Latent Variable Model for Multi-modal Translation

Calixto, I.; Rios, M.; Aziz, W.

Published in:
The 57th Annual Meeting of the Association for Computational Linguistics

DOI:
10.18653/v1/P19-1642

Citation for published version (APA):
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 K´ad´ar, 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 K´ad´ar (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 K´ad´ar, 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 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 sentence \( x^n \) into a target sequence \( y^n \). The main difference is the presence of an image \( v \) which illustrates the sentence pair \( x^n, y^n \). We do not model images directly, but instead an 2048-dimensional vector of pre-activations of a ResNet-50’s maxpool5 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 \( y^n \) are independently generated given a common stochastic embedding \( z \). The generative story is as follows. We observe a source sentence \( x^n \) and draw an embedding \( z \) from a latent Gaussian model,

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

\[
\mu = f_\mu(x^n; \theta) \\
\sigma = f_\sigma(x^n; \theta)
\]

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 \( \sigma \in \mathbb{R}_{>0}^c \), respectively. We then proceed to draw the image features from a Gaussian observation model,

\[
V|z \sim \mathcal{N}(\nu, \sigma^2 I) \\
\nu = f_\nu(z; \theta)
\]

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

\[
Y_j|x^n, z, y_{<j} \sim \text{Cat}(\pi_j) \\
\pi_j = f_{\pi}(x^n, y_{<j}, z; \theta)
\]

Similarly to standard NMT, in MMT we wish to generate a translation by drawing each target word in context from a Categorical observation model,

\[
Y_j|x^n, z, y_{<j} \sim \text{Cat}(\pi_j) \\
\pi_j = f_{\pi}(x^n, y_{<j}, z; \theta)
\]

Note that in effect we model a joint distribution

\[
p_\theta(y^n, v, z|x^n) = p_\theta(z|x^n)p_\theta(v|z)p_\theta(y^n|x^n, z)
\]

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

\[
P_\theta(y^n, v|x^n) = \int p_\theta(z|x^n)p_\theta(v|z)p_\theta(y^n|x^n, z)dz.
\]

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

\[\text{Locations have support on the entire real space, thus we use linear activations, scales must be strictly positive, thus we use a softplus activation.}\]
while stochastic embeddings (Eikema and Aziz, 2019, a) VMMT variational inference (VI) (Jordan et al., 1999), y
Inference Parameter estimation for our model is explicitly in the near future. In particular, our inference model lowerbound (ELBO) is expressed in terms of an on the log-likelihood function. This evidence lowerbound, for two Gaussians the proximate posterior is a Gaussian distribution

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

\[ \mu = g_u(x^n_1, y^n_1, v; \lambda) \]

\[ s = g_s(x^n_1, y^n_1, v; \lambda) \]

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 \( \mu \in \mathbb{R}^c \) and a variational scale \( s \in \mathbb{R}^{c>0} \). 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:

\[ z = \mu + \epsilon \odot s \quad \text{where} \quad \epsilon \sim \mathcal{N}(0, I). \]

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, \( f_\sigma \) is a standard sequence-to-sequence architecture with attention and a softmax output. We build upon OpenNMT (Klein et al., 2017), which we modify by adding a reparametrisation network to \( f_\sigma \) and a location-scale reparametrisation network to the output of the decoder. This allows us to build a generative model of the translation.

\[ y^n_1, v|x_1^m \] 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, 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 \( (x^n_1, y^n_1, v) \).

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

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

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 ap-
We use OpenNMT to implement all our models. 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 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 $x^m_1$ 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 $\hat{0}$.

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 (NMT, VMMT$_F$, 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 (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-
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.

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU4↑</th>
<th>METEOR↑</th>
<th>chrF↑</th>
<th>BEER↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMT</td>
<td>35.0 (0.4)</td>
<td>54.9 (0.2)</td>
<td>61.0 (0.2)</td>
<td>65.2 (0.1)</td>
</tr>
<tr>
<td>Imagination</td>
<td>36.8 (0.8)</td>
<td>55.8 (0.4)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Model G</td>
<td>36.5</td>
<td>56.0</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>VMMT_F</td>
<td>37.7 (0.4)</td>
<td>56.0 (0.1)</td>
<td>62.1 (0.1)</td>
<td>66.6 (0.1)</td>
</tr>
<tr>
<td>VMMT_C</td>
<td>37.5 (0.3)</td>
<td>55.7 (0.1)</td>
<td>61.9 (0.1)</td>
<td>66.5 (0.1)</td>
</tr>
</tbody>
</table>

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.

