

# Understanding the Effect of Textual Adversaries in Multimodal Machine Translation

**Koel Dutta Chowdhury**

Dept. of Language Science and Technology  
Saarland University, Germany  
koelddc@lst.uni-saarland.de

**Desmond Elliott**

Department of Computer Science  
University of Copenhagen  
de@di.ku.dk

## Abstract

It is assumed that multimodal machine translation systems are better than text-only systems at translating phrases that have a direct correspondence in the image. This assumption has been challenged in experiments demonstrating that state-of-the-art multimodal systems perform equally well in the presence of randomly selected images, but, more recently, it has been shown that masking entities from the source language sentence during training can help to overcome this problem. In this paper, we conduct experiments with both visual and textual adversaries in order to understand the role of incorrect textual inputs to such systems. Our results show that when the source language sentence contains mistakes, multimodal translation systems do not leverage the additional visual signal to produce the correct translation. We also find that the degradation of translation performance caused by textual adversaries is significantly higher than by visual adversaries.

## 1 Introduction

There has been a surge of interest in tackling machine translation problems using additional information, such as a image or video context. It has been claimed that systems trained on a combination of visual and textual inputs produce better translations than systems trained using only textual inputs (Specia et al., 2016; Elliott et al., 2017). However, these claims have been the subject of debate in the literature: Elliott (2018) argued that the additional visual input is not necessarily used by demonstrating that the performance of a system did not change when it was evaluated with randomly selected images, and Grönroos et al. (2018) observed that their models were insensitive to being evaluated with an “averaged” visual vector, as opposed to the expected visual vector. More recently, Caglayan et al. (2019) presented experiments in which the colour and entity tokens (e.g.

*blue* or *woman*) were masked during the training of a multimodal translation model. They found that training the model under these conditions resulted in the system relying on the visual modality to recover the masked words during evaluation. Although, their results show that the visual modality can be used to recover the masked tokens in the source sentences, it is not clear if these systems will perform similarly when there is a *mismatch* between the textual and visual concepts.

In this paper, we explore the effect of textual adversaries in multimodal machine translation. We construct hard negative textual adversaries, which contradict the original meaning, in order to explore the robustness of systems to textual adversaries. The textual adversaries are based on minimal manipulations to the sentences, for example:

- (1) a. Two people walking on the beach.  
b. \*Two people walking on the grass.

The adversarial sentence (1b) still retains most aspects of the original sentence but it depicts a completely unrelated scene. In our experiments, we study how significantly these types of textual perturbations affect the performance of multimodal translation systems. If a system is sufficiently modelling the visual modality, we expect it to ignore this type of perturbation, and to produce the correct translation by leveraging the visual input.

The main contribution of this paper is an evaluation of multimodal translation systems in the presence of adversarial textual data. This evaluation is based on four types of textual adversaries described in Section 2. We evaluate the effect of these adversaries on three state-of-the-art systems, and we also probe the visual awareness of these models by exposing them to randomly selected images. Our results show that although these systems are not greatly affected by the visual adver-


	Type	Original	Adversarial
	Num	<u>Two</u> people walking on the beach.	<i>Four</i> people walking on the beach.
	Noun	Two people walking on the <u>beach</u> .	Two people walking on the <i>grass</i> .
	NP	<u>Two people walking on the beach</u> .	<i>The beach</i> walking on <i>two people</i> .
	Prep	Two people walking <u>on</u> the beach.	Two people walking <i>through</i> the beach.

Figure 1: Examples of adversarial textual samples that we use to attack the multimodal translation models. The underlined text denotes the words or phrases that are perturbed to create the adversarial example.

saries, they are substantially affected by the textual adversaries.

## 2 Generating Textual Adversaries

We define *visual term* as a word or phrase that can be expected to be clearly illustrated in an image. In our experiments, we evaluate the performance of multimodal translation systems by modifying a *visual term* in a sentence to create a textual adversary. We create four types of adversarial samples following the methodology introduced in Young et al. (2014); Hodosh and Hockenmaier (2016); Shi et al. (2018)<sup>1</sup> The adversaries are constructed from syntactic analyses of the sentences using POS tagging, chunking, and dependency parses from the SpaCy toolkit (Honnibal and Johnson, 2015). Figure 1 presents an overview and examples of each type of adversary.

**Replace Numeral (Num):** Our simplest adversary is to replace the numeral in a sentence with a different quantity. We detect the tokens in a sentence that represent numbers (based on their part-of-speech tags) and replace them with alternative numerals. In addition, we treat the indefinite articles “a” and “an” as the numeral “one” because they are typically used as numerals in image captions. Furthermore, subsequent noun phrase chunks are either singularized or pluralized accordingly. We expect that this will have a small effect on translation quality unless the adversary introduces a serious inconsistency with the image.

