
Detecting concrete visual tokens for multimodal machine translation

Braeden Bowen

Vipin Vijayan

Scott Grigsby

PAR Government Systems Corporation, Dayton, OH*

bowen_braeden@bah.com

vipin255@gmail.com

grigsby_scott@bah.com

Timothy Anderson

Jeremy Gwinnup

Air Force Research Laboratory 711HPW/RHWTE, Dayton, OH

timothy.anderson.20@us.af.mil

jeremy.gwinnup.1@us.af.mil

Abstract

The challenge of visual grounding and masking in multimodal machine translation (MMT) systems has encouraged varying approaches to the detection and selection of visually-grounded text tokens for masking. We introduce new methods for the detection of visually and contextually relevant (concrete) tokens from source sentences, including detection with natural language processing (NLP), detection with object detection, and a joint detection-verification technique. We also introduce new methods for selection of detected tokens, including shortest n tokens, longest n tokens, and *all* detected concrete tokens. We utilize the GRAM MMT architecture to train models against synthetically collated multimodal datasets of source images with masked sentences, showing performance improvements and improved usage of visual context during translation tasks over the baseline model.

1 Introduction

The challenge of multimodal machine translation (MMT) is to design a system that automatically translates text from one language to another while utilizing other modalities (e.g., image, video, audio) as inputs to assist in translation (Caglayan et al., 2016).

Prior work has shown that translation ambiguities and missing textual information can be supplied by contextually-relevant images, aiding in multilingual translation (Lala and Specia, 2018; Caglayan et al., 2019; Wu et al., 2021). For example, the noun “bank” is ambiguous and contextually dependent in English (“financial institution” or “river edge”) but unambiguous in French (“*banque*” or “*rive*”) (Futeral et al., 2023). The hypothesis for MMT research is that these translation ambiguities can be resolved with the inclusion of image context.

In practice, not every sentence has semantic ambiguities, missing information, or relevant visual context; it is therefore beneficial to ensure that ambiguous text is visually and contextually relevant to an associated image (Zhou et al., 2018).

To enforce reliance on image context for translation tasks, some MMT models mask tokens from text inputs (Caglayan et al., 2019; Sato et al., 2023). While most early masking iterations randomly selected tokens for masking, more recent efforts have sought to mask tokens based on contextual relevance to a given image (Tan and Bansal, 2020), increasing the usefulness of the image in resolving ambiguity. Still, those methods tend to ignore deterministic selection of relevant tokens, opting to randomly select from a pool of viable tokens.

While these approaches have displayed performance improvements over text-only and random

* Now doing business as Booz Allen Hamilton Corporation.

masking models, these methods generally do not take into account the relevance of a masked token. Therefore, we hypothesize that more intentional selection and masking of **concrete** (i.e., visually and contextually relevant) text tokens will improve visual grounding and increase model usage of multimodal context.

In order to select visually and contextually relevant tokens, we explore a combination of natural language processing (NLP) techniques and object detection models and examine deterministic methods for token selection from the set of available detections.

Using these techniques, we collate multimodal datasets based on the Multi30k dataset (Elliott et al., 2016); the resulting datasets are triplets of source sentences with masked concrete tokens, unmasked target sentences, and associated images.

When masking concrete text tokens from source sentences, we find improvements in both usage of visual information in translation and in performance on evaluation challenges, including CoM-MuTE scores of up to 0.67 and BLEU scores of up to 46.2.

2 Related Works

2.1 Masking for Visual Grounding

In a text-only modality, Devlin et al. (2019) randomly masked text tokens during pre-training of a bidirectional transformer encoder-decoder and found performance improvements against other text-only models.

Zhou et al. (2018) utilized jointly-encoded unmasked text and image embeddings to visually ground entire source sentences to images. Using a visual-text attention mechanism on the embeddings, they extracted words that shared semantic context with the images.

