# Neural Machine Translation with Deep Attention

Biao Zhang, Deyi Xiong and Jinsong Su

**Abstract**—Deepening neural models has been proven very successful in improving the model's capacity when solving complex learning tasks, such as the machine translation task. Previous efforts on deep neural machine translation mainly focus on the encoder and the decoder, while little on the attention mechanism. However, the attention mechanism is of vital importance to induce the translation correspondence between different languages where shallow neural networks are relatively insufficient, especially when the encoder and decoder are deep. In this paper, we propose a deep attention model (DeepAtt). Based on the low-level attention information, DeepAtt is capable of automatically determining what should be passed or suppressed from the corresponding encoder layer so as to make the distributed representation appropriate for high-level attention and translation. We conduct experiments on NIST Chinese-English, WMT English-German and WMT English-French translation tasks, where, with 5 attention layers, DeepAtt yields very competitive performance against the state-of-the-art results. We empirically find that with an adequate increase of attention layers, DeepAtt tends to produce more accurate attention weights. An in-depth analysis on the translation of important context words further reveals that DeepAtt significantly improves the faithfulness of system translations.

**Index Terms**—Deep attention network, neural machine translation (NMT), attention-based sequence-to-sequence learning, natural language processing

## 1 Introduction

Recent advances in deep learning enable the end-to-end neural machine translation (NMT) system to achieve very promising results on various language pairs [1], [2], [3], [4], [5]. Unlike conventional statistical machine translation (SMT), NMT is a single, large neural network which heavily relies on an *encoder-decoder* framework, where the encoder transforms source sentence into distributed vectors from which the decoder generates the corresponding target sentence word by word [2]. However, to achieve state-of-the-art performance, researchers often resort to deep NMT so as to enhance its capacity in capturing source and target semantics [3], [4], [5].

Previous efforts on deep NMT mainly focus on the encoder and the decoder. For example, Wu et al. [3] use residual connections to train 8 encoder and 8 decoder layers; Zhou et al. [4] propose fast-forward connections and train a NMT with a depth of 16; Wang et al. [5] propose linear associative units and apply it on 4 layers encoder and 4 layers decoder. Intuitively, the deep encoder is able to summarize and represent source semantics more accurately, and the deep decoder can memorize much longer translation history and dependency. Although these deep models benefit NMT significantly, they all build only upon a single-layered attention network, which might be insufficient in modeling translation correspondence between different languages thus hindering the performance of NMT systems.

The attention mechanism [2] aims to dynamically detect translation-relevant source words for predicting next target word according to the partial translation. It acts as the *translation model* in SMT, bridging the gap between encoder and decoder, which requires complex reasoning ability and is very crucial to the faithfulness of translation. To improve its capacity, we propose a deep attention model (DeepAtt). Figure 1 shows the overall architecture. With one more encoding layer, DeepAtt stacks one more attention layer. In this way, the low-level attention layer is able to provide translation-aware information to the high-level attention layer. This enables the higher layer to automatically determine what should be passed or suppressed from the corresponding encoder layer which in turn, makes the learned distributed representation more appropriate for the next target word prediction. Besides, DeepAtt retains the hierarchy of the encoder, and sets up the layer-wise interaction between the encoder and the decoder. This can help the decoder to capture more accurate source semantics since only one attention layer often induces inadequate attentions [6].

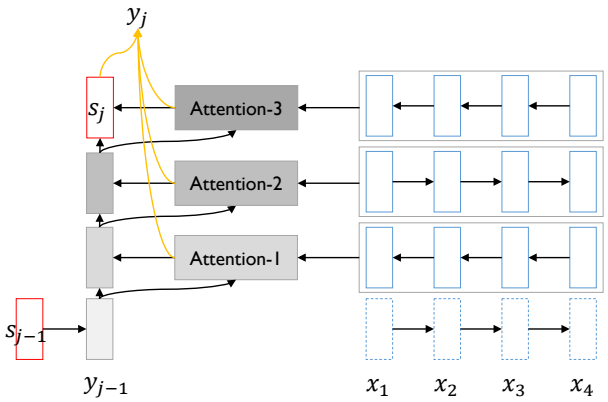


Fig. 1: Illustration of the proposed DeepAtt. We use blue and red color to indicate the source and target side respectively. The yellow and gray color denotes the information flow for target word prediction and attention respectively. Notice that we draw the encoder on the right and the decoder on the left for clarity.


- *J. Su is the corresponding author.*
  *E-mail: jssu@xmu.edu.cn.*
- *B. Zhang and J. Su are with the Software School, Xiamen University, Xiamen 361005, China.*
  *E-mail: zb@stu.xmu.edu.cn; jssu@xmu.edu.cn.*
- *D. Xiong is with the Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou 215006, China.*
  *E-mail: dyxiong@suda.edu.cn.*


*Manuscript received xxx xx, xxx; revised August xx, xxx.*

DeepAtt is deep on not only the encoder and the decoder, but also the attention mechanism. This deep attention architecture significantly improves the connection strength between the encoder and decoder, enabling more complex reasoning operations during translation. To testify its effectiveness, we conduct a series of experiments on machine translation tasks. On NIST Chinese-English task, our model achieves the best performance in terms of BLEU score compared with existing work using the same training data. We quantitatively analyze the attention weights of each attention layer in terms of alignment error rate, and find that with an adequate increase of attention layers, DeepAtt produces more accurate attention weights. We further check whether our model yields more adequate translation. Experiment results show that the translation quality of important context words (e.g. noun, verb, adjective) is indeed improved. On the WMT14 English-German and English-French (using 12M corpus only) task, our single model, with 5 attention layers, achieves a BLEU score of 24.73 and 38.56 respectively, both comparable to the state-of-the-art.

Our main contributions are summarized as follows:

- We propose a novel deep attention mechanism which operates in a hierarchical manner and allows the encoder to interact with the decoder layer by layer. The hierarchy architecture ensures that source-side semantics can be fully utilized to generate the next target word. The layer-wise interaction, , on the other hand, enables the decoder to selectively pick essential untranslated source words for the prediction of the next target word.
- We develop a novel deep encoding schema which alternates forward and backward RNNs with skip connections to the source input at each layer. The alternation helps capture more accurate source semantics via integrating both history and future source-side information. The skip connection, on the other hand, makes the gradient propagation more fluent so as to enable feasible model optimization.
- We conducted a series of experiments on NIST Chinese-English, WMT14 English-German and WMT14 English-French translation tasks. The proposed model yields consistent and significant improvements over several strong baselines, and achieves results comparable to various state-of-the-art NMT systems.

## 2 Related Work

Our work is closely related with two lines of research: *the attention mechanism* and *the deep NMT*.

Observing that the use of a fixed-length vector is insufficient in summarizing source-side semantics, Bahdanau et al. [2] propose the attention mechanism. Luong et al. [7] explore several efficient architectures for this mechanism, introducing the global and local attention models. Zhang et al. [8] propose that the recurrent neural network can be used as an alternative to the attention network. Recently, Zhang et al. [9] introduce a GRU gate to the attention model so as to improve the discriminative ability of the learned attention vectors. Vaswani et al. [10] propose a multi-head attention network with scaled-dot operation, expecting each head to capture a particular aspect of the source-target interaction. Zhang et al. [11] develop an average attention model which greatly simplifies the decoder-part self-attention mechanism using solely cumulative average operation. Rather than developing more flexible attention models, we treat these exiting models as our basic unit. Although we employ the model of Bahdanau et al. [2] in our experiments, our DeepAtt can be easily adapted to other attention models.

