Cross-lingual Visual Pre-training for Multimodal Machine Translation

Ozan Caglayan[1], **Menekse Kuyu**[2], **Mustafa Sercan Amac**[2], **Pranava Madhyastha**[1]
Erkut Erdem[2], **Aykut Erdem**[3] and **Lucia Specia**[1,4,5]
Imperial College London[1], Hacettepe University[2], Koç University[3]
University of Sheffield[4], ADAPT - Dublin City University[5]
o.caglayan@ic.ac.uk, meneksekuyu@gmail.com, sercanamac@gmail.com, pranava@ic.ac.uk
erkut@cs.hacettepe.edu.tr, aerdem@ku.edu.tr, l.specia@ic.ac.uk

Abstract

Pre-trained language models have been shown to improve performance in many natural language tasks substantially. Although the early focus of such models was single language pre-training, recent advances have resulted in cross-lingual and visual pre-training methods. In this paper, we combine these two approaches to learn visually-grounded cross-lingual representations. Specifically, we extend the translation language modelling (Lample and Conneau, 2019) with masked region classification and perform pre-training with three-way parallel vision & language corpora. We show that when fine-tuned for multimodal machine translation, these models obtain state-of-the-art performance. We also provide qualitative insights into the usefulness of the learned grounded representations.

1 Introduction

Pre-trained language models (Peters et al., 2018; Devlin et al., 2019) have been proven valuable tools for contextual representation extraction. Many studies have shown their effectiveness in discovering linguistic structures (Tenney et al., 2019), which is useful for a wide variety of NLP tasks (Talmor et al., 2019; Kondratyuk and Straka, 2019; Petroni et al., 2019). These positive results led to further exploration of (i) cross-lingual pre-training (Lample and Conneau, 2019; Conneau et al., 2020; Wang et al., 2020) through the use of multiple mono-lingual and parallel resources, and (ii) visual pre-training where large-scale image captioning corpora are used to induce grounded vision & language representations (Lu et al., 2019; Tan and Bansal, 2019; Li et al., 2020a; Su et al., 2020; Li et al., 2020b). The latter is usually achieved by extending the masked language modelling (MLM) objective (Devlin et al., 2019) with auxiliary vision & language tasks such as masked region classification and image sentence matching.

In this paper, we present the first attempt to bring together cross-lingual and visual pre-training. Our visual translation language modelling (VTLM) objective combines the translation language modelling (TLM) (Lample and Conneau, 2019) with masked region classification (MRC) (Chen et al., 2020; Su et al., 2020) to learn grounded cross-lingual representations. Unlike most of the prior work that use classification or retrieval based downstream evaluation, we focus on the *generative* task of multimodal machine translation (MMT), where images accompany captions during translation (Sulubacak et al., 2020). Once pre-trained, we transfer the VTLM encoder to a Transformer-based (Vaswani et al., 2017) MMT and fine-tune it for the MMT task. To our knowledge, this is also the first attempt of pre-training & fine-tuning for MMT, where the current state of the art mostly relies on training multimodal sequence-to-sequence systems from scratch (Calixto et al., 2016; Caglayan et al., 2016; Libovický and Helcl, 2017; Elliott and Kádár, 2017; Caglayan et al., 2017; Yin et al., 2020).

Our findings highlight the effectiveness of cross-lingual visual pre-training: when fine-tuned on the English→German direction of the Multi30k dataset (Elliott et al., 2016), our MMT model surpasses our constrained MMT baseline by about 10 BLEU and 8 METEOR points. The rest of the paper is organised as follows: §2 describes our pre-training and fine-tuning protocol, §3 presents our quantitative and qualitative analyses, and §4 concludes the paper with pointers for future work.

2 Method

We propose Visual Translation Language Modelling (VTLM) objective to learn multimodal cross-lingual representations. In what follows, we first describe the TLM objective (Lample and Conneau, 2019) and then introduce the modifications required to extend it to VTLM.

