UMONS Submission for WMT18 Multimodal Translation Task

Jean-Benoit Delbrouck and **Stéphane Dupont**
TCTS Lab, University of Mons, Belgium
{jean-benoit.delbrouck, stephane.dupont}@umons.ac.be

Abstract

This paper describes the UMONS solution for the Multimodal Machine Translation Task presented at the third conference on machine translation (WMT18). We explore a novel architecture, called deepGRU, based on recent findings in the related task of Neural Image Captioning (NIC). The models presented in the following sections lead to the best METEOR translation score for both constrained (English, image) → German and (English, image) → French sub-tasks.

1 Introduction

In the field of Machine Translation (MT), the efficient integration of multimodal information still remains a challenging task. It requires combining diverse modality vector representations with each other. These vector representations, also called context vectors, are computed in order the capture the most relevant information in a modality to output the best translation of a sentence.

To investigate the effectiveness of information obtained from images, a multimodal neural machine translation (MNMT) shared task (Specia et al., 2016) has been introduced to the community.[1] Even though soft attention models had been extensively studied in MNMT (Delbrouck and Dupont, 2017a; Caglayan et al., 2016; Calixto et al., 2017), the most successful recent work (Caglayan et al., 2017a) focused on using the max-pooled features extracted from a convolutional network to modulate some components of the system (i.e. the target embeddings). Convolutional features or attention maps recently showed some success (Delbrouck and Dupont, 2017b) in a encoder-based attention model conditioned on the source encoder representation. Both model types lead to similar results, the latter being slightly complex and taking longer to train. One similar feature they share is that the proposed models remain relatively small. Indeed, the number of trainable parameters seems "upper bounded" due to the number of unique training examples being limited (cfr. Section 4). Heavy or complex attention models on visual features showed premature convergence and restricted scalability.

The model proposed by the University of Mons (UMONS) in 2018 is called DeepGRU, a novel idea based on the previously investigated conditional GRU (cGRU).[2] We enrich the architecture with three ideas borrowed from the closely related NIC task: a third GRU as bottleneck function, a multimodal projection and the use of gated tanh activation. We make sure to keep the overall model light, efficient and rapid to train. We start by describing the baseline model in Section 2 followed by the three aforementioned NIC upgrades which make up our deepGRU model in Section 3. Finally, we present the data made available by the Multimodal Machine Translation Task in Section 4 and the results in section 5, then engage a quick discussion in Section 6.

2 Baseline Architecture

Given a source sentence $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M)$ and an image I, an attention-based encoder-decoder model (Bahdanau et al., 2014) outputs the translated sentence $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N)$. If we denote $\boldsymbol{\theta}$ as the model parameters, then $\boldsymbol{\theta}$ is learned by maximizing the likelihood of the observed sequence $\boldsymbol{Y}$ or in other words by minimizing the cross entropy loss. The objective function

[1] http://www.statmt.org/wmt18/multimodal-task.html

[2] https://github.com/nyu-dl/dl4mt-tutorial/blob/master/docs/cgru.pdf

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 643–647

https://doi.org/10.18653/v1/W18-6444

is given by:

$$\mathcal{L}(\boldsymbol{\theta}) = -\sum_{t=1}^{n} \log p_{\boldsymbol{\theta}}(\boldsymbol{y}_t | \boldsymbol{y}_{<t}, I, X) \quad (1)$$

Three main components are involved: an encoder, a decoder and an attention model.

Encoder At every time-step t, an encoder creates an annotation $\boldsymbol{h}_t$ according to the current embedded word $\boldsymbol{x}_t$ and internal state $\boldsymbol{h}_{t-1}$:

$$\boldsymbol{h}_t = f_{\text{enc}}(\boldsymbol{x}'_t, \boldsymbol{h}_{t-1}) \quad (2)$$

Every word $\boldsymbol{x}_t$ of the source sequence $\boldsymbol{X}$ is an index in the embedding matrix $\boldsymbol{E}^x$ so that the following formula maps the word to the f_{enc} size S:

$$\boldsymbol{x}'_t = \boldsymbol{W}^x \boldsymbol{E}^x x_t \quad (3)$$

The total size of the embeddings matrix $\boldsymbol{E}^x$ depends on the source vocabulary size $|\mathcal{Y}_s|$ and the embedding dimension d such that $\boldsymbol{E}^x \in \mathbb{R}^{|\mathcal{Y}_s| \times d}$. The mapping matrix $\boldsymbol{W}^x$ also depends on the embedding dimension because $\boldsymbol{W}^x \in \mathbb{R}^{d \times S}$.

