ViTA: Visual-Linguistic Translation by Aligning Object Tags

Kshitij Gupta **Devansh Gautam** **Radhika Mamidi**
International Institute of Information Technology Hyderabad
{kshitij.gupta,devansh.gautam}@research.iiit.ac.in,
radhika.mamidi@iiit.ac.in

Abstract

Multimodal Machine Translation (MMT) enriches the source text with visual information for translation. It has gained popularity in recent years, and several pipelines have been proposed in the same direction. Yet, the task lacks quality datasets to illustrate the contribution of visual modality in the translation systems. In this paper, we propose our system under the team name *Volta* for the Multimodal Translation Task of WAT 2021[1] (Nakazawa et al., 2021) from English to Hindi. We also participate in the textual-only subtask of the same language pair for which we use mBART, a pretrained multilingual sequence-to-sequence model. For multimodal translation, we propose to enhance the textual input by bringing the visual information to a textual domain by extracting object tags from the image. We also explore the robustness of our system by systematically degrading the source text. Finally, we achieve a BLEU score of 44.6 and 51.6 on the test set and challenge set of the multimodal task.

1 Introduction

Machine Translation deals with the task of translation between language pairs and has been an active area of research in the current stage of globalization. In the task of multimodal machine translation, the problem is further extended to incorporate visual modality in the translations. The visual cues help build a better context for the source text and are expected to help in cases of ambiguity.

With the help of visual grounding, the machine translation system has scope for becoming more robust by mitigating noise from the source text and relying on the visual modality as well.

In the current landscape of multimodal translation, one of the issues is the limited datasets available for the task. Another contributing factor is that often the images add irrelevant information to the sentences, which may act as noise instead of an added feature. The available datasets, like Multi30K (Elliott et al., 2016), are relatively smaller when compared to large-scale text-only datasets (Bahdanau et al., 2015). The scarcity of such datasets hinders building robust systems for multimodal translation.

To address these issues, we propose to bring the visual information to a textual domain and fine-tune a high resource unimodal translation system to incorporate the added information in the input. We add the visual information by extracting the object classes by using an object detector and add them as tags to the source text. Further, we use mBART, a pretrained multilingual sequence-to-sequence model, as the base architecture for our translation system. We fine-tune the model on a textual-only dataset released by Kunchukuttan et al. (2018) consisting of 1,609,682 parallel sentences in English and Hindi. Further, we fine-tune it on the training set enriched with the object tags extracted from the images. We achieve state-of-the-art performance on the given dataset. The code for our proposed system is available at https://github.com/kshitij98/vita.

The main contributions of our work are as follows:

- We explore the effectiveness of fine-tuning mBART to translate English sentences to Hindi in the text-only domain.
- We further propose a multimodal system for translation by enriching the input with the object tags extracted from the images using an object detector.
- We explore the robustness of our system by a thorough analysis of the proposed pipelines

[1] http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html

Proceedings of the 8th Workshop on Asian Translation, pages 166–173
Bangkok, Thailand (online), August 5-6, 2021.

by systematically degrading the source text and finally give a direction for future work.

The rest of the paper is organized as follows. We discuss prior work related to multimodal translation. We describe our systems for the textual-only and multimodal translation tasks. Further, we report and compare the performance of our models with other systems from the leaderboard. Lastly, we conduct a thorough error analysis of our systems and conclude with a direction for future work.

2 Related Work

Earlier works in the field of machine translation largely used statistical or rule-based approaches, while neural machine translation has gained popularity in the recent past. Kalchbrenner and Blunsom (2013) released the first deep learning model in this direction, and later works utilize transformer-based approaches (Vaswani et al., 2017; Song et al., 2019; Conneau and Lample, 2019; Edunov et al., 2019; Liu et al., 2020) for the problem.

Multimodal translation aims to use the visual modality with the source text to help create a better context of the source text. Specia et al. (2016) first conducted a shared task on the problem and released the dataset, Multi30K (Elliott et al., 2016). It is an extended German version of Flickr30K (Young et al., 2014), which was further extended to French and Czech (Elliott et al., 2017; Barrault et al., 2018). For multimodal translation between English and Hindi, Parida et al. (2019) propose a subset of Visual Genome dataset (Krishna et al., 2017) and provide parallel sentences for each of the captions.

