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 FC_7 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 Kádár, 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 hyperparameters 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 ex-

---

tract 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* $\mathbf{S}$ where each element s_i is a vector of dimension $C = 2 \times R$:

$$\mathbf{S} = \begin{bmatrix} \text{GRU}_{\text{Forw}}(\overrightarrow{X}) \\ \text{GRU}_{\text{Back}}(\overleftarrow{X}) \end{bmatrix} \in \mathbb{R}^{M \times C}$$

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:

$$h_0 = \tanh\left(\mathbf{W_{init}} \cdot \frac{1}{M}\sum_i^M s_i\right), s_i \in \mathbf{S} \quad (1)$$

Attention At each decoding timestep t, an unnormalized attention score g_i is computed for each source annotation s_i using the first GRU's hidden state h_t and s_i itself:
($\mathbf{W}_a \in \mathbb{R}^C$, $\mathbf{W_s} \in \mathbb{R}^{C \times C}$ and $\mathbf{W_h} \in \mathbb{R}^{C \times R}$)

$$g_i = \mathbf{W}_a^{\text{T}} \tanh(\mathbf{W_s}s_i + b_s + \mathbf{W_h}h_t) + b_a \quad (2)$$

The context vector c_t is a weighted sum of s_i and its respective attention probability α_i obtained using a softmax operation over all the unnormalized scores:

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

$$c_t = \sum_i^M \alpha_i s_i$$

The final hidden state $\widetilde{h}_t$ 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 $\widetilde{h}_t$ and the context vector c_t, the latter two transformed with $\mathbf{W_{dec}}$ and $\mathbf{W_{ctx}}$ respectively:

$$o_t = \tanh(y_{t-1} + \mathbf{W_{dec}}\widetilde{h}_t + \mathbf{W_{ctx}}c_t)$$

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

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:

$$g'_j = \mathrm{W'_a}^{\mathrm{T}} \tanh\left(\mathbf{W_s}'s'_j + b'_s + \mathbf{W_h}'h_t\right) + b'_a$$

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

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

$$v_t = \sum_j^{196} \beta_j s'_j$$

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:

$$h_0 = \tanh\left(\mathbf{W_{img}} \cdot V\right)$$

(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) :

$$e_0 = h_0 = \tanh\left(\mathbf{W_{img}} \cdot V\right)$$

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

$$s_i = s_i \odot \tanh\left(\mathbf{W_{img}} \cdot V\right)$$

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

$$y_j = y_j \odot \tanh\left(\mathbf{W_{img}} \cdot V\right)$$

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

$$h_0 = \tanh\left(\mathbf{W_{img}} \cdot V\right)$$
$$s_i = s_i \odot \tanh\left(\mathbf{W'_{img}} \cdot V\right)$$
$$y_j = y_j \odot \tanh\left(\mathbf{W''_{img}} \cdot V\right)$$

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 $\mathbf{S}$ and pre-softmax activations o_t with dropout probabilities of $(0.3, 0.5, 0.5)$ respectively. $((0.2, 0.4, 0.4)$ for

En→De Flickr	# Params	Test2016 ($\mu \pm \sigma$/Ensemble) BLEU	METEOR	Test2017 ($\mu \pm \sigma$/Ensemble) BLEU	METEOR
Caglayan et al. (2016a)	62.0M	29.2	48.5		
Huang et al. (2016)	-	36.5	54.1		
Calixto et al. (2017a)	213M	36.5	55.0		
Calixto et al. (2017b)	-	37.3	55.1		
Elliott and Kádár (2017)	-	36.8	55.8		
Baseline NMT	4.6M	38.1 ± 0.8 / 40.7	57.3 ± 0.5 / 59.2	30.8 ± 1.0 / 33.2	51.6 ± 0.5 / 53.8
(D1) fusion-conv	6.0M	37.0 ± 0.8 / 39.9	57.0 ± 0.3 / 59.1	29.8 ± 0.9 / 32.7	51.2 ± 0.3 / 53.4
(D2) dec-init-ctx-trg-mul	6.3M	38.0 ± 0.9 / 40.2	57.3 ± 0.3 / 59.3	30.9 ± 1.0 / 33.2	51.4 ± 0.3 / 53.7
(D3) dec-init	5.0M	38.8 ± 0.5 / 41.2	57.5 ± 0.2 / 59.4	31.2 ± 0.7 / 33.4	51.3 ± 0.3 / 53.2
(D4) encdec-init	5.0M	38.2 ± 0.7 / 40.6	57.6 ± 0.3 / 59.5	31.4 ± 0.4 / 33.5	<u>51.9</u> ± 0.4 / 53.7
(D5) ctx-mul	4.6M	38.4 ± 0.3 / 40.4	<u>57.8</u> ± 0.5 / 59.6	31.1 ± 0.7 / 33.5	<u>51.9</u> ± 0.2 / 53.8
(D6) trg-mul	4.7M	37.8 ± 0.9 / 41.0	<u>57.7</u> ± 0.5 / **60.4**	30.7 ± 1.0 / 33.4	<u>52.2</u> ± 0.4 / **54.0**

