English-to-Low-Resource Translation: A Multimodal Approach for Hindi, Malayalam, Bengali, and Hausa

Ali Hatami[1,4] and **Shubhanker Banerjee**[2,4] and **Mihael Arcan**[3] and
Bharathi Raja Chakravarthi[1,4] and **Paul Buitelaar**[1,4] and **John Philip McCrae**[1,2,4]
[1]Insight SFI Research Centre for Data Analytics, Data Science Institute, University of Galway
[2]ADAPT Centre, Data Science Institute, University of Galway
[3]Lua Health, Galway, Ireland
[4]School of Computer Science, University of Galway
ali.hatami@insight-centre.org

Abstract

Multimodal machine translation leverages multiple data modalities to enhance translation quality, particularly for low-resourced languages. This paper uses a multimodal model that integrates visual information with textual data to improve translation accuracy from English to Hindi, Malayalam, Bengali, and Hausa. This approach employs a gated fusion mechanism to effectively combine the outputs of textual and visual encoders, enabling more nuanced translations that consider both language and contextual visual cues. The model's performance was evaluated against the text-only machine translation model based on BLEU, ChrF2 and TER. Experimental results demonstrate that the multimodal approach consistently outperforms the text-only baseline, highlighting the potential of integrating visual information in low-resourced language translation tasks.

1 Introduction

In recent years, neural network-based translation models have been widely used in translation tasks, demonstrating remarkable performance in terms of fluency and precision compared to previous generations of machine translation systems (Cho et al., 2014). The Transformer model, in particular, has shown significant improvements in machine translation tasks. A crucial component of the Transformer model is the cross-attention mechanism, which enhances the model's ability to capture semantic dependencies by combining self-attention—allowing source words to interact with one another—with attention mechanisms that involve target words (Vaswani et al., 2017).

Despite the broader context focus in text-only translation models, understanding the input text remains a challenge. In natural language, lexical ambiguity (Rios Gonzales et al., 2017) occurs when a single word has multiple meanings or interpretations, complicating text comprehension. For example, in the domain of finance and economics, the word "bank" almost always refers to a financial institution rather than the side of a river.

Multimodal Machine Translation (MMT), a subarea of NMT, has been introduced to utilise visual information from other modalities, such as images, to translate an aligned sentence in a source language into a target language. Recent studies (Yao and Wan, 2020; Zhao et al., 2022; Wang and Xiong, 2021) demonstrate the potential of leveraging multimodal information, alongside textual content, to enhance translation quality. Visual cues, as an additional source of information, can provide valuable insights that complement textual information, enabling MMT models to better understand and produce more accurate and contextually appropriate translations. The concept behind MMT is to integrate visual information to help disambiguate input words, detect the correct scenes in the source language, and select the appropriate translation in the target language (Hatami et al., 2022). MMT is particularly beneficial when dealing with low-resource languages where there is not sufficient parallel data to train the model.

This paper aims to explore the benefit of using visual information in translating English into four different low-resource languages, Hindi, Malayalam, Bengali and Hausa. We used a gated fusion approach to integrate textual and visual information in the encoder and generate the text in the target language on the decoder side. In the baseline, we train the model on the input text without considering the aligned image. For the multimodal model, we trained four different models for each language. We explain our methodology in Section 3, our experimental setup in Section 4, results in Section 5, and we conclude our findings in Section 6.

2 Related Work

There are various approaches proposed to integrate visual information with text-only translation models. These approaches typically utilise a visual

Proceedings of the Ninth Conference on Machine Translation, pages 815–822

attention mechanism in either the decoder or encoder to capture the relationships between words in a sentence and image features. The common method involves extracting visual information by employing Convolutional Neural Networks (CNN) and then integrating this information with textual features.

Regarding visual features, existing studies on MMT employ two types of visual features: global and local visual features. Global features represent the entire image as a single vector without attention to the spatial layout of the image. On the other hand, local features describe an image as a sequence of equally sized patches (Calixto et al., 2017). Local features are extracted from multiple points in the image and are more robust to clutter than global features (Lisin et al., 2005). CNNs can be used to extract both global and local features from the image (Zheng et al., 2019).

Global image features are used in the encoder in addition to word sequences (Huang et al., 2016). Alternatively, they can be used to initialise the hidden parameters of the encoder and decoder of a RNN (Calixto and Liu, 2017). Element-wise multiplication was used to initialise the hidden states of the encoder/decoder in the attention-based model (Caglayan et al., 2017). Visual attention mechanism was employed to link visual and corresponding text semantically (Zhou et al., 2018).

