

# Latent Variable Model for Multi-modal Translation

Iacer Calixto

Miguel Rios

Wilker Aziz

ILLC

The University of Amsterdam

{iacer.calixto,m.riosgaona,w.aziz}@uva.nl

## Abstract

In this work, we propose to model the interaction between visual and textual features for multi-modal neural machine translation (MMT) through a latent variable model. This latent variable can be seen as a multi-modal stochastic embedding of an image and its description in a foreign language. It is used in a target-language decoder and also to predict image features. Importantly, our model formulation utilises visual and textual inputs during training but does not require that images be available at test time. We show that our latent variable MMT formulation improves considerably over strong baselines, including a multi-task learning approach (Elliott and Kádár, 2017) and a conditional variational auto-encoder approach (Toyama et al., 2016). Finally, we show improvements due to (i) predicting image features in addition to only conditioning on them, (ii) imposing a constraint on the KL term to promote models with non-negligible mutual information between inputs and latent variable, and (iii) by training on additional target-language image descriptions (i.e. synthetic data).

## 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 three years there have been shared tasks (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018) 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) learns to rank image features using language data in a multi-task learning (MTL) framework, therefore sharing parameters between a translation (generative) and a sentence-image ranking model (discriminative). 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:<sup>1</sup>

- we propose a novel multi-modal NMT model

<sup>1</sup>Code and pre-trained models available in [https://github.com/iacercalixto/variational\\_mmt](https://github.com/iacercalixto/variational_mmt).

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 describe our variational MMT models. In §3, we introduce the data sets we used and report experiments and assess how our models compare to prior work. In §4, we position our approach with respect to the literature. Finally, in §5 we draw conclusions and provide avenues for future work.

## 2 Variational Multi-modal NMT

Similarly to standard NMT, in MMT we wish to translate a source sequence  $x_1^m \triangleq \langle x_1, \dots, x_m \rangle$  into a target sequence  $y_1^n \triangleq \langle y_1, \dots, y_n \rangle$ . The main difference is the presence of an image  $v$  which illustrates the sentence pair  $\langle x_1^m, y_1^n \rangle$ . We do not model images directly, but instead an 2048-dimensional vector of pre-activations of a ResNet-50’s pool5 layer (He et al., 2015).

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.

**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,

$$\begin{aligned} Z|x_1^m &\sim \mathcal{N}(\boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma}^2)) \\ \boldsymbol{\mu} &= f_\mu(x_1^m; \theta) \\ \boldsymbol{\sigma} &= f_\sigma(x_1^m; \theta), \end{aligned} \quad (1)$$

where  $f_\mu(\cdot)$  and  $f_\sigma(\cdot)$  map from a source sentence to a vector of locations  $\boldsymbol{\mu} \in \mathbb{R}^c$  and a vector of scales  $\boldsymbol{\sigma} \in \mathbb{R}_{>0}^c$ , respectively. We then proceed to

draw the image features from a Gaussian observation model,

$$\begin{aligned} V|z &\sim \mathcal{N}(\boldsymbol{\nu}, \varsigma^2 I) \\ \boldsymbol{\nu} &= f_\nu(z; \theta), \end{aligned} \quad (2)$$

where  $f_\nu(\cdot)$  maps from  $z$  to a vector of locations  $\boldsymbol{\nu} \in \mathbb{R}^o$ , 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,

$$\begin{aligned} Y_j|x_1^m, z, y_{<j} &\sim \text{Cat}(\boldsymbol{\pi}_j) \\ \boldsymbol{\pi}_j &= f_\pi(x_1^m, y_{<j}, z; \theta), \end{aligned} \quad (3)$$

where  $f_\pi(\cdot)$  maps  $z$ ,  $x_1^m$ , and a prefix translation  $y_{<j}$  to the parameters  $\boldsymbol{\pi}_j$  of a categorical distribution over the target vocabulary. Functions  $f_\mu(\cdot)$ ,  $f_\sigma(\cdot)$ ,  $f_\nu(\cdot)$ , 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.<sup>2</sup>

Note that in effect we model a joint distribution

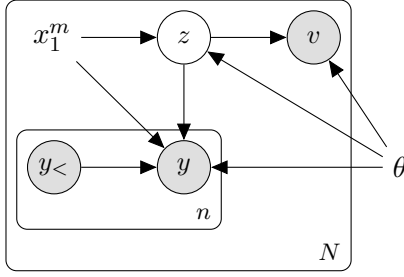
$$\begin{aligned} p_\theta(y_1^n, v, z|x_1^m) &= \\ p_\theta(z|x_1^m)p_\theta(v|z)P_\theta(y_1^n|x_1^m, z) \end{aligned} \quad (4)$$

consisting of three 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

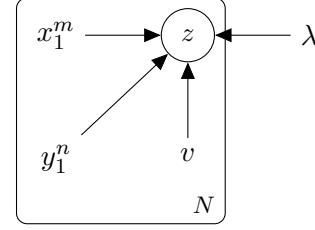
$$\begin{aligned} P_\theta(y_1^n, v|x_1^m) &= \\ \int p_\theta(z|x_1^m)p_\theta(v|z)P_\theta(y_1^n|x_1^m, z)dz. \end{aligned} \quad (5)$$