Based on the success achieved in computer vision [12], [13], deep neural networks have become a pretty hot spot in the NMT community, such as [3], [4], [5]. These studies differ significantly from ours in the following two aspects. First, their major focus is to enable flexible optimization since training a deep neural network is very difficult. To this end, Wu et al. [3] leverage the residual connection; Zhou et al. [4] propose a fast-forward connection, while Wang et al. [5] introduce a linear associative unit. Second, their deep architecture lies in the encoder and the decoder, ignoring the single-layered attention network which is still shallow. In contrast, our model is also deep in the attention network, making the deep encoder and deep decoder couple more tightly and the training more flexible.

Our work brings these two lines of research together. In this respect, Yang et al. [14] propose stacked attention networks to learn to answer natural language questions from images. The difference is that they apply the attention only on a single encoding layer and compose the attended vectors using the adding operation. Although their model works fine on CNN-based image question answering task, this simple architecture and operation is relatively insufficient for machine translation. Very recently, Gehring et al. [15] and Gangi et al. [16], independent of our work, propose a multi-step attention. In comparison, our DeepAtt has more compact network structure, and is more feasible for optimization. Through experiments, we observe that our model is superior to their stacked multi-layered attention-based decoder.

## 3 The Model

Unlike SMT, NMT models the translation procedure by directly mapping the source sentence $\mathbf{x} = \{x_1, \ldots, x_n\}$ to its target translation $\mathbf{y} = \{y_1, \ldots, y_m\}$. As shown in Figure 1, this is achieved via three components: *encoder*, *decoder* and *attention mechanism*. We describe these components in succession.

The *encoder* aims at summarizing and representing source semantics such that the decoder can recover them using the target language. Given a source sentence $\mathbf{x}$, we design our encoder as follows (see the blue color in Figure 1):

$$\mathbf{h}_i^k = \begin{cases} \overrightarrow{\mathbf{h}}_i^k = f_{enc}(\overrightarrow{\mathbf{h}}_{i-1}^k, \mathbf{c}_i^k, \mathbf{E}_{x_i}) & \text{if } k \text{ is even} \\ \overleftarrow{\mathbf{h}}_i^k = f_{enc}(\overleftarrow{\mathbf{h}}_{i+1}^k, \mathbf{c}_i^k, \mathbf{E}_{x_i}) & \text{otherwise} \end{cases} \tag{1}$$

where $\mathbf{h}_i^k \in \mathbb{R}^{d_h}$ is the hidden representation of source word $x_i$ in the $k$-th encoder layer, and $\mathbf{E}_{x_i} \in \mathbb{R}^{d_w}$ is the source word embedding. $\mathbf{c}_i^k \in \mathbb{R}^{d_h}$ denotes the context representation in source position $i$ which tells the encoder the unobserved information in the future. Formally,

$$\begin{aligned} &\mathbf{c}_i^k = \mathbf{h}_i^{k-1}, \\ &\text{where } \mathbf{h}_i^0 = \overrightarrow{\mathbf{h}}_i^0 = \text{GRU}(\overrightarrow{\mathbf{h}}_{i-1}^0, \mathbf{E}_{x_i}) \end{aligned} \tag{2}$$

With such a serpentine manner, our encoder is aware of what has been read so far $\overrightarrow{\mathbf{h}}_{i-1}^k / \overleftarrow{\mathbf{h}}_{i+1}^k$, what will be read next $\mathbf{c}_i^k$, and what is the current input word $\mathbf{E}_{x_i}$ so that the future and history information can be fully incorporated into the learned source word representations. To enable this full integration, we design the encoding function $f_{enc}$ using two-level hierarchy (take the first case in Eq. (1) as example):

$$\begin{aligned} \overrightarrow{\mathbf{h}}_i &= \text{GRU}_{higher}(\widetilde{\mathbf{h}}_i, \mathbf{c}_i) \\ \widetilde{\mathbf{h}}_i &= \text{GRU}_{lower}(\overrightarrow{\mathbf{h}}_{i-1}, \mathbf{E}_{x_i}) \end{aligned} \tag{3}$$

Intuitively, the low-level GRU model [17] provides a special short-cut connection to the encoding function such that our deep encoder can be feasibly optimized. After encoding, the source sentence is converted into real-valued hidden representations: $\mathbf{H}^k = \{\mathbf{h}_1^k, \ldots, \mathbf{h}_n^k\}$ (where $1 \leq k \leq K$, $K$ denotes the number of encoder layers). The higher the encoder layer is, the more abstract meanings $\mathbf{H}^k$ represents.

The *decoder* aims at leveraging these encoded source semantics $\{\mathbf{H}^k\}_{k=1}^K$ to generate not only faithful but also fluent translation. Generally, it is a conditional recurrent neural language model which models the translation probability based on the following chain rule:

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) &= \prod_{j=1}^{m} p(y_j|\mathbf{x}, \mathbf{y}_{<j}) \\ &= \prod_{j=1}^{m} \text{softmax}\left(g(\mathbf{E}_{y_{j-1}}, \mathbf{s}_j, \{\mathbf{a}_j^k\}_{k=1}^K)\right) \end{aligned} \tag{4}$$

where $\mathbf{y}_{<j} = \{y_1, \cdots, y_{j-1}\}$ denotes a partial translation. $\mathbf{E}_{y_{j-1}} \in \mathbb{R}^{d_w}$ is the embedding of previously generated target word $y_{j-1}$, $\mathbf{a}_j^k \in \mathbb{R}^{d_h}$ is the translation-sensitive attention vector produced by the $k$-th attention layer and $g(\cdot)$ is a highly non-linear function. $\mathbf{s}_j \in \mathbb{R}^{d_h}$ is the $j$-th target-side hidden state, which is usually calculated in a recurrent manner:

$$\mathbf{s}_j = f_{dec}(\mathbf{s}_{j-1}, \mathbf{E}_{y_{j-1}}, \{\mathbf{H}^k\}_{k=1}^K) \tag{5}$$

As shown in Figure 1, DeepAtt is highly coupled with this decoding function. Formally, we decompose $f_{dec}$ as follows:

$$\mathbf{s}_j = \widetilde{\mathbf{s}}_j^K \tag{6}$$

$$\widetilde{\mathbf{s}}_j^k = \text{GRU}(\widetilde{\mathbf{s}}_j^{k-1}, \mathbf{a}_j^k) \tag{7}$$

$$\text{where } \widetilde{\mathbf{s}}_j^0 = \text{GRU}(\mathbf{s}_{j-1}, \mathbf{E}_{y_{j-1}}), \quad \mathbf{a}_j^k = \text{Att}(\widetilde{\mathbf{s}}_j^{k-1}, \mathbf{H}^k) \tag{8}$$

Notice that $\mathbf{a}^k$ relies on $\widetilde{\mathbf{s}}^{k-1}$, while $\widetilde{\mathbf{s}}^k$ relies on both $\widetilde{\mathbf{s}}^{k-1}$ and $\mathbf{a}^k$. In this way, the low-level attention information $\widetilde{\mathbf{s}}^{k-1}$ can help to determine what should be extracted from the corresponding encoding layer $\mathbf{H}^k$, and the extracted information $\mathbf{a}^k$, in turn, amends the expressibility of $\widetilde{\mathbf{s}}^k$ in translation correspondence between source and target sentences for predicting the next target word.

