Abstract

In this work, we propose to model the interaction between visual and textual features for multi-modal neural machine translation through a latent variable model. This latent variable can be seen as a stochastic embedding and it is used in the target-language decoder and also to predict image features. Importantly, even though in our model formulation we capture correlations between visual and textual features, we do not require that images be available at test time. We show that our latent variable MMT formulation improves considerably over strong baselines, including the multi-task learning approach of Elliott and Kádár (2017) and the conditional variational auto-encoder approach of Toyama et al. (2016). Finally, in an ablation study we show that (i) predicting image features in addition to only conditioning on them and (ii) imposing a constraint on the minimum amount of information encoded in the latent variable slightly improved translations.

1 Introduction

Multi-modal machine translation (MMT) is an exciting novel take on machine translation (MT) where we are interested in learning to translate sentences in the presence of visual input (mostly images). In the last two years there have been two shared tasks (Specia et al., 2016; Elliott et al., 2017) where many research groups proposed different techniques to integrate images into MT, e.g. Caglayan et al. (2017); Libovický and Helcl (2017).

Most MMT models expand neural machine translation (NMT) architectures (Sutskever et al., 2014; Bahdanau et al., 2015) to additionally condition on an image in order to compute the likelihood of a translation in context. This gives the model a chance to exploit correlations in visual and language data, but also means that images must be available at test time. An exception to this rule is the work of Toyama et al. (2016) who exploit the framework of conditional variational auto-encoders (CVAEs) (Sohn et al., 2015) to decouple the encoder used for posterior inference at training time from the encoder used for generation at test time. Rather than conditioning on image features, the model of Elliott and Kádár (2017) exploits the potential in predicting such features from language data in a multi-task learning (MTL) framework. This similarly exploits correlations between the two modalities and has the advantage that images are also not necessary at test time.

In this work, we also aim at translating without images at test time, yet learning a visually grounded translation model. To that end, we resort to probabilistic modelling instead of multi-task learning and estimate a joint distribution over translations and images. In a nutshell, we propose to model the interaction between visual and textual features through a latent variable. This latent variable can be seen as a stochastic embedding which is used in the target-language decoder, as well as to predict image features. Our experiments show that this joint formulation improves over an MTL approach (Elliott and Kádár, 2017), which does model both modalities but not jointly, and over the CVAE of Toyama et al. (2016), which uses image features to condition an inference network but crucially does not model the images.

The main contributions of this paper are:

- we propose a novel multi-modal NMT model that incorporates image features through latent variables in a deep generative model.
- our latent variable MMT formulation improves considerably over strong baselines, and compares favourably to the state-of-the-art.
we exploit correlations between both modalities at training time through a joint generative approach and do not require images at prediction time.

The remainder of this paper is organised as follows. In §2, we introduce the data sets we used. In §3, we describe our different variational MMT models. In §4, we report experiments and assess how our models compare to prior work. In §5, we position our approach with respect to the literature. Finally, in §6 we draw conclusions and provide avenues for future work.

2 Dataset

The Flickr30k dataset (Young et al., 2014) consists of images from Flickr and their English descriptions. We use the Multi30k dataset (Elliott et al., 2016), which consists of an expansion of Flickr30k where for each image, one of its English descriptions was translated into German by a professional translator. Training, validation and test sets contain 29k, 1014 and 1k images respectively, each accompanied by the original English sentence and its translation into German.

In order to obtain features for images, we use ResNet-50 (He et al., 2015) pre-trained on the ImageNet dataset (Russakovsky et al., 2015). We extract pool15 features, i.e. 2048-dimensional pre-activations of the last layer of the network, and report experiments using these as our image features.

3 Variational Multi-modal NMT

Similarly to standard NMT, in MMT we wish to translate a source sequence \( x_1^m \triangleq (x_1, \cdots, x_m) \) into a target sequence \( y_1^n \triangleq (y_1, \cdots, y_n) \). The main difference is the presence of an image \( v \) which illustrates the sentence pair \( (x_1^m, y_1^n) \). We do not model images directly, but instead an \( o \)-dimensional vector of pre-activations of a ResNet-50’s pool15 layer.

In our variational MMT models, image features are assumed to be generated by transforming a stochastic latent embedding \( z \), which is also used to inform the RNN decoder in translating source sentences into a target language.

3.1 Variational MMT Models

In this section we present our generative model as well as efficient parameter estimation.

