All Information is Valuable: Question Matching over Full Information Transmission Network

Le Qi[1], Yu Zhang[1]*, Qingyu Yin[1], Guidong Zheng[2], Junjie Wen[2], Jinlong Li[2], and Ting Liu[1]

[1]Research Center for Social Computing and Information Retrieval,
Harbin Institute of Technology, Harbin, China
[2]AI Lab of China Merchants Bank
{lqi, zhangyu, qyyin, tliu}@ir.hit.edu.cn
{zhengguidong, wenjunjieee, lucida}@cmbchina.com

Abstract

Question matching is the task of identifying whether two questions have the same intent. For better reasoning the relationship between questions, existing studies adopt multiple interaction modules and perform multi-round reasoning via deep neural networks. In this process, there are two kinds of critical information that are commonly employed: the representation information of original questions and the interactive information between pairs of questions. However, previous studies tend to transmit only one kind of information, while failing to utilize both kinds of information simultaneously. To address this problem, in this paper, we propose a Full Information Transmission Network (FITN) that can transmit both representation and interactive information together in a simultaneous fashion. More specifically, we employ a novel memory-based attention for keeping and transmitting the interactive information through a global interaction matrix. Besides, we apply an original-average mixed connection method to effectively transmit the representation information between different reasoning rounds, which helps to preserve the original representation features of questions along with the historical hidden features. Experiments on two standard benchmarks demonstrate that our approach outperforms strong baseline models.

1 Introduction

Question Matching (QM) aims to identify whether two questions have the same intent, which is widely applied in Question Answering (QA) applications such as community QA and intelligent customer services. Typically, QM is regarded as a semantic matching task (Hu et al., 2021). To correctly infer the relationship of a given question pair, there are two kinds of information that should be considered: the representation information of questions that captures the semantics of the texts, and the interactive information between questions that contains critical hints for relationship reasoning.

For better detecting the relationship between question pairs, it's far from being enough to conduct only one single round of reasoning. Existing methods commonly resort to multiple interaction modules to do deep reasoning, where each module is generally composed of an encoding layer (can be omitted (Gong et al., 2018)) to update the representation information of questions and an interaction layer for capturing the interactive information between questions (Kim et al., 2019; Hu et al., 2021). In such a multi-round reasoning procedure, both the representation and interactive information in history rounds play a vital role in guiding the future inference. However, previous studies either only transmit the representation information (Kim et al., 2019) or only the interactive information (Gong et al., 2018), while failing to utilize both kinds of information simultaneously.

As shown in Figure 1 (i), when performing multi-round reasoning, if a model only transmits the representation information, the interactive information between questions will then be simply utilized to generate the representation of questions for future rounds. Consequently, the critical hints for relationship reasoning conveyed by interactive information are abandoned and cannot be directly used for future inferences. On the other side, if a model only transmits the interactive information, it is equivalent to conduct multi-round reasoning with only one single pass on question pairs, as shown in Figure 1 (ii). Admittedly, missing the representation information of original questions may lead to understanding deviation and thus bring cascading errors. Therefore, as shown in Figure 1 (iii), to better perform reasoning between question pairs, a desirable solution should be able to transmit both the representation and interactive information from historical rounds to the current round simultaneously.

To address the aforementioned problems, in this

* Corresponding author.

Findings of the Association for Computational Linguistics: NAACL 2022, pages 1431 - 1440

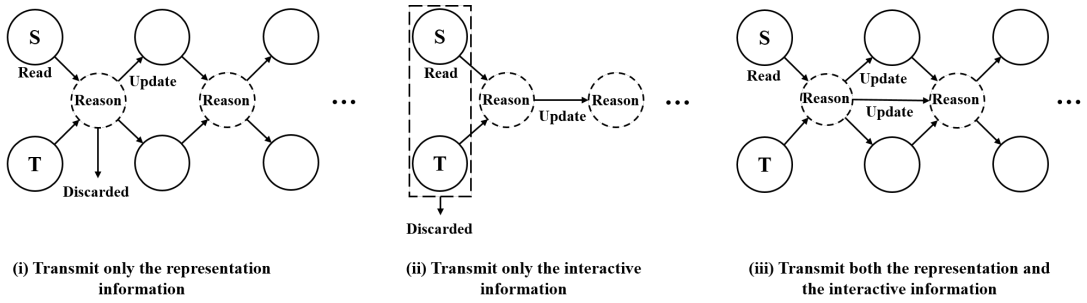


Figure 1: Comparison between different reasoning process of the relationship between questions S and T when transmitting different kinds of information.

paper, we propose a **F**ull **I**nformation **T**ransmission **N**etwork (FITN) that learns to transmit both the representation and interactive information between each round of reasoning. In particular, we propose a novel **Mem**ory-based **Att**ention (Mem-Att) to transmit the interactive information between question pairs. In the Mem-Att, we maintain a global interaction matrix as a memory for keeping the interactive information and do inference on top of it. Compared with traditional attentions that calculate the alignment score directly, the proposed interaction matrix keeps rich interactive information and is more stable in the update process due to its redundancy. Thanks to the global interaction matrix, each round of inference could benefit from the historical interactive information and the whole reasoning procedure is progressive.