We use $\text{Att}(\cdot)$ to denote the *attention mechanism*, which extracts a fixed-length vector $\mathbf{a}^k$ from varied-length source representations $\mathbf{H}^k$. Currently, there are several alternatives [2], [7], [8], [9], [11], among which we employed the most widely-used one [2]. This can be summarized as follows:

$$\begin{aligned} &\mathbf{a}_j^k = \sum_i \alpha_{ji}^k \mathbf{h}_i^k \\ &\text{where } \alpha_{ji}^k = \text{softmax}\left(\exp\left(v_a^T \tanh(W_a \widetilde{\mathbf{s}}_j^{k-1} + U_a \mathbf{h}_i^k)\right)\right) \end{aligned} \tag{9}$$

Our model is deep not only on the encoder and decoder, but also on the attention mechanism. To optimize our model, we used the most popular maximum likelihood objective via stochastic gradient descent algorithm.

## 4 Experiments

We evaluated DeepAtt mainly on the NIST Chinese-English task. Besides, we also provided results on the WMT14 English-German and English-French task.

### 4.1 Setup

**Datasets.** The training data for NIST Chinese-English task consists of 1.25M sentence pairs, with 27.9M Chinese words and 34.5M English words respectively. This data is a combination of LDC2002E18, LDC2003E07, LDC2003E14, Hansards portion of LDC2004T07, LDC2004T08 and LDC2005T06. We chose the NIST 2005 dataset as the development set, and the NIST 2002, 2003, 2004, 2006 and 2008 datasets as our test sets. There are 878, 919, 1788, 1082 and 1664 sentences in NIST 2002, 2003, 2004, 2005, 2006, 2008 dataset respectively.

For English-German task, we used the same subset of the WMT 2014 training corpus as in [3], [4], [5], [7]. This training data consists of 4.5M sentence pairs with 116M English words and 110M German words respectively.[1] We used the news-test 2013 as the development set, and the news-test 2014 as the test set.

For English-French task, we also used the WMT 2014 training data. The whole training corpus consists of around 36M sentence pairs, from which we selected 12M sentence pairs for training so as to meet our computational capability. The selection algorithm strictly follows the previous work[2]. Finally, our training data contain 304M English words and 348M French words. We used the combination of news-test 2012 and news-test 2013 as the development set, and the news-test 2014 as the test set.

**Evaluation.** We used the case-insensitive and case-sensitive BLEU-4 metric [18] to evaluate translation quality of Chinese-English and English-German, English-French task respectively. For all tasks, we tokenized the reference and evaluated the performance using *multi-bleu.perl*.[3] We performed paired bootstrap sampling [19] for significance test.

### 4.2 Model Settings

We adopted similar settings as Bahdanau et al. [2]. For Chinese-English task, we extracted the most frequent 30K words from two corpora as the source and target vocabulary, covering approximately 97.7% and 99.3% of each corpora respectively. With respect to *Moses*, we used all the 1.25M sentence pairs (without length limitation). We trained a 4-gram language model on the target portion of training data using the SRILM[4] toolkit with modified Kneser-Ney smoothing. We word-aligned the training corpus using GIZA++[5] toolkit with the option "*grow-diag-final-and*". We employed the default lexical reordering model with the type "*wbe-msd-bidirectional-fe-allff*". All other parameters were kept as the default settings.

For English-German task, we applied the byte pair encoding compression algorithm [20] to reduce the vocabulary size as well as to deal with rich morphology. For both languages, we preserved 16K sub-words as the vocabulary. We also tested a big setting with 30K sub-words extracted as the vocabulary. Similarly, for English-French task, we preserved 40K sub-words in the source and target vocabulary, respectively.

We used $d_w = 620$ dimensional word embeddings and $d_h = 1000$ dimensional hidden states for both the source and target languages. All non-recurrent parts were randomly initialized with

1. The preprocessed data can be found and downloaded from http://nlp.stanford.edu/projects/nmt/.

2. http://www-lium.univ-lemans.fr/šchwenk/cslm_joint_paper/

3. https://github.com/moses-smt/mosesdecoder/tree/master/scripts/generic/multi-bleu.perl

4. http://www.speech.sri.com/projects/srilm/download.html

5. http://www.fjoch.com/GIZA++.html

| System | MT05 | MT02 | MT03 | MT04 | MT06 | MT08 | ALL |
|---|---|---|---|---|---|---|---|
| *Moses* | 31.70 | 33.61 | 32.63 | 34.36 | 31.00 | 23.96 | 31.03 |
| *RNNSearch [2]* | 34.72 | 37.95 | 35.23 | 37.32 | 33.56 | 26.12 | 34.06 |
| *DeepAtt-1* | 36.44 | 40.12 | 37.63 | 39.83 | 35.44 | 27.34 | $36.12^{\Uparrow ++}$ |
| *DeepAtt-2* | 36.90 | 39.71 | 37.79 | 39.93 | 35.95 | 27.87 | $36.34^{\Uparrow ++}$ |
| *DeepAtt-3* | 36.75 | 40.53 | 38.12 | 40.14 | 36.14 | 28.12 | $36.65^{\Uparrow ++}$ |
| *DeepAtt-4* | 37.87 | 40.99 | 39.10 | 40.77 | 37.14 | 28.44 | $37.34^{\Uparrow ++}$ |
| *DeepAtt-5* | **38.82** | 41.00 | 39.07 | **41.09** | **37.37** | 28.52 | $37.50^{\Uparrow ++}$ |
| *DeepAtt-6* | 38.29 | **41.40** | **39.23** | 40.66 | 37.20 | **28.99** | $\mathbf{37.51}^{\Uparrow ++}$ |

TABLE 1: Case-insensitive BLEU scores on the Chinese-English translation task. *DeepAtt-$k$*: the proposed model using "$k$" attention layers, "$k$" encoder layers and "$k$" decoder layers (i.e. $K =$"$k$"). *RNNSearch*: a vanilla NMT system using 1-layer encoder and 1-layer decoder with 1-layer attention. **ALL** = total BLEU score on all test sets. We highlight the best results in bold for each test set. All neural models were trained with Adadelta optimizer. "↑/⇑": significantly better than *Moses* ($p < 0.05/p < 0.01$); "+/++": significantly better than *RNNSearch* ($p < 0.05/p < 0.01$).

| System | Layer-1 | Layer-2 | Layer-3 | Layer-4 | Layer-5 | Layer-6 | ALL |
|---|---|---|---|---|---|---|---|
| *RNNSearch* | 50.83 | - | - | - | - | - | 50.83 |
| *DeepAtt-1* | 45.25 | - | - | - | - | - | 45.25 |
| *DeepAtt-2* | 54.43 | 52.53 | - | - | - | - | 48.09 |
| *DeepAtt-3* | 67.27 | 43.99 | 96.28 | - | - | - | 47.23 |
| *DeepAtt-4* | 77.64 | 52.16 | 76.86 | 53.30 | - | - | 45.38 |
| *DeepAtt-5* | 60.71 | 49.22 | 64.39 | 78.86 | 96.44 | - | **44.69** |
| *DeepAtt-6* | 60.42 | 54.45 | 53.91 | 73.18 | 97.50 | 95.65 | 46.01 |

TABLE 2: AER scores of word alignments. The lower the score, the better the alignment quality. **ALL** = overall AER scores on all attention layers.

zero mean and standard deviation of 0.01, except the recurrent parameters which were initialized with random orthogonal matrices. During decoding, we used the beam-search algorithm, and set the beam size to 10.