Table 1: Flickr En→De results: underlined METEOR scores are from systems significantly different (p-value ≤ 0.05) than the baseline using the approximate randomization test of *multeval* for 5 runs. **(D6)** is the official submission of LIUM-CVC.

En→Fr.) An L_2 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 compara-

ble 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.

En→De	MSCOCO ($\mu \pm \sigma$/Ensemble) BLEU	METEOR
Baseline NMT	26.4 ± 0.2 / 28.7	46.8 ± 0.7 / 48.9
(D1) fusion-conv	25.1 ± 0.7 / 28.0	46.0 ± 0.6 / 48.0
(D2) dec-init-ctx-trg-mul	26.3 ± 0.9 / 28.8	46.5 ± 0.4 / 48.5
(D3) dec-init	26.8 ± 0.5 / 28.8	46.5 ± 0.6 / 48.4
(D4) encdec-init	27.1 ± 0.9 / 29.4	47.2 ± 0.6 / **49.2**
(D5) ctx-mul	27.0 ± 0.7 / 29.3	47.1 ± 0.7 / 48.7
(D6) **trg-mul**	26.4 ± 0.9 / 28.5	<u>47.4</u> ± 0.3 / 48.8

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.

En→Fr	Test2016 ($\mu \pm \sigma$ / Ensemble)		Test2017 ($\mu \pm \sigma$ / Ensemble)	
	BLEU	METEOR	BLEU	METEOR
Baseline NMT	52.5 ± 0.3 / 54.3	69.6 ± 0.1 / 71.3	50.4 ± 0.9 / 53.0	67.5 ± 0.7 / 69.8
(F1) NMT + nol2reg	52.6 ± 0.8 / 55.3	69.6 ± 0.6 / 71.7	50.0 ± 0.9 / 52.5	67.6 ± 0.7 / 70.0
(F2) fusion-conv	53.5 ± 0.8 / 56.5	70.4 ± 0.6 / 72.8	51.6 ± 0.9 / 55.5	68.6 ± 0.7 / 71.7
(F3) dec-init	54.5 ± 0.8 / 56.7	71.2 ± 0.4 / 73.0	52.7 ± 0.9 / 55.5	69.4 ± 0.7 / 71.9
(F4) ctx-mul	54.6 ± 0.8 / 56.7	71.4 ± 0.6 / 73.0	52.6 ± 0.9 / 55.7	69.5 ± 0.7 / 71.9
(F5) trg-mul	54.7 ± 0.8 / 56.7	71.3 ± 0.6 / 73.0	52.7 ± 0.9 / 55.5	69.5 ± 0.7 / 71.7
ens-nmt-7	54.6	71.6	53.3	70.1
ens-mmt-6	**57.4**	**73.6**	**55.9**	**72.2**

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 differ-

ent 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.

En→Fr	MSCOCO ($\mu \pm \sigma$ / ensemble)	
	BLEU	METEOR
Baseline NMT	41.2 ±1.2 / 43.3	61.3 ±0.9 / 63.3
(F1) NMT + nol2reg	40.6 ±1.2 / 43.5	61.1 ±0.9 / 63.7
(F2) fusion-conv	43.2 ±1.2 / 45.9	63.1 ±0.9 / 65.6
(F3) dec-init	43.3 ±1.2 / 46.2	63.4 ±0.9 / 66.0
(F4) ctx-mul	43.3 ±1.2 / 45.6	63.4 ±0.9 / 65.4
(F5) trg-mul	43.5 ±1.2 / 45.5	63.2 ±0.9 / 65.1
ens-nmt-7	43.6	63.4
ens-mmt-6	45.9	65.9

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

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* http://arxiv.org/abs/1607.06450.

