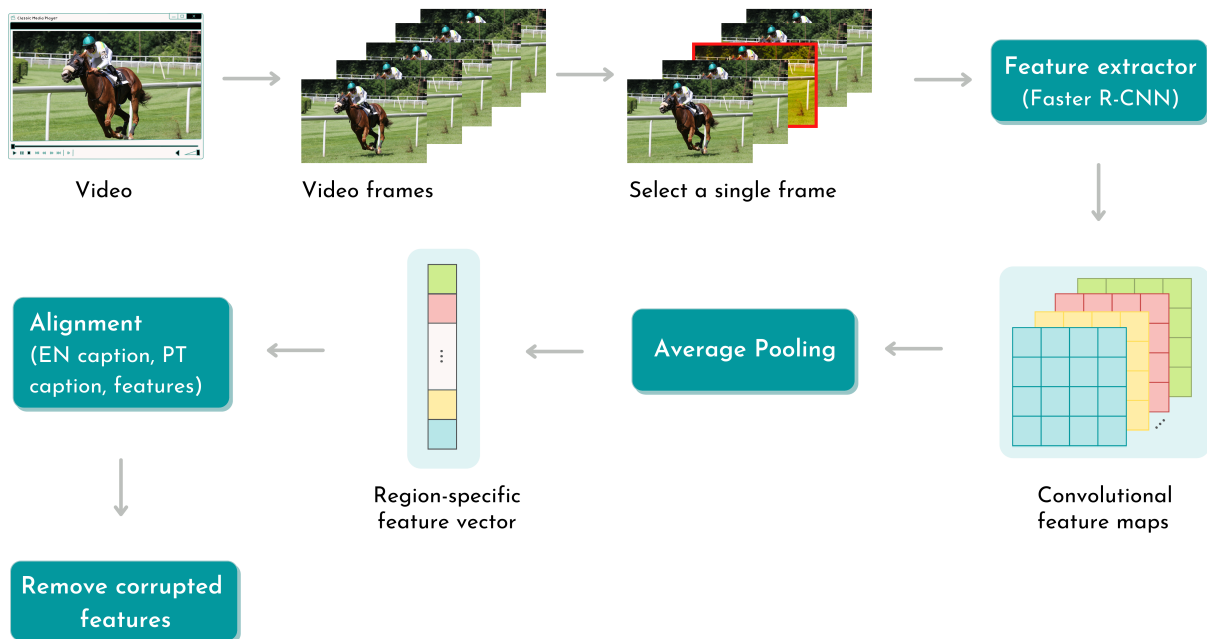


Figure 3: Pre-processing steps

of them – a single image – to perform the feature extraction. Therefore, for each video segment, the middle frame was selected and convolutional feature maps were extracted from the 36 most reliable regions using the Faster R-CNN model (Ren et al., 2015) pre-trained on the Open Images dataset (Kuznetsova et al., 2020).

Finally, each feature map was average pooled to obtain a region-specific feature vector (Caglayan et al., 2021). To perform feature extraction, we split the images so that it would be possible to carry out this process on different machines, in order to speed it up.¹

¹We used the Google Cloud Platform c2-standard-8 vir-

The next step was the bilingual and multimodal alignment, that is, the features obtained in the feature extraction process were associated with their corresponding English and Portuguese texts. Furthermore, we developed a script to disregard features that could be corrupted as VTLM does not support non-existent or non-loadable features. Thus, after identifying all valid features, the training, testing and validation sets of How2 were modified in order to eliminate the segments that had corrupted features.

We also used the MOSES scripts² to preprocess the dataset and then we applied byte pair encoding (BPE) (Sennrich et al., 2016) to convert tokens into subwords. After pre-processing, we changed the VTLM pre-training, fine-tuning and decoding scripts in order to adapt VTLM to the Portuguese language, making Portuguese-English the default language pair. As a result, it was possible for the projected region features and the English-Portuguese translation pairs to be processed in a single stream by the VTLM model.

4. Experiments

The experimentation was performed following the same steps of (Caglayan et al., 2021).

4.1. Pre-training

For pre-training, we used a set from the How2 corpus that contains 155k features and their corresponding text in English and Portuguese. The pre-training was conducted for 690 epochs, using a single NVIDIA GeForce GTX 1070 GPU, and best checkpoints were selected with respect to validation set accuracy. It took about four days and four hours to finish pre-training.

Similar to the original VTLM settings (Caglayan et al., 2021), we set the model dimension to 512, the feed-forward layer dimension to 2048, the number of layers to 6 and the number of attention heads to 8. Moreover, the model parameters are also randomly initialised and we used Adam (Kingma and Ba, 2014) with the mini-batch size set to 32 and the learning rate set to 0.0001. The dropout (Srivastava et al., 2014) rate was set to 0.1 in all layers.

