ICANN19
28th International Conference on Artificial Neural Networks

Dependent Multilevel Interaction Network for Natural language Inference

$Yun\ Li^1, Yan\ \ Yang^1, Yong\ Deng^1, Qinmin\ Vivian\ Hu^2, Chengcai\ Chen^3, Liang\ He^1$ and $Zhou\ Yu^4$
1 East China Normal University, Shanghai, China 2 Ryerson University, Toronto, Ontario, Canada 3 Xiaoi Research, Shanghai, China 4 University of California, Davis, USA
Contact Author: yanyang@cs.ecnu.edu.cn

Introduction

Natural language inference (NLI) is a challenging natural language processing (NLP) task which requires one to determine whether the logical relationship between two sentences is among entailment (the hypothesis must be true if the premise is true), contradiction (the hypothesis must be false if the premise is true) and neutral (neither entailment nor contradiction). Generally, NLI is also related to many other NLP tasks under the paradigm of semantic matching of two sentences. An essential challenge is to capture the semantic relevance of the two sentences.

In this paper, we propose a new interaction model, named Dependent Multilevel Interaction (DMI) network, which models multiple interactions between a premise and a hypothesis to capture more comprehensive information.

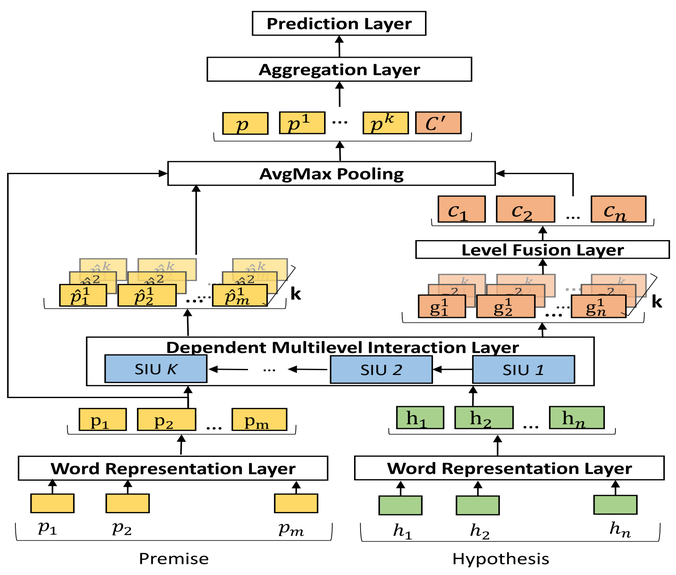


Fig. 1. Dependent Multilevel Interaction Model

Proposed Model

Our proposed DMI including two important layers: dependent multilevel interaction layer and level fusion layer. The framework of the DMI network is presented in Figure 1.

Dependent Multilevel Interaction Layer

This layer provides multiple interactions by adopting a serial of single-interaction Units and cascading them together to capture comprehensive information.

Single-interaction Unit Single-interaction unit provides an interaction between a premise and a hypothesis, which combines the attention mechanism and the comparison module. The structure is shown in Figure 2.

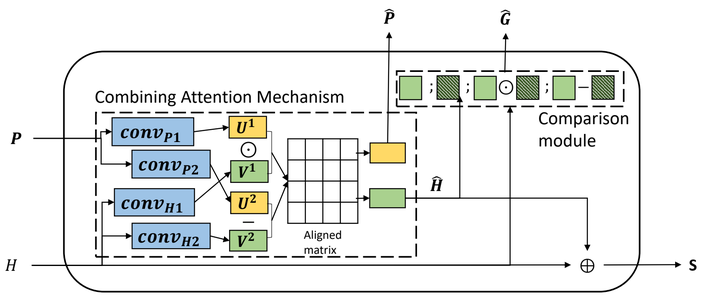


Fig. 2. The Structure of Single-interaction Unit (SIU)

Combining Attention Mechanism

$$U^1, U^2, V^1, V^1 = \left[conv_{p1}(P), conv_{p2}(P), conv_{H1}(H), conv_{H2}(H)\right]$$
$$s[i;j] = F([U^1 \odot V^1; U^2 - V^2]), i \in [1,m], j \in [1,n]$$

$$\hat{P}_i = \sum_j \frac{\exp(s_{ij}^T)}{\sum_k \exp(s_{ik})} H_j\ , \qquad \hat{H}_i = \sum_j \frac{\exp(s_{ij}^T)}{\sum_k \exp(s_{ik})} P_j$$

Comparison Module

$$G = \left[H; \hat{H}; H - \hat{H}; H \odot \hat{H}\right], \qquad S = H + \hat{H}$$