Ive et al. (2019) combined these approaches, randomly *and* manually masking ambiguous and gender-neutral words from source texts to force their MMT model to utilize visual information on evaluation tasks. This work showed that the model was able to use image context to recover from missing, inaccurate, or ambiguous textual context.

Caglayan et al. (2019) used image descriptions from the Flickr30k-Entities dataset (Plummer et al., 2015) to dynamically mask “visually depictable entities” and color descriptors from source sentences,

but noted a degradation in performance on the Multi30k test sets (Elliott et al., 2016). In contrast, Wang and Xiong (2021) found that masking *irrelevant* objects improved performance on MMT evaluation tasks, suggesting that state-of-the-art MMT models are ineffectively utilizing visual information.

A meta-analysis by Wu et al. (2021) found that many reported improvements in MMT performance are the result of regularization effects, not model interpolation of multimodal features; similarly, Zhuang et al. (2023) found that while visual grounding can improve performance in word learning, these improvements are only marginal. However, they also found that training sets with less textual information and fewer direct co-occurrences of visual words more effectively utilize visual information, suggesting that the relationship between text and image context is still viable.

2.2 Token Selection for Visual Grounding

In practice, many sentences have more than one visually grounded token; in these cases, available tokens must be dynamically selected for masking. The standard method is to randomly select viable tokens (Devlin et al., 2019); however, recent work in masked language modeling (MLM) has shown that informed selection of masked tokens may improve performance (Sato et al., 2023).

Other work has given consideration to the length of source segments in text masking (Xiao et al., 2023) and to the number of tokens selected (Joshi et al., 2020), but little work has been done to select tokens deterministically (e.g., by token length).

3 Approach

We perform improved visual grounding by detecting concrete tokens in source sentences. We explore three detection techniques to identify concrete text tokens (Section 3.1) and four selection techniques to appropriately select the identified concrete text tokens (Section 3.2.1). We then collate permutations of synthetic MMT datasets by masking the selected concrete tokens from source sentences and aligning each sentence with its original dataset image pair. We then train an MMT model (Section 3.3) on these datasets, expanding on work by Vijayan et al. (2024) and Caglayan et al. (2019).



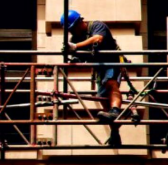
	<p>SRC: A girl in sunglasses walks by a red car.</p> <p>DT1: <i>girl, red car, sunglasses, girl</i></p> <p>DT2: <i>sunglasses, A girl</i></p> <p>DT3: <i>girl, red car, sunglasses, girl</i></p> <p>MSK: A girl in <unk> walks by a red car.</p>
	<p>SRC: Young boy kicks a red and white soccer ball on a grassy field.</p> <p>DT1: <i>field, young boy, ball, young, white soccer ball, boy, grassy field</i></p> <p>DT2: <i>soccer, field, young boy, grassy field, young, boy, white soccer ball, ball</i></p> <p>DT3: <i>field, young boy, ball, young, white soccer ball, boy, grassy field</i></p> <p>MSK: Young boy kicks a red and <unk> <unk> <unk> on a grassy field.</p>
	<p>SRC: A construction worker fits metal pipes together.</p> <p>DT1: <i>construction worker, worker, pipes</i></p> <p>DT2: <i>construction worker, worker, construction, pipes</i></p> <p>DT3: <i>construction worker, worker, pipes</i></p> <p>MSK: A <unk> <unk> fits metal pipes together.</p>

Figure 1: Multi30k source pairs (image, **SRC**) with results from each detection technique (**DT**) and an example masked source text (**MSK**). **DT1** represents the *NLTK* technique; **DT2** represents the *MDETR Detection* technique; **DT3** represents the *Joint Detection* technique. The masked sentence **MSK** represents a possible masked sentence based on the bold tokens in the **DT3** detections.

3.1 Detection of Concrete Tokens

As Caglayan et al. (2019) found, masking visually relevant objects from a source text can force the model to utilize image context to fill in the artificially-created gap in lexical/semantic understanding. We hypothesize that for a given text-image pair, the masking of text tokens that are directly relevant to the image (i.e., “concrete” tokens), will improve visual grounding, increasing model correlation of image inputs during downstream translation tasks.

