Probing the Need for Visual Context in Multimodal Machine Translation

Ozan Caglayan
LIUM, Le Mans University
ozan.caglayan@univ-lemans.fr

Pranava Madhyastha
Imperial College London
pranava@imperial.ac.uk

Lucia Specia
Imperial College London
l.specia@imperial.ac.uk

Loïc Barrault
LIUM, Le Mans University
loic.barrault@univ-lemans.fr

Abstract

Current work on multimodal machine translation (MMT) has suggested that the visual modality is either unnecessary or only marginally beneficial. We posit that this is a consequence of the very simple, short and repetitive sentences used in the only available dataset for the task (Multi30K), rendering the source text sufficient as context. In the general case, however, we believe that it is possible to combine visual and textual information in order to ground translations. In this paper we probe the contribution of the visual modality to state-of-the-art MMT models by conducting a systematic analysis where we partially deprive the models from source-side textual context. Our results show that under limited textual context, models are capable of leveraging the visual input to generate better translations. This contradicts the current belief that MMT models disregard the visual modality because of either the quality of the image features or the way they are integrated into the model.

1 Introduction

Multimodal Machine Translation (MMT) aims at designing better translation systems which take into account auxiliary inputs such as images. Initially organized as a shared task within the First Conference on Machine Translation (WMT16) (Specia et al., 2016), MMT has so far been studied using the Multi30K dataset (Elliott et al., 2016), a multilingual extension of Flickr30K (Young et al., 2014) with translations of the English image descriptions into German, French and Czech (Elliott et al., 2017; Barrault et al., 2018).

The three editions of the shared task have seen many exciting approaches that can be broadly categorized as follows: (i) multimodal attention using convolutional features (Caglayan et al., 2016; Calixto et al., 2016; Libovický and Helcl, 2017; Helcl et al., 2018) (ii) cross-modal interactions with spatially-unaware global features (Calixto and Liu, 2017; Ma et al., 2017; Caglayan et al., 2017a; Madhyastha et al., 2017) and (iii) the integration of regional features from object detection networks (Huang et al., 2016; Grönroos et al., 2018). Nevertheless, the conclusion about the contribution of the visual modality is still unclear: Grönroos et al. (2018) consider their multimodal gains "modest" and attribute the largest gain to the usage of external parallel corpora. Lala et al. (2018) observe that their multimodal word-sense disambiguation approach is not significantly different than the monomodal counterpart. The organizers of the latest edition of the shared task concluded that the multimodal integration schemes explored so far resulted in marginal changes in terms of automatic metrics and human evaluation (Barrault et al., 2018). In a similar vein, Elliott (2018) demonstrated that MMT models can translate without significant performance losses even in the presence of features from unrelated images.

These empirical findings seem to indicate that images are ignored by the models and hint at the fact that this is due to representation or modeling limitations. We conjecture that the most plausible reason for the linguistic dominance is that – at least in Multi30K – the source text is sufficient to perform the translation, eventually preventing the visual information from intervening in the learning process. To investigate this hypothesis, we introduce several input degradation regimes (Section 2) and revisit state-of-the-art MMT models (Section 3) to assess their behavior under degraded regimes. We further probe the visual sensitivity by deliberately feeding features from unrelated images. Our results (Section 4) show that MMT models successfully exploit the visual modality when the linguistic context is scarce, but indeed tend to be less sensitive to this modality when exposed to complete sentences.

Proceedings of NAACL-HLT 2019, pages 4159–4170
Minneapolis, Minnesota, June 2 - June 7, 2019. ©2019 Association for Computational Linguistics

2 Input Degradation

In this section we propose several degradations to the input language modality to simulate conditions where sentences may miss crucial information. We denote a set of translation pairs by $\mathcal{D}$ and indicate degraded variants with subscripts. Both the training and the test sets are degraded.

Color Deprivation. We consistently replace *source* words that refer to colors with a special token `[v]` ($\mathcal{D}_{\text{C}}$ in Table 1). Our hypothesis is that a monomodal system will have to rely on source-side contextual information and biases, while a multimodal architecture could potentially capitalize on color information extracted by exploiting the image and thus obtain better performance. This affects 3.3% and 3.1% of the words in the training and the test set, respectively.

Entity Masking. The Flickr30K dataset, from which Multi30K is derived, has also been extended with coreference chains to tag mentions of *visually depictable* entities in image descriptions (Plummer et al., 2015). We use these to mask out the head nouns in the *source* sentences ($\mathcal{D}_{\text{N}}$ in Table 1). This affects 26.2% of the words in both the training and the test set. We hypothesize that a multimodal system should heavily rely on the images to infer the missing parts.