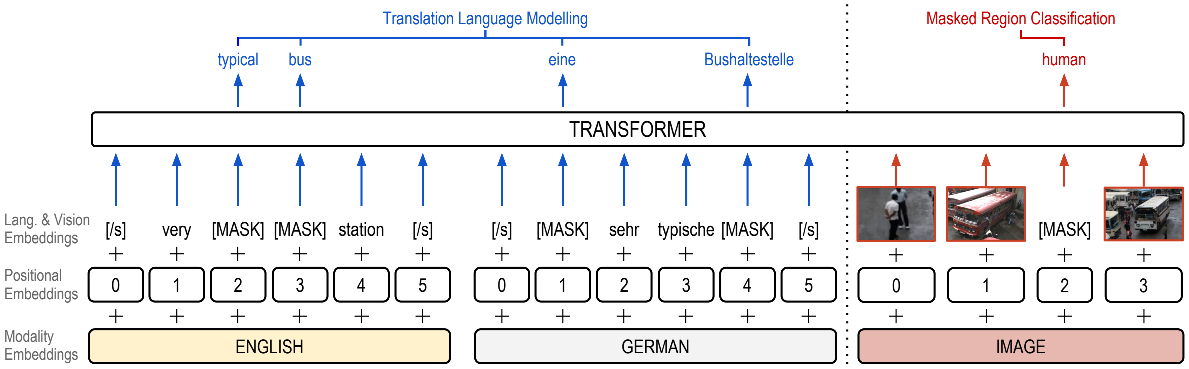

Figure 1: The architecture of the proposed model: VTLM extends the TLM (Lample and Conneau, 2019) (left side of the dotted line) with regional image features. Masking applies on both linguistic and visual tokens.

2.1 Translation language modelling

The TLM objective is based on Transformer networks and assumes the availability of parallel corpora during training. It defines the input x as the concatenation of m-length source language sentence $s^{(1)}_{1:m}$ and n-length target language sentence $s^{(2)}_{1:n}$:

$$x = \left[s^{(1)}_1, \cdots, s^{(1)}_m, s^{(2)}_1, \cdots, s^{(2)}_n\right]$$

For a given input, TLM follows (Devlin et al., 2019), and selects a random set of input tokens $y = \{s^{(l)}_1, \ldots, s^{(l)}_k\}$ for masking. Let us denote the masked input sequence with $\tilde{x}$, and the ground-truth targets for masked positions with $\hat{y}$. TLM employs the masked language modelling (MLM) objective to maximise the log-probability of correct labels $\hat{y}$, conditioned on the masked input $\tilde{x}$:

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

where θ are the model parameters. We keep the standard hyper-parameters for masking, i.e. 15% of inputs are randomly selected for masking, from which 80% are replaced with the `[MASK]` token, 10% are replaced with random tokens from the vocabulary, and 10% are left intact.

2.2 Visual translation language modelling

VTLM extends the TLM by adding the visual modality alongside the translation pairs (Figure 1). Therefore, we assume the availability of sentence pair & image triplets and redefine the input as:

$$x = \left[s^{(1)}_1, \cdots, s^{(1)}_m, s^{(2)}_1, \cdots, s^{(2)}_n, v_1, \cdots, v_o\right]$$

where $\{v_1, \cdots, v_o\}$ are features extracted from a Faster R-CNN model (Ren et al., 2015) pre-trained on the Open Images dataset (Kuznetsova et al., 2018).[1] Specifically, we extract convolutional feature maps from $o = 36$ most confident regions, and average pool each of them to obtain a region-specific feature vector $v_i \in \mathbb{R}^{1536}$. Each region i is also associated with a detection label $\hat{v}_i$ provided by the extractor. Before encoding, the feature vectors and their bounding box coordinates are projected into the language embedding space.

The final model processes translation pairs and projected region features in a **single-stream** fashion (Su et al., 2020; Li et al., 2020a), and combines the TLM loss with the masked region classification (MRC) loss as follows:

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

Masking. 15% random masking ratio is applied separately to both language and visual streams, and the $\hat{v}$ above now denotes the correct region labels for the masked feature positions. Different from previous work that zeroes out masked regions (Tan and Bansal, 2019; Su et al., 2020), VTLM replaces their projected feature vectors with the `[MASK]` token embedding.[2] Similar to textual masking, 10% of the random masking amounts to using regional features randomly sampled from all images in the batch, and the remaining 10% of regions are left intact.

2.3 Pre-training

VTLM requires a three-way parallel multimodal corpus, which does not exist in large-scale. To ad-

[1] The "*faster rcnn inception resnet v2 atrous oid v4*" model from TensorFlow.

[2] Although this choice is mostly practical, we hypothesise that using the same signal for both language and visual masking can be beneficial for grounding.

dress this, we extend[3] the **Conceptual Captions** (CC) (Sharma et al., 2018) dataset with German translations. CC is a large-scale collection of $\sim$3.3M images retrieved from the Internet, with noisy *alt-text* captions in English. The translation of English captions into German was automatically performed using an existing NMT model (Ng et al., 2019) provided[4] in the Fairseq (Ott et al., 2019) toolkit. Since some of the images are no longer accessible, the final corpus' size is reduced to $\sim$3.1M triplets. We used byte pair encoding (BPE) (Sennrich et al., 2016) to learn a joint 50k BPE model on the CC dataset. The pre-training was conducted for 1.5M steps, using a single RTX2080-Ti GPU, and best checkpoints were selected with respect to validation set accuracy.

