Gated-Attention Readers for Text Comprehension

Bhuwan Dhingra∗ Hanxiao Liu∗ Zhilin Yang
William W. Cohen Ruslan Salakhutdinov
School of Computer Science
Carnegie Mellon University
{bdhingra,hanxiaol,zhiliny,wcohen,rsalakhu}@cs.cmu.edu

Abstract

In this paper we study the problem of answering cloze-style questions over documents. Our model, the Gated-Attention (GA) Reader, integrates a multi-hop architecture with a novel attention mechanism, which is based on multiplicative interactions between the query embedding and the intermediate states of a recurrent neural network document reader. This enables the reader to build query-specific representations of tokens in the document for accurate answer selection. The GA Reader obtains state-of-the-art results on three benchmarks for this task—the CNN & Daily Mail news stories and the Who Did What dataset. The effectiveness of multiplicative interaction is demonstrated by an ablation study, and by comparing to alternative compositional operators for implementing the gated-attention.

1 Introduction

A recent trend to measure progress towards machine reading is to test a system’s ability to answer questions about a document it has to comprehend. Towards this end, several large-scale datasets of cloze-style questions over a context document have been introduced recently, which allow the training of supervised machine learning systems (Hermann et al., 2015; Hill et al., 2016; Onishi et al., 2016). Such datasets can be easily constructed automatically and the unambiguous nature of their queries provides an objective benchmark to measure a system’s performance at text comprehension.

Deep learning models have been shown to outperform traditional shallow approaches on text comprehension tasks (Hermann et al., 2015). The success of many recent models can be attributed primarily to two factors: (1) Multi-hop architectures (Weston et al., 2015; Sordoni et al., 2016; Shen et al., 2016), allow a model to scan the document and the question iteratively for multiple passes. (2) Attention mechanisms (Chen et al., 2016; Hermann et al., 2015) borrowed from the machine translation literature (Bahdanau et al., 2014), allow the model to focus on appropriate subparts of the context document. Intuitively, the multi-hop architecture allows the reader to incrementally refine token representations, and the attention mechanism re-weights different parts in the document according to their relevance to the query.

The effectiveness of multi-hop reasoning and attentions have been explored orthogonally so far in the literature. In this paper, we focus on combining both in a complementary manner, by designing a novel attention mechanism which gates the evolving token representations across hops. More specifically, unlike existing models where the query attention is applied either token-wise (Hermann et al., 2015; Kadlec et al., 2016; Chen et al., 2016; Hill et al., 2016) or sentence-wise (Weston et al., 2015; Sukhbaatar et al., 2015) to allow weighted aggregation, the Gated-Attention (GA) module proposed in this work allows the query to directly interact with each dimension of the token embeddings at the semantic-level, and is applied layer-wise as information filters during the multi-hop representation learning process. Such a fine-grained attention enables our model to learn conditional token representations w.r.t. the given question, leading to accurate answer selections.

We show in our experiments that the proposed GA reader, despite its relative simplicity, consis-
tently improves over a variety of strong baselines on three benchmark datasets. Our key contribution, the GA module, provides a significant improvement for large datasets. Qualitatively, visualization of the attentions at intermediate layers of the GA reader shows that in each layer the GA reader attends to distinct salient aspects of the query which help in determining the answer.

2 Related Work

The cloze-style QA task involves tuples of the form \((d, q, a, C)\), where \(d\) is a document (context), \(q\) is a query over the contents of \(d\), in which a phrase is replaced with a placeholder, and \(a\) is the answer to \(q\), which comes from a set of candidates \(C\). In this work we consider datasets where each candidate \(c \in C\) has at least one token which also appears in the document. The task can then be described as: given a document-query pair \((d, q)\), find \(a \in C\) which answers \(q\). Below we provide an overview of representative neural network architectures which have been applied to this problem.

LSTMs with Attention: Several architectures introduced in Hermann et al. (2015) employ LSTM units to compute a combined document-query representation \(g(d, q)\), which is used to rank the candidate answers. These include the DeepLSTM Reader which performs a single forward pass through the concatenated \((document, query)\) pair to obtain \(g(d, q)\); the Attentive Reader which first computes a document vector \(d(q)\) by a weighted aggregation of words according to attentions based on \(q\), and then combines \(d(q)\) and \(q\) to obtain their joint representation \(g(d(q), q)\); and the Impatient Reader where the document representation is built incrementally. The architecture of the Attentive Reader has been simplified recently in Stanford Attentive Reader, where shallower recurrent units were used with a bilinear form for the query-document attention (Chen et al., 2016).

Attention Sum: The Attention-Sum (AS) Reader (Kadlec et al., 2016) uses two bidirectional GRU networks (Cho et al., 2015) to encode both \(d\) and \(q\) into vectors. A probability distribution over the entities in \(d\) is obtained by computing dot products between \(q\) and the entity embeddings and taking a softmax. Then, an aggregation scheme named pointer-sum attention is further applied to sum the probabilities of the same entity, so that frequent entities the document will be favored compared to rare ones. Building on the AS Reader, the Attention-over-Attention (AoA) Reader (Cui et al., 2017) introduces a two-way attention mechanism where the query and the document are mutually attentive to each other.