Progressive Masking. A *progressively* degraded variant $\mathcal{D}_k$ replaces all but the first k tokens of *source* sentences with `[v]`. Unlike the color deprivation and entity masking, masking out suffixes does not guarantee systematic removal of visual context, but rather simulates an increasingly low-resource scenario. Overall, we form 16 degraded variants $\mathcal{D}_k$ (Table 1) where $k \in \{0, 2, \ldots, 30\}$. We stop at $\mathcal{D}_{30}$ since 99.8% of the sentences in Multi30K are shorter than 30 words with an average sentence length of 12 words. $\mathcal{D}_0$ – where the only remaining information is the source sentence length – is an interesting case from two perspectives: a neural machine translation (NMT) model trained on it resembles a target language model, while an MMT model becomes an image captioner with access to "expected length information".

Visual Sensitivity. Inspired by Elliott (2018), we experiment with *incongruent decoding* in order to understand how sensitive the multimodal systems are to the visual modality. This is achieved by explicitly violating the test-time semantic congruence across modalities. Specifically, we feed the visual features in *reverse* sample order to break image-sentence alignments. Consequently, a model capable of integrating the visual modality would likely deteriorate in terms of metrics.

$\mathcal{D}$	**a**	**lady**	**in**	**a**	**blue**	**dress**	**singing**
$\mathcal{D}_{\text{C}}$	a	lady	in	a	`[v]`	dress	singing
$\mathcal{D}_{\text{N}}$	a	`[v]`	in	a	blue	`[v]`	singing
$\mathcal{D}_4$	a	lady	in	a	`[v]`	`[v]`	`[v]`
$\mathcal{D}_2$	a	lady	`[v]`	`[v]`	`[v]`	`[v]`	`[v]`
$\mathcal{D}_0$	`[v]`	`[v]`	`[v]`	`[v]`	`[v]`	`[v]`	`[v]`

Table 1: An example of the proposed input degradation schemes: $\mathcal{D}$ is the original sentence.

3 Experimental Setup

Dataset. We conduct experiments on the English→French part of Multi30K. The models are trained on the concatenation of the *train* and *val* sets (30K sentences) whereas *test2016 (dev)* and *test2017 (test)* are used for early-stopping and model evaluation, respectively. For *entity masking*, we revert to the default Flickr30K splits and perform the model evaluation on *test2016*, since *test2017* is not annotated for entities. We use word-level vocabularies of 9,951 English and 11,216 French words. We use Moses (Koehn et al., 2007) scripts to lowercase, normalize and tokenize the sentences with hyphen splitting. The hyphens are stitched back prior to evaluation.

Visual Features. We use a ResNet-50 CNN (He et al., 2016) trained on ImageNet (Deng et al., 2009) as image encoder. Prior to feature extraction, we center and standardize the images using ImageNet statistics, resize the shortest edge to 256 pixels and take a center crop of size 256x256. We extract spatial features of size 2048x8x8 from the final convolutional layer and apply L_2 normalization along the depth dimension (Caglayan et al., 2018). For the non-attentive model, we use the 2048-dimensional global average pooled version (pool5) of the above convolutional features.

Models. Our baseline NMT is an attentive model (Bahdanau et al., 2014) with a 2-layer bidirectional GRU encoder (Cho et al., 2014) and a 2-layer conditional GRU decoder (Sennrich et al., 2017). The second layer of the decoder receives the output of the attention layer as input.

	$\mathcal{D}$	$\mathcal{D}_C$
NMT	70.6 ± 0.5	68.4 ± 0.1
INIT	70.7 ± 0.2	**68.9** ± 0.1
HIER	70.9 ± 0.3	**69.0** ± 0.3
DIRECT	70.9 ± 0.2	**68.8** ± 0.3

Table 2: Baseline and color-deprivation METEOR scores: bold systems are significantly different from the NMT system within the same column (p-value ≤ 0.03).

For the MMT model, we explore the basic multimodal attention (DIRECT) (Caglayan et al., 2016) and its hierarchical (HIER) extension (Libovický and Helcl, 2017). The former linearly projects the concatenation of textual and visual context vectors to obtain the multimodal context vector, while the latter replaces the concatenation with another attention layer. Finally, we also experiment with encoder-decoder initialization (INIT) (Calixto and Liu, 2017; Caglayan et al., 2017a) where we initialize both the encoder and the decoder using a non-linear transformation of the pool5 features.

Hyperparameters. The encoder and decoder GRUs have 400 hidden units and are initialized with 0 except the multimodal INIT system. All embeddings are 200-dimensional and the decoder embeddings are tied (Press and Wolf, 2016). A dropout of 0.4 and 0.5 is applied on source embeddings and encoder/decoder outputs, respectively (Srivastava et al., 2014). The weights are decayed with a factor of $1e{-}5$. We use ADAM (Kingma and Ba, 2014) with a learning rate of $4e{-}4$ and mini-batches of 64 samples. The gradients are clipped if the total norm exceeds 1 (Pascanu et al., 2013). The training is early-stopped if *dev* set METEOR (Denkowski and Lavie, 2014) does not improve for ten epochs. All experiments are conducted with *nmtpytorch*[1] (Caglayan et al., 2017b).