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 hyper-parameters 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↔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 (M30kT).

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 reduced variance (though note this is based on only 4 independent runs in our case, and 3 independent runs of Imagination). 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.

In Table 2, we report results when translating the Multi30k test2017 and the ambiguous MSCOCO test sets. Note that standard deviations for the conditional model VMMT_C are considerably higher than those obtained for model VMMT_F. We investigated the issue further and found out that one of the runs of VMMT_C performed considerably higher than the others.
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_C are considerably higher than those obtained for model VMMT_F. This is partly due to the fact that one of the runs of VMMT_C underperformed compared to the other three.

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU4↑</th>
<th>METEOR↑</th>
<th>chrF↑</th>
<th>BEER↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi30k 2017 test set</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VMMT_F</td>
<td>30.1 (0.3)</td>
<td>49.9 (0.3)</td>
<td>57.2 (0.4)</td>
<td>62.2 (0.3)</td>
</tr>
<tr>
<td>VMMT_C</td>
<td>26.1 (6.6)</td>
<td>45.4 (7.3)</td>
<td>52.2 (8.4)</td>
<td>58.6 (5.8)</td>
</tr>
<tr>
<td>Ambiguous MSCOCO 2017 test set</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VMMT_F</td>
<td>25.5 (0.5)</td>
<td>44.8 (0.2)</td>
<td>52.0 (0.3)</td>
<td>58.3 (0.2)</td>
</tr>
<tr>
<td>VMMT_C</td>
<td>21.8 (5.6)</td>
<td>41.2 (6.3)</td>
<td>47.4 (7.6)</td>
<td>55.3 (5.2)</td>
</tr>
</tbody>
</table>

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_F either performs slightly (Table 1) or considerably better (Table 2) than the conditional model VMMT_C. We speculate this is partly 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.

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_C). 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_F and VMMT_C 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. All three models NMT, VMMT_F and VMMT_C, are further fine-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 a model should be pre-trained using the additional noisy back-translated descriptions, and note that both VMMT_F and VMMT_C 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 ±1 std over 4 independent runs.

![Figure 2: Validation set BLEU scores per number of pre-trained epochs for models VMMT_C and VMMT_F 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 ±1 std over 4 independent runs.](image-url)

---

4One 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.
In one of our ablation experiments, 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 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 numbers 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.

### 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>B. (θ)</th>
<th>M. (θ)</th>
<th># train sents.</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMT</td>
<td>37.7 (0.5)</td>
<td>56.0 (0.3)</td>
<td></td>
</tr>
<tr>
<td>VMMTF</td>
<td>38.4 (0.6)</td>
<td>0.7</td>
<td>38.6 (0.4)</td>
</tr>
<tr>
<td>VMMTC</td>
<td>38.4 (0.2)</td>
<td>0.7</td>
<td>56.3 (0.2)</td>
</tr>
<tr>
<td>Imagination</td>
<td>37.8 (0.7)</td>
<td>57.1 (0.2)</td>
<td>654K</td>
</tr>
</tbody>
</table>

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).

There are no additional images because the comparable Multi30k consists of additional German descriptions for the same 29K images already in the translated Multi30k.
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 VMMTC trained on limited data could translate archway correctly; the NMT baseline translates it as “scheibe” (disk) and “bogen” (bow), and VMMTC also incorrectly translates it as “bogen” (bow). However, VMMTCC 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 VMMTF translates it as “bucht” (bay) or “wellen” (waves) and VMMTC always as “bucht” (bay).

addition 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 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 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 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; Libovichy 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 $x^m_1$, 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 Toyama et al. (2016), image features are used as input to the inference model $q_λ(z|x^m_1, y^n_1, v)$ 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 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. $p(z|x^m_1)$ 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 RNN 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 multimodal 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


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


A Model Architecture

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

![Illustration of multi-modal machine translation](image)

