Adversarial Evaluation of Multimodal Machine Translation

Desmond Elliott[*]
Department of Computer Science
University of Copenhagen
de@di.ku.dk

Abstract

The promise of combining vision and language in multimodal machine translation is that systems will produce better translations by leveraging the image data. However, inconsistent results have lead to uncertainty about whether the images actually improve translation quality. We present an adversarial evaluation method to directly examine the utility of the image data in this task. Our evaluation measures whether multimodal translation systems perform better given either the congruent image or a random incongruent image, in addition to the correct source language sentence. We find that two out of three publicly available systems are sensitive to this perturbation of the data, and recommend that all systems pass this evaluation in the future.

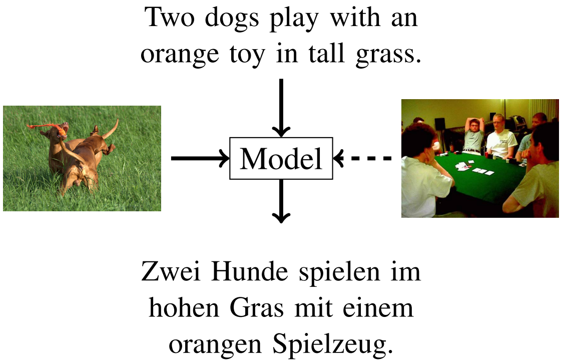

Two dogs play with an orange toy in tall grass.
Model
Zwei Hunde spielen im hohen Gras mit einem orangen Spielzeug.

Figure 1: An adversarial evaluation for multimodal translation. We measure the difference in performance when a model sees a congruent image (left) or an incongruent image (right).

1 Introduction

Multimodal machine translation is the task of translating sentences situated in a visual context, such as captioned images on social media. The core argument of this area of research is that we can produce better translations by exploiting both the source language sentence and the visual context (Elliott et al., 2015; Hitschler et al., 2016). There is some evidence to support this argument for human translation: Frank et al. (2018) found that 13% of the German evaluation data in the Multi30K dataset (Elliott et al., 2016) needed at least one post-edit to reflect the joint meaning of the visual and linguistic context. However, the evidence that visual context helps computational models is less clear. Consider the three teams that submitted contrastive multimodal and text-only variants of their systems to the 2017 Multimodal Translation Shared Task (Elliott et al., 2017): the University of Le Mans' multimodal system outperformed their text-only variant (Caglayan et al., 2017); the Oregon State University text-only system outperformed their multimodal variant (Ma et al., 2017); and the performance of the Charles University systems depended on the language pair (Libovický and Helcl, 2017). In light of these results, we need a better understanding of the role of visual context in multimodal translation systems.

We propose an adversarial evaluation Method to determine whether multimodal translation systems are aware of the visual context. We introduce a measure of *image awareness* to quantify the difference in performance in two settings: (i) when a system is presented with congruent visual data; (ii) when it is presented with incongruent visual data. In both settings, a system is presented with the correct source language sentence. See Figure 1 for an illustration of our evaluation. We hypothesise that if a system is aware of the visual context, i.e. it is actually using the image for translation, then the system will perform better when it is presented with the congruent visual data than incongruent visual data. Our evaluation is related

[*] Work carried out at the University of Edinburgh.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2974–2978
Brussels, Belgium, October 31 - November 4, 2018. ©2018 Association for Computational Linguistics

to the foiled image captions evaluation, in which the performance of an image captioning system is measured when a single word is replaced with an incorrect, but similar word (Shekhar et al., 2017); the main difference is that we *replace* the visual data instead of manipulating the text. Our work is also related to a study of question-answering systems, in which additional text was appended to the end of a document (Jia and Liang, 2017). They found that these additional text segments distracted QA systems from producing the correct answer. In contrast, our evaluation does not manipulate the textual data, instead we replace the original visual input with a random distractor.

We evaluate three publicly available multimodal translation systems with our adversarial evaluation. The main finding of this paper is that one publicly available multimodal translation system is not aware of the congruent image data. This finding raises doubts about whether state-of-the-art multimodal translation systems actually use the visual context to produce better translations. We conclude this paper by discussing whether this is likely to be due to problems with the data or with the model architectures.

2 Adversarial Evaluation

2.1 Image Awareness

We propose an adversarial evaluation method for multimodal machine translation. This method measures how a system performs when it is presented with the correct text data and either the congruent image or with an incongruent image. In this section we define two image awareness functions to measure whether a multimodal translation system is aware of the congruent visual data.