Generative model

We propose a generative model of translation and image generation where both the image \( v \) and the target sentence \( y_1^n \) are independently generated given a common stochastic embedding \( z \). The generative story is as follows. We observe a source sentence \( x_1^m \) and draw an embedding \( z \) from a latent Gaussian model,

\[
Z | x_1^m \sim \mathcal{N}(\mu, \text{diag}(\sigma^2))
\]

\[
\mu = f_{\mu}(x_1^m; \theta)
\]

\[
\sigma = f_{\sigma}(x_1^m; \theta),
\]

where \( f_{\mu}(\cdot) \) and \( f_{\sigma}(\cdot) \) map from a source sentence to a vector of locations \( \mu \in \mathbb{R}^c \) and a vector of scales \( \sigma \in \mathbb{R}^{c \times c} \), respectively. We then proceed to draw the image features from a Gaussian observation model,

\[
V | z \sim \mathcal{N}(\nu, \varsigma^2 I)
\]

\[
\nu = f_{\nu}(z; \theta),
\]

where \( f_{\nu}(\cdot) \) maps from \( z \) to a vector of locations \( \nu \in \mathbb{R}^c \), and \( \varsigma \in \mathbb{R}_{>0} \) is a hyperparameter of the model (we use \( 1 \)). Conditioned on \( z \) and on the source sentence \( x_1^m \), and independently of \( v \), we generate a translation by drawing each target word in context from a Categorical observation model,

\[
Y_j | x_1^m, z, y_{<j} \sim \text{Cat}(\pi_j)
\]

\[
\pi_j = f_{\pi}(x_1^m, y_{<j}, z; \theta),
\]

where \( f_{\pi}(\cdot) \) maps \( z, x_1^m \), and a prefix translation \( y_{<j} \) to the parameters \( \pi_j \) of a categorical distribution over the target vocabulary. Functions \( f_{\mu}(\cdot), f_{\nu}(\cdot), f_{\sigma}(\cdot), \) and \( f_{\pi}(\cdot) \) are implemented as neural networks whose parameters are collectively denoted by \( \theta \). In particular, implementing \( f_{\pi}(\cdot) \) is as simple as augmenting a standard NMT architecture (Bahdanau et al., 2015; Luong et al., 2015), i.e. encoder-decoder with attention, with an additional input \( z \) available at every time-step. All other functions are single-layer MLPs that transform the average encoder hidden state to the dimensionality of the corresponding Gaussian variable followed by an appropriate activation.\(^1\)

Note that in effect we model a joint distribution

\[
p_{\theta}(y_1^n, v, z | x_1^m) = p_{\theta}(z | x_1^m)p_{\theta}(v | z)P_{\theta}(y_1^n | x_1^m, z)
\]

\(^1\)Locations have support on the entire real space, thus we use linear activations, scales must be strictly positive, thus we use a softplus activation.
This is in direct contrast with Rezende et al., (right) is a graphical depiction of the inference, (left) is a graphical depiction of the generative model of target text and image features (left), and inference model (right).

Figure 1: Generative model of target text and image features (left), and inference model (right).

consisting of two components which we parameterise directly. As there are no observations for $z$, we cannot estimate these components directly. We must instead marginalise $z$ out, which yields the marginal

$$P_\theta(y^n_1, v|x^n_1) = \int p_\theta(z|x^n_1)p_\theta(v|z)p_\theta(y^n_1|x^n_1, z)dz.$$ (5)

An important statistical consideration about this model is that even though $y^n_1$ and $v$ are conditionally independent given $z$, they are marginally dependent. This means that we have designed a data generating process where our observations $y^n_1, v|x^n_1$ are not assumed to have been independently produced. This is in direct contrast with multi-task learning or joint modelling without latent variables—for an extended discussion see (Eikema and Aziz, 2018, § 3).

Finally, Figure 1 (left) is a graphical depiction of the generative model: shaded circles denote observed random variables, unshaded circles indicate latent random variables, deterministic quantities are not circled; the internal plate indicates iteration over time-steps, the external plate indicates iteration over the training data. Note that deterministic parameters $\theta$ are global to all training instances, while stochastic embeddings $z$ are local to each tuple $(x^n_1, y^n_1, v)$.