Several approaches have been proposed to improve the quality of the visual modality in Multimodal Machine Translation (MMT). For instance, a multimodal Transformer-based self-attention mechanism was introduced to encode relevant information in images (Yao and Wan, 2020). A graph-based multimodal fusion encoder was developed to capture various relationships between modalities (Yin et al., 2020). Additionally, a translate-and-refine mechanism was implemented using images in a second-stage decoder to refine a text-only Neural Machine Translation (NMT) model, particularly for handling ambiguous words. A latent variable model was also employed to extract the multimodal relationships between image and text modalities (Calixto et al., 2019).

Recent methods aim to reduce noise in visual information and select visual features relevant to the text. For example, object-level visual modelling has been used to mask irrelevant objects and specific words in the source text to enhance visual feature learning (Wang and Xiong, 2021). Object detection in the image encoder has been employed to extract visual features from object regions within an image, which are then applied to a doubly-attentive decoder model (Zhao et al., 2022).

In this paper, we adopt the gated fusion MMT model (Wu et al., 2021), which integrates visual and textual representations through a gate mechanism. This gated fusion mechanism allows the model to adjust the amount of visual information that contributes to the translation process.

3 Methodology

The objective of our experiments is to evaluate the impact of visual features on translation quality in low-resource languages. Following Wu et al. (2021), we conduct experiments to assess both the text-only Transformer and the gated fusion multimodal Transformer (gated fusion MMT) using the shared task data for Hindi, Bengali, Malayalam, and Hausa. In this section, we provide descriptions of the model architectures mentioned above.

3.1 Text-only Machine Translation

For the text-only translation model, we use the training and development sets for Hindi, Bengali, Malayalam, and Hausa to train the Transformer-based model. This model serves as our baseline for evaluating the multimodal model. The text-only Transformer architecture was introduced by Vaswani et al. (2017). It consists of an encoder-decoder structure, where both the encoder and decoder are composed of stacked layers of self-attention, and feed-forward neural networks.

First, we tokenize the sentences into subwords in the training, development, and test sets. We then train four translation models on the tokenized sentences for these language pairs. Tokenization helps the model better learn the language and handle out-of-vocabulary words, especially in low-resource languages. During the inference step, we translate the tokenized test sentences from English into the four low-resource languages.

3.2 Multimodal Machine Translation

For the multimodal model, we use the gated fusion approach (Wu et al., 2021) to fuse both textual and visual information. Gated fusion MMT incorporates visual information into the translation process in a controlled and interpretable manner using a gating mechanism. The textual component is similar to the text-only model, with tokenized sentences

Dataset	Hindi	Bengali	Malayalam	Hausa
Training Set	28,932	28,930	29,000	28,930
Development Set (D-Test)	998	998	1,000	998
Evaluation Set (E-Test)	1,595	1,595	1,600	1,595
Challenge Test Set (C-Test)	1,400	1,400	1,400	1,400
Total	32,925	32,923	33,000	32,923

Table 1: Number of sentences of Visual Genome dataset for Hindi, Bengali, Malayalam and Hausa.

fed into the model. On the visual side, each sentence is paired with an image, and for each image, we have the coordinates of the rectangular region corresponding to the part of the image that relates to the sentence (see Figure 1).

For each language, we trained two models: one that considers the entire image and another that considers only the specific rectangular region. We use the pre-trained ResNet-101 CNN (He et al., 2016) to extract visual features from the images. In this study, we extract visual representations from both the whole image and the designated rectangular region, which is aligned with the text caption. The motivation for using the partial image (rather than the full image) is that objects outside the rectangular region may be irrelevant to the text caption and could potentially degrade translation model performance (Hatami et al., 2023).

Both the textual and visual representations are fed into the gated fusion model, allowing it to be trained based on both modalities. We then use these multimodal models to translate test sentences that are aligned with images. More detailed information about the multimodal models can be found in Section 4.2.2.

4 Experimental Setup

4.1 Dataset

The Hindi Visual Genome (HVG) (Parida et al., 2019), Bengali Visual Genome (BVG) (Sen et al., 2022), Malayalam Visual Genome (MVG) (Parida et al., 2019), and Hausa Visual Genome (HaVG) (Abdulmumin et al., 2022) datasets are multimodal datasets designed for English-to-Hindi, English-to-Bengali, English-to-Malayalam, and English-to-Hausa machine translation, respectively (Figure 1). These datasets, based on the original Visual Genome dataset, contain real-world images annotated with region-specific captions. The captions have been translated into the respective languages through a combination of automated translation and manual post-editing by native speakers to ensure contextual accuracy.

