A Deep Memory-based Architecture for Sequence-to-Sequence Learning

Fandong Meng[1]* **Zhengdong Lu**[2] **Zhaopeng Tu**[2] **Hang Li**[2] **and Qun Liu**[1]
[1]Institute of Computing Technology, Chinese Academy of Sciences
{mengfandong, liuqun}@ict.ac.cn
[2]Noah's Ark Lab, Huawei Technologies
{Lu.Zhengdong, Tu.Zhaopeng, HangLi.HL}@huawei.com

Abstract

We propose DeepMemory, a novel deep architecture for sequence-to-sequence learning, which performs the task through a series of nonlinear transformations from the representation of the input sequence (e.g., a Chinese sentence) to the final output sequence (e.g., translation to English). Inspired by the recently proposed Neural Turing Machine (Graves et al., 2014), we store the intermediate representations in stacked layers of memories, and use read-write operations on the memories to realize the nonlinear transformations between the representations. The types of transformations are designed in advance but the parameters are learned from data. Through layer-by-layer transformations, DeepMemory can model complicated relations between sequences necessary for applications such as machine translation between distant languages. The architecture can be trained with normal back-propagation on sequence-to-sequence data, and the learning can be easily scaled up to a large corpus. DeepMemory is broad enough to subsume the state-of-the-art neural translation model in (Bahdanau et al., 2015) as its special case, while significantly improving upon the model with its deeper architecture. Remarkably, DeepMemory, being purely neural network-based, can achieve performance comparable to the traditional phrase-based machine translation system Moses with a small vocabulary and a modest parameter size.

1 Introduction

Sequence-to-sequence learning is a fundamental problem in natural language processing, with many important applications such as machine translation (Bahdanau et al., 2015; Sutskever et al., 2014), part-of-speech tagging (Collobert et al., 2011; Vinyals et al., 2014) and dependency parsing (Chen & Manning, 2014). Recently, there has been significant progress in development of technologies for the task using purely neural network-based models. Without loss of generality, we consider machine translation in this paper. Previous efforts on neural machine translation generally fall into two categories:

- **Encoder-Decoder:** As illustrated in left panel of Figure 1, models of this type first summarize the source sentence into a fixed-length vector by the encoder, typically implemented with a recurrent neural network (RNN) or a convolutional neural network (CNN), and then unfold the vector into the target sentence by the decoder, typically implemented with a RNN (Auli et al., 2013; Kalchbrenner & Blunsom, 2013; Cho et al., 2014; Sutskever et al., 2014);

- **Attention-Model:** with RNNsearch (Bahdanau et al., 2015; Luong et al., 2015) as representative, it represents the source sentence as a sequence of vectors after a RNN (e.g., a bi-directional RNN (Schuster & Paliwal, 1997)), and then simultaneously conducts dynamic alignment with a gating neural network and generation of the target sentence with another RNN, as illustrated in right panel of Figure 1.

Empirical comparison between the approaches indicates that the attention-model is more efficient than the encoder-decoder approach: it can achieve comparable results with far less parameters and training instances (Jean et al., 2015). This superiority in efficiency comes mainly from the mechanism of dynamic alignment, which avoids the need to represent the entire source sentence with a fixed-length vector (Sutskever et al., 2014).

*The work was done when the first author worked as intern at Noah's Ark Lab, Huawei Technologies.
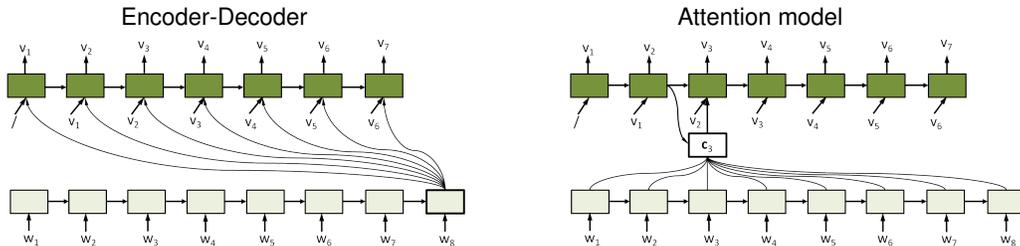
Figure 1: Two types of neural machine translators. Note the pictorial illustrations may deviate from individual models e.g., Sutskever et al. (2014), on modeling details.

1.1 DEEP MEMORY-BASED ARCHITECTURE

Both encoder-decoders and attention models can be reformalized in the language of Neural Turing Machines (NTM) (Graves et al., 2014), by replacing different forms of representations as content in memories and the operations on them as basic neural net-controlled read-write actions, as illustrated in the left panel of Figure 2. This is clear after realizing that the attention mechanism (Bahdanau et al., 2015) is essentially a special case of reading (in particular with content-based addressing) in NTM on the memory that contains the representation of source sentence. More importantly, under this new view, the whole process becomes transforming the source sentence and putting it into memory (vector or array of vectors), and reading from this memory to further transform it into the target sentence. This architecture is intrinsically shallow in terms of the transformations on the sequence as an object, with essentially one hidden "layer", as illustrated in the left panel of Figure 2. Note that although RNN (as encoding/decoding or equivalently as controller in NTM) can be infinitely deep, this depth is merely for dealing with the temporal structure *within* the sequence. On the other hand, many sequence-to-sequence tasks, e.g, translation, are intrinsically complex and calls for more complex and powerful transformation mechanism than that in encoder-decoder and attention models.
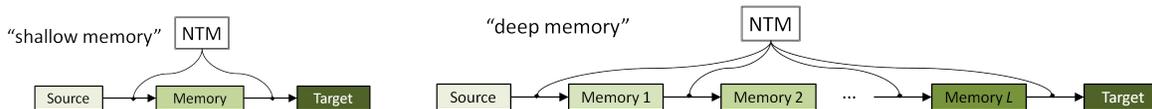

Figure 2: Space holder for NTM view of things

For this reason, we propose a novel deep memory-based architecture, named DEEPMEMORY, for sequence-to-sequence learning. As shown in the right panel of Figure 2, DEEPMEMORY carries out the task through a series of non-linear transformations from the input sequence, to different levels of intermediate memory-based representations, and eventually to the final output sequence. DEEPMEMORY is essentially a customized and deep version of NTM with multiple stages of operations controlled by a program, where the choices of "layers" and types of read/write operations between layers are tailored for a particular task.