4.2. Fine-tuning

The encoder and the decoder of Transformer-based MMT and NMT models are initialized with weights from VTLM, and fine-tuned with a smaller learning rate. The fine-tuning was conducted for 54 epochs for the MMT model and 84 epochs for the NMT model.

The same hyperparameters as the pre-training phase were used, except for the batch size and the learning rate, which were decreased to 16 and 1e-5, respectively.

tual machine, with 8 CPUs and 32 GB RAM, and a local machine equipped with an NVIDIA GeForce GTX 1070 GPU, 8 CPUs and 16 GB RAM. Even so, the feature extraction process took approximately 600 hours.

²<https://github.com/moses-smt/mosesdecoder>

For evaluation, we used the models with the lowest validation set perplexity to decode translations with beam size equal to 8.

4.3. Baselines

An equivalent process was performed with a TLM model. The TLM architecture corresponds to the VTLM architecture (3.1) without regional image features. Pre-training was conducted for 553 epochs using the same settings as VTLM, apart from the batch size that was decreased to 16, and fine-tuning was conducted for 61 epochs for the MMT model and 103 epochs for the NMT model.

For comparison, we also trained *from scratch* models without transferring weights from the pre-trained TLM or VTLM models. These models were trained only on the MT dataset and the training was conducted for 190 epochs for the MMT model and 225 epochs for the NMT model.

5. Results

The trained models were evaluated for the multimodal machine translation (MMT) and neural machine translation (NMT) tasks. Table 1 shows BLEU and METEOR scores across valid and test sets of How2. It is important to highlight that METEOR is the official metric for MMT; it was the metric used in the WMT competition (2016-2018) for multimodal machine translation task.

Similar to the original proposal (Caglayan et al., 2021), the results show the impact of cross-lingual visual pre-training on the final performance. The MMT model outperforms the MMT baseline by approximately 14 BLEU and 14 METEOR points when fine-tuned for multimodal machine translation.

Moreover, for the models trained from scratch (Baseline Transformers), MMT is inferior to NMT by about 5 BLEU and 7 METEOR points, but when pre-trained TLM/VTLM checkpoints are fine-tuned for MT, the difference between the MMT and NMT models scores diminishes or becomes non-existent.

Compared to the best results obtained by Caglayan et al. (2021) with German-English VTLM – 44.0 BLEU and 61.3 METEOR on the Multi30k test set for the MMT task – we achieved higher scores for Portuguese-English VTLM – 51.8 BLEU and 78.04 METEOR on the How2 test set for the MMT task. However, it is important to point out that a direct comparison is not possible here due to language and corpus differences.

5.1. Qualitative Analysis

Some examples of texts translated by each model are presented in Table 2.

In the first case, the MMT baseline misses the translation of the source words “consulting” and “Coral Gables”, while both VTLM and TLM translate them correctly, obtaining a better performance (about 80 BLEU points above the baseline). This indicates the

		Test		Valid	
		BLEU	METEOR	BLEU	METEOR
Baseline Transformers	MMT	37.57	63.51	38.34	63.60
	NMT	43.58	70.62	43.28	70.18
TLM: Pre-train and fine-tune on How2	MMT	51.99	77.52	52.19	77.87
	NMT	50.61	77.67	50.72	78.01
VTLM: Pre-train and fine-tune on How2	MMT	51.80	78.04	52.44	78.25
	NMT	52.20	78.20	52.81	78.70

Table 1: BLEU and METEOR values for baseline transformers (text only), TLM (text only) and VTLM (text and image) for NMT and MMT tasks.

efficacy of pre-training on the performance of the models.

In the second example, the difference between VTLM and TLM scores is greater. VTLM reaches a BLEU score of 100.0 and TLM obtains a BLEU score of 37.8, mostly due to the incorrect translation of the words “brown” and “whole”. Therefore, we observe that the regional visual features can help the model understand the context, resulting in a more accurate translation.

In the third case, the image has extra objects disassociated with the sentence, such as the hand and other components in the screen. As a result, the image can bring irrelevant information to the text, which may introduce noise and affect the translation quality. The translation of the VTLM model illustrates this possible disadvantage, as the model performs worse than the TLM model by about 19 BLEU points.

Furthermore, in the fourth example, none of the models accurately translated the source text. The underlying reason is that the word “grooming” in the reference sentence only appears a few times in the training set (25 times in a set of 3,304,534 tokens). Nevertheless, there is a gap between the translations of the three models. For example, the translation of the TLM model shows a greater degree of inaccuracy compared to the translation of the VTLM model, and the result of the baseline is “What kind of treatment our horse likes this horse.”, which is farther away from the correct result.