We present three techniques for detection of concrete tokens: NLP with NLTK (Section 3.1.1), object detection with MDETR (Section 3.1.2), and joint NLTK/MDETR detection and grounding (Section 3.1.3). While techniques one and two respectively use text and image context, method three uses contextual information from both modalities to make decisions about which text tokens are concrete.

3.1.1 Detection with NLTK

The first concrete token detection approach is to parse sentences for nouns and noun phrases that are likely to represent visual context. By masking to-

kens that are critical to comprehension and translation of the text, we can encourage the model to learn with visual context.

The Natural Language Toolkit (NLTK) (Loper and Bird, 2002) includes the WordNet corpus (Fellbaum and Miller, 1998), an English-language lexical database that provides structured relationships between cognitive synonyms (“synsets”) for nouns, verbs, adjectives, and adverbs. Specifically, WordNet defines a directed acyclic graph (DAG) for each of these parts of speech (POS), containing synonyms, troponyms, antonyms, and meronyms (Figure 2). Critically, these relational graphs establish affiliations between English words, their definitions, and their related parent categories (i.e., “hypernyms”).

Starting with specific synonyms and troponyms (e.g., “sedan”, “hatchback”, “SUV”) and traversing the DAG upwards, WordNet collapses definitions and synsets into their associated hypernym classes (e.g., “car”, “vehicle”) until it reaches a root hypernym (e.g., “physical_entity”, “entity”). Using recursive graph traversal, we can select any node in the DAG and parse its hypernyms upward until we reach either a root hypernym or a parent hy-

pernym on which we can base an estimate of the root hypernym (e.g., “object” generally maps to “physical_entity”).

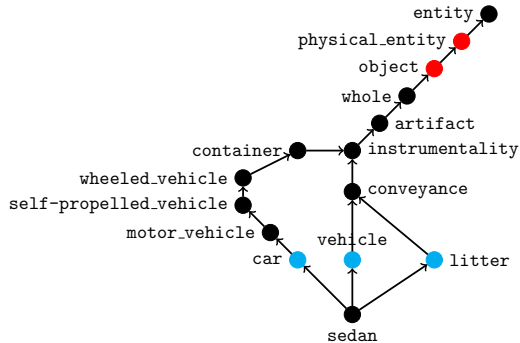


Figure 2: An example hypernym graph. The original token, sedan, its three synset entries (labeled in blue), and its associated concrete hypernyms (labeled in red).

Concrete Hypernyms	Abstract Hypernyms
physical_entity	abstract_entity
physical_object	abstraction
stuff	
object	
person	
unit	
whole	

Table 1: Labeled WordNet (Fellbaum and Miller, 1998) hypernyms. A token is classified as concrete or abstract if any of the above hypernyms are in its DAG.

Given that there exists only a small cluster of root and high-level parent hypernyms for nouns in WordNet, we can classify the hypernym DAG of any noun or noun phrase as “concrete” or “abstract” based on these high-level hypernyms (Table 1).

While this method provides a simple concrete/abstract classifier for text tokens, it introduces additional complications. Although most DAG nodes have multiple child hyponyms (e.g., “car” may have “sedan” and “hatchback”), some have multiple cognitive synonyms, as English words often have multiple equally likely definitions. For a

given node, each of its “definitions” will appear as an entry into its synset; for example, the English noun “link” has nine values in its WordNet synset, ranging from “URL” to “channel for communication” to “element of a chain.” These varied definitions may branch to different root hypernyms, impacting the classification based on which definition is chosen (Table 1).

To compensate, we consider each entry in a word’s synset and extract a ratio of concrete/abstract definitions, which more comprehensively projects a token’s likelihood of being concrete. We perform recursive graph traversal for each entry and retain the percent of concrete entries as a “concreteness score.” To then classify the original word as abstract or concrete, we establish a threshold of 33% likelihood and only accept words above that concreteness score.

