

An empirical study on the effectiveness of images in Multimodal Neural Machine Translation

Anonymous EMNLP submission

Abstract

In state-of-the-art Neural Machine Translation (NMT), an attention mechanism is used during decoding to enhance the translation. At every step, the decoder uses this mechanism to focus on different parts of the source sentence to gather the most useful information before outputting its target word. Recently, the effectiveness of the attention mechanism has also been explored for multimodal tasks, where it becomes possible to focus both on sentence parts and image regions that they describe. In this paper, we compare several attention mechanism on the multimodal translation task (English, image \rightarrow German) and evaluate the ability of the model to make use of images to improve translation. We surpass state-of-the-art scores on the Multi30k data set, we nevertheless identify and report different misbehavior of the machine while translating.

1 Introduction

In machine translation, neural networks have attracted a lot of research attention. Recently, the attention-based encoder-decoder framework (Sutskever et al., 2014; Bahdanau et al., 2014) has been largely adopted. In this approach, Recurrent Neural Networks (RNNs) map source sequences of words to target sequences. The attention mechanism is learned to focus on different parts of the input sentence while decoding. Attention mechanisms have shown to work with other modalities too, like images, where they are able to learn to attend the salient parts of an image, for instance when generating text captions (Xu et al., 2015). For such applications, Convolutional Neural Networks (CNNs) such as Deep Residual (He et al.,

2016) have shown to work best to represent images.

Multimodal models of texts and images empower new applications such as visual question answering or multimodal caption translation. Also, the grounding of multiple modalities against each other may enable the model to have a better understanding of each modality individually, such as in natural language understanding applications.

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 to 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 machine translation shared task (Specia et al., 2016) has been addressed to the MT community¹. The best results of NMT model were those of Huang et al. (2016) who used LSTM fed with global visual features or multiple regional visual features followed by rescoring. Recently, Calixto et al. (2017) proposed a doubly-attentive decoder that outperformed this baseline with less data and without rescoring.

Our paper is structured as follows. In section 2, we briefly describe our NMT model as well as the conditional GRU activation used in the decoder. We also explain how multi-modalities can be implemented within this framework. In the following sections (3 and 4), we detail three attention mechanisms and explain how we tweak them to work as well as possible with images. Finally, we report and analyze our results in section 5 then conclude in section 6.

¹<http://www.statmt.org/wmt16/multimodal-task.html>

2 Neural Machine Translation

In this section, we detail the neural machine translation architecture by Bahdanau et al. (2014), implemented as an attention-based encoder-decoder framework with recurrent neural networks (§2.1). We follow by explaining the conditional GRU layer (§2.2) - the gating mechanism we chose for our RNN - and how the model can be ported to a multimodal version (§2.3).

2.1 Text-based NMT

Given a source sentence $X = (x_1, x_2, \dots, x_M)$, the neural network directly models the conditional probability $p(Y|X)$ of its translation $Y = (y_1, y_2, \dots, y_N)$. The network consists of one encoder and one decoder with one attention mechanism. The encoder computes a representation C for each source sentence and a decoder generates one target word at a time and by decomposing the following conditional probability :

$$\log p(Y|X) = \sum_{t=1}^n \log p(y_t | y < t, C) \quad (1)$$

Each source word x_i and target word y_i are a column index of the embedding matrix E_X and E_Y . The encoder is a bi-directional RNN with Gated Recurrent Unit (GRU) layers (Chung et al., 2014; Cho et al., 2014), where a forward RNN $\vec{\Psi}_{\text{enc}}$ reads the input sequence as it is ordered (from x_1 to x_M) and calculates a sequence of forward hidden states $(\vec{h}_1, \vec{h}_2, \dots, \vec{h}_M)$. A backward RNN $\overleftarrow{\Psi}_{\text{enc}}$ reads the sequence in the reverse order (from x_M to x_1), resulting in a sequence of backward hidden states $(\overleftarrow{h}_M, \overleftarrow{h}_{M-1}, \dots, \overleftarrow{h}_1)$. We obtain an annotation for each word x_i by concatenating the forward and backward hidden state $h_t = [\vec{h}_t; \overleftarrow{h}_t]$. Each annotation h_t contains the summaries of both the preceding words and the following words. The representation C for each source sentence is the sequence of annotations $C = (h_1, h_2, \dots, h_M)$.

