LIUM-CVC Submissions for WMT17 Multimodal Translation Task
Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes Garcia-Martinez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, Joost van de Weijer

To cite this version:
Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes Garcia-Martinez, Fethi Bougares, et al.. LIUM-CVC Submissions for WMT17 Multimodal Translation Task. SECOND CONFERENCE ON MACHINE TRANSLATION, 2017, Copenhagen, Denmark. pp.432 - 439. hal-01742382

HAL Id: hal-01742382
https://hal-univ-lemans.archives-ouvertes.fr/hal-01742382
Submitted on 26 Mar 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
LIUM-CVC Submissions for WMT17 Multimodal Translation Task

Ozan Caglayan†, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault
LIUM, University of Le Mans
†ozancag@gmail.com
FirstName.LastName@univ-lemans.fr

Marc Masana, Luis Herranz and Joost van de Weijer
CVC, Universitat Autonoma de Barcelona
{joost,mmasana,lherranz}@cvc.uab.es

Abstract
This paper describes the monomodal and multimodal Neural Machine Translation systems developed by LIUM and CVC for WMT17 Shared Task on Multimodal Translation. We mainly explored two multimodal architectures where either global visual features or convolutional feature maps are integrated in order to benefit from visual context. Our final systems ranked first for both En→De and En→Fr language pairs according to the automatic evaluation metrics METEOR and BLEU.

1 Introduction
With the recent advances in deep learning, purely neural approaches to machine translation, such as Neural Machine Translation (NMT), (Sutskever et al., 2014; Bahdanau et al., 2014) have received a lot of attention because of their competitive performance (Toral and Sánchez-Cartagena, 2017). Another reason for the popularity of NMT is its flexible nature allowing researchers to fuse auxiliary information sources in order to design sophisticated networks like multi-task, multi-way and multi-lingual systems to name a few (Luong et al., 2015; Johnson et al., 2016; Firat et al., 2017).

Multimodal Machine Translation (MMT) aims to achieve better translation performance by visually grounding the textual representations. Recently, a new shared task on Multimodal Machine Translation and Crosslingual Image Captioning (CIC) was proposed along with WMT16 (Specia et al., 2016). In this paper, we present MMT systems jointly designed by LIUM and CVC for the second edition of this task within WMT17.

Last year we proposed a multimodal attention mechanism where two different attention distributions were estimated over textual and image representations using shared transformations (Caglayan et al., 2016a). More specifically, convolutional feature maps extracted from a ResNet-50 CNN (He et al., 2016) pre-trained on the ImageNet classification task (Russakovsky et al., 2015) were used to represent visual information. Although our submission ranked first among multimodal systems for CIC task, it was not able to improve over purely textual NMT baselines in neither tasks (Specia et al., 2016). The winning submission for MMT (Caglayan et al., 2016a) was a phrase-based MT system rescored using a language model enriched with FC7 global visual features extracted from a pre-trained VGG-19 CNN (Simonyan and Zisserman, 2014).

State-of-the-art results were obtained after WMT16 by using a separate attention mechanism for different modalities in the context of CIC (Caglayan et al., 2016b) and MMT (Calixto et al., 2017a). Besides experimenting with multimodal attention, Calixto et al. (2017a) and Libovický and Helcl (2017) also proposed a gating extension inspired from Xu et al. (2015) which is believed to allow the decoder to learn when to attend to a particular modality although Libovický and Helcl (2017) report no improvement over baseline NMT.

There have also been attempts to benefit from different types of visual information instead of relying on features extracted from a CNN pre-trained on ImageNet. One such study from Huang et al. (2016) extended the sequence of source embeddings consumed by the RNN with several regional features extracted from a region-proposal...
network (Ren et al., 2015). The architecture thus predicts a single attention distribution over a sequence of mixed-modality representations leading to significant improvement over their NMT baseline.

More recently, a radically different multi-task architecture called Imagination (Elliott and Kadár, 2017) is proposed to learn visually grounded representations by sharing an encoder between two tasks: a classical encoder-decoder NMT and a visual feature reconstruction using as input the source sentence representation.

This year, we experiment\(^1\) with both convolutional and global visual vectors provided by the organizers to better exploit multimodality (Section 3). Data preprocessing for both English→{German,French} and training hyper-parameters are detailed respectively in Section 2 and Section 4. The results based on automatic evaluation metrics are reported in Section 5. The paper ends with a discussion in Section 6.

