Distilling Translations with Visual Awareness

Julia Ive[1]**, Pranava Madhyastha**[2] and **Lucia Specia**[2]
[1]DCS, University of Sheffield, UK
[2]Department of Computing, Imperial College London, UK
`j.ive@sheffield.ac.uk`
`{pranava,l.specia}@imperial.ac.uk`

Abstract

Previous work on multimodal machine translation has shown that visual information is only needed in very specific cases, for example in the presence of ambiguous words where the textual context is not sufficient. As a consequence, models tend to learn to ignore this information. We propose a translate-and-refine approach to this problem where images are only used by a second stage decoder. This approach is trained jointly to generate a good first draft translation and to improve over this draft by (i) making better use of the target language textual context (both left and right-side contexts) and (ii) making use of visual context. This approach leads to the state of the art results. Additionally, we show that it has the ability to recover from erroneous or missing words in the source language.

1 Introduction

Multimodal machine translation (MMT) is an area of research that addresses the task of translating texts using context from an additional modality, generally static images. The assumption is that the visual context can help ground the meaning of the text and, as a consequence, generate more adequate translations. Current work has focused on datasets of images paired with their descriptions, which are crowdsourced in English and then translated into different languages, namely the Multi30K dataset (Elliott et al., 2016).

Results from the most recent evaluation campaigns in the area (Elliott et al., 2017; Barrault et al., 2018) have shown that visual information can be helpful, as humans generally prefer translations generated by multimodal models than by their text-only counterparts. However, previous work has also shown that images are only needed in very specific cases (Lala et al., 2018). This is also the case for humans. Frank et al. (2018)

(see Figure 1) concluded that visual information is needed by humans in the presence of the following: **incorrect or ambiguous** source words and **gender-neutral words** that need to be marked for gender in the target language. In an experiment where human translators were asked to first translate descriptions based on their textual context only and then revise their translation based on a corresponding image, they report that these three cases accounted for 62-77% of the revisions in the translations in two subsets of Multi30K.

Ambiguities are very frequent in Multi30K, as in most language corpora. Barrault et al. (2018) shows that in its latest test set, 358 (German) and 438 (French) instances (out of 1,000) contain at least one word that has more than one translation in the training set. However, these do not always represent a challenge for translation models: often the text context can easily disambiguate words (see baseline translation in Figure 4(a)); additionally, the models are naturally biased to generate the most frequent translation of the word, which by definition is the correct one in most cases.

The need to gender-mark words in a target language when translating from English can be thought of as a disambiguation problem, except that the text context is often less telling and the frequency bias plays ends up playing a bigger role (see baseline translation in Figure 4(c)). This has been shown to be a common problem in neural machine translation (Vanmassenhove et al., 2018; Font and Costa-Jussà, 2019), as well as in areas such as image captioning (Hendricks et al., 2018) and co-reference resolution (Zhao et al., 2018).

Incorrect source words are common in Multi30K, as in many other crowdsourced or user-generated dataset. In this case the context may not be enough (see DE translation in Figure 1(c)). We posit that models should be robust to such a type of noise and note that similar treatment would be

EN: Three children in <u>football</u> uniforms are playing <u>football</u>.

DE: Drei Kinder in <u>Fußballtrikots</u> spielen <u>Fußball</u>.

PE: Drei Kinder in <u>Footballtrikots</u> spielen <u>Football</u>.

(a) **Ambiguous** word *football* translated as *soccer* (*Fußball*)

EN: A baseball <u>player</u> in a black shirt just tagged a player in a white shirt.

DE: Ein Baseballspieler in einem schwarzen Shirt fängt <u>einen Spieler</u> in einem weißen Shirt.

PE: <u>Eine Baseballspielerin</u> in einem schwarzen Shirt fängt <u>eine Spielerin</u> in einem weißen Shirt.

(b) **Gender-neutral** word *player* translated as male player (*Spieler*)

EN: A woman wearing a white <u>shirt</u> works out on an elliptical machine.

DE: Eine Frau in einem weißen <u>Shirt</u> trainiert auf einem Crosstrainer.

PE: Eine Frau in einem weißen <u>Pullover</u> trainiert auf einem Crosstrainer.

(c) **Inaccurate English** word *shirt* instead of *sweater* or *pullover*

Figure 1: Examples of lexical and gender ambiguity, and inaccurate English description where post-edits (PE) required the image to correct human translation from English (EN) to German (DE).

required for out of vocabulary (OOV) words, i.e. correct words that are unknown to the model.

We propose an approach that takes into account the strengths of a text-only baseline model and only **refines** its translations when needed. Our approach is based on **deliberation networks** (Xia et al., 2017) to jointly learn to generate draft translations and refine them based on left and right side target context as well as structured visual information. This approach outperforms previous work.

In order to further probe how well our models can address the three problems mentioned above, we perform a controlled experiment where we minimise the interference of the frequency bias by masking ambiguous and gender-related words, as well as randomly selected words (to simulate noise and OOV). This experiment shows that our multimodal refinement approach outperforms the text-only one in more complex linguistic setups.