The decoder is an RNN that uses a conditional GRU (cGRU, more details in §2.2) with an attention mechanism to generate a word y_t at each time-step t . The cGRU uses its previous hidden state s_{t-1} , the whole sequence of source annotations C and the previously decoded symbol y_{t-1} in order to update its hidden state s_t :

$$s_t = \text{cGRU}(s_{t-1}, y_{t-1}, C) \quad (2)$$

In the process, the cGRU also computes a time-dependent context vector c_t . Both s_t and c_t are further used to decode the next symbol. We use a deep output layer (Pascanu et al., 2014) to compute a vocabulary-sized vector :

$$o_t = L_o \tanh(L_s s_t + L_c c_t + L_w E_Y[y_{t-1}]) \quad (3)$$

where L_o, L_s, L_c, L_w are model parameters. We can parameterize the probability of decoding each word y_t as:

$$p(y_t | y_{t-1}, s_t, c_t) = \text{Softmax}(o_t) \quad (4)$$

The initial state of the decoder s_0 at time-step $t = 0$ is initialized by the following equation :

$$s_0 = f_{\text{init}}(h_M) \quad (5)$$

where f_{init} is a feedforward network with one hidden layer.

2.2 Conditional GRU

The conditional GRU ² consists of two stacked GRU activations called REC₁ and REC₂ and an attention mechanism f_{att} in between (called ATT in the footnote paper). At each time-step t , REC₁ firstly computes a hidden state proposal s_t based on the previous hidden state s_{t-1} and the previously emitted word y_{t-1} :

$$\begin{aligned} z'_t &= \sigma(W'_z E_Y[y_{t-1}] + U'_z s_{t-1}) \\ r'_t &= \sigma(W'_r E_Y[y_{t-1}] + U'_r s_{t-1}) \\ \underline{s}'_t &= \tanh(W'_s E_Y[y_{t-1}] + r'_t \odot (U'_s s_{t-1})) \\ s'_t &= (1 - z'_t) \odot \underline{s}'_t + z'_t \odot s_{t-1} \end{aligned} \quad (6)$$

Then, the attention mechanism computes c_t over the source sentence using the annotations sequence C and the intermediate hidden state proposal s'_t :

$$c_t = f_{\text{att}}(C, s'_t) \quad (7)$$

() Finally, the second recurrent cell REC₂, computes the hidden state s_t of the cGRU by looking at the intermediate representation s'_t and context vector c_t :

$$\begin{aligned} z_t &= \sigma(W_z c_t + U_z s'_t) \\ r_t &= \sigma(W_r c_t + U_r s'_t) \\ \underline{s}_t &= \tanh(W_s c_t + r_t \odot (U_s s'_t)) \\ s_t &= (1 - z_t) \odot \underline{s}_t + z_t \odot s'_t \end{aligned} \quad (8)$$

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

2.3 Multimodal NMT

Recently, Calixto et al. (2017) proposed a doubly attentive decoder (referred as the "MNMT" model in the author's paper) which can be seen as an expansion of the attention-based NMT model proposed in the previous section. Given a sequence of second a modality annotations $I = (a_1, a_2, \dots, a_L)$, we also compute a new context vector based on the same intermediate hidden state proposal s'_t :

$$\mathbf{i}_t = f'_{\text{att}}(I, s'_t) \quad (9)$$

This new time-dependent context vector is an additional input to a modified version of REC2 which now computes the final hidden state s_t using the intermediate hidden state proposal s'_t and both time-dependent context vectors c_t and \mathbf{i}_t :

$$\begin{aligned} \mathbf{z}_t &= \sigma(\mathbf{W}_z c_t + \mathbf{W}_z \mathbf{i}_t + \mathbf{U}_z s'_t) \\ \mathbf{r}_t &= \sigma(\mathbf{W}_r c_t + \mathbf{W}_r \mathbf{i}_t + \mathbf{U}_r s'_t) \\ \underline{s}_t &= \tanh(\mathbf{W} c_t + \mathbf{W} \mathbf{i}_t + \mathbf{r}_t \odot (\mathbf{U} s'_t)) \\ s_t &= (1 - \mathbf{z}_t) \odot \underline{s}_t + \mathbf{z}_t \odot s'_t \end{aligned} \quad (10)$$

The probabilities for the next target word (from equation 3) also takes into account the new context vector \mathbf{i}_t :

$$L_o \tanh(L_s s_t + L_c c_t + L_i \mathbf{i}_t + L_w E_Y[y_{t-1}]) \quad (11)$$

where L_i is a new trainable parameter.