The model is trained through standard SGD algorithm with a mini-batch size of 80 sentences. We updated the learning rate using the Adadelta algorithm [21] ($\epsilon = 10^{-6}$ and $\rho = 0.95$). We clipped the norm of model gradient to make it no more than 5.0 so as to avoid gradient explosion issue. Dropout was also applied on the output layer to avoid over-fitting. We set the dropout rate to be 0 for Chinese-English task, and 0.2 for English-German and English-French task. In addition, following recent advances in deep learning community [3], [10], we employed Adam algorithm [22] ($\epsilon = 10^{-8}$ and $\beta_1 = 0.9, \beta_2 = 0.999$) in some cases. Without clear declaration, we used the Adadelta for experiment.

We implemented all our models based on the open-sourced *dl4mt* system.[6]. All NMT systems were trained on a GeForce GTX 1080 based on the computational framwork Theano where up to 6 attention layers were tested due to the physical memory limit of our GPU. The training of our DeepAtt costs around 5 days on the Chinese-English task, around 3 weeks on the English-German task and around 6 weeks on the English-French task when $K$ is set to be 5.

### 4.3 Results on Chinese-English Translation

Table 1 shows the translation results of different systems. No matter how many attention layers are used, DeepAtt always significantly outperforms both Moses and RNNSearch, with gains of up to 6.48 and 3.45 BLEU points respectively. Besides, as the number of attention layers increases, the overall translation performance is improved. Specifically, when there are 6 attention layers, DeepAtt achieves 37.51 BLEU score on all test sets. This suggests that the deep attention architecture benefits the NMT system.

6. https://github.com/nyu-dl/dl4mt-tutorial/blob/master/session3/nmt.py

We also observed that compared to DeepAtt-6, DeepAtt-5 requires less training time with no significant performance degradation. Thus, we conducted the following experiments using 5 attention layers by default, unless mentioned otherwise.

### 4.4 Analysis on Chinese-English Translation

The major difference between DeepAtt and other deep NMTs lies in the multiple attention layers. Therefore, we first quantitatively evaluated the quality of the learned attention weights at different layers. To this end, we employed the alignment error rate (AER) metric [23] and used the evaluation dataset from Liu and Sun [24], which contains 900 manually aligned Chinese-English sentence pairs [6]. Table 2 summarizes the results. With respect to the overall AER score, we observed that there are no consistent improvements as the attention layers deepen. However, all DeepAtt models achieve lower (better) AER scores than the RNNSearch, especially the DeepAtt-5 which yields the lowest 44.69 AER score. This indicates that deepening the attention layers can help improve the alignment quality, which typically contributes significantly to the translation performance.

With respect to the AER score across different attention layers (take DeepAtt-5 as example), we find that the score decreases at first, then raises sharply ($60.71 \rightarrow 49.22 \rightarrow 96.77$). This suggests that DeepAtt first seeks the translation-relevant source words, and then pay more attention to the other words. We argue that this phenomenon, to some extent, is consistent with human's procedure of translation. That is, a human needs to determine which source word to translate at first, then checks broader context to confirm its meaning, and finally finds out adequate target translations.

High quality word alignment plays an important role in the translation of significant context words (e.g. noun, verb, adjective). As DeepAtt produces better attention weights, we dug into the translations and investigated whether the translation quality of context words can be improved. We assigned parts of speech to

| Metric | RNNSearch | | | | DeepAtt | | | |
|---|---|---|---|---|---|---|---|---|
| | NN | VB | NN&VB | NN&VB&JJ | NN | VB | NN&VB | NN&VB&JJ |
| *BLEU-1* | 61.97 | 55.75 | 59.18 | 60.66 | 65.38 | 58.87 | 62.52 | 64.07 |
| *BLEU-2* | 44.33 | 28.62 | 29.80 | 33.15 | 45.84 | 33.24 | 33.09 | 36.45 |
| *BLEU-3* | 35.66 | 15.01 | 15.01 | 17.92 | 36.98 | 21.21 | 17.41 | 20.38 |
| *BLEU-4* | 11.40 | - | 7.71 | 9.68 | 17.07 | - | 9.27 | 11.50 |

TABLE 3: Case-insensitive BLEU scores on specific context words. **NN** = noun, **VB** = verb, **JJ** = adjective, **NN&VB** = noun and verb. "—" indicates the BLEU score is zero.

| | |
|---|---|
| *Source* | 津巴布韦 总统 穆加贝 在 3月 9日 至 11日 的 总统 选举 中 再次 当选 , 但 西方 国家 指责 选举 存在 严重 作弊 行为 , 缺乏 公正性 和 自由性 , 因此 拒绝 承认 选举 结果 , 并 扬言 将 对 津巴布韦 进一步 实施 制裁 。 |
| | jīn bā bù wéi zǒng tǒng mù jiā bèi zài 3 yuè 9 rì zhì 11 rì de zǒng tǒng xuǎn jǔ zhōng zài cì dāng xuǎn , dàn xī fāng guó jiā zhǐ zé xuǎn jǔ cún zài yán zhòng zuò bì háng wéi , quē fá gōng zhèng xìng hé zì yóu xìng , yīn cǐ jù jué chéng rèn xuǎn jǔ jié guǒ , bìng yáng yán jiāng duì jīn bā bù wéi jìn yī bù shí shī zhì cái 。 |
| *Reference* | *zimbabwean president mugabe was reelected at the presidential election held from march 9 to 11 , but western countries , alleging serious cheatings and lack of fairness and freedom in the election , refused to recognize the election results and threatened further sanctions against zimbabwe .* |
| *Moses* | *zimbabwean president mugabe 9 to 11 march in the presidential election again elected , however , western countries criticize election , there exist serious lack of fairness and cheating at UNK and therefore refused to recognize the election results , and vowed to zimbabwe further sanctions .* |
| *RNNSearch* | *in his election in the presidential election on march 9 - 11 , zimbabwe president UNK was re - elected during the presidential election on march 9 th .* |
| *DeepNMT* | *in the presidential election from march 9 to the 11 th of march - 11 , zimbabwean president mugabe , however , refused to acknowledge the election results and threatened to further impose sanctions against zimbabwe .* |
| *VDeepAtt* | *in the presidential election from march 9 th to the 11 th of the presidential election from march 9 th to the 11 th , western countries have refused to recognize the election results and threatened to further impose sanctions against zimbabwe .* |
| *WideAtt* | *during his election in the presidential election on the 9 th to 11 march , zimbabwe 's president mugabe was re - elected in the presidential election on march 9 - 11 , but he refused to acknowledge the election results and threatened to further impose sanctions against zimbabwe .* |
| *DeepAtt* | *zimbabwean president mugabe was re - elected in the presidential election on 9 - 11 march , but western countries have accused the election of serious UNK and lack fairness and UNK , thus refusing to admit the election result and threatened to further impose sanctions on zimbabwe .* |

TABLE 4: Examples generated by different systems. The translation of DeepAtt is more accurate in expressing the meanings of source sentences. Important phrases are highlighted in red color.