Mulit-hop Architectures: Memory Networks (MemNets) were proposed in Weston et al. (2015), where each sentence in the document is encoded to a memory by aggregating nearby words. Attention over the memory slots given the query is used to compute an overall memory and to renew the query representation over multiple iterations, allowing certain types of reasoning over the salient facts in the memory and the query. Neural Semantic Encoders (NSE) (Munkhdalai & Yu, 2017a) extended MemNets by introducing a write operation which can evolve the memory over time during the course of reading. Iterative reasoning has been found effective in several more recent models, including the Iterative Attentive Reader (Sordoni et al., 2016) and ReasoNet (Shen et al., 2016). The latter allows dynamic reasoning steps and is trained with reinforcement learning.

Other related works include Dynamic Entity Representation network (DER) (Kobayashi et al., 2016), which builds dynamic representations of the candidate answers while reading the document, and accumulates the information about an entity by max-pooling; EpiReader (Trischler et al., 2016) consists of two networks, where one proposes a small set of candidate answers, and the other reranks the proposed candidates conditioned on the query and the context; Bi-Directional Attention Flow network (BiDAF) (Seo et al., 2017) adopts a multi-stage hierarchical architecture along with a flow-based attention mechanism; Bajgar et al. (2016) showed a 10% improvement on the CBT corpus (Hill et al., 2016) by training the AS Reader on an augmented training set of about 14 million examples, making a case for the community to exploit data abundance. The focus of this paper, however, is on designing models which exploit the available data efficiently.

3 Gated-Attention Reader

Our proposed GA readers perform multiple hops over the document (context), similar to the Memory Networks architecture (Sukhbaatar et al., 2015). Multi-hop architectures mimic the multi-step comprehension process of human readers, and have shown promising results in several recent models for text comprehension (Sordoni et al.,
The contextual representations in GA readers, namely the embeddings of words in the document, are iteratively refined across hops until reaching a final attention-sum module (Kadlec et al., 2016) which maps the contextual representations in the last hop to a probability distribution over candidate answers.

The attention mechanism has been introduced recently to model human focus, leading to significant improvement in machine translation and image captioning (Bahdanau et al., 2014; Mnih et al., 2014). In reading comprehension tasks, ideally, the semantic meanings carried by the contextual embeddings should be aware of the query across hops. As an example, human readers are able to keep the question in mind during multiple passes of reading, to successively mask away information irrelevant to the query. However, existing neural network readers are restricted to either attend to tokens (Hermann et al., 2015; Chen et al., 2016) or entire sentences (Weston et al., 2015), with the assumption that certain sub-parts of the document or entire sentences (Weston et al., 2015), with the assumption that certain sub-parts of the document or entire sentences (Weston et al., 2015), with the assumption that certain sub-parts of the document are more important than others. In contrast, we propose a finer-grained model which attends to components of the semantic representation being built up by the GRU. The new attention mechanism, called gated-attention, is implemented via multiplicative interactions between the query and the contextual embeddings, and is applied per hop to act as fine-grained information filters during the multi-step reasoning. The filters weigh individual components of the vector representation of each token in the document separately.

The design of gated-attention layers is motivated by the effectiveness of multiplicative interaction among vector-space representations, e.g., in various types of recurrent units (Hochreiter & Schmidhuber, 1997; Wu et al., 2016) and in relational learning (Yang et al., 2014; Kiros et al., 2014). While other types of compositional operators are possible, such as concatenation or addition (Mitchell & Lapata, 2008), we find that multiplication has strong empirical performance (section 4.3), where query representations naturally serve as information filters across hops.

3.1 Model Details

Several components of the model use a Gated Recurrent Unit (GRU) (Cho et al., 2015) which maps an input sequence \( X = [x_1, x_2, \ldots, x_T] \) to an output sequence \( H = [h_1, h_2, \ldots, h_T] \) as follows:

\[
\begin{align*}
    r_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r), \\
    z_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z), \\
    \tilde{h}_t &= \tanh(W_h x_t + U_h (r_t \odot h_{t-1}) + b_h), \\
    h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t.
\end{align*}
\]

where \( \odot \) denotes the Hadamard product or the element-wise multiplication. \( r_t \) and \( z_t \) are called the reset and update gates respectively, and \( h_t \) the candidate output. A Bi-directional GRU (Bi-GRU) processes the sequence in both forward and backward directions to produce two sequences \([h_1^f, h_2^f, \ldots, h_T^f] \) and \([h_1^b, h_2^b, \ldots, h_T^b] \), which are concatenated at the output

\[
\text{GRU}(X) = [h_1^b \| h_T^b, \ldots, h_T^f \| h_1^f]
\]

where GRU(X) denotes the full output of the Bi-GRU obtained by concatenating each forward state \( h_t^f \) and backward state \( h_t^b \) at step \( t \) given the input \( X \). Note GRU(X) is a matrix in \( \mathbb{R}^{2n_h \times T} \) where \( n_h \) is the number of hidden units in GRU.

Let \( X^{(i)} = [x_1^{(i)}, x_2^{(i)}, \ldots, x_{|D|}^{(i)}] \) denote the token embeddings of the document, which are also inputs at layer \( 1 \) for the document reader below, and \( Y = [y_1, y_2, \ldots, y_{|Q|}] \) denote the token embeddings of the query. Here \( |D| \) and \( |Q| \) denote the document and query lengths respectively.