In the field of multimodal NMT, the second modality is usually an image computed into feature maps with the help of a CNN. The annotations a_1, a_2, \dots, a_L are spatial features (i.e. each annotation represents features for a specific region in the image). We follow the same protocol for our experiments and describe it in section 5.

3 Attention-based Models

We evaluate three models of the image attention mechanism f'_{att} of equation 7. They have in common the fact that at each time step t of the decoding phase, all approaches first take as input the annotation sequence I to derive a time-dependent context vector that contain relevant information in the image to help predict the current target word y_t . Even though these models differ in how the time-dependent context vector is derived, they share the same subsequent steps. For each mechanism, we propose two hand-picked illustrations showing where the attention is placed in an image.

3.1 Soft attention

Soft attention has firstly been used for syntactic constituency parsing by Vinyals et al. (2015) but has been widely used for translation tasks ever since. One should note that it slightly differs from Bahdanau et al. (2014) where their attention takes as input the previous decoder hidden state instead of the current (intermediate) one as shown in equation 7. This mechanism has also been successfully investigated for the task of image description generation (Xu et al., 2015) where a model generates an image's description in natural language. It has been used in multimodal translation as well (Calixto et al., 2017), for which it constitutes a state-of-the-art.

The idea of the soft attentional model is to consider all the annotations when deriving the context vector \mathbf{i}_t . It consists of a single feed-forward network used to compute an expected alignment e_t between modality annotation a_l and the target word to be emitted at the current time step t . The inputs are the modality annotations and the intermediate representation of REC1 s'_t :

$$e_{t,l} = v^T \tanh(\mathbf{U}_a s'_t + \mathbf{W}_a a_l) \quad (12)$$

The vector e_t has length L and its l -th item contains a score of how much attention should be put on the l -th annotation in order to output the best word at time t . We compute normalized scores to create an attention mask α_t over annotations:

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{j=1}^L \exp(e_{t,j})} \quad (13)$$

$$\mathbf{i}_t = \sum_{i=1}^L \alpha_{t,i} \mathbf{a}_i \quad (14)$$

Finally, the modality time-dependent context vector \mathbf{i}_t is computed as a weighted sum over the annotation vectors (equation 14). In the above expressions, v^T , \mathbf{U}_a and \mathbf{W}_a are trained parameters.

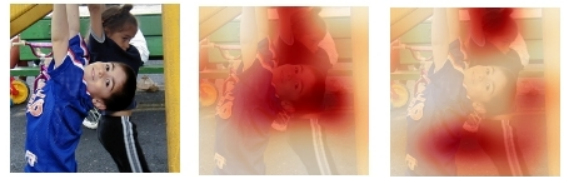


Figure 1: Die beiden Kinder spielen auf dem Spielplatz.

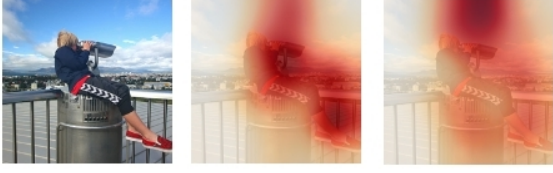


Figure 2: Ein Junge sitzt auf und blickt aus einem Mikroskop .

3.2 Hard Stochastic attention

This model is a stochastic and sampling-based process where, at every timestep t , we are making a hard choice to attend only one annotation. This corresponds to one spatial location in the image. Hard attention has previously been used in the context of object recognition (Mnih et al., 2014; Ba et al., 2015) and later extended to image description generation (Xu et al., 2015). In the context of multimodal NMT, we can follow Xu et al. (2015) because both our models involve the same process on images.