Meanwhile, to effectively transmit the representation information of questions, we introduce an interesting connection method, namely the **O**riginal-**A**verage **Mixed Connection** (OA-mixed Connection). Instead of feeding only the hidden features from the last reasoning round, when performing reasoning at the current round, we regard both the hidden features and the original representation embeddings of questions as the input. Such a connection method enables our model an ability to explicitly utilize the entire rich information of original texts when inference. In addition, the OA-mixed Connection employs the average operation over hidden features from the last two rounds to build the input hidden feature for the current reasoning round. Compared with the residual connection (He et al., 2016) that treats the information in each round equally, the average connection pays more attention to the information in the nearer rounds, and thus brings better discrimination ability.

We evaluate our proposed method on the Quora and LCQMC benchmarks. Experimental results show that FITN outperforms the non-pretrained baselines with considerable margins. Furthermore, compared with pre-trained models (small ones with comparable parameter sizes as FITN), our FITN also achieves better performance, which reveals the advantage of proposed method under resource-constrained conditions. All these illustrates the effectiveness of our method.

In sum, our major contributions are three-fold:

- We propose the Full Information Transmission Network (FITN) that can better utilize the historical information, capturing both the representation and interactive information of questions for question matching.
- We propose the memory-based attention for keeping and transmitting the interactive information and the original-average mixed connection to fully utilize the original embedding features of texts and historical hidden features.
- We evaluate the proposed FITN on two benchmark datasets, where considerable improvements are gained over strong baseline models.

2 Methodology

In this section, we introduce our proposed full information transmission network (FITN) in detail. As shown in Figure 2, FITN comprises three modules: the embedding module, the interaction module and the prediction module. In FITN, we first embed each question in the embedding module, then do inference in the interaction module and finally predict their relationship in the prediction module.

We denote two input questions as $S = \{s_1, s_2, ..., s_m\}$ and $T = \{t_1, t_2, ..., t_n\}$ where s_i/t_j is the i^{th}/j^{th} token of question S/T and m/n is the token length of S/T.

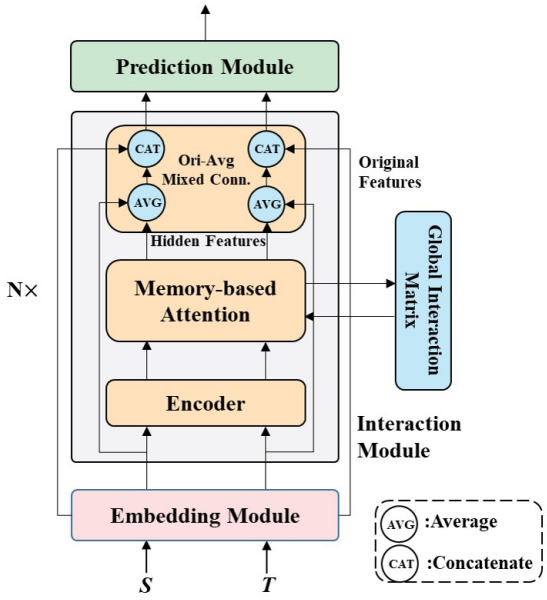


Figure 2: Architecture of the FITN model.

2.1 Embedding Module

In the embedding module, we apply the word embedding along with the character embedding to embed tokens in each question. The character embedding is randomly initialized and then processed by a convolutional neural network (CNN) with a max-pooling operation. Formally, the final representation e_{s_i} of token s_i is calculated as follows:

$$e_{s_i} = [\mathrm{Emb}(s_i); \mathrm{ChConv}(s_i)] \tag{1}$$

where $[;]$ denotes the concatenation operation, Emb is the word embedding and ChConv is the character-level CNN. Each word in S and T is treated in the same procedure and then S and T can be represented as $E_S \in \mathbb{R}^{m \times d_e}$ and $E_T \in \mathbb{R}^{n \times d_e}$.

2.2 Interaction Module

The interaction module is the core of our FITN, composed of N same-structured blocks for doing N rounds of inference. Each block contains 3 components: the encoding layer, the memory-based attention layer and the original-average mixed connection layer. We denote I_S^l and I_T^l as the inputs of the l^{th} block, where $I_S^0 = E_S$ and $I_T^0 = E_T$.