**Inference** Parameter estimation for our model is challenging due to the intractability of the marginal likelihood function (5). We can however employ variational inference (VI) (Jordan et al., 1999), in particular amortised VI (Kingma and Welling, 2014; Rezende et al., 2014), and estimate parameters to maximise a lowerbound

$$\mathbb{E}_{q_\lambda(z|x^n_1, y^n_1, v)}[\log p_\theta(v)|z] + \log p_\theta(y^n_1|x^n_1, z)] - \text{KL}(q_\lambda(z|x^n_1, y^n_1, v)||p(z|x^n_1)).$$ (6)

on the log-likelihood function. This evidence lowerbound (ELBO) is expressed in terms of an inference model $q_\lambda(z|x^n_1, y^n_1, v)$ which we design having tractability in mind. In particular, our approximate posterior is a Gaussian distribution

$$q_\lambda(z|x^n_1, y^n_1, v) = \mathcal{N}(z|\mu, \text{diag}(\sigma^2))$$

$$\mu = g_\theta(x^n_1, y^n_1, v, \lambda)$$

$$\sigma = g_s(x^n_1, y^n_1, v, \lambda)$$ (7)

parametrised by an inference network, that is, an independently parameterised neural network (whose parameters we denote collectively by $\lambda$) which maps from observations, in our case a sentence pair and an image, to a variational location $\mu \in \mathbb{R}^c$ and a variational scale $\sigma \in \mathbb{R}^{c \times 0}$. Figure 1 (right) is a graphical depiction of the inference model.

Location-scale variables (e.g. Gaussians) can be reparametrised, i.e. we can obtain a latent sample via a deterministic transformation of the variational parameters and a sample from the standard Gaussian distribution:

$$z = \mu + \epsilon \odot \sigma \text{ where } \epsilon \sim \mathcal{N}(0, I).$$ (8)

This reparametrisation enables backpropagation through stochastic units (Kingma and Welling,
We use feed-forward neural networks with a single ReLU hidden layer and a linear output. Scales are strictly positive, thus \( f_a \) and \( g_a \) are designed similarly, but with softplus output. For the generative model, \( f_v \) and \( f_p \) transform the average source-language encoder hidden state. The inference model conditions additionally on the complete target sentence, and has therefore a target-language bidirectional LSTM encoder. Then \( g_u \) and \( g_s \) transform a concatenation of the average source-language encoder hidden state, the average target-language bidirectional encoder hidden state, and the image features.

**Fixed Gaussian prior** We have just presented our variational MMT model in its full generality— we refer to that model as VMMT\(_C\). However, keeping in mind that MMT datasets are rather small, it is desirable to simplify some of our model’s components. In particular, the estimated latent Gaussian model (1) can be replaced by a fixed standard Gaussian prior, i.e., \( Z \sim \mathcal{N}(0, I) \)— we refer to this model as VMMT\(_F\). Along with this change it is convenient to modify the inference model to condition on \( x_1^m \) alone, which allow us to use the inference model for both training and prediction. Importantly this also sidesteps the need for a target-language bidirectional LSTM encoder, which leaves us a smaller set of inference parameters \( \lambda \) to estimate. Interestingly, this model does not rely on features from \( v \), instead only using it as learning signal through the objective in (6), which is in direct contrast with the model of Toyama et al. (2016).

## 4 Experiments

### 4.1 Experimental settings

Our encoder is a 2-layer 500D bidirectional RNN with GRU, the latent embedding \( z \), source, and target word embeddings are also 500D each, and trained jointly with the model. We use OpenNMT to implement all our models (Klein et al., 2017). All model parameters are initialised sampling from a uniform distribution \( \mathcal{U}(-0.1, +0.1) \) and bias vectors are initialised to 0.

Visual features are obtained by feeding images to the pre-trained ResNet-50 and using the activations of the pool15 layer (He et al., 2015). We apply dropout with a probability of 0.5 in the encoder bidirectional RNN, the image features, the decoder RNN, and before emitting a target word.

All models are trained using the Adam optimiser (Kingma and Ba, 2014) with an initial learning rate of 0.002 and minibatches of size 40, where each training instance consists of one English sentence, one German sentence and one image (MMT). Models are trained for up to 40 epochs and we perform model selection based on BLEU4, and use the best performing model on the validation set to translate test data. Moreover, we halt training if the model does not improve BLEU4 scores on the validation set for 10 epochs or more. We report mean and standard deviation over 4 independent runs for all models we trained ourselves (NMT, VMMT\(_C\), VMMT\(_F\)), and other baseline results are the ones reported in the authors’ publications (Toyama et al., 2016; Elliott and Kádár, 2017).

