

# Attention-based CNN Matching Net

Tzu-Chien Liu\*, Yu-Hsueh Wu\*, Hung-Yi Lee

\*These authors contributed to the work equally and should be regarded as co-first authors

Graduate Institute of Communication Engineering, National Taiwan University

{b01901153, b01901062, hungyilee}@ntu.edu.tw

## Abstract

In this paper, we introduce attention-based CNN matching net (ACM-Net), an end-to-end neural network for question answering. ACM-Net matches between the given passage, query and multiple answer choices, and then it extracts features from passage and choices based on query information. We also propose a two-staged CNN architecture and a query-based attention mechanism in our model. These two component can effectively find out the most important parts in passage according to the query. Finally, we extract features from those important parts and find out the most possible answer choice. We conduct this model on the MovieQA dataset [1] using Plot Synopses only, and achieve 79.99% accuracy which is the state of the art on the dataset.

## 1. Introduction

Many machine learning models in question answering tasks involve matching mechanism. For example, in factoid question answering such as SQuAD [2], one needs to match between query and corpus in order to find out the most possible fragment as answer. In multiple choice question answering, such as MC Test [3], matching mechanism can also help making the correct decision.

The easiest way of matching is to calculate the cosine similarity between two vectors. It is generally done by two step: First, encode text into word vectors, sentence vectors or paragraph vectors. Second, simply calculate the cosine similarity between target vectors. This method performs well when applied to word-level matching. However, as for matching between sentences or paragraphs, a single vector is not sufficient to encode all the important information. An alternative way to perform sentence-based or paragraph-based matching is using "compare-aggregate" model. Literally, compare-aggregate first compares vectors at word-level, and then aggregates along sequence of words. Wang and Jiang's proposed a general compare-aggregate [4] framework that performs word-level comparison using multiple techniques followed by aggregation with convolutional neural network. In their work, they show that a general compare-aggregate framework can effectively match two sequences through a wide range.

Although "compare-aggregate" matching mechanism performs well on multiple question answering tasks, it tends to aggregate passively through the sequence rather than take the importance of each element into account. That is, "compare-aggregate" model considers all the sequential contents equally. Therefore, we introduce attention technique into our network and re-design the aggregation mechanism in order to deal with aforementioned deficiency.

In this paper, we propose Attention-based CNN Matching Net (ACM-Net). Our model consists of three components: 1)

The similarity mapping layer which convert the input passage, query and choice into feature representation and perform a similarity operation to each other. 2) The attention-based CNN matching network composed of a two-staged CNN focusing on word-level and sentence-level matching respectively. 3) The prediction layer which makes the final decision.

The main contributions of this work are three-fold. First, we introduce a two-staged CNN architecture which integrates information from word-level to sentence-level, and then from sentence-level to passage-level. This architecture avoids the sequence from being too long, and reasonably extracts feature stage by stage. Second, we introduce attention mechanism into this net. We use specially designed CNN structure and attention mechanism to recognize the pattern of similarity map and eventually identify specific syntactic structure of queries. By transforming passage-query feature into attention maps and applying it to passage-choice matching result, we reasonably give weight to every word in the passage. Lastly, our model reaches 79.99% accuracy on the MovieQA dataset which yields top 1 result on this dataset.

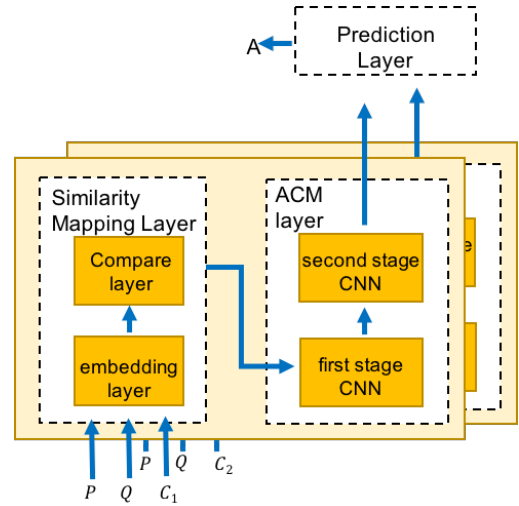


Figure 1: ACM net overview,  $P$  denotes paragraph,  $Q$  denotes query,  $C$  denotes one of choices