An important statistical consideration about this model is that even though  $y_1^n$  and  $v$  are conditionally independent given  $z$ , they are *marginally dependent*. This means that we have designed a data generating process where our observations

<sup>2</sup>Locations have support on the entire real space, thus we use linear activations, scales must be strictly positive, thus we use a softplus activation.



(a) VMMTC: given the source text  $x_1^m$ , we model the joint likelihood of the translation  $y_1^n$ , the image (features)  $v$ , and a stochastic embedding  $z$  sampled from a conditional latent Gaussian model. Note that the stochastic embedding is the sole responsible for assigning a probability to the observation  $v$ , and it helps assign a probability to the translation.



(b) Inference model for VMMTC: to approximate the true posterior we have access to both modalities (text  $x_1^m$ ,  $y_1^n$  and image  $v$ ).

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

$y_1^n, v | x_1^m$  are not assumed to have been independently produced.<sup>3</sup> This is in direct contrast with multi-task learning or joint modelling without latent variables—for an extended discussion see (Eikema and Aziz, 2019, § 3.1).

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  $\langle x_1^m, y_1^n, v \rangle$ .

**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

$$\mathbb{E}_{q_\lambda(z|x_1^m, y_1^n, v)} [\log p_\theta(v|z) + \log P_\theta(y_1^n|x_1^m, z)] - \text{KL}(q_\lambda(z|x_1^m, y_1^n, v) || p_\theta(z|x_1^m)) \quad (6)$$

on the log-likelihood function. This evidence lowerbound (ELBO) is expressed in terms of an *inference model*  $q_\lambda(z|x_1^m, y_1^n, v)$  which we design having tractability in mind. In particular, our *ap-*

*proximate posterior* is a Gaussian distribution

$$q_\lambda(z|x_1^m, y_1^n, v) = \mathcal{N}(z|\mathbf{u}, \text{diag}(\mathbf{s}^2))$$

$$\mathbf{u} = g_u(x_1^m, y_1^n, v; \lambda) \quad (7)$$

$$\mathbf{s} = g_s(x_1^m, y_1^n, v; \lambda)$$

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  $\mathbf{u} \in \mathbb{R}^c$  and a variational scale  $\mathbf{s} \in \mathbb{R}_{>0}^c$ . 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:

$$z = \mathbf{u} + \epsilon \odot \mathbf{s} \quad \text{where } \epsilon \sim \mathcal{N}(0, I) \quad (8)$$

This reparametrisation enables backpropagation through stochastic units (Kingma and Welling, 2014; Titsias and Lázaro-Gredilla, 2014). In addition, for two Gaussians the KL term in the ELBO (6) can be computed in closed form (Kingma and Welling, 2014, Appendix B). Altogether, we can obtain a reparameterised gradient estimate of the ELBO, we use a single sample estimate of the first term, and count on stochastic gradient descent to attain a local optimum of (6).

**Architecture** All of our parametric functions are neural network architectures. In particular,  $f_\pi$  is a standard sequence-to-sequence architecture with attention and a softmax output. We build upon OpenNMT (Klein et al., 2017), which we modify

<sup>3</sup>This is an aspect of the model we aim to explore more explicitly in the near future.

slightly by providing  $z$  as additional input to the target-language decoder at each time step. Location layers  $f_\mu$ ,  $f_\nu$  and  $g_u$ , and scale layers  $f_\sigma$  and  $g_s$ , are feed-forward networks with a single ReLU hidden layer. Furthermore, location layers have a linear output while scale layers have a softplus output. For the generative model,  $f_\mu$  and  $f_\sigma$  transform the average source-language encoder hidden state. We let the inference model condition on source-language encodings without updating them, and we use a target-language bidirectional LSTM encoder in order to also condition on the complete target sentence. 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  $\text{VMMT}_C$ . 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  $\text{VMMT}_F$ . 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).

### 3 Experiments

Our encoder is a 2-layer 500D bidirectional RNN with GRU, the source and target word embeddings are 500D, and all are 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  $\vec{0}$ .

Visual features are obtained by feeding images to the pre-trained ResNet-50 and using the activations of the `pool5` 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 ( $\text{NMT}$ ,  $\text{VMMT}_F$ ,  $\text{VMMT}_C$ ), 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, lower-casing, and converting words to subword tokens using a bilingual BPE model with 10k merge operations (Sennrich et al., 2016b). 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).