4 Results

We train all systems three times each with different random initialization in order to perform significance testing with *multeval* (Clark et al., 2011). Throughout the section, we always report the mean over three runs (and the standard deviation) of the considered metrics. We decode the translations with a beam size of 12.

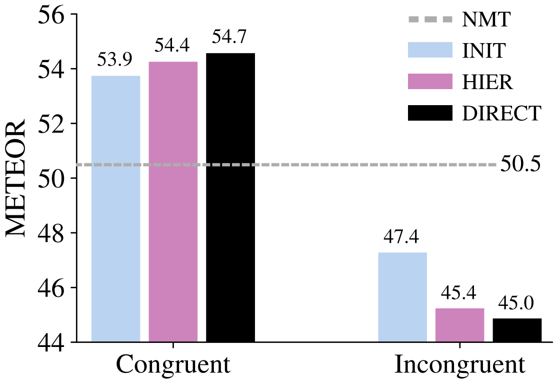


Figure 1: Entity masking: all masked MMT models are significantly better than the masked NMT (dashed). Incongruent decoding severely worsens all systems. The vanilla NMT baseline is 75.9[2].

We first present *test2017* METEOR scores for the baseline NMT and MMT systems, when trained on the full dataset $\mathcal{D}$ (Table 2). The first column indicates that, although MMT models perform slightly better on average, they are not significantly better than the baseline NMT. We now introduce and discuss the results obtained under the proposed degradation schemes. Please refer to Table 5 and the appendix for qualitative examples.

4.1 Color Deprivation

Unlike the inconclusive results for $\mathcal{D}$, we observe that all MMT models are significantly better than NMT when color deprivation is applied ($\mathcal{D}_C$ in Table 2). If we further focus on the subset of the test set subjected to color deprivation (247 sentences), the gain increases to 1.6 METEOR for HIER. For the latter subset, we also computed the average color accuracy per sentence and found that the attentive models are 12% better than the NMT (32.5→44.5) whereas the INIT model only brings 4% (32.5→36.5) improvement. This shows that more complex MMT models are better at integrating visual information to perform better.

4.2 Entity Masking

The gains are much more prominent with entity masking, where the degradation occurs at a larger scale: Attentive MMT models show up to 4.2 METEOR improvement over NMT (Figure 1). We observed a large performance drop with *incongruent decoding*, suggesting that the visual modality is

[1] github.com/lium-lst/nmtpytorch

[2] Since entity masking uses Flickr30K splits (Section 3) rather than our splits, the scores are not comparable to those from other experiments in this paper.

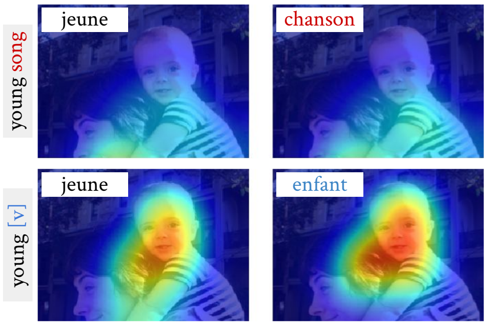


Figure 2: Baseline MMT (top) translates the misspelled "son" while the masked MMT (bottom) correctly produces "enfant" (child) by focusing on the image.

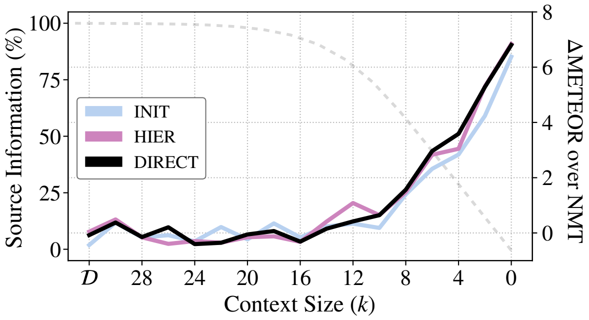


Figure 3: Multimodal gain in absolute METEOR for *progressive masking*: the dashed gray curve indicates the percentage of non-masked words in the training set.

	+ Gain (⇃ Incongruence Drop)		
	INIT	HIER	DIRECT
Czech	+1.4 (⇃ 2.9)	+1.7 (⇃ 3.5)	+1.7 (⇃ 4.1)
German	+2.1 (⇃ 4.7)	+2.5 (⇃ 5.9)	+2.7 (⇃ 6.5)
French	+3.4 (⇃ 6.5)	+3.9 (⇃ 9.0)	+4.2 (⇃ 9.7)

Table 3: *Entity masking* results across three languages: all MMT models perform significantly better than their NMT counterparts (p-value ≤ 0.01). The incongruence drop applies on top of the MMT score.