We preprocess our data by tokenizing, lowercasing, and converting words to subword tokens using a bilingual BPE model with 10k merge operations (Sennrich et al., 2016). We quantitatively evaluate translation quality using case-insensitive and tokenized outputs in terms of BLEU4 (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), chrF3 (Popović, 2015), and BEER (Stanojević and Sima’an, 2014). By using these, we hope to include word-level metrics which are traditionally used by the MT community (i.e. BLEU and METEOR), as well as more recent metrics which operate at the character level and that better correlate with human judgements of translation quality (i.e. chrF3 and BEER) (Bojar et al., 2017).
We compare our work against three different baselines. The first one is a standard text-only sequence-to-sequence NMT model with attention (Luong et al., 2015), trained from scratch using hyperparameters described in Section 4.1. The second baseline is the variational multi-modal MT model Model G proposed by Toyama et al. (2016), where global image features are used as additional input to condition an inference network. Finally, a third baseline is the Imagination model of Elliott and Kádár (2017), a multi-task MMT model which uses a shared source-language encoder RNN and is trained in two tasks: to translate from English into German and on image-sentence ranking (English→image).

### 4.3 Results

We now report on experiments conducted with models trained on the Multi30k data set to translate from English into German.

In Table 1, we compare our variational MMT models—VMMT_{C} for the general case with a conditional Gaussian latent model, and VMMT_{F} for the simpler case of a fixed Gaussian prior—to the three baselines described above. The general trend is that both formulations of our VMMT improve with respect to all three baselines. We note an improvement in BLEU and METEOR mean scores compared to the Imagination model (Elliott and Kádár, 2017), as well as a much smaller variance (our results are based on 4 independent runs, whereas Imagination’s are based on 3 runs). Both models VMMT_{F} and VMMT_{C} outperform Model G according to BLEU and perform comparably according to METEOR, especially since results reported by (Toyama et al., 2016) are based on a single run.

Moreover, we also note that both our models outperform the text-only NMT baseline according to all four metrics, and by 1%–1.4% according chrF3 and BEER, both being metrics well-suited to measure the quality of translations into German and generated with subwords units.

Finally, one interesting finding is that the fixed-prior model VMMT_{F} performs slightly better than the conditional model VMMT_{C} according to all four metrics studied. We speculate this is due to VMMT_{F}’s simpler parameterisation, after all, we have just about 29k training instances to estimate two sets of parameters (θ and λ) and the more complex VMMT_{C} requires an additional bidirectional LSTM encoder for the target text.

### 4.4 Ablative experiments

In our ablation studies we are interested in finding out how important the prediction of the image features are to the overall model, and also if and how the KL term in the loss affects translations.

**No image log-likelihood** In these experiments, we wish to explore whether gains observed in Table 1 derive from the fact there is a latent variable $z$ used to augment the decoder, i.e. a source of stochasticity that would make the decoder more robust to noise, or from the fact that this latent variable $z$ is used to predict image features $v$.

We remove the image log-likelihood from the loss function used at training, effectively making our models latent-variable neural machine translation models, but no longer multi-modal (similarly to the probabilistic formulation in Toyama et al. (2016)). Concretely, instead of the loss described...
in Equation (6), the new loss function is given by:

\[
\log P_\theta(y^n_1|x^n_1) \geq \mathbb{E}_{q_\lambda(z|x^n_1)} \left[ \log P_\theta(y^n_1|x^n_1, z) \right] - \text{KL}(q_\lambda(z|x^n_1)||p(z)).
\]

(9)

From Table 2 below, we notice that using the image log-likelihood in the loss slightly improves mean BLEU scores in model VMMT_F, while not affecting VMMT_C. Nevertheless, the variance of the multiple runs of both models increased significantly (i.e. more than doubled) by removing the image log-likelihood from the loss, suggesting that it might function as a regulariser of the model.

| Model  | Number of free bits (KL) | BLEU4↑ |
|--------|-------------------------|--------|
| VMMT_C | 0                       | 38.5 ± 0.2 |
|        | 1                       | 38.3 ± 0.3 |
|        | 2                       | 38.2 ± 0.2 |
|        | 4                       | 36.8 ± 2.6 |
|        | 8                       | 38.6 ± 0.2 |
| VMMT_F | 0                       | 38.3 ± 0.2 |
|        | 1                       | 38.1 ± 0.3 |
|        | 2                       | 38.4 ± 0.4 |
|        | 4                       | 38.4 ± 0.4 |
|        | 8                       | 35.7 ± 3.1 |