Our main **contributions** are: (i) a novel approach to MMT based on deliberation networks and structured visual information which gives state of the art results (Sections 3.2 and 5.1); (ii) a frequency bias-free investigation on the need for visual context in MMT (Sections 4.2 and 5.2); and (iii) a thorough investigation on different vi-

sual representations for transformer-based architectures (Section 3.3).

2 Related work

MMT: Approaches to MMT vary with regards to how they represent images and how they incorporate this information in the models. Initial approaches use RNN-based sequence to sequence models (Bahdanau et al., 2015) enhanced with a single, global image vector, extracted as one of the layers of a CNN trained for object classification (He et al., 2016), often the penultimate or final layer. The image representation is integrated into the MT models by initialising the encoder or decoder (Elliott et al., 2015; Caglayan et al., 2017; Madhyastha et al., 2017); element-wise multiplication with the source word annotations (Caglayan et al., 2017); or projecting the image representation and encoder context to a common space to initialise the decoder (Calixto and Liu, 2017). Elliott and Kádár (2017) and Helcl et al. (2018) instead model the source sentence and reconstruct the image representation jointly via multi-task learning.

An alternative way of exploring image representations is to have an attention mechanism

(Bahdanau et al., 2015) on the output of the last convolutional layer of a CNN (Xu et al., 2015). The layer represents the activation of K different convolutional filters on evenly quantised $N \times N$ spatial regions of the image. Caglayan et al. (2017) learn the attention weights for both source text and visual encoders, while Calixto et al. (2017); Delbrouck and Dupont (2017) combine both attentions independently via a gating scalar, and Libovický and Helcl (2017); Helcl et al. (2018) apply a hierarchical attention distribution over two projected vectors where the attention for each is learnt independently.

Helcl et al. (2018) is the closest to our work: we also use a doubly-attentive transformer architecture and explore spatial visual information. However, we differ in two main aspects (Section 3): (i) our approach explores additional textual context through a second pass decoding process and uses visual information only at this stage, and (ii) in addition to convolutional filters we use object-level visual information. The latter has only been explored to generate a single global representation (Grönroos et al., 2018) and used for example to initialise the encoder (Huang et al., 2016). We note that translation refinement is different translation re-ranking from a text-only model based on image representation (Shah et al., 2016; Hitschler et al., 2016; Lala et al., 2018), since the latter assumes that the correct translation can already be produced by a text-only model.

Caglayan et al. (2019) investigate the importance and the contribution of multimodality for MMT. They perform careful experiments by using input degradation and observe that, specially under limited textual context, multimodal models exploit the visual input to generate better translations. Caglayan et al. (2019) also show that MMT systems exploit visual cues and obtain correct translations even with typographical errors in the source sentences. In this paper, we build upon this idea and investigate the potential of visual cues for refining translation.

Translation refinement: The idea of treating machine translation as a two step approach dates back to statistical models, e.g. in order to improve a draft sentence-level translation by exploring document-wide context through hill-climbing for local refinements (Hardmeier et al., 2012). Iterative refinement approaches have also been proposed that start with a draft translation and then predict discrete substitutions based on an attention mechanism (Novak et al., 2016), or using non-autoregressive methods with a focus on speeding up decoding (Lee et al., 2018). Translation refinement can also be done through learning a separate model for automatic post-editing (Niehues et al., 2016; Junczys-Dowmunt and Grundkiewicz, 2017; Chatterjee et al., 2018), but this requires additional training data with draft translations and their correct version.

An interesting approach is that of deliberation networks, which jointly train an encoder and first and second stage decoders (Xia et al., 2017). The second stage decoder has access to both left and right side context and this has been shown to improve translation (Xia et al., 2017; Hassan et al., 2018). We follow this approach as it offers a very flexible framework to incorporate additional information in the second stage decoder.

3 Model

We base our model on the **transformer architecture** (Vaswani et al., 2017) for neural machine translation. Our implementation is a multi-layer encoder-decoder architecture that uses the `tensor2tensor`[1] (Vaswani et al., 2018) library. The encoder and decoder blocks are as follows:

Encoder Block ($\mathcal{E}$): The encoder block comprises of 6 layers, with each containing two sub-layers of multi-head self-attention mechanism followed by a fully connected feed forward neural network. We follow the standard implementation and employ residual connections between each layer, as well as layer normalisation. The output of the encoder forms the encoder memory which consists of contextualised representations for each of the source tokens ($M_{\mathcal{E}}$).

Decoder Block ($\mathcal{D}$): The decoder block also comprises of 6 layers. It contains an additional sublayer which performs multi-head attention over the outputs of the encoder block. Specifically, decoding layer d_{l_i} is the result of a) multi-head attention over the outputs of the encoder which in turn is a function of the encoder memory and the outputs from the previous layer: $A_{\mathcal{D} \to \mathcal{E}} = f(M_{\mathcal{E}}, d_{l_{i-1}})$ where, the keys and values are the encoder outputs and the queries correspond to the decoder input, and b) the multi-head self attention

---

[1] https://github.com/tensorflow/tensor2tensor

which is a function of the generated outputs from the previous layer: $A_{\mathcal{D}} = f(d_{l_{i-1}})$.