	$\mathcal{D}_4$	$\mathcal{D}_6$	$\mathcal{D}_{12}$	$\mathcal{D}_{20}$	$\mathcal{D}$
DIRECT	**32.3**	**42.2**	**64.5**	**70.1**	**70.9**
Incongruent Dec.	⇃ 6.4	⇃ 5.5	⇃ 1.4	⇃ 0.7	⇃ 0.7
Blinding	⇃ 3.9	⇃ 2.9	⇃ 0.4	⇃ 0.5	⇃ 0.3
NMT	⇃ 3.7	⇃ 2.6	⇃ 0.6	⇃ 0.2	⇃ 0.3

Table 4: The impact of incongruent decoding for *progressive masking*: all METEOR differences are against the DIRECT model. The blinded systems are both trained and decoded using incongruent features.

now much more important than previously demonstrated (Elliott, 2018). A comparison of attention maps produced by the baseline and masked MMT models reveals that the attention weights are more consistent in the latter. An interesting example is given in Figure 2 where the masked MMT model attends to the correct region of the image and successfully translates a dropped word that was otherwise a spelling mistake ("son"→"song").

Czech and German. In order to understand whether the above observations are also consistent across different languages, we extend the *entity masking* experiments to German and Czech parts of Multi30K. Table 3 shows the gain of each MMT system with respect to the NMT model and the subsequent drop caused by incongruent decoding[3]. First, we see that the multimodal benefits clearly hold for German and Czech, although the gains are lower than for French[4]. Second, when we compute the average drop from using incongruent images across all languages, we see how conservative the INIT system is (⇃ 4.7) compared to HIER (⇃ 6.1) and DIRECT (⇃ 6.8). This raises a follow-up question as to whether the hidden state initialization eventually loses its impact throughout the recurrence where, as a consequence, the only modality processed is the text.

4.3 Progressive Masking

Finally, we discuss the results of the progressive masking experiments for French. Figure 3 clearly shows that as the sentences are progressively degraded, all MMT systems are able to leverage the visual modality. When the multimodal task becomes image captioning at k=0, MMT models improve over the language-model counterpart by ∼7 METEOR. Further qualitative examples show that the systems perform surprisingly well by producing visually plausible sentences (see Table 5 and the Appendix).

To get a sense of the visual sensitivity, we pick the DIRECT models trained on four degraded variants and perform *incongruent decoding*. We notice that as the amount of linguistic information increases, the gap narrows down: the MMT system gradually becomes less perplexed by the incongruence or, put in other words, less sensitive to the visual modality (Table 4).

[3] For example, the INIT system for French (Figure 1) surpasses the baseline (50.5) by reaching 53.9 (+3.4), which ends up at 47.4 (⇃ 6.5) after incongruent decoding.

[4] This is probably due to the morphological richness of DE and CS which is suboptimally handled by word-level MT.

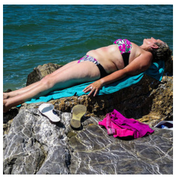
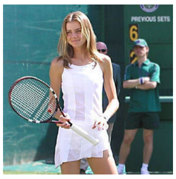
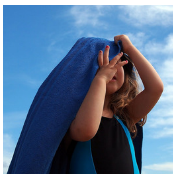
	SRC :	an older woman in [v] [v] [v] [v] [v] [v] [v] [v] [v] [v] [v]
	NMT :	une femme âgée avec un t-shirt blanc et des lunettes de soleil est assise sur un banc
		(an older woman with a white t-shirt and sunglasses is sitting on a bank)
	MMT :	une femme âgée en **maillot de bain rose** est assise sur un **rocher au bord de l'eau**
		(an older woman with a pink swimsuit is sitting on a rock at the seaside)
	REF :	une femme âgée **en bikini** bronze sur **un rocher au bord de l'océan**
		(an older woman in bikini is tanning on a rock at the edge of the ocean)
	SRC :	a young [v] in [v] holding a tennis [v]
	NMT :	un jeune garçon en bleu tenant une raquette de tennis
		(a young boy in blue holding a tennis racket)
	MMT :	**une** jeune **femme** en **blanc** tenant une raquette de tennis
	REF :	**une** jeune **femme** en **blanc** tenant une raquette de tennis
		(a young girl in white holding a tennis racket)
	SRC :	little girl covering her face with a [v] towel
	NMT :	une petite fille couvrant son visage avec une serviette blanche
		(a little girl covering her face with a white towel)
	MMT :	une petite fille couvrant son visage avec une serviette **bleue**
	REF :	une petite fille couvrant son visage avec une serviette **bleue**
		(a little girl covering her face with a blue towel)

Table 5: Qualitative examples from progressive masking, entity masking and color deprivation, respectively. Underlined and bold words highlight the bad and good lexical choices. MMT is an attentive system.

We then conduct a contrastive "blinding" experiment where the DIRECT models are not only fed with incongruent features at decoding time but also trained with them from scratch. The results suggest that the blinded models learn to ignore the visual modality. In fact, their performance is equivalent to NMT models.