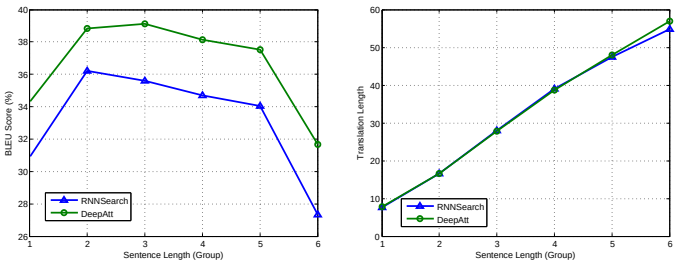


Fig. 2: BLEU score and translation length on different length groups of source sentences.

each word in the references and translations using the *Stanford POS Tagger*[7], and evaluated the translation quality of noun (NN), verb (VB) and adjective (JJ) alone. We report BLEU from 1 to 4, and show the results in Table 3. Obviously, DeepAtt leads to remarkable improvements over RNNSearch on all context words. Specifically, on "NN", DeepAtt outperforms RNNSearch by 5.67 BLEU-4 points, while on "VB", DeepAtt achieves a gain of 6.2 BLEU-3 points. These significant improvements strongly indicate that DeepAtt connects the encoder and decoder more tightly so as to enable the translations more faithful.

A common challenge for NMT system is the translation of long source sentences. The above analysis reveals that DeepAtt generates more faithful translation. We further verified this point on long sentences. Following Bahdanau et al. [2], we grouped sentences of similar lengths together and computed BLEU score and average length of translations in each group.[8] Figure 2 shows the overall results. We observe that the performance of RNNSearch drops sharply when the length of source sentence exceeds 50. Compared with RNNSearch, DeepAtt yields consistent and significant improvements on all groups. Specifically, DeepAtt obtains a gain of up to 4 BLEU points on the longest group. Surprisingly, the translation length of DeepAtt is almost the same as that of RNNSearch. This suggests that DeepAtt achieves much better translation performance without changing the length of translation, demonstrating the ability of DeepAtt in dealing with long-range dependencies as well as generating faithful translations.

### 4.5 Translation Analysis

Following the above analysis, we further provide some translation examples to verify whether our model indeed generates more fluent and faithful translation. We show the instances in Table 4.

As a traditional statistical system that relies heavily on large-scale phrase pairs, the Moses succeeds in generating faithful translations, which, however, tend to lack of fluency. For example, the sentence "*zimbabwean president mugabe 9 to 11 march*

7. https://nlp.stanford.edu/software/tagger.shtml

8. We divide our test sets into 6 disjoint groups according to the length of source sentences ((0, 10), [10, 20), [20, 30), [30, 40), [40, 50), [50, -)), each of which has 680, 1923, 1839, 1189, 597 and 378 sentences respectively.

| System | #Enc | #Dec | #Att | MT05 | MT02 | MT03 | MT04 | MT06 | MT08 | ALL |
|---|---|---|---|---|---|---|---|---|---|---|
| *Existing End-to-End NMT Systems* | | | | | | | | | | |
| *Coverage [5]* | 1 | 1 | 1 | 34.91 | - | 34.49 | 38.34 | 34.25 | - | - |
| *MemDec [5]* | 1 | 1 | 1 | 35.91 | - | 36.16 | 39.81 | 35.98 | - | - |
| *DeepLAU [5]* | 4 | 4 | 1 | 38.07 | - | 39.35 | 41.15 | 37.29 | - | - |
| *VRNMT [25]* | 1 | 1 | 1 | 36.82 | - | 38.08 | 41.07 | 36.72 | - | - |
| *ABDNMT [26]* | 1 | 1 | 1 | 38.84 | - | 40.02 | 42.32 | 38.38 | - | - |
| *Our End-to-End NMT systems* | | | | | | | | | | |
| *RNNSearch* | 1 | 1 | 1 | 34.72 | 37.95 | 35.23 | 37.32 | 33.56 | 26.12 | 34.06 |
| *DeepNMT* | 5 | 5 | 5 | 36.44 | 39.29 | 37.89 | 39.65 | 35.37 | 27.63 | $36.02^{++}$ |
| *VDeepAtt* | 5 | 5 | 5 | 37.15 | 39.71 | 38.36 | 40.48 | 36.29 | 28.00 | $36.69^{++}$ |
| *WideAtt* | 5 | 1 | 1 | 35.66 | 38.60 | 37.01 | 38.49 | 34.66 | 26.06 | $35.00^{++}$ |
| *MHeadNMT* | 1 | 1 | 1 | 36.23 | 39.61 | 35.89 | 39.45 | 35.70 | 27.80 | $36.09^{++}$ |
| *UDeepAtt* | 5 | 5 | 5 | 36.71 | 39.93 | 37.78 | 39.38 | 36.03 | 28.25 | $36.30^{++}$ |
| *DeepAtt* | 5 | 5 | 5 | 38.82 | 41.00 | 39.07 | 41.09 | 37.37 | 28.52 | $37.50^{++}$ |
| *DeepAtt + LN* | 5 | 5 | 5 | 40.19 | 42.20 | 40.24 | 42.13 | 38.59 | 30.15 | $38.78^{++}$ |
| *DeepAtt + LN + Adam* | 5 | 5 | 5 | 44.16 | 45.70 | 44.17 | 46.82 | 43.12 | 34.16 | $43.08^{++}$ |
| *DeepAtt + LN + Adam (4 model ensemble)* | 5 | 5 | 5 | **46.17** | **47.61** | **47.30** | **49.14** | **45.94** | **36.64** | $\mathbf{45.58}^{++}$ |

TABLE 5: Case-insensitive BLEU scores of advanced systems on the Chinese-English translation task. "*#Enc*" = number of encoder layers, "*#Dec*" = number of decoder layers and "*#Att*" = number of attention layers. "—" indicates no result is provided in [5]. "*Adam*" = model is optimized with Adam optimizer, if specified. "*LN*" = length normalization during decoding.

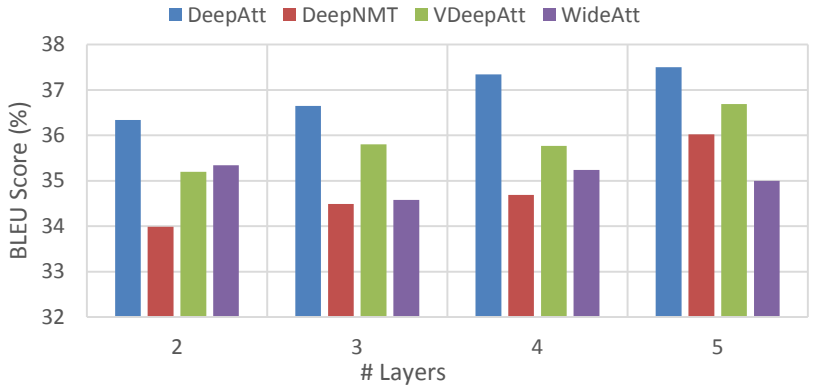


Fig. 3: BLEU score of different systems on all test sets under different numbers of layers.