Through layer-by-layer stacking of transformations on memory-based representations, DEEPMEMORY generalizes the notion of inter-layer nonlinear mapping in neural networks, and therefore introduces a powerful new deep architecture for sequence-to-sequence learning. The aim of DEEPMEMORY is to learn the representation of sequence better suited to the task (e.g., machine translation) through layer-by-layer transformations. Just as in deep neural network (DNN), we expect that stacking relatively simple transformations can greatly enhance the expressing power and the efficiency of DEEPMEMORY, especially in handling translation between languages with vastly different nature (e.g., Chinese and English) and sentences with complicated structures. DEEPMEMORY naturally subsumes current neural machine translation models (Bahdanau et al., 2015; Sutskever et al., 2014) as special cases, but more importantly it accommodates many deeper alternatives with more modeling power, which are empirically superior to the current shallow architectures on machine translation tasks.

Although DEEPMEMORY is initially proposed for machine translation, it can be adapted for other tasks that require substantial transformations of sequences, including paraphrasing, reasoning (Peng et al., 2015), and semantic parsing (Yin et al., 2015). Also, in defining the layer-by-layer transformations, we can go beyond the read-write operations proposed in (Graves et al., 2014) and design differentiable operations for the specific structures of the task (e.g., in (Yin et al., 2015)).

2

RoadMap We will first discuss in Section 2 the read-write operations as a new form of non-linear transformation, as the building block of DEEPMEMORY. Then in Section 3, we stack the transformations together to get the full DEEPMEMORY architecture, and discuss several architectural variations of it. In Section 4 we report our empirical study of DEEPMEMORY on a Chinese-English translation task.

2 READ-WRITE AS A NONLINEAR TRANSFORMATION

We start with discussing read-write operations between two pieces of memory as a generalized form of nonlinear transformation. As illustrated in Figure 3 (left panel), this transformation is between two *non-overlapping* memories, namely R-memory and W-memory, with W-memory being initially blank. A controller operates the read-heads to get the values from R-memory ("reading"), which are then sent to the write-head for modifying the values at specific locations in W-memory ("writing"). After those operations are completed, the content in R-memory is considered transformed and written to W-memory. These operations therefore define a transformation from one representation (in R-memory) to another (in W-memory), which is pictorially noted in the right panel of Figure 3.
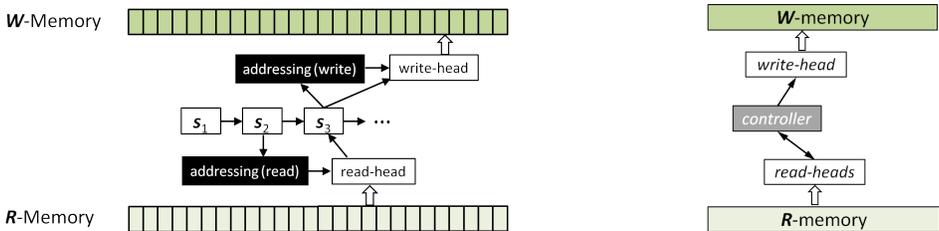

Figure 3: Read-write as an nonlinear transformation.

These basic components are more formally defined below, following those in a generic NTM (Graves et al., 2014), with however important modifications for the nesting architecture, implementation efficiency and description simplicity.

Memory: a memory is generally defined as a matrix with potentially infinite size, while here we limit ourselves to pre-determined (pre-claimed) $N \times d$ matrix, with N memory locations and d values in each location. In our implementation of DEEPMEMORY, N is always instance-dependent and is pre-determined by the algorithm[1]. Memories of different layers generally have different d. Now suppose for one particular instance (index omitted for notational simplicity), the system reads from the R-memory ($\mathcal{M}^{\mathrm{R}}$, with N_{R} units) and writes to W-memory (denoted $\mathcal{M}^{\mathrm{W}}$, with N_{W} units)

$$R\text{-memory: } \mathcal{M}^{\mathrm{R}} = \{\mathbf{x}_1^{\mathrm{R}}, \mathbf{x}_2^{\mathrm{R}}, \cdots, \mathbf{x}_{N_{\mathrm{R}}}^{\mathrm{R}}\}, \quad W\text{-memory: } \mathcal{M}^{\mathrm{W}} = \{\mathbf{x}_1^{\mathrm{W}}, \mathbf{x}_2^{\mathrm{W}}, \cdots, \mathbf{x}_{N_{\mathrm{W}}}^{\mathrm{W}}\},$$

with $\mathbf{x}_n^{\mathrm{R}} \in \mathbb{R}^{d_{\mathrm{R}}}$ and $\mathbf{x}_n^{\mathrm{W}} \in \mathbb{R}^{d_{\mathrm{W}}}$.

Read/write heads: a read-head gets the values from the corresponding memory, following the instructions of the controller, which also influences the controller in feeding the state-machine. DEEPMEMORY allows multiple read-heads for one controller, with potentially different addressing strategies (see Section 2.1 for more details). A write-head simply takes the instruction from controller and modifies the values at specific locations.