The mechanism f_{att} is now a function that returns a sampled intermediate latent variables $\gamma_{t,i}$ based upon a multinoulli distribution parameterized by α :

$$\gamma_t \sim \text{Multinoulli}(\{\alpha_{1,\dots,L}\}) \quad (15)$$

where $\gamma_{t,i}$ an indicator one-hot variable which is set to 1 if the i -th annotation (out of L) is the one used to compute the context vector i_t :

$$p(\gamma_{t,l} = 1 | \gamma < t, I) = \alpha_{t,l} \quad (16)$$

$$i_t = \sum_{i=1}^L \gamma_{t,i} a_i \quad (17)$$

Context vector i_t is now seen as the random variable of this distribution. We define the variational lower bound $\mathcal{L}(\gamma)$ on the marginal log evidence $\log p(y|I)$ of observing the target sentence y given modality annotations I .

$$\begin{aligned} \mathcal{L}(\gamma) &= \sum_{\gamma} p(\gamma|I) \log p(y|\gamma, I) \\ &\leq \log \sum_{\gamma} p(\gamma|I) p(y|\gamma, I) \\ &= \log p(y|I) \end{aligned} \quad (18)$$

The learning rules can be derived by taking derivatives of the above variational free energy

$\mathcal{L}(\gamma)$ with respect to the model parameter \mathbf{W} :

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = \sum_{\gamma} p(\gamma|I) \left[\frac{\partial \log p(y|\gamma, I)}{\partial \mathbf{W}} + \log p(y|\gamma, I) \frac{\partial \log p(\gamma|I)}{\partial \mathbf{W}} \right] \quad (19)$$

In order to propagate a gradient through this process, the summation in equation 19 can then be approximated using Monte Carlo based sampling defined by equation 16:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}} \approx \frac{1}{N} \sum_{n=1}^N \left[\frac{\partial \log p(y|\tilde{\gamma}^n, I)}{\partial \mathbf{W}} + \log p(y|\tilde{\gamma}^n, I) \frac{\partial \log p(\tilde{\gamma}^n|I)}{\partial \mathbf{W}} \right] \quad (20)$$

To reduce variance of the estimator in equation 20, we use a moving average baseline estimated as an accumulated sum of the previous log likelihoods with exponential decay upon seeing the k -th mini-batch:

$$b_k = 0.9 \times b_{k-1} + 0.1 \times \log p(y|\tilde{\gamma}_k, I) \quad (21)$$

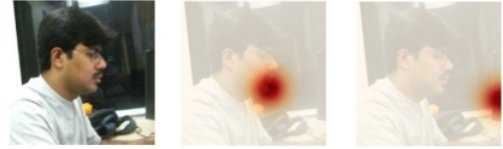


Figure 3: Ein Mann sitzt neben einem Computerbildschirm .

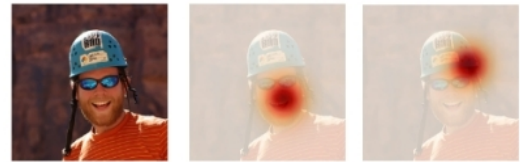


Figure 4: Ein Mann in einem orangefarbenen Hemd und mit Helm .

3.3 Local Attention

In this section, we propose a local attentional mechanism that chooses to focus only on a small

subset of the image annotations. Local Attention has been used for text-based translation (Luong et al., 2015) and is inspired by the selective attention model of Gregor et al. (2015) for image generation. Their approach allows the model to select an image patch of varying location and zoom. Local attention uses instead the same "zoom" for all target positions and still achieved good performance. This model can be seen as a trade-off between the soft and hard attentional models. The model picks one patch in the annotation sequence (one spatial location) and selectively focuses on a small window of context around it. Even though an image can't be seen as a temporal sequence, we still hope that the model finds points of interest and selects the useful information around it. This approach has an advantage of being differentiable whereas the stochastic attention requires more complicated techniques such as variance reduction and reinforcement learning to train as shown in section 3.2. The soft attention has the drawback to attend the whole image which can be difficult to learn, especially because the number of annotations L is usually large (presumably to keep a significant spatial granularity).