Although both English and Hindi are spoken by a large number of people around the world, there has been limited research in this direction. Dutta Chowdhury et al. (2018) created a synthetic dataset for multimodal translation of the language pair and further used the system proposed by Calixto and Liu (2017). Later, Sanayai Meetei et al. (2019) work with the same architecture on the multimodal translation task in WAT 2019. Laskar et al. (2019) used a doubly attentive RNN-based encoder and decoder architecture (Calixto and Liu, 2017; Calixto et al., 2017). Laskar et al. (2020) also proposed a similar architecture and pretrained on a large textual parallel dataset (Kunchukuttan et al., 2018) in their system.

	Train	Valid	Test	Challenge
#sentence pairs	28,930	998	1,595	1,400
Avg. #tokens (source)	4.95	4.93	4.92	5.85
Avg. #tokens (target)	5.03	4.99	4.92	6.17

Table 1: The statistics of the provided dataset. The average number of tokens in the source and target language are reported for all the sentence pairs.

3 System Overview

In this section, we describe the systems we use for the task.

3.1 Dataset Description

We use the dataset provided by the shared task organizers (Parida et al., 2019), which consists of images and their associated English captions from Visual Genome (Krishna et al., 2017) along with the Hindi translations of the captions. The dataset also provides a challenge test which consists of sentences where there are ambiguous English words, and the image can help in resolving the ambiguity. The statistics of the dataset are shown in Table 1. We use the provided dataset splits for training our models.

We also use the dataset released by Kunchukuttan et al. (2018) which consists of parallel sentences in English and Hindi. We use the training set, which contains 1,609,682 sentences, for training our systems.

3.2 Model

We fine-tune mBART, which is a multilingual sequence-to-sequence denoising auto-encoder that has been pre-trained using the BART (Lewis et al., 2020) objective on large-scale monolingual corpora of 25 languages, including both English and Hindi. The pre-training corpus consists of 55,608 million English tokens (300.8 GB) and 1,715 million Hindi tokens (20.2 GB). Its architecture is a standard sequence-to-sequence Transformer (Vaswani et al., 2017), with 12 encoder and decoder layers each and a model dimension of 1024 on 16 heads resulting in ~680 million parameters. To train our systems efficiently, we prune mBART's vocabulary by removing the tokens which are not present in the provided dataset or the dataset released by Kunchukuttan et al. (2018).

3.2.1 mBART

We fine-tune mBART for text-only translation from English to Hindi and feed the English sentences

Model	Test Set			Challenge Set		
	BLEU	RIBES	AMFM	BLEU	RIBES	AMFM
Text-only Translation						
CNLP-NITS-PP	37.01	0.80	0.81	37.16	0.77	0.80
ODIANLP	40.85	0.79	0.81	38.50	0.78	0.80
NLPHut	42.11	0.81	0.82	43.29	0.82	0.83
mBART (ours)	**44.12**	**0.82**	**0.84**	**51.66**	**0.86**	**0.88**
Multimodal Translation						
CNLP-NITS	40.51	0.80	0.82	33.57	0.75	0.79
iitp	42.47	0.81	0.82	37.50	0.79	0.81
CNLP-NITS-PP	39.46	0.80	0.82	39.28	0.79	0.83
ViTA (ours)	**44.64**	**0.82**	**0.84**	**51.60**	**0.86**	**0.88**

Table 2: Performance of our proposed systems on the test and challenge set.

to the encoder and decode Hindi sentences. We first fine-tune the model on the dataset released by Kunchukuttan et al. (2018) for 30 epochs, and then fine-tune it on the Hindi Visual Genome dataset for 30 epochs.

3.2.2 ViTA

We again fine-tune mBART for multimodal translation from English to Hindi but add the visual information of the image to the text by adding the list of object tags detected from the image. We feed the English sentences along with the list of object tags to the encoder and decode Hindi sentences. For feeding the data to the encoder, we concatenate the English sentence, followed by a separator token '##', followed by the object tags which are separated by ','. We use Faster R-CNN with ResNet-101-C4 backbone[2] (Ren et al., 2015) to detect the list of objects present in the image. We sort the objects by their confidence scores and choose the top ten objects.