#### 3.1 Datasets

The Flickr30k dataset (Young et al., 2014) consists of images from Flickr and their English descriptions. We use the *translated Multi30k* ( $\text{M30k}_T$ ) dataset (Elliott et al., 2016), i.e. an extension 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 addition to the test set released for the first run of the multimodal translation shared task (Elliott et al., 2016), henceforth `test2016`, we also use `test2017` released for the next run of this shared task (Elliott et al., 2017).

Since this dataset is very small, we also investigate the effect of including more in-domain data to train our models. To that purpose, we use addi-

Model	BLEU4 $\uparrow$	METEOR $\uparrow$	chrF $\uparrow$	BEER $\uparrow$
NMT	35.0 (0.4)	54.9 (0.2)	<u>61.0 (0.2)</u>	<u>65.2 (0.1)</u>
Imagination	<u>36.8 (0.8)</u>	55.8 (0.4)	–	–
Model G	36.5	<b>56.0</b>	–	–
VMMT <sub>F</sub>	<b>37.7 (0.4)</b> $\uparrow 0.9$	<b>56.0 (0.1)</b> $\uparrow 0.0$	<b>62.1 (0.1)</b> $\uparrow 1.1$	<b>66.6 (0.1)</b> $\uparrow 1.4$
VMMT <sub>C</sub>	37.5 (0.3) $\uparrow 0.7$	55.7 (0.1) $\downarrow 0.3$	61.9 (0.1) $\uparrow 0.9$	66.5 (0.1) $\uparrow 1.3$

Table 1: Results of applying variational MMT models to translate the Multi30k 2016 test set. For each model, we report the mean and standard deviation over 4 independent runs where models were selected using validation BLEU4 scores. Best mean baseline scores per metric are underlined and best overall results (i.e. means) are in bold. We highlight in green/red the improvement brought by our models compared to the best baseline mean score.

tional 145K monolingual German descriptions released as part of the Multi30k dataset to the task of image description generation (Elliott et al., 2016). We refer to this dataset as *comparable Multi30k* (M30k<sub>C</sub>). Descriptions in the *comparable Multi30k* were collected independently of existing English descriptions and describe the same 29K images as in the M30k<sub>T</sub> dataset.

In order to obtain features for images, we use ResNet-50 (He et al., 2015) pre-trained on ImageNet (Russakovsky et al., 2015). We report experiments using  $p_{0015}$  features as our image features, i.e. 2048-dimensional pre-activations of the last layer of the network.

In order to investigate how well our models generalise, we also evaluate our models on the ambiguous MSCOCO test set (Elliott et al., 2017) which was designed with example sentences that are hard to translate without resorting to visual context available in the accompanying image.

Finally, we use a 50D latent embedding  $z$  in our experiments with the translated Multi30k data, whereas in our ablative experiments and experiments with the comparable Multi30k data, we use a 500D stochastic embedding  $z$ .

### 3.2 Baselines

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 above. 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 $\leftrightarrow$ image).

### 3.3 Translated Multi30k

We now report on experiments conducted with models trained to translate from English into German using the *translated Multi30k* data set (M30k<sub>T</sub>).

In Table 1, we compare our variational MMT models—VMMT<sub>C</sub> for the general case with a conditional Gaussian latent model, and VMMT<sub>F</sub> 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 reduced variance (though note this is based on only 4 independent runs in our case, and 3 independent runs of Imagination). Both models VMMT<sub>F</sub> and VMMT<sub>C</sub> 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.

In Table 2, we report results when translating the Multi30k<sub>test2017</sub> and the ambiguous MSCOCO test sets. Note that standard deviations for the conditional model VMMT<sub>C</sub> are considerably higher than those obtained for model VMMT<sub>F</sub>. We investigated the issue further and found out that one of the runs of VMMT<sub>C</sub> performed considerably

Model	BLEU4 $\uparrow$	METEOR $\uparrow$	chrF $\uparrow$	BEER $\uparrow$
Multi30k 2017 test set				
VMMT <sub>F</sub>	<b>30.1 (0.3)</b>	<b>49.9 (0.3)</b>	<b>57.2 (0.4)</b>	<b>62.2 (0.3)</b>
VMMT <sub>C</sub>	26.1 (6.6)	45.4 (7.3)	52.2 (8.4)	58.6 (5.8)
Ambiguous MSCOCO 2017 test set				
VMMT <sub>F</sub>	<b>25.5 (0.5)</b>	<b>44.8 (0.2)</b>	<b>52.0 (0.3)</b>	<b>58.3 (0.2)</b>
VMMT <sub>C</sub>	21.8 (5.6)	41.2 (6.3)	47.4 (7.6)	55.3 (5.2)

Table 2: Results of applying variational MMT models to translate the Multi30k 2017 and the ambiguous MSCOCO test sets. For each model, we report the mean and standard deviation over 4 independent runs where models were selected using validation BLEU4 scores. Best overall results (i.e. means) are in bold. Note that standard deviations for the conditional model VMMT<sub>C</sub> are considerably higher than those obtained for model VMMT<sub>F</sub>. This is partly due to the fact that one of the runs of VMMT<sub>C</sub> underperformed compared to the other three.

worse than the others; this caused the mean scores to be much lower and also increased the variance significantly.