2.2.1 Encoding Layer

We encode two questions through a Bi-LSTM encoder to extract the contextual representation of each token in questions, shown as:

$$H_S^l = \mathrm{BiLSTM}^{\mathrm{l}}(I_S^l) \tag{2}$$

$$H_T^l = \mathrm{BiLSTM}^{\mathrm{l}}(I_T^l) \tag{3}$$

where $H_S^l \in \mathbb{R}^{m \times d_h}$ and $H_T^l \in \mathbb{R}^{m \times d_h}$ are the hidden representations of I_S^l and I_T^l in the l^{th} round, respectively.

2.2.2 Memory-based Attention Layer

As shown in Figure 3, we maintain a global interaction matrix for keeping and transmitting the interactive information in the memory-based attention (Mem-Att) layer. The global interaction matrix is treated as a memory, which keeps all the historical interactive information and will be updated when getting the new one. For each pair of tokens, we keep an interactive vector instead of an attention score in the global interaction matrix. The interactive vector keeps richer information and is more stable in the update process due to its redundancy.

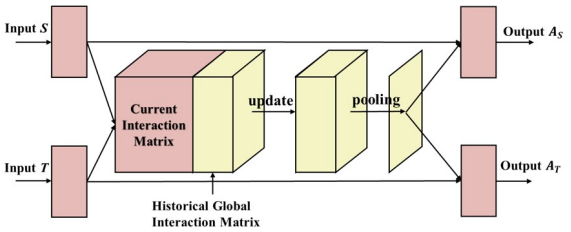


Figure 3: Architecture of the Mem-Att.

In each round, we firstly update the global interaction matrix and then do attention based on this matrix. In this way, the interactive information in history can be transmitted into the current round and provides assistance on the soft-alignment and inference between the two questions.

Global Interaction Matrix Update The global interaction matrix is updated through two steps: current interaction matrix calculation and global interaction matrix combination.

Current Interaction Matrix Calculation The current interaction matrix in the l^{th} round $M^l \in \mathbb{R}^{n \times m \times d_h}$ is calculated as follows:

$$M^l = H_S^l \odot H_T^l \tag{4}$$

For each pair of tokens s_i and t_j in the question S and T, the interaction vector $M_{i,j}^l \in \mathbb{R}^{d_h}$ in M^l is calculated through the element-wise multiplication operation, shown as:

$$M_{i,j}^l = H_{s_i}^l \circ H_{t_j}^l \tag{5}$$

where $\circ$ is the element-wise multiplication.

Global Interaction Matrix Combination After that, we combine the current interaction matrix M^l and the global interaction matrix $\bar{M}^{l-1}$ in the previous round and feed the concatenation result of them into a fully-connected layer with a non-linear activation function as the global interaction matrix $\bar{M}^l \in \mathbb{R}^{n \times m \times d_m}$ in the l^{th} round:

$$\bar{M}^l = \begin{cases} F(M^l) & l = 0 \\ F([M^l; \bar{M}^{l-1}]) & l > 0 \end{cases} \quad (6)$$

where $\bar{M}^l_{i,j} \in \mathbb{R}^{d_m}$ in $\bar{M}$ is calculated as:

$$\bar{M}^l_{i,j} = f(w^l_m[\bar{M}^{l-1}_{i,j}; M^l_{i,j}] + b^l_m) \quad (7)$$

where $[;]$ is vector concatenation across row, $w^l_m \in \mathbb{R}^{(d_h+d_m)\times d_m}$ and $b^l_m \in \mathbb{R}^{d_m}$ correspond to the weight and bias respectively.

Attention over Interaction Matrix Next, we do inference and alignment through the global interaction matrix. We firstly adopt a dense-pooling method to extract an attention map from the global interaction matrix. More specifically, we utilize a fully-connected layer with a nonlinear function to convert each vector into the attention value. Each element $Att^l_{i,j}$ in the attention map $Att^l \in \mathbb{R}^{m \times n}$ is calculated as:

$$Att^l_{i,j} = f(w^l_p \bar{M}^l_{i,j} + b^l_p) \quad (8)$$

where $w^l_p \in \mathbb{R}^{d_m \times 1}$ and $b^l_p \in \mathbb{R}$ correspond to the weight and bias, respectively. Then, the attentive representation $A^l_{s_i}$ of s_i in the l^{th} round is weighted summed by $H^l_{t_j}$, where the weights are calculated by the softmax operation over $Att^l_{i,j}$:

$$A^l_{s_i} = \sum_{j=1}^{n} \mathrm{softmax}(Att^l_{i,j}) H^l_{t_j} \quad (9)$$

Finally, we calculate the average and the difference between the attentive representation $A_{S/T}$ and the contextual representation $H_{S/T}$, concatenate the results with themselves together, and then feed the concatenation result into a fully-connected layer to get the outputs of the block.