More formally, at every decoding step t , the model first generates an aligned position p_t . Context vector i_t is derived as a weighted sum over the annotations within the window $[p_t - D; p_t + D]$ where D is a fixed model parameter chosen empirically³. These selected annotations correspond to a squared region in the attention maps around p_t . The attention mask α_t is of size $2D + 1$. The model predicts p_t as an aligned position in the annotation sequence (referred as Predictive alignment (local-m) in the author's paper) according to the following equation:

$$p_t = S \cdot \text{sigmoid}(v^T \tanh(U_a s'_t)) \quad (22)$$

where v^T and U_a are both trainable model parameters and S is the annotation sequence length $|I|$. Because of the sigmoid, $p_t \in [0, S]$. We use equation 12 and 13 respectively to compute the expected alignment vector e_t and the attention mask α_t . In addition, a Gaussian distribution centered around p_t is placed on the alphas in order to favor

annotations near p_t :

$$\alpha_{t,i} = \alpha_{t,i} \exp\left(-\frac{(i - p_t)^2}{2\sigma^2}\right) \quad (23)$$

where standard deviation $\sigma = \frac{D}{2}$. We obtain context vector i_t by following equation 14.

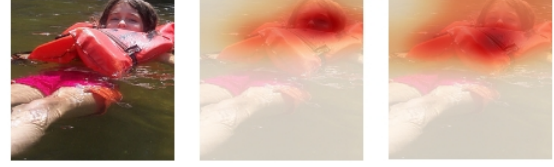


Figure 5: Ein Mädchen mit einer Schwimmweste schwimmt im Wasser .

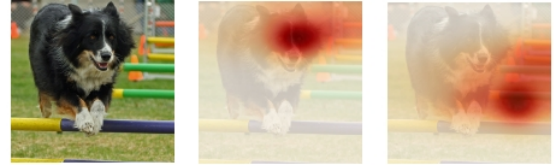


Figure 6: Ein kleiner schwarzer Hund springt über Hindernisse .

4 Image attention optimization

Three optimizations can be added to the attention mechanism regarding the image modality. All lead to a better use of the image by the model and improved the translation scores overall.

At every decoding step t , we compute a gating scalar $\beta_t \in [0, 1]$ according to the previous decoder state s_{t-1} :

$$\beta_t = \sigma(W_\beta s_{t-1} + b_\beta) \quad (24)$$

It is then used to compute the time-dependent image context vector :

$$i_t = \beta_t \sum_{l=1}^L \alpha_{t,l} a_l \quad (25)$$

Xu et al. (2015) empirically found it to put more emphasis on the objects in the image descriptions generated with their model.

We also double the output size of trainable parameters U_a , W_a and v^T in equation 12 when it comes to compute the expected annotations over the image annotation sequence. More

³We pick $D = |a_i|/4 = 49$

formally, given the image annotation sequence $I = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_L)$, $\mathbf{a}_i \in \mathbb{R}^D$, the tree matrices are of size $D \times 2D$, $D \times 2D$ and $2D \times 1$ respectively. We noticed a better coverage of the objects in the image by the alpha weights.

Lastly, we use a grounding attention inspired by [Delbrouck and Dupont \(2017\)](#). The mechanism merge each spatial location \mathbf{a}_i in the annotation sequence I with the initial decoder state \mathbf{s}_0 obtained in equation 5 with non-linearity :

$$I' = (f(\mathbf{a}_1 + \mathbf{s}_0), f(\mathbf{a}_2 + \mathbf{s}_0), \dots, f(\mathbf{a}_L + \mathbf{s}_0)) \quad (26)$$

where f is tanh function. The new annotations go through a L2 normalization layer followed by two 1×1 convolutional layers (of size $D \rightarrow 512, 512 \rightarrow 1$ respectively) to obtain $L \times 1$ weights, one for each spatial location. We normalize the weights with a softmax to obtain a soft attention map α . Each annotation \mathbf{a}_i is then weighted according to its corresponding α_i :

$$I = (\alpha_1 \mathbf{a}_1, \alpha_2 \mathbf{a}_2, \dots, \alpha_L \mathbf{a}_L) \quad (27)$$

This method can be seen as the removal of unnecessary information in the image annotations according to the source sentence. This attention is used on top of the others - before decoding - and is referred as "grounded image" in Table 1.