3.1.2 Detection with MDETR

While the NLTK approach can quickly and efficiently select concrete tokens from a sentence, it incorrectly assumes that every concrete token in the sentence is relevant to its associated image. Contextually linking an irrelevant concrete token to a given image could negatively impact model performance, especially if the token has high commonality in a dataset. As a second approach to concrete token detection, we utilize an object detection model to select concrete tokens. Rather than relying solely on the text processing for detection, we inspect the image itself for object classes relevant to the text.

For this approach, we use MDETR (Kamath et al., 2021), an end-to-end object detection model. Rather than relying exclusively on pre-defined object classes, MDETR uses NLP techniques alongside a pre-trained detection model (Carion et al., 2020) to perform object detection and image classification based on the input tokens. Given a text-image pair (Figure 3), the model assigns each text token an object classification, confidence score, and bounding box. To maximize the number of detectable tokens, we pass an entire Multi30k sentence into the MDETR model and filter out detections with low confidence scores, retaining only the tokens with a high confidence of correlation to the image. While Kamath et al. (2021) filter all outputs with confidence less than 0.7, we filter at 0.85; after analyzing performance at threshold increments between 0.5 and 0.95, we found that this threshold ensured the most balanced object confidence.



SRC: Cooking hot peppers in the cold winter!
DT2: “cooking hot peppers in the cold winter”
DT3: “pepper”

Figure 3: Multi30k source pair (image, **SRC**) with results from the MDETR (**DT2**, top image) and Joint (**DT3**, bottom image) detection techniques. MDETR query strings, bounding boxes, and confidence scores are shown. In this example, supplying the entire source sentence as text input to the MDETR object detection model incorrectly identifies the peppers being cooked, while querying only the word “pepper” increases the model’s confidence and more closely identifies the region containing the query.

3.1.3 Detection with Joint Visual Grounding

While the MDETR technique is less likely than the NLTK technique to improperly select text tokens as visually-grounded, the pre-trained MDETR model will always attempt to assign a bounding box to some text token, often resulting in outputs with high confidence but incorrect alignment. In practice, providing extended textual context (i.e., entire captions

or sentences) further exacerbates this problem (Figure 3).

Therefore, we are left with two techniques with contrasting weaknesses: NLTK ignores image context, and MDETR misinterprets textual context. To mitigate these issues, we present a conjoined detection technique that “verifies” the presence of NLTK-detected concrete tokens within an image using MDETR, ensuring that concrete tokens are visually grounded in the image.

Like the MDETR technique, the joint technique parses text-image pairs (unlike the NLTK technique, which is image-agnostic). The source sentence is first processed by the NLTK technique, which returns the noun and noun phrase tokens that met or surpassed the concrete threshold. Each of those tokens is paired with a copy of the source image and passed into the MDETR technique, which performs object detection and filters out all tokens whose resulting confidence is below the confidence threshold. This simultaneously reduces the probability of incorrect alignment by the object detection model and ensures that text tokens are visually grounded, resulting in a set of linguistically concrete and visually-grounded text tokens with high probability of relevance to the source image. Masking these explicitly-relevant tokens will force model reliance on image context.

3.2 Synthetic Dataset Collation

Because most current work in MMT focuses on the Multi30k dataset (Elliott et al., 2016), an image-caption dataset consisting of 30,014 images with English sentences and corresponding multilingual translations, we collate synthetic datasets of masked sentence-image pairs from Multi30k.

We use each detection technique (Section 3.1) to detect concrete tokens and align them to their original dataset image. From these masked sentence-image pairs, we collate a series of MMT datasets in which a maximum of two concrete tokens are masked from each sentence and associated with the relevant image from the original dataset, resulting in training and validation sets that are at most twice as large as the original Multi30k sets.