Controller: The core to the controller is a state machine, implemented as a RNN Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997), with state at time t denoted as $\mathbf{s}_t$ (as illustrated in Figure 3). With $\mathbf{s}_t$, the controller determines the reading and writing at time t, while the return of reading in turn takes part in updating the state. For simplicity, only one reading and writing is allowed at one time step, but more than one read-heads are allowed. The main equations for controllers are then

$$\begin{aligned} \text{READ VECTOR:} \quad & \mathbf{r}_t = F_{\mathrm{R}}(\mathcal{M}^{\mathrm{R}}, \mathbf{s}_t;\ \Theta_{\mathrm{R}}) \\ \text{WRITE VECTOR:} \quad & \mathbf{v}_t = F_{\mathrm{W}}(\mathbf{s}_t;\ \Theta_{\mathrm{W}}) \\ \text{STATE UPDATE:} \quad & \mathbf{s}_{t+1} = F_{\mathrm{D}}(\mathbf{s}_t, \mathbf{r}_t;\ \Theta_{\mathrm{D}}) \end{aligned}$$

---

[1]It is possible to let the controller learn to determine the length of the memory, but that does not yield better performance on our tasks and is therefore omitted here.

where $F_D(\cdot)$, $F_R(\cdot)$ and $F_W(\cdot)$ are respectively the operators for dynamics, reading and writing[2], parameterized by Θ_D, Θ_R, and Θ_W. In DEEPMEMORY, 1) it is only allowed to read from memory of lower layer and write to memory of higher layer, and 2) reading memory can only be performed after finishing the writing to it.

The above read-write operations transform the representation in R-memory to the new representation in W-memory, while the design choice of read-write specifies the inner structure of the memories in W-memory. The transformation is therefore jointly specified by the read-write strategies (e.g., the addressing described in Section 2.1) and the parameters learned in a supervised fashion (described later in Section 3.2). Memory with the designed inner structure, in this particular case vector array with instance specific length, offers more flexibility than a fixed-length vector in representing sequences. This representational flexibility is particularly advantageous when combined with proper reading-writing strategy in defining nonlinear transformations for sequences, which will serve as the building block of the deep architecture.

2.1 ADDRESSING

2.1.1 ADDRESSING FOR READING

Location-based Addressing With location-based addressing (L-addressing), the reading is simply $\mathbf{r}_t = \mathbf{x}_t^R$. Notice that with L-addressing, the state machine automatically runs on a clock determined by the spatial structure of R-memory. Following this clock, the write-head operates the same number of times. One important variant, as suggested in (Bahdanau et al., 2015; Sutskever et al., 2014), is to go through R-memory backwards after the forward reading pass, where the controller RNN has the same structure but is parameterized differently.

Content-based Addressing With a content-based addressing (C-addressing), the return at t is

$$\mathbf{r}_t = F_R(\mathcal{M}^R, \mathbf{s}_t; \Theta_R) = \sum_{n=1}^{N_R} \tilde{g}(\mathbf{s}_t, \mathbf{x}_n^R; \Theta_R)\mathbf{x}_n^R, \quad \text{and} \quad \tilde{g}(\mathbf{s}_t, \mathbf{x}_n^R; \Theta_R) = \frac{g(\mathbf{s}_t, \mathbf{x}_n^R; \Theta_R)}{\sum_{n'=1}^{N_R} g(\mathbf{s}_t, \mathbf{x}_{n'}^R; \Theta_R)},$$

where $g(\mathbf{s}_t, \mathbf{x}_n^R; \Theta_R)$, implemented as a DNN, gives an un-normalized "affliiation" score for unit $\mathbf{x}_n^R$ in R-memory. Clearly it is related to the attention mechanism introduced in (Bahdanau et al., 2015) for machine translation and general attention models discussed in (Gregor et al., 2015) for computer vision. Content-based addressing offers the following two advantages in representation learning:

1. it can focus on the right segment of the representation, as demonstrated by the automatic alignment observed in (Bahdanau et al., 2015), therefore better preserving the information in lower layers;
2. it provides a way to alter the spatial structure of the sequence representation on a large scale, for which the re-ordering in machine translation is an intuitive example.

Hybrid Addressing: With hybrid addressing (H-addressing) for reading, we essentially use two read-heads (can be easily extended to more), one with L-addressing and the other with C-addressing. At each time t, the controller simply concatenates the return of two individual read-heads as the final return:

$$\mathbf{r}_t = [\mathbf{x}_t^R, \ \sum_{n=1}^{N_R} g(\mathbf{s}_t, \mathbf{x}_n^R; \Theta_R)\mathbf{x}_n^R].$$

It is worth noting that with H-addressing, the tempo of the state machine will be determined by the L-addressing read-head, and therefore creates W-memory of the same number of locations in writing. As shown later, H-addressing can be readily extended to allow read-heads to work on different memories.

2.1.2 ADDRESSING FOR WRITING

Location-based Addressing With L-addressing, the writing is simple. At any time t, only the tth location in W-memory is updated: $\mathbf{x}_t^W = \mathbf{v}_t \stackrel{\text{def}}{=} F_W(\mathbf{s}_t; \Theta_W)$, which will be kept unchanged afterwards. For both location- and content-based addressing, $F_W(\mathbf{s}_t; \Theta_W)$ is implemented as a DNN with weights Θ_W.

Content-based Addressing In a way similar to C-addressing for reading, the units to write is determined through a gating network $g(\mathbf{s}_t, \mathbf{x}_{n,t}^W; \Theta_W)$, where the values in W-memory at time t is given by

$$\Delta_{n,t} = \tilde{g}(\mathbf{s}_t, \mathbf{x}_{n,t}^W; \Theta_W)F_W(\mathbf{s}_t; \Theta_W), \quad \mathbf{x}_{n,t}^W = (1 - \alpha_t)\mathbf{x}_{n,t-1}^W + \alpha_t\Delta_{n,t}, \ n = 1, 2, \cdots, N_W,$$

---

[2]Note that our definition of writing is slightly different from that in (Graves et al., 2014).