$$U^l_S = [H^l_S; A^l_S; (H^l_S + A^l_S)/2; H^l_S - A^l_S] \quad (10)$$

$$U^l_T = [H^l_T; A^l_T; (H^l_T + A^l_T)/2; H^l_T - A^l_T] \quad (11)$$

$$O^l_S = f(w^l_f U^l_S + b^l_f) \quad (12)$$

$$O^l_T = f(w^l_f U^l_T + b^l_f) \quad (13)$$

where $O^l_S \in \mathbb{R}^{m \times d_h}$, $O^l_T \in \mathbb{R}^{n \times d_h}$, $w^l_f \in \mathbb{R}^{4d_h \times d_h}$, and $b^l_f \in \mathbb{R}^{d_h}$ are the weight and bias respectively.

2.2.3 Original-Average Mixed Connection Layer

Finally, we transmit the representation information through the original-average mixed connectivity pattern (OA-mixed connection) in this layer. The question representation input to each round of inference can be divided into two parts: the original features from the initial embedding of questions and the hidden features extracted from previous inference rounds. Both of them play a vital role in each round of inference, where the original features can lead the model to make inference in the right direction, and the hidden features contain deeper contextual and interactive information. Besides, the hidden features can be seen as the information enhancement of the original features. Formally, the whole process can be shown as:

$$I^l = \begin{cases} I_E & l = 0 \\ [I_E; I^l_H] & l > 0 \end{cases} \quad (14)$$

where $I^l \in \mathbb{R}^{m \times (d_e + d_h)}$ $(l > 0)$ is the l-th round input, I_E is the initial embedding, and I^l_H is the l-th round hidden input, calculated as:

$$I^l_H = \begin{cases} O^0 & l = 1 \\ (O^{l-1} + I^{l-1}_H)/2 & l > 1 \\ = \frac{O^{l-1}}{2} + \cdots + \frac{O^1}{2^{l-1}} + \frac{O^0}{2^{l-1}} & \end{cases} \quad (15)$$

where O^l are the hidden outputs of the interaction module before the average connection.

Here, instead of the residual connection, we apply the average connection to capture the hidden features. Compared with the residual connection that treats the information in each round equally, the average connection pay more attention to the information in the nearer rounds. Besides, the residual connection's summation operation may cause the variance of the vectors in the hidden part to go larger as the layers deepen. In comparison, the average connection can balance the variance between the two parts of the question representation.

2.3 Prediction Module

The final representations of the two questions in the interaction module are the last block's next inputs I^{N+1}_S and I^{N+1}_T. To extract a proper representation for each question, we apply the max-pooling operation over them, i.e.:

$$V_S = max(I^{N+1}_S) \quad (16)$$

$$V_T = max(I^{N+1}_T) \quad (17)$$

Table 1: Experimental results on the Quora and LCQMC datasets. Para. denotes the number of parameters. The evaluation metric of Quora is accuracy (%), and that of LCQMC is accuracy (%) and F1. The results are average scores using 5 different seeds along with the standard deviation.

Type	Model	Para.	Quora	LCQMC
Non-pretrained	BiMPM(Wang et al., 2017)	1.6m	88.2	83.3/84.9
	DIIN (Gong et al., 2018)	4.4m	89.1	-/-
	CSRAN (Tay et al., 2018)	-	89.2	-/-
	RE2 (Yang et al., 2019)	2.8m	89.4	-/ -
	Enhanced-RCNN (Peng et al., 2020)	7.3m	89.5	-/-
	TIM-W (Zhou et al., 2020)	-	89.6	-/-
	DRCN (Kim et al., 2019)	6.7m	90.2	-/-
	GMN (Chen et al., 2020)	-	-	84.6/86.0
	LET (Lyu et al., 2021)	-	-	84.8/86.1
	COIN (Hu et al., 2021)	6.5m	89.4	85.6/86.5
Pre-trained	AlBERT-tiny (Lan et al., 2020)	4.1m	-	85.3/86.3
	BERT-tiny (Turc et al., 2019)	4.4m	87.2	-/-
	BERT-mini (Turc et al., 2019)	11.3m	88.8	-/-
	AlBERT-base (Lan et al., 2020)	11.7m	90.0	86.3/87.0
Ours	**FITN**	**2.5m**	**90.6±0.1**	**86.0±0.5/87.1±0.4**

where $V_S, V_T \in \mathbb{R}^{d_h+d_e}$ and max extracts the maximum value in each column of the inputs.

Finally, we concatenate V_S and V_T to get the feature vector V and feed the feature vector V into a two-layer feed-forward network with one hidden layer and one softmax layer to make the prediction.