## 2. ACM-Net

In this question answering task, a reading passage, a query and several answer choices are given.  $P$  denotes the passage,  $Q$  denotes query and  $C$  denotes one of the multiple choices. The target of the model is to choose a correct answer  $A$  from multi-

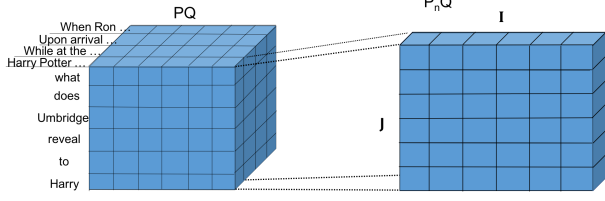


Figure 2: Similarity map between paragraph  $P$  and query  $Q$ .  $I$  denotes the length of each sentence  $P_n$ ,  $J$  denotes the length of query  $Q$

ple choices based on informations of  $P$  and  $Q$ .

Fig.1 is the pipeline overview of ACM Net. First, we use embedding layer to transforms  $P$ ,  $Q$ , and  $C$  into word embedding. Then the compare layer generates passage-query similarity map  $PQ$  and passage-choice similarity map  $PC$ . The following part is the main component of ACM-Net consisting of two-staged CNN architecture. The first stage projects word-level feature into sentence-level, and the second stage projects sentence-level feature into passage-level. Moreover, we apply query-based attention mechanism to each stage on the basis of  $PQ$  feature at word level and sentence level respectively. After ACM layer, we obtain each choice answer feature. Finally, a prediction layer collects output information from every choice feature and returns the most possible predicted answer.

## 2.1. Similarity Mapping Layer

Similarity Mapping Layer is composed of two part: embedding layer and compare layer. Given a passage  $P$  with  $N$  sentences, a query  $Q$ , and a choice  $C$ , the embedding layer transforms every words in  $P$ ,  $Q$  and  $C$  into word embedding<sup>1</sup>:

$$\begin{aligned} P &= \{p_n^i\}_{i=1, n=1}^{I, N} \\ Q &= \{q^j\}_{j=1}^J \\ C &= \{c^k\}_{k=1}^K \end{aligned} \quad (1)$$

$I$  is the length of a sentence in passage<sup>2</sup>, and  $J$  and  $K$  are the length of query and length of one single choice respectively<sup>3</sup>.  $p_n^i$ ,  $q^j$  and  $c^k$  are word embeddings. Word embedding can be obtained by any type of embedding technique, such as recurrent neural network[5], Sequence-to-sequence model[6], Word2vec[7], etc. In our work, we simply use pre-trained GloVe word vectors[8] as the embedding without any further modification or training.

After word embedding step, We want to acquire similarity map which tells us location relationship between passage and query, passage and choices. We use compare layer to compare each passage sentence  $P_n$  to  $Q$  and  $C$  at word level separately as Fig.2 and Fig.3 show.

$$\begin{aligned} P_nQ &= \{\cos(p_n^i, q^j)\}_{i=1, j=1}^{I, J} \\ P_nC &= \{\cos(p_n^i, c^k)\}_{i=1, k=1}^{I, K} \end{aligned} \quad (2)$$

That is, we compare each word in sentences of passage to each word in query and choice. We use cosine similarity as the comparing method here. This step creates

<sup>1</sup>Both query and choice are considered as a sentence.

<sup>2</sup>By padding, all the sentences in all the passages have the same length.

<sup>3</sup>We also make all the queries have the same length  $J$ , and all the choices have the same length  $K$  by padding.