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

Ozan Caglayan, Walid Aransa, Yaxing Wang, Marc Masana, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, and Joost van de Weijer. 2016a. Does multimodality help human and machine for translation and image captioning? In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 627–633. http://www.aclweb.org/anthology/W/W16/W16-2358.pdf.

Ozan Caglayan, Loïc Barrault, and Fethi Bougares. 2016b. Multimodal attention for neural machine translation. *CoRR* abs/1609.03976. http://arxiv.org/abs/1609.03976.

Ozan Caglayan, Mercedes García-Martínez, Adrien Bardet, Walid Aransa, Fethi Bougares, and Loïc Barrault. 2017. Nmtpy: A flexible toolkit for advanced neural machine translation systems. *arXiv preprint arXiv:1706.00457* http://arxiv.org/abs/1706.00457.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017a. Doubly-attentive decoder for multimodal neural machine translation. *arXiv preprint arXiv:1702.01287* http://arxiv.org/abs/1702.01287.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017b. Incorporating global visual features into attention-based neural machine translation. *arXiv preprint arXiv:1701.06521* http://arxiv.org/abs/1701.06521.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR* abs/1412.3555. http://arxiv.org/abs/1412.3555.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '11, pages 176–181. http://dl.acm.org/citation.cfm?id=2002736.2002774.

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.

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*. Association for Computational Linguistics, Berlin, Germany, pages 70–74. http://anthology.aclweb.org/W16-3210.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves multimodal translation. *CoRR* abs/1705.04350. http://arxiv.org/abs/1705.04350.

Orhan Firat and Kyunghyun Cho. 2016. Conditional gated recurrent unit with attention mechanism. http://github.com/nyu-dl/dl4mt-tutorial/blob/master/docs/cgru.pdf.

Orhan Firat, Kyunghyun Cho, Baskaran Sankaran, Fatos T. Yarman Vural, and Yoshua Bengio. 2017. Multi-way, multilingual neural machine translation. *Comput. Speech Lang.* 45(C):236–252. https://doi.org/10.1016/j.csl.2016.10.006.

---

[3] http://m2cr.univ-lemans.fr

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. PMLR, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256. http://proceedings.mlr.press/v9/glorot10a.html.

K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pages 770–778. https://doi.org/10.1109/CVPR.2016.90.

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*. Association for Computational Linguistics, Berlin, Germany, pages 639–645. http://www.aclweb.org/anthology/W16-2360.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Vigas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. Technical report, Google. https://arxiv.org/abs/1611.04558.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* http://arxiv.org/abs/1412.6980.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Stroudsburg, PA, USA, StatMT '07, pages 228–231. http://dl.acm.org/citation.cfm?id=1626355.1626389.

Jindrich Libovický and Jindrich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. *CoRR* abs/1704.06567. http://arxiv.org/abs/1704.06567.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114* http://arxiv.org/abs/1511.06114.

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*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '02, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*. JMLR.org, ICML'13, pages III–1310–III–1318. http://dl.acm.org/citation.cfm?id=3042817.3043083.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*. MIT Press, Cambridge, MA, USA, NIPS'15, pages 91–99. http://dl.acm.org/citation.cfm?id=2969239.2969250.

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. https://doi.org/10.1007/s11263-015-0816-y.

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)*. Association for Computational Linguistics, Berlin, Germany, pages 1715–1725. http://www.aclweb.org/anthology/P16-1162.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* http://arxiv.org/abs/1409.1556.

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*. Association for Computational Linguistics, Berlin, Germany, pages 543–553. http://www.aclweb.org/anthology/W/W16/W16-2346.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15(1):1929–1958. http://dl.acm.org/citation.cfm?id=2627435.2670313.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*. MIT Press, Cambridge, MA, USA, NIPS'14, pages 3104–3112. http://dl.acm.org/citation.cfm?id=2969033.2969173.

Theano Development Team. 2016. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints* abs/1605.02688. http://arxiv.org/abs/1605.02688.

Antonio Toral and Víctor M. Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, pages 1063–1073. http://www.aclweb.org/anthology/E17-1100.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*. JMLR Workshop and Conference Proceedings, pages 2048–2057. http://jmlr.org/proceedings/papers/v37/xuc15.pdf.