3.1.1 Multi-Hop Architecture

Fig. 1 illustrates the Gated-Attention (GA) reader. The model reads the document and the query over \( K \) horizontal layers, where layer \( k \) receives the contextual embeddings \( X^{(k-1)} \) of the document from the previous layer. The document embeddings are transformed by taking the full output of a document Bi-GRU (indicated in blue in Fig. 1):

\[
D^{(k)} = \text{GRU}_D^{(k)} (X^{(k-1)})
\]

At the same time, a layer-specific query representation is computed as the full output of a separate query Bi-GRU (indicated in green in Figure 1):

\[
Q^{(k)} = \text{GRU}_Q^{(k)} (Y)
\]

Next, Gated-Attention is applied to \( D^{(k)} \) and \( Q^{(k)} \) to compute inputs for the next layer \( X^{(k)} \).

\[
X^{(k)} = \text{GA}(D^{(k)}, Q^{(k)})
\]

where GA is defined in the following subsection.
3.1.2 Gated-Attention Module

For brevity, let us drop the superscript $k$ in this subsection as we are focusing on a particular layer. For each token $d_i$ in $D$, the GA module forms a token-specific representation of the query $\tilde{q}_i$ using soft attention, and then multiplies the query representation element-wise with the document token representation. Specifically, for $i = 1, \ldots, |D|$:  

$$
\alpha_i = \text{softmax}(Q^T d_i) \quad (5)
$$

$$
\tilde{q}_i = Q\alpha_i 
$$

$$
x_i = d_i \odot \tilde{q}_i \quad (6)
$$

In equation (6) we use the multiplication operator to model the interactions between $d_i$ and $\tilde{q}_i$. In the experiments section, we also report results for other choices of gating functions, including addition $x_i = d_i + \tilde{q}_i$ and concatenation $x_i = d_i \| \tilde{q}_i$.

3.1.3 Answer Prediction

Let $q^{(K)}_\ell = q^T_\ell \| q^b_{\ell-\ell+1}$ be an intermediate output of the final layer query Bi-GRU at the location $\ell$ of the cloze token in the query, and $D^{(K)} = \text{GRU}_D^{(K)}(X^{(K-1)})$ be the full output of final layer document Bi-GRU. To obtain the probability that a particular token in the document answers the query, we take an inner-product between these two, and pass through a softmax layer:

$$
s = \text{softmax}((q^{(K)}_\ell)^T D^{(K)}) \quad (7)
$$

where vector $s$ defines a probability distribution over the $|D|$ tokens in the document. The probability of a particular candidate $c \in C$ as being the answer is then computed by aggregating the probabilities of all document tokens which appear in $c$ and renormalizing over the candidates:

$$
\Pr(c|d, q) \propto \sum_{i \in I(c,d)} s_i \quad (8)
$$

where $I(c, d)$ is the set of positions where a token in $c$ appears in the document $d$. This aggregation operation is the same as the pointer sum attention applied in the AS Reader (Kadlec et al., 2016).

Finally, the candidate with maximum probability is selected as the predicted answer:

$$
a^* = \arg\max_{c \in C} \Pr(c|d, q). \quad (9)
$$

During the training phase, model parameters of GA are updated w.r.t. a cross-entropy loss between the predicted probabilities and the true answers.

3.1.4 Further Enhancements

Character-level Embeddings: Given a token $w$ from the document or query, its vector space representation is computed as $x = L(w)\|C(w)$. $L(w)$ retrieves the word-embedding for $w$ from a lookup table $L \in \mathbb{R}^{|V| \times n_l}$, whose rows hold a vector for each unique token in the vocabulary. We also utilize a character composition model $C(w)$ which generates an orthographic embedding of the token. Such embeddings have been previously shown to be helpful for tasks like Named Entity Recognition (Yang et al., 2016) and dealing with OOV tokens at test time (Dhingra et al., 2016). The embedding $C(w)$ is generated by taking the final outputs $z^f_{nc}$ and $z^b_{nc}$ of a Bi-GRU applied to embeddings from
a lookup table of characters in the token, and applying a linear transformation:

\[ z = z^i_n \parallel a_n, \]

\[ C(w) = Wz + b \]

**Question Evidence Common Word Feature (qe-comm):** Li et al. (2016) recently proposed a simple token level indicator feature which significantly boosts reading comprehension performance in some cases. For each token in the document we construct a one-hot vector \( f_i \in \{0, 1\}^2 \) indicating its presence in the query. It can be incorporated into the GA reader by assigning a feature lookup table \( F \in \mathbb{R}^{n_F \times 2} \) (we use \( n_F = 2 \)), taking the feature embedding \( e_i = f_i^T F \) and appending it to the inputs of the last layer document BiGRU as, \( x_i^{(K)} \parallel f_i \) for all \( i \). We conducted several experiments both with and without this feature and observed some interesting trends, which are discussed below. Henceforth, we refer to this feature as the qe-comm feature or just feature.