3 Experiments

We evaluate our FITN on two QM benchmarks: the Quora (English dataset) (Iyer et al., 2017) and LCQMC (Chinese dataset) (Liu et al., 2018). The Quora dataset contains over 400k question pairs collecting from Quora, an English community question answering (cQA) website, and the data splits (380K/10K/10K) are provided in BIMPM (Wang et al., 2017). The LCQMC collects over 260k question pairs from a Chinese cQA website called BaiduKnows (240K/8K/12K).

3.1 Implementation Details

In the original FITN, we initialize the word embedding with 300d Fasttext vectors (Bojanowski et al., 2017) for the English task and 300d Word2Vec vectors trained in Baidu Encyclopedia (Qiu et al., 2018) for the Chinese task, respectively. We randomly initialize the character embedding with a 25d vector and extract a 50d character representation by CNN. Then, we conduct three rounds of inference and set the hidden size of each layer to 100d in the interaction module. Finally, we set 500 hidden units for the 2-layer FFN in the prediction module. We apply an Adam (Kingma and Ba, 2015) optimizer with a learning rate of 1e-3. We train 100 epochs on the Quora dataset and 20 epochs on the LCQMC dataset. We run 5 times with 5 different randomly selected seeds and report the mean value with the standard deviation selected according to the best performance in the development set.

3.2 Experimental Results

The main experimental results are shown in Table 1. We compare our FITN with non-pretrained models at first. In particular, we employ the baselines including: 1) DIIN(Gong et al., 2018): a CNN-based model that employs a DenseNet on the interaction matrix; 2) RE2 (Yang et al., 2019): a CNN-based model with the augmented residual connection; 3) Enhanced RCNN (Peng et al., 2020): a model that encodes sentences by multi-layer CNNs and adopts the attention-based RNNs for relationship inference; 4) TIM-W (Zhou et al., 2020): a model based on deep mutual information estimation; 5) DRCN (Kim et al., 2019): a co-attention BiLSTM model with dense-connection; 6) GMN (Zhou et al., 2020): a neural graph matching model; 7) LET (Lyu et al., 2021): a transformer-based model that employs the external linguistic knowledge derived from Graph Attention Networks; and 8) COIN (Hu et al., 2021): A CNN-based model

with a deep context-aware cross-attention based interaction module.

We can see that our model outperforms all these baselines on the two benchmarks. More specifically, our model beats DIIN because we can keep updating the representation information based on the historical representation information during iterations. Compared with DRCN, our FITN utilizes the historical interactive information for inference and in return acquires performance improvements with fewer inference rounds. Besides, the historical interactive information can also benefit our model on deeper inference. Therefore, the performance of our model is unsurprisingly better than RE2.

In addition, to further verify the effectiveness of our FITN under restricted computing resources, we compare our FITN with 4 publicly available tiny pre-trained models, which are distilled from large pre-trained models (BERT-tiny and BERT-mini (Turc et al., 2019) that are distilled from BERT-base (Devlin et al., 2019)) or directly pre-trained by large-scale datasets (AlBERT-tiny and AlBERT-base (Lan et al., 2020)). As shown in Table 1, our model can achieve competitive or even better performance than pre-trained models with similar model size. It demonstrates that our FITN can be a desirable choice compared with pre-trained models in resource-constrained scenarios.

3.3 Analysis

In this subsection, we firstly verify the effectiveness of our proposed Mem-Att and OA-mixed connection, then show the impact of inference rounds on model performance, and finally further analyze the Mem-Att by a statistical analysis and a case study.

3.3.1 Effectiveness of the Mem-Att

We compare Mem-Att with three attention mechanisms to verify the ability of Mem-Att to maintain richer interactive information and leverage historical interactive information to aid future inference, containing 1) the scaled dot product attention (Dot-Att); 2) the scaled weighted dot product attention (wDot-Att); and 3) the interactive attention (Inter-Att), a variance of Mem-Att, which is only based on the current interaction matrix. The functions of these attentions are shown as following:

$$Att = \begin{cases} Pool_{att}(F(M^c)) & \text{Inter-Att} \\ \frac{ST^T}{\sqrt{d}} & \text{Dot-Att} \\ \frac{SWT^T}{\sqrt{d}} & \text{wDot-Att} \end{cases} \quad (18)$$

Table 2: Comparison experiments about different attention mechanisms on the Quora dataset.

Attentions	Dev Acc.	Test Acc.
Dot-Att	88.4	88.0
wDot-Att	88.8	88.3
Inter-Att	90.4	90.1
Mem-Att	**90.8**	**90.6**

Table 3: Comparison experiments about different connectivity patterns on the Quora dataset.

Patterns	Dev Acc.	Test Acc.
Direct	90.3	90.0
Dense	90.4	90.3
Residual	90.5	90.2
OA-mixed	**90.8**	**90.6**

where d is the dimension of the question representation, $S \in \mathbb{R}^{d \times m}$, $T \in \mathbb{R}^{d \times n}$ and $W \in \mathbb{R}^{d \times d}$.