3.2.1 Token Selection Techniques

During the dataset collation process, a single sentence may have $n > 2$ available concrete tokens; in this case, additional consideration must be given

to which tokens are selected for inclusion in the dataset. The standard method has generally been to randomly select from the available tokens (Devlin et al., 2019), but recent work in masked language modeling (MLM) has shown that more informed selection of masked tokens may actually improve performance (Sato et al., 2023).

To examine this, we implement two deterministic token selection techniques, selecting the *n* **longest** and **shortest** tokens (by number of characters) respectively for each sentence. We compare these techniques to a **random** selection of *n* tokens and an **unrestricted** selection which ignores the $n=2$ normalization and accepts all available concrete tokens.

3.3 GRAM Model

As the basis for our multimodal translation architecture, we utilize the GRAM architecture (Vijayan et al., 2024). GRAM modifies the FAIR WMT19 (Ng et al., 2019) text-only model, an encoder/decoder-based transformer architecture (Vaswani et al., 2017), by adding additional multimodal components to create an MMT model.

To process text input, GRAM uses the same byte-pair encoding (BPE) and vocabulary dictionary as the FAIR WMT19 model (Ng et al., 2019). Masked sentences are BPE-encoded and fed as standard text inputs to the MMT model. We mask by replacing each token with an <unk> token, as that token is the closest to a mask token available in the FAIR WMT19 model (Ng et al., 2019). Our method expands on prior work by Tang et al. (2022) and Wu et al. (2021) while increasing the requirements for a token to be visually grounded to an image.

To process image input, the GRAM model uses CLIP, a pre-trained text-only translation model alongside a pre-trained vision encoder, a perceiver resampler, and vision-text cross-attention layers (Radford et al., 2021). While the original GRAM paper utilizes the ViT-L/14@336px CLIP model, we noted better results within our evaluation framework when using the RN50x4 CLIP model; we present those results below (Section 4.2). This vision encoder converts input images into image embeddings, enabling the perceiver resampler to convert those embeddings into a fixed number of vision tokens. Vision tokens and corresponding text embeddings are interleaved into vision-text cross-attention lay-

ers within the transformer encoder, creating mappings from both the text and the image embeddings onto a sequence of joint representations. Finally, the transformer decoder ingests this sequence and outputs probabilities for the next output text token in the target sequence.

The number of parameters in the original text-only Transformer is 269,746,176; the number of parameters in the RN50x4 CLIP vision encoder is 101,520,396, for a total of 371,266,572 parameters in our GRAM model. Additionally, our GRAM perceiver resampler contains 87,137,080 parameters.

4 Results and Discussion

4.1 Experimental Framework

We train the GRAM models on unique permutations of synthetically collated datasets representing each combination of detection (**NLTK**, **MDETR**, **Joint**) (Section 3.1) and selection (**unrestricted**, **restricted-long**, **restricted-short**, **restricted-random**) (Section 3.2.1) techniques. We compare the resulting trained versions to the GRAM model trained on a unmasked dataset of original sentences.

Most current work in MMT focuses on the Multi30k dataset; because of its prevalence in other MMT works, we utilize the Multi30k dataset for collation of our training datasets. We then evaluate the GRAM models on the Multi30k 2016, 2017, and COCO test sets using BLEU4 scores.

We also evaluate the GRAM model with an additional metric, Contrastive Multilingual Multimodal Translation Evaluation (CoMMuTE). Futeral et al. (2023) proposed the CoMMuTE dataset to evaluate both performance on translation tasks and usage of visual information by MMT models. In the ensemble CoMMuTE evaluation, the model is given two images, a lexically or semantically ambiguous English sentence, and a target language translation that resolves the ambiguity according to one of the two images. The task involves determining which of the two images the sentence pairs best match. The evaluation is made using the perplexity of the model output, and the resulting CoMMuTE score is calculated using the model’s determination of accuracy across 100 text-image pairs.