Let x be a source language sentence, y be a target language sentence, v be the congruent image, and $\bar{v}$ be an incongruent image. Image awareness is calculated using an evaluable performance measure $\mathcal{E}$. The overall image awareness of a model $\mathcal{M}$ on an evaluation dataset $\mathcal{D}$ is:

$$\Delta\text{-Awareness} = \frac{1}{|\mathcal{D}|} \sum_{i}^{|\mathcal{D}|} a_{\mathcal{M}}(x_i, y_i, v_i, \bar{v}_i) \quad (1)$$

The image awareness of a model $\mathcal{M}$ for a single instance $\mathrm{a}_{\mathcal{M}}(x_i, y_i, v_i, \bar{v}_i)$ is given by:

$$a_{\mathcal{M}}(x_i, y_i, v_i, \bar{v}_i) = \mathcal{E}(x_i, y_i, v_i) - \mathcal{E}(x_i, y_i, \bar{v}_i) \quad (2)$$

Under this definition, the output of the evaluable performance measure should be higher in the presence of the congruent data than the incongruent data, i.e. $\mathcal{E}(x_i, y_i, v_i) > \mathcal{E}(x_i, y_i, \bar{v}_i)$.[1] If this is the case, on average, then the overall image awareness of a model Δ-Awareness is positive. This can only happen when model outputs are evaluated more favourably in the presence of the the congruent image data than the incongruent image data.

2.2 Model-internal awareness Δ_I

A model-internal image measure of awareness is the difference in the probability assigned to the target language sentence y in the congruent and incongruent conditions. This is model-internal because it has the same form as the maximum-likelihood objective used to train the translation model. In this case, $\mathcal{E} = p(y|x, \cdot)$, and the difference in performance for a single instance is:

$$a_{\mathcal{M}} = \Delta_I = p(y_i|x_i, v_i) - p(y_i|x_i, \bar{v}_i) \quad (3)$$

2.3 Model-external awareness Δ_E

A model-external awareness measure could be a text-similarity evaluation or human judgement. In this paper, we use the Meteor text-similarity score (Denkowski and Lavie, 2014) because it naturally decomposes to the sentence level, and it is already the de-facto evaluation metric for multimodal machine translation (Specia et al., 2016). Let $\mathcal{E}$ be any text-similarity scoring function T that decomposes to the sentence level. The difference in performance for a single instance is defined as:

$$a_{\mathcal{M}} = \Delta_E = T(x_i, y_i, v_i) - T(x_i, y_i, \bar{v}_i) \quad (4)$$

3 Systems Evaluation

We evaluate the image awareness of three pretrained multimodal translations systems that we received by direct correspondence:

decinit: The initial state of the decoder network is set with a learned transformation of the visual data (Caglayan et al., 2017).

trgmul: The target language word embeddings are modulated by an element-wise multiplication with a learned transformation of the visual data (Caglayan et al., 2017).

[1] This assumes that a higher score means better performance for the performance measure $\mathcal{E}$. Swap the order of the operands if lower performance means better.

	C	**I**	Δ_E-Awareness
trgmul	57.3	57.3 ± 0.2	-0.001 ± 0.002
decinit	57.0	56.8 ± 0.1	0.003 ± 0.001
hierattn	55.0	53.3 ± 0.3	0.019 ± 0.003

Table 1: Corpus-level Meteor scores in the **C**ongruent and **I**ncongruent settings, along with the Meteor-awareness results. Incongruent and Δ_E-Awareness scores are the mean and standard deviation of five permutations of the visual data.

hierattn: The decoder network learns to selectively attend to a combination of the source language and the visual data (Libovický and Helcl, 2017).

Each system was trained on the 29,000 English–German–image tuples in the Multi30K dataset (Elliott et al., 2016). We evaluate the image awareness of these systems using the 1,014 tuples in the validation data, which is typically used for model selection. We select the incongruent images $\bar{v}$ by randomly shuffling the order in which the images v are associated with the source language text x. In our evaluation, we report the mean and standard deviation of randomly shuffling the image data five times. The code to evaluate your own system is publicly available.[2]

3.1 Statistical test

To determine if a model passes the proposed evaluation, we conduct a non-parametric Wilcoxon signed-rank test of the following hypothesis:

> H_1: Congruent images *improve* the quality of multimodal translation compared to incongruent images.

> H_0: Congruent images *make no difference* to the quality of multimodal translation compared to incongruent images.

We conduct this statistical test using the pairs of values that are calculated in the process of computing the the image awareness scores (Eq. 2), i.e. $\mathcal{E}(x_i, y_i, v_i)$ and $\mathcal{E}(x_i, y_i, \bar{v}_i)$.