3.1 Deliberation networks

Deliberation networks (Hassan et al., 2018; Xia et al., 2017) build on the standard sequence to sequence architecture to add an additional decoder block (in our case, with 3 layers – see Figure 2). The additional decoder (also referred to as second-pass decoder) is conditioned on the source and sampled outputs from the standard transformer decoder (the first-pass decoder). More concretely, the second-pass decoder ($\mathcal{D}'$) at layer d'_l consists of $A_{\mathcal{D}'}$, $A_{\mathcal{D}' \rightarrow \mathcal{E}}$, $A_{\mathcal{D}' \rightarrow \mathcal{D}}$, where, $A_{\mathcal{D}'}$ and $A_{\mathcal{D}' \rightarrow \mathcal{E}}$ is similar to the standard deliberation architecture multi-head attention over the encoder memory and self attention respectively while, $A_{\mathcal{D}' \rightarrow \mathcal{D}}$ is the multi-head attention over outputs O_d from the first-pass decoder ($\mathcal{D}$) ($A_{\mathcal{D} \rightarrow \mathcal{E}} = f(O_d, d'_{l_{i-1}})$).[2] In our experiments, we obtain samples as a set of translations from the first-pass decoder using beam-search. Given a translation candidate, O_d consists of the first-pass decoder's hidden layer before softmax concatenated with the embeddings of the resultant words.
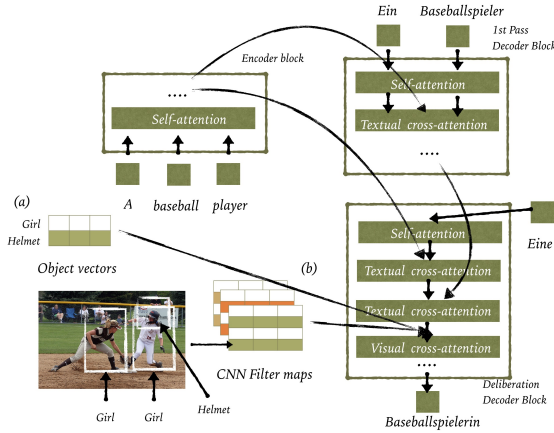

Figure 2: Our deliberation architecture: The second-pass decoder is conditioned on the source and samples output from the first-pass decoder. The second-pass decoder has access to (a) the object based features represented by embeddings, or (b) spacial image features.

3.2 Multimodal transformer & deliberation

Our multimodal transformer models follow one of the two formulations below for conditioning translations on image information:

Additive image conditioning (AIC): A projected image vector is added to each of the outputs of the encoder. The projections matrices are parameters that are jointly learned with the model.

Attention over image features (AIF): The model attends over image features, as in Helcl et al. (2018), where the decoder block now contains an additional cross-attention sub-layer $A_{\mathcal{D}' \rightarrow \mathcal{V}}$ which attends to the visual information ($\mathcal{V}$). The keys and values correspond to the visual information.

Within the deliberation network framework, based on the previously discussed observation (Section 1) that images are only needed in a small number of cases, we propose to add visual cross-attention only to the second-pass decoder block (see Figure 2).

3.3 Image features

Motivated by previous work that indicates the importance of structured information from images (Caglayan et al., 2017; Wang et al., 2018; Madhyastha et al., 2018), we focus on structural forms of image representations, including the spatially aware feature maps from CNNs and information extracted from automatic object detectors.

Spatial image features: We use spatial feature maps from the last convolutional layer of a pretrained ResNet-50 (He et al., 2016) CNN-based image classifier for every image.[3] These feature maps contain output activations for various filters while preserving spatial information. They have been used in various vision to language tasks including image captioning (Xu et al., 2015) and multimodal machine translation (Section 2). Our formulation for the integration of these features into the deliberation network is shown in Figure 2, setup (b). We use the the AIF setup and refer to models that use the representation as att.

Object-based image features: We use a bag-of-objects representation where the objects are obtained using an off-shelf object detector (Kuznetsova et al., 2018) based on the Open Images dataset. This representations is a sparse 545-dimensional vector with the frequency of each (545) given object in an image. This is inspired by previous research that investigates the potential of object-based information for vision to language tasks (Mitchell et al., 2012; Wang et al., 2018). We

---

use the the AIC setup and refer to models that use the representation as `sum`.

Object-based embedding features: The bag-of-objects representations makes it hard to exploit object-to-object similarity, since visual representations of different objects can be very different. To mitigate this, we propose a simple extension using bag-of-object *embeddings*. We represent each object using the pre-trained GLoVe-based (Pennington et al., 2014) 50-dimensional word vectors for their categories (e.g. *woman*). We use the the AIF based setup and refer to models that use the representation as `obj` (Figure 2 setup (a)).