For training the model, we first fine-tune the model on the dataset released by Kunchukuttan et al. (2018). Since this is a text-only dataset, we do not add any object tag information. Afterward, we fine-tune the model on Hindi Visual Genome dataset, where each sentence has been concatenated with object tags. Initially, we mask ~15% of the tokens in each sentence to incentivize the model to use the object tags along with the text and fine-tune the model on masked sentences along with object tags for 30 epochs. Finally, we train the model for 30 more epochs on Hindi Visual Genome dataset with unmasked sentences and object tags.

3.3 Experimental Setup

We implement our systems using the implementation of mBART available in the fairseq library[3] (Ott et al., 2019). We fine-tune on 4 Nvidia GeForce RTX 2080 Ti GPUs with an effective batch size of 1024 tokens per GPU. We use the Adam optimizer ($\epsilon = 10^{-6}, \beta_1 = 0.9, \beta_2 = 0.98$) (Kingma and Ba, 2015) with 0.1 attention dropout, 0.3 dropout, 0.2 label smoothing and polynomial decay learning rate scheduling. We validate the models every epoch and select the best checkpoint after each training based on the best validation BLEU score. To train our systems efficiently, we prune the vocabulary of our model by removing the tokens which do not appear in any of the datasets mentioned in the previous section. While decoding, we use beam search with a beam size of 5.

4 Results and Discussion

The BLEU score (Papineni et al., 2002) is the official metric for evaluating the performance of the models in the leaderboard. The leaderboard further uses RIBES (Isozaki et al., 2010) and AMFM (Banchs and Li, 2011) metrics for the evaluations. We report the performance of our models after tokenizing the Hindi outputs using `indic-tokenizer`[4] in Table 2.

It can be seen that our model is able to generalize well on the challenge set as well and performs better than other systems by a large margin. To

[2]We use the implementation available in Detectron2 (https://github.com/facebookresearch/detectron2).

[3]https://github.com/pytorch/fairseq

[4]https://github.com/ltrc/indic-tokenizer

English Sentence	A large pipe extending from the wall of the court.
Hindi Translation	कोर्ट की दीवार से निक्ली हुई एक बड़ी पाइप
Object Tags	building, man, flowers, shorts, racket, hat, court, shoe, shirt, window
mBART output	**अदालत** की दीवार से विस्तारित एक बड़ा पाइप
ViTA output	**कोर्ट** की दीवार से विस्तारित एक बड़ा पाइप

Figure 1: A translation example from the challenge set which illustrates the advantage of using ViTA to resolve ambiguities. mBART is translating the word court to judicial court, while ViTA translates it to tennis court.

English Sentence	A person riding a motorcycle.
Masked Sentence	A <mask> riding a <mask>.
Object Tags	helmet, building, sign, man, shirt, bike, flowers, barrier, tree, wheel
mBART output	एक आदमी **घोड़े** की सवारी करता है
ViTA output	एक आदमी एक **बाइक** की सवारी कर रहा है

Figure 2: The effect of object tags on an entity masked input from the test set. ViTA is able to use the context built from the object tags to predict a motorcycle, while mBART is predicting a horse instead.

	Train	Valid	Test	Challenge
#entities in text	29,583	1,028	1,631	1,592
#objects tags in images	253,051	8,679	13,855	12,507
#entities in object tags	13,959	498	758	442
%entities in object tags	47.18%	48.44%	46.47%	27.76%

Table 3: We show the overlap between the entities in the text and the object tags detected using Faster R-CNN model. The entities were identified using the `en_core_web_sm` model from the spaCy library[5].

	No masking	Entity Masking	Degradation %
mBART	44.2	15.1	65.8
ViTA	**44.6**	22.5	49.6
ViTA-gt	43.6	**25.4**	**41.7**

Table 4: The effect of entity masking on the BLEU score of the proposed models on the test set.

further analyze the results, we find a few cases in the challenge set wherein ViTA is able to resolve ambiguities, and an example is illustrated in Figure 1. Yet, the performance of the models is very similar across the textual-only and multimodal domains, and there are no significant improvements observed in the multimodal system.

4.1 Degradation

Although there is no significant improvement in the multimodal systems over the textual-only models, Caglayan et al. (2019) explore the robustness of multimodal systems by systematically degrading the source text for translations. We employ a similar approach and degrade the source text to compare our systems.

