Multimodal Neural Machine Translation System for English to Bengali

Shantipriya Parida[*], Subhadarshi Panda[†], Satya Prakash Biswal[♣], Ketan Kotwal[*], Arghyadeep Sen[♠], Satya Ranjan Dash[♠], Petr Motlicek[*]

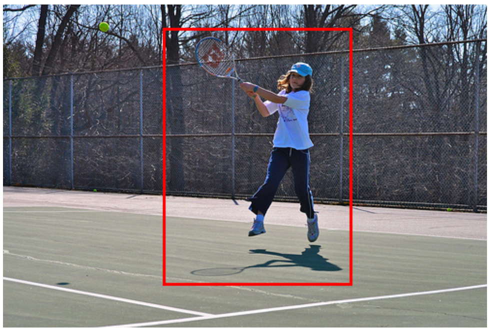
[*]Idiap Research Institute, Martigny, Switzerland
`{firstname.lastname}@idiap.ch`
[†]Graduate Center, City University of New York, USA
`spanda@gradcenter.cuny.edu`
[♣]The University Of Chicago, Chicago, USA
`sbiswal@chicagobooth.edu`
[♠]KIIT University, Bhubaneswar, India
`{2081012,sdashfca}@kiit.ac.in`

Abstract

Multimodal Machine Translation (MMT) systems utilize additional information from other modalities beyond text to improve the quality of machine translation (MT). The additional modality is typically in the form of images. Despite proven advantages, it is indeed difficult to develop an MMT system for various languages primarily due to the lack of a suitable multimodal dataset. In this work, we develop an MMT for English→Bengali using a recently published *Bengali Visual Genome* (BVG) dataset that contains images with associated bilingual textual description. Through a comparative study of the developed MMT system *vis-a-vis* a Text-to-text translation, we demonstrate that the use of multimodal data not only improves the translation performance improvement in BLEU score of +1.3 on the development set, +3.9 on the evaluation test, and +0.9 on the challenge test set but also helps to resolve ambiguities in the pure text description. As per best of our knowledge, our English-Bengali MMT system is the first attempt in this direction, and thus, can act as a baseline for the subsequent research in MMT for low resource languages.

1 Introduction

Over the last decade, deep neural networks (DNN) achieved state-of-the-art results for many tasks including computer vision, natural language processing, and speech processing—which encouraged researchers to design a system that will get benefit from the fusion of multiple modalities (Caglayan et al., 2016).



English Text: A girl playing tennis.
Bengali Text: একটি মেয়ে টেনিস খেলছে

Figure 1: A sample from the BVG dataset: an image with a specific region marked and its description in English and Bengali.

Multimodal Translation refers to the extraction of information from more than one modality where it is assumed that alternative views would be used for input data (Sulubacak et al., 2020; Yao and Wan, 2020; Elliott, 2018). The tasks and applications in multimodal translation involve translation of image captions, translation of video content, translation of spoken language, and others. These applications exploit more than one modality such as translation from video content includes audio and visual modality, and translation of image captions includes visual modality and caption text. Although there are different opinions on the performance of machine translation using visual modality but under limited resources, visual input generates better translation (Caglayan et al., 2019). It is observed that multimodal translation systems do

Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021),
in conjuction with RANLP-2021, September 7, 2021. pages 31–39
https://doi.org/10.26615/978-954-452-073-1_006

not leverage the visual signal to produce the correct translation in case of mistakes in the source language sentence (Chowdhury and Elliott, 2019).

When images are considered as an additional modality, the recent research can be divided into two major approaches based on their utilization of image features for the MMT: processing the global image features (Calixto and Liu, 2017), and processing the object tags derived from the images (Gupta et al., 2021). Cross-lingual visual pre-training which learns multimodal cross-lingual representations also found effective in MMT (Caglayan et al., 2021).

Bengali (also known as *Bangla*) is an Indo-Aryan language widely spoken in India and Bangladesh and considered as the 6-th most spoken language of the world with approximately 230 million speakers. Although English-to-Bengali text-only parallel corpora (Haddow and Kirefu, 2020; Ramesh et al., 2021) are available for building MT systems (Hasan et al., 2019, 2020; Parida et al., 2020); the multimodal dataset for Bengali did not exist. Thus, English-Bengali MMT systems have not been developed until now. Recently, the first English-Bengali multimodal dataset: Bengali Visual Genome (BVG) has been published (Sen et al., (in press) —which has facilitated research and development of corresponding multimodal as well as image captioning tasks.