2 Data

We use the Multi30k (Elliott et al., 2016) dataset provided by the organizers which contains 29000, 1014 and 1000 English→{German,French} image-caption pairs respectively for training, validation and Test2016 (the official evaluation set of WMT16 campaign) set. Following task rules we normalized punctuations, applied tokenization and lowercasing. A Byte Pair Encoding (BPE) model (Sennrich et al., 2016) with 10K merge operations is learned for each language pair resulting in 5234→7052 tokens for English→German and 5945→6547 tokens for English→French respectively.

We report results on Flickr Test2017 set containing 1000 image-caption pairs and the additional ambiguous MSCOCO test set (Elliott et al., 2017) of 461 image-caption pairs.

Image Features We experimented with several types of visual representation using deep features extracted from convolutional neural networks (CNN) trained on large visual datasets. Following the current state-of-the-art in visual representation, we used a network with the ResNet-50 architecture (He et al., 2016) trained on the ImageNet dataset (Russakovsky et al., 2015) to extract two types of features: the 2048-dimensional features from the pool5 layer and the 14x14x1024 features from the res4f.relu layer. Note that the former is a global feature while the latter is a feature map with roughly localized spatial information.

3 Architecture

Our baseline NMT is an attentive encoder-decoder (Bahdanau et al., 2014) variant with a Conditional GRU (CGRU) (Firat and Cho, 2016) decoder.

Let us denote source and target sequences \(X\) and \(Y\) with respective lengths \(M\) and \(N\) as follows where \(x_i\) and \(y_j\) are embeddings of dimension \(E\):

\[
X = (x_1, \ldots, x_M)
\]
\[
Y = (y_1, \ldots, y_N)
\]

**Encoder** Two GRU (Chung et al., 2014) encoders with \(R\) hidden units each, process the source sequence \(X\) in forward and backward directions. Their hidden states are concatenated to form a set of \(source\ annotations\) \(S\) where each element \(s_i\) is a vector of dimension \(C = 2 \times R\):

\[
S = \begin{bmatrix} GRU_{\text{Forw}}(\hat{X}) \\ GRU_{\text{Back}}(\bar{X}) \end{bmatrix} \in \mathbb{R}^{M \times C}
\]

Both encoders are equipped with layer normalization (Ba et al., 2016) where each hidden unit adaptively normalizes its incoming activations with a learnable gain and bias.

**Decoder** A decoder block namely CGRU (two stacked GRUs where the hidden state of the first GRU is used for attention computation) is used to estimate a probability distribution over target tokens at each decoding step \(t\).

The hidden state \(h_0\) of the CGRU is initialized using a non-linear transformation of the average source annotation:

\[
h_0 = \tanh \left( W_{\text{init}} \cdot \frac{1}{M} \sum_{i} s_i \right), s_i \in S
\]

**Attention** At each decoding timestep \(t\), an un-normalized attention score \(g_t\) is computed for each source annotation \(s_i\) using the first GRU’s hidden state \(h_t\) and \(s_i\) itself:

\[
(W_a \in \mathbb{R}^C, W_s \in \mathbb{R}^{C \times C} \text{ and } W_h \in \mathbb{R}^{C \times R})
\]

\[
g_t = W_a^T \tanh (W_s s_i + b_s + W_h h_t) + b_a
\]

\(1\)A detailed tutorial for reproducing the results of this paper is provided at https://github.com/lium-lst/wmt17-mnt.
The context vector $c_t$ is a weighted sum of $s_i$ and its respective attention probability $\alpha_i$ obtained using a softmax operation over all the unnormalized scores:

$$\alpha_i = \text{softmax} \left( \left\{ g_1, g_2, \ldots, g_M \right\}_i \right)$$

$$c_t = \sum_i \alpha_i s_i$$

The final hidden state $\tilde{h}_t$ is computed by the second GRU using the context vector $c_t$ and the hidden state of the first GRU $h_t$.