5 Discussion and Conclusions

We presented an in-depth study on the potential contribution of images for multimodal machine translation. Specifically, we analysed the behavior of state-of-the-art MMT models under several degradation schemes in the Multi30K dataset, in order to reveal and understand the impact of textual predominance. Our results show that the models explored are able to integrate the visual modality if the available modalities are complementary rather than redundant. In the latter case, the primary modality (text) sufficient to accomplish the task. This dominance effect corroborates the seminal work of Colavita (1974) in Psychophysics where it has been demonstrated that visual stimuli dominate over the auditory stimuli when humans are asked to perform a simple audiovisual discrimination task. Our investigation using source degradation also suggests that visual grounding can increase the robustness of machine translation systems by mitigating input noise such as errors in the source text. In the future, we would like to devise models that can learn when and how to integrate multiple modalities by taking care of the complementary and redundant aspects of them in an intelligent way.

Acknowledgments

This work is a follow-up on the research efforts conducted within the "Grounded sequence-to-sequence transduction" team of the JSALT 2018 Workshop. We would like to thank Jindřich Libovický for contributing the hierarchical attention to *nmtpytorch* during the workshop. We also thank the reviewers for their valuable comments.

Ozan Caglayan and Loïc Barrault received funding from the French National Research Agency (ANR) through the CHIST-ERA M2CR project under the contract ANR-15-CHR2-0006-01. Lucia Specia and Pranava Madhyastha received funding from the MultiMT (H2020 ERC Starting Grant No. 678017) and MMVC (Newton Fund Institutional Links Grant, ID 352343575) projects.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *Computing Research Repository*, arXiv:1409.0473. Version 7.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 308–327, Belgium, Brussels. Association for Computational Linguistics.

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. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 432–439, Copenhagen, Denmark. Association for Computational Linguistics.

Ozan Caglayan, Adrien Bardet, Fethi Bougares, Loïc Barrault, Kai Wang, Marc Masana, Luis Herranz, and Joost van de Weijer. 2018. LIUM-CVC submissions for WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 603–608, Belgium, Brussels. Association for Computational Linguistics.

Ozan Caglayan, Loïc Barrault, and Fethi Bougares. 2016. Multimodal attention for neural machine translation. *Computing Research Repository*, arXiv:1609.03976.

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.

Iacer Calixto, Desmond Elliott, and Stella Frank. 2016. DCU-UvA multimodal MT system report. In *Proceedings of the First Conference on Machine Translation*, pages 634–638, Berlin, Germany. Association for Computational Linguistics.

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 992–1003, Copenhagen, Denmark. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 176–181, Stroudsburg, PA, USA. Association for Computational Linguistics.

Francis B. Colavita. 1974. Human sensory dominance. *Perception & Psychophysics*, 16 (2):409–412.

J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

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. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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.

Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, and Raúl Vázquez. 2018. The MeMAD submission to the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation*, pages 609–617, Belgium, Brussels. Association for Computational Linguistics.

K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Jindřich Helcl, Jindřich Libovický, and Dusan Varis. 2018. CUNI system for the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 622–629, Belgium, Brussels. Association for Computational Linguistics.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 639–645, Berlin, Germany. Association for Computational Linguistics.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *Computing Research Repository*, arXiv:1412.6980.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

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, Volume 2: Shared Task Papers*, pages 630–637, Belgium, Brussels. Association for Computational Linguistics.

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 (Volume 2: Short Papers)*, pages 196–202, Vancouver, Canada. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Pranava Swaroop Madhyastha, Josiah Wang, and Lucia Specia. 2017. Sheffield MultiMT: Using object posterior predictions for multimodal machine translation. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 470–476, Copenhagen, Denmark. Association for Computational Linguistics.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1310–1318, Atlanta, Georgia, USA. PMLR.

B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. 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.

Ofir Press and Lior Wolf. 2016. Using the output embedding to improve language models. *Computing Research Repository*, arXiv:1608.05859.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

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.

A Qualitative Examples

In this appendix, we provide further translation examples for color deprivation (Table 6), entity masking (Table 7) and progressive masking (Table 8). Specifically for the entity masking experiments, we also give further examples to showcase the behavior of the visual attention in Figure 4 and Figure 5.

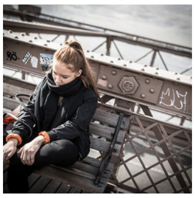
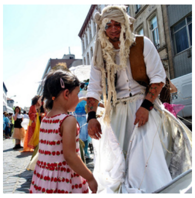
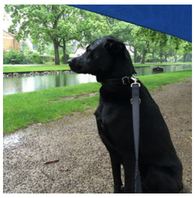
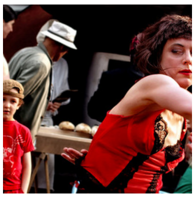
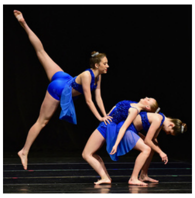
	SRC:	a girl in [v] is sitting on a bench
	NMT:	pink
	Init:	pink
	Hier:	**black**
	Direct:	**black**
	SRC:	a man dressed in [v] talking to a girl
	NMT:	black
	Init:	black
	Hier:	**white**
	Direct:	**white**
	SRC:	a [v] dog sits under a [v] umbrella
	NMT:	brown / **blue**
	Init:	**black / blue**
	Hier:	**black / blue**
	Direct:	**black / blue**
	SRC:	a woman in a [v] top is dancing as a woman and boy in a [v] shirt watch
	NMT:	blue / blue
	Init:	blue / blue
	Hier:	**red / red**
	Direct:	**red / red**
	SRC:	three female dancers in [v] dresses are performing a dance routine
	NMT:	white
	Init:	white
	Hier:	white
	Direct:	**blue**

Table 6: Color deprivation examples from the English→French models: bold indicates correctly predicted cases. The colors generated by the models are shown in English for the sake of clarity.