The MVG, HVG, BVG, and HaVG datasets are divided into training, development, evaluation, and challenge test sets, as outlined in Table 1.

Training Set: The training sets for Malayalam, Hindi, Bengali, and Hausa contain 29,000, 28,932, 28,930, and 28,930 image-caption pairs, respectively. Each pair consists of an image, a selected region in the image, and its corresponding English and Malayalam/Hindi/Bengali/Hausa captions. The captions have been manually refined to align with the visual context of the images.

Development Set (D-Test): The development sets contain 1,000 image-caption pairs in the Malayalam dataset and 998 pairs in the Hindi, Bengali, and Hausa datasets. These sets are used to validate and fine-tune model performance during the training process.

Evaluation Set (E-Test): The evaluation sets include 1,600 image-caption pairs in the Malayalam dataset and 1,595 pairs in the Hindi, Bengali, and Hausa datasets. These sets are used for evaluating model performance on unseen data, providing a benchmark for generalization capabilities.

Challenge Test Set (C-Test): The challenge test sets for all four languages consist of 1,400 image-caption pairs. These sets are designed to focus on ambiguous English words that require visual context to resolve their meaning in Malayalam, Hindi, Bengali, or Hausa. The ambiguous words were identified based on embedding similarity, and the corresponding images help disambiguate their meaning, providing a robust test for multimodal translation systems (Hatami et al., 2024).

4.2 Machine Translation Models

4.2.1 Text-only Translation Model

A text-only Transformer model serves as the baseline in our experiment, utilizing only the textual captions of images for translation. The model is trained using the OpenNMT toolkit (Klein et al., 2018) on the Visual Genome dataset for English-

Coordinates of the rectangular region: 20, 150, 325, 121

Text caption of the region:

- English: many giraffes at a zoo
- Hindi: एक चिड़ियाघर में कई जिराफ
- Bengali: একটি চিড়িয়াখানায় অনেক জিরাফ
- Malayalam: ഒരു മൃഗശാലയിലെ നിരവധി ജിറാഫുകൾ
- Hausa: rakuman ruwa da yawa a gidan namun daji

Coordinates of the rectangular region: 61, 191, 437, 182

Text caption of the region:

- English: fruit stand outside market
- Hindi: बाजार के बाहर फल स्टैंड
- Bengali: বাজারের বাইরে ফলের স্ট্যান্ড
- Malayalam: ഫ്രൂട്ട് സ്റ്റാൻഡ് മാർക്കറ്റിന് പുറത്താണ്
- Hausa: 'ya'yan itace suna tsaye a waje da kasuwa

Figure 1: Examples from the Visual Genome dataset show English caption of the rectangular region (solid red line) with translation in Hindi, Bengali, Malayalam and Hausa.

to-Hindi, Bengali, Malayalam, and Hausa translations. It comprises a 6-layer Transformer architecture with attention mechanisms in both the encoder and decoder stages, trained for 50k steps.

The encoder processes a sequence of tokens (words or subword units) and generates context-aware representations for each token. The decoder generates the output sequence (e.g., translated text) by leveraging the encoded representations from the encoder along with the previously generated tokens. It employs multi-head self-attention and feed-forward layers, incorporating additional attention mechanisms to effectively focus on the encoded input. The core innovation of the Transformer is the **self-attention mechanism**, which computes attention scores across all tokens in the sequence, creating weighted representations that capture contextual relationships between tokens.

Since the Transformer model does not inherently process sequences in a fixed order, as recurrent neural networks (RNNs) do, it uses **positional encodings** to inject information about the position of tokens in the sequence. These positional encodings are added to the input embeddings, enabling the model to differentiate between tokens based on their positions within the sequence. To enhance its ability to capture different types of relationships between tokens, the Transformer employs **multi-head attention**. This involves splitting the self-attention process into multiple parallel attention heads, each learning a different set of attention weights. The outputs from all heads are then concatenated and linearly transformed to provide a richer, more comprehensive representation of the input sequence.

SentencePiece (Kudo and Richardson, 2018) is employed to segment words into subword units, offering a language-independent approach to tokenization without requiring pre-processing steps, thereby enhancing the model's adaptability and versatility in handling raw text.

4.2.2 Multimodal Machine Translation

In the MMT model, we adopt the gated fusion MMT model (Wu et al., 2021), which fuses visual and text representations by employing a gate mechanism. Gated fusion is a mechanism used to integrate visual information from images with textual information from source sentences during the translation process. The main idea behind gated fusion is to control the amount of visual information that is blended into the textual representation using a gating matrix.