4 Experimental settings

4.1 Data

We build and test our MMT models on the **Multi30K** dataset (Elliott et al., 2016). Each image in Multi30K contains one English (EN) description taken from Flickr30K (Young et al., 2014) and human translations into German (DE), French (FR) and Czech (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018). The dataset contains 29,000 instances for training, 1,014 for development, and 1,000 for test. We only experiment with German and French, which are languages for which we have in-house expertise for the type of analysis we present. In addition to the official Multi30K test set (test 2016), we also use the test set from the latest WMT evaluation competition, test 2018 (Barrault et al., 2018).[4]

4.2 Degradation of source

In addition to using the Multi30K dataset as is (standard setup), we probe the ability of our models to address the three linguistic phenomena where additional context has been proved important (Section 1): ambiguities, gender-neutral words and noisy input. In a controlled experiment where we aim to remove the influence of frequency biases, we degrade the source sentences by masking words through three strategies to replace words by a placeholder: random source words, ambiguous source words and gender unmarked source words. The procedure is applied to the train, validation and test sets. For the resulting dataset generated for each setting, we compare models having access to text-only context versus

additional text and multimodal contexts. We seek to get insights into the contribution of each type of context to address each type of degradation.

Random content words In this setting (RND) we simulate erroneous source words by randomly dropping source content words. We first tag the entire source sentences using the `spacy` toolkit (Honnibal and Montani, 2017) and then drop nouns, verbs, adjectives and adverbs and replace these with a default `BLANK` token. By focusing on content words, we differ from previous work that suggests that neural machine translation is robust to non-content word noise in the source (Klubička et al., 2017).

Ambiguous words In this setting (AMB), we rely on the MLT dataset (Lala et al., 2018) which provides a list of source words with multiple translations in the Multi30k training set. We replace ambiguous words with the `BLANK` token in the source language, which results in two language-specific datasets.

Person words In this setting (PERS), we use the Flickr Entities dataset (Plummer et al., 2017) to identify all the words that were annotated by humans as corresponding to the category person.[5] We then replace such source words with the `BLANK` token.

The statistics of the resulting datasets for the three degradation strategies are shown in Table 1. We note that RND and PERS are the same for language pairs as the degradation only depends on the source side, while for AMB the words replaced depend on the target language.

setup	% sent.	avg. blanks per sent.
RND	100	1.5
AMB DE	83	2
AMB FR	77	1.8
PERS	92	1.6

Table 1: Statistics of datasets after applying source degradation strategies

4.3 Models

Based on the models described in Section 3 we experiment with eight variants: (a) baseline transformer model (`base`); (b) `base` with AIC

---

[4]The pre-processed datasets provided by the organisers were used without additional pre-processing.

[5]We pre-processed the initial dataset to remove noise. We also add the gender-marked pronouns *he*, *she*, *her* and *his* to the person word list.

(`base+sum`); (c) `base` with AIF using spacial (`base+att`) or object based (`base+obj`) image features; (d) standard deliberation model (`del`); (e) deliberation models enriched with image information: `del+sum`, `del+att` and `del+obj`.

4.4 Training

In all cases, we optimise our models with cross entropy loss. For deliberation network models, we first train the standard transformer model until convergence, and use it to initialise the encoder and first-pass decoder. For each of the training samples, we follow (Xia et al., 2017) and obtain a set of 10-best samples from the first pass decoder, with a beam search of size 10. We use these as the first-pass decoder samples. We use Adam as optimiser (Kingma and Ba, 2014) and train the model until convergence.[6]

	model	test 2016		test 2018	
		M	B	M	B
DE	MMT (Helcl et al., 2018)	53.1	38.4	-	-
	base	54.5	36.4	45.0	26.5
	base+sum	54.2	35.9	45.0	26.4
	base+att	54.5	36.9	45.3	27.2
	base+obj	54.5	36.4	45.0	26.7
	del	**55.5***	37.7	**46.3***	27.7
	del+sum	**55.2***	37.3	**46.3***	27.7
	del+att	**55.1***	37.2	**46.1***	27.4
	del+obj	**55.6***	38.0	**46.5***	27.6
FR	MMT (Helcl et al., 2018)	75.0	60.6	-	-
	base	73.7	59.0	56.4	37.0
	base+sum	73.9	59.2	56.6	37.1
	base+att	73.5	58.7	56.1	36.2
	base+obj	72.9	57.3	55.8	36.3
	del	**74.6***	60.1	**57.2***	37.8
	del+sum	**74.3***	59.6	**56.9***	37.2
	del+att	73.7†	59.2	56.3†	36.9
	del+obj	**74.4***	59.8	**57.0***	37.4

Table 2: Results for the test sets 2016 and 2018. M denotes METEOR, B – BLEU; * marks statistically significant changes for METEOR (p-value ≤ 0.05) as compared to `base`, † – as compared to `del`. Bold highlights statistically significant improvements. We report previous state of the art results for multimodal models from (Helcl et al., 2018).

5 Results

In this section we present results of our experiments, first in the original dataset without any

---

[6]We built on the `tensor2tensor` implementation of deliberation nets in `https://github.com/ustctf/delibnet` using the `transformer_big` parameters with a learning rate of 0.05 with 8K warmup steps for both the first and the second-pass decoders, and early stopping with the patience of 10 epochs based on the validation BLEU score.

source degradation (Section 5.1) and then in the setup with various source degradation strategies (Section 5.2).

