Context-aware Evaluation of Machine Translation (CAEMT) Systems

Sami Ul Haq
Dublin City University

Sheila Castilho
Dublin City University

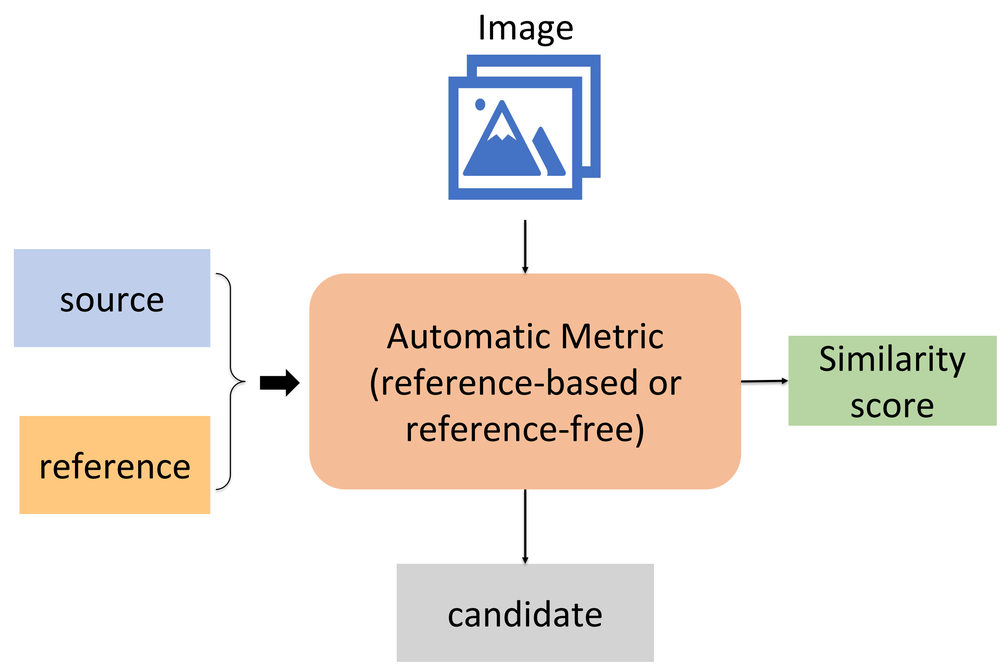
Yvette Graham
Trinity College Dublin

Current Evaluation Metrics

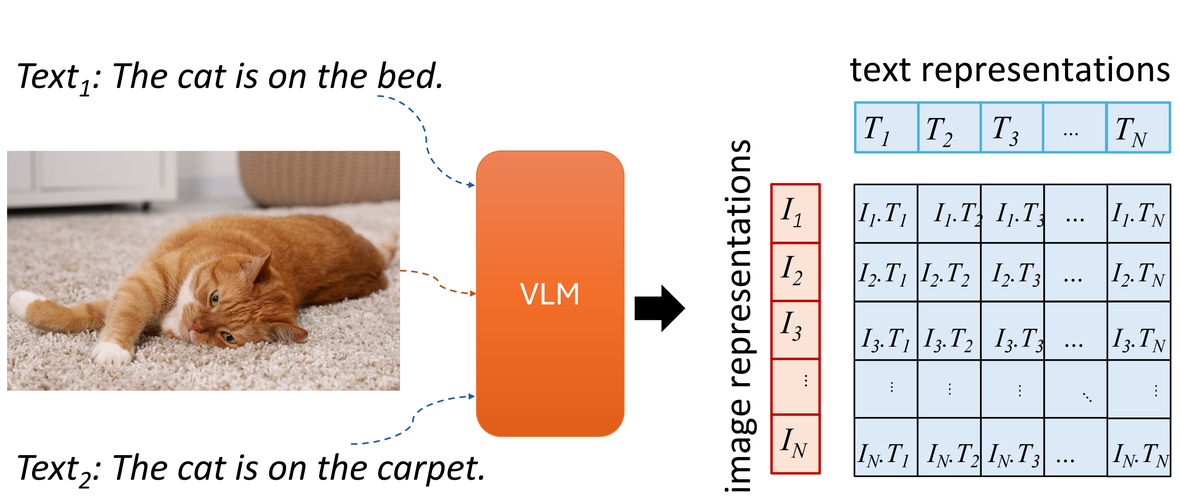
- Traditional evaluation metrics (e.g., BLEU[1]) rely on **surface-level** n-gram overlap.
- Recent metrics (e.g., BERTScore[2]) address this limitation by leveraging continuous word embeddings to compute similarity in **semantic space**.
- However, token level embedding distance may **overestimate similarity** (e.g., between *cat* and *dog*).
- We propose **CAEMT**—which incorporates **cross-modal semantic similarity** from both textual and visual (image) modalities to enhance the reliability of MT evaluation.