4 Experiments and Results

4.1 Datasets

We evaluate the GA reader on five large-scale datasets recently proposed in the literature. The first two, CNN and Daily Mail news stories\(^2\) consist of articles from the popular CNN and Daily Mail websites (Hermann et al., 2015). A query over each article is formed by removing an entity from the short summary which follows the article. Further, entities within each article were anonymized to make the task purely a comprehension one. N-gram statistics, for instance, computed over the entire corpus are no longer useful in such an anonymized corpus.

The next two datasets are formed from two different subsets of the Children’s Book Test (CBT)\(^3\) (Hill et al., 2016). Documents consist of 20 contiguous sentences from the body of a popular children’s book, and queries are formed by deleting a token from the 21st sentence. We only focus on subsets where the deleted token is either a common noun (CN) or named entity (NE) since simple language models already give human-level performance on the other types (cf. (Hill et al., 2016)).

The final dataset is Who Did What\(^4\) (WDW) (Onishi et al., 2016), constructed from the LDC English Gigaword newswire corpus. First, article pairs which appeared around the same time and with overlapping entities are chosen, and then one article forms the document and a cloze query is constructed from the other. Missing tokens are always person named entities. Questions which are easily answered by simple baselines are filtered out, to make the task more challenging. There are two versions of the training set—a small but focused “Strict” version and a large but noisy “Relaxed” version. We report results on both settings which share the same validation and test sets. Statistics of all the datasets used in our experiments are summarized in the Appendix (Table 5).

4.2 Performance Comparison

Tables 1 and 3 show a comparison of the performance of GA Reader with previously published results on WDW and CNN, Daily Mail, CBT datasets respectively. The numbers reported for GA Reader are for single best models, though we compare to both ensembles and single models from prior work. GA Reader-- refers to an earlier version of the model, unpublished but described in a preprint, with the following differences—(1) it does not utilize token-specific attentions within the GA module, as described in equation (5), (2) it does not use a character composition model, (3) it is initialized with word embeddings pretrained on the corpus itself rather than GloVe. A detailed analysis of these differences is studied in the next section. Here we present 4 variants of the latest GA Reader, using combinations of whether the qe-comm feature is used (+feature) or not, and whether the word lookup table \( L(w) \) is updated during training or fixed to its initial value. Other hyperparameters are listed in Appendix A.

Interestingly, we observe that feature engineering leads to significant improvements for WDW and CBT datasets, but not for CNN and Daily Mail datasets. We note that anonymization of the latter datasets means that there is already some feature engineering (it adds hints about whether a token is an entity), and these are much larger than the other four. In machine learning it is common to see the effect of feature engineering diminish with increasing data size. Similarly, fixing the word embeddings provides an improvement for the WDW...
Table 1: Validation/Test accuracy (%) on WDW dataset for both “Strict” and “Relaxed” settings. Results with “†” are cf previously published works.

<table>
<thead>
<tr>
<th>Model</th>
<th>Strict</th>
<th>Relaxed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Val</td>
<td>Test</td>
</tr>
<tr>
<td>Human †</td>
<td>–</td>
<td>84</td>
</tr>
<tr>
<td>Attentive Reader †</td>
<td>–</td>
<td>53</td>
</tr>
<tr>
<td>AS Reader †</td>
<td>–</td>
<td>57</td>
</tr>
<tr>
<td>Stanford AR †</td>
<td>–</td>
<td>64</td>
</tr>
<tr>
<td>NSE †</td>
<td>66.5</td>
<td>66.2</td>
</tr>
<tr>
<td>GA-- †</td>
<td>–</td>
<td>57</td>
</tr>
<tr>
<td>GA (update (L(w)))</td>
<td>67.8</td>
<td>67.0</td>
</tr>
<tr>
<td>GA (fix (L(w)))</td>
<td>68.3</td>
<td>68.0</td>
</tr>
<tr>
<td>GA (+feature, update (L(w)))</td>
<td>70.1</td>
<td>69.5</td>
</tr>
<tr>
<td>GA (+feature, fix (L(w)))</td>
<td><strong>71.6</strong></td>
<td><strong>71.2</strong></td>
</tr>
</tbody>
</table>

Table 2: Top: Performance of different gating functions. Bottom: Effect of varying the number of hops \(K\). Results on WDW without using the qe-comm feature and with fixed \(L(w)\).

<table>
<thead>
<tr>
<th>Gating Function</th>
<th>Accuracy</th>
<th>Val</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum</td>
<td>64.9</td>
<td>64.5</td>
<td></td>
</tr>
<tr>
<td>Concatenate</td>
<td>64.4</td>
<td>63.7</td>
<td></td>
</tr>
<tr>
<td>Multiply</td>
<td><strong>68.3</strong></td>
<td><strong>68.0</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>K</th>
<th>Val</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (AS) †</td>
<td>–</td>
<td>57</td>
</tr>
<tr>
<td>2</td>
<td>65.6</td>
<td>65.6</td>
</tr>
<tr>
<td>3</td>
<td><strong>68.3</strong></td>
<td>68.0</td>
</tr>
<tr>
<td>4</td>
<td><strong>68.3</strong></td>
<td><strong>68.2</strong></td>
</tr>
</tbody>
</table>

and CBT, but not for CNN and Daily Mail. This is not surprising given that the latter datasets are larger and less prone to overfitting.