The primary objective of this paper is to develop an MMT system for Bengali where the multimodal input is provided as an image and its description in English. We have used the BVG dataset to demonstrate our MMT system. The BVG consists of image descriptions (or captions) in the bilingual corpus for a specific rectangular region in the image as shown in Figure 1. The bounded box region information (X, Y, width, height) for each of the images is provided in the dataset. The MMT system uses both text and the associated image to build the model to translate into the target Bengali text. We extracted the object tags as image features. Then the object tags are appended to the original English sentence which is then translated using mBART (Liu et al., 2020), a multilingual sequence to sequence model trained on millions of unsupervised multilingual sentences. We also perform a comparative study between the English-Bengali Text-to-text translation system and the built MMT system.

2 Related Work

There is limited research conducted in the domain of multimodal machine translation for Indian languages. Other than Hindi, no MMT system is available in other Indian languages due to the unavailability of the multimodal dataset for translation.

The Flickr30k dataset with Hindi description is used for multimodal NMT task by Dutta Chowdhury et al. (2018). They attempted to conduct multimodal translation from Hindi to English and examined whether visual image features can improve translation performance. They used synthetic Hindi descriptions for the Flickr30k dataset and provided validation and test corpus of English translations of the Flickr30k dataset. Similarly, Madaan et al. (2020) considered the Flickr30k dataset and asked five different crowd workers to provide Hindi translation of an image from the dataset and generated English captions with evaluating the quality of the translation.

Laskar et al. (2020) used Hindi Visual Genome 1.1 dataset (Parida et al., 2019) and used OpenNMT-py to build text-only NMT and multimodal NMT. They had used pre-trained CNN with VGG19 for extracting local and global features from the images for the multimodal translation. The multimodal NMT performs better as compared to text-only NMT.

3 Description of the MMT System

In this section, we describe the multimodal translation system developed for English →Bengali using the multimodal data which consists of images accompanying text. Our model is adapted from ViTA (Gupta et al., 2021)[1] which uses mBART (Liu et al., 2020), a multilingual sequence-to-sequence denoising auto-encoder that has been pre-trained using the BART objective (Lewis et al., 2020). Gupta et al. (2021) built a English→Hindi

[1] https://github.com/kshitij98/vita

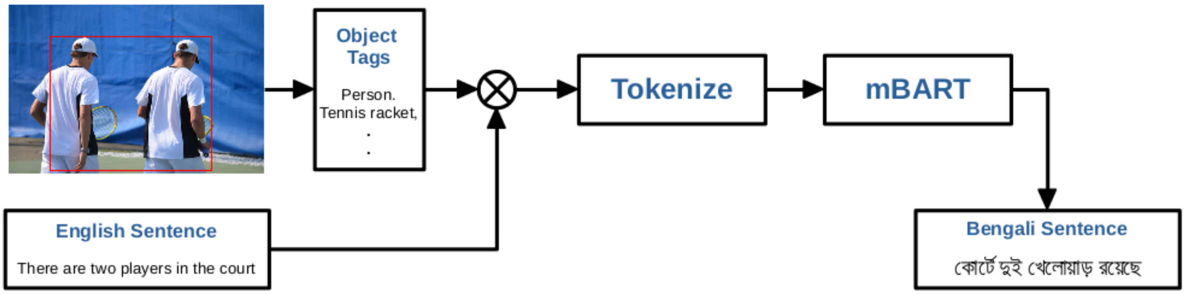


Figure 2: Multimodal machine translation. The object tags extracted from images along with the English source text input to the mBART to generate the Bengali translation output.

multimodal translation system by utilizing the object tags extracted from the images of the Hindi Visual Genome multimodal dataset (Parida et al., 2019) which is a dataset similar to the Bengali Visual Genome (see Section 4.1).

Similar to the ViTA approach, we first derive the list of object tags for a given image using the pretrained Faster R-CNN with ResNet-101-C4 backbone. Based on their confidence scores, we pick the top 10 object tags. In cases where less than 10 object tags are detected, we consider all the tags. The object tags are then concatenated to the English sentence which needs to be translated to Bengali. The concatenation is done using the special token '##' as the separator. The separator is followed by comma-separated object tags. Adding objects enables the model to utilize visual concepts which may not be readily available in the original sentence. The English sentences along with the object tags are fed to the encoder of the mBART model. The mBART's decoder generates the Bengali translations autoregressively. The block diagram of the multimodal translation using object tags is shown in Figure 2.