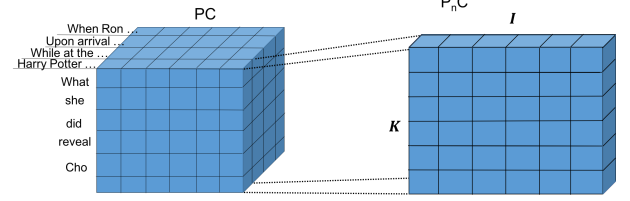


Figure 3: Similarity map between paragraph  $P$  and choice  $C$ .  $I$  denotes the length of each sentence  $P_n$ ,  $K$  denotes the length of choice  $C$

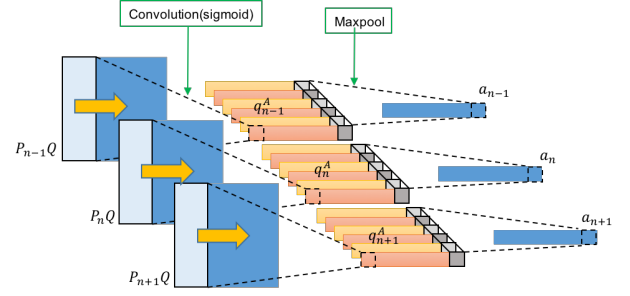


Figure 4: First stage CNN attention part.

two similarity map, passage-query similarity map  $PQ = [P_1Q, P_2Q, \dots, P_NQ] \in \mathbb{R}^{N \times J \times I}$  and passage-choice similarity map  $PC = [P_1C, P_2C, \dots, P_NC] \in \mathbb{R}^{N \times K \times I}$ .

## 2.2. ACM Layer

We propose an attention convolutional matching layer to integrate two similarity maps given above. That is, ACM Layer is used to learn the location relationship pattern. It contains a two-staged CNN combined with query-based attention mechanism. Each stage comprises two major part: attention map and output representation.

### 2.2.1. Attention Map of First Stage

Fig.4 shows the architecture of the attention map in first stage CNN. We choose  $n^{th}$  sentence slice  $P_nQ \in \mathbb{R}^{J \times I}$  in  $PQ$ , and apply CNN to it using the convolution kernel  $W_1^A \in \mathbb{R}^{J \times l \times d}$ , where superscript  $A$  denotes attention map, subscript 1 denotes the first stage CNN. Symbol  $d$  and  $l$  represent width of kernel and number of kernel respectively. The generated feature  $q_n^A \in \mathbb{R}^{l \times (I-d+1)}$  is as follow:

$$q_n^A = \text{sigmoid}(W_1^A * P_nQ + b_1^A) \quad (3)$$

where  $b_1^A \in \mathbb{R}^l$  is the bias. With  $W_1^A$  covering whole query and several words in the passage, convolution kernels would learn the query syntactic structure and give weight to each passage's location. That's why we use sigmoid function as activation function in this stage. Furthermore, we perform maxpooling to  $q_n^A$  perpendicularly in order to find the largest weight between different kernels in the same location, using maxpool kernel shaped  $l$ , and then generate word-level attention map  $a_n \in \mathbb{R}^{I-d+1}$  for each sentence.

### 2.2.2. Output Representation of First Stage

In this stage, we want to acquire passage's sentence feature based on query and choice respectively. We apply CNN to  $P_nC$  to aggregate pattern of location relationship and acquire

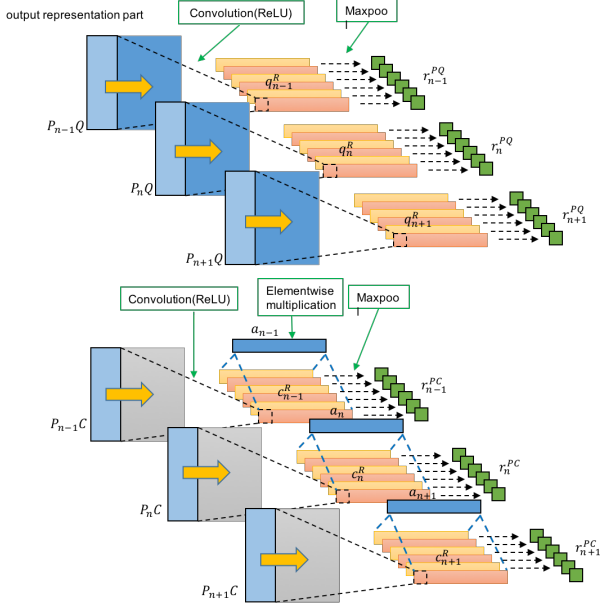


Figure 5: First stage CNN representation part.  $a_n$  is the word-level attention from First stage CNN attention.