*in the presidential election again elected*" suffers from serious disorder problem as well as missing-predicate problem. In contrast, the translations of all NMT systems exhibit incredible fluency. Nevertheless, different NMT systems are faced with different challenges.

On one hand, these models sometimes prefer to avoid translating some important source clauses, which is a well-known under-translation problem [6]. For example, RNNSearch fails to translate the source clause " 但(*but*) 西方(*western*) 国家(*countries*) 指责(*alleging*) 选举(*the election*) 存在(*in*) 严重(*serious*) 作弊(*cheating*) 行为(*behavior*) , 缺乏(*lack of*) 公正性(*fairness*) 和(*and*) 自由性(*freedom*) , 因此(*thus*) 拒绝(*refused to*) 承认(*recognize*) 选举(*the election*) 结果(*results*) , 并(*and*) 扬言(*threatened*) 将(*will*) 对(*against*) 津巴布韦(*zimbabwe*) 进一步(*further*) 实施(*carry out*) 制裁(*sanctions*) 。".[9] It seems that the shallow model has difficulties in extracting and transforming the source semantics, which can also be reflected on its poor alignment quality. Deepening the model is a promising way to alleviate this problem, as we observe that all deep models can recover more source meaning into the translations. However, except our model, other deep models still neglect several source clauses during transformation.

9. Words in bracket are word-by-word English translations.

On the other hand, some common source clauses can be translated repeatedly, which is a well-known over-translation problem [6]. This is because if the model fails to capture the source semantics, it may try to translate the recognized part over and over. As an example, the sub-translation "*9-11 march*" appears several times in all the NMT systems except ours. Additionally, both DeepNMT and WideNMT mistakenly produce "*zimbabwe 's president mugabe*" rather than "*western countries*" as the subject that "*refused to acknowledge the election results and threatened to further impose sanctions against zimbabwe* .". All these strongly demonstrate that deepening the model alone is not sufficient enough to correctly convey the meaning of the source sentences.

Our DeepAtt, although its generated translations are not perfect either, handles these problems much better. We contribute this to the proposed attention architecture that is more capable of dealing with the underlying semantics of source sentences.

### 4.6 More Comparisons on Chinese-English Translation

Except for the Moses and RNNSearch, we provide the following existing systems:

- **Coverage** [6]: A RNNSearch with a coverage vector to keep track of the translated and un-translated source words.
- **MemDec** [31]: A RNNSearch whose decoder is enhanced with an external memory.

| System | Architecture | Vocab | BLEU |
|---|---|---|---|
| Buck et al. [27] | Winning WMT14 system phrase-based + large LM | - | 20.7 |
| *Existing end-to-end NMT systems* | | | |
| Jean et al. [28] | RNNSearch + unk replace + large vocab | 500K | 19.40 |
| Luong et al. [7] | LSTM with 4 layers + dropout + local att. + unk replace | 50K | 20.90 |
| Shen et al. [29] | RNNSearch (GroundHog) + MRT + PosUnk | 50K | 20.45 |
| Zhou et al. [4] | LSTM with 16 layers + Fast-Forward connections | 80K | 20.60 |
| Wu et al. [3] | LSTM with 8 layers + Word | 80K | 23.10 |
| Wu et al. [3] | LSTM with 8 layers + RL-refined WPM | 32K | 24.60 |
| Wang et al. [5] | RNNSearch with 4 layers + LAU | 80K | 22.10 |
| Wang et al. [5] | RNNSearch with 4 layers + LAU + PosUnk | 80K | 23.80 |
| Gehring et al. [15] | CNN with 15 layers + Multi-step Attention + BPE | 40K | 25.16 |
| Cheng et al. [30] | RNN with 2 layers + adversarial stability training + BPE | 30K | 25.26 |
| Gangi et al. [16] | RNN with 10 layers + SR + BPE | 32K | 24.98 |
| Vaswani et al. [10] | Attention with 6 layers + WPM + base | 32K | 27.30 |
| Wang et al. [5] | RNNSearch with 4 layers + LAU + PosUnk (8 model ensemble) | 80K | 26.10 |
| Wu et al. [3] | LSTM with 8 layers + RL-refined WPM (8 model ensemble) | 32K | 26.20 |
| Gehring et al. [15] | CNN with 15 layers + Multi-step Attention + BPE (8 model ensemble) | 40K | 26.43 |
| *Our end-to-end NMT systems* | | | |
| *this work* | DeepAtt with 5 layers + BPE | 16K | 24.22 |
| | DeepAtt with 5 layers + BPE + Adam | 30K | 24.73 |
| | DeepAtt with 5 layers + BPE + Adam (4 model ensemble) | 30K | **26.45** |

TABLE 6: Case-sensitive BLEU scores on WMT14 English-German translation task. "*unk replace*" and "*PosUnk*" denotes the approach of handling rare words in Jean et al. [28] and Luong et al. [7] respectively. "*RL*" and "*WPM*" represent the reinforcement learning optimization and wordpiece model used in Wu et al. [3], respectively. "*LAU*" and "*MRT*" denote the linear associative unit and the minimum risk training proposed by Wang et al. [5] and Shen et al. [29] respectively. "*BPE*" denotes the byte pair encoding algorithm in Sennrich et al. [20]. "*SR*" indicates the weakly-recurrent model proposed by Gangi et al. [16].

- **DeepLAU** [5]: A deep RNNSearch with linear associative units to reduce the gradient propagation length inside the recurrent unit.
- **VRNMT** [25]: A RNNSearch equipped with recurrent latent variable to capture semantic variance during decoding.
- **ABDNMT** [26]: A RNNSearch enhanced with a bidirectional decoding procedure. This model uses a two-stage translation.

Besides, we also implement several closely-related models:

- **DeepNMT**: A vanilla deep NMT model with 5 encoder and 5 decoder layers, but only one attention layer. In practice, we used the same encoder as our DeepAtt.
- **VDeepAtt**: A vanilla design of DeepAtt-5. The difference lies in the decoder, where VDeepAtt simply stacks multiple attention-based decoder layers [15] rather than coupling these attention layers into one recurrent unit as in DeepAtt. Formally, VDeepAtt employs 5 stacked conditional recurrent decoders to predict the next target word:

$$\widetilde{\mathbf{s}}_j^k = \text{GRU}(\mathbf{s}_{j-1}^k, \mathbf{s}_j^{k-1}) \tag{10}$$

$$\mathbf{s}_j^k = \text{GRU}(\widetilde{\mathbf{s}}_j^k, \mathbf{a}_j^k) \tag{11}$$

$$\mathbf{a}_j^k = \text{Att}(\widetilde{\mathbf{s}}_j^k, \mathbf{H}^k) \tag{12}$$