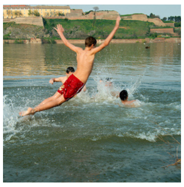
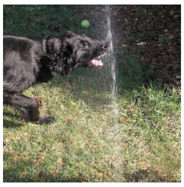
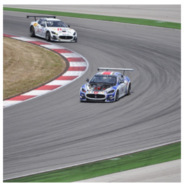
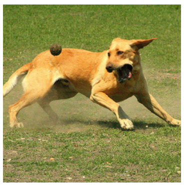
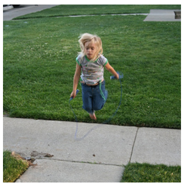
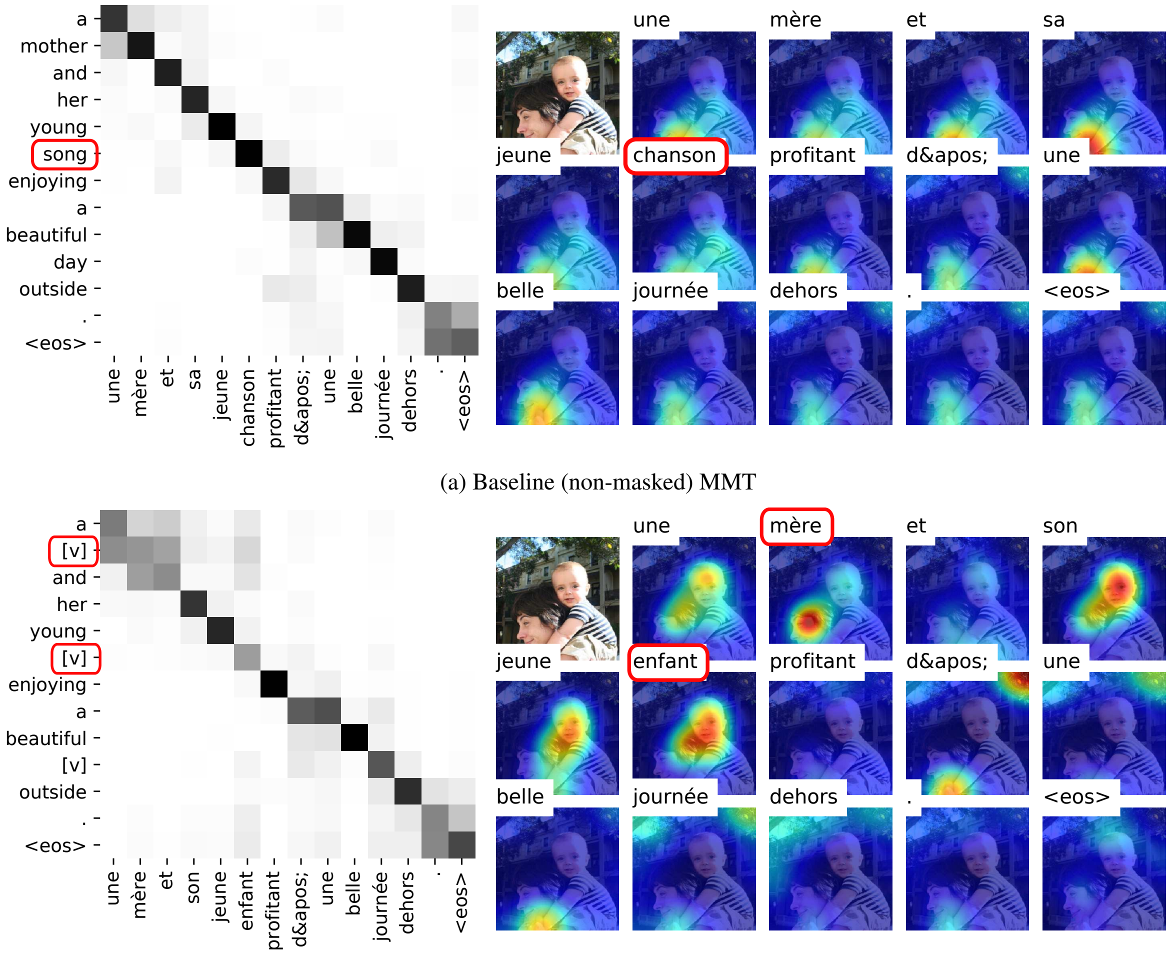
	SRC : a [v] in a red [v] plays in the [v] NMT : un garçon en t-shirt rouge joue dans la neige *(a boy in a red t-shirt plays in the snow)* MMT : un garçon en **maillot de bain** rouge joue dans **l'eau** REF : un garçon en **maillot de bain** rouge joue dans **l'eau** *(a boy in a red swimsuit plays in the water)*
	SRC : a [v] drinks [v] outside on the [v] NMT : un homme boit du vin dehors sur le trottoir *(a man drinks wine outside on the sidewalk)* MMT : un **chien** boit de **l'eau** dehors sur **l'herbe** REF : un **chien** boit de **l'eau** dehors sur **l'herbe** *(a dog drinks water outside on the grass)*
	SRC : two [v] are driving on a [v] NMT : deux hommes font du vélo sur une route *(two men riding bicycles on a road)* MMT : deux **voitures roulent sur une piste** *(two cars driving on a track/circuit)* REF : deux **voitures roulent sur** un circuit
	SRC : a [v] turns on the [v] to pursue a flying [v] NMT : un homme tourne sur la plage pour attraper un frisbee volant *(a man turns on the beach to catch a flying frisbee)* MMT : un **chien** tourne sur **l'herbe** pour attraper un frisbee volant *(a dog turns on the grass to catch a flying frisbee)* REF : un **chien** tourne sur **l'herbe** pour poursuivre une balle en l'air *(a dog turns on the grass to chase a ball in the air)*
	SRC : a [v] jumping [v] on a [v] near a parking [v] NMT : un homme sautant à cheval sur une plage près d'un parking *(a man jumping on a beach near a parking lot)* MMT : une **fille** sautant à la **corde** sur un **trottoir** près d'un parking REF : une **fille** sautant à la **corde** sur un **trottoir** près d'un parking *(a girl jumping rope on a sidewalk near a parking lot)*