Gupta et al. (2021) applied ViTA for English to Hindi translation by using the mBART-25 model which has been pre-trained using the BART objective (Lewis et al., 2020). For this pre-training, only multilingual unsupervised data spanning 25 languages was used (Liu et al., 2020). Then they finetune the model for the machine translation task using 1.6 million English-Hindi parallel sentences. Finally, they finetune the model on the English-Hindi multimodal data with the addition of object tags by (a) first masking out 15% of the English tokens in the input and then (b) with no masking. For translating from English to Bengali, however, we do not perform a large-scale machine translation pre-training using a million training examples. We instead use the pre-trained mBART-50[2] model finetuned on the machine translation task in a one-to-many setup using multilingual data which contains merely 4487 English-Bengali parallel sentences (Tang et al., 2020). We take this pre-trained model and train it further on the machine translation task using the Bengali Visual Genome multimodal data by adding object tags to the English source sentences. Because of the scarcity of the pre-training machine translation English-Bengali parallel data (nearly 320 times smaller) as compared English-Hindi parallel data, our system represents a low-resource scenario.

Although the main approach we use is similar to the ViTA method by Gupta et al. (2021), we state the one modification in our implementation and the reason behind it. The original ViTA method stochastically masks out 15% of the tokens in the input English sentence. This is done to incentivize the model to utilize the object tags while generating the Bengali translation and not rely only on the English sentence. However, in our experiments, we do not mask out 15% of the English tokens. This is done because we already see gains above the text-only results without masking. Using masking can potentially improve the multimodal translation scores further.

[2] The mBART-50 model is trained using the same objective as the mBART-25 model. The difference is that the former supports 50 languages one of which is Bengali. Bengali is not supported by the latter.

4 Experiment Details

In this section, we first describe the dataset used followed by details of model training configurations.

4.1 BVG Dataset

Our experiments have been carried out using the BVG (Sen et al., (in press) dataset. The BVG dataset is segmented under four parts as the Training set, Development test set (D-Test), Evaluation test set (E-Test), and Challenge test set (C-Test). The BVG dataset statistics are shown in Table 1.

4.2 Translation using image and text modality

We used the large mBART one-to-many pre-trained model[3] in Huggingface Transformers library (Wolf et al., 2020). We did not freeze any model parameters during fine-tuning, therefore, the number of trainable parameters was 610M. The fine-tuning did not fit in the memory of a 28 GB GPU so we decreased the batch size to 1 and trained on a 48 GB GPU which was successful. The training time per epoch was 170 min. The model was fine-tuned for a maximum of 30 epochs. Adam optimizer (Kingma and Ba, 2014) was used with a learning rate of 1e-4. The training was stopped early if the development BLEU score did not improve for 5 consecutive epochs. The decoding beam size was set to 5. Model checkpoints were saved after every epoch and the best checkpoint was selected based on the development BLEU score.

4.3 Translation using text modality only

To demonstrate the impact of using image signals in the form of object tags, we conducted the experiments described in the previous section but without using any object tags. We did not modify any other configuration to ensure a fair comparison. We also note that adding object tags results in a large increase of tokens in each sentence. As a result, while not using object tags we observed that the training time per epoch reduced to 60 min, that is, the training was nearly 3 times faster.

4.4 Translation using Text-to-text transformer

In addition to the mBART pre-trained models, we also experimented with training a plain transformer model (Vaswani et al., 2017) from scratch. We first trained sentencepiece subword units (Kudo and Richardson, 2018) setting maximum vocabulary size to 8k. The vocabulary was learned jointly on the source and target sentences of the Bengali Visual Genome training dataset. The implementation was done using PyTorch (Paszke et al., 2019). The number of encoder and decoder layers was set to 3 each and the number of heads was set to 8. The hidden size was set to 128, along with the dropout value of 0.1. We initialized the model parameters using Xavier initialization (Glorot and Bengio, 2010) and used the Adam optimizer (Kingma and Ba, 2014) with a learning rate of $5e - 4$ for optimizing model parameters. Gradient clipping was used to clip gradients greater than 1. The training was stopped when the development loss did not improve for 5 consecutive epochs. For generating translations, we used greedy decoding and generated tokens auto-regressively till the end-of-sentence token was generated or the maximum translation length was reached, which was set to 100.