Comparing with prior work, on the WDW dataset the basic version of the GA Reader outperforms all previously published models when trained on the Strict setting. By adding the qe-comm feature the performance increases by 3.2% and 3.5% on the Strict and Relaxed settings respectively to set a new state of the art on this dataset. On the CNN and Daily Mail datasets the GA Reader leads to an improvement of 3.2% and 4.3% respectively over the best previous single models. They also outperform previous ensemble models, setting a new state of the art for both datasets. For CBT-NE, GA Reader with the qe-comm feature outperforms all previous single and ensemble models except the AS Reader trained on the much larger BookTest Corpus (Baiggar et al., 2016). Lastly, on CBT-CN the GA Reader with the qe-comm feature outperforms all previously published single models except the NSE, and AS Reader trained on a larger corpus. For each of the 4 datasets on which GA achieves the top performance, we conducted one-sample proportion tests to test whether GA is significantly better than the second-best baseline. The p-values are 0.319 for CNN, <0.00001 for DailyMail, 0.028 for CBT-NE, and <0.00001 for WDW. In other words, GA statistically significantly outperforms all other baselines on 3 out of those 4 datasets at the 5% significance level. The results could be even more significant under paired tests, however we did not have access to the predictions from the baselines.

4.3 GA Reader Analysis

In this section we do an ablation study to see the effect of Gated Attention. We compare the GA Reader as described here to a model which is exactly the same in all aspects, except that it passes document embeddings \(D(k)\) in each layer directly to the inputs of the next layer without using the GA module. In other words \(X^{(k)} = D^{(k)}\) for all \(k > 0\). This model ends up using only one query GRU at the output layer for selecting the answer from the document. We compare these two variants both with and without the qe-comm feature on CNN and WDW datasets for three subsets of the training data - 50%, 75% and 100%. Test set accuracies for these settings are shown in Figure 2. On CNN when tested without feature engineering, we observe that GA provides a significant boost in performance compared to without GA. When tested with the feature it still gives an improvement, but the improvement is significant only with 100% training data. On WDW-Strict, which is a third of the size of CNN, without the feature we see an improvement when using GA versus without using GA, which becomes significant as the training set size increases. When tested with the feature it still gives an improvement, but the improvement is significant only with 100% training data. On WDW-Strict, which is a third of the size of CNN, without the feature we see an improvement when using GA versus without using GA, which becomes significant as the training set size increases. When tested with the feature on WDW, for a small data size without GA does better than with GA, but as the dataset size increases they become equivalent. We conclude that GA provides a boost in the absence of feature engineering, or as the training set size increases.

Next we look at the question of how to gate intermediate document reader states from the query, i.e. what operation to use in equation 6.
Table 3: Validation/Test accuracy (%) on CNN, Daily Mail and CBT. Results marked with “†” are cf previously published works. Results marked with “‡” were obtained by training on a larger training set. Best performance on standard training sets is in bold, and on larger training sets in italics.