choice-based sentence feature. Also, we apply CNN to  $P_nQ$  to acquire query-based sentence feature. CNN architecture of output representation part Fig.5 is similar to which of attention map part, but we use different kernels  $W_1^R \in \mathbb{R}^{l \times K \times d}$  and different bias  $b_1^R \in \mathbb{R}^l$ :

$$\begin{aligned} q_n^R &= \text{ReLU}(W_1^R * P_nQ + b_1^R) \\ c_n^R &= \text{ReLU}(W_1^R * P_nC + b_1^R) \end{aligned} \quad (4)$$

where the superscript  $R$  denotes output representation.<sup>4</sup> We apply  $W_1^R$  on  $P_nC$  and  $P_nQ$  then finally generates  $q_n^R$  and  $c_n^R \in \mathbb{R}^{l \times (I-d+1)}$  using eq.4. We then multiply  $c_n^R$  by the word-level attention map  $a_n$  which comes from stage 2.2.1 element-wise through the first dimension. At last, we maxpool  $q_n^R$  and  $c_n^R$  horizontally with kernel shape  $(I-d+1)$  to get the query-based sentence feature  $r_n^{PQ}$  and choice-based sentence feature  $r_n^{PC} \in \mathbb{R}^l$ .

### 2.2.3. Attention Map of Second Stage

Fig.6 is the architecture of attention map in the second stage CNN. Based on first stage query-based sentence feature from section 2.2.2, we want to acquire sentence-level attention map. The input of this stage is  $r^{PQ} = [r_1^{PQ}, r_2^{PQ}, \dots, r_N^{PQ}]$ , which will be further refined by CNN with kernel  $W_2^A \in \mathbb{R}^{l \times d \times l}$  and generates intermediate feature  $\hat{q}^A \in \mathbb{R}^{l \times (N-d+1)}$ .

$$\hat{q}^A = \text{sigmoid}(W_2^A * r^{PQ} + b_2^A) \quad (5)$$

Then, same as attention map of first stage, we maxpool  $\hat{q}^A$  with kernel shaped  $l$ , and obtain sentence-level attention map  $\hat{a} \in \mathbb{R}^{N-d+1}$ .

### 2.2.4. Output Representation of Second Stage

Output representation part of the second stage in Fig.7 has two input, sentence-level attention map  $\hat{a}$  and sentence-level feature

<sup>4</sup>different choices share same  $W_1^R$  and  $b_1^R$ .

$r^{PC} = [r_1^{PC}, r_2^{PC}, \dots, r_N^{PC}]$ . The equations here are similar to those previously mentioned. As follow:

$$\begin{aligned} \hat{c}^R &= \text{ReLU}(W_2^R * r^{PC} + b_2^R) \\ \hat{r} &= \{\max(\hat{c}_t^R \cdot \hat{a})\}_{t=1}^l \end{aligned} \quad (6)$$

where  $W_2^R \in \mathbb{R}^{l \times l \times d}$ ,  $b_2^R \in \mathbb{R}^l$ , and  $\hat{c}^R \in \mathbb{R}^{l \times (N-d+1)}$ . Output representation of certain choice  $\hat{r} \in \mathbb{R}^l$  is the final output of ACM layer.

## 2.3. Prediction Layer

Prediction Layer is the final part of ACM-Net. We use  $\hat{r}_m \in \mathbb{R}^l$  to represent the final output representation of the  $m^{th}$  choice. In order to find out the most correct choice, we simply pass  $\hat{r}_m$  to two fully-connected layer and compute probability for each choice using softmax as follows:

$$R = \{\hat{r}_m\}_{m=1}^M \quad (7)$$

$$p(m|R) = \text{softmax}(W^o(\tanh(W^p R + b^p)) + b^o) \quad (8)$$

where  $W^p \in \mathbb{R}^{l \times l}$ ,  $b^p \in \mathbb{R}^l$ ,  $W^o \in \mathbb{R}^l$ , and  $b^o \in \mathbb{R}$ .