The encoder function f_{enc} is a bi-directional GRU (Cho et al., 2014). The following equations define a single GRU block (called f_{gru} for future references) :

$$\begin{aligned}
\boldsymbol{z}_t &= \sigma\left(\boldsymbol{x}'_t + \boldsymbol{W}^z \boldsymbol{h}_{t-1}\right) \\
\boldsymbol{r}_t &= \sigma\left(\boldsymbol{x}'_t + \boldsymbol{W}^r \boldsymbol{h}_{t-1}\right) \\
\underline{\boldsymbol{h}}_t &= \tanh\left(\boldsymbol{x}'_t + \boldsymbol{r}_t \odot (\boldsymbol{W}^h \boldsymbol{h}_{t-1})\right) \\
\boldsymbol{h}'_t &= (1 - \boldsymbol{z}_t) \odot \underline{\boldsymbol{h}}_t + \boldsymbol{z}_t \odot \boldsymbol{h}_{t-1}
\end{aligned} \quad (4)$$

where $\boldsymbol{h}'_t \in \mathbb{R}^S$. Our encoder consists of two GRUs, one is reading the source sentence from 1 to M and the second from M to 1. The final encoder annotation $\boldsymbol{h}_t$ for timestep t becomes the concatenation of both GRUs annotations $\boldsymbol{h}'_t$. Therefore, the encoder set of annotations $\boldsymbol{H}$ is of size $M \times 2S$.

Decoder At every time-step t, a decoder outputs probabilities $\boldsymbol{p}_t$ over the target vocabulary $\mathcal{Y}_d$ according to previously generated word $\boldsymbol{y}_{t-1}$, internal state $\boldsymbol{s}_{t-1}$ and image I:

$$y_t \sim \boldsymbol{p}_t = f_{\text{bot}}\big(\ f_{\text{dec}}(\boldsymbol{y}_{t-1}, \boldsymbol{s}_{t-1}, I)\ \big) \quad (5)$$

Every word $\boldsymbol{y}_t$ of the target sequence $\boldsymbol{Y}$ is an index in the embedding matrix $\boldsymbol{E}^y$ so that the following formula maps the word in the f_{dec} size D:

$$\boldsymbol{y}'_t = \boldsymbol{W}^y \boldsymbol{E}^y \boldsymbol{y}_{t-1} \quad (6)$$

The decoder function f_{dec} is a conditional GRU (cGRU). The following equations describes a cGRU cell :

$$\begin{aligned}
\boldsymbol{s}'_t &= f_{\text{gru}_1}(\boldsymbol{y}'_t, \boldsymbol{s}_{t-1}) \\
\boldsymbol{c}_t &= f_{\text{att}}(\boldsymbol{s}'_t, I, \boldsymbol{H}) \\
\boldsymbol{s}_t &= f_{\text{gru}_2}(\boldsymbol{s}'_t, \boldsymbol{c}_t)
\end{aligned} \quad (7)$$

where f_{att} is the visual attention module over the set of source annotation $\boldsymbol{H}$ and pooled vector $\boldsymbol{v}$ of ResNet-50 features extracted from image I. More precisely, our attention model is the product between the so-called soft attention over the M source annotations $\boldsymbol{h}_{\{0,\dots,M-1\}}$ and the linear transformation over pooled vector v of image I :

$$\boldsymbol{a}'_t = \boldsymbol{W}^a \tanh(\boldsymbol{W}^{\text{s}} \boldsymbol{s}'_t + \boldsymbol{W}^{\text{H}} \boldsymbol{H}) \quad (8)$$

$$\boldsymbol{a}_t = \text{softmax}(\boldsymbol{a}'_t) \quad (9)$$

$$\boldsymbol{c}'_t = \sum_{i=0}^{M-1} \boldsymbol{a}_{t_i} \boldsymbol{h}_i \quad (10)$$

$$\boldsymbol{v}_t = \tanh(\boldsymbol{W}^{\text{img}} I) \quad (11)$$

$$\boldsymbol{c}_t = \boldsymbol{W}^c \boldsymbol{c}'_t \odot \boldsymbol{v}_t \quad (12)$$

The bottleneck function f_{bot} projects the cGRU output into probabilities over the target vocabulary. It is defined so:

$$\boldsymbol{b}_t = \tanh(\boldsymbol{W}^{\text{bot}}[\boldsymbol{y}_{t-1}, \boldsymbol{s}_t, \boldsymbol{c}_t]) \quad (13)$$

$$y_t \sim \boldsymbol{p}_t = \text{softmax}(\boldsymbol{W}^{\text{proj}} \boldsymbol{b}_t) \quad (14)$$

where $[\cdot, \cdot]$ denotes the concatenation operation.