5 Result and Discussion

We have used the popular machine translation metric BLEU (Papineni et al., 2002) for the automatic evaluation, computed using sacreBLEU toolkit (Post, 2018).

The development BLEU scores during training are shown in Figure 3.

The development BLEU score increases as the training progress. The mBART based scores (both Text-to-text and Multimodal) reach a notably high BLEU score even after one epoch of training. This is because of the prior knowledge acquired from pre-training, which is missing in the case of the Text-to-text transformer.

The MMT results on the D-Test, E-Test, and C-Test are shown in Table 2. The C-Test scores are consistently lower than D-Test and E-Test scores, indicating that the C-Test consists of more challenging segments which are harder to translate to Bengali. The Text-to-

[3] https://huggingface.co/facebook/mbart-large-50-one-to-many-mmt

Dataset	#Sentences	#Tokens	
		EN	BN
Train	28930	143156	113993
D-Test	998	4922	3936
E-Test	1595	7853	6408
C-Test	1400	8186	6657

Table 1: Statistics of BVG for experiments. The number of tokens for English (EN) and Bengali (BN) for each set are reported.

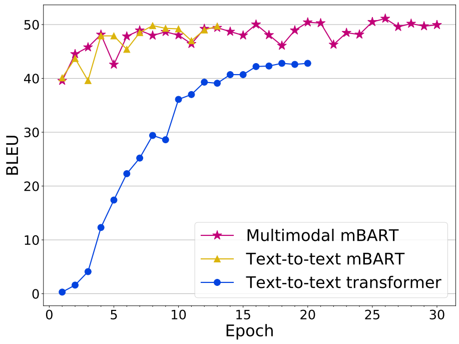


Figure 3: Development BLEU scores during training.

text mBART performs better as compared to the Text-to-text transformer model. The multimodal mBART performs the best overall.

The object tags added to the original English sentences provide more context about the image and enable the generation of a better translation, which is indicated by the higher overall BLEU scores. The improvement is seen in translating C-Test as well.

We performed the comparison of MMT system with the best Text-to-text translation system without using any image features. There is a performance improvement (BLEU score) of **+1.3** on D-Test, **+3.9** on E-Test, and **+0.9** on C-Test. Apart from performance, MMT systems help to resolve ambiguities as shown in the Table 3.

The MMT system can correctly translate the ambiguous word *court* which the Text-to-text MT system fails. We compared the translation output between both Text-to-text and MMT systems and observed the MMT system produces a better translation, correct word order, no ambiguity as compared with the Text-to-text MT system.

6 Manual Evaluation

To validate the automatic scoring, we manually annotated 100 randomly selected sentences from the C-Test set as translated by the Text-to-text machine translation system and MMT system. The annotation was performed by the native Bengali speaker. Bengali Captions in the MMT translation outcomes fall under five different sets where some of them are translated perfectly without any issue, some of them are very close to perfect, some of them have parts of speech or grammar issues, some of them have ambiguity in meanings, and some of them have lack of words than original annotation.

In this annotation, each annotated segment gets exactly one label from the following set (Parida and Bojar, 2018):

Flawless for translations without any error (typesetting issues with diacritic marks due to different tokenization are ignored),

Good for translations which are generally OK and complete but need a small correction,

Partly Correct for cases where a part of the segment is correct but some words are mistranslated,

Ambiguity for segments where the MT system “misunderstood” a word’s meaning, and

Incomplete for segments that run well but stop too early, missing some content words. This category also includes the relatively rare cases where the Text-to-text or MMT system produced just a single word, unrelated to the source.

The manual evaluation results are summarized in following graph Figure 4. The MMT system generates more flawless and good translation output as compared to the Text-to-text system (see Figure 4). The Text-to-text system obtained more partial correct and incomplete translation output. It observed that still

MT System	D-Test BLEU	E-Test BLEU	C-Test BLEU
Text-to-text transformer	42.8	35.6	17.2
Text-to-text mBART	49.8	39.6	25.9
Multimodal mBART	**51.1**	**43.5**	**26.8**

Table 2: Text only and multimodal translation performance on the BVG dataset.