5.1 Standard setup

Table 2 shows the results of our main experiments on the 2016 and 2018 test sets for French and German. We use Meteor (Denkowski and Lavie, 2014) as the main metric, as in the WMT tasks (Barrault et al., 2018). We compare our transformer baseline to transformer models enriched with image information, as well as to the deliberation models, with or without image information.

We first note that our multimodal models achieve the state of the art performance for transformer networks (constrained models) on the English-German dataset, as compared to (Helcl et al., 2018). Second, our deliberation models lead to significant improvements over this baseline across test sets (average ΔMETEOR $= 1$, ΔBLEU $= 1$).

Transformer-based models enriched with image information (`base+sum`, `base+att` and `base+obj`), on the other hand, show no major improvements with respect to the `base` performance. This is also the case for deliberation models with image information (`del+sum`, `del+att`, `del+obj`), which do not show significant improvement over the vanilla deliberation performance (`del`).

However, as it has been shown in the WMT shared tasks on MMT (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018), automatic metrics often fail to capture nuances in translation quality, such as, the ones we expect the visual modality to help with, which – according to human perception – lead to better translations. To test this assumption in our settings, we performed human evaluation involving professional translators and native speakers of both French and German (three annotators).

The annotators were asked to rank randomly selected test samples according to how well they convey the meaning of the source, given the image (50 samples per language pair per annotator). For each source segment, the annotator was shown the outputs of three systems: `base+att`, the current MMT state-of-the-art (Helcl et al., 2018), `del` and `del+obj`. A rank could be assigned from 1 to 3, allowing ties (Bojar et al., 2017). Annotators could assign zero rank to all translations if they

	EN:	Two men work under the hood of a white <u>race car</u>.
	base+att:	Zwei Männer arbeiten unter der Motorhaube eines weißen <u>Rennens</u>.
	del:	Zwei Männer arbeiten unter der Motorhaube eines weißen <u>Autos</u>.
	del+obj:	Zwei Männer arbeiten unter der Motorhaube eines weißen <u>Rennwagen</u>.
	DE:	Zwei Männer arbeiten unter der Haube eines weißen <u>Rennautos</u>.

(a) `base+att` translates *race car* with *Rennen* (race), `del` with *Auto* (car) and `del+obj` with *Rennwagen* (race car).
Objects: land, vehicle, car, wheel

	EN:	A young child holding an oar <u>paddling</u> a blue kayak <u>in a body of water</u>.
	base+att:	Un jeune enfant tenant une rame dans un kayak bleu.
	del:	Un jeune enfant tenant une rame dans un kayak bleu <u>sur un plan d'eau</u>.
	del+obj:	Un jeune enfant tenant une rame dans un kayak bleu <u>pagayant</u> <u>sur un plan d'eau</u>.
	FR:	Un jeune enfant avec une rame <u>pagayant</u> dans un kayak bleu <u>sur un plan d'eau</u>.

(b) `del` and `del+obj` translate *in a body of water* with *sur un plan d'eau* (on a body of water), missing in `base+att`. `del+obj` translates the word *paddling* with *pagayant* (paddling). **Objects**: paddle, canoe

Figure 3: Examples of improvements of `del` and `del+obj` over `base+att` for test set 2016 for French and German. Underlined words represent some of the improvements.

were judged incomprehensible.

Following the common practice in WMT (Bojar et al., 2017), each system was then assigned a score which reflects the proportion of times it was judged to be better or equal other systems.

Table 3 shows the human evaluation results. They are consistent with the automatic evaluation results when it comes to the preference of humans towards the deliberation-based setups, but show a more positive outlook regarding the addition of visual information (del+obj over del) for French.

lang	base+att	del	del+obj
DE	0.35	**0.62**	0.59
FR	0.41	0.6	**0.67**

Table 3: Human ranking results: normalised rank (micro-averaged). Bold highlights best results.

Manual inspection of translations suggests that deliberation setups tend to improve both the grammaticality and adequacy of the first pass outputs. For German, the most common modifications performed by the second-pass decoder are substitutions of adjectives and verbs (for test 2016, 15% and 12% respectively, of all the edit distance operations). Changes to adjectives are mainly grammatical, changes to verbs are contextual (e.g., changing *laufen* to *rennen*, both verbs mean run, but the second refers to running very fast). For

French, 15% of all the changes are substitutions of nouns (for test 2016). These are again very contextual. For example, the French word *travailleur* (worker) is replaced by *ouvrier* (manual worker) in the contexts where tools, machinery or buildings are mentioned. For our analysis we used again spacy.

The information on detected objects is particularly helpful for specific adequacy issues. Figure 3 demonstrates some such cases. In the first case, the base+att model misses the translation of *race car*: the German word *Rennen* translates only the word *race*. del introduces the word *car* (*Auto*) into the translation. Finally, del+obj correctly translates the expression *race car* (*Rennwagen*) by exploiting the object information. For French, del translates the source part *in a body of water*, missing from the base+att translation. del+obj additionally translated the word *paddling* according to the detected object Paddle.