Finally, one interesting finding is that all four metrics indicate that the fixed-prior model VMMT<sub>F</sub> either performs slightly (Table 1) or considerably better (Table 2) than the conditional model VMMT<sub>C</sub>. We speculate this is partly due to VMMT<sub>F</sub>'s simpler parameterisation, after all, we have just about 29k training instances to estimate two sets of parameters ( $\theta$  and  $\lambda$ ) and the more complex VMMT<sub>C</sub> requires an additional bidirectional LSTM encoder for the target text.

### 3.4 Back-translated Comparable Multi30k

Since the *translated Multi30k* dataset is very small, we also investigate the effect of including more in-domain data to train our models. For that purpose, we use additional 145K monolingual German descriptions released as part of the *comparable Multi30k* dataset (M30k<sub>C</sub>). We train a text-only NMT model to translate from German into English using the original 29K parallel sentences in the *translated Multi30k* (without images), and apply this model to back-translate the 145K German descriptions into English (Sennrich et al., 2016a).

In this set of experiments, we explore how pre-training models NMT, VMMT<sub>F</sub> and VMMT<sub>C</sub> using both the *translated* and *back-translated comparable Multi30k* affects results. Models are pre-trained on mini-batches with a one-to-one ratio of *translated* and *back-translated* data.<sup>4</sup> All three models NMT, VMMT<sub>F</sub> and VMMT<sub>C</sub>, are further fine-

<sup>4</sup>One pre-training epoch corresponds to about 290K examples, i.e. we up-sample the smaller *translated Multi30k* data set to achieve the one-to-one ratio.

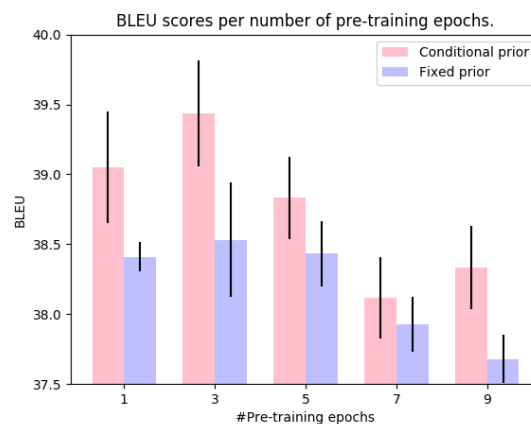


Figure 2: Validation set BLEU scores per number of pre-trained epochs for models VMMT<sub>C</sub> and VMMT<sub>F</sub> pre-trained using the *comparable Multi30k* and *translated Multi30k* data sets. The height of a bar represents the mean and the black vertical lines indicate  $\pm 1$  std over 4 independent runs.

tuned on the translated Multi30k until convergence, and model selection using BLEU is only applied during fine-tuning and not at the pre-training stage.

In Figure 2, we inspect for how many epochs should a model be pre-trained using the additional noisy back-translated descriptions, and note that both VMMT<sub>F</sub> and VMMT<sub>C</sub> reach best BLEU scores on the validation set when pre-trained for about 3 epochs. As shown in Figure 2, we note that when using additional noisy data VMMT<sub>C</sub>, which uses a conditional prior, performs considerably better than its counterpart VMMT<sub>F</sub>, which has a fixed prior. These results indicate that VMMT<sub>C</sub> makes better use of additional synthetic data than VMMT<sub>F</sub>. Some of the reasons that explain these results are (i) the conditional prior  $p(z|x)$  can learn

Model	BLEU4 $\uparrow$	METEOR $\uparrow$	# train sents.
NMT	37.7 (0.5)	56.0 (0.3)	
VMMT <sub>F</sub>	<b>38.4 (0.6)</b> $\uparrow$ 0.7	56.0 (0.3) $\uparrow$ 0.0	145K
VMMT <sub>C</sub>	<b>38.4 (0.2)</b> $\uparrow$ 0.7	<b>56.3 (0.2)</b> $\uparrow$ 0.3	
Imagination	37.8 (0.7)	57.1 (0.2)	654K

Table 3: Results for models pre-trained using the *translated* and *comparable Multi30k* to translate the Multi30k test set. We report the mean and standard deviation over 4 independent runs. Our best overall results are highlighted in bold, and we highlight in green/red the improvement/decrease brought by our models compared to the baseline mean score. We additionally show results for the Imagination model trained on  $4\times$  more data (as reported in the authors’ paper).

to be sensitive to whether  $x$  is gold-standard or synthetic, whereas  $p(z)$  cannot; (ii) in the conditional case the posterior approximation  $q(z|x, y, v)$  can directly exploit different patterns arising from a gold-standard versus a synthetic  $\langle x, y \rangle$  pair; and finally (iii) our synthetic data is made of *target-language* gold-standard image descriptions, which help train the inference network’s target-language BiLSTM encoder.