Table 2: Results of applying VMMT models trained with and without an image log-likelihood as part of the loss to translate the Multi30k development set.

| Model  | Image log-likelihood? | BLEU4↑ |
|--------|-----------------------|--------|
| VMMT_C | —                     | 38.6 ± 0.4 |
|        | y                     | 38.6 ± 0.2 |
| VMMT_F | —                     | 38.0 ± 0.7 |
|        | y                     | 38.3 ± 0.2 |

Table 3: Results of applying VMMT models trained with different numbers of free bits in the KL (Kingma et al., 2016) to translate the Multi30k development set.

5 Related work

Multi-modal machine translation was only recently addressed by the MT community by means of a shared task (Specia et al., 2016; Elliott et al., 2017). Nevertheless, recently there has been research involving both variational and deterministic multi-modal machine translation models.

Fully supervised MMT models. All submissions to the two runs of the multi-modal MT shared tasks (Specia et al., 2016; Elliott et al., 2017) are fully supervised models, i.e. they either directly model the conditional probability of a translation given source sentence and images \( P(y^n_1|x^n_1, v) \), or given the source sentence only \( P(y^n_1|x^n_1) \).

Perhaps the first MMT model proposed prior to these shared tasks is that of Hitschler et al. (2016), who used image features to re-rank translations of image descriptions generated by a phrase-based statistical MT model (PBSMT) and reported sig-
significant improvements. Shah et al. (2016) propose a similar model where image logits are used to re-rank the output of PBMSMT. Global image features, i.e., features computed over an entire image (such as pool5 ResNet-50 features used in this work), have been directly used as “words” in the source sentence, to initialise encoder RNN hidden states, or as additional information used to initialise the decoder RNN states (Huang et al., 2016; Libovický et al., 2016; Calixto and Liu, 2017). On the other hand, spatial visual features, i.e., local features that encode different parts of the image separately in different vectors, have been used in doubly-attentive models where there is one attention mechanism over the source RNN hidden states and another one over the image features (Caglayan et al., 2016; Calixto et al., 2017).

Finally, Caglayan et al. (2017) proposed to interact image features with target word embeddings, more specifically to perform an element-wise multiplication of the (projected) global image features and the target word embeddings before feeding the target word embeddings into their decoder GRU. They reported significant improvements by using image features to gate target word embeddings and won the 2017 Multi-modal MT shared task (Elliott et al., 2017).

**Multi-task MMT models.** Multi-task learning MMT models are easily applicable to translate sentences without images (at test time), which is an advantage over the above-mentioned deterministic models.

Luong et al. (2016) proposed a multi-task approach where a model is trained using two tasks and a shared decoder: the main task is to translate from German into English and the secondary task is to generate English descriptions given an image. They show improvements in the main translation task when also training for the secondary image description task. Their model is large, i.e., a 4-layer encoder LSTM and a 4-layer decoder LSTM, and their best setup uses a ratio of 0.05 image description generation training data samples in comparison to translation training data samples. Elliott and Kádár (2017) propose an MTL model trained to do translation (English→German) and sentence-image ranking (English++image), using a standard word cross-entropy and margin-based losses as its task objectives, respectively. Their model uses the pre-trained GoogleNet v3 CNN (Szegedy et al., 2016) to extract pool5 features, and has a 1-layer source-language bidirectional GRU encoder and a 1-layer GRU decoder.

**Variational MMT models.** Toyama et al. (2016) proposed a variational MMT model that is likely the most similar model to the one we put forward in this work. They build on the variational neural MT (VNMT) model of Zhang et al. (2016), which is a conditional latent model where a Gaussian-distributed prior of $z$ is parameterised as a function of the the source sentence $x^m_1$, i.e., $p(z|x^m_1)$, and both $x^m_1$ and $z$ are used at each time step in an attentive decoder RNN, $P(y_j|x^m_1, z, y_{<j})$.