5.2 Source degradation setup

Results of our source degradation experiments are shown in Table 4. A first observation is that – as with the standard setup – the performance of our deliberation models is overall better than that of the base models. The results of the multimodal models differ for German and French. For German, del+obj is the most successful configuration and shows statistically significant improve-

	model	RND				AMB				PERS			
		test 2016		test 2018		test 2016		test 2018		test 2016		test 2018	
		M	B	M	B	M	B	M	B	M	B	M	B
DE	base	45.6	27.1	37.7	20.0	48.4	30.1	38.9	21.0	47.0	28.6	40.3	22.2
	del	44.6*	25.1	36.8*	18.1	47.7	29.0	38.0*	19.0	47.5	29.0	40.9	22.0
	del+sum	45.7†	27.2	38.1†	19.9	46.9*†	27.9	37.2*†	18.7	**48.1***	29.8	**41.1***	22.4
	del+obj	**46.5*†**	28.1	**39.0*†**	20.7	**49.8*†**	31.3	**40.0*†**	21.3	**48.1***	29.4	**41.6*†**	23.4
FR	base	59.3	43.4	46.3	28.1	66.4	51.2	49.2	30.4	63.9	48.6	50.3	31.7
	del	**61.0***	45.3	**47.1***	28.4	**67.3***	52.2	**50.2***	31.3	**64.5***	49.3	**51.2***	32.4
	del+sum	**60.4***	44.4	**47.5***	29.3	**67.7***	52.8	**50.4***	31.5	**65.0***	49.7	**51.1***	32.1
	del+obj	**61.3***	45.4	**47.9*†**	29.4	**67.7***	52.6	**50.5***	31.7	**65.0***	49.5	**50.9***	32.2

Table 4: Results for the test sets 2016 and 2018 for the three degradation configurations: RND, AMB and PERS. M denotes METEOR, B – BLEU; * marks statistically significant changes as computed for METEOR (p-value ≤ 0.05) as compared to base, † – as compared to del. Bold highlights statistically significant improvements over base.

ments over base for all setups. Moreover, for RND and AMB, it shows statistically significant improvements over del. However, especially for RND and AMB, del and del+sum are either the same or slightly worse than base.

For French, all the deliberation models show statistically significant improvements over base (average ΔMETEOR $= 1$, ΔBLEU $= 1.1$), but the image information added to del only improve scores significantly for test 2018 RND.

This difference in performances for French and German is potentially related to the need of more significant restructurings while translating from English into German.[7] This is where a more complex del+obj architecture is more helpful. This is especially true for RND and AMB setups where blanked words could also be verbs, the part-of-speech most influenced by word order differences between English and German (see the decreasing complexity of translations for del and del+obj for the example (c) in Figure 4).

To get an insight into the contribution of different contexts to the resolution of blanks, we performed manual analysis of examples coming from the English-German base, del and del+obj setups (50 random examples per setup), where we count correctly translated blanks per system.

The results are shown in Table 5. As expected, they show that the RND and AMB blanks are more difficult to resolve (at most 40% resolved as compared to 61% for PERS). Translations of the majority of those blanks tend to be guessed by the

setup	base	del	del+obj	gold
RND	22	23	**24**	79
AMB	29	25	**33**	88
PERS	43	46	**51**	84

Table 5: Results of human annotation of blanked translations (English-German). We report counts of blanks resolved by each system, as well as total source blank count for each selection (50 sentences selected randomly).

textual context alone (especially for verbs). Image information is more helpful for PERS: we observe an increase of 10% in resolved blanks for del+obj as compared to del. However, for PERS the textual context is still enough in the majority of the cases: models tend to associate men with sports or women with cooking and are usually right (see Figure 4 example (c)).

The cases where image helps seem to be those with rather generic contexts: see Figure 4 (b) where *enjoying a summer day* is not associated with any particular gender and make other models choose *homme* (man) or *femme* (woman), and only base+obj chooses *enfant* (child) (the option closest to the reference).

In some cases detected objects are inaccurate or not precise enough to be helpful (e.g., when an object Person is detected) and can even harm correct translations.

6 Conclusions

We have proposed a novel approach to multi-modal machine translation which makes better use of context, both textual and visual. Our results show that further exploring textual context through deliberation networks already leads

	EN:	Three farmers harvest rice out in a rice <u>field</u>.
	`base`:	Drei Bauern ernten sich mit einem <u>Reisfeld</u>.
	`del`:	Drei Bauern ernten Reis mit einem <u>Reisfeld</u>.
	`del+obj`:	Drei Bauern ernten sich mit einem <u>Reishut</u> auf.
	DE:	Drei Farmer ernten Reis auf einem <u>Feld</u>.

(a) Example of a blank resolved by the textual context for AMB: *field* translated as *Reisfeld* (rice field) by `base`. `del+obj` incorrectly translated the blank into *Reishut* (rice hat) due to detected objects. **Objects**: person, clothing, mammal

	EN:	The <u>boy</u> is outside enjoying a summer day.
	`base`:	<u>L'homme</u> profite d'une journée d'été.
	`del`:	<u>La femme</u> profite d'une journée d'été.
	`del+obj`:	<u>L'enfant</u> profite d'une journée d'été.
	FR:	<u>Le garçon</u> est dehors, profitant d'une journée d'été.