<table>
<thead>
<tr>
<th>Model</th>
<th>CNN Val</th>
<th>CNN Test</th>
<th>Daily Mail Val</th>
<th>Daily Mail Test</th>
<th>CBT-NE Val</th>
<th>CBT-NE Test</th>
<th>CBT-CN Val</th>
<th>CBT-CN Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Humans (query) †</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>52.0</td>
<td>–</td>
<td>64.4</td>
</tr>
<tr>
<td>Humans (context + query) †</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>81.6</td>
<td>–</td>
<td>81.6</td>
</tr>
<tr>
<td>LSTMs (context + query) †</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>51.2</td>
<td>41.8</td>
<td>–</td>
<td>62.6</td>
</tr>
<tr>
<td>Deep LSTM Reader †</td>
<td>55.0</td>
<td>57.0</td>
<td>63.3</td>
<td>62.2</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Attentive Reader †</td>
<td>61.6</td>
<td>63.0</td>
<td>70.5</td>
<td>69.0</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Impatient Reader †</td>
<td>61.8</td>
<td>63.8</td>
<td>69.0</td>
<td>68.0</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>MemNets †</td>
<td>63.4</td>
<td>66.8</td>
<td>–</td>
<td>–</td>
<td>70.4</td>
<td>66.6</td>
<td>64.2</td>
<td>63.0</td>
</tr>
<tr>
<td>AS Reader †</td>
<td>68.6</td>
<td>69.5</td>
<td>75.0</td>
<td>73.9</td>
<td>73.8</td>
<td>68.6</td>
<td>68.8</td>
<td>63.4</td>
</tr>
<tr>
<td>DER Network †</td>
<td>71.3</td>
<td>72.9</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Stanford AR (relabeling) †</td>
<td>73.8</td>
<td>73.6</td>
<td>77.6</td>
<td>76.6</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Iterative Attentive Reader †</td>
<td>72.6</td>
<td>73.3</td>
<td>–</td>
<td>–</td>
<td>75.2</td>
<td>68.6</td>
<td>72.1</td>
<td>69.2</td>
</tr>
<tr>
<td>EpiReader †</td>
<td>73.4</td>
<td>74.0</td>
<td>–</td>
<td>–</td>
<td>75.3</td>
<td>69.7</td>
<td>71.5</td>
<td>67.4</td>
</tr>
<tr>
<td>AoA Reader †</td>
<td>73.1</td>
<td>74.4</td>
<td>–</td>
<td>–</td>
<td>77.8</td>
<td>72.0</td>
<td>72.2</td>
<td>69.4</td>
</tr>
<tr>
<td>ReasoNet †</td>
<td>72.9</td>
<td>74.7</td>
<td>77.6</td>
<td>76.6</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>NSE †</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>78.2</td>
<td>73.2</td>
<td>74.3</td>
<td>71.9</td>
</tr>
<tr>
<td>BiDAF †</td>
<td>76.3</td>
<td>76.9</td>
<td>80.3</td>
<td>79.6</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>MemNets (ensemble) †</td>
<td>66.2</td>
<td>69.4</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>AS Reader (ensemble) †</td>
<td>73.9</td>
<td>75.4</td>
<td>78.7</td>
<td>77.7</td>
<td>76.2</td>
<td>71.0</td>
<td>71.1</td>
<td>68.9</td>
</tr>
<tr>
<td>Stanford AR (relabeling,ensemble) †</td>
<td>77.2</td>
<td>77.6</td>
<td>80.2</td>
<td>79.2</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Iterative Attentive Reader (ensemble) †</td>
<td>75.2</td>
<td>76.1</td>
<td>–</td>
<td>–</td>
<td>76.9</td>
<td>72.0</td>
<td>74.1</td>
<td>71.0</td>
</tr>
<tr>
<td>EpiReader (ensemble) †</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>76.6</td>
<td>71.8</td>
<td>73.6</td>
<td>70.6</td>
</tr>
<tr>
<td>AS Reader (+BookTest) † ‡</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>80.5</td>
<td>76.2</td>
<td>83.2</td>
<td>80.8</td>
</tr>
<tr>
<td>AS Reader (+BookTest,ensemble) † ‡</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>82.3</td>
<td>78.4</td>
<td>85.7</td>
<td>83.7</td>
</tr>
<tr>
<td>GA--</td>
<td>73.0</td>
<td>73.8</td>
<td>76.7</td>
<td>75.7</td>
<td>74.9</td>
<td>69.0</td>
<td>69.0</td>
<td>63.9</td>
</tr>
<tr>
<td>GA (update $L(w)$)</td>
<td>77.9</td>
<td>77.9</td>
<td>81.5</td>
<td>80.9</td>
<td>76.7</td>
<td>70.1</td>
<td>69.8</td>
<td>67.3</td>
</tr>
<tr>
<td>GA (fix $L(w)$)</td>
<td>77.9</td>
<td>77.8</td>
<td>80.4</td>
<td>79.6</td>
<td>77.2</td>
<td>71.4</td>
<td>71.6</td>
<td>68.0</td>
</tr>
<tr>
<td>GA (+feature, update $L(w)$)</td>
<td>77.3</td>
<td>76.9</td>
<td>80.7</td>
<td>80.0</td>
<td>77.2</td>
<td>73.3</td>
<td>73.0</td>
<td>69.8</td>
</tr>
<tr>
<td>GA (+feature, fix $L(w)$)</td>
<td>76.7</td>
<td>77.4</td>
<td>80.0</td>
<td>79.3</td>
<td><strong>78.5</strong></td>
<td><strong>74.9</strong></td>
<td><strong>74.4</strong></td>
<td><strong>70.7</strong></td>
</tr>
</tbody>
</table>

Table 3 (top) shows the performance on WDW dataset for three common choices – \( \text{sum}(x = d + q) \), \( \text{concatenate}(x = d \| q) \) and \( \text{multiply}(x = d \odot q) \). Empirically we find element-wise multiplication does significantly better than the other two, which justifies our motivation to “filter” out document features which are irrelevant to the query.

At the bottom of Table 2 we show the effect of varying the number of hops \( K \) of the GA Reader on the final performance. We note that for \( K = 1 \), our model is equivalent to the AS Reader without any GA modules. We see a steep and steady rise in accuracy as the number of hops is increased from \( K = 1 \) to \( 3 \), which remains constant beyond that. This is a common trend in machine learning as model complexity is increased, however we note that a multi-hop architecture is important to achieve a high performance for this task, and provide further evidence for this in the next section.

### 4.4 Ablation Study for Model Components

Table 4 shows accuracy on WDW by removing one component at a time. The steepest reduction is observed when we replace pretrained GloVe vectors with those pretrained on the corpus itself. GloVe vectors were trained on a large corpus of about 6 billion tokens (Pennington et al., 2014), and provide an important source of prior knowl-
Figure 2: Performance in accuracy with and without the Gated-Attention module over different training sizes. *p*-values for an exact one-sided McNemar’s test are given inside the parentheses for each setting.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Val</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>GA</td>
<td>68.3</td>
<td>66.9</td>
<td>66.9</td>
</tr>
<tr>
<td>GA--†</td>
<td></td>
<td></td>
<td>57</td>
</tr>
<tr>
<td>−char</td>
<td>65.7</td>
<td>65.0</td>
<td></td>
</tr>
<tr>
<td>−token-attentions (eq. 5)</td>
<td>64.0</td>
<td>62.5</td>
<td></td>
</tr>
<tr>
<td>−glove, +corpus</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Ablation study on WDW dataset, without using the qe-comm feature and with fixed \( L(w) \). Results marked with † are cf Onishi et al. (2016).