3 DeepGRU

The deepGRU decoder (Delbrouck and Dupont, 2018) is a variant of the cGRU decoder.

Gated hyperbolic tangent First, we make use of the gated hyperbolic tangent activation (Teney et al., 2017) instead of tanh. This non-linear layer

implements a function $f_{\text{ght}} : x \in \mathbb{R}^n \rightarrow y \in \mathbb{R}^m$ with parameters defined as follows:

$$\begin{aligned} \boldsymbol{y}' &= \tanh(\boldsymbol{W}^t \boldsymbol{x} + b) \\ \boldsymbol{g} &= \sigma(\boldsymbol{W}^g \boldsymbol{x} + b) \\ \boldsymbol{y} &= \boldsymbol{y}' \odot \boldsymbol{g} \end{aligned} \tag{15}$$

where $\boldsymbol{W}^x, \boldsymbol{W}^g \in \mathbb{R}^{n \times m}$. We apply this gating system for equation 11 and 13.

GRU bottleneck When working with small dimensions, one can afford to replace the computation of $\boldsymbol{b}_t$ of equation 13 by a new gru block f_{gru}:

$$\boldsymbol{b}_t^v = f_{ght}\big(\boldsymbol{W}_{\text{bot}}^v\big(f_{\text{gru}_3}([\boldsymbol{y}_{t-1}, \boldsymbol{s}'_t, \boldsymbol{v}_t], \boldsymbol{s}_t)\big)\big) \tag{16}$$

The GRU bottleneck can be seen as a new block f_{gru_3} encoding the visual information $\boldsymbol{v}_t$ with its surrounding context ($\boldsymbol{y}_{t-1}$ and $\boldsymbol{s}'_t$). Therefore, equation 12 is not computed with $\boldsymbol{v}_t$ anymore so that the second block f_{gru_2} encodes the textual information only.

Multimodal projection Because we now have a linguistic GRU block and a visual GRU block, we want both representations to have their own projection to compute the candidate probabilities. Equation 13 and 14 becomes:

$$\boldsymbol{b}_t^t = f_{ght}(\boldsymbol{W}_{\text{bot}}^t \, \boldsymbol{s}_t) \tag{17}$$

$$y_t \sim \boldsymbol{p}_t = \text{softmax}(\boldsymbol{W}_{\text{proj}}^t \, \boldsymbol{b}_t^t + \boldsymbol{W}_{\text{proj}}^v \, \boldsymbol{b}_t^v) \tag{18}$$

where $\boldsymbol{b}_t^v$ comes from equation 16. Note that we use the gated hyperbolic tangent for equation 16 and 17.

4 Data and settings

The Multi30K dataset (Elliott et al., 2016) is provided by the challenge. For each image, one of the English descriptions was selected and manually translated into German and French by a professional translator. As training and development data, 29,000 and 1,014 triples are used respectively. We use the three available test sets to score our models. The Flickr Test2016 and the Flickr Test2017 set contain 1000 image-caption pairs and the ambiguous MSCOCO test set (Elliott et al., 2017) 461 pairs. For the WMT18 challenge, a new Flickr Test2018 set of 1,071 sentences is released without the German and French gold translations.

Marices of the model are initialized using the Xavier method (Glorot and Bengio, 2010) and the gradient norm is clipped to 5. We chose ADAM (Kingma and Ba, 2014) as the optimizer with a learning rate of 0.0004 and batch-size 32. To marginally reduce our vocabulary size, we use the byte pair encoding (BPE) algorithm on the train set to convert space-separated tokens into sub-words (Sennrich et al., 2016). With 10K merge operations, the resulting vocabulary sizes of each language pair are: 5204 $\rightarrow$ 7067 tokens for English$\rightarrow$ German and 5835$\rightarrow$ 6577 tokens for English$\rightarrow$French.

We use the following regularization methods: we apply dropout of 0.3 on source embeddings $\boldsymbol{x}'$, 0.5 on source annotations $\boldsymbol{H}$ and 0.5 on both bottlenecks $\boldsymbol{b}_t^t$ and $\boldsymbol{b}_t^v$. We also stop training when the METEOR score does not improve for 10 evaluations on the validation set (i.e. one validation is performed every 1000 model updates).

The dimensionality of the various settings and layers is as follows:
Embedding size d is 128, encoder and decoder GRU size S is 256, embedding layers are: $[\boldsymbol{W}^x, \boldsymbol{W}^y \in \mathbb{R}^{128 \times 256}$, $\boldsymbol{H} = M \times 512$, $\boldsymbol{E}^x \in \mathbb{R}^{\mathcal{Y}_s \times 128}, \boldsymbol{E}^y \in \mathbb{R}^{\mathcal{Y}_d \times 128}]$.