**Output** The probability distribution over the target tokens is conditioned on the previous token embedding $y_{t-1}$, the hidden state of the decoder $\tilde{h}_t$ and the context vector $c_t$, the latter two transformed with $\mathbf{W}_{\text{dec}}$ and $\mathbf{W}_{\text{ctx}}$ respectively:

$$o_t = \tanh (y_{t-1} + \mathbf{W}_{\text{dec}} \tilde{h}_t + \mathbf{W}_{\text{ctx}} c_t)$$

$$P(y_t|y_{t-1}, \tilde{h}_t, c_t) = \text{softmax}(\mathbf{W}_o o_t)$$

### 3.1 Multimodal NMT

#### 3.1.1 Convolutional Features

The **fusion-conv** architecture extends the CGRU decoder to a multimodal decoder (Caglayan et al., 2016b) where convolutional feature maps of 14x14x1024 are regarded as 196 spatial annotations $s'_j$ of 1024-dimension each. For each spatial annotation, an unnormalized attention score $g'_j$ is computed (Equation 2) except that the weights and biases are specific to the visual modality and thus not shared with the textual attention:

$$g'_j = \mathbf{W}^T \tanh \left( \mathbf{W}_s \mathbf{s}'_j + b'_s + \mathbf{W}_h \tilde{h}_t \right) + b'_a$$

The visual context vector $v_t$ is computed as a weighted sum of the spatial annotations $s'_j$ and their respective attention probabilities $\beta_j$:

$$\beta_j = \text{softmax} \left( \left\{ g'_1, g'_2, \ldots, g'_{196} \right\}_j \right)$$

$$v_t = \sum_j \beta_j s'_j$$

The output of the network is now conditioned on a multimodal context vector which is the concatenation of the original context vector $c_t$ and the newly computed visual context vector $v_t$.

#### 3.1.2 Global pool5 Features

In this section, we present 5 architectures guided with global 2048-dimensional visual representation $V$ in different ways. In contrast to the baseline NMT, the decoder’s hidden state $h_0$ is initialized with an all-zero vector unless otherwise specified.

- **dec-init** initializes the decoder with $V$ by replacing Equation 1 with the following:

  $$h_0 = \tanh (\mathbf{W}_{\text{img}} \cdot V)$$

(Calixto et al., 2017b) previously explored a similar configuration (IMG$_d$) where the decoder is initialized with the sum of global visual features extracted from FC7 layer of a pre-trained VGG-19 CNN and the last source annotation.

- **encdec-init** initializes the bi-directional encoder and the decoder with $V$ where $e_0$ represents the initial state of encoder (Note that in the baseline NMT, $e_0$ is an all-zero vector):

  $$e_0 = h_0 = \tanh (\mathbf{W}_{\text{img}} \cdot V)$$

- **ctx-mul** modulates each source annotation $s_i$ with $V$ using element-wise multiplication:

  $$s_i = s_i \odot \tanh (\mathbf{W}_{\text{img}} \cdot V)$$

- **trg-mul** modulates each target embedding $y_j$ with $V$ using element-wise multiplication:

  $$y_j = y_j \odot \tanh (\mathbf{W}_{\text{img}} \cdot V)$$

- **dec-init-ctx-trg-mul** combines the latter two architectures with **dec-init** and uses separate transformation layers for each of them:

  $$h_0 = \tanh (\mathbf{W}_{\text{img}} \cdot V)$$

  $$s_i = s_i \odot \tanh (\mathbf{W}_{\text{img}} \cdot V)$$

  $$y_j = y_j \odot \tanh (\mathbf{W}_{\text{img}} \cdot V)$$

### 4 Training

We use ADAM (Kingma and Ba, 2014) with a learning rate of $4e-4$ and a batch size of 32. All weights are initialized using Xavier method (Glorot and Bengio, 2010) and the total gradient norm is clipped to 5 (Pascanu et al., 2013). Dropout (Srivastava et al., 2014) is enabled after source embeddings $X$, source annotations $S$ and pre-softmax activations $o_t$ with dropout probabilities of (0.3, 0.5, 0.5) respectively. ((0.2, 0.4, 0.4) for
An L₂ regularization term with a factor of 1e−5 is also applied to avoid overfitting unless otherwise stated. Finally, we set E=128 and R=256 (Section 3) respectively for embedding and GRU dimensions.

All models are implemented and trained with the nmtpy framework\(^2\) (Caglayan et al., 2017) using Theano v0.9 (Theano Development Team, 2016). Each experiment is repeated with 5 different seeds to mitigate the variance of BLEU (Papineni et al., 2002) and METEOR (Lavie and Agarwal, 2007) and to benefit from ensembling. The training is early stopped if validation set METEOR does not improve for 10 validations performed per 1000 updates. A beam-search with a beam size of 12 is used for translation decoding.

5 Results

All results are computed using multeval (Clark et al., 2011) with tokenized sentences.