  where the $j$-th target hidden state in the $k$-th layer $\mathbf{s}_j^k$ depends on the previous hidden state in the same layer $\mathbf{s}_{j-1}^k$, the current hidden state in the last layer $\mathbf{s}_j^{k-1}$ and the source representation $\mathbf{H}^k$. The start point for the hidden representation $\mathbf{s}_j^0 = \mathbf{E}_{y_{j-1}}$.
- **WideAtt**: Rather than stacking multiple attention layers, WideAtt concatenates the multiple encoded source representations and attends to it using only one decoder layer. In summary, WideAtt uses 5 encoder layers and 1 decoder layer with 1 attention layer. The source representation applied for decoding is calculated as follows:

$$\mathbf{H} = \text{concate}(\left\{\mathbf{H}^1, ..., \mathbf{H}^K\right\}) \tag{13}$$

  where $\mathbf{H}^k$ is defined as in Eq. (1).
- **MHeadNMT**: A vanilla RNNSearch system, which utilizes a multi-head attention network described in [10] rather than the vanilla attention mechanism [2]. We used 8 heads for experiment.
- **UDeepAtt**: The same model as DeepAtt-5, except that each encoder layer follows the same direction, rather than the alternative forward and backward architecture. Formally, the encoder of UDeepAtt operates as follows:

$$\mathbf{h}_i^k = \overrightarrow{\mathbf{h}}_i^k = f_{enc}(\overrightarrow{\mathbf{h}}_{i-1}^k, \mathbf{c}_i^k, \mathbf{E}_{x_i}) \tag{14}$$

Table 5 shows the results. For our NMT systems, all deep models outperform RNNSearch significantly, demonstrating the modeling capacity of deep neural networks as well as the solidness of this line of research. Among these systems, WideAtt yields the worst performance. This indicates that concatenating the multi-layered source representations makes the NMT shallow, and finally results in the loss of valuable capacity in modeling translation. Compared with DeepNMT, VDeepAtt achieves better performance with a gain of 0.67 BLEU points. Since the main difference between DeepNMT and VDeepAtt lies in that VDeepAtt applies multiple attention layers, we believe that deep attention is a feasible and effective direction. Enhanced with our proposed attention architecture, DeepAtt obtains another gain of 0.81 BLEU points over VDeepAtt, which suggests both the effectiveness and efficiency of our DeepAtt architecture, considering that DeepAtt has more compact structure than VDeepAtt, enabling much efficient gradient propagation inside the decoder.

Compared with UDeepAtt, DeepAtt achieves a clear improvement of 1.2 BLEU points. The only difference between these two models is that we alternate the encoding direction between consecutive encoder layers. A benefit from this alternation is that future information can be fully mixed with history information, thus enabling DeepAtt to produce more accurate source representations. We also compared our model with the multi-head attention

| System | Architecture | Data | Vocab | BLEU |
|---|---|---|---|---|
| *Existing end-to-end NMT systems* | | | | |
| Jean et al. [28] | RNNSearch + unk replace + large vocab | 12M | 500K | 34.11 |
| Luong et al. [32] | LSTM with 6 layers + PosUnk | 12M | 40K | 32.70 |
| Shen et al. [29] | RNNSearch + MRT + PosUnk | 12M | 30K | 34.23 |
| Zhou et al. [4] | LSTM with 16 layers + Fast-Forward connections | 36M | 80K | 37.70 |
| Wu et al. [3] | LSTM with 8 layers + WPM | 36M | 32K | 38.95 |
| Wang et al. [5] | RNNSearch with 4 layers + LAU + PosUnk | 12M | 30K | 35.10 |
| Gehring et al. [15] | CNN with 15 layers + Multi-step Attention + BPE | 36M | 40K | 40.51 |
| Vaswani et al. [10] | Attention with 6 layers + WPM + base | 36M | 32K | 38.10 |
| Wu et al. [3] | LSTM with 8 layers + WPM (8 model ensemble) | 36M | 32K | 40.35 |
| Gehring et al. [15] | CNN with 15 layers + Multi-step Attention + BPE (8 model ensemble) | 36M | 40K | 41.44 |
| *Our end-to-end NMT systems* | | | | |
| *this work* | DeepAtt with 5 layers + BPE + Adam | 12M | 40K | 38.56 |
| | DeepAtt with 5 layers + BPE + Adam (4 model ensemble) | 12M | 40K | **39.88** |

TABLE 7: Case-sensitive BLEU scores on WMT14 English-French translation task. "*unk replace*" and "*PosUnk*" denotes the approach of handling rare words in Jean et al. [28] and Luong et al. [7] respectively. "*RL*" and "*WPM*" represent the reinforcement learning optimization and wordpiece model used in Wu et al. [3], respectively. "*LAU*" and "*MRT*" denote the linear associative unit and the minimum risk training proposed by Wang et al. [5] and Shen et al. [29] respectively. "*BPE*" denotes the byte pair encoding algorithm in Sennrich et al. [20].

network [10]. Results show that MHeadNMT yields a gain of 2.03 BLEU points over the RNNSearch, indicating that capturing different aspects of the source-target interaction is beneficial for translation. Nevertheless, deepening the attention network with compact structures as in our DeepAtt can reach better performance, achieving a gain of 1.41 BLEU points over MHeadNMT.

In order to have a fair comparison with the existing systems, we apply the length normalization during translation.[10] To the best of our knowledge, DeepLAU [5] reported the best BLEU scores using the 1.25M training data. However, our DeepAtt outperforms all these systems significantly. Enhanced with the Adam optimizer, our model reaches an overall BLEU score of 43.08, a strong improvement over the one trained with Adadelta by a great margin of 4.3 BLEU points. We further performed model ensemble. Using 4 well-trained model under different random seeds, our model resets the state-of-the-art results on this task, where the overall BLEU score increases to 45.58. Besides, except for NMT with 5 layers, we also compared different models under other numbers of layers, which is shown in Figure 3. We observe that with the increase of layers, NMT models produce better results, and no matter how many layers are used, DeepAtt always outperforms other related models and achieves the best result. All these demonstrate the modeling power of our deep attention architecture.

### 4.7 Results on English-German Translation

Table 6 shows the results on English-German translation. We also show existing systems comparable to ours including the winning system in WMT14 [27], a phrase-based system whose language models were trained on a huge monolingual text, the Common Crawl corpus. Obviously, current WMT14 performance is led by deep NMT systems. For example, Wu et al. [3] reported 24.61 BLEU score with 8 LSTM layers, and Wang et al. [5] generated 23.80 BLEU score with 4 GRU+LAU layers. Very recently, the state-of-the-art is refreshed by Gehring et al. [15] using 15 CNN layers and becomes 25.12, which is further broken through by Vaswani et al. [10] and reaches 27.30.

10. Even with length normalization, the comparison is not completely fair. Although all systems use the same training data, the existing systems are tuned on NIST 02, while ours is tuned on NIST 05. However, we believe this is not the key.

Our model achieves 24.73 BLEU score, a very competitive result against the RNN-based and CNN-based systems above. Under similar model settings, the GNMT [3] yields 24.36 BLEU score (0.37 BLEU points lower than our model) with various non-trivial tricks such as coverage penalty, specific length normalization, fine-tunning and the RL-refined model. Although Gehring et al. [15] achieved 25.16, they used 40K sub-words and 15 layers, several times larger than those of our mode. We also performed model ensemble to enhance the translation performance. By initializing with different random seeds, we trained 4 different models whose ensemble pushed the BLEU score to 26.45, making our model outperform both GNMT [3], LAU-NMT [5] and CNN-NMT [15].