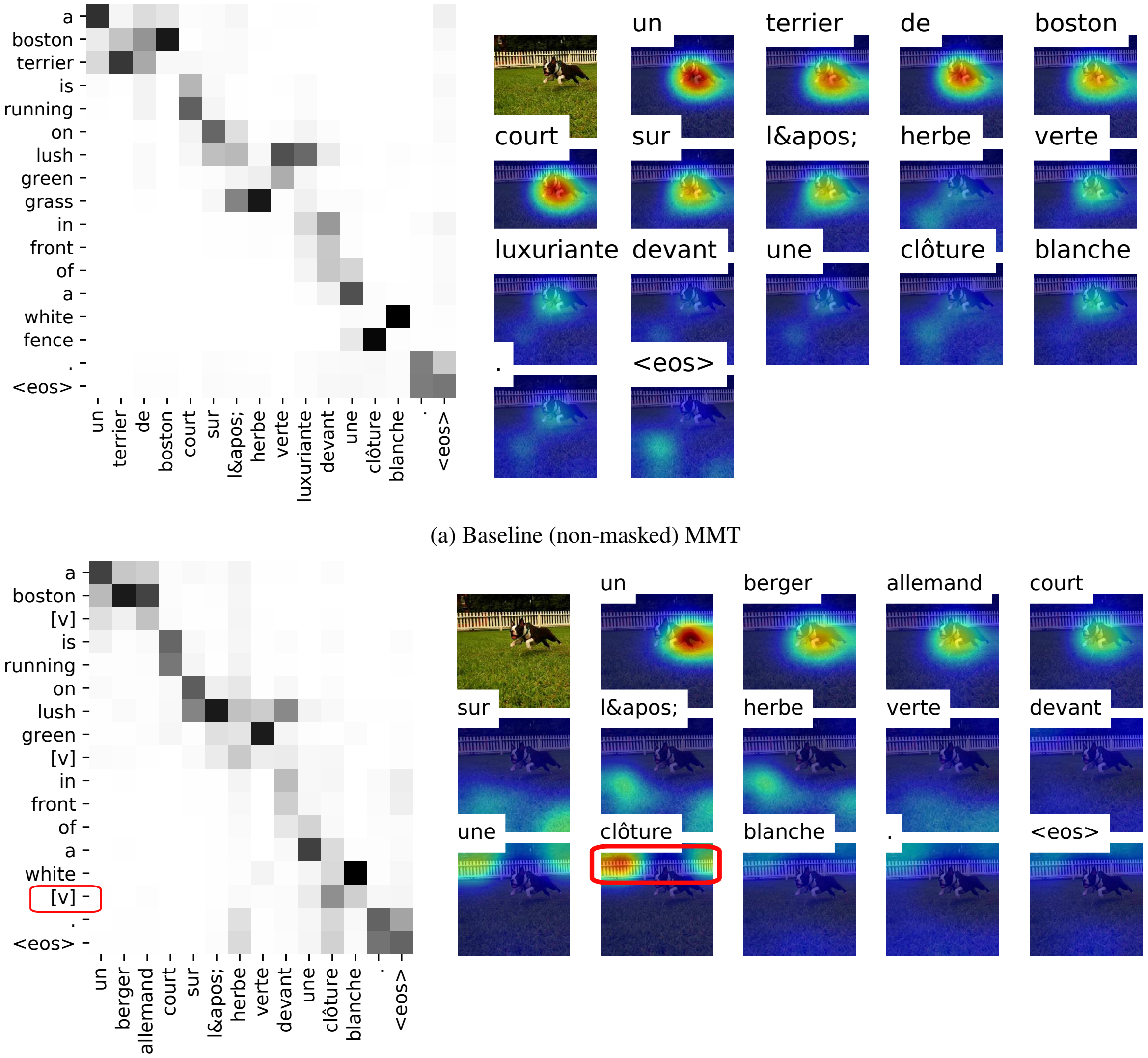
Table 7: Entity masking examples from the English→French models: underlined and bold words highlight bad and good lexical choices, respectively. English translations are provided in parentheses. MMT is an attentive model.