6. Conclusion

We present the work performed with VTLM (Visual Translation Language Modelling) to test its generalization ability for other languages and corpora. Unlike previous studies, we did not apply the widely used Multi30k dataset. Instead, we used the multimodal and multilingual corpus How2 in order to focus on the Brazilian Portuguese language – which is a low-resource language that has little attention in the current landscape of multimodal translation – and introduce a more challenging scenario, where the text associated with each image/video frame is not a simple description of it, but a subtitle that may not be related to its corresponding frame.

We applied the necessary pre-processing steps on the corpus so that it could be used to investigate the ef-

fectiveness of the model for a new language pair. We also compared the performance of the trained models to some baselines and showed the impact of adding the visual modality alongside the translation pairs by analyzing a few differences between the translations obtained by each model.

On multimodal machine translation, we reassert the efficacy of cross-lingual visual pretraining. Our trained model reached a BLEU score of 51.8 and a METEOR score of 78.0 on the Portuguese-English How2 test set, outperforming the MMT baseline by about 14 BLEU and 14 METEOR. We also reported higher scores in comparison to the original English-German VTLM, establishing suitability of the model to this new language pair and indicating a generalization ability for other languages. In addition, the good performance of the model showed that it can overcome the additional challenges that come with using a corpus that is a collection of videos (not images).

The main contributions of this work are: (i) the generation of new linguistic-computational resources for Brazilian Portuguese, processing scripts and the processed corpus/database generated for the experiments; (ii) the adaptation of VTLM to a new language pair and to more challenging circumstances in relation to the image-text relationship, as well as (iii) the relevant results of the experimentation. The source code is publicly available³. We expect this work can encourage the development of multimodal cross-lingual language models for low-resource languages.

In future work, we plan to incorporate a selecting method to remove irrelevant parts of the image features, as well as explore the impact of variability across languages to achieve a better comprehension of multimodality in language processing.

7. Acknowledgements

The work described in this paper is supported by grant #2020/15995-1, São Paulo Research Foundation (FAPESP).

³https://github.com/LALIC-UFSCar/VTLM_English-Portuguese





	<p>Source: Consultoria de imagem e etiqueta em Coral Gables, Flórida. Reference: Image and Etiquette Consulting in Coral Gables, Florida.</p>
	<p>Baseline: Image and Etiquette Etiquette ette: Florida, Florida. 17.57 BLEU TLM: Image and Etiquette Consulting in Coral Gables, Florida. 100.0 BLEU VTLM: Image and Etiquette Consulting in Coral Gables, Florida. 100.0 BLEU</p>
	<p>Source: E então algo como arroz integral ou pão de trigo integral. Reference: And then something like brown rice or whole wheat bread.</p>
	<p>Baseline: And then something like full rice or full trigger. 29.38 BLEU TLM: And then something like full rice or full wheat bread. 37.82 BLEU VTLM: And then something like brown rice or whole wheat bread. 100.0 BLEU</p>
	<p>Source: E você pode usar quantas dessas faixas de bateria dentro da sua janela de arranjos lógicos. Reference: And you can use as many of these drum tracks within your logic arrange window.</p>
	<p>Baseline: And you can use how many of these drum tracks inside your clock window. 42.83 BLEU TLM: And you can use as many of these drum tracks within your logic arrange window. 100.0 BLEU VTLM: And you can use as many of these drum tracks within your logic plugs. 81.07 BLEU</p>
	<p>Source: Que tipo de tratamento o vosso cavalo gosta. Reference: The kind of grooming that your horse likes.</p>
	<p>Baseline: What kind of treatment our horse likes this horse. 6.67 BLEU TLM: What kind of treatment our horse likes. 15.25 BLEU VTLM: What kind of treatment your horse likes. 36.28 BLEU</p>

Table 2: Translation examples of different MMT models: MMT Baseline Transformer, TLM and VTLM.

8. Bibliographical References

- Barrault, L., Bougares, F., Specia, L., Lala, C., Elliott, D., and Frank, S. (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, October. Association for Computational Linguistics.
- Caglayan, O., Barrault, L., and Bougares, F. (2016). Multimodal attention for neural machine translation.
- Caglayan, O., Kuyu, M., Amac, M. S., Madhyastha, P., Erdem, E., Erdem, A., and Specia, L. (2021). Cross-lingual Visual Pre-training for Multimodal Machine Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Short Papers*, online, April. Association for Computational Linguistics.
- Calixto, I. and Liu, Q. (2017). Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 992–1003, Copenhagen, Denmark, September. Association for Computational Linguistics.
- Calixto, I., Elliott, D., and Frank, S. (2016). DCU-UvA multimodal MT system report. In *Proceed-*