In Table 3, we show results when applying VMMT<sub>F</sub> and VMMT<sub>C</sub> to translate the Multi30k test set. Both models and the NMT baseline are pre-trained on the *translated* and the *back-translated comparable Multi30k* data sets, and are selected according to validation set BLEU scores. For comparison, we also include results for Imagination (Elliott and Kádár, 2017) when trained on the *translated Multi30k*, the WMT News Commentary English-German dataset (240K parallel sentence pairs) and the MSCOCO image description dataset (414K German descriptions of 83K images, i.e. 5 descriptions for each image). In contrast, our models observe 29K images (i.e. the same as the models evaluated in Section 3.3) plus 145K German descriptions only.<sup>5</sup>

### 3.5 Ablative experiments

In our ablation we are interested in finding out to what extent the model makes use of the latent space, i.e. how important is the latent variable.

**KL free bits** A common issue when training latent variable models with a strong decoder is having

<sup>5</sup>There are no additional images because the *comparable Multi30k* consists of additional German descriptions for the same 29K images already in the *translated* Multi30k.

Model	Number of free bits (KL)	BLEU4 $\uparrow$
VMMT <sub>F</sub>	0	38.3 (0.2)
	1	38.1 (0.3)
	2	<b>38.4 (0.4)</b>
	4	<b>38.4 (0.4)</b>
VMMT <sub>C</sub>	8	35.7 (3.1)
	0	38.5 (0.2)
	1	38.3 (0.3)
	2	38.2 (0.2)
VMMT <sub>C</sub>	4	36.8 (2.6)
	8	<b>38.6 (0.2)</b>

Table 4: Results of applying VMMT models trained with different numbers of free bits in the KL (Kingma et al., 2016) to translate the Multi30k validation set.

the true posterior collapse to the prior and the KL term in the ELBO vanish to zero. In practice, that would mean the model has virtually not used the latent variable  $z$  to predict image features  $v$ , but mostly as a source of stochasticity in the decoder. This can happen because the model has access to informative features from the source bi-LSTM encoder and need not learn a difficult mapping from observations to latent representations predictive of image features.

For that reason, we wish to measure how well can we train latent variable MMT models while ensuring that the KL term in the loss (Equation (6)) does not vanish to zero. We use the *free bits* heuristic (Kingma et al., 2016) to impose a constraint on the KL, which in turn promotes models with non-negligible mutual information between inputs and latent variables (Alemi et al., 2018).

In Table 4, we see the results of different models trained using different number of free bits in the KL component. We note that including free bits improves translations slightly, but note that finding the optimal number of free bits requires hyper-parameter search.

### 3.6 Discussion

In Table 5 we show how our different models translate two examples of the M30k test set. In the first example (id#801), training on additional back-translated data improves variational models but not the NMT baseline, whereas in the second example (id#873) differences between baseline and variational models still persist even when training on



Model		Example #801		Example #873
source reference		a man on a bicycle pedals through an <b>archway</b> . ein mann fährt auf einem fahrrad durch einen <b>torbogen</b> .		a man throws a fishing net into the <b>bay</b> . ein mann wirft ein fischernetz in die <b>bucht</b> .
NMT		<b>M30k<sub>T</sub></b>		<b>M30k<sub>T</sub></b>
VMMT <sub>F</sub> VMMT <sub>C</sub>		ein mann auf einem fahrrad fährt durch eine <b>scheibe</b> . ein mann auf einem fahrrad fährt durch einen <b>torbogen</b> . ein mann auf einem fahrrad fährt durch einen <b>bogen</b> .		ein mann wirft ein fischernetz in die <b>luft</b> . ein mann wirft ein fischernetz in die <b>bucht</b> . ein mann wirft ein fischernetz in die <b>bucht</b> .
NMT VMMT <sub>F</sub> VMMT <sub>C</sub>		<b>M30k<sub>T</sub> + back-translated M30k<sub>C</sub></b>		<b>M30k<sub>T</sub> + back-translated M30k<sub>C</sub></b>
		ein mann auf einem fahrrad fährt durch einen <b>bogen</b> . ein mann auf einem fahrrad fährt durch einen <b>torbogen</b> . ein mann auf einem fahrrad fährt durch einen <b>torbogen</b> .		ein mann wirft ein fischernetz ins <b>meer</b> . ein mann wirft ein fischernetz in den <b>wellen</b> . ein mann wirft ein fischernetz in die <b>bucht</b> .