(b) Example of a blank resolved by the multimodal context for PERS. The textual context is too generic and `del+obj` uses the detected objects to correctly translate *boy* into *l'enfant* (child). **Objects**: clothing, face, tree, boy, jeans

	EN:	Dirt <u>biker</u> makes a sloping turn in a forest during the fall.
	`base`:	<u>Geländemotorradfahrer</u> macht in einem Wald eine Kurve.
	`del`:	<u>Geländemotorradfahrer</u> macht in einem Herbst während Zuschauer eine Kurve.
	`del+obj`:	<u>Geländemotorradfahrer</u> macht in einem Herbst eine Kurve.
	DE:	Ein <u>offroad-biker</u> fährt im Herbst durch eine steile Kurve.

(c) Example of a blank resolved by the textual context for PERS. *biker* correctly translated into the Masc. form *Geländemotorradfahrer* (dirt biker) by `base`. **Objects**: person, tree, bike, helmet

Figure 4: Examples of resolved blanks for test set 2016. Underlined text denotes blanked words and their translations. Object field indicates the detected objects.

to better results than the previous state of the art. Adding visual information, and in particular structural representations of this information, proved beneficial when input text contains noise and the language pair requires substantial restructuring from source to target. Our findings suggest that the combination of a deliberation approach and information from additional modalities is a promising direction for machine translation that is robust to noisy input. Our code and pre-processing scripts are available at `https://github.com/ImperialNLP/MMT-Delib`.

Acknowledgments

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of International Conference on Learning Representations (ICLR)*.

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: Shared Task Papers*, pages 304–323. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214. 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. 2017. LIUM-CVC submissions for WMT17 multimodal translation task. In *Proceedings of the Second Conference on Machine Translation*, pages 432–439. Association for Computational Linguistics.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 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.

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. Association for Computational Linguistics.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924. Association for Computational Linguistics.

Rajen Chatterjee, Matteo Negri, Raphael Rubino, and Marco Turchi. 2018. Findings of the WMT 2018 shared task on automatic post-editing. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 723–738, Belgium, Brussels. Association for Computational Linguistics.

Jean-Benoit Delbrouck and Stéphane Dupont. 2017. An empirical study on the effectiveness of images in multimodal neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 910–919. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

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, 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 *5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany.

Desmond Elliott and Àkos Kádár. 2017. Imagination improves multimodal translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 130–141. Asian Federation of Natural Language Processing.

Joel Escudé Font and Marta R. Costa-Jussà. 2019. Equalizing gender biases in neural machine translation with word embeddings techniques. *CoRR*, abs/1901.03116.

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):393413.

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: Shared Task Papers*, pages 603–611. Association for Computational Linguistics.

Christian Hardmeier, Joakim Nivre, and Jörg Tiedemann. 2012. Document-wide decoding for phrase-based statistical machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1179–1190. Association for Computational Linguistics.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic Chinese to English news translation. *CoRR*, abs/1803.05567.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 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: Shared Task Papers*, pages 616–623. Association for Computational Linguistics.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Computer Vision – ECCV 2018*, pages 793–811. Springer International Publishing.

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, Berlin, Germany. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

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. Association for Computational Linguistics.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2017. An exploration of neural sequence-to-sequence architectures for automatic post-editing. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 120–129. Asian Federation of Natural Language Processing.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Filip Klubička, Antonio Toral, and Víctor M Sánchez-Cartagena. 2017. Fine-grained human evaluation of neural versus phrase-based machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108(1):121–132.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. 2018. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*.

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.

Jason Lee, Elman Mansimov, and Kyunghyun Cho. 2018. Deterministic non-autoregressive neural sequence modeling by iterative refinement. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182. 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.

Pranava 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*, pages 470–476. Association for Computational Linguistics.

Pranava Swaroop Madhyastha, Josiah Wang, and Lucia Specia. 2018. End-to-end image captioning exploits multimodal distributional similarity. *CoRR*, abs/1809.04144.

Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Xufeng Han, Alyssa Mensch, Alex Berg, Tamara Berg, and Hal Daume III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756, Avignon, France. Association for Computational Linguistics.

Jan Niehues, Eunah Cho, Thanh-Le Ha, and Alex Waibel. 2016. Pre-translation for neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1828–1836. The COLING 2016 Organizing Committee.

Roman Novak, Michael Auli, and David Grangier. 2016. Iterative refinement for machine translation. *CoRR*, abs/1610.06602.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. volume 123, pages 74–93, Hingham, MA, USA. Kluwer Academic Publishers.