(a) Baseline (non-masked) MMT

(b) Entity-masked MMT

Figure 4: Attention example from entity masking experiments: (a) Baseline MMT translates the misspelled "son" (song → chanson) while (b) the masked MMT achieves a correct translation ([v] → enfant) by exploiting the visual modality.



(a) Baseline (non-masked) MMT

(b) Entity-masked MMT

Figure 5: Attention example from entity masking experiments where *terrier, grass* and *fence* are dropped from the source sentence: (a) Baseline MMT is not able to shift attention from the salient *dog* to the *grass* and *fence*, (b) the attention produced by the masked MMT first shifts to the background area while translating “on lush green `[v]`” then focuses on the *fence*.

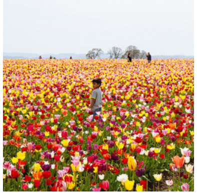
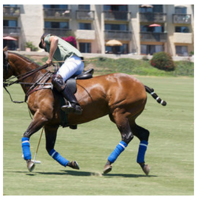
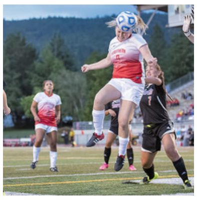
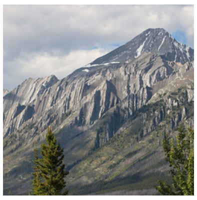
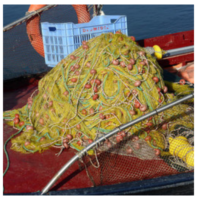
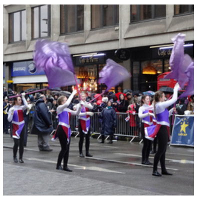
	SRC: a child [v] [v] [v] [v] [v] [v] NMT: un enfant avec des lunettes de soleil en train de jouer au tennis *(a child with sunglasses playing tennis)* MMT: un enfant **est debout dans un champ de fleurs** *(a child is standing in field of flowers)* REF: un enfant **dans un champ de tulipes** *(a child in a field of tulips)*
	SRC: a jockey riding his [v] [v] NMT: un jockey sur son vélo *(a jockey on his bike)* MMT: un jockey sur son **cheval** REF: un jockey sur son **cheval** *(a jockey on his horse)*
	SRC: girls are playing a [v] [v] [v] NMT: des filles jouent à un jeu de cartes *(girls are playing a card game)* MMT: des filles jouent un **match de football** REF: des filles jouent un **match de football** *(girls are playing a football match)*
	SRC: trees are in front [v] [v] [v] [v] [v] NMT: des vélos sont devant un bâtiment en plein air *(bicycles are in front of an outdoor building)* MMT: des **arbres** sont devant la **montagne** *(trees are in front of the mountain)* REF: des **arbres** sont devant une grande **montagne** *(trees are in front of a big mountain)*
	SRC: a fishing net on the deck of a [v] [v] NMT: un filet de pêche sur la terrasse d'un bâtiment *(a fishing net on the terrace of a building)* MMT: un filet de pêche sur le **pont d'un bateau** *(a fishing net on the deck of a boat)* REF: un filet de pêche sur le **pont d'un bateau** rouge *(a fishing net on the deck of a red boat)*
	SRC: girls wave purple flags [v] [v] [v] [v] [v] [v] [v] NMT: des filles en t-shirts violets sont assises sur des chaises dans une salle de classe *(girls in purple t-shirts are sitting on chairs in a classroom)* MMT: des filles en costumes violets **dansent dans une rue en ville** *(girls in purple costumes dance on a city street)* REF: des filles agitent des drapeaux violets tandis qu'elles défilent dans la rue *(girls wave purple flags as they parade down the street)*

Table 8: English→French progressive masking examples: underlined and bold words highlight bad and good lexical choices, respectively. English translations are provided in parentheses. MMT is an attentive model.