Detection	Selection	Score			
		CoMMuTE	Multi30k BLEU4 (en-de)		
			2016	2017	COCO
	Futeral et al. (2023)	0.59	43.3	38.3	35.7
	Vijayan et al. (2024)	0.61	46.5	43.6	39.1
Unmasked		0.5	45.0	42.0	38.2
NLTK	Unrestricted	0.55	45.7	41.9	39.2
NLTK	Restricted-Longest	0.62	46.0	<u>42.5</u>	37.8
NLTK	Restricted-Shortest	0.63	46.0	42.0	37.9
NLTK	Restricted-Random	0.67	<u>46.2</u>	41.4	37.8
MDETR	Unrestricted	0.56	<u>46.0</u>	<u>42.4</u>	<u>38.4</u>
MDETR	Restricted-Longest	0.63	45.7	41.7	38.0
MDETR	Restricted-Shortest	0.63	45.0	41.2	36.9
MDETR	Restricted-Random	0.63	45.6	42.2	37.6
Joint	Unrestricted	0.52	45.5	42.4	<u>38.9</u>
Joint	Restricted-Longest	<u>0.63</u>	<u>45.8</u>	<u>42.6</u>	38.8
Joint	Restricted-Shortest	0.61	45.4	42.0	37.9
Joint	Restricted-Random	0.61	45.5	42.4	37.5

Table 2: Selected performance results of our model against the CoMMuTE and Multi30k test sets. The best result by column is indicated in **bold**; the best result for each detection technique is underlined. Results as reported by GRAM (Vijayan et al., 2024) and VGAMT (Futeral et al., 2023) are included for reference.

Detection	Concrete %	Unique Detections
NLTK	99.51	5,393
MDETR	99.92	6,674
Joint	99.49	4,761

Table 3: Unique concrete token detections and percent of Multi30k sentences with detected tokens by detection technique.

4.2 Results

We review the performance of the model variants trained using the synthetic Multi30k datasets (Section 3.2) on the above evaluation metrics. We train 13 variants, consisting of one unmasked baseline and 12 models representing each combination of detection (Section 3.1) and selection (Section 3.2.1) techniques.

4.3 Detection Results

We introduced three distinct methods for detection of concrete text tokens: the NLTK technique (Section 3.1.1), which parses nouns and noun phrases from sentences, the MDETR technique (Section

3.1.2), which inputs sentences as queries to an object detection model, and the Joint technique (Section 3.1.3). Each technique generates the same output structure: multimodal datasets of sentences masked concrete tokens and matching images. We hypothesize that masking concrete tokens with these techniques will improve performance on evaluation metrics. We further hypothesize that the Joint technique will be more selective with its detections than its component NLTK and MDETR techniques, and will thus utilize image context more efficiently and critically.

We found that all three techniques consistently extracted relevant tokens from the text: each technique extracted concrete tokens from over 99% of Multi30k sentences (Table 3). The MDETR detection technique was the most successful, extracting 23.8% and 40.2% more unique concrete tokens than the NLTK and Joint techniques, respectively. This resulted in the MDETR technique masking the highest concentration of original Multi30k sentences (114 and 120 sentences more than NLTK and Joint, respectively).

Increased rates of detection did not correlate with better performance, though. All tested models outperformed the unmasked (baseline) dataset in CoMMuTE and BLEU scores, but in contrast to our hypothesis the NLTK technique outperformed both the MDETR and Joint techniques both in CoMMuTE and BLEU score (Table 2). The Joint technique, which we hypothesized would improve on its component techniques, consistently underperformed against the others. This is especially true in the Joint Unrestricted model, which only improved its CoMMuTE score by 0.02 and its BLEU score 0.5 over the baseline. We suggest that the Joint technique was actually hindered by its strict selection process, leading to a much smaller pool of objects to mask from. Conversely, the MDETR technique tended to over-select longer, rarely-used, or irrelevant tokens (Figure 3), contributing to the larger masking percentages but the lower overall performance. The success of the NLTK technique over the others was its “middle ground” approach, classifying concrete tokens more liberally than the Joint technique but more consistently than the MDETR technique.

