Towards Grounded Evaluation of Multimodal Machine Translation Systems

Sami Ul Haq and Sheila Castilho
ADAPT Centre
Dublin City University, Ireland
{firstname.lastname}@adaptcentre.ie

Abstract

Multimodal Machine Translation (MMT) represents a promising research direction for improving the translation of lexically ambiguous text by leveraging auxiliary inputs such as images. However, the lack of robust evaluation methods for assessing such models remains a significant bottleneck to progress in the field. Therefore, we introduce the task of visually grounded evaluation of MMT systems, bringing visual modality directly into the assessment process. To facilitate this, we present MuTE (Multimodal Machine Translation Evaluation), a novel dataset constructed by aligning high-quality human evaluations with public image datasets. We further propose a simple evaluation method based on Visual Language Models (VLMs) to quantify translation quality using relevant images as a grounding signal. The resources produced have several use cases beyond MMT, including the evaluation of multilingual image captioning, information retrieval, and cross-modal reasoning.

To facilitate reproducibility, the source code, dataset[1], and a working demo[2] developed during this research are made publicly available.

1 Introduction

Traditional machine translation (MT) relies exclusively on textual data, overlooking valuable information available from other modalities. MMT addresses this by integrating complementary inputs such as images to improve translation quality, particularly in cases where source text is ambiguous (Baltrušaitis et al., 2018).

Several MMT models (Helcl et al., 2018; Calixto et al., 2017; Ive et al., 2019; Lin et al., 2020) have been proposed, demonstrating strengths over text-only baselines. However, Lala et al. (2018) and Raunak et al. (2019) observe that there is no significant performance gain with multimodal systems, at least when measured via automatic evaluation techniques. Wu et al. (2021) further argued that the visual modality primarily serves as a form of regularization during training rather than meaningfully complementing textual input. In response, Li et al. (2021) and Futeral et al. (2022) have questioned the nature of existing evaluation benchmarks and training datasets based on Multi30K (Elliott et al., 2016)—characterizing them as simple, short and biased—suggesting they represent a major bottleneck in MMT progress.

In this research, we aim to overcome these limitations in MMT evaluation by proposing a vision-grounded assessment framework. The goal is to incorporate image content when evaluating translation outputs, in contrast to existing approaches that rely solely on similarity measures at the word or embedding level. Our contributions are: (i) a method for constructing a high-quality Multimodal MT Evaluation (MuTE) dataset by pairing existing human quality annotations with corresponding images, and (ii) a Visually Grounded Evaluation (VGE) that combines similarity scores of source-candidate-image pairs, bridging visual and textual modalities across languages.

We build and evaluate these resources for English-to-German, French, Spanish, and Czech translation.

2 Multimodal MT Evaluation (MuTE) dataset

2.1 Generating the MuTE dataset

The Workshop on Machine Translation (WMT)[3] annually produces extensive human evaluation data to compare and rank participating systems in shared

[1] https://github.com/sami-haq99/mmt_dataset

[2] https://huggingface.co/spaces/samiulhaq/Multilingual_Cross_model_semantic_relevance?logs=container

[3] https://www2.statmt.org/

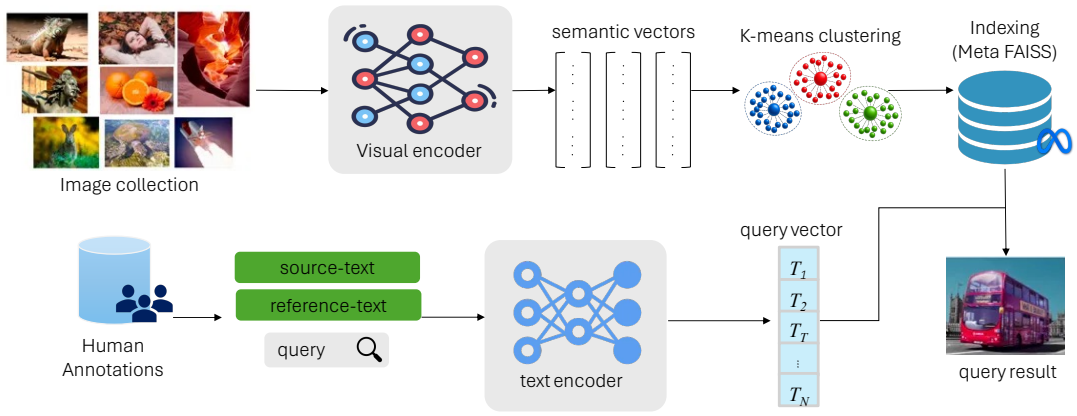

Figure 1: The MuTE dataset generation pipeline. Multimodal-multilingual embedding models extract image features for indexing via the FAISS API; these are then aligned with multilingual queries initiated from human-annotated MT evaluation sets.

tasks. Due to the high cost of data collection, similar evaluation benchmarks are currently unavailable for MMT. We make use of WMT evaluation resources to align human assessments with image representations via VLMs to obtain Source-Reference-Image triplets. Figure 1 illustrates the MuTE dataset generation process. The dataset is constructed using the following steps:

Visual Feature Extraction → Clustering and Indexing → Multilingual Alignment → Automatic & Human Evaluation/ Filtering.