5 Experiments

For this experiments on Multimodal Machine Translation, we used the Multi30K dataset ([Elliott et al., 2016](#)) which is an extended version of the Flickr30K Entities. For each image, one of the English descriptions was selected and manually translated into German by a professional translator. As training and development data, 29,000 and 1,014 triples are used respectively. A test set of size 1000 is used for metrics evaluation.

5.1 Training and model details

All our models are build on top of the nematus framework ([Sennrich et al., 2017](#)). The encoder is a bidirectional RNN with GRU, one 1024D single-layer forward and one 1024D single-layer backward RNN. Word embeddings for source and target language are of 620D and trained jointly with the model. Word embeddings and other non-recurrent matrices are initialized by sampling

from a Gaussian $\mathcal{N}(0, 0.01^2)$, recurrent matrices are random orthogonal and bias vectors are all initialized to zero.

To create the image annotations used by our decoder, we used a ResNet-50 pre-trained on ImageNet and extracted the features of size $14 \times 14 \times 1024$ at its res4f layer ([He et al., 2016](#)). In our experiments, our decoder operates on the flattened 196×1024 (i.e $L \times D$). We also apply dropout with a probability of 0.5 on the embeddings, on the hidden states in the bidirectional RNN in the encoder as well as in the decoder. In the decoder, we also apply dropout on the text annotations \mathbf{h}_i , the image features \mathbf{a}_i , on both modality context vector and on all components of the deep output layer before the readout operation. We apply dropout using one same mask in all time steps ([Gal and Ghahramani, 2016](#)).

We also normalize and tokenize English and German descriptions using the Moses tokenizer scripts ([Koehn et al., 2007](#)). We use the byte pair encoding algorithm on the train set to convert space-separated tokens into subwords ([Sennrich et al., 2016](#)), reducing our vocabulary size to 9226 and 14957 words for English and German respectively.

All variants of our attention model were trained with ADADELTA ([Zeiler, 2012](#)), with mini-batches of size 80 for our monomodal (text-only) NMT model and 40 for our multimodal NMT. We apply early stopping for model selection based on BLEU4 : training is halted if no improvement on the development set is observed for more than 20 epochs. We use the metrics BLEU4 ([Papineni et al., 2002](#)), METEOR ([Denkowski and Lavie, 2014](#)) and TER ([Snover et al., 2006](#)) to evaluate the quality of our models' translations.

5.2 Quantitative results

We notice a nice overall progress over [Calixto et al. \(2017\)](#) multimodal baseline, especially when using the stochastic attention. With improvements of +1.51 BLEU and -2.2 TER on both precision-oriented metrics, the model shows a strong similarity of the n-grams of our candidate translations with respect to the references. The more recall-oriented metrics METEOR scores

Model	Test Scores					
	BLEU↑		METEOR↑		TER↓	
Monomodal (text only)						
Caglayan et al. (2016)	32.50		49.2			
Calixto et al. (2017)	33.70		52.3		46.7	
NMT	34.11	↑ +0.41	52.4	↑ +0.1	46.2	↓ -0.5
Multimodal						
Caglayan et al. (2016)	27.82		45.0		-	
Huang et al. (2016)	36.50		54.1		-	
Calixto et al. (2017)	36.50		55.0		43.7	
Soft attention	37.10	↑ +0.60	54.8	↓ -0.2	42.8	↓ -0.9
Local attention	37.55	↑ +1.05	54.8	↓ -0.2	42.4	↓ -1.3
Stochastic attention	38.01	↑ +1.51	55.4	↑ +0.4	41.5	↓ -2.2
Soft attention + grounded image	37.62	↑ +1.12	55.3	↑ +0.3	41.8	↓ -1.9
Stochastic attention + grounded image	38.17	↑ +1.67	55.4	↑ +0.4	41.5	↓ -2.2

Table 1: Results on the 1000 test triples of the Multi30K dataset. We pick Calixto et al. (2017) scores as baseline and report our results accordingly (green for improvement and red for deterioration). In each of our experiments, Soft attention is used for text. The comparison is hence with respect to the attention mechanism used for the image modality.

are roughly the same across our models which is expected because all attention mechanisms share the same subsequent step at every time-step t , i.e. taking into account the attention weights of previous time-step $t - 1$ in order to compute the new intermediate hidden state proposal and therefore the new context vector i_t . Again, the largest improvement is given by the hard stochastic attention mechanism (+0.4 METEOR): because it is modeled as a decision process according to the previous choices, this may reinforce the idea of recall. We also remark interesting improvements when using the grounded mechanism, especially for the soft attention. The soft attention may benefit more of the grounded image because of the wide range of spatial locations it looks at, especially compared to the stochastic attention. This motivates us to dig into more complex grounding techniques in order to give the machine a deeper understanding of the modalities.