4

where $\mathbf{x}_{n,t}^{\mathrm{W}}$ stands for the values of the nth location in W-memory at time t, α is the forgetting factor (similarly defined as in (Graves et al., 2014)), $\tilde{g}$ is the normalized weight (with unnormalized score implemented also with a DNN) given to the nth location at time t.

2.2 Types of Nonlinear Transformations

As the most "conventional" special case, if we use L-addressing for both reading and writing, we actually get the familiar structure in units found in RNN with stacked layers (Pascanu et al., 2014). Indeed, as illustrated in the figure right to the text, this read-write strategy will invoke a relatively local dependency based on the original spatial order in R-memory. It is not hard to show that we can recover some deep RNN model in (Pascanu et al., 2014) after stacking layers of read-write operations like this. This deep architecture actually partially accounts for the great performance of the Google neural machine translation model (Sutskever et al., 2014).
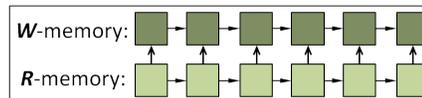
The C-addressing, however, be it for reading and writing, offers a means of major reordering on the units, while H-addressing can add into it the spatial structure of the lower layer memory. In this paper, we consider four types of transformations induced by combinations of the read and write addressing strategies, listed pictorially in Figure 4. Notice that 1) we only include one combination with C-addressing for writing since it is computationally expensive to optimize when combined with a C-addressing reading (see Section 3.2 for some analysis) , and 2) for one particular read-write strategy there are still a fair amount of implementation details to be specified, which are omitted due to the space limit. One can easily design different read/write strategies, for example a particular way of H-addressing for writing.
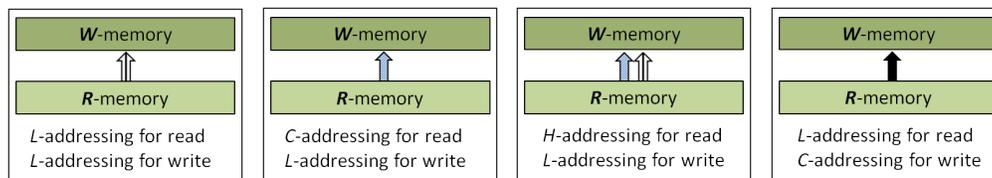
Figure 4: Examples of read-write strategies.

3 DeepMemory: Stacking Them Together

As illustrated in Figure 5 (left panel), the stacking is straightforward: we can just apply a transformation on top of another, with the W-memory in lower layer being the R-memory of upper layer. The entire deep architecture of DeepMemory, with diagram in Figure 5 (right panel), can be therefore defined accordingly. Basically, it starts with a symbol sequence (Layer-0), then moves to the sequence of word embeddings (Layer-1), through layers of transformation to reach the final intermeidate layer (Layer-L), which will be read by the output layer.

The operations in output layer, relying on another LSTM to generate the target sequence, are similar to a memory read-write, with the following two differences:

- it predicts the symbols for the target sequence, and takes the "guess" as part of the input to update the state of the generating LSTM, while in a memory read-write, there is no information flow from higher layers to the controller;

- since the target sequence in general has different length as the top-layer memory, it takes only pure C-addressing reading and relies on the built-in mechanism of the generating LSTM to stop (i.e., after generating a `End-of-Sentence` token).

Memory of different layers could be equipped with different read-write strategies, and even for the same strategy, the configurations and learned parameters are in general different. This is in contrast to DNNs, for which the transformations of different layers are more homogeneous (mostly linear transforms with nonlinear activation function). A sensible architecture design in combining the nonlinear transformations can greatly affect the performance of the model, on which however little is known and future research is needed.
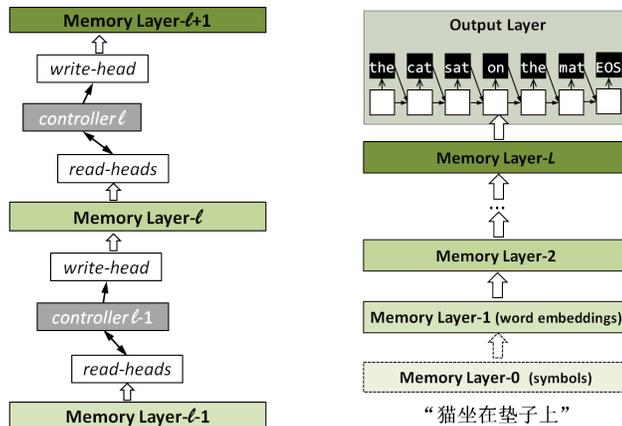
Figure 5: Illustration of stacked layers of memory (left) and the overall diagram of DEEPMEMORY (right).

3.1 CROSS-LAYER READING

In addition to the generic read-write strategies in Section 2.2, we also introduce the cross-layer reading into DEEPMEMORY for more modeling flexibility. In other words, for writing in any Layer-ℓ, DEEPMEMORY allows reading from more than one layers lower than ℓ, instead of just Layer-$\ell-1$. More specifically, we consider the following two cases.

MEMORY-BUNDLE: A MEMORY-BUNDLE, as shown in Figure 6 (left panel), concatenates the units of two *aligned* memories in reading, regardless of the addressing strategy. Formally, the nth location in the bundle of memory Layer-ℓ' and Layer-ℓ'' would be $\mathbf{x}_n^{(\ell'+\ell'')} = [(\mathbf{x}_n^{\ell'})^\top, (\mathbf{x}_n^{\ell''})^\top]^\top$. Since it requires strict alignment between the memories to put together, MEMORY-BUNDLE is usually on layers created with spatial structure of same origin (see Section 3.3 for examples).
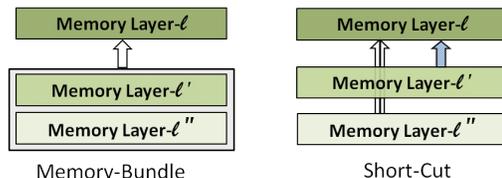

Figure 6: Cross-layer reading.