Settings. We use a small version of the TLM (Lample and Conneau, 2019)[5] and set the model dimension, feed-forward layer dimension, number of layers and number of attention heads to $d = 512$, $f = 2048$, $l = 6$ and $h = 8$, respectively. We randomly initialise model parameters, instead of using pre-trained LM checkpoints such as BERT or XLM. We use Adam (Kingma and Ba, 2014) with the mini-batch size and the learning rate set to 64 and 0.0001, respectively. The dropout (Srivastava et al., 2014) rate is set to 0.1 in all layers. The pre-training is done for 1.5M steps using a single RTX2080-Ti GPU, and best checkpoints are selected with respect to validation accuracy.

2.4 Baseline MT models and fine-tuning

Our experimental protocol consists of **initialising** the encoder and the decoder of Transformer-based NMT and MMT models with weights from TLM/VTLM, and **fine-tuning** them with a smaller learning rate. The architectural difference between the NMT and the MMT models is that the latter encodes 36 regional visual features as part of the source sequence, similar to the VTLM (§ 2.2). As a natural baseline, we train constrained (trained only on the MT dataset) models without transferring weights from the pre-trained TLM/VTLM models. We refer to these models as **from-scratch**. For the fine-tuning experiments, we train three runs with different seeds. For evaluation, we use the models with the lowest validation set perplexity to decode translations with beam size equal to 8.

Dataset. We use the standard MMT corpus **Multi30k** (Elliott et al., 2016) for both fine-tuning and from-scratch runs. It contains 30k image descriptions from Flickr30k (Young et al., 2014) and their human translations in German for training, along with three test sets of 1K samples each: the original and the most in-domain **2016** test set, as well as **2017** and **COCO** test sets created using images and descriptions collected from sources other than Flickr.

Settings. For fine-tuning, we use the same hyperparameters as the pre-training phase, apart from decreasing the learning rate to $1e{-}5$. For MT models that are trained from scratch, we increase the dropout rate to 0.4 and linearly warm up the learning rate from $1e{-}7$ to $1e{-}4$ during the first 4,000 iterations. *Inverse square-root* annealing is applied after 4,000 iterations.

3 Results

3.1 Machine translation

Table 1 reports METEOR and BLEU scores across three different test sets of Multi30k. First, we observe that the MMT system trained from scratch is consistently worse than its NMT counterpart. However, the gap disappears when pre-trained TLM/VTLM checkpoints are fine-tuned for MT. This suggests that pre-training may be necessary for *single-stream* multimodal encoding, where the number of regions (36) outnumbers the avg. number of source tokens (13 for Multi30k).

Second, we see that the best performances are obtained when models are first pre-trained on the *three-way* parallel Conceptual Captions (CC) dataset. To validate this further, we train a baseline NMT on the concatenation of Multi30k and CC (NMT+CC) and an MMT that uses only Multi30k for **both** pre-training and fine-tuning. The results clearly show that these systems lag behind the ones pre-trained on CC.

We also experimented with an alternative pre-training strategy where we do not mask visual regions. Interestingly, this alternative MMT in Table 1 reveals that not masking visual regions during pre-training yields slightly better results overall. This is equivalent to letting the model predict the object labels from a multimodal input where words are stochastically masked but regional features are kept intact. Overall, MMT fine-tuning on VTLM sets a new state of the art across all Multi30k test

[3] `https://hucvl.github.io/VTLM`

[4] The `transformer.wmt19.en-de` model.