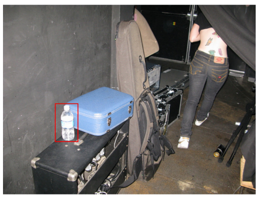
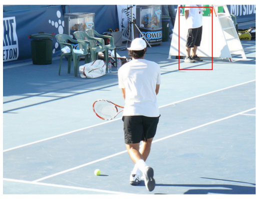
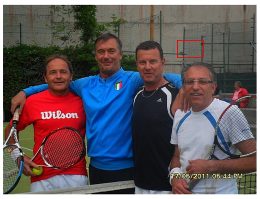
Input Image	Input Caption	Text-to-text Result	MMT Result
	The water bottle on the stand	স্ট্যান্ডে জলের বোতল	স্ট্যান্ডে জলের বোতল
		"Water bottle on the stand"	"Water bottle on the stand"
	Two people waiting to cross	দুজন লোক ক্রস অপেক্ষা করছে	দুজন লোক ক্রস অপেক্ষা করছে
		"Two people are waiting cross"	"Two people are waiting cross"
	Man standing on a tennis court	টেনিস কোর্টে দাঁড়িয়ে লোক	টেনিস কোর্টে দাঁড়িয়ে লোক
		"Man standing on a tennis court"	"Man standing on a tennis court"
	stamp on boy's left hand	ছেলেটির বাম হাতে স্ট্যাঙ্ক	ছেলেটির বাম হাতে স্ট্যাম্প
		"Stank on boy's left hand" (incorrect Bengali word 'Stank' obtained in T2T translation)	"Stamp on boy's left hand" (correct Bengali word 'stamp' obtained in MMT translation)
	fence around the court	আদালতের চারদিকে বেড়া	কোর্টের চারপাশে বেড়া
		"Fence around the court" (court is translated by T2T as *Judicial Court* in Bengali)	"Fence around the court" (court is translated by MMT as *Tennis Court* in Bengali)

Table 3: Samples of Text-to-text and Multimodal Translation obtained from the Text-to-text mBART and the Multimodal mBART systems. First two columns from left provide the input image and its corresponding English caption. The third and fourth columns are the Bengali captions generated by Text-only and Multimodal translation systems. For each Bengali caption, we also provide the English translation.

there are ambiguities exist in the translation output of both systems. Some translation samples are shown in Figure 5.

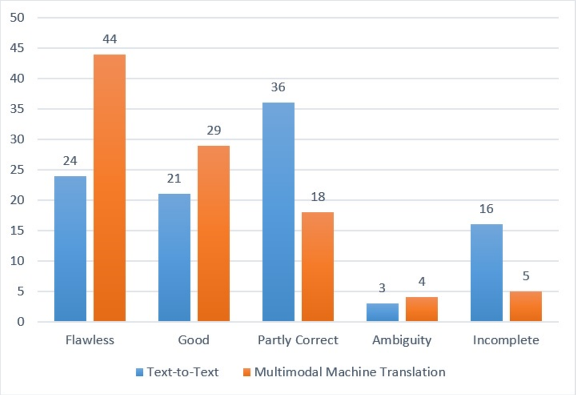


Figure 4: Manual evaluation summary. Out of 100 translation samples, five categories are chosen for observing Text-to-text and Multimodal Machine Translation Accuracy

7 Conclusion

In this paper, we build an English-Bengali MMT system utilizing bi-lingual text and the associated images which improves the translation quality (based on an automatic evaluation) and resolve ambiguities. Our work helps to build a better English-Bengali MT system and encourages researchers to explore the MMT system for Bengali. The future work includes exploring other state-of-the-art MMT systems on the BVG dataset and performs a comparison analysis (Tamura et al., 2020; Caglayan et al., 2021; Tan et al., 2020; Liu et al., 2021).

Acknowledgments

The authors Shantipriya Parida and Petr Motlicek were supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 833635 (project ROXANNE: Real-time network, text, and speaker analytics for combating organized crime, 2019-2022).

The authors do not see any significant ethical or privacy concerns that would prevent the processing of the data used in the study. The datasets do contain personal data, and these are processed in compliance with the GDPR and national law.