## 3. Experiments

### 3.1. Implementation details

In the preprocessing step, we used pre-trained GloVe vectors for word embeddings, and they would not be updated during training; We padded sentence number in each passage to 101, all word number in each sentence to 100. Word number of queries and choices are padded to 50. For all kernels of CNN,  $W_1^A, W_2^A, W_1^R, W_2^R$ , each of which has three different kernel width  $d = \{1, 3, 5\}$ ; each of them has same kernel number  $l = 128$ . We applied dropout in each CNN layer with dropout rate 0.8. We used Adam [9] optimizer to optimize our model with initial learning rate 0.001.

### 3.2. Experimental Result

#### 3.2.1. MovieQA Result

We mainly focus on the MovieQA dataset to train and evaluate our model. MovieQA dataset aims to evaluate automatic story comprehension from both video and text. The data set consists of almost 15,000 multiple choice question answers. Diverse information in this dataset like plots, scripts, sub-title and video captions can be used to infer answers. In our task, only plot informations are used. This challenging dataset is suitable to evaluate ACM-Net because movie plots are longer than normal reading comprehension task. Each question comes with a set of five highly plausible choices, only one of which is correct; In the MovieQA benchmark, there are 1958 QA pairs in the val set and 3138 QA pairs in the test set.

We used ensemble model in this dataset. The ensemble model consists of eight training runs models with identical structure and hyper-parameter. In the val set, we achieve 77.6% accuracy with single model and 79.0% accuracy with ensemble model. In the test set, as the Table 1. shows, our model achieves 79.99 % accuracy with ensemble model and is the state of the art.

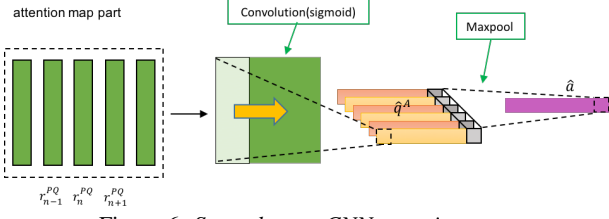


Figure 6: Second stage CNN attention part.

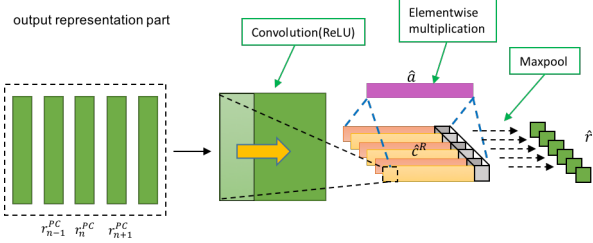


Figure 7: Second stage CNN representation part.  $\hat{a}$  is the sentence-level attention from Second stage CNN attention.

### 3.2.2. MCTest Result

We also applied our model to MCTest dataset which requires machines to answer multiple-choice reading comprehension questions about fictional stories. The original paper describe that a baseline method using a combination of a sliding window score and a distance based . They achieve 66.7% and 56.7% on MC500 and MC160 separately. Because of the restricted training set and development set, we train our model on MovieQA training set and apply the result to test MCTest dataset. On MCTest dataset, we still outperform baseline and achieve 68.1% accuracy on MC160 and 61.5% accuracy on MC500.

### 3.3. Discussion

ACM-Net is a powerful network focusing on multiple choice QA task. It matches between passage and choices based on query information. One of the most important idea in ACM-Net is two-staged attention map. The first attention map is at word level, representing the importance of each word in paragraph to a certain question; the second attention map, however is at sentence level, representing the importance of each sentence in paragraph to a certain question.

In this section, we designed several experiments to test how two-stage mechanism and attention maps impact on our model.

#### 3.3.1. Two-stage Effect Experiment

In this experiment, we focused on the difference between one-stage ACM-Net and two-stage ACM-Net. For one-stage ACM-

Table 1: MovieQA result [1], [4]

Models	dev set	test set
Cosine Word2Vec	46.4	45.63
Cosine TFIDF	47.6	47.36
SSCB TFIDF	48.5	-
Compare Aggregate	72.1	72.9
ACM-Net	77.6	75.84
Convnet Fusion	-	77.63
<b>ACM-Net(ensemble)</b>	<b>79.0</b>	<b>79.99</b>

Net, we didn't split an entire passage into sentences. That is, the shape of passage-query similarity map  $PQ$  and passage-choice similarity map  $PC$  are 2D rather than 3D. We convolved them directly on word-level and output passage feature without second-stage involved. The result is shown on table 2. The result shows that the modified one staged ACM-Net reaches 66.8% accuracy on validation set, which is ten percent lower than 78.1%, the original ACM-Net accuracy on validation set.