2.1.1 Visual Feature Extraction

We compile 1.3 million real-world images from public datasets (LAION, Flickr30K, MSCOCO, and Visual Genome). To enable text-visual alignment, we extract visual embeddings using a VLM and store them alongside their relative file paths for subsequent processing.

2.1.2 Clustering and Indexing

We use the FAISS[4] `IndexIVFPQ` structure to efficiently manage and query the image embeddings. To accelerate retrieval, K-means clustering partitions the high-dimensional vector space into 4,096 Voronoi cells. This partitioning allows the system to restrict its search to the most relevant clusters, avoiding an exhaustive scan of the entire database.

2.1.3 Multilingual Alignment

In this step, we use multilingual representations (where source and target embeddings are averaged) obtained by encoding text with a shared encoder to retrieve relevant images. Specifically, source texts and reference translations from the WMT dataset are encoded as queries. Since the encoder is inherently multilingual and multimodal, it projects text from any language alongside visual data into a shared semantic vector space. During retrieval, the FAISS index performs nearest-neighbour search using cosine similarity between the text query vector and the stored image vectors, returning the most semantically similar images (See Appendix A.1).

2.1.4 Automatic & Human Evaluation/Filtering

To test and validate the text-image alignment and retrieval, we benchmarked our approach against MSCOCO test set. The automatic evaluation measures the similarity between retrieved and ground-truth reference images, while human annotators rated the relevance of text-image pairs on an SQM scale, across three dimensions: source-only, target-only, and joint retrieval. As shown in the results in Table A.2, joint queries consistently perform better, thereby mitigating potential ambiguity in either the source or target text.

3 Visually Grounded Evaluation

Building on the MuTE dataset, the evaluation of MMT systems can now incorporate both textual and visual signals. We propose to compute cross-modal semantic relevance by comparing embeddings of the source sentence, candidate translation, and associated image, as formulated in equation below:

$$\text{VGE}_{\text{score}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{S_{\text{text}}^{(i)} + \lambda^{(i)} \cdot S_{\text{vis}}^{(i)}}{1+\lambda^{(i)}}\right) \quad (1)$$

[4] https://github.com/facebookresearch/faiss

where N is the total number of samples, S_{text} and S_{vis} represent text-text and text-image similarity respectively, and $\lambda^{(i)}$ is the relevance gate that controls the contribution of visual information relative to linguistic similarity.

References

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. *arXiv preprint arXiv:1702.01287*.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Matthieu Futeral, Cordelia Schmid, Ivan Laptev, Benoît Sagot, and Rachel Bawden. 2022. Tackling ambiguity with images: Improved multimodal machine translation and contrastive evaluation. *arXiv preprint arXiv:2212.10140*.

Jindřich Helcl, Jindřich Libovickỳ, and Dušan Variš. 2018. Cuni system for the wmt18 multimodal translation task. *arXiv preprint arXiv:1811.04697*.

Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Distilling translations with visual awareness. *arXiv preprint arXiv:1906.07701*.

Chiraag Lala, Pranava Swaroop Madhyastha, Carolina Scarton, and Lucia Specia. 2018. Sheffield submissions for WMT18 multimodal translation shared task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 624–631, Belgium, Brussels. Association for Computational Linguistics.

Jiaoda Li, Duygu Ataman, and Rico Sennrich. 2021. Vision Matters When It Should: Sanity Checking Multimodal Machine Translation Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8556–8562, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. 2020. Dynamic context-guided capsule network for multimodal machine translation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1320–1329.

Vikas Raunak, Sang Keun Choe, Quanyang Lu, Yi Xu, and Florian Metze. 2019. On leveraging the visual modality for neural machine translation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 147–151, Tokyo, Japan. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 6153–6166, Online. Association for Computational Linguistics.

A Appendix

A.1 Example

Table 1 presents examples from the MuTE dataset. The first instance demonstrates highly relevant image retrieval for a joint English–French query. The second image, while a close semantic match, diverges slightly from the source text (two hot-dogs not represented by image).

Table 1: Examples from MuTE dataset.

	**En:** two surfers looking at the dark stormy looking sky. **Fr:** Deux surfeurs regardent le ciel qui paraît sombre et orageux .
	**En:** Several people standing around and a woman being handed two hot dogs. **Fr:** Plusieurs personnes debout autour et une femme recevait deux hot-dogs.

A.2 Automatic Evaluation Results

Table 2: Human and automatic evaluation (cosine similarity) for reference and candidate images retrieved using different query languages. Spearman's correlation indicates moderate but statistically significant agreement between human and automatic judgments.

Query	**Avg. Score**	**Corr.**	p**-value**
English	0.842	0.45	10^{-23}
German	0.830		
French	0.830	-	
English-German	0.840	-	
English-French	0.840	0.46	10^{-11}

Avg. Score represents the average cosine similarity, and Corr. denotes Spearman's correlation coefficient between human and automatic ratings.