[5] https://github.com/facebookresearch/XLM

	2016		2017		COCO	
	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU
Best RNN-MMT (Caglayan, 2019)						
	58.7	39.4	52.9	32.6	–	–
Graph-based Transformers MMT (Yin et al., 2020)						
	57.6	39.8	51.9	32.2	37.6	28.7
Ensemble RNN-MMT (Delbrouck and Dupont, 2018)						
	59.6	40.3	–	–	–	–
Unconstrained Transformers MMT (Helcl et al., 2018)						
	59.1	42.7	–	–	–	–
Our Baseline Transformers **(from scratch)**						
NMT	56.4	37.6	51.3	30.9	47.2	27.5
+CC	58.8	39.5	55.6	36.2	51.5	33.0
MMT	55.4	35.2	49.5	27.7	46.2	25.4
VTLM: Pre-train and fine-tune on **Multi30k**						
MMT	59.0	40.2	53.5	32.7	49.3	28.9
TLM: Pre-train on **CC** – fine-tune on **Multi30k**						
NMT	60.7	43.1	56.5	37.6	53.3	34.8
	60.5 ± 0.21	42.5 ± 0.46	56.4 ± 0.10	37.3 ± 0.38	53.1 ± 0.13	34.6 ± 0.17
MMT	60.3	41.9	56.7	37.6	53.3	34.3
	60.2 ± 0.08	41.7 ± 0.18	56.5 ± 0.16	37.5 ± 0.10	53.0 ± 0.20	34.1 ± 0.14
VTLM: Pre-train on **CC** – fine-tune on **Multi30k**						
NMT	61.2	43.3	56.9	37.2	53.7	35.1
	60.5 ± 0.46	42.5 ± 0.53	56.4 ± 0.34	37.0 ± 0.16	53.1 ± 0.42	34.6 ± 0.40
MMT	60.8	42.7	57.1	**38.1**	53.1	34.2
	60.6 ± 0.15	42.6 ± 0.14	56.9 ± 0.20	37.7 ± 0.43	53.0 ± 0.05	33.9 ± 0.19
VTLM: Alternative (0% visual masking during pre-training)						
MMT	**61.3**	**44.0**	**57.2**	38.0	**53.8**	**35.2**
	60.9 ± 0.30	43.3 ± 0.59	57.1 ± 0.07	37.6 ± 0.31	53.6 ± 0.17	35.1 ± 0.09

Table 1: Quantitative comparison of experiments: when the mean and the standard deviation is reported, the single numbers appearing above, denote the maximum across three different runs.

sets.[6] We leave the exploration of visual region masking for the MRC task as future work and proceed with the alternative variant in the following experiments.

Encoder attention parameters. When fine-tuning the TLM for MT, the default XLM implementation randomly initialises the decoder's missing encoder attention parameters. In our experiments, we noticed that **copying** those parameters from the TLM self-attention layers substantially improves the results up to 2.2 BLEU.

3.2 Explicit masking

Here, we will evaluate the extent to which the visual information is taken into account (i) when TLM/VTLM predicts masked tokens, and (ii) when the fine-tuned NMT and MMT models are forced to translate source sentences with missing visual entities. For the latter, we use Flickr30k entities (Plummer et al., 2015) to mask head nouns in 2016 test set sentences, similar to Caglayan et al. (2019).

Last-word masking. In this experiment, we measure the target word prediction accuracy, when last tokens[7] of input caption pairs are systematically masked during evaluation. Table 2 suggests

[6] We exclude Grönroos et al. (2018) as their improvements (45.5 BLEU) were not due to multi-modality but rather to other modifications such as heavy parallel data augmentation, domain fine-tuning, and ensembling.

[7] We pre-process the sentences to ensure that they do not end with punctuation marks, which would make the task easier for masked punctuation.

	VALID			TEST		
	EN	DE	BOTH	EN	DE	BOTH
TLM	89.0	87.3	55.2	88.5	86.3	53.6
VTLM	⇑ **0.9**	⇑ **1.4**	⇑ **5.0**	⇑ **1.1**	⇑ **2.2**	⇑ **5.8**
+shuf	⇓ **1.0**	⇓ **0.2**	⇓ **7.7**	⇓ **1.3**	⇓ **0.3**	⇓ **7.4**

Table 2: Masked *last-word* prediction accuracies: VTLM gains are with respect to TLM, whereas the incongruent (+shuf) drops are relative to VTLM.

	MASK	REMOVE
TLM→NMT	31.44	27.38
TLM→MMT	⇓ **0.43**	⇓ **0.26**
VTLM→NMT	31.27	27.63
VTLM→MMT	⇑ **1.65**	⇑ **0.65**

Table 3: Entity masking on 2016 test set: results are BLEU averages of three fine-tuned MT systems.

that the visual information is much more helpful (i.e. up to 6% accuracy improvement) when last tokens are masked in both English and German captions. However, if one caption is available, it provides enough context for cross-lingual prediction. Finally, when we shuffle (+shuf) the test set features to introduce incongruence (Elliott, 2018), we see that the VTLM model deteriorates substantially. This confirms that the accuracy improvements are not due to side-effects of experimentation noise, such as regularisation or random seed related effects.