Note that even though our baseline NMT model is basically the same as Calixto et al. (2017), our experiments results are slightly better. This is probably due to the different use of dropout and subwords. We also compared our results to Caglayan et al. (2016) because our multimodal models are nearly identical with the major ex-

ception of the gating scalar (cfr. section 4). This motivated some of our qualitative analysis and hesitation towards the current architecture in the next section.

5.3 Qualitative results

For space-saving and ergonomic reasons, we only discuss about the hard stochastic and soft attention, the latter being a generalization of the local attention.

As we can see in Figure 7, the soft attention model is looking roughly at the same region of the image for every decoding step t . Because the words "hund" (dog), "wald" (forest) or "weg" (way) in left image are objects, they benefit from a high gating scalar. As a matter of fact, the attention mechanism has learned to detect the objects within a scene (at every time-step, whichever word we are decoding as shown in the right image) and the gating scalar has learned to decide whether or not we have to look at the picture (or more accurately whether or not we are translating an object). Without this scalar, the translation scores undergo a massive drop (as seen in Caglayan et al. (2016)) which means that the attention mechanisms don't really understand the more complex relationships between objects, what is really happening in the scene. Surprisingly, the

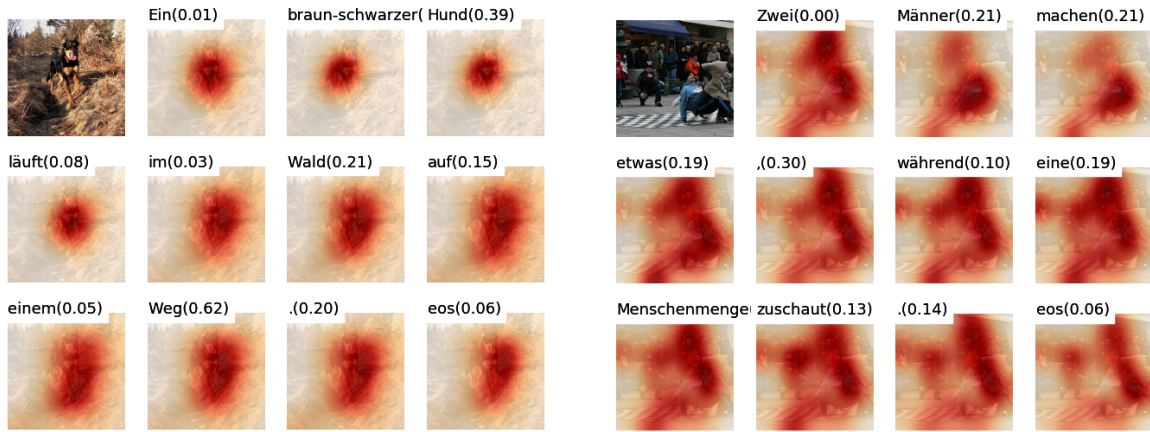


Figure 7: Representative figures of the soft-attention behavior discussed in §5.3

gating scalar happens to be really low in the stochastic attention mechanism: a significant amount of sentences don't have a summed gating scalar ≥ 0.10 . The model totally discards the image in the translation process.