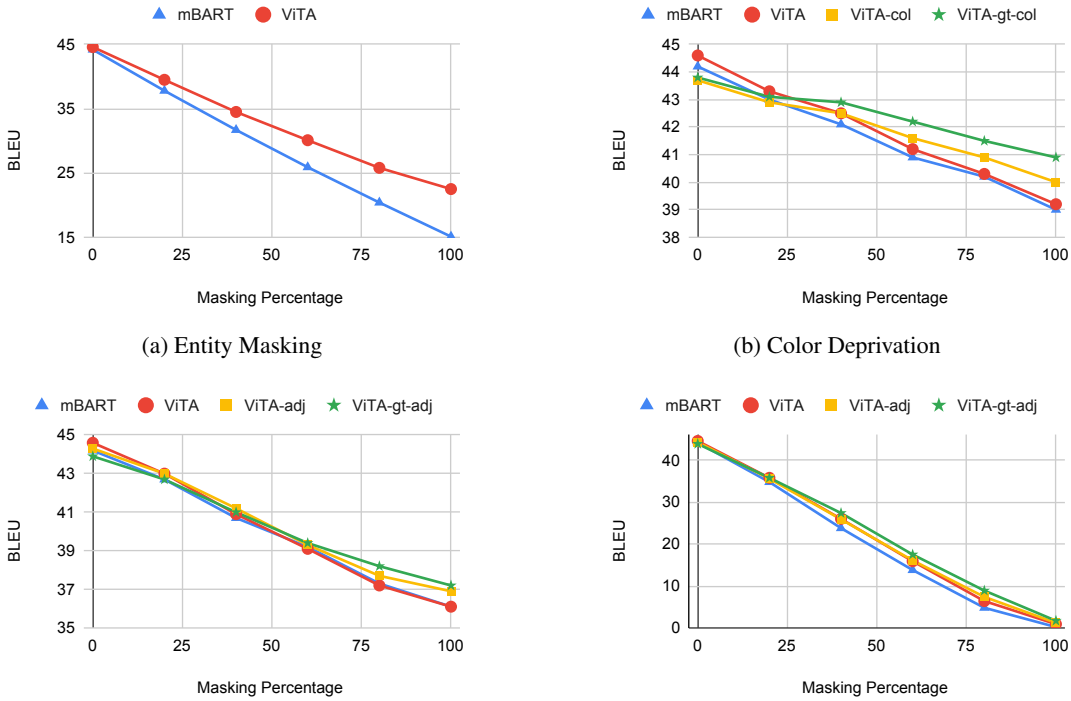
4.1.1 Entity masking

The goal of entity masking is to mask out the visually depictable entities in the source text so that the multimodal systems can make use of the visual cues in the image. To identify such entities, we use the `en_core_web_sm` model in spaCy[5] to predict the nouns in the sentence. The statistics of the tagged entities can be seen in Table 3.

We progressively increase the percentage of masked entities to better compare the degradation of our systems and it can be seen in Figure 3a. The final degraded values are reported in Table 4. Since the masked entities can also be predicted by using only the textual context of the sentence, we similarly add a training step of masking ~15% tokens while training mBART for a valid comparison. An example of the performance of our systems on an entity masked input is illustrated in Figure 2.

As an upper bound to the scope of our system, we propose `ViTA-gt`, which uses the ground-truth object labels from the Visual Genome dataset. Since the number of annotated objects is large, we filter them by removing the objects far from the image region.

[5] https://spacy.io/



(a) Entity Masking

(b) Color Deprivation

(c) Adjective Masking

(d) Random Masking

Figure 3: BLEU score comparison of the proposed models by increasing the masking percentage in the source text.

	No masking	Color Deprivation	Degradation %
mBART	44.2	39.0	11.8
ViTA	**44.6**	39.2	12.1
ViTA-col	43.7	40.0	8.5
ViTA-gt-col	43.8	**40.9**	**6.6**

Table 5: The effect of color deprivation on the BLEU score of the proposed models on the test set.

	No masking	Adjective Masking	Degradation %
mBART	44.2	36.1	18.3
ViTA	**44.6**	36.1	19.1
ViTA-adj	44.3	36.9	16.7
ViTA-gt-adj	43.9	**37.2**	**15.3**

Table 6: The effect of adjective masking on the BLEU score of the proposed models on the test set.