**Replace Noun Head (Noun):** We extract the list of all concrete noun heads (Zwicky, 1985) from the COCO dataset (Chen et al., 2015) and swap them with the noun heads in our data. We compute concreteness<sup>2</sup> following Turney et al. (2011) and only consider words with concreteness

<sup>1</sup>The code to recreate these textual adversaries or new adversaries is available at <https://github.com/koelddc/Textual-adversaries-generation>

<sup>2</sup>The degree of concreteness in a word’s context is correlated with the likelihood that the word is used in a literal sense and not metaphorically (Turney et al., 2011).

measure  $\theta > 0.6$ . We use WordNet (Miller, 1998) heuristic hypernymy rules to replace noun heads with terms that are semantically different.

- (2)
  - a. The girl plays with the LEGOs.
  - b. The girl plays with the bricks.
  - c. \*The girl plays with the giraffes.

If our aim is to create an adversarial sentence, given 2(a), then 2(b) is too semantically similar and does not create a good adversarial example. However, (2c) creates a better adversarial example because “giraffes” are more semantically different to “LEGOs” than “bricks”. We hypothesize that the system should heavily rely on the information contained in the visual model and discard these errors to produce correct translation.

**Switch Noun Phrases (NP):** For each sentence, the position of the extracted noun phrases are switched. In the example in Figure 1, we refer to *two people* and *the beach* respectively as the partitive first noun phrase ( $NP_1$ ) and second noun phrase ( $NP_2$ ). The position of  $NP_1$  and  $NP_2$  are switched. As a result, the new sentence depicts a different scene. Such examples allow us to evaluate whether the models can identify important changes in word-order.

**Replace Preposition (Prep):** Finally, we detect the prepositions used in a sentence and randomly replace them with different prepositions. The translation system should be least sensitive to this type of adversary because it typically results in the smallest change in the meaning of the sentence, as compared to switching the noun phrases.

## 3 Experiments

We use settings similar to that of Elliott (2018) in order to make the evaluation of textual adversaries comparable to that of visual adversaries. Each system in this analysis is trained on the 29,000 English-German-image triplets in the *translation* data in the Multi30K dataset (Elliott et al., 2016). The analysis is performed on the Multi30K Test

	Original	Visual	Textual			
			Num	Noun	NP	Prep
decinit	<b>51.5</b>	+0.4	-14.0	-11.1	-11.0	-5.7
trgmul	<b>52.1</b>	+0.2	-14.8	-11.2	-11.2	-5.8
hierattn	<b>48.2</b>	-2.0	-13.2	-9.4	-11.2	-5.4
Text-only	51.5	–	-14.6	-10.4	-10.5	-6.3

Table 1: The differences in Corpus-level Meteor scores for the English–German Multi30K Test 2017 data for the different adversaries compared to the systems evaluated on the Original text and images. Visual: evaluation on the correct text but adversarial images. Textual: evaluation on the four different textual adversaries and the correct images. Text-only: performance of a text-only translation model with adversarial sentences.

2017 split (Elliott et al., 2017). The predicted translations are evaluated against human references using Meteor 1.5 (Denkowski and Lavie, 2014). The translations of the sentences with textual adversaries are evaluated against the gold standard, and not what the model *should* predict, given the adversarial input.

In this analysis, we evaluate the performance of three off-the-shelf multimodal systems: **decinit** uses a learned transformation of a global 2048D visual feature vector is used to initialise the decoder hidden state (Caglayan et al., 2017a). In **trgmul**, the target language word embeddings and 2048D visual representations are interacted through element-wise multiplication (Caglayan et al., 2017a). In **hierattn**, the decoder learns to selectively attend to a combination of the source language and a  $7 \times 7 \times 512$  volume of spatial-location preserving visual features (Libovický and Helcl, 2017). We also evaluate an attention-based text-only NMT system (Bahdanau et al., 2014) trained on only the English–German sentences in Multi30K. The model uses a conditional GRU decoder (Firat and Cho, 2016) with attention over a GRU encoder (Cho et al., 2014), as implemented in `nmtpytorch` (Caglayan et al., 2017b).

**Visual Adversaries:** Visual concepts and their relationships with the text are expected to provide rich supervision to multimodal translation systems. In addition to evaluating the robustness of these systems to textual adversaries, we also determine the interplay with visual adversaries. We pair each caption with a randomly sampled image from the test data to break the alignment between learned word semantics and visual concepts.