The comparison results are shown in Table 2. With an intuition that the 3D interaction matrix can keep richer interactive information than the 2D attention map, the performance of the Int-Att is unsurprisingly better than those of the wDot-Att and the Dot-Att, which demonstrates that richer interactive information can bring benefit to the model on conducting more proper inference. Furthermore, the performance of the Mem-Att is better than that of the Int-Att, which reflects that the historical interactive information can provide assistance on the current and the future inference.

3.3.2 Effectiveness of the OA-mixed Connection

To illustrate the advantage of the OA-mixed connection, we compare our method with the following connective patterns: 1) residual connection (He et al., 2016); 2) dense connection (Huang et al., 2017), and 3) direct connection that directly treats the output of the previous round as the input.

As shown in Table 3, the direct connection unsurprisingly performs worst. These results show that the historical representation information provides benefits for the current round of inference and it is critical to design advanced connectivity patterns to effectively transmit important information between different reasoning rounds. Moreover, our OA-mixed connection beats both the residual

connection and the dense connection. We attribute it to the fact that our method can preserve the entire information of original texts. Meanwhile, the average connection we proposed can help the model to focus more on the information conveyed by the surrounding reasoning rounds. All these bring rich information and helpful hints to determine the relationships between the question pair.

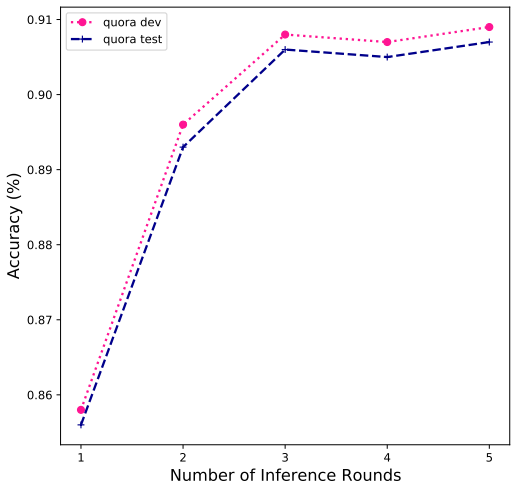


Figure 4: The accuracy curve for different rounds of inference on the Quora dataset.

3.3.3 Impact of the Inference Rounds

In this part, we design a comparison experiment to demonstrate the impact of the inference rounds. We set the inference round in our FITN from 1 to 5 and compare their performance on Quora's development and test set. The comparison result is shown in Figure 4. Obviously, as the number of the inference round increases, our model's accuracy increases, verifying the utility of the multi-round inference. However, the increasing trend of the accuracy gradually slows down as the number of inference rounds grows. Continue stacking layers may not bring further significant improvements. We attribute this to the model capturing enough information from a limited multi-round inference under the assistance of our proposed Mem-Att and OA-mixed connection. There is no need to stack too many inference modules.

3.3.4 Analysis of the Mem-Att

In order to further analyze how the Mem-Att works, we compare our Mem-Att with the Dot-Att and conduct a statistical analysis along with a case study to verify that the Mem-Att can pay higher attention to the critical word pairs and the inference round

Table 4: Statistical analysis on the Quora dataset. "Mean±std" denotes the mean value and the standard deviation of the attention distribution. R1., R2., and R3. denote the attention in the round 1, 2, and 3 respectively.

Metric	Mem-Att	Dot-Att
R1. Mean±std	0.0967±0.0026	0.0967±0.0017
R2. Mean±std	0.0967±0.0028	0.0967±0.0003
R3. Mean±std	0.0967±0.0029	0.0967±0.0013
Pearson(R2,R1)	0.7327	0.5456
Pearson(R3,R2)	0.8521	0.5390

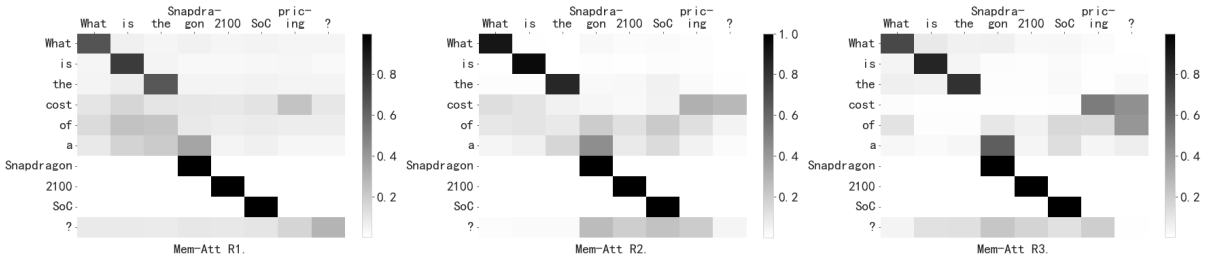
in the Mem-Att is progressive.