#### 3.3.2. Attention Effect Experiment

In this experiment, our target is to validate the effect of query-based attention in ACM-Net. We modified three different structures from original ACM layer below:

1) For the first one, we modified ACM layer in section 2.2 and removed both sentence-level attention map and word-level attention map part from it. However, this modified model would have a deficiency of query information. Therefore, we concatenated the final output representation of  $PQ$  and  $PC$  together before prediction layer. The experiment result is shown on the Table 2. The result is almost ten percent less than the original one. 2) For the second one, we only removed sentence-level attention in section 2.2.3 from ACM layer and kept word-level attention in the model. 3) For the last one, instead of removing sentence-level attention, we removed word-level attention from ACM layer.

The result is shown in Table 2. We can see that ACM-Net(with only word-level attention) performs better than ACM-Net(without attention); ACM-Net(with only sentence-level attention) performs better than ACM-Net(with only word-level attention); And original ACM-Net which contains both word-level and sentence-level attention does the best job among all. Thus, not only word-level attention but also sentence-level attention can contribute to the performance of ACM-Net. However, sentence-level attention seems to play a more important role.

Table 2: Experiment result

Models	dev set	test set
One stage ACM-Net	66.8	-
ACM-Net(no attention)	69.6	-
ACM-Net(only word-level attention)	72.5	-
ACM-Net(only sentence-level attention)	75.1	-
ACM-Net(single)	77.6	75.84
<b>ACM-Net(ensemble)</b>	<b>79.0</b>	<b>79.99</b>

#### 3.3.3. Attention Effect Discussion

Figure.8 is the visualization of two attention maps and their corresponding question. The upper half of Figure.8 is the attention map at sentence level. We picked the sentence with the largest attention value as target sentence and examine it. Thus, we could get the lower half of Figure.8, which shows the attention map at word level on the target sentence. We used a question from movie, Harry Potter, as an example. The result shows that the sentence with the largest attention value is exactly where the correct answer comes from. It turns out that sentence-level attention map can successfully find out which sentence contains the information of correct answer. As for word-level attention map, we can easily see that the attention map focus mainly on the end of target sentence, which is obviously more important for answering this question.