### 3.1 Results

In Table 1 we present the corpus-level Meteor scores for the text-only and multimodal systems when evaluated on the original data and the difference in performance when evaluating these models using the different adversaries. For visual adversaries, we confirm previously reported results of no substantial performance losses for the translations generated by the **trgmul** and **decinit** systems with visual features from unrelated images (Elliott, 2018). The **hierattn** model, however, is affected by the incongruent images, result in a 2.0 Meteor point drop in performance, indicating that the attention-based model is sensitive to the relevance of the visual input. In the case of the textual adversaries, all models suffer a significant drop in Meteor score for all types of adversary, with numeral replacements producing the largest differences. (This was a surprising result but we believe it is partially due to unseen numerals, e.g. “Seventeen” being mapped to the UNK token.) The **hierattn** model is least affected by noun and numeral replacements, and all three models are similarly affected by the noun phrase shuffle and prepositional swap adversaries. The text-only translation model is similarly affected by the textual adversaries, with the exception of the prepositional swap adversary, which has a more marked affect on performance than in the multimodal models.

In addition to the standard evaluation measures, we estimate the lexical diversity of the translations by calculating the type-to-token ratio (Templin, 1957, TTR) of the system outputs when evaluated with the congruent or incongruent visual inputs.<sup>3</sup>

<sup>3</sup>TTR has previously been used to estimate the quality of machine translation system outputs (Bentivogli et al., 2016).

	<b>Congruent</b>	<b>Incongruent</b>
decinit	0.1659	0.1655
trgmul	0.1703	0.1692
hierattn	0.1399	0.1352

Table 2: Type-to-token ratios of the system outputs given congruent and incongruent visual context.

	<b>95 % Confidence Interval</b>
Original	140.01 - 210.62
Num	335.03 - 490.07
Noun	388.02 - 511.52
NP	490.24 - 816.45
Prep	443.94 - 736.79

Table 3: The 95% confidence interval of the sentence-level perplexity of the original and each textual adversarial data samples, as estimated by GPT-2.

In our experiments, the multimodal systems were trained on the congruent image-sentence pairs so any difference in lexical diversity is likely to be due to the visual component of the respective models. However, the results in Table 2 indicate that there is no meaningful difference in TTR when the models are evaluated with the congruent or incongruent visual inputs.

### 3.2 Discussion

Given substantial decreases in Meteor score of the translations, we conducted an analysis to estimate the well-formedness of the adversarial sentences. To this end, we measure the perplexity of the perturbed sentences in each textual adversarial category using the pre-trained GPT-2 language model (Radford et al., 2018) and further average them to compute 95% confidence intervals for each category. From Table 3, we observe that the boundaries of the intervals are not overlapping, indicating statistically significant differences in distribution between the adversarial categories and the original sample<sup>4</sup>.

**Qualitative Analysis:** Figure 2 shows examples of translations under textual adversarial conditions for the **hierattn** system. We also show the output of the same system given the original text data. In these examples, we see that the system produces incorrect translations with respect to ei-

<sup>4</sup>The higher perplexities for the adversarial samples were, in part, due to incorrect grammatical conjugations.

ther the sentence or the image. In NUM, pluralizing “A” to “Two” causes the model to generate an *unknown word*<sup>5</sup> “Japan” instead of “Halloween”. The translation model is likely to have good representations of “A” and “two” because these words occur frequently in the training data, but it fails to distinguish between singulars and plurals, resulting in an incorrect translation. In PREP, swapping “in” for “up” causes the model to make an *incorrect lexical choice* “fische” (“fish”) instead of “waterfall”, which is incorrect, given the image. This example shows that a small lexical error can have a catastrophic effect on the output. This may be because the semantics of spatial relations are not diverse enough in Multi30K. In NOUN, replacing “man” with “city” causes the model to generate an output containing the *mistranslated unit* “Stadt”(“city”), although a man is clearly visible in the image. This implies that addition visual signals is not always helpful in the obvious situations where we wish to translate direct visual terms. In NP, we see that the systems fail to fully capture the information contained in the image, resulting in *under-translation*. However, unlike the output in the adversarial condition, which did not translate the important visual concept “people”, the model with the original sentence translates “People” into “Menschen”. An inspection of the training data shows that there are sentences that describe ‘people fishing’, therefore the model may be exploiting the distribution in the training data.

Overall, this analysis shows that the visual modality does not help the system to recover the correct translation, given textual adversaries.

## 4 Conclusion

In this paper, we study the potential contribution of each modality for the task of multimodal machine translation. We evaluated the performance of three multimodal translations system with adversarial source language sentences that share some aspects of the correct caption. Our evaluation offers new insights on the limitations of these systems. The results indicate that the systems are primarily performing text-based translations, which is supported by the observation that the visual adversaries do not harm the systems as much as their textual counterparts. However, the textual adversaries sometimes resulted in ungrammatical sentences, which may be addressed by adopt-

<sup>5</sup>We use the error taxonomy from Vilar et al. (2006).