We combine the k=5 separate p values from each test using Fisher's method and reject the null hypothesis H_0 if the result of the χ^2 test with $2k$ degrees of freedom is $p \leq 0.005$.[3]

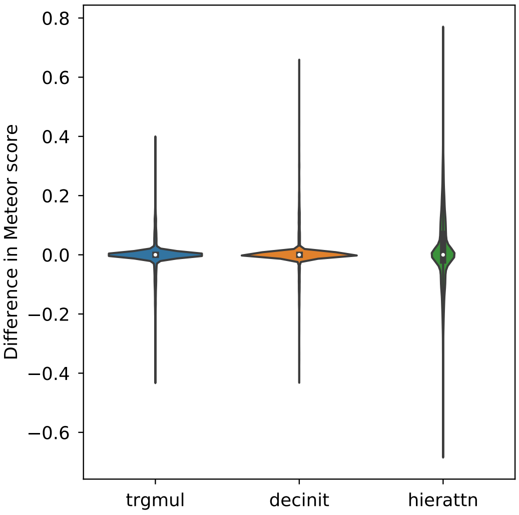


Figure 2: Violin plots of the Meteor-awareness scores for evaluated models. The white dot marks the median value, the thick gray bar shows the interquartile range, and the thin gray bar is the 95% confidence interval. The width of the plots show the kernel density estimate of the distributions.

3.2 Results

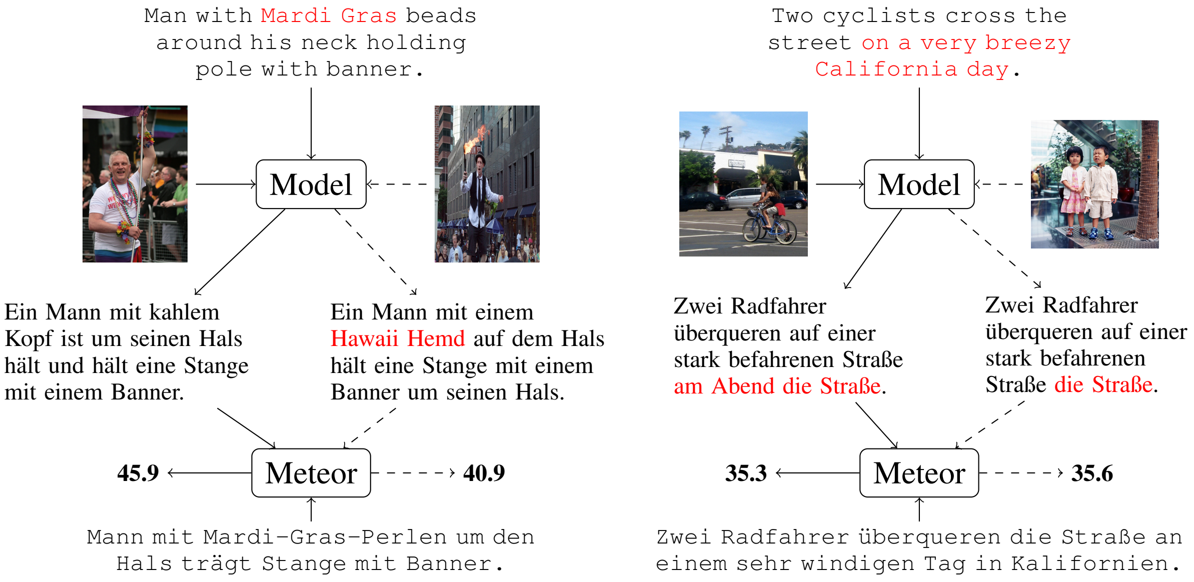
Table 1 shows the corpus-level results of a Meteor-based evaluation and the Meteor-awareness evaluation. We find that images improve the quality of the **hierattn** system (χ^2 = 136.74, $p < 0.0001$), and images also improve the quality of the **decinit** system (χ^2 = 32.79, $p = 0.0003$). Images make no difference to the quality of the translations generated by the **trgmul** system (χ^2 = 8.98, $p = 0.533$). To complement these tests, Figure 2 shows violin plots of the Meteor-awareness scores. These show that the translations generated by the **trgmul** and **decinit** systems are most likely to result in no difference in Meteor score between the congruent and incongruent conditions.

We now turn our attention to the results of the probability-awareness evaluation. Images improve the quality of the **trgmul** system (χ^2 = 52.55, $p < 0.0001$), and images also improve the quality of the **hierattn** system (χ^2 = 622.03, $p <$ 0.0001). Images make no difference to the quality of **decinit** system (χ^2 = 6.49, p = 0.772).

Figure 3 shows examples of translations pro-

[2] http://github.com/elliottd/awareness

[3] We use a stricter threshold to reduce the chance of false positive findings (Benjamin et al., 2018).



Man with Mardi Gras beads around his neck holding pole with banner.

Ein Mann mit kahlem Kopf ist um seinen Hals hält und hält eine Stange mit einem Banner.

Ein Mann mit einem Hawaii Hemd auf dem Hals hält eine Stange mit einem Banner um seinen Hals.

Mann mit Mardi-Gras-Perlen um den Hals trägt Stange mit Banner.

(a) Congruent is better than Incongruent

Two cyclists cross the street on a very breezy California day.