23% of tested models underperformed the original GRAM model (Vijayan et al., 2024) on CoMMuTE metrics, 15.4% performed identically, and the remaining 53.8% outperformed. All tested models underperformed the original GRAM model in Multi30k 2016/2017 BLEU metrics. One model (NLTK Unrestricted) outperformed the original GRAM model in the Multi30k COCO metric, but the improvement is well within a margin for normalization effects. We suggest that the performance disparity between models in these Multi30k BLEU metrics is due to dataset size: the original GRAM model was pre-trained trained on the Conceptual Captions dataset (Sharma et al., 2018) of 2,878,999 text-image pairs, resulting in synthetic datasets nearly 100 times larger than those used here. Despite this, the majority of models outperformed GRAM in CoMMuTE metrics, achieving scores of up to 0.67.

In general, we also note an inverse relationship between CoMMuTE and BLEU performance: that is, when CoMMuTE scores increase, BLEU scores tend to decrease. For example, the MDETR Unrestricted model notched the highest average BLEU score across all three Multi30k metrics, but

had the second-lowest CoMMuTE score.

4.4 Selection Results

Critical to the synthetic dataset collation system is the process of selecting concrete tokens for masking. Prior efforts have generally selected tokens at random (Ive et al., 2019); we introduced three additional techniques (Section 3.2.1), longest-token selection, shortest-token selection, and unrestricted selection, and test each against a baseline of randomly-selected concrete tokens. We hypothesize that the presented token selection techniques will outperform the baseline of random selection; specifically, we hypothesize that longest-token and unrestricted selection will encourage additional usage of visual context and thus improve CoMMuTE score, and that shortest-token selection will minimize the number of token predictions required by the model (Section 3.3) and thus improve BLEU score.

We found that while all tested selection techniques (Section 3.2.1) outperformed the unmasked baseline, comparative performance between techniques are less conclusive. When paired with the NLTK detection technique, the random selection technique outperformed the others in CoMMuTE and Multi30k 2016 BLEU scores. When paired with the MDETR metric, none of the restricted selection techniques had any impact on CoMMuTE score. When paired with the Joint detection technique, the longest-token selection technique improved CoMMuTE and Multi30k 2016/2017 BLEU scores.

Contrary to our hypothesis, the deterministic token selection techniques did not consistently outperform the random selection technique. The most consistent results were with the unrestricted selection technique, which significantly degraded CoMMuTE performance but tended to improve BLEU performance (especially in the COCO BLEU metric, where it outperformed all other tested models). Shortest-token selection also tended to follow these patterns of performance degradation, but not as substantially: its NLTK and Joint detection variants performed identically on the Multi30k 2017 and COCO BLEU metrics and performed near the bottom of results for the CoMMuTE and 2016 BLEU metrics across all three detection techniques.

Each of these findings runs counter to our hypotheses in this area, suggesting that token selection at this scale has less impact on model perfor-

mance than we expected; in fact, random or pseudo-random token selection of the identified concrete tokens may actually improve performance over deterministic methods.

4.5 Future Work

Given the high percentage of visually-grounded tokens in the Multi30k training set, future work should consider the techniques against both larger MMT datasets and MMT datasets with lower concentrations of visually-grounded tokens (e.g., Conceptual Captions). Similarly, future work should consider synthetically collated datasets that combine elements of multiple multimodal datasets (e.g., images from Conceptual Captions, sentences from Multi30k), including synthetic datasets created from text-only datasets.

Additionally, future work should compare baseline scores for tokens selected completely at random to more accurately gauge the efficacy of object token selection.

Finally, future work should consider a more deterministic way to classify the concreteness of a token with NLP, including selection of definitions based on contextual awareness.

5 Conclusion

The continued challenge of visual grounding and masking in MMT systems has encouraged varying approaches to the detection and selection of visually-grounded text tokens for masking (Caglayan et al., 2019; Wu et al., 2021).