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

In Figure 3, we illustrate generative and inference networks for models VMMT_C and 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):

$$
\begin{align*}
    f_i &= \text{emb}(x_i; \theta_{\text{emb-x}}), \\
    h^0_0 &= \vec{0}, \\
    \overrightarrow{h}_i &= \text{LSTM}(h_{i-1}, f_i; \theta_{\text{lstmb-x}}), \\
    \overleftarrow{h}_i &= \text{LSTM}(h_{i+1}, f_i; \theta_{\text{lstmb-x}}), \\
    h_i &= [\overrightarrow{h}_i, \overleftarrow{h}_i],
\end{align*}
$$

where $\text{emb}$ is the source look-up matrix, trained jointly with the model, and $h^n_0$ 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:

$$
\begin{align*}
    w_j &= \text{emb}(y_j; \theta_{\text{emb-y}}), \\
    s_0 &= \tanh(\text{affine}(h^n_0; \theta_{\text{init-y}})), \\
    s_j &= \text{LSTM}(s_{j-1}, [w_j, z]; \theta_{\text{lstmy}}), \\
    c_{i,j} &= \text{attention}(h^1_m, s^n_j; \theta_{\text{attn}}), \\
    f_\theta(z, x, y_{<j}) &= \text{softmax}(\text{affine}(s_j, c_{j}; \theta_{\text{out-y}})),
\end{align*}
$$

where the attention mechanism is a bilinear attention (Luong et al., 2015), and the generative parameters are $\theta = \{\theta_{\text{emb-x,y}}, \theta_{\text{lstmy}}, \theta_{\text{inn}}, \theta_{\text{attn}}, \theta_{\text{out-y}}\}$.

**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:

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

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

**Conditional prior VMMT_C** Given a source sentence $x^m_0$, we draw an embedding $z$ from a latent Gaussian model:

$$
\begin{align*}
    Z | x^m_0 &\sim N(\mu, \text{diag}(\varsigma^2)), \\
    \mu &= \text{MLP}(h^n_0; \theta_{\text{latent}}), \\
    \varsigma &= \text{softplus}(\text{MLP}(h^n_0; \theta_{\text{latent}})),
\end{align*}
$$

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 $\varsigma \in \mathbb{R}_{>0}$, respectively.

**Fixed prior VMMT_F** In the MMT model VMMT_F, we simply have a draw from a standard Normal prior:

$$
Z \sim N(0, I).
$$

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 (VMMT_C or VMMT_F).
**Conditional prior VMMT\(_C\)**  Model VMMT\(_C\)’s approximate posterior \(q_\lambda(z|x^m_1, y^n_1, v)\) is a Gaussian distribution:

\[
Z|x^m_1, y^n_1, v \sim \mathcal{N}(u, \text{diag}(s^2); \lambda).
\]

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:

\[
\begin{align*}
    h^m_1 &= \text{detach}(\text{BiLSTM}(x^m_1; \theta_{\text{emb-x}, \text{lstm-f-x}, \text{lstm-b-x}})), \\
    w^n_1 &= \text{detach}(\text{emb}(y^n_1; \theta_{\text{emb-y}})), \\
    h_x &= \text{avg}(\text{affine}(h^m_1; \lambda_x)), \\
    h_y &= \text{avg}(\text{BiLSTM}(w^n_1; \lambda_y)), \\
    h_v &= \text{MLP}(v; \lambda_v), \\
    h_{\text{all}} &= [h_x, h_y, h_v], \\
    u &= \text{MLP}(h_{\text{all}}; \lambda_{\text{mu}}), \\
    s &= \text{softplus}(\text{MLP}(h_{\text{all}}; \lambda_{\text{sigma}})),
\end{align*}
\]

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\(_F\)**  Model VMMT\(_F\)’s approximate posterior \(q_\lambda(z|x^m_1)\) is also a Gaussian:

\[
Z|x^m_1 \sim \mathcal{N}(u, \text{diag}(s^2); \lambda),
\]

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:

\[
\begin{align*}
    h^m_1 &= \text{detach}(\text{BiLSTM}(x^m_1; \theta_{\text{emb-x}, \text{lstm-f-x}, \text{lstm-b-x}})), \\
    h_x &= \text{avg}(\text{affine}(h^m_1; \lambda_x)), \\
    u &= \text{MLP}(h_x; \lambda_{\text{mu}}), \\
    s &= \text{softplus}(\text{MLP}(h_x; \lambda_{\text{sigma}})),
\end{align*}
\]

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:

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