Totally, we achieve three features from a SIU structure:

$$\hat{P}, G, S = SIU(P, H)$$

Levels dependency To reduce information redundancy and enhance the dependency between the adjacent interactions, we update the input H_i of SIU_i as follows:

$$\mathrm{H}_i = \begin{cases} H, & if\ i == 1 \\ S^{i-1}, & if\ i > 1 \end{cases}$$

Level Fusion Layer

To fuse features from each level, we make some operations on G:

$$G' = max(split(F(G)))$$

where *F(·)* is a standard projection layer with *ReLU* activation function, *split(·)* is a function that splits the input vector at the last axis, max(·) is to reduce dimensions of vectors by choosing the max numerical in each axis.
Now we use G' to replace G. Based on G' , we use an attention mechanism to get a wide representation C:

$$A = softmax(ReLU(G'W_1)W_2)$$
$$C = \sum_k A_{ik} G'_{ik}$$

Where W_1, W_2 are trainable parameters.
Finally, we employ a BiGRU to encode C and then use an avg-max pooling to obtain a fixed vector of level-comparison:

$$C' = [avgPooling(BiGRU(C)); maxPooling(BiGRU(C))]$$

Experimental Result

We evaluate our model over the SicTail and SNLI datasets and achieve competitive scores. The ablation study shows the effectiveness of each components in our model.

Model	Dev acc	Test acc
1. Majority Class [9]	63.3	60.3
2. Ngram [9]	65.0	70.6
3. ESIM [21]	70.5	70.6
4. Decomposable Att [14]	75.4	72.3
5. DGEM [9]	79.6	77.3
6. CAFE [20]	-	83.3
7. DMI(our model)	89.6	**85.5**

Table 1: Results on SciTail dataset

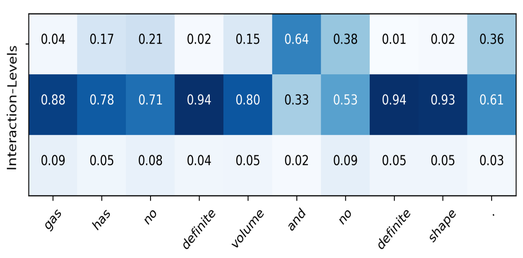
Model	Dev acc	Test acc
1. Handcrafted features [1]	99.7	78.2
2. TBCNN [13]	83.3	82.1
3. Gated-Att BiLSTM [2]	90.5	85.5
4. DiSAN [18]	91.1	85.6
5. ReSAN [19]	92.6	86.3
6. Attention LSTM [17]	85.3	83.5
7. Decomposable Att [14]	89.5	86.3
8. BiMPM [21]	90.9	87.5
9. DIIN [6]	91.2	88.0
10. DR-BiLSTM [5]	94.1	88.5
11. CAFE [20]	89.8	88.5
12. KIM [3]	94.1	88.6
13. ESIM [16]	92.6	88.0
14. ESIM+ELMo [16]	91.6	**88.7**
15. DMI (our model)	94.7	**88.7**

Table 2: Results on SNLI dataset

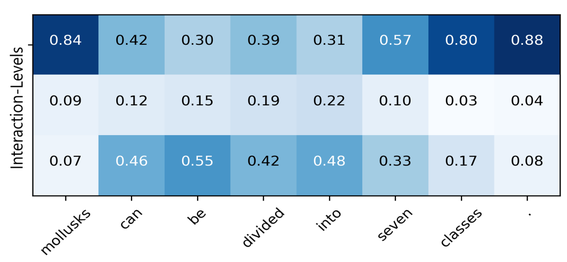
Model	Dev acc
1. Full model	89.64
2. w/o character feature	88.56
3. w/o POS feature	89.14
4. w/o EM feature	88.74
5. w/o parameters transfer	88.60
6. w/o dot product in combining attention	88.88
7. w/o subtraction in combining attention	88.50
8. w/o combining attention	89.03
9. w/o aggregation layer	87.65

Table 3: Ablation study Results on SciTail dataset

Attention visualization in dependent multilevel interaction



(a) **Premise:** department of education gas does not have a definite shape and it does not have a definite volume. **Hypothesis:** gas has no definite volume and no definite shape. **Label:** entailment



(b) **Premise:** divide the class into four teams of six or seven students. **Hypothesis:** mollusks can be divided into seven classes. . **Label:** neutral

Conclusion

In this work, we propose a dependent multilevel interaction model that provides multiple interactions by cascading a serial of single-interaction units (SIUs). Each SIU includes a novel attention mechanism and comparison module to model the interaction between the premise and the hypothesis. Experiments on two benchmark datasets demonstrate the efficacy of our model. In the future, we hope to improve the scalability of our model and apply it to other NLI tasks, such as machine reading comprehension and answer selection.

Other contact: hli@ica.stc.sh.cn

978-3-030-30489-8, ICANN 2019, Part IV, LNCS 11730