Zwei Radfahrer überqueren auf einer stark befahrenen Straße am Abend die Straße.

Zwei Radfahrer überqueren auf einer stark befahrenen Straße die Straße.

Zwei Radfahrer überqueren die Straße an einem sehr windigen Tag in Kalifornien.

(b) Incongruent is better than Congruent

Figure 3: Examples of the difference in Meteor awareness for the **hierattn** system. In each example, the source sentence is shown at the top and the reference sentence is shown at the bottom, both in `Typewriter` font. The congruent image is on the left, and the incongruent image is on the right.

duced by the **hierattn** system for sentences paired with congruent / incongruent images. Figure 3 (a) shows an example with high positive difference in Meteor score. The incongruent image causes the translation system to refer to an unseen Hawaiian shirt. In neither setting does the system translate the phrase "Mardi Gras". Figure 3 (b) shows an example with a negative difference in Meteor score. The congruent image results in a long translation with poor coverage of the reference, which Meteor punishes more severely than the shorter translation arising from the incongruent image. In neither setting does the model translate the prepositional phrase "on a very breezy California day".

4 Discussion

4.1 Data problems

We posit that the current Multi30K training data does not necessarily require systems to use the visual context to solve the translation task. Elliott et al. (2016) note that the German translation data was produced without showing the translators the images, and Frank et al. (2018) found that 13% of the Multi30K test data needed to be post-edited to reflect the joint semantics of both modalities. We recommend that entirety of the German Multi30K training data should be post-edited so that future systems are more likely to require a joint understanding of the visual and linguistic context.[4] We note that a similar issue was found in a visual question answering dataset, resulting in the creation of a new "balanced" dataset (Goyal et al., 2017).

4.2 The role of model architectures

The key difference between the systems evaluated in this paper is how they use the visual context. The **hierattn** system learns a timestep-dependent context vector over a location-preserving 3D volume of image features, whereas the **trgmul** and **decinit** systems use an average-pool of the 3D location-preserving features. In our evaluation, the only system that is aware of the congruent image data for both types of image-awareness is the **hierattn** system that learns a spatial context over the image. Learning to attend to specific regions of the image may prove to be crucial to improving translations with visual context.

5 Conclusion

We proposed an adversarial evaluation method to determine whether multimodal translation systems are aware of the visual context. This evaluation method measures the difference in the perfor-

[4] We planned to repeat the adversarial evaluation with the Multi30K French data, which *was* created by showing the annotators the images. However, we did not receive pre-trained models for all the systems for English–French translation.

mance of a system given the congruent or an incongruent image as additional context. We found that two out of three publicly available multimodal translation systems were improved by the congruent visual context, when compared to the incongruent visual context. We encourage researchers to use this method to evaluate their own systems. Future work includes augment existing multimodal translation models with an additional adversarial objective that forces the model to perform better in the presence of the congruent image than a random incongruent image. We will also apply this evaluation method to other tasks that use additional context, e.g. images in visual-question answering, or part-of-speech tags in neural machine translation.

Acknowledgements

Barry Haddow asked if we knew whether the additional image data actually improved the quality of multimodal machine translation. I claimed that there was indisputable evidence, given the results of the human evaluation at the 2017 Multimodal Translation Shared Task. I am not sure if I agree with myself any more.

Desmond Elliott is supported by an Amazon Research Award. Stella Frank, Ákos Kádár, Emiel van Miltenburg, and the Edinburgh NLP group provided valuable feedback on this paper.

References

Daniel J Benjamin, James O Berger, Magnus Johannesson, Brian A Nosek, E-J Wagenmakers, Richard Berk, Kenneth A Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, et al. 2018. Redefine statistical significance. *Nature Human Behaviour*, 2(1):6.

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.

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, Baltimore, Maryland, USA.

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*, pages 215–233, Copenhagen, Denmark.

Desmond Elliott, Stella Frank, and Eva Hasler. 2015. Multi-Language Image Description with Neural Sequence Models. *CoRR*, abs/1510.04709.

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.

Stella Frank, Desmond Elliott, and Lucia Specia. 2018. Assessing multilingual multimodal image description: Studies of native speaker preferences and translator choices. *Natural Language Engineering*, 24(3):393–413.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition*.

Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. 2016. Multimodal Pivots for Image Caption Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2399–2409.

Robin Jia and Percy Liang. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031.

Jindřich Libovický and Jindřich Helcl. 2017. Attention Strategies for Multi-Source Sequence-to-Sequence Learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 196–202, Vancouver, Canada.

Mingbo Ma, Dapeng Li, Kai Zhao, and Liang Huang. 2017. OSU Multimodal Machine Translation System Report. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 465–469, Copenhagen, Denmark.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. FOIL it! Find One mismatch between Image and Language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265, Vancouver, Canada.

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*, pages 543–553, Berlin, Germany.