It is also worth to mention that we use a ResNet trained on 1.28 million images for a classification tasks. The features used by the attention mechanism are strongly object-oriented and the machine could miss important information for a multimodal translation task. We believe that the robust architecture of both encoders $\{\overleftarrow{\Psi}_{enc}, \overrightarrow{\Psi}_{enc}\}$ combined with a GRU layer and word-embeddings took care of the right translation for relationships between objects and time-dependencies. Yet, we noticed a common misbehavior for all our multimodal models: if the attention loose track of the objects in the picture and "gets lost", the model still takes it into account and somehow overrides the information brought by the text-based annotations. The translation is then totally mislead. We illustrate with an example:

Ref:	Ein Kind sitzt auf den Schultern einer Frau und klatscht .
Mono:	Ein Kind sitzt auf den Schultern einer Frau und schläft .
Soft:	Ein Kind , das sich auf der Schultern eines Frau reitet , fährt auf den Schultern .
Hard:	Ein Kind in der Haltung , während er auf den Schultern einer Frau fährt .

The monomodal translation has a sentence-level BLEU of 82.16 whilst the soft attention and hard

stochastic attention scores are of 16.82 and 34.45 respectively. Figure 8 shows the attention maps for both mechanism. Nevertheless, one has to concede that the use of images indubitably helps the translation as shown in the score tabular.



Figure 8: Wrong detection for both Soft attention (top) and Hard stochastic attention (bottom)

6 Conclusion and future work

We have tried different attention mechanism and tweaks for the image modality. We showed improvements and encouraging results overall on the Flickr30K Entities dataset. Even though we identified some flaws of the current attention mechanisms, we can conclude pretty safely that images are an helpful resource for the machine in a translation task. We are looking forward to try out richer and more suitable features for multimodal translation (ie. dense captioning features). Another interesting approach would be to use visually grounded word embeddings to capture visual notions of semantic relatedness.

800	850
801	851
802	852
803	853
804	854
805	855
806	856
807	857
808	858
809	859
810	860
811	861
812	862
813	863
814	864
815	865
816	866
817	867
818	868
819	869
820	870
821	871
822	872
823	873
824	874
825	875
826	876
827	877
828	878
829	879
830	880
831	881
832	882
833	883
834	884
835	885
836	886
837	887
838	888
839	889
840	890
841	891
842	892
843	893
844	894
845	895
846	896
847	897
848	898
849	899

References

- Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. 2015. Multiple object recognition with visual attention. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- 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, Yaxing Wang, Marc Masana, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, and Joost van de Weijer. 2016. 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>.
- Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. *CoRR* abs/1702.01287. <http://arxiv.org/abs/1702.01287>.
- 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)*. Association for Computational Linguistics, Doha, Qatar, pages 1724–1734. <http://www.aclweb.org/anthology/D14-1179>.
- Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling.
- Jean-Benoît Delbrouck and Stéphane Dupont. 2017. Multimodal compact bilinear pooling for multimodal neural machine translation. *arXiv preprint arXiv:1703.08084* <https://arxiv.org/pdf/1703.08084.pdf>.
- 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 pages 70–74.
- Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems 29 (NIPS)*.
- Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. 2015. Draw: A recurrent neural network for image generation. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, Lille, France, volume 37 of *Proceedings of Machine Learning Research*, pages 1462–1471. <http://proceedings.mlr.press/v37/gregor15.html>.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- 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, Berlin, Germany*.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '07, pages 177–180. <http://dl.acm.org/citation.cfm?id=1557769.1557821>.
- Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.
- Volodymyr Mnih, Nicolas Heess, Alex Graves, and koray kavukcuoglu. 2014. Recurrent models of visual attention. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc., pages 2204–2212. <http://papers.nips.cc/paper/5542-recurrent-models-of-visual-attention.pdf>.
- 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, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. How to construct deep recurrent neural networks.
- Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel L'aubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: a Toolkit for Neural Machine

Translation. In *Proceedings of the Demonstrations at the 15th Conference of the European Chapter of the 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)*. <http://www.aclweb.org/anthology/P16-1162>.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*. pages 223–231.

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>.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. pages 3104–3112.

Oriol Vinyals, Łukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2015. [Grammar as a foreign language](#). In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, Curran Associates, Inc., pages 2773–2781. <http://papers.nips.cc/paper/5635-grammar-as-a-foreign-language.pdf>.

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 David Blei and Francis Bach, editors, *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>.

Matthew D. Zeiler. 2012. [ADADELTA: an adaptive learning rate method](#). *CoRR* abs/1212.5701. <http://arxiv.org/abs/1212.5701>.