Attention matrices: $[\boldsymbol{W}^s \in \mathbb{R}^{256 \times 512}, \boldsymbol{W}^H \in \mathbb{R}^{512 \times 512}, \boldsymbol{W}^a \in \mathbb{R}^{512 \times 1}, \boldsymbol{W}^c \in \mathbb{R}^{512 \times 256}, \boldsymbol{W}^{\text{img}} \in \mathbb{R}^{2048 \times 256}]$.

Bottleneck matrices: $[\boldsymbol{W}_{\text{bot}}^t, \boldsymbol{W}_{\text{bot}}^v \in \mathbb{R}^{256 \times 128}]$ and projection matrices: $[\boldsymbol{W}_{\text{proj}}^t, \boldsymbol{W}_{\text{proj}}^v \in \mathbb{R}^{128 \times \mathcal{Y}_d}]$. Weights $\boldsymbol{E}^y$ and $\boldsymbol{W}_{\text{proj}}^t$ are tied.

The size of gated hyperbolic tangent weights $\boldsymbol{W}^t, \boldsymbol{W}^g$ depends on their respective application.

5 Results

Our models performance are evaluated according to the following automated metrics: BLEU-4 (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014). We decode with a beam-search of size 12 and use model ensembling of size 5 for German and 6 for French. We used the

nmtpytorch (Caglayan et al., 2017b) framework for all our experiments. We also release our code.[3]

Test sets	BLEU	METEOR
Test 2016 Flickr		
FR-Baseline	59.08	74.73
FR-DeepGRU	62.49 +3.41	76.83 +2.10
DE-Baseline	38.43	58.37
DE-DeepGRU	40.34 +1.91	59.58 +1.21
Test 2017 Flickr		
FR-Baseline	51.86	72.75
FR-DeepGRU	55.13 +3.27	71.52 +1.98
DE-Baseline	30.80	52.33
DE-DeepGRU	32.57 +1.77	53.60 +1.27
Test 2017 COCO		
FR-Baseline	43.31	64.39
FR-DeepGRU	46.16 +2.85	65.79 +1.40
DE-Baseline	26.30	48.45
DE-DeepGRU	29.21 +2.91	49.45 +1.00
Test 2018 Flickr		
FR-DeepGRU	39.40	60.17
DE-DeepGRU	31.10	51.64

6 Conclusion and future work

The full leaderboard scores [4] shows close results and it seems that everybody converges towards the same translation quality score. A few questions arise. Did we reach —to some extent— the full potential of images related to the information they can provide? Should we try and add traditional machine translation techniques such as post-edition, since images have been exploited successfully? Another major step forward would be to successfully develop strong and stable models using convolutional features, the latter having 98 times more features than the max-pooled ones.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost van de Weijer. 2017a. Lium-cvc submissions for wmt17 multimodal translation task. In *Proceedings of the Second Conference on Machine Translation*, pages 432–439. Association for Computational Linguistics.

Ozan Caglayan, Walid Aransa, Yaxing Wang, Marc Masana, Mercedes Garcia-Martinez, Fethi Bougares, Loic Barrault, and Joost van de Weijer. 2016. Does multimodality help human and machine for translation and image captioning? *arXiv preprint arXiv:1605.09186*.

Ozan Caglayan, Mercedes García-Martínez, Adrien Bardet, Walid Aransa, Fethi Bougares, and Loïc Barrault. 2017b. Nmtpy: A flexible toolkit for advanced neural machine translation systems. *Prague Bull. Math. Linguistics*, 109:15–28.

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. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Çalar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Jean-Benoit Delbrouck and Stéphane Dupont. 2017a. An empirical study on the effectiveness of images in multimodal neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 910–919, Copenhagen, Denmark. Association for Computational Linguistics.

Jean-Benoit Delbrouck and Stéphane Dupont. 2017b. Modulating and attending the source image during encoding improves multimodal translation. *CoRR*, abs/1712.03449.

Jean-Benoit Delbrouck and Stéphane Dupont. 2018. Bringing back simplicity and lightliness into neural image captioning. *CoRR*.

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*.

D. Elliott, S. Frank, K. Sima'an, and L. Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74.

[3] https://github.com/jbdel/WMT18_MNMT

[4] https://competitions.codalab.org/competitions/19917

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*, Copenhagen, Denmark.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS10). Society for Artificial Intelligence and Statistics*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics.

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*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.

Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. 2017. Tips and tricks for visual question answering: Learnings from the 2017 challenge. *CoRR*, abs/1708.02711.