Table 5: Translations for examples 801 and 873 of the M30k test set. In the first example, neither the NMT baseline (with or without back-translated data) nor model VMMT<sub>C</sub> (trained on limited data) could translate **archway** correctly; the NMT baseline translates it as “scheibe” (disk) and “bogen” (bow), and VMMT<sub>C</sub> also incorrectly translates it as “bogen” (bow). However, VMMT<sub>C</sub> translates without errors when trained on additional back-translated data, i.e. “torbogen” (archway). In the second example, the NMT baseline translates **bay** as “luft” (air) or “meer” (sea), whereas VMMT<sub>F</sub> translates it as “bucht” (bay) or “wellen” (waves) and VMMT<sub>C</sub> always as “bucht” (bay).

additional back-translated data.

## 4 Related work

Even though there has been growing interest in variational approaches to machine translation (Zhang et al., 2016; Schulz et al., 2018; Shah and Barber, 2018; Eikema and Aziz, 2019) and to tasks that integrate vision and language, e.g. image description generation (Pu et al., 2016; Wang et al., 2017), relatively little attention has been dedicated to variational models for multi-modal translation. This is partly due to the fact that 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; Barrault et al., 2018). Nevertheless, we now discuss relevant variational and deterministic multi-modal MT models in the literature.

**Fully supervised MMT models.** All submissions to the three runs of the multi-modal MT shared tasks (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018) model conditional probabilities directly without latent variables.

Perhaps the first MMT model proposed prior to these shared tasks is that of Hirschler 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 significant improvements. Shah et al. (2016) propose a similar model where image logits are used to re-rank the output of PBSMT. Global image features, i.e. features computed over an entire image (such as  $p_{0015}$  ResNet-50 features used in this work), have been directly used as “tokens” 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 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 set up 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  $x_1^m$ , i.e.  $p(z|x_1^m)$ , and both  $x_1^m$  and  $z$  are used at each time step in an attentive decoder RNN,  $P(y_j|x_1^m, z, y_{<j})$ .

In Toyama et al. (2016), image features are used as input to the inference model  $q_\lambda(z|x_1^m, y_1^n, v)$  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 image features in all our generative models, and propose modelling them as random observed outcomes 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.  $p(z|x_1^m)$  and  $p(z)$ , whereas their model is always conditional. Interestingly, we found in our experiments that fixed-prior models perform slightly better than conditional ones under limited training data.

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 nor object detections.

## 5 Conclusions and Future work

We have proposed a latent variable model for multi-modal neural machine translation and have shown benefits from both modelling images and promoting use of latent space. We also show that in the absence of enough data to train a more complex inference network a simple fixed prior suffices, whereas

when more training data is available (even noisy data) a conditional prior is preferable. Importantly, our models compare favourably to the state-of-the-art.

In future work we will explore other generative models for multi-modal MT, as well as different ways to directly incorporate images into these models. We are also interested in modelling different views of the image, such as global vs. local image features, and also in using larger image collections and modelling images directly, i.e. pixel intensities.

## Acknowledgements

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

## References

- Alexander A. Alemi, Ben Poole, Ian Fischer, Joshua V. Dillon, Rif A. Saurous, and Kevin Murphy. 2018. Fixing a Broken ELBO. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, pages 159–168.
- 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.
- Loïc Barrault, Fethi Bougares, Lucia Specia, Chirag 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, Belgium, Brussels. Association for Computational Linguistics.
- 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. 2019. **Auto-encoding variational neural machine translation.** In *4th Workshop on Representation Learning for NLP*.
- 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*, 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 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.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. **Long short-term memory.** *Neural Comput.*, 9(8):1735–1780.
- 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.
- 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 Lukasz Kaiser. 2016. 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.
- Maja Popović. 2015. [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.
- Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. 2016. [Variational autoencoder for deep learning of images, labels and captions](#). In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2352–2360. Curran Associates, Inc.
- 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.
- Philip Schulz, Wilker Aziz, and Trevor Cohn. 2018. [A stochastic decoder for neural machine translation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1243–1252. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. [Improving Neural Machine Translation Models with Monolingual Data](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. [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.
- Harshil Shah and David Barber. 2018. [Generative neural machine translation](#). In S. Bengio, H. Wallach, H. Larochelle, K. Grauman,

- N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 1346–1355. Curran Associates, Inc.
- 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 Xinchun 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.
- Michalis Titsias and Miguel Lázaro-Gredilla. 2014. Doubly stochastic variational bayes for non-conjugate inference. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1971–1979.
- Joji Toyama, Masanori Misono, Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. 2016. [Neural machine translation with latent semantic of image and text](#). *CoRR*, abs/1611.08459.
- Liwei Wang, Alexander Schwing, and Svetlana Lazebnik. 2017. [Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space](#). In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5756–5766. Curran Associates, Inc.
- 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.
- Biao Zhang, Deyi Xiong, jinsong su, Hong Duan, and Min Zhang. 2016. [Variational neural machine translation](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 521–530, Austin, Texas. Association for Computational Linguistics.