Figure 3 shows an example from the validation set of WDW dataset (several more are in the Appendix). In each figure, the left and middle plots visualize attention over the query (equation 5) for candidates in the document after layers 1 & 2 respectively. The right plot shows attention over candidates in the document of cloze placeholder (XXX) in the query at the final layer. The full document, query and correct answer are shown at the bottom.

A generic pattern observed in these examples is that in intermediate layers, candidates in the document (shown along rows) tend to pick out salient tokens in the query which provide clues about the cloze, and in the final layer the candidate with the highest match with these tokens is selected as the answer. In Figure 3 there is a high attention of the correct answer on financial regulatory standards in the first layer, and on us president in the second layer. The incorrect answer, in contrast, only attends to one of these aspects, and hence receives a lower score in the final layer despite the n-gram overlap it has with the cloze token in the query. Importantly, different layers tend to focus on different tokens in the query, supporting the hypothesis that the multi-hop architecture of GA Reader is able to combine distinct pieces of information to answer the query.

5 Conclusion

We presented the Gated-Attention reader for answering cloze-style questions over documents. The GA reader features a novel multiplicative gating mechanism, combined with a multi-hop architecture. Our model achieves the state-of-the-art performance on several large-scale benchmark datasets with more than 4% improvements over competitive baselines. Our model design is backed up by an ablation study showing statistically significant improvements of using Gated Attention as information filters. We also showed empirically that multiplicative gating is superior to addi-
tion and concatenation operations for implementing gated-attentions, though a theoretical justification remains part of future research goals. Analysis of document and query attentions in intermediate layers of the reader further reveals that the model iteratively attends to different aspects of the query to arrive at the final answer. In this paper we have focused on text comprehension, but we believe that the Gated-Attention mechanism may benefit other tasks as well where multiple sources of information interact.

Acknowledgments

We would like to thank Eduard Hovy and Teruko Mitamura for useful comments on this work. This work was funded by NSF under CCF1414030 and Google Research.

References


Ryan Kiros, Richard Zemel, and Ruslan R Salakhutdinov. A multiplicative model for learning distributed text-based attribute representations. In *Advances in


Tsendsuren Munkhdalai and Hong Yu. Neural semantic encoders. EACL, 2017a.


A Implementation Details

Our model was implemented using the Theano (Theano Development Team, 2016) and Lasagne\textsuperscript{5} Python libraries. We used stochastic gradient descent with ADAM updates for optimization, which combines classical momentum and adaptive gradients (Kingma & Ba, 2015). The batch size was 32 and the initial learning rate was $5 \times 10^{-4}$ which was halved every epoch after the second epoch. The same setting is applied to all models and datasets. We also used gradient clipping with a threshold of 10 to stabilize GRU training (Pascanu et al., 2013). We set the number of layers $K$ to be 3 for all experiments. The number of hidden units for the character GRU was set to 50. The remaining two hyperparameters—size of document and query GRUs, and dropout rate—were tuned on the validation set, and their optimal values are shown in Table 6. In general, the optimal GRU size increases and the dropout rate decreases as the corpus size increases.

The word lookup table was initialized with 100d GloVe vectors\textsuperscript{6} (Pennington et al., 2014) and OOV

\textsuperscript{5}https://lasagne.readthedocs.io/en/latest/
\textsuperscript{6}http://nlp.stanford.edu/projects/glove/
tokens at test time were assigned unique random vectors. We empirically observed that initializing with pre-trained embeddings gives higher performance compared to random initialization for all datasets. Furthermore, for smaller datasets (WDW and CBT) we found that fixing these embeddings to their pretrained values led to higher test performance, possibly since it avoids overfitting. We do not use the character composition model for CNN and Daily Mail, since their entities (and hence candidate answers) are anonymized to generic tokens. For other datasets the character lookup table was randomly initialized with $25d$ vectors. All other parameters were initialized to their default values as specified in the Lasagne library.

**B Attention Plots**
Figure 4: Layer-wise attention visualization of GA Reader trained on WDW-Strict. See text for details.

DOC: england pace bowler ajmal shahzad has warned australia that his fearless team-mates are ready to deal their rivals a psychological blow in the forthcoming one-day series. shahzad has been called into england's squad for five matches against australia as cover for the injured ryan sidebottom and he senses an extremely positive mood in the dressing room. the 24-year-old admits england are desperate to warm up for the ashes in australia later this year by giving ricky ponting's men another beating after defeating them in the recent icc world twenty20 final in the caribbean. and shahzad, the first british-born asian to play for yorkshire, is confident andrew Strauss's side won't back down against the ultra-aggressive australia. "it will be a step up for the lads, but everybody is focused and ready for it," shahzad said. "there's no fear, no nerves - and it's nice to be part of that kind of dressing room." sidebottom's hamstring problem has given shahzad his opportunity in the one-day squad and he is determined to seize the opportunity to stake his claim for a permanent place by impressing against the australians. "it's just nice to be called up, and i hope if my chance comes i can grasp it with both hands," he said. "a lot's happened to me in the last six months. if the chance comes - ryan's got a niggle, but i do not think it's too bad. i've just got to put in a decent performance and have my name in the hat for the games to come." but it's just nice to be involved, and know you're there or thereabouts. i just hope i can get the nod."  