[1]Papineni, Kishore, et al. "Bleu: a Method for Automatic Evaluation of Machine Translation."
[2]Zhang, Tianyi, et al. "BERTScore: Evaluating Text Generation with BERT."

Our proposed approach (CAEMT)



Proposed workflow of CAEMT



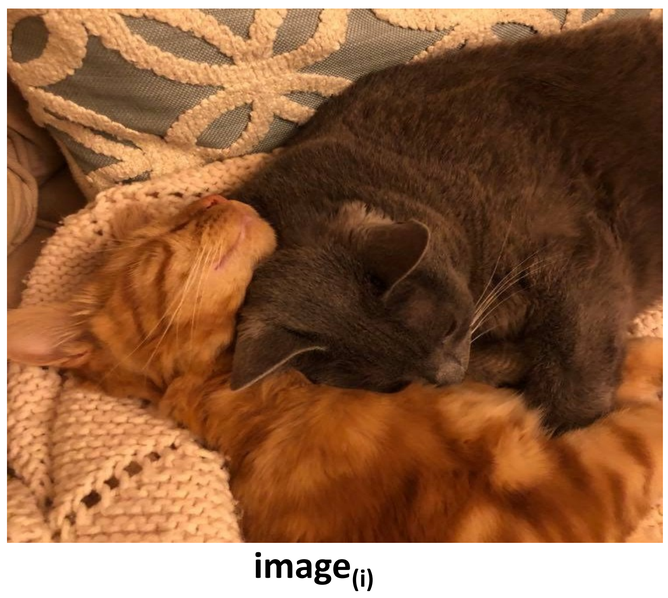
- Use of **Visual-Language Models** (e.g., Jina[3], CLIP[4]) to generate text and image representations.
- Cross-lingual cross-modal weighted **cosine similarity** between image and text for **reference-based** evaluation:

$$\omega * cos(\boldsymbol{v}, \boldsymbol{t_i}, \boldsymbol{t_j})$$

- Word matching in a **semantic space** (such as BERTScore) using word embeddings.
- Also**, reference-free** evaluation, directly comparing source with candidate text, using image as ground truth.

[3]Koukounas, Andreas, et al. "jina-clip-v2: Multilingual multimodal embeddings for text and images." *arXiv preprint arXiv:2412.08802* (2024).
[4]Hessel, Jack, et al. "CLIPScore: A Reference-free Evaluation Metric for Image Captioning." *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021.

CAEMT vs traditional metrics: initial results



image$_{(i)}$

Source$_{(s)}$:
Eine graue Katze kuschelt mit einer orangefarbenen Katze auf einer Decke.

Candidate$_{(c)}$:
a calico cat is cuddling with an orange dog on a blanket.

Reference$_{(r)}$:
a grey cat is cuddling with an orange cat on a blanket.

Metric*	Score	
BLEU$_{c, r}$	↑ 59.20	✗
TER$_{c, r}$	↓ 16.70	✗
BERTSCORE$_{c, r}$	↑ 94.70	✗
COMET-22$_{c, r}$	↑ 84.10	✗
CAEMT$_{r, i}$	↑ **0.43**	✓
CAEMT$_{c,i}$	↓ **0.37**	✓
CAEMT$_{c, r, i}$	↑**0.49**	✓
CAEMT$_{c, s}$	↓0.77	✗

*For CAEMT we have used multilingual multimodal jina-clip-v2 model with transformer API to calculate the text-text and text-image cosine similarity . For other metrics we used online MATEO tool (https://mateo.ivdnt.org).

This research was conducted with the financial support of Science Foundation Ireland at ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology at DCU [13/RC/2106_P2].

Ollscoil Chathair Bhaile Átha Cliath
Dublin City University

Research Ireland Centre for Research Training in Digitally-Enhanced Reality

Scan for examples, code etc.