LIUM-CVC Submissions for WMT18 Multimodal Translation Task
Ozan Caglayan, Adrien Bardet, Fethi Bougares, Loïc Barrault, Kai Wang,
Marc Masana, Luis Herranz, Joost van de Weijer

To cite this version:

HAL Id: hal-02498133
https://hal.archives-ouvertes.fr/hal-02498133
Submitted on 4 Mar 2020

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 WMT18 Multimodal Translation Task
Ozan Caglayan, Adrien Bardet,
Fethi Bougares, Loïc Barrault
LIUM, Le Mans University
FirstName.LastName@univ-lemans.fr

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

Abstract
This paper describes the multimodal Neural Machine Translation systems developed by LIUM and CVC for WMT18 Shared Task on Multimodal Translation. This year we propose several modifications to our previous multimodal attention architecture in order to better integrate convolutional features and refine them using encoder-side information. Our final constrained submissions ranked first for English→French and second for English→German language pairs among the constrained submissions according to the automatic evaluation metric METEOR.

1 Introduction
In this paper, we present the neural machine translation (NMT) and multimodal NMT (MMT) systems developed by LIUM and CVC for the third edition of the shared task. Several lines of work have been conducted since the introduction of the shared task on MMT in 2016 (Specia et al., 2016). The majority of last year submissions including ours (Caglayan et al., 2017a) were based on the integration of global visual features into various parts of the NMT architecture (Elliott et al., 2017). Apart from these, hierarchical multimodal attention (Helcl and Libovický, 2017) and multi-task learning (Elliott and Kádár, 2017) were also explored by the participants.

This year we decided to revisit the multimodal attention (Caglayan et al., 2016) since our previous observations about qualitative analysis of the visual attention was not satisfying. In order to improve the multimodal attention both qualitatively and quantitatively, we experiment with several refinements to it: first, we try to use different input image sizes prior to feature extraction and second we normalize the final convolutional feature maps to assess its impact on the final MMT performance. In terms of architecture, we propose to refine the visual features by learning an encoder-guided early spatial attention. In overall, we find that normalizing feature maps is crucial for the multimodal attention to obtain a comparable performance to monomodal NMT while the impact of the input image size remains unclear. Finally, with the help of the refined attention, we obtain modest improvements in terms of BLEU (Papineni et al., 2002) and METEOR (Lavie and Agarwal, 2007).

The paper is organized as follows: data preprocessing, model details and training hyperparameters are detailed respectively in section 2 and section 3. The results based on automatic evaluation metrics are reported in section 4. Finally the paper ends with a conclusion in section 5.

2 Data
We use Multi30k (Elliott et al., 2016) dataset provided by the organizers which contains 29000, 1014, 1000 and 1000 English→{German,French} sentence pairs respectively for train, dev, test2016 and test2017. A new training split of 30014 pairs is formed by concatenating the train and dev splits. Early-stopping is performed based on METEOR computed over the test2016 set and the final model selection is done over test2017. Punctuation normalization, lowercasing and aggressive hyphen splitting were applied to all sentences prior to training. A Byte Pair Encoding (BPE) model (Sennrich et al., 2016) with 10K merge operations is jointly learned on English-German and English-French resulting in vocabularies of 5189-7090 and 5830-6608 subwords respectively.

2.1 Visual Features
Since Multi30k images involve much more complex region-level relationships and scene compositions compared to ImageNet (Russakovsky et al., 2015),...
object classification task, we explore different input image sizes to quantify its impact in the context of MMT since rescaling the input image has a direct effect on the size of the receptive fields of the CNN. After normalizing the images using ImageNet mean and standard deviation, we resize and crop the images to 224x224 and 448x448. Features are then extracted from the final convolutional layer (res5c_relu) of a pre-trained ResNet50 (He et al., 2016) CNN. This led to feature maps $V \in \mathbb{R}^{2048 \times h \times w}$ where the spatial dimensionality $w$ is 7 or 14.

2.1.1 Feature Normalization

We conjecture that transferring ReLU features from a CNN into a model that only makes use of bounded non-linearities like sigmoid and tanh, can saturate the non-linear neurons in the very early stages of training if their weights are not carefully initialized. Instead of tuning the initialization, we experiment with $L_2$ normalization over the channel dimension so that each feature vector $x \in \mathbb{R}^{2048}$ has an $L_2$ norm of 1.

3 Models

In this section we will describe our baseline NMT and multimodal NMT systems. All models use 128 dimensional embeddings and GRU (Cho et al., 2014) layers with 256 hidden states. Dropout (Srivastava et al., 2014) is applied over source embeddings $x_s$, encoder states $H^{enc}$ and pre-softmax activations $o_t$. We also apply $L_2$ regularization with a factor of $1e-5$ on all parameters except biases. The parameters are initialized using the method proposed by He et al. (2015) and optimized with Adam (Kingma and Ba, 2014). The total gradient norm is clipped to 1 (Pascanu et al., 2013). We use batches of size 64 and an initial learning rate of $4$. All systems are implemented using the PyTorch version of nmtpy (Caglayan et al., 2017b).