QRY: <beg> yorkshire fast bowler ajmal shahzad has been called up to england's one-day international squad after xxx suffered a hamstring injury. <end>  
ANS: ryan sidebottom

DOC: result sunday from the open 23, a (euro) 512,750 (€ 697,400) atp world tour indoor hardcourt event at palais des sports (seeds in parentheses): singles final michael llodra, france, def. julien Benneteau (8), france, 6-3, 6-4. doubles final michael llodra and julien Benneteau, france (2), def. julian knowle, austria and robert lindstedt, sweden (1), 6-4, 6-3.

QRY: <beg> france's michael llodra beat his compatriot and doubles partner xxx. <end>  
ANS: julien Benneteau
Figure 5: Layer-wise attention visualization of GA Reader trained on WDW-Strict. See text for details.
Figure 6: Layer-wise attention visualization of GA Reader trained on WDW-Strict. See text for details.

European Union officials complained on Tuesday about the lack of gas flow from Russia through Ukraine to Europe after Russia resumed early gas supplies under a three-way deal signed on the previous day. European Commission President Jose Manuel Barroso spoke by phone to Russian Prime Minister Vladimir Putin, expressing disappointment over the lack of natural gas flowing to Europe. EU monitors on the ground reported that only very little gas is flowing through the pipeline.

Barroso voiced his disappointment with both the level of gas flowing to Europe and the lack of access of our monitors to dispatch centers. According to his aide, Putin promised him to take a look into what he complained of on the same day. Czech Prime Minister Miloš Topolánek, who spoke on the phone to his Ukrainian counterpart Yulia Tymoshenko about the matter, said a press release from the Czech EU presidency. Tymoshenko informed Topolánek, who asked about the causes and circumstances of the delay in supplies, of some technical difficulties, saying that more specifically the pressure of gas arriving from the Russian is too low. The Czech Prime Minister recommended her to immediately contact the Eurogaz experts who are ready to assist Ukraine with technical problems. Tymoshenko promised to act on this offer.

On Tuesday morning to let gas flow to Europe via Ukraine after cutting off gas supplies to Europe on Wednesday amid a pricing dispute with Ukraine, the cutoff left a number of European countries in lack of heating gas amid freezing weather.

Aide said. <end>

European Commission Chief Jose Manuel Barroso called XXX to express the EU’s disappointment at the weak flow of Russian gas via Ukraine to the bloc.

ANS: Vladimir Putin

The president of the Palestinian Authority Mahmoud Abbas held talks with French President Nicolas Sarkozy on Friday, and slammed the Israeli plan of constructing more settlement buildings on the West Bank as unacceptable. "This is not acceptable," said Abbas after meeting with Sarkozy in the Elysee palace. It was reported that Israeli Prime Minister Benjamin Netanyahu planned to approve the construction of new homes buildings on the West Bank before considering a freeze on settlement activities. "We want a freeze on settlement and the launch of negotiations on the final phase of it," Abbas said. "This was the main subject of our talks," according to a statement from the Elysee Palace, "talks between the two leaders were aimed at starting again the peace process within the Palestinian territories, as well as discussing regional issues." During the meeting, Sarkozy emphasized the urgency of the resumption of a negotiation process between Israel and Palestine. This has been Abbas' third official visit to France since 2007.

AB: Benjamin Netanyahu
Figure 7: Layer-wise attention visualization of GA Reader trained on WDW-Strict. See text for details.

DOC: If there was a shred of doubt where the world cup was built and won for Spain this year, it was removed Monday night when Barcelona destroyed Real Madrid, 5-0. In teeming rain, host Barcelona simply outplayed rivals, which led the Spanish league until Monday. Barcelona's lineup contained eight home-bred players, six of them world champions. Xavi Hernandez and Pedro Rodriguez scored. David Villa scored twice, and Jeffren Suarez scored the fifth as a substitute. The loss ended Madrid's 26-match unbeaten streak. Lionel Messi was once did not score, but Messi struck a post, and he was involved in three of the goals. Indeed, by taking a deeper role and taking considerable brutal tackles, he epitomized Barcelona's collective will to work for one another.

QRY: <beg> Barcelona cruised into the knockout round of the champions league Wednesday by beating Panathinaikos 3-0 after two goals from Pedro Rodriguez and another from XXX. <end>
ANS: Lionel Messi

DOC: Dinara Safina has barely avoided becoming the first No. 1-seeded woman to lose in the first round at the U.S. Open. Safina overcame 11 double-faults and 48 unforced errors to come back and beat 16th-ranked Oliva Rogowska of Australia 6-7 (5), 6-2, 6-4 in Arthur Ashe Stadium on Tuesday. Safina, the younger sister of two-time major champion Marat Safin, moved up to No. 1 in the rankings in April — and is assured of staying there no matter what happens at Flushing Meadows, the Russian reached the finals at the Australian Open and French Open this year, losing both. Rogowska is 18, received a wild-card invitation into the U.S. Open and has won one grand slam match. She has never defeated anyone ranked better than 47th.

QRY: <beg> World number-one Dinara Safina would love to silence her critics and justify her ranking by winning her first grand slam singles title at the U.S. Open in the same way XXX did in 2000. <end>
ANS: Marat Safin