SHORT-CUT Unlike MEMORY-BUNDLE, SHORT-CUT allows reading from layers with potentially different inner structures by using multiple read-heads, as shown in Figure 6 (right panel). For example, one can use a C-addressing read-head on memory Layer-ℓ' and a L-addressing read-head on Layer-ℓ'' for the writing to memory Layer-ℓ with $\ell', \ell'' < \ell$.

3.2 OPTIMIZATION

For any designed architecture, the parameters to be optimized include $\{\Theta_D, \Theta_R, \Theta_W\}$ for each controller, the parameters for the LSTM in the output layer, and the word-embeddings. Since the reading from each memory can only be done after the writing on it completes, the "feed-forward" process can be described in two scales: 1) the flow from memory of lower layer to memory of higher layer, and 2) the forming of a memory at each layer controlled by the corresponding state machine. Accordingly in optimization, the flow of "correction signal" also propagates at two scales:

- On the "cross-layer" scale: the signal starts with the output layer and propagates from higher layers to lower layers, until Layer-1 for the tuning of word embedding;

- On the "within-layer" scale: the signal back-propagates through time (BPTT) controlled by the corresponding state-machine (LSTM). In optimization, there is a correction for each reading or writing on each location in a memory, making the C-addressing more expensive than L-addressing for it in general involves all locations in the memory at each time t.

The optimization can be done via the standard back-propagation (BP) aiming to maximize the likelihood of the target sequence. In practice, we use the standard stochastic gradient descent (SGD) and mini-batch (size 80) with learning rate controlled by AdaDelta (Zeiler, 2012).

6

3.3 ARCHITECTURAL VARIATIONS OF DEEPMEMORY

We discuss four representative special cases of DEEPMEMORY: ARC-I, II, III and IV, as novel deep architectures for machine translation. We also show that current neural machine translation models like RNNsearch can be described in the framework of DEEPMEMORY as a relatively shallow case.

ARC-I The first proposal, including two variants (ARC-I_{HYB} and ARC-I_{LOC}), is designed to demonstrate the effect of C-addressing reading between intermediate memory layers, with diagram shown in the figure right to the text. Both variants employ a L-addressing reading from memory Layer-1 (the embedding layer) and L-addressing writing to Layer-2. After that, ARC-I_{HYB} writes to Layer-3 (L-addressing) based on its H-addressing reading (two read-heads) on Layer-2, while ARC-I_{LOC} uses L-addressing to read from Layer-2. Once Layer-3 is formed, it is then put together with Layer-2 for a MEMORY-BUNDLE, from which the output layer reads (C-addressing) for predicting the target sequence. MEMORY-BUNDLE, with its empirical advantage over single layers (see Section 4.2), is also used in other three architectures for generating the target sequence or forming intermediate layers.

ARC-II As an architecture similar to ARC-I_{HYB}, ARC-II is designed to investigate the effect of H-addressing reading from different layers of memory (or SHORT-CUT in Section 3.1). It uses the same strategy as ARC-I_{HYB} in generating memory Layer-1 and 2, but differs in generating Layer-3, where ARC-II uses C-addressing reading on Layer-2 but L-addressing reading on Layer-1. Once Layer-3 is formed, it is then put together with Layer-2 as a MEMORY-BUNDLE, which is then read by the output layer for predicting the target sequence.

ARC-III We intend to use this design to study a deeper architecture and more complicated addressing strategy. ARC-III follows the same way as ARC-II to generate Layer-1, Layer-2 and Layer-3. After that it uses two read-heads combined with a L-addressing write to generate Layer-4, where the two read-heads consist of a L-addressing read-head on Layer-1 and a C-addressing read-head on the memory bundle of Layer-2 and Layer-3. After the generation of Layer-4, it puts Layer-2, 3 and 4 together for a bigger MEMORY-BUNDLE to the output layer. ARC-III, with 4 intermediate layers, is the deepest among the four special cases.

Arc-IV This proposal is designed to study the efficacy of C-addressing writing in forming intermediate representation. It employs a L-addressing reading from memory Layer-1 and L-addressing writing to Layer-2. After that, it uses a L-addressing reading on Layer-2 to write to Layer-3 with C-addressing. For C-addressing writing to Layer-3, all locations in Layer-3 are randomly initialized. Once Layer-3 is formed, it is then bundled with Layer-2 for the reading (C-addressing) of the output layer.
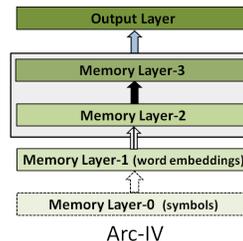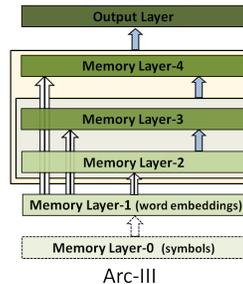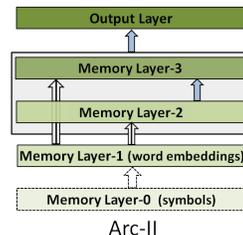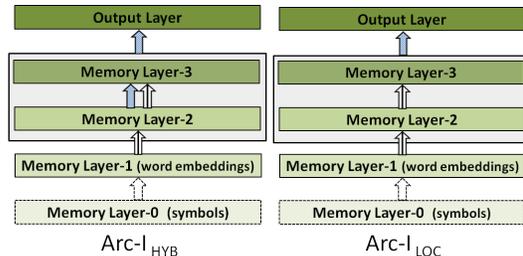
3.3.1 RELATION TO OTHER NEURAL MACHINE TRANSLATORS

As pointed out earlier, RNNsearch (Bahdanau et al., 2015) with its automatic alignment, is a special case of DEEPMEMORY with shallow architecture. As pictorially illustrated in Figure 7, it employs L-addressing reading on memory Layer-1 (the embedding layer), and L-addressing writing to Layer-2, which then is read (C-addressing) by the output layer to generate the target sequence. As shown in Figure 7, Layer-2 is the only intermediate layer created by nontrivial read-write operations.