The source sentence x is fed into a vanilla Transformer encoder to obtain a textual representation H_{text} of dimension $T \times d$ [1]. The image z is processed using a pre-trained ResNet-101 CNN (He et al., 2016), which has been trained on the ImageNet dataset (Russakovsky et al., 2014), to extract a 2048-dimensional average-pooled visual representation, denoted as $Embed_{image}(z)$. The visual representation $Embed_{image}(z)$ is projected to the

[1] T is the number of tokens (words) in the input sentence, and d is the dimensionality of the representation

English → Hindi	BLEU ↑	ChrF2 ↑	TER ↓
Text-only MT	38.26	58.65	42.54
Multimodal MT (entire image)	**39.65***	**59.34***	**41.92***
Multimodal MT (partial image)	38.64	58.84	42.62
English → Bengali	BLEU ↑	ChrF2 ↑	TER ↓
Text-only MT	39.85	64.32	39.24
Multimodal MT (entire image)	**41.92***	**65.96***	**38.37***
Multimodal MT (partial image)	39.45	64.75	39.65
English → Malayalam	BLEU ↑	ChrF2 ↑	TER ↓
Text-only MT	28.94	58.74	54.87
Multimodal MT (entire image)	**32.34***	**61.15***	**53.94***
Multimodal MT (partial image)	28.76	58.63	54.58
English → Hausa	BLEU ↑	ChrF2 ↑	TER ↓
Text-only MT	39.86	61.21	47.59
Multimodal MT (entire image)	**41.25***	**62.94***	**46.48***
Multimodal MT (partial image)	38.31	60.87	47.62

Table 2: BLEU, ChrF2 and TER scores for text-only and multimodal models for English to Hindi, Bengali, Malayalam and Hausa on the test set (* represents a statistically significant result compared to the baseline text-only model at a significance level of $p < 0.05$).

same dimension as H_{text} using a weight matrix W_z, denoted as:

$$\text{Embed}_{\text{image}}(z) = W_z \text{ResNet}_{\text{pool}}(z)$$

where W_z is a learned projection matrix.

To determine the amount of visual information to fuse with the textual representation, a gating matrix Λ of dimension $T \times d$ is generated ($[0, 1]^{T \times d}$). This matrix is computed using a sigmoid function applied to both the projected visual representation and the textual representation:

$$\Lambda = \sigma\left(W_\Lambda \text{Embed}_{\text{image}}(z) + U_\Lambda H_{\text{text}}\right)$$

where W_Λ and U_Λ are learned parameters, and σ is the sigmoid function. The gating matrix Λ makes the fusion process interpretable, as it controls how much visual context is used in translation. A larger value in Λ indicates that the model is relying more on the visual context, while a smaller value indicates a stronger reliance on the textual representation alone.

The final representation H that combines both textual and visual information is given by:

$$H = H_{\text{text}} + \Lambda \text{Embed}_{\text{image}}(z)$$

This fused representation H is then passed into the Transformer decoder for generating the target translation.

4.3 Evaluation Metrics

We use three evaluation metrics: BLEU (Papineni et al., 2002), ChrF2 (Popović, 2015), and TER (Snover et al., 2006). BLEU assesses the precision of translation by comparing candidate translations to reference translations based on *n-grams*. ChrF2 evaluates the similarity between character *n-grams* in machine-generated and reference translations, particularly beneficial for languages with complex writing systems. TER quantifies the number of edits needed to align machine translations with human-generated references. We conduct statistical significance testing using the *sacreBLEU*[2] toolbox.

[2] https://github.com/mjpost/sacrebleu

5 Results and Discussion

In this section, we present the results of our experiments, where we trained our models on the Visual Genome dataset and evaluated the translation quality using the BLEU, ChrF2, and TER metrics. We compare the translation quality of our proposed models with text-only baseline models, where the text-only NMT model was trained solely on text captions without images, across test sets for four languages.The MMT models were trained on both text captions and original images with entire images and just considering the coordinates of a part of the image related to the caption (partial image).

The results in Table 2 demonstrate the performance of both text-only and multimodal models across four language pairs: English to Hindi, Bengali, Malayalam, and Hausa. For English to Hindi, the MMT model that utilizes the entire image outperforms the text-only model, achieving a BLEU score of 39.65, ChrF2 score of 59.34, and TER score of 41.92. These improvements are statistically significant over the text-only MT model at $p < 0.05$, highlighting the benefit of incorporating visual context into the translation process. Similar trends are observed for English to Bengali, where the entire image-based MMT achieves a BLEU score of 41.92, a ChrF2 score of 65.96, and a TER score of 38.37, all of which are significantly better than the text-only model.