Statistical Analysis We conduct the statistical analysis on the development set of Quora and compare our Mem-Att with the Dot-Att. We calculate the mean value and the standard deviation of the attention distributions in each inference round to observe the distribution characteristics. Then, we calculate the Pearson correlation coefficient (Benesty et al., 2009) to quantify the relevance between two attention distributions in adjacent rounds. We take the average of the above metrics among all samples as the final metrics.

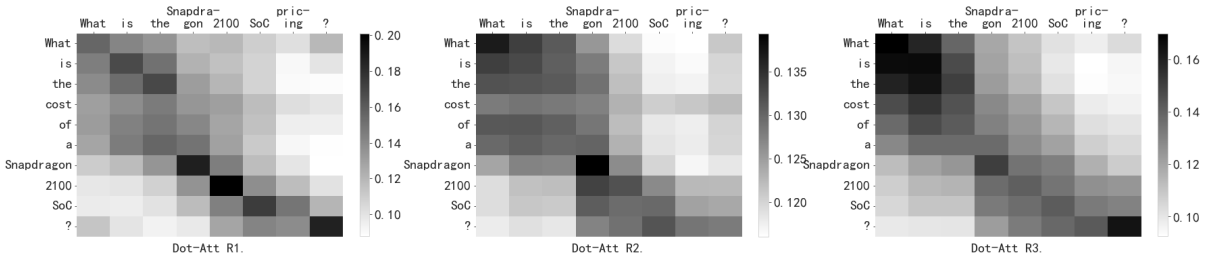
As shown in Table 4, the standard deviation of the attention distributions in the Mem-Att is larger than that in the Dot-Att and the distribution of the Dot-Att tends to be uniform. It demonstrates that our Mem-Att is more discrete and pays more attention to the specific token pairs. Besides, the Pearson correlation coefficient between the attention distributions of the Mem-Att in adjacent rounds is higher than that in the Dot-Att, which denotes that the inference between adjacent rounds has more relevance in the Mem-Att. The inference procedure in the Mem-Att is progressive.

Case Study Then, we take a pair of similar questions "What is the cost of a Snapdragon 2100 SoC ?" and "What is the Snapdragon 2100 SoC pricing ?" as an example and visualize the attention distributions of the Mem-Att and the Dot-Att in each round of inference. Here, the Mem-Att predicts right and the Dot-Att predicts wrong.

As shown in Figure 5, both the Dot-Att and the Mem-Att can align all pairs of the same word in the first inference round, where the Mem-Att focuses more on these word pairs than the Dot-Att. The distribution of the Mem-Att is more concentrated than that of the Dot-Att, which denotes that



(a) The heat maps of Mem-Att.



(b) The heat maps of Dot-Att.

Figure 5: The heat maps of Mem-Att and Dot-Att. Take "What is the cost of a Snapdragon 2100 SoC ?" and "What is the Snapdragon 2100 SoC pricing ?" as an example.

the Mem-Att has obvious tendency to pay attention. With the increase in the number of inference rounds, the Mem-Att's distribution does not tend to be uniform. Furthermore, the change of the Mem-Att's distribution is continuous, where the Mem-Att gradually deepens its focus on "cost" and "pricing". It demonstrates that the inference in the Mem-Att is progressive. The Mem-Att can gradually align word pairs with similar semantics.

4 Related Work

Question Matching can be regarded as a semantic matching task, which core lies in how to model the vector representation of texts (Shen et al., 2018; Reimers and Gurevych, 2019; Gao et al., 2021) and reason about the semantic relationship between text pairs. ESIM (Chen et al., 2017) encodes texts through BiLSTM or TreeLSTM (Socher et al., 2013) and applies the co-attention to extract fine-grained alignment information for inference. BiMPM (Wang et al., 2017) matches texts from multiple perspectives by multiple kinds of attentions. For better inference, many studies tend to employ deeper models. DIIN (Gong et al., 2018) applies a dense-net on the interaction matrix extracted from two texts for deep inference. DRCN (Kim et al., 2019) iterates one same block multiple times for multi-turn inference. TIM-W (Zhou et al., 2020) is based on deep mutual information estimation. ADIN (Liang et al., 2019) performs multiple rounds of asynchronous reasoning for the NLI task. In comparison, our FITN performs better due to the better utilization of historical information.