5.1 En→De

Table 1 summarizes BLEU and METEOR scores obtained by our systems. It should be noted that since we trained each system with 5 different seeds, we report results obtained by ensembling 5 runs as well as the mean/deviation over these 5 runs. The final system to be submitted is selected based on ensemble Test2016 METEOR.

First of all, multimodal systems which use global pool5 features generally obtain comparable scores which are better than the baseline NMT in contrast to fusion-conv which fails to improve over it. Our submitted system (D6) achieves an ensembling score of 60.4 METEOR which is 1.2 better than NMT. Although the improvements are smaller, (D6) is still the best system on Test2017 in terms of ensembling/mean METEOR scores. One interesting point to be stressed at this level is that in terms of mean BLEU, (D6) performs worse than baseline on both test sets. Similarly, (D3) which has the best BLEU on Test2016, is the worst system on Test2017 according to METEOR. This is clearly a discrepancy between these metrics where an improvement in one does not necessarily yield an improvement in the other.

Table 2: MSCOCO En→De results: the best Flickr system trg-mul (Table 1) has been used for this submission as well.

For the MSCOCO set no held-out set for model selection was available. Therefore, we submitted the system (D6) with best METEOR on Flickr Test2016.
Table 3: Flickr En→Fr results: Scores are averages over 5 runs and given with their standard deviation (σ) and the score obtained by ensembling the 5 runs. *ens-nmt-7* and *ens-mmt-6* are the submitted ensembles which correspond to the combination of 7 monomodal and 6 multimodal (global pool5) systems, respectively.

After scoring all the available systems (Table 2) we observe that (D4) is the best system according to ensemble metrics. This can be explained by the out-of-domain/ambiguous nature of MSCOCO where best generalization performance on Flickr is not necessarily transferred to this set.

Overall, (D4), (D5) and (D6) are the top systems according to METEOR on Flickr and MSCOCO test sets.

### 5.2 En→Fr

Table 5.1 shows the results of our systems on the official test set of last year (Test2016) and this year (Test2017). F1 is a variant of the baseline NMT without $L_2$ regularization. F2 is a multimodal system using convolutional feature maps as visual features while F3 to F5 are multimodal systems using pool5 global visual features. We note that all multimodal systems perform better than monomodal ones.

Compared to the MMT 2016 results, we can see that the fusion-conv (F2) system with separate attention over both modalities achieve better performance than monomodal systems. The results are further improved by systems F3 to F5 which use pool5 global visual features. We conjecture that the way of integrating the global visual features into these systems does not seem to affect the final results since they all perform equally well on both test sets.

The submitted systems are presented in the last two lines of Table 5.1. Since we did not have all 5 runs with different seeds ready by the submission deadline, heterogeneous ensembles of different architectures and different seeds were considered. *ens-nmt-7* (contrastive monomodal submission) and *ens-mmt-6* (primary multimodal submission) correspond to ensembles of 7 monomodal and 6 multimodal (pool5) systems respectively. *ens-mmt-6* benefits from the heterogeneity of the included systems resulting in a slight improvement of BLEU and METEOR.

Table 4: MSCOCO En→Fr results: **ens-mmt-6**, the best performing ensemble on Test2016 corpus (see Table 5.1) has been used for this submission as well.

Results on the ambiguous dataset extracted from MSCOCO are presented in Table 4. We can observe a slightly different behaviour compared to the results in Table 5.1. The systems using the convolutional features are performing equally well compared to those using pool5 features. One should note that no specific tuning was performed for this additional task since no specific validation data was provided.
6 Conclusion

We have presented the LIUM-CVC systems for English to German and English to French Multimodal Machine Translation evaluation campaign. Our systems were ranked first for both tasks in terms of automatic metrics. Using the pool5 global visual features resulted in a better performance compared to multimodal attention architecture which makes use of convolutional features. This might be explained by the fact that the attention mechanism over spatial feature vectors cannot capture useful information from the extracted features maps. Another explanation for this is that source sentences contain most necessary information to produce the translation and the visual content is only useful to disambiguate a few specific cases. We also believe that reducing the number of parameters aggressively to around 5M allowed us to avoid overfitting leading to better scores in overall.

Acknowledgments

This work was supported by the French National Research Agency (ANR) through the CHIST-ERA M2CR project3, under the contract number ANR-15-CHR2-0006-01 and by MINECO through APCIN 2015 under the contract number PCIN-2015-251.

References


http://m2cr.univ-lemans.fr