- ings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pages 634–638, Berlin, Germany, August. Association for Computational Linguistics.
- Calixto, I., Rios, M., and Aziz, W. (2019). Latent variable model for multi-modal translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6392–6405, Florence, Italy, July. Association for Computational Linguistics.
- Chen, Y.-C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. (2020). UNITER: Universal image-text representation learning.
- Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. In H. Wallach, et al., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics.
- Elliott, D. and Kádár, Á. (2017). Imagination improves multimodal translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 130–141, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.
- Elliott, D., Frank, S., Sima'an, K., and Specia, L. (2016). Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.
- Elliott, D., Frank, S., Barrault, L., Bougares, F., and Specia, L. (2017). Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233, Copenhagen, Denmark, September. Association for Computational Linguistics.
- Garg, S., Vu, T., and Moschitti, A. (2020). TANDA: Transfer and adapt pre-trained transformer models for answer sentence selection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:7780–7788, Apr.
- Gupta, K., Gautam, D., and Mamidi, R. (2021). ViTA: Visual-linguistic translation by aligning object tags. In *Proceedings of the 8th Workshop on Asian Translation, WAT@ACL/IJCNLP 2021, Online*, pages 166–173. Association for Computational Linguistics.
- Huang, P.-Y., Liu, F., Shiang, S.-R., Oh, J., and Dyer, C. (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, August. Association for Computational Linguistics.
- Huang, P.-Y., Patrick, M., Hu, J., Neubig, G., Metze, F., and Hauptmann, A. (2021). Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models.
- Ive, J., Madhyastha, P., and Specia, L. (2019). Distilling translations with visual awareness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6538, Florence, Italy, July. Association for Computational Linguistics.
- Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12.
- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M. S., and Li, F.-F. (2016). Visual genome: Connecting language and vision using crowdsourced dense image annotations.
- Kunchukuttan, A., Mehta, P., and Bhattacharyya, P. (2018). The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).
- Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Mallocci, M., Kolesnikov, A., and et al. (2020). The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, Mar.
- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). ALBERT: A lite BERT for self-supervised learning of language representations.
- Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. (2019). VisualBERT: A simple and performant baseline for vision and language. In *Arxiv*.
- Libovický, J. and Helcl, J. (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, July. Association for Computational Linguistics.
- Lin, J., Yang, A., Zhang, Y., Liu, J., Zhou, J., and Yang, H. (2021). InterBERT: Vision-and-language interaction for multi-modal pretraining.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized bert pretraining approach.
- Liu, P., Cao, H., and Zhao, T. (2021). Gumbel-attention for multi-modal machine translation. *CoRR*, abs/2103.08862.
- Long, Q., Wang, M., and Li, L. (2021). Generative imagination elevates machine translation. In *Proceedings of the 2021 Conference of the North Amer-*

- ican Chapter of the Association for Computational Linguistics: *Human Language Technologies*, pages 5738–5748, Online, June. Association for Computational Linguistics.
- Lu, J., Batra, D., Parikh, D., and Lee, S. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In H. Wallach, et al., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Ni, M., Huang, H., Su, L., Cui, E., Bharti, T., Wang, L., Zhang, D., and Duan, N. (2021). M3P: Learning universal representations via multitask multilingual multimodal pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3977–3986, June.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual BERT?
- Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.
- Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In C. Cortes, et al., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.
- Rothe, S., Narayan, S., and Severyn, A. (2020). Leveraging pre-trained checkpoints for sequence generation tasks.
- Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.
- Sharma, P., Ding, N., Goodman, S., and Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.
- Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.
- Specia, L., Frank, S., Sima'an, K., and Elliott, D. (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, August. Association for Computational Linguistics.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.
- Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., and Dai, J. (2020). VL-BERT: Pre-training of generic visual-linguistic representations.
- Tan, H. and Bansal, M. (2019). LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Yao, S. and Wan, X. (2020). Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350, Online, July. Association for Computational Linguistics.
- Yin, Y., Meng, F., Su, J., Zhou, C., Yang, Z., Zhou, J., and Luo, J. (2020). A novel graph-based multimodal fusion encoder for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3035, Online, July. Association for Computational Linguistics.
- Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.
- Zhou, M., Cheng, R., Lee, Y. J., and Yu, Z. (2018). A visual attention grounding neural model for multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3643–3653, Brussels, Belgium, October-November. Association for Computational Linguistics.

9. Language Resource References

- Ramon Sanabria and Ozan Caglayan and Shruti Palaskar and Desmond Elliott and Loïc Barrault and Lucia Specia and Florian Metz. (2018). *How2: A Large-scale Dataset for Multimodal Language Understanding*.