On the other hand, the connection between DEEPMEMORY and encoder-decoder architectures is less obvious since they usually require the reading from only the last cell (i.e., for a fixed-length vector representation) between certain layers. More specifically, Sutskever et al. (2014) can be viewed as DeepMemory with stacking layers of L-addressing read-write (described in Section 2.2) for both the encoder and decoder part, while the two are actually connected through last hidden states of the LSTMs of the corresponding layers.

7
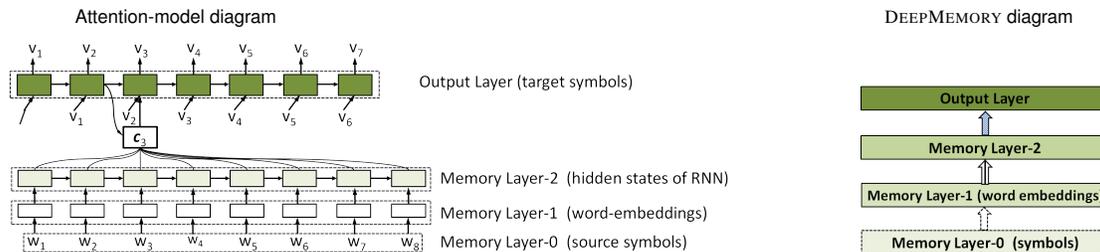
Figure 7: RNNsearch as a special case of DEEPMEMORY.

4 EXPERIMENTS

We report our empirical study on applying DEEPMEMORY to Chinese-to-English translation. Our training data consist of 1.25M sentence pairs extracted from LDC corpora, with 27.9M Chinese words and 34.5M English words respectively. We choose NIST 2002 (MT02) dataset as our development set, and the NIST 2003 (MT03), 2004 (MT04) and 2005 (MT05) datasets as our test sets. We use the case-insensitive 4-gram NIST BLEU score as our evaluation metric, and *sign-test* (Collins et al., 2005) as statistical significance test. In training of the neural networks, we limit the source and target vocabularies to the most frequent 16K words in Chinese and English, covering approximately 95.8% and 98.3% of the two corpora respectively.

We compare our method with two state-of-the-art SMT and NMT[3] models:

- **Moses** (Koehn et al., 2007): an open source phrase-based translation system with default configuration and a 4-gram language model trained on the target portion of training data with (Stolcke et al., 2002);

- **RNNsearch** (Bahdanau et al., 2015): an attention-based NMT model with default setting (RNNsearch$_{DEFAULT}$), as well as an optimal re-scaling of the model (on sizes of both embedding and hidden layers, with about 50% more parameters) (RNNsearch$_{BEST}$).

For a fair comparison, 1) the output layer in each DEEPMEMORY variant is implemented as Gated Recurrent Units (GRU) in (Bahdanau et al., 2015), and 2) all the DEEPMEMORY architectures are designed to have the same embedding size as in RNNsearch$_{DEFAULT}$ with parameter size less or comparable to RNNsearch$_{BEST}$.

4.1 RESULTS

The main results of different models are given in Table 1. RNNsearch (best) is about 1.7 points behind Moses in BLEU on average, which is consistent with the observations made by other authors on different machine translation tasks (Bahdanau et al., 2015; Jean et al., 2015). Remarkably, some sensible designs of DEEPMEMORY (e.g., ARC-II) can already achieve performance comparable to Moses, with only 42M parameters, while RNNsearch$_{BEST}$ has 46M parameters.

Clearly all DEEPMEMORY architectures yield performance significantly better than (ARC-I$_{HYB}$, ARC-II & ARC-III) or comparable (ARC-I$_{LOC}$ & ARC-IV) to the NMT baselines. Among them, ARC-II outperforms the best setting of NMT baseline (RNNsearch$_{BEST}$), by about 1.5 BLEU on average with less parameters.

SYSTEMS	MT03	MT04	MT05	AVERAGE	PARAMETERS #
RNNsearch$_{DEFAULT}$	29.02	31.25	28.32	29.53	31M
RNNsearch$_{BEST}$	30.28	31.72	28.52	30.17	46M
ARC-I$_{LOC}$	28.98	32.02	29.53*	30.18	54M
ARC-I$_{HYB}$	30.14	32.70*	29.40*	30.75	54M
ARC-II	**31.27***	33.02*	**30.63***	31.64	42M
ARC-III	30.15	**33.46***	29.49*	31.03	53M
ARC-IV	29.88	32.00	28.76	30.21	48M
Moses	31.61	33.48	30.75	31.95	–

Table 1: BLEU-4 scores (%) of NMT baselines: RNNsearch$_{DEFAULT}$ and RNNsearch$_{BEST}$, DEEPMEMORY architectures (ARC-I, II, III and IV), and phrase-based SMT system (Moses). The "*" indicates that the results are significantly (p<0.05) better than those of the RNNsearch$_{BEST}$.

---

[3]There are recent progress on aggregating multiple models or enlarging the vocabulary(e.g., in (Jean et al., 2015)), but here we focus on the generic models.