### 4.8 Results on English-French Translation

Table 7 summarizes the translation performance of different NMT systems. Unlike the above translation tasks, this task provides a training corpus of 12M sentence pairs, around three times and ten times larger than that of English-German and Chinese-English translation task respectively. Overall, our single model achieves a BLEU score of 38.56, and its ensemble using 4 well-trained models improves the score to 39.88. Both results are competitive against both RNN-based and CNN-based systems.

Among systems trained with 12M sentence pairs, our model is the best, outperforming the previous best model, i.e Wang et al. [5] (35.10), by a great margin of 3.46 BLEU points. When using the full 36M sentence pairs, GNMT [3] yeilds a BLEU score of 38.95, Transformer [10] achieves 38.10, and CNN-NMT [15] reaches 40.15. By contrast, our model, using only 12M training data, is able to generate translations with a BLEU score of 38.56, demonstrating our model's excellent capability in modeling translation relationship.

## 5 Conclusion and Future Work

In this paper, we have presented a deep attention model (DeepAtt) for NMT systems. Through multiple stacked attention layers with each layer paying attention to a corresponding encoder layer, DeepAtt enables low-level attention information to guide what should be passed or suppressed from the encoder layer so as to make the learned distributed representations appropriate for high-level translation tasks. Our model is simple to implement

and flexible to train. Experiments on both NIST Chinese-English, WMT14 English-German and English-French translation tasks demonstrated the effectiveness of our model in improving both the translation and alignment quality.

In the future, we want to testify DeepAtt on other tasks, e.g., summarization. Additionally, our model is not limited to current attention unit. As mentioned in Section 2, we are also interested in adapting DeepAtt to other more complex attention models.

## ACKNOWLEDGMENT

The authors were supported by National Natural Science Foundation of China (Nos. 61672440 and 61622209), the Fundamental Research Funds for the Central Universities (Grant No. ZK1024), and Scientific Research Project of National Language Committee of China (Grant No. YB135-49). Biao Zhang greatly acknowledges the support of the Baidu Scholarship. We also thank the reviewers for their insightful comments.

## REFERENCES

[1] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *CoRR*, vol. abs/1409.3215, 2014.
[2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. of ICLR*, 2014.
[3] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," *CoRR*, vol. abs/1609.08144, 2016.
[4] J. Zhou, Y. Cao, X. Wang, P. Li, and W. Xu, "Deep recurrent models with fast-forward connections for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 371–383, 2016.
[5] M. Wang, Z. Lu, J. Zhou, and Q. Liu, "Deep Neural Machine Translation with Linear Associative Unit," *ArXiv e-prints*, May 2017.
[6] Z. Tu, Z. Lu, Y. Liu, X. Liu, and H. Li, "Coverage-based neural machine translation," *CoRR*, vol. abs/1601.04811, 2016.
[7] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. of EMNLP*, September 2015, pp. 1412–1421.
[8] B. Zhang, D. Xiong, and J. Su, "Recurrent neural machine translation," *CoRR*, vol. abs/1607.08725, 2016.
[9] B. Zhang, D. Xiong, and J. Su, "A GRU-Gated Attention Model for Neural Machine Translation," *ArXiv e-prints*, Apr. 2017.
[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5998–6008.
[11] B. Zhang, D. Xiong, and J. Su, "Accelerating neural transformer via an average attention network," in *Proc. of ACL*. Melbourne, Australia: Association for Computational Linguistics, July 2018.
[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.
[13] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Proc. of NIPS*, ser. NIPS'15, 2015, pp. 2377–2385.
[14] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola, "Stacked attention networks for image question answering," *CoRR*, vol. abs/1511.02274, 2015.
[15] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional Sequence to Sequence Learning," *ArXiv e-prints*, May 2017.
[16] M. Antonino Di Gangi and M. Federico, "Deep Neural Machine Translation with Weakly-Recurrent Units," *ArXiv e-prints*, May 2018.
[17] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling." *CoRR*, 2014.
[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proc. of ACL*, 2002, pp. 311–318.
[19] P. Koehn, "Statistical significance tests for machine translation evaluation," in *Proc. of EMNLP*, 2004.
[20] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. of ACL*, Berlin, Germany, August 2016, pp. 1715–1725.
[21] M. D. Zeiler, "ADADELTA: an adaptive learning rate method," *CoRR*, vol. abs/1212.5701, 2012.
[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[23] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Comput. Linguist.*, vol. 29, no. 1, pp. 19–51, Mar. 2003.
[24] Y. Liu and M. Sun, "Contrastive unsupervised word alignment with non-local features," *CoRR*, vol. abs/1410.2082, 2014.
[25] J. Su, S. Wu, D. Xiong, Y. Lu, X. Han, and B. Zhang, "Variational recurrent neural machine translation," *arXiv preprint arXiv:1801.05119*, 2018.
[26] X. Zhang, J. Su, Y. Qin, Y. Liu, R. Ji, and H. Wang, "Asynchronous bidirectional decoding for neural machine translation," *CoRR*, vol. abs/1801.05122, 2018. [Online]. Available: http://arxiv.org/abs/1801.05122
[27] C. Buck, K. Heafield, and B. van Ooyen, "N-gram counts and language models from the common crawl," in *Proc. of LREC*, Reykjavik, Iceland, May 2014, pp. 3579–3584.
[28] S. Jean, K. Cho, R. Memisevic, and Y. Bengio, "On using very large target vocabulary for neural machine translation," in *Proc. of ACL-IJCNLP*, July 2015, pp. 1–10.
[29] S. Shen, Y. Cheng, Z. He, W. He, H. Wu, M. Sun, and Y. Liu, "Minimum risk training for neural machine translation," *CoRR*, vol. abs/1512.02433, 2015.
[30] Y. Cheng, Z. Tu, F. Meng, J. Zhai, and Y. Liu, "Towards Robust Neural Machine Translation," *ArXiv e-prints*, May 2018.
[31] M. Wang, Z. Lu, H. Li, and Q. Liu, "Memory-enhanced decoder for neural machine translation," in *Proc. of EMNLP*, Austin, Texas, November 2016, pp. 278–286.
[32] T. Luong, I. Sutskever, Q. Le, O. Vinyals, and W. Zaremba, "Addressing the rare word problem in neural machine translation," in *Proc. of ACL-IJCNLP*, July 2015, pp. 11–19.

**Biao Zhang** received his Bachelor degree in Software Engineering from Xiamen University, and is a graduate student in the School of Software at Xiamen University now. He is supervised by Prof. Hong Duan and Prof. Jinsong Su. His major research interests are natural language processing and deep learning.

**Deyi Xiong** is a Professor at Soochow University. Previously, he was a Research Scientist at the Institute for Infocomm Research of Singapore from 2007–2013. He completed his Ph.D. in computer science at the Institute of Computing Technology of the Chinese Academy of Sciences in 2007. His research interests are in the area of natural language processing, including parsing and statistical machine translation.

**Jinsong Su** was born in 1982, he received the Ph.D. degree in Chinese Academy of Sciences. He is now an associate professor of Software School in Xiamen University. His research interests include natural language processing and statistical machine translation.