**Original:** A group of young people dressed up for Halloween.  
**Baseline:** Eine Gruppe junger Menschen verkleidet.  
 NUM: Two groups of young people dressed up for halloween.  
 NMT: Zwei Frauen vor einem Glasgebäude.  
 MMT: Zwei Gruppen von jungen Menschen in Japan.  
**Reference :** Eine Gruppe junger Leute verkleidet sich für Halloween.



**Original:** A man paddles an inflatable canoe.  
**Baseline:** Ein Mann paddelt in einem aufblasbaren Kanu.  
 NOUN: A city paddles an inflatable canoe.  
 NMT: Ein Bewölkter küssen über die Absperrung.  
 MMT: Eine Stadt paddelt in einem aufblasbaren Kanu.  
**Reference:** Ein Mann paddelt in einem aufblasbaren Kanu.



**Original:** People fishing off a pier.  
**Baseline:** Menschen beim Angeln.  
 NP: A pier fishing off people.  
 NMT: Ein Bewölkter küssen über die Absperrung.  
 MMT: Ein Pier beim Angeln.  
**Reference:** Leute fischen an einem Pier.



**Original:** A beautiful waterfall in the middle of a forest.  
**Baseline:** Ein schöner Wasserfall in der Mitte eines Waldes.  
 PREP: A beautiful waterfall up the middle of a forest.  
 NMT: Zwei Frauen vor einem Glasgebäude.  
 MMT: Eine schöne Fische in einem Wald.  
**Reference:** Ein schöner Wasserfall mitten im Wald.

Figure 2: Examples of translations produced by the **hierattn** multimodal translation system. **Baseline:** the system output given the **Original** image-caption pair. NUM / NOUN / NP / PREP: The adversarial caption with the underlined replacement. **NMT:** the output of a text-only translation system, given the adversarial input. **MMT:** the output of the **hierattn** system, given the adversarial input.

ing recently-proposed neural perturbation models (Alzantot et al., 2018). We will also put more emphasis on the specific *visual term* in the image, aligning them with corresponding mention in the source data, and we plan on developing models with an max-margin ranking loss that forces the model to distinguish important differences (Huang et al., 2018) between the true image-sentence pair and well-formed adversarial perturbed sentences.

## Acknowledgements

We thank Mareike Hartmann, Ákos Kádár, Mitja Nikolaus, Kai Pierre Waelti, Thiago Castro Ferreira and the reviewers for their feedback and comments on this paper.

## References

- Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. [Generating natural language adversarial examples](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. *arXiv preprint arXiv:1608.04631*.

- Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost Van de Weijer. 2017a. Lium-cvc submissions for wmt17 multimodal translation task. *arXiv preprint arXiv:1707.04481*.
- Ozan Caglayan, Mercedes García-Martínez, Adrien Bardet, Walid Aransa, Fethi Bougares, and Loïc Barrault. 2017b. *Nmtpy: A flexible toolkit for advanced neural machine translation systems*. *Prague Bull. Math. Linguistics*, 109:15–28.
- Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. *arXiv preprint arXiv:1903.08678*.
- Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.
- Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.
- 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.
- Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*.
- Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*.
- Orhan Firat and Kyunghyun Cho. 2016. Conditional gated recurrent unit with attention mechanism. *System BLEU baseline*, 31.
- Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, et al. 2018. The memad submission to the wmt18 multimodal translation task. *arXiv preprint arXiv:1808.10802*.
- Micah Hodosh and Julia Hockenmaier. 2016. Focused evaluation for image description with binary forced-choice tasks. In *Proceedings of the 5th Workshop on Vision and Language*, pages 19–28.
- Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378.
- Jiaji Huang, Yi Li, Wei Ping, and Liang Huang. 2018. Large margin neural language model. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1191.
- Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. *arXiv preprint arXiv:1704.06567*.
- George Miller. 1998. *WordNet: An electronic lexical database*. MIT press.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.
- Haoyue Shi, Jiayuan Mao, Tete Xiao, Yuning Jiang, and Jian Sun. 2018. Learning visually-grounded semantics from contrastive adversarial samples. *arXiv preprint arXiv:1806.10348*.
- Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 543–553.
- Mildred C Templin. 1957. Certain language skills in children; their development and interrelationships.
- Peter D Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 680–690. Association for Computational Linguistics.
- David Vilar, Jia Xu, D'Haro Luis Fernando, and Hermann Ney. 2006. Error analysis of statistical machine translation output. In *LREC*, pages 697–702.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.
- Arnold M Zwicky. 1985. Heads. *Journal of linguistics*, 21(1):1–29.