4.2 Discussion

About Depth: A more detailed comparison between RNNsearch (two layers), Arc-II (three layers) and Arc-III) (four layers), both quantitatively and qualitatively, suggests that with deep architectures are essential to the superior performance of DeepMemory. Although the deepest architecture Arc-III is about 0.6 BLEU behind the Arc-II, its performance on long sentences is significant better. Figure 8 shows the BLEU scores of generated translations on the test sets with respect to the length of the source sentences. In particular, we test the BLEU scores on sentences longer than $\{0, 10, 20, 30, 40, 50, 60\}$ in the merged test set of MT03, MT04 and MT05. Clearly, on sentences with length >30, Arc-III yields consistently higher BLEU scores than Arc-II. This observation is further confirmed by our observations of translation quality (see Appendix) and is consistent with our intuition that DeepMemory with its multiple layers of transformation, is especially good at modeling the transformations of representations essential to machine translation on relatively complicated sentences.
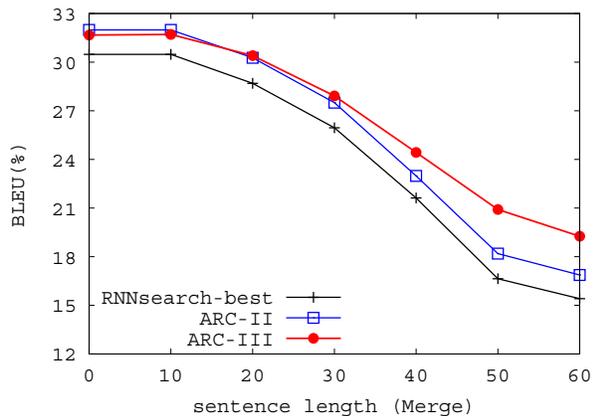

Figure 8: The BLEU scores of generated translations on the merged three test sets with respect to the lengths of source sentences. The numbers on X-axis of the figure stand for sentences *longer than* the corresponding length, e.g., 30 for source sentences with > 30 words.

About C-addressing Read: Further comparison between Arc-I$_{\mathrm{HYB}}$ and Arc-I$_{\mathrm{LOC}}$ (similar parameter sizes) suggests that C-addressing reading plays an important role in learning a powerful transformation between intermediate representations, necessary for translation between language pairs with vastly different syntactical structures. This conjecture is further verified by the good performances of Arc-II and Arc-III, both of which have C-addressing read-heads in their intermediate memory layers. However, if memory Layer-ℓ+1 is formed with only C-addressing read from memory Layer-ℓ, and serves as the only going to later stages, the performance is usually less satisfying. Comparison of this design with H-addressing (results omitted here) suggests that another read-head with L-addressing can prevent the transformation from going astray by adding the tempo from a memory with a clearer temporal structure.

About C-addressing Write: The BLEU scores of Arc-IV are lower than that of Arc-II but comparable to that of RNNsearch$_{\mathrm{BEST}}$, suggesting that writing with C-addressing alone yields reasonable representation. A closer look shows that although Arc-IV performs poorly on very long sentences (e.g., source sentences with over 60 words), it does fairly well on sentences with normal length. More specifically, on source sentences with length ≤ 40, it outperforms RNNsearch$_{\mathrm{BEST}}$ with 0.79 BLEU points. One possible explanation is that our particular implementation of C-addressing for writing in Arc-IV (Section 3.3) relies heavily on the randomly initialized content and is hard to optimize, especially when the structure of the sentence is complex, which might need to be "guided" by another write-head or some smart initialization.

About Cross-layer Read: As another observation, cross-layer reading almost always helps. The performances of Arc-I, II, III and IV unanimously drop after removing the Memory-Bundle and Short-Cut (results omitted here), even after the broadening of memory units to keep the parameter size unchanged. It might be due to the flexibility gained in mixing different addressing modes and representations of different stages.

5 Conclusion

We propose DeepMemory, a novel architecture for sequence-to-sequence learning, which is stimulated by the recent work of Neural Turing Machine (Graves et al., 2014) and Neural Machine Translation (Bahdanau et al., 2015). DeepMemory builds its deep architecture for processing sequence data on the basis of a series of transformations induced by the read-write operations on a stack of memories. This new architecture significantly improves the expressive power of models in sequence-to-sequence learning, which is verified by our empirical study on a benchmark machine translation task.

REFERENCES

Michael Auli, Michel Galley, Chris Quirk, and Geoffrey Zweig. Joint language and translation modeling with recurrent neural networks. In *Proceedings of EMNLP*, pp. 1044–1054, 2013.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*, 2015.

Danqi Chen and Christopher D Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of EMNLP*, pp. 740–750, 2014.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of EMNLP*, pp. 1724–1734, 2014.

Michael Collins, Philipp Koehn, and Ivona Kučerová. Clause restructuring for statistical machine translation. In *Proceedings of ACL*, pp. 531–540, 2005.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537, 2011.

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

Karol Gregor, Ivo Danihelka, Alex Graves, and Daan Wierstra. DRAW: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. On using very large target vocabulary for neural machine translation. In *ACL-IJNLP*, 2015.

Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of EMNLP*, pp. 1700–1709, 2013.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL on interactive poster and demonstration sessions*, pp. 177–180, Prague, Czech Republic, June 2007.

Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, 2015.

Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. How to construct deep recurrent neural networks. In *Proceedings of ICLR*, 2014.

Baolin Peng, Zhengdong Lu, Hang Li, and Kam-Fai Wong. Towards neural network-based reasoning. *arXiv preprint arXiv:1508.05508*, 2015.

Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*, 45(11):2673–2681, 1997.