## A Model Architecture

Once again, we wish to translate a source sequence  $x_1^m \triangleq \langle x_1, \dots, x_m \rangle$  into a target sequence  $y_1^n \triangleq \langle y_1, \dots, y_n \rangle$ , and also predict image features  $v$ .

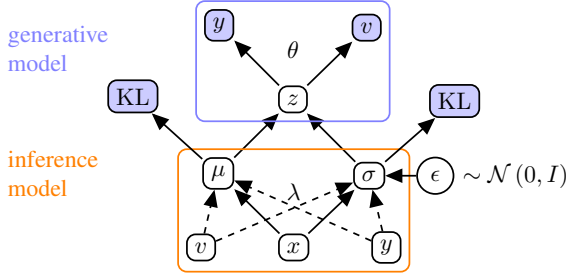


Figure 3: Illustration of multi-modal machine translation generative and inference models. The conditional model  $\text{VMMT}_C$  includes dashed arrows; the fixed prior model  $\text{VMMT}_F$  does not, i.e. its inference network only uses  $x$ .

In Figure 3, we illustrate generative and inference networks for models  $\text{VMMT}_C$  and  $\text{VMMT}_F$ .

### A.1 Generative model

**Source-language encoder** The source-language encoder is deterministic and implemented using a 2-layer bidirectional Long Short-Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997):

$$\begin{aligned} f_i &= \text{emb}(x_i; \theta_{\text{emb-x}}), \\ h_0 &= \vec{0}, \\ \vec{h}_i &= \text{LSTM}(h_{i-1}, f_i; \theta_{\text{lstmf-x}}), \\ \overleftarrow{h}_i &= \text{LSTM}(h_{i+1}, f_i; \theta_{\text{lstm-b-x}}), \\ h_i &= [\vec{h}_i, \overleftarrow{h}_i], \end{aligned} \quad (9)$$

where  $\text{emb}$  is the source look-up matrix, trained jointly with the model, and  $h_1^m$  are the final source hidden states.

**Target-language decoder** Now we assume that  $z$  is given, and will discuss how to compute it later on. The translation model consists of a sequence of draws from a Categorical distribution over the target-language vocabulary (independently from image features  $v$ ):

$$Y_j | z, x, y_{<j} \sim \text{Cat}(f_\theta(z, x, y_{<j})),$$

where  $f_\theta$  parameterises the distribution with an attentive encoder-decoder architecture:

$$\begin{aligned} w_j &= \text{emb}(y_j; \theta_{\text{emb-y}}), \\ s_0 &= \tanh(\text{affine}(h_1^m; \theta_{\text{init-y}})), \\ s_j &= \text{LSTM}(s_{j-1}, [w_j, z]; \theta_{\text{lstm-y}}), \\ c_{i,j} &= \text{attention}(h_1^m, s_1^n; \theta_{\text{attn}}), \end{aligned}$$

$$f_\theta(z, x, y_{<j}) = \text{softmax}(\text{affine}([s_j, c_j]; \theta_{\text{out-y}})),$$

where the attention mechanism is a bi-linear attention (Luong et al., 2015), and the generative parameters are  $\theta = \{\theta_{\text{emb-}\{x,y\}}, \theta_{\text{lstm}\{f,b\}-x}, \theta_{\text{init-y}}, \theta_{\text{lstm-y}}, \theta_{\text{attn}}, \theta_{\text{out-y}}\}$ .

**Image decoder** We do not model images directly, but instead as a 2048-dimensional feature vector  $v$  of pre-activations of a ResNet-50’s `pool5` layer. We simply draw image features from a Gaussian observation model:

$$\begin{aligned} V | z &\sim \mathcal{N}(\nu, \varsigma^2 I), \\ \nu &= \text{MLP}(z; \theta), \end{aligned} \quad (10)$$

where a multi-layer perceptron (MLP) maps from  $z$  to a vector of locations  $\nu \in \mathbb{R}^o$ , and  $\varsigma \in \mathbb{R}_{>0}$  is a hyper-parameter of the model (we use 1).