4.1.2 Color deprivation

The goal of color deprivation is to similarly mask tokens that are difficult to predict without the visual context of the image. To identify the colors in the source text, we maintain a list of colors and check whether the words in the sentence are present in the list. Similar to entity masking, we progressively increase the percentage of masked colors in the dataset to compare our systems. The comparison of our systems can be seen in Figure 3b. The final values of color deprivation are reported in Table 5.

As an upper bound to the scope of our system, we believe that colors can further be added to the object tags to help build a more robust system. As an added experiment, we propose `ViTA-col` by using the ground-truth annotations from the Visual Genome dataset and adding colors to our predicted object tags, which are present in the ground-truth objects as well. As a part of future work, we would like to extend our system to predict the colors from the image itself. We further experiment with `ViTA-gt-col`, which uses ground-truth objects with added colors in the input.

4.1.3 Adjective Masking

Similar to color deprivation, we propose adjective masking as several of the adjectives are visually depictable, and the degradation comparison should not be limited to just entities and colors. We predict the adjectives in the sentence by using the POS tagging model `en_core_web_sm` from spaCy library.

The performance of our models is compared in Figure 3c. The final values are reported in Table 6.

As an upper bound to the scope of our system, we propose to add all the adjectives to their corresponding object tags in the input. We propose

ViTA-adj by adding the ground truth adjectives annotated in the Visual Genome dataset to the object tags which are also predicted by our object detector. We also propose ViTA-gt-adj, which uses the ground-truth objects with their corresponding adjectives. The objects which are from the image region are removed to mitigate the noise added by the large number of objects in the annotations.

4.1.4 Random Masking

For a general robustness comparison of our models, we remove the limitation of manually masking the source sentences and progressively mask the text by random sampling.

The performance of our models is compared in Figure 3d.

5 Conclusion

We propose a multimodal translation system and utilize the textual-only pre-training of a neural machine translation system, mBART, by extracting object tags from the image. Further, we explore the robustness of our proposed multimodal system by systematically degrading the source texts and observe improvements from the textual-only counterpart. We also explore the shortcomings of the currently available object detectors and use ground-truth annotations in our experiments to show the scope of our methodology. The addition of colors and adjectives further adds to the robustness of the system and can be explored further in the future.

References

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

Rafael E. Banchs and Haizhou Li. 2011. AM-FM: A semantic framework for translation quality assessment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 153–158, Portland, Oregon, USA. Association for Computational Linguistics.

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.

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 and Qun Liu. 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. Association for Computational Linguistics.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924, Vancouver, Canada. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Koel Dutta Chowdhury, Mohammed Hasanuzzaman, and Qun Liu. 2018. Multimodal neural machine translation for low-resource language pairs using synthetic data. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 33–42, Melbourne. Association for Computational Linguistics.

Sergey Edunov, Alexei Baevski, and Michael Auli. 2019. Pre-trained language model representations for language generation. 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 4052–4059, Minneapolis, Minnesota. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 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. 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. Association for Computational Linguistics.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on*

Empirical Methods in Natural Language Processing, pages 944–952, Cambridge, MA. Association for Computational Linguistics.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2020. Multimodal neural machine translation for English to Hindi. In *Proceedings of the 7th Workshop on Asian Translation*, pages 109–113, Suzhou, China. Association for Computational Linguistics.

Sahinur Rahman Laskar, Rohit Pratap Singh, Partha Pakray, and Sivaji Bandyopadhyay. 2019. English to Hindi multi-modal neural machine translation and Hindi image captioning. In *Proceedings of the 6th Workshop on Asian Translation*, pages 62–67, Hong Kong, China. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, and Sadao Oda, Yusuke Kurohashi. 2021. Overview of the 8th workshop on Asian translation. In *Proceedings of the 8th Workshop on Asian Translation*, Bangkok, Thailand. 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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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.

Shantipriya Parida, Ondřej Bojar, and Satya Ranjan Dash. 2019. Hindi Visual Genome: A Dataset for Multimodal English-to-Hindi Machine Translation. *Computación y Sistemas*, 23(4):1499–1505. Presented at CICLing 2019, La Rochelle, France.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99.

Loitongbam Sanayai Meetei, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2019. WAT2019: English-Hindi translation on Hindi visual genome dataset. In *Proceedings of the 6th Workshop on Asian Translation*, pages 181–188, Hong Kong, China. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

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, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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.