In their model, image features are used as input to the inference model $q_\lambda(z|x^m_1, y^m_1, v)$ that approximates the posterior over the latent variable, but otherwise are not modelled and not used in the generative network. Differently from their work, we use the image features in all generative models we propose modelling them as random outcomes and predicting them directly while still being able to use our model to translate without images at test time. In the conditional case, we further use image features for posterior inference. Additionally, we also investigate both conditional and fixed priors, i.e., $p(z|x^m_1)$ and $p(z)$, respectively, whereas their model is always conditional. Interestingly, we found in our experiments that fixed-prior models perform slightly better than conditional ones. Toyama et al. (2016) uses the pre-trained VGG19 CNN (Simonyan and Zisserman, 2015) to extract FC7 features, and additionally experiment with using additional features from object detections obtained with the Fast RCNN network (Girshick, 2015). One more difference between their work and ours is that we only use the ResNet-50 network to extract pool5 features, and no additional pre-trained CNN or object detections.

6 Conclusions and Future work

In this paper, we have proposed a latent variable model for multi-modal neural machine translation. We show that predicting image features is important and leads to improvements compared to using these same features just to condition the model. An ablative study shows that fine-tuning hyperparameters such as the number of free bits in the KL divergence is also important, which makes navigating the space of and finding the best hyperparame-
This is a work in progress, and in future work we will explore other generative models for multimodal MT, as well as different ways to incorporate images into these models.

Acknowledgements

This work is supported by the Dutch Organisation for Scientific Research (NWO) VICI Grant nr. 277-89-002.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations, ICLR 2015*, San Diego, California.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the wmt17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. 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, Copenhagen, Denmark. Association for Computational Linguistics.

Ozan Caglayan, Loïc Barrault, and Fethi Bougares. 2016. Multimodal attention for neural machine translation. *CoRR*, abs/1609.03976.

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.

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, Vancouver, Canada. 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*.

Bryan Eikema and Wilker Aziz. 2018. Auto-encoding variational neural machine translation. *arXiv preprint arXiv:1807.10564*.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation. In *Proceedings of the Second Conference on Machine Translation*, 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, VL@ACL 2016*, Berlin, Germany.

Desmond Elliott and Ákos 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, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Ross Girshick. 2015. Fast R-CNN. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV ’15, pages 1440–1448, Washington, DC, USA. IEEE Computer Society.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385*.

Julian Hitschler, Shigehiko Schamon, and Stefan Riezler. 2016. *MachineTrialsPatriots* 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.

Po-Yao Huang, Frederick Liu, Sz-Rung Shi-ang, 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.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. 1999. An introduction to variational methods for graphical models. Machine Learning, 37(2):183–233.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. CoRR, abs/1412.6980.

Diederik P. Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. 2016. Improved variational inference with inverse autoregressive flow. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, Advances in Neural Information Processing Systems 29, pages 4743–4751. Curran Associates, Inc.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In International Conference on Learning Representations.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In Proc. ACL.

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.

Jindřich Libovický, Jindřich Helcl, Marek Tlustý, Ondřej Bojar, and Pavel Pecina. 2016. CUNI System for WMT16 Automatic Post-Editing and Multimodal Translation Tasks. In Proceedings of the First Conference on Machine Translation, pages 646–654, Berlin, Germany.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Łukasz Kaiser. 2015. Multi-Task Sequence to Sequence Learning. In Proceedings of the International Conference on Learning Representations (ICLR), 2016, San Juan, Puerto Rico.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1412–1421, Lisbon, Portugal.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, pages 311–318, Philadelphia, Pennsylvania.

Daniele Picozzi and Marco Kiefer. 2014. chrf: character n-gram f-score for automatic mt evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014, pages 1278–1286.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), 115(3):211–252.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725, Berlin, Germany.
Kashif Shah, Josiah Wang, and Lucia Specia. 2016. SHEF-Multimodal: Grounding Machine Translation on Images. In Proceedings of the First Conference on Machine Translation, pages 660–665, Berlin, Germany.

K. Simonyan and A. Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, Advances in Neural Information Processing Systems 28, pages 3483–3491. Curran Associates, Inc.

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, WMT 2016, colocated with ACL 2016, pages 543–553, Berlin, Germany.

Miloš Stanojević and Khalil Sima’an. 2014. Fitting sentence level translation evaluation with many dense features. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 202–206, Doha, Qatar. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. V. Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, Advances in Neural Information Processing Systems 27, pages 3104–3112. Curran Associates, Inc.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2016. Rethinking the inception architecture for computer vision. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2818–2826.