Flawless/Good:
Window on second level
দ্বিতীয় মেঝেতে উইন্ডোটি Gloss: Window on second level
Blue tennis court area
নীল টেনিস কোর্ট অঞ্চল Gloss: Blue tennis court area
Three baseball players on baseball field.
তিনটি বেসবল খেলোয়াড় মাঠে রয়েছে Gloss: Three baseball players on baseball field.
Partly Correct:
Railing on second floor of building
দ্বিতীয় তলায় রেলিংয়ে জেলিং Gloss: Gelling on second floor railing (output does not convey the intended meaning in the target language)
The hand is going to press a button
হাত জন একটি বোতামে যাচ্ছে Gloss: The hand is going to a button ('press' is completely omitted from the translated outcome)
Ambiguity:
Brown tennis court floor
আদালতে ধূসর রং Gloss: Brown color on Court Floor (Confusing Tennis Court with Judicial)
Street sign on a pole in English and Chinese
ইংলিশ এবং চীনা ভাষায় একটি মেরুতে রাস্তার চিহ্ন Gloss: Street sign on a pole in English and Chinese (confusing street-side pole with 'south pole and north pole' in Bengali)

Figure 5: Sample segment translations and their manual classification taken from MMT translation of Challenge Test Set.

References

Ozan Caglayan, Loïc Barrault, and Fethi Bougares. 2016. Multimodal attention for neural machine translation. *arXiv preprint arXiv:1609.03976*.

Ozan Caglayan, Menekse Kuyu, Mustafa Sercan Amac, Pranava Swaroop Madhyastha, Erkut Erdem, Aykut Erdem, and Lucia Specia. 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: Main Volume*, pages 1317–1324.

Ozan Caglayan, Pranava Swaroop 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.

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *EMNLP*.

Koel Dutta Chowdhury and Desmond Elliott. 2019. Understanding the effect of textual adversaries in multimodal machine translation. In *Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pages 35–40.

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.

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.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy. PMLR.

Kshitij Gupta, Devansh Gautam, and Radhika Mamidi. 2021. Vita: Visual-linguistic translation by aligning object tags. *arXiv preprint arXiv:2106.00250*.

Barry Haddow and Faheem Kirefu. 2020. Pmindia–a collection of parallel corpora of languages of india. *arXiv preprint arXiv:2001.09907*.

Md Arid Hasan, Firoj Alam, Shammur Absar Chowdhury, and Naira Khan. 2019. Neural machine translation for the bangla-english language pair. In *2019 22nd International Conference on Computer and Information Technology (ICCIT)*, pages 1–6. IEEE.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M Sohel Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for bengali-english machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2612–2623.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

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.

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.

Xiao Liu, Jing Zhao, Shiliang Sun, Huawen Liu, and Hao Yang. 2021. Variational multimodal machine translation with underlying semantic alignment. *Information Fusion*, 69:73–80.

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. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Aman Madaan, Shruti Rijhwani, Antonios Anastasopoulos, Yiming Yang, and Graham Neubig. 2020. Practical comparable data collection for low-resource languages via images. *arXiv preprint arXiv:2004.11954*.

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.

Shantipriya Parida and Ondřej Bojar. 2018. Translating short segments with nmt: A case study in english-to-hindi.

Shantipriya Parida, Ondřej Bojar, and Satya Ranjan Dash. 2019. Hindi visual genome: A dataset for multimodal english-to-hindi machine translation. *arXiv preprint arXiv:1907.08948*.

Shantipriya Parida, Petr Motlicek, Amulya Ratna Dash, Satya Ranjan Dash, Debasish Kumar Mallick, Satya Prakash Biswal, Priyanka Pattnaik, Biranchi Narayan Nayak, and Ondřej Bojar. 2020. Odianlp's participation in wat2020. In *Proceedings of the 7th Workshop on Asian Translation*, pages 103–108.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, et al. 2021. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *arXiv preprint arXiv:2104.05596*.

Arghyadeep Sen, Shantipriya Parida, Ketan Kotwal, Subhadarshi Panda, Ondřej Bojar, and Satya Ranjan Dash. (in press) 2021. Bengali visual genome: A multimodal datasetfor machine translation and image captioning. In *Proceedings of 9th International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA)*. Springer.

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*, 34(2):97–147.

Hiroto Tamura, Tosho Hirasawa, Masahiro Kaneko, and Mamoru Komachi. 2020. Tmu japanese-english multimodal machine translation system for wat 2020. In *Proceedings of the 7th Workshop on Asian Translation*, pages 80–91.

Liang Tan, Lin Li, Yifeng Han, Dong Li, Kaixi Hu, Dong Zhou, and Peipei Wang. 2020. An empirical study on ensemble learning of multimodal machine translation. In *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, pages 63–69. IEEE.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350.