For English to Malayalam, the entire image-based multimodal model also shows clear advantages, with a BLEU score of 32.34, ChrF2 of 61.15, and TER of 53.94, outperforming the text-only model on all metrics. Finally, in the case of English to Hausa, the entire image-based multimodal MT model again demonstrates superior performance, achieving a BLEU score of 41.25, ChrF2 of 62.94, and TER of 46.48, compared to the text-only model. Across all language pairs, the partial image-based multimodal models do not consistently outperform the text-only models, suggesting that complete visual context is necessary for achieving the best translation quality.

6 Conclusion

This paper demonstrates the significant advantages of employing a multimodal machine translation approach that integrates visual information with textual data, especially in the case of low-resourced languages like Hindi, Malayalam, Bengali, and Hausa. The results indicate that the gated fusion MMT model enhances translation accuracy and provides a more nuanced understanding of context, leading to improved performance over traditional text-only models. By leveraging visual context, we can address the challenges faced in translating low-resourced languages, highlighting the importance of incorporating diverse data modalities to enrich the translation process.

Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Numbers SFI/12/RC/2289_P2 (Insight) and 13/RC/2106_P2 (ADAPT), co-funded by the European Regional Development Fund. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

Idris Abdulmumin, Satya Ranjan Dash, Musa Abdullahi Dawud, Shantipriya Parida, Shamsuddeen Muhammad, Ibrahim Sa'id Ahmad, Subhadarshi Panda, Ondřej Bojar, Bashir Shehu Galadanci, and Bello Shehu Bello. 2022. Hausa visual genome: A dataset for multi-modal English to Hausa machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6471–6479, Marseille, France. European Language Resources Association.

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.

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.

Iacer Calixto, Miguel Rios, and Wilker Aziz. 2019. Latent variable model for multi-modal translation. In

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 6392–6405, Florence, Italy. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Ali Hatami, Mihael Arcan, and Paul Buitelaar. 2024. Enhancing translation quality by leveraging semantic diversity in multimodal machine translation. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 154–166, Chicago, USA. Association for Machine Translation in the Americas.

Ali Hatami, Paul Buitelaar, and Mihael Arcan. 2022. Analysing the correlation between lexical ambiguity and translation quality in a multimodal setting using WordNet. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 89–95, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Ali Hatami, Paul Buitelaar, and Mihael Arcan. 2023. A filtering approach to object region detection in multimodal machine translation. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 393–405, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 639–645, Berlin, Germany. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander Rush. 2018. OpenNMT: Neural machine translation toolkit. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 177–184, Boston, MA. Association for Machine Translation in the Americas.

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.

D.A. Lisin, M.A. Mattar, M.B. Blaschko, E.G. Learned-Miller, and M.C. Benfield. 2005. Combining local and global image features for object class recognition. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, pages 47–47.

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 Multi-modal English to Hindi Machine Translation. *Computación y Sistemas*, 23(4):1499–1505. Presented at CICLing 2019, La Rochelle, France.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark. Association for Computational Linguistics.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. 2014. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575.

Arghyadeep Sen, Shantipriya Parida, Ketan Kotwal, Subhadarshi Panda, Ondřej Bojar, and Satya Ranjan Dash. 2022. Bengali Visual Genome: A Multimodal Dataset for Machine Translation and Image Captioning. In *Intelligent Data Engineering and Analytics*, pages 63–70. Springer.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Dexin Wang and Deyi Xiong. 2021. Efficient object-level visual context modeling for multimodal machine translation: Masking irrelevant objects helps grounding. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 2720–2728. AAAI Press.

Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6153–6166. 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, Online. Association for Computational Linguistics.

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 *Annual Meeting of the Association for Computational Linguistics*.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2022. Region-attentive multimodal neural machine translation. *Neurocomputing*, 476:1–13.

Yufeng Zheng, Jun Huang, Tianwen Chen, Yang Ou, and Wu Zhou. 2019. CNN classification based on global and local features. In *Real-Time Image Processing and Deep Learning 2019*, volume 10996, page 109960G. International Society for Optics and Photonics, SPIE.

Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. 2018. A visual attention grounding neural model for multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3643–3653, Brussels, Belgium. Association for Computational Linguistics.