We introduced three new techniques for detection of concrete tokens from source sentences: detection with natural language processing (NLP), detection with object detection, and joint NLP/object detection. We also introduced deterministic methods for the selection of detected tokens, including longest and shortest n tokens.

Finally, we utilized the GRAM MMT architecture (Vijayan et al., 2024) to train models against synthetically collated datasets of masked sentences and associated images. We showed performance improvement over the baseline models and elevated use of visual context during translation tasks.

Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government. Cleared for public release on 12 Feb 2024. Originator reference number RH-24-125351. Case number AFRL-2024-0803.

Acknowledgements

Thanks to AFRL SCREAM Lab and AFRL 711HPW/RHWTE for their help in this project.

Funding: This work was supported by the AIMMIER project via Infoscitex Corporation (IST) and Air Force Research Laboratory (AFRL) under Air Force contract FA8650-20-D-6207.

References

- Caglayan, O., Aransa, W., Wang, Y., Masana, M., García-Martínez, M., Bougares, F., Barrault, L., and van de Weijer, J. (2016). Does multimodality help human and machine for translation and image captioning? In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. Association for Computational Linguistics.
- Caglayan, O., Madhyastha, P., Specia, L., and Barrault, L. (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.
- Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Elliott, D., Frank, S., Sima'an, K., and Specia, L. (2016). Multi30k: Multilingual english-german image descriptions.
- Fellbaum, C. and Miller, G. A. (1998). *WordNet: An Electronic Lexical Database*. The MIT Press.
- Futeral, M., Schmid, C., Laptev, I., Sagot, B., and Bawden, R. (2023). Tackling ambiguity with images:

- Improved multimodal machine translation and contrastive evaluation. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the ACL (Volume 1: Long Papers)*, pages 5394–5413, Toronto, Canada. Association for Computational Linguistics.
- Ive, J., Madhyastha, P., and Specia, L. (2019). Distilling Translations with Visual Awareness. *ArXiv*, abs/1906.07701.
- Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2020). SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.
- Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., and Carion, N. (2021). Mdetr – modulated detection for end-to-end multi-modal understanding.
- Lala, C. and Specia, L. (2018). Multimodal lexical translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Loper, E. and Bird, S. (2002). NLTK: The natural language toolkit. *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*.
- Ng, N., Yee, K., Baeovski, A., Ott, M., Auli, M., and Edunov, S. (2019). Facebook FAIR’s WMT19 News Translation Task Submission.
- Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Sato, J., Caseli, H., and Specia, L. (2023). Choosing what to mask: More informed masking for multimodal machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 244–253, Toronto, Canada. Association for Computational Linguistics.
- Sharma, P., Ding, N., Goodman, S., and Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.
- Tan, H. and Bansal, M. (2020). Vokenization: Improving language understanding with contextualized, visual-grounded supervision.
- Tang, Z., Zhang, X., Long, Z., and Fu, X. (2022). Multimodal neural machine translation with search engine based image retrieval. In *Proceedings of the 9th Workshop on Asian Translation*, pages 89–98, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.
- Vijayan, V., Bowen, B., Grigsby, S., Anderson, T., and Gwinnup, J. (2024). Adding multimodal capabilities to a text-only translation model. *arXiv:2403.03045 [cs.CL]*.
- Wang, D. and Xiong, D. (2021). Efficient Object-Level Visual Context Modeling for Multimodal Machine Translation: Masking Irrelevant Objects Helps Grounding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):2720–2728.
- Wu, Z., Kong, L., Bi, W., Li, X., and Kao, B. (2021). Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation.
- Xiao, Y., Xu, R., Wu, L., Li, J., Qin, T., Liu, Y.-T., and Zhang, M. (2023). Amom: Adaptive masking over masking for conditional masked language model.
- Zhou, M., Cheng, R., Lee, Y. J., and Yu, Z. (2018). A visual attention grounding neural model for multimodal machine translation.
- Zhuang, C., Fedorenko, E., and Andreas, J. (2023). Visual grounding helps learn word meanings in low-data regimes.