Andreas Stolcke et al. SRILM-an extensible language modeling toolkit. In *Proceedings of ICSLP*, volume 2, pp. 901–904, 2002.

Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pp. 3104–3112, 2014.

Oriol Vinyals, Lukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. Grammar as a foreign language. *arXiv preprint arXiv:1412.7449*, 2014.

Pengcheng Yin, Zhengdong Lu, Hang Li, and Kao Ben. Neural enquirer: Learning to query tables. *arXiv preprint arXiv:1512.00965*, 2015.

Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

APPENDIX: ACTUAL TRANSLATION EXAMPLES

In appendix we give some example translations from DEEPMEMORY, more specically, ARC-II and ARC-III, and compare them against the reference and the translation given by RNNsearch. We will focus on long sentences with relatively complicated structures.

EXAMPLE TRANSLATION OF ARC-II

Source	反对派 国会 议员 昨天 拒绝 就 政府 提出 的 宪法 修正案 进行 表决
Reference	yesterday opposition parliament members refused to vote on the constitutional amendment bill proposed by the government
RNNsearch	the members of the opposition parliament have refused to accept the decision on the amendment of the constitution
DeepMemory (Arc-II)	the opposition parliament members yesterday refused to vote on the constitutional amendment proposed by the government
Source	谢瓦尔德纳泽 在 讲话 中 强调 了 在 此 问题 上 与 俄 美 两 国 进行 合作 和 交换 情报 的 重要性 。
Reference	in his address, shevardnadze stressed the importance of the exchange of intelligence with russia and the united states on this matter.
RNNsearch	in his speech , UNK stressed the importance of cooperation and exchange between the two countries on this question .
DeepMemory (Arc-II)	in his speech , UNK stressed the importance of cooperation and exchange of intelligence between russia and the united states on this issue .
Source	博尔久扎 指出 , 独联体 集体 安全 条约 组织 快速 反应 部队 的 建立 有助于 遏制 中亚 地区 的 极端 组织 势力
Reference	bordyuzha noted that the creation of a fast reaction force by the cis collective security treaty organization would help contain the strength of extremist groups in central asia
RNNsearch	UNK pointed out that the establishment of the organization of the commonwealth of independent states (cis) is conducive to the maintenance of extremist organizations in the central asian region
DeepMemory (Arc-II)	UNK noted that the establishment of a rapid reaction force in the organization of collective security of the commonwealth of independent states is conducive to curbing the extremist forces in the central asian region
Source	英格兰 说 , 五角大楼 是 在 一 、 两 天 前 确认 这 名 囚犯 身份 的 , 国务院 将 就 这 一 决定 通报 囚犯 所 属 的 国家 并 安排 有关 释放 事宜 。
Reference	england said the pentagon verified the identity of this prisoner one or two days ago and that the state department will notify the home country of the prisoner of this decision and make arrangements for his release.
RNNsearch	england said that the pentagon was established in a and two days ago that the state council would inform the state of the decision and arrangements for the release of the detainees .
DeepMemory (Arc-II)	england said that the pentagon had confirmed the identity of the prisoners before one or two days ago , and the state council would notify the countries involved in the decision and arrange the release of the prisoners .

EXAMPLE TRANSLATION OF ARC-III

Source	美国 副国务卿 阿米塔吉 今天 对 叙利亚 发出 警告 , 指称 大马士革 与 华府 未来 关系 的 发展 , 端 看 叙利亚 是否 愿 在 伊拉克 及 黎巴嫩 问题 上 与 美国 合作 。
Reference	us deputy secretary of state armitage issued warnings to syria today, saying the development of future relations between damascus and washington depends on whether syria wants to cooperate with the united states on the issues of iraq and lebanon.
RNNsearch	us deputy secretary of state richard armitage issued a warning to syria today that the development of damascus , washington , and washington could see if syria is willing to cooperate with the united states in iraq and lebanon .
DeepMemory (Arc-III)	us deputy secretary of state richard armitage issued a warning today that the development of the damascus - washington relations in the future will depend on whether syria is willing to cooperate with the united states on iraq and lebanon issues .
Source	俄罗斯 国营 石油 公司 正 与 俄罗斯 国营 瓦斯 公司 进行 合并 , 加上 尤干斯克 石油 天然气 公司 后 , 三 者 合计 将 掌控 俄罗斯 石油 总 生产 的 五分之一 以及 全部 天然气 生产 , 一举 成为 全球 规模 最 大 能源 集团 。
Reference	the russian state petroleum company is merging with the russian state gas company, and with the inclusion of yuganskneftegaz, the combined three will account for a fifth of russia's oil output and all its natural gas production, making it the largest global energy group.
RNNsearch	the russian state - run oil company is merging with the state - owned gas company , and the state - owned gas corp .
DeepMemory (Arc-III)	russia 's state - owned petroleum company is merging with the russian state - owned gas company , and the third country will have to control the one - fifth of russian oil production and to become the biggest energy group in the world .
Source	希 土 两 国 同 为 北约 成员国 , 但 双方 在 爱琴海 领海 、 领空 以及 岛屿 大陆架 的 划分 和 塞浦路斯 等 问题 上 存在 利益 冲突 和 矛盾 , 也 影响 了 两 国 关系 的 正常化 。
Reference	both greece and turkey are nato members, but the two have conflicts of interest and clashes over maritime and air territories and continental shelf boundary disputes in the aegean sea and the issue of cyprus, which have affected the normalization of their bilateral relations.
RNNsearch	the two countries have maintained conflicts and contradictions in the fields of territorial waters , airspace , and islands , as well as contradictions and conflicts .
DeepMemory (Arc-III)	greece and turkey are the member countries of the nato , but the conflict of interests and contradictions between the two countries on the demarcation of the territorial waters , airspace , and the islands of the islands also adversely affect the normalization of relations between the two countries .