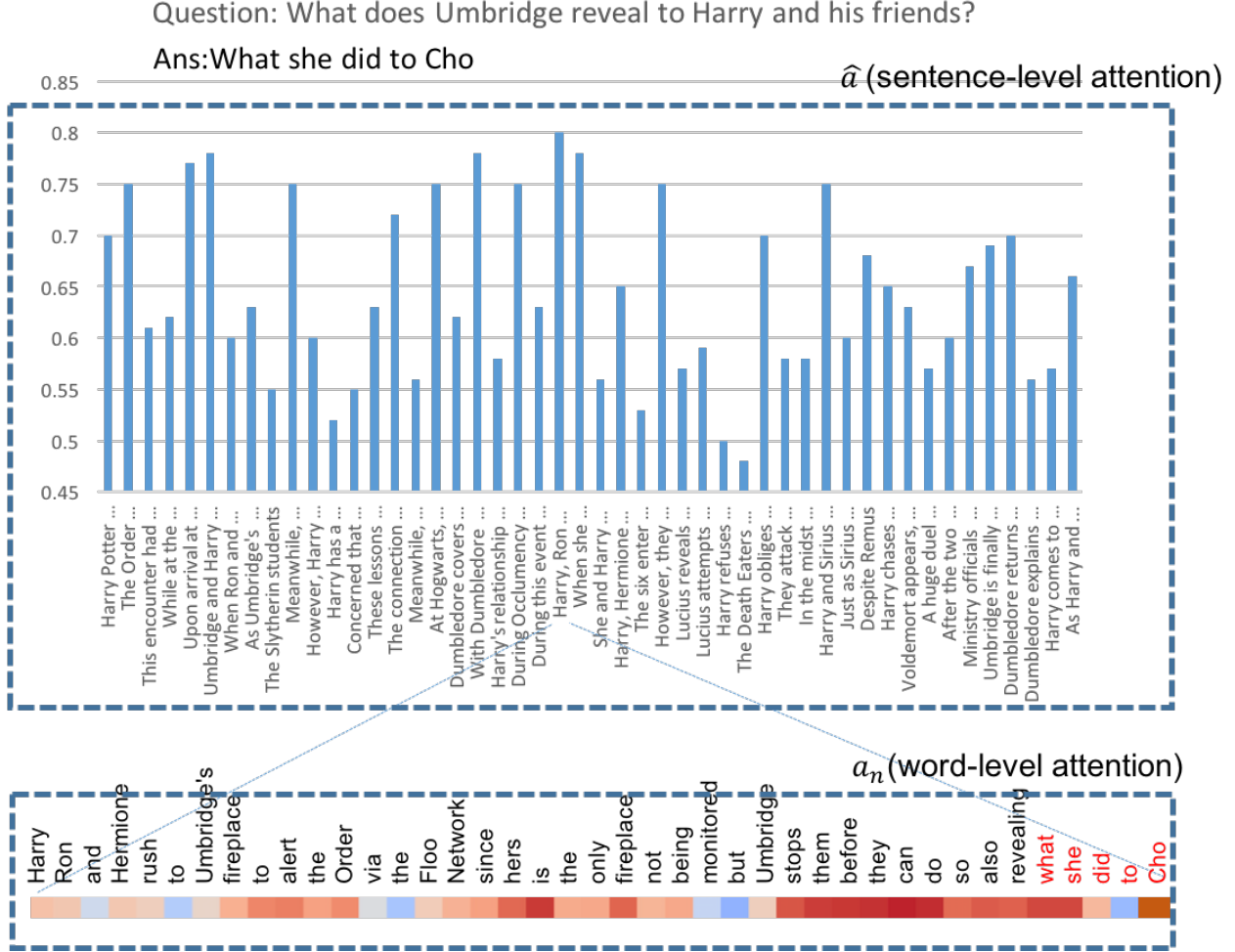


Figure 8: Visualization of the attention map.  $\hat{a}$  is the sentence-level attention from Second stage CNN attention.  $a_n$  is the word-level attention from First stage CNN attention.

#### 4. Conclusion

In this paper, we present an efficient matching mechanism on multiple choice question answering task. We introduce two-staged CNN to match passage and choice on word level and sentence level. In addition, we use query-based CNN attention to enhance matching effect.

The power of the model is verified on MovieQA dataset, which yielded the state of the art result on the dataset. In the future, we are now working on training our model based on our own trained embedding with TF-IDF [10] weighting. Furthermore, we would like to test our model on open-answer task like SQuAD by seeing the whole corpus as an “answer pool” and solve it like multiple choice question.

#### 5. References

- [1] M. Tapaswi, Y. Zhu, R. Stiefelhaagen, A. Torralba, R. Urtasun, and S. Fidler, “Movieqa: Understanding stories in movies through question-answering,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [2] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “Squad: 100,000+ questions for machine comprehension of text,” *arXiv preprint arXiv:1606.05250*, 2016.
- [3] M. Richardson, C. J. Burges, and E. Renshaw, “Mctest: A challenge dataset for the open-domain machine comprehension of text,” in *EMNLP*, vol. 3, 2013, p. 4.
- [4] S. Wang and J. Jiang, “A compare-aggregate model for matching text sequences,” *arXiv preprint arXiv:1611.01747*, 2016.
- [5] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, “Recurrent neural network based language model,” in *Inter-speech*, vol. 2, 2010, p. 3.
- [6] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *Advances in neural information processing systems*, 2014, pp. 3104–3112.
- [7] Y. Goldberg and O. Levy, “word2vec explained: deriving mikolov et al.’s negative-sampling word-embedding method,” *arXiv preprint arXiv:1402.3722*, 2014.
- [8] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference*

*on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

- [9] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [10] J. Ramos *et al.*, “Using tf-idf to determine word relevance in document queries,” in *Proceedings of the first instructional conference on machine learning*, vol. 242, 2003, pp. 133–142.