**Conditional prior  $\text{VMMT}_C$**  Given a source sentence  $x_1^m$ , we draw an embedding  $z$  from a latent Gaussian model:

$$\begin{aligned} Z | x_1^m &\sim \mathcal{N}(\mu, \text{diag}(\sigma^2)), \\ \mu &= \text{MLP}(h_1^m; \theta_{\text{latent}}), \\ \sigma &= \text{softplus}(\text{MLP}(h_1^m; \theta_{\text{latent}})), \end{aligned} \quad (11)$$

where Equations (11) and (12) employ two multi-layer perceptrons (MLPs) to map from a source sentence (i.e. source hidden states) to a vector of locations  $\mu \in \mathbb{R}^c$  and a vector of scales  $\sigma \in \mathbb{R}_{>0}^c$ , respectively.

**Fixed prior  $\text{VMMT}_F$**  In the MMT model  $\text{VMMT}_F$ , we simply have a draw from a standard Normal prior:

$$Z \sim \mathcal{N}(0, I).$$

All MLPs have one hidden layer and are implemented as below (eqs. (10) to (12)):

$$\text{MLP}(\cdot) = \text{affine}(\text{ReLU}(\text{affine}(\cdot; \theta))); \theta).$$

### A.2 Inference model

The inference network shares the source-language encoder with the generative model and differs depending on the model ( $\text{VMMT}_C$  or  $\text{VMMT}_F$ ).

**Conditional prior VMMT<sub>C</sub>** Model VMMT<sub>C</sub>'s approximate posterior  $q_\lambda(z|x_1^m, y_1^n, v)$  is a Gaussian distribution:

$$Z|x_1^m, y_1^n, v \sim \mathcal{N}(\mathbf{u}, \text{diag}(\mathbf{s}^2); \lambda).$$

We use two bidirectional LSTMs, one over source- and the other over target-language words, respectively. To reduce the number of model parameters, we re-use the entire source-language BiLSTM and the target-language embeddings in the generative model but prevent updates to the generative model's parameters by blocking gradients from being back-propagated (Equation 9). Concretely, the inference model is parameterised as below:

$$\begin{aligned} \mathbf{h}_1^m &= \text{detach}(\text{BiLSTM}(x_1^m; \theta_{\text{emb-x, lstm-f-x, lstm-b-x}})), \\ \mathbf{w}_1^n &= \text{detach}(\text{emb}(y_1^n; \theta_{\text{emb-y}})), \\ \mathbf{h}_x &= \text{avg}(\text{affine}(\mathbf{h}_1^m; \lambda_x)), \\ \mathbf{h}_y &= \text{avg}(\text{BiLSTM}(\mathbf{w}_1^n; \lambda_y)), \\ \mathbf{h}_v &= \text{MLP}(\mathbf{v}; \lambda_v), \\ \mathbf{h}_{\text{all}} &= [\mathbf{h}_x, \mathbf{h}_y, \mathbf{h}_v], \\ \mathbf{u} &= \text{MLP}(\mathbf{h}_{\text{all}}; \lambda_{\text{mu}}), \\ \mathbf{s} &= \text{softplus}(\text{MLP}(\mathbf{h}_{\text{all}}; \lambda_{\text{sigma}})), \end{aligned}$$

where the set of the inference network parameters are  $\lambda = \{\lambda_x, \lambda_y, \lambda_v, \lambda_{\text{mu}}, \lambda_{\text{sigma}}\}$ .

**Fixed prior VMMT<sub>F</sub>** Model VMMT<sub>F</sub>'s approximate posterior  $q_\lambda(z|x_1^m)$  is also a Gaussian:

$$Z|x_1^m \sim \mathcal{N}(\mathbf{u}, \text{diag}(\mathbf{s}^2); \lambda),$$

where we re-use the source-language BiLSTM from the generative model but prevent updates to its parameters by blocking gradients from being back-propagated (Equation 9). Concretely, the inference model is parameterised as below:

$$\begin{aligned} \mathbf{h}_1^m &= \text{detach}(\text{BiLSTM}(x_1^m; \theta_{\text{emb-x, lstm-f-x, lstm-b-x}})), \\ \mathbf{h}_x &= \text{avg}(\text{affine}(\mathbf{h}_1^m; \lambda_x)), \\ \mathbf{u} &= \text{MLP}(\mathbf{h}_x; \lambda_{\text{mu}}), \\ \mathbf{s} &= \text{softplus}(\text{MLP}(\mathbf{h}_x; \lambda_{\text{sigma}})), \end{aligned}$$

where the set of the inference network parameters are  $\lambda = \{\lambda_x, \lambda_{\text{mu}}, \lambda_{\text{sigma}}\}$ .

Finally, all MLPs are implemented as below:

$$\text{MLP}(\cdot) = \text{affine}(\text{ReLU}(\text{affine}(\cdot; \lambda)); \lambda).$$