Thanks to the knowledge obtained from massive data, pre-trained models can greatly improve the performance of semantic matching, such as BERT (Devlin et al., 2019) AlBERT (Lan et al., 2020). However, the complexity of the model and the time consumption of reasoning are greatly increased, making them not suitable to resource-constrained scenarios. Enhanced-RCNN (Peng et al., 2020) compares itself with BERT in inference speed and accuracy. Although the performance is relative low, its inference speed is 10 times faster than BERT-base. Under the resource-constrained condition, directly using publicly available tiny pre-trained models is another solution. These models are commonly pre-trained with large-scale corpus (like AlBERT-tiny and AlBERT-base (Lan et al., 2020)) or distilled from large pre-trained models (like BERT-tiny (Turc et al., 2019)). Compared with these tiny pre-trained models, our FITN achieves better performance.

5 Conclusion

In this paper, we study the task of question matching and propose a Full Information Transmission

Network (FITN) that can utilize both the historical representation and the historical interactive information together in a simultaneous fashion. Specifically, the FITN employs a memory-based attention to keep and transmit the historical interactive information and an original-average mixed connectivity pattern to transmit the representation information. Experimental results on two benchmarks show that our FITN takes advantage of both kinds of information and outperforms strong baselines with considerable margin.

Acknowledgement

This work is supported by the Key Development Program of the Ministry of Science and Technology (No.2019YFF0303003), the National Natural Science Foundation of China (No.61976068) and "Hundreds, Millions" Engineering Science and Technology Major Special Project of Heilongjiang Province (No.2020ZX14A02).

Ethical Considerations

This paper proposes a general model for question matching. This paper neither introduces any social/ethical bias to the model nor amplify any bias in the data. In all the experiments, we use public datasets and consist their intended use. We build our algorithms using public code bases (PyTorch). We do not foresee any direct social consequences or ethical issues.

References

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In Noise reduction in speech processing, pages 1–4. Springer.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146.

Lu Chen, Yanbin Zhao, Boer Lyu, Lesheng Jin, Zhi Chen, Su Zhu, and Kai Yu. 2020. Neural graph matching networks for Chinese short text matching. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6152–6158, Online. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6894–6910.

Yichen Gong, Heng Luo, and Jian Zhang. 2018. Natural language inference over interaction space. In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pages 770–778. IEEE Computer Society.

Zhe Hu, Zuohui Fu, Yu Yin, and Gerard de Melo. 2021. Context-aware interaction network for question matching. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3846–3853.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely connected convolutional networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 2261–2269. IEEE Computer Society.

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. 2017. First quora dataset release: Question pairs. data. quora. com.

Seonhoon Kim, Inho Kang, and Nojun Kwak. 2019. Semantic sentence matching with densely-connected recurrent and co-attentive information. In The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, pages 6586–6593. AAAI Press.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In

3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Di Liang, Fubao Zhang, Qi Zhang, and Xuanjing Huang. 2019. Asynchronous deep interaction network for natural language inference. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2692–2700, Hong Kong, China. Association for Computational Linguistics.

Xin Liu, Qingcai Chen, Chong Deng, Huajun Zeng, Jing Chen, Dongfang Li, and Buzhou Tang. 2018. LCQMC:a large-scale Chinese question matching corpus. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1952–1962, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Boer Lyu, Lu Chen, Su Zhu, and Kai Yu. 2021. Let: Linguistic knowledge enhanced graph transformer for chinese short text matching. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 13498–13506.

Shuang Peng, Hengbin Cui, Niantao Xie, Sujian Li, Jiaxing Zhang, and Xiaolong Li. 2020. Enhanced-rcnn: An efficient method for learning sentence similarity. In WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020, pages 2500–2506. ACM / IW3C2.

Yuanyuan Qiu, Hongzheng Li, Shen Li, Yingdi Jiang, Renfen Hu, and Lijiao Yang. 2018. Revisiting correlations between intrinsic and extrinsic evaluations of word embeddings. In Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, pages 209–221. Springer.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. 2018. Disan: Directional self-attention network for rnn/cnn-free language understanding. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pages 5446–5455. AAAI Press.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Yi Tay, Anh Tuan Luu, and Siu Cheung Hui. 2018. Co-stack residual affinity networks with multi-level attention refinement for matching text sequences. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4492–4502, Brussels, Belgium. Association for Computational Linguistics.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: The impact of student initialization on knowledge distillation. CoRR, abs/1908.08962.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017, pages 4144–4150. ijcai.org.

Runqi Yang, Jianhai Zhang, Xing Gao, Feng Ji, and Haiqing Chen. 2019. Simple and effective text matching with richer alignment features. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4699–4709, Florence, Italy. Association for Computational Linguistics.

Xixi Zhou, Chengxi Li, Jiajun Bu, Chengwei Yao, Keyue Shi, Zhi Yu, and Zhou Yu. 2020. Matching text with deep mutual information estimation. arXiv preprint arXiv:2003.11521.