3.1 Baseline NMT

Let us denote the length of the source sentence $\{x_1, \ldots, x_S\}$ and the target sentence $\{y_1, \ldots, y_T\}$ by $S$ and $T$ respectively. The source sentence is first encoded with a 2-layer bidirectional GRU to obtain the set of hidden states:

$$H^{enc} \leftarrow Enc(\{x_1, \ldots, x_S\}), H^{enc} \in \mathbb{R}^{S \times 512}$$

The decoder is a 2-layer conditional GRU (CGRU) (Sennrich et al., 2017) with tied embeddings (Press and Wolf, 2016). CGRU is a stacked 2-layer recurrence block with the attention mechanism in the middle. We use feed-forward attention (Bahdanau et al., 2014) which encapsulates a learnable layer. The first decoder (which is initialized with a zero vector) receives the previous target embeddings as inputs (equation 1). At each timestep of the decoding stage, the attention mechanisms produces a context vector $c_t^{txt}$ (equation 2) that becomes the input to the second GRU (equation 3). Finally, the probability over the target vocabulary is conditioned over a transformation of the final hidden state $h_t^{dec2}$ (equation 4, 5).

3.2 Multimodal Attention (MA)

Our baseline multimodal attention (MA) system (Caglayan et al., 2016) applies a spatial attention mechanism (Xu et al., 2015) over the visual features. At each timestep $t$ of the decoding stage,
a multimodal context vector $c_t$ is computed and given as input to the second decoder (equation 3):

$$c_t = W_f [c^{txt}_t; W_{vis} c^{vis}_t]$$  \hspace{1cm} (6)$$
$$c^{vis}_t = \text{ATT}_{vis}(V, h^{dec}_t)$$  \hspace{1cm} (7)$$

Previous analysis showed that the attention over the visual features is inconsistent and weak. We argue that this is because of the diluted relevant visual information, and the competition with the far more relevant source text information.

### 3.3 Filtered Attention (FA)

In order to enhance the visual attention, we propose an extension to the multimodal attention where the objective is to filter the convolutional feature maps using the last hidden state of the source language encoder (Figure 1). We conjecture that a learnable masking operation over the convolutional feature maps can help the decoder-side visual attention mechanism by filtering out regions irrelevant to translation and focus on the most important part of the visual input. The filtered convolutional feature map $\tilde{V}$ is computed as follows:

$$\beta^{pre} = \text{ConvAtt}(\text{Tile}(h^{enc}_S; V))$$  \hspace{1cm} (8)$$
$$\tilde{V} = \beta^{pre} \odot V, \beta^{pre} \in R^{1 \times w \times w}$$  \hspace{1cm} (9)$$

$\text{ConvAtt}$ block is inspired from previous works in visual question answering (VQA) (Yang et al., 2016; Kazemi and Elqursh, 2017). It basically computes a spatial attention distribution $\beta^{pre}$ which we further use to mask the actual convolutional features $V$. The filtered $\tilde{V}$ replaces $V$ in the equation 7 instead of being pooled into a single visual embedding in contrast to VQA models.

### Feature Normalization

We can see from Table 1 that without $L_2$ normalization, multimodal attention is not able to reach the performance of baseline NMT. Applying the normalization consistently improves the results for all input sizes by around ~2 points in BLEU and METEOR. From now on, we only present systems trained with normalized features.

### Image Size

Although the impact of doubling the image width and height at the input seems marginal (Table 2), we switch to 448x448 images to benefit from the slight gains which are consistent across both attention variants.

### 4 Results

We train each model 4 times using different seeds and report mean and standard deviation for the final results using multeval (Clark et al., 2011)
Table 3: EN→DE results: Filtered attention is statistically different than the NMT ($p \leq 0.02$).

Table 4: EN→FR results: multimodal systems are not able to improve over NMT in terms of automatic metrics.

Table 5: Official test2018 results (†: Unconstrained, ⋆: Constrained.)

5 Conclusion

MMT task consists of translating a source sentence into a target language with the help of an image representing the source sentence. The different level of relevance of both input modalities makes it a difficult task where the image should be used with parsimony. With the aim of improving the attention over visual input, we introduced a filtering technique to allow the network to ignore irrelevant parts of the image that should not be considered during decoding. This is done by using an attention-like mechanism between the source sentence and the convolutional feature maps. Results show that this mechanism significantly improves the results for English→German on one of the test sets. In the future, we plan to qualitatively analyze the spatial attention and try to improve it further.

Acknowledgments

This work was supported by the French National Research Agency (ANR) through the CHISTERA M2CR project\(^4\), under the contract number ANR-15-CHR2-0006-01 and by MINECO through APCIN 2015 under the contract number PCIN-2015-251.

References


\(^4\)http://m2cr.univ-lemans.fr