Entity masking in MT. We devise two ways of masking entities i.e. we either replace them with the [MASK] token or remove them entirely so that the masking phenomena is *not known* to the model. The results in Table 3 show that MMT models can recover the missing source context to some extent, only when they are pre-trained using the proposed VTLM objective. In other words, the grounding ability can only be acquired when visual modality is present for both pre-training and fine-tuning. The gap between MASK and REMOVE also seems to highlight the importance of reserving a source position even it is corrupted/masked.

3.3 Visual attention in MMT

Here we take the MMT decoder's cross-attention layers and measure the attention mass they attribute to regional features in the input embeddings. Although the encoder's self-attention layers produce increasingly mixed contextual embeddings as we move towards the top layers, Brunner et al. (2020) show that the final layer states still encode corresponding input embeddings to some extent. With this assumption at hand, Figure 2 shows the average attention mass attributed to the first 36 (visual) top-layer encoding states, by each cross-attention layer in the decoder. We find these results to be in *agreement* with the quantitative metrics (Table 1), with VTLM-MMT assigning substantially more attention to these positions, compared to TLM-MMT and MMT from scratch.

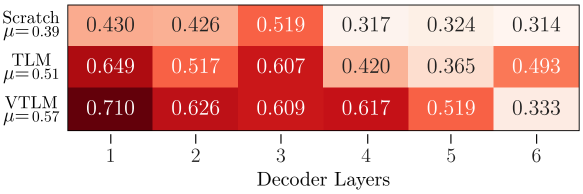

Figure 2: Cross-attention mass over the visual portion of input sequences, averaged across the 2016 test set.

4 Conclusions

We proposed a novel cross-lingual visual pre-training approach and tested its efficacy for multimodal machine translation. Our pre-training approach extends the TLM framework (Lample and Conneau, 2019) with regional features and performs masked language modelling and masked region classification on a three-way parallel corpus. We show that this leads to substantial improvements compared to multimodal machine translation with cross-lingual pre-training only or without pre-training at all. As future work, we consider exploring *more informed* masking strategies for visual regions and investigating the impact of visual masking probability for the MRC pre-training task for downstream MMT performance.

Acknowledgments

This work was supported in part by TUBA GEBIP fellowship awarded to Erkut Erdem, and the MMVC project funded by TUBITAK and the British Council via the Newton Fund Institutional Links grant programme (grant ID 219E054 and 352343575). Lucia Specia, Pranava Madhyastha and Ozan Caglayan also received support from MultiMT (H2020 ERC Starting Grant No. 678017).

References

Gino Brunner, Yang Liu, Damian Pascual Ortiz, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2020. On Identifiability in Transformers. In *ICLR*.

Ozan Caglayan. 2019. *Multimodal Machine Translation*. Theses, Université du Maine.

Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost van de Weijer. 2017. LIUM-CVC submissions for WMT17 multimodal translation task. In *Proceedings of the Second Conference on Machine Translation*, pages 432–439, Copenhagen, Denmark. Association for Computational Linguistics.

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

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, Desmond Elliott, and Stella Frank. 2016. DCU-UvA multimodal MT system report. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 634–638, Berlin, Germany. Association for Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *Computer Vision – ECCV 2020*, pages 104–120, Cham. Springer International Publishing.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jean-Benoit Delbrouck and Stéphane Dupont. 2018. UMONS Submission for WMT18 Multimodal Translation Task. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 643–647, Belgium, Brussels. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

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, Berlin, Germany. Association for Computational Linguistics.

Desmond Elliott and Ákos Kádár. 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. Asian Federation of Natural Language Processing.

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

Jindřich Helcl, Jindřich Libovický, and Dušan Variš. 2018. CUNI system for the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 616–623, Belgium, Brussels. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. 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 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. 2018. The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. *CoRR*, abs/1811.00982.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020a. Unicoder-vl: A universal encoder for vision and language by cross-modal pretraining. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11336–11344. AAAI Press.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020b. Oscar: Object-Semantics Aligned Pretraining for Vision-Language Tasks. In *Computer Vision – ECCV 2020*, pages 121–137, Cham. Springer International Publishing.

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

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision (ICCV)*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc.

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.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 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 (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pretraining of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann. 2020. Multimodal machine translation through visuals and speech. *Machine Translation*, pages 1–51.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 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.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *arXiv preprint arXiv:2002.10957*.

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A novel graph-based multi-modal fusion encoder for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3035, Online. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 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.