Kashif Shah, Josiah Wang, and Lucia Specia. 2016. SHEF-Multimodal: Grounding machine translation on images. In *First Conference on Machine Translation, Volume 2: Shared Task Papers*, WMT, pages 657–662. 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. Association for Computational Linguistics.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference*

on Empirical Methods in Natural Language Processing, pages 3003–3008. Association for Computational Linguistics.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2Tensor for neural machine translation. pages 193–199.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Josiah Wang, Pranava Swaroop Madhyastha, and Lucia Specia. 2018. Object counts! Bringing explicit detections back into image captioning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana. Association for Computational Linguistics.

Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. 2017. Deliberation networks: Sequence generation beyond one-pass decoding. In *Advances in Neural Information Processing Systems 30*, pages 1784–1794. Curran Associates, Inc.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057. PMLR.

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. 2:67–78.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20. Association for Computational Linguistics.

A   Appendices

EN: A <u>bride</u> and <u>groom</u> kiss under the <u>bride</u>'s veil.

base: Ein <u>Mann</u> und eine <u>Frau</u> küssen sich unter den Blicken der <u>Frau</u>.

del: Ein <u>Mann</u> und eine <u>Frau</u> küssen sich unter dem <u>Brautschleier</u>.

del+obj: Ein <u>Mann</u> und eine <u>Frau</u> küssen sich unter den <u>hin</u>.

DE: Eine <u>Braut</u> und <u>Bräutigam</u> küssen sich unter dem <u>Brautschleier</u> .

(a) PERS example: *bride* and *groom* translated are correctly translated by base into *Frau* (wife) and *Mann* (husband). **Objects**: face, woman, dress

EN: A brown dog <u>runs</u> down the sandy beach.

base: Ein brauner Hund <u>läuft</u> an einem sandigen Strand.

del: Ein brauner Hund <u>rennt</u> den Sandstrand hinunter.

del+obj: Ein brauner Hund <u>läuft</u> an einem sandigen Strand hinunter.

FR: Ein brauner Hund <u>läuft</u> über den Sandstrand.

(b) AMB example: *runs* is correctly translated by base into *läuft*. **Objects**: dog

Figure 5: Examples of blanks for test set 2016 that were correctly resolved by the textual context. The underlined words denote blanked words and their translations.
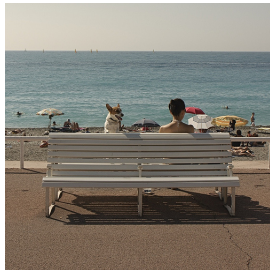
EN: A woman and a dog sit on a white <u>bench</u> near a beach.

base: Eine Frau und ein Hund sitzen an einem weißen <u>Strand</u> nahe einem Strand.

del: Eine Frau und ein Hund sitzen auf einem weißen <u>Sofa</u> in der nähe eines Strands.

del+obj: Eine Frau und ein Hund sitzen auf einer weißen <u>Bank</u> nahe einem Strand.

DE: Eine Frau und eine Hund sitzen auf einer weißen <u>Bank</u> in der nähe eines Strandes.

(a) RND example: the blank *bench* is correctly translated by del+obj into *Bank* due to the detected object Bench. **Objects**: person, dog, bench

EN: Two <u>men</u> dressed in green are preparing food in a restaurant.

base: Deux <u>femmes</u> vêtues de vert préparent des aliments dans un restaurant.

del: Deux <u>femmes</u> vêtues de vert préparent de la nourriture dans un restaurant.

del+obj: Deux <u>asiatiques</u> en vert préparent de la nourriture dans un restaurant.

FR: Deux <u>hommes</u> habillés en vert préparent de la nourriture dans un restaurant.

(b) PERS example. *men* correctly translated into *asiatiques* (asians) by del+obj. **Objects**: person, clothing, man, food, cake

Figure 6: Examples of blanks for test set 2016 that were correctly resolved by the multimodal context. The underlined words denote blanked words and their translations.

EN: A guy give a kiss to a guy also.

base: Ein <u>Mann</u>, der sich vor, um eine <u>Frau</u> zu knüssen .

del: Ein <u>Mann</u>, der sich vor, um eine <u>Frau</u> zu küssen.

del+obj: Ein <u>Mann</u>, der einem kuss küsst, um eine <u>Frau</u> zu küssen.

DE: Ein <u>Typ</u> küsst einen anderen <u>Typ</u> .

(a) PERS example: the second mention of *guy* is consistently translated into *Frau* (woman). **Objects**: clothing, man, face

EN: A group of students sit and <u>listen</u> to the <u>speaker</u>.

base: Eine Gruppe von Studenten sitzt und <u>schaut</u> nach <u>rechts</u> .

del: Eine Gruppe Schüler sitzt und <u>schaut</u> nach <u>rechts</u>.

del+obj: Eine Gruppe Schüler sitzt und <u>schaut</u> zu rechts auf das <u>Wasser</u>.

DE: Eine Gruppe von Studenten sitzt und <u>hört</u> der <u>Sprecherin</u> <u>zu</u>.

(b) AMB example. The blanks *listen* and *speaker* are consistently translated into *schaut* (look) and *rechts* (right) or *Wasser* (water). **Objects**: person, clothing, man, food, cake

Figure 7: Examples of unresolved blanks. The underlined words denote blanked words and their translations.