Context-Aware Interaction Network for Question Matching

Zhe Hu[1], Zuohui Fu[2], Yu Yin[3], and Gerard de Melo[4]

[1]Baidu Inc, [2]Rutgers University
[3]Northeastern University, [4]HPI/University of Potsdam

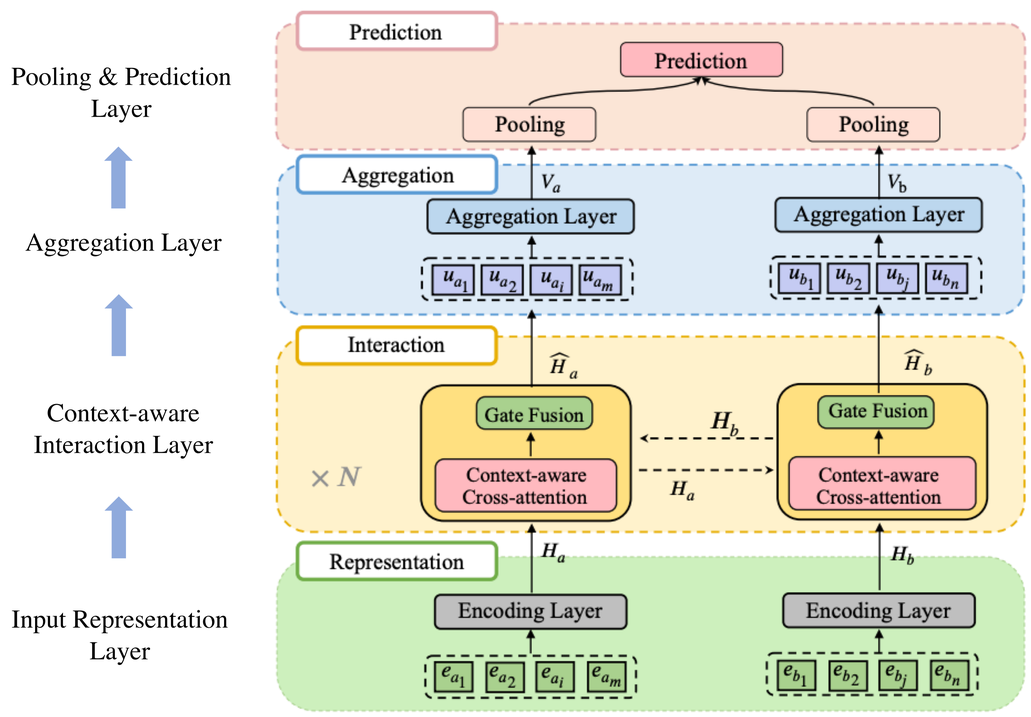
TASK

- Question matching aims to predict the semantic relationship given two questions

How do I know if my phone is tapped ?
→ How do I check if my phone is tapped ? *duplicate*
→ How does a landline call a cell phone ? *non-duplicate*

METHOD

- We propose COIN: a COntext-aware Interaction Network



Context-Aware Cross-Attention Layer

- We want to integrate **contextual features** C
- We apply a **self-alignment layer** to aggregate pertinent contextual information

Original cross-attention

Step 1: compute attention matrix:

$$\mathbf{E}_{ij} = \mathrm{Att}(\mathbf{h}_{a_i}, \mathbf{h}_{b_j}) = FFN(\mathbf{h}_{a_i})^{\mathrm{T}} FFN(\mathbf{h}_{b_j})$$

Step 2: compute alignments

$$\mathbf{a}_i = \mathrm{softmax}(\mathbf{E}_{i:}), \quad \mathbf{b}_j = \mathrm{softmax}(\mathbf{E}_{:j})$$

$$\mathbf{h}'_{b_j} = \sum_{k=1}^{m} \mathbf{b}_{kj} \mathbf{h}_{a_k}, \quad \mathbf{h}'_{a_i} = \sum_{k=1}^{n} \mathbf{a}_{ik} \mathbf{h}_{b_k}$$

Context-aware cross-attention

$$\mathbf{C}_a = \mathrm{Layer}_{\mathrm{self-align}}(\mathbf{H}_a)$$
$$\mathbf{C}_b = \mathrm{Layer}_{\mathrm{self-align}}(\mathbf{H}_b)$$
$$\mathbf{E}^{c}_{ij} = \mathrm{Att}_{\mathrm{context}}(\mathbf{h}_{a_i}, \mathbf{h}_{b_j}, \mathbf{c}_{a_i}, \mathbf{c}_{b_j}) = FFN(\mathbf{h}_{a_i} + \mathbf{c}_{a_i})^{\mathrm{T}} FFN(\mathbf{h}_{b_j} + \mathbf{c}_{b_j})$$

Contextual features (C_a & C_b) of two sequences are directly considered when computing the attention matrix

This mirrors human behavior that people tend to first read each sentence and pay attention to the salient contents, and then compare and match two sentences.

Gated Fusion Layer

- update sequence representations by blending alignments

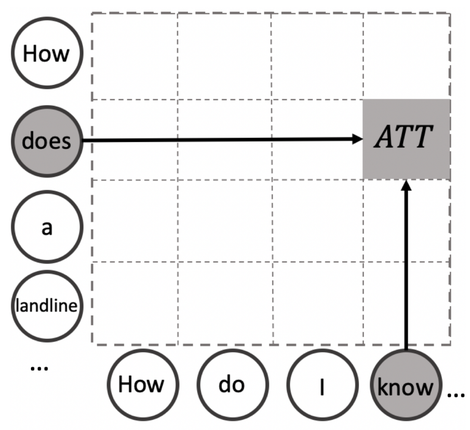
$$\boldsymbol{f}_i = \sigma(\mathbf{W}_1 \mathbf{h}_{a_i} + \mathbf{W}_2 \widetilde{\mathbf{h}}_{a_i} + \mathbf{b}_g) \quad \leftarrow \text{gate}$$
$$\widehat{\mathbf{h}}_{a_i} = \boldsymbol{f}_i \odot \mathbf{h}_{a_i} + (\mathbf{1} - \boldsymbol{f}_i) \odot \widetilde{\mathbf{h}}_{a_i} \quad \leftarrow \text{updated representation}$$

$$\widetilde{\mathbf{h}}^1_{a_i} = G_1([\mathbf{h}_{a_i}; \mathbf{h}'_{a_i}])$$
$$\widetilde{\mathbf{h}}^2_{a_i} = G_2([\mathbf{h}_{a_i}; \mathbf{h}_{a_i} - \mathbf{h}'_{a_i}])$$
$$\widetilde{\mathbf{h}}^3_{a_i} = G_3([\mathbf{h}_{a_i}; \mathbf{h}_{a_i} \odot \mathbf{h}'_{a_i}])$$
$$\widetilde{\mathbf{h}}_{a_i} = \mathrm{ReLU}(\mathbf{W}_f[\widetilde{\mathbf{h}}^1_{a_i}; \widetilde{\mathbf{h}}^2_{a_i}; \widetilde{\mathbf{h}}^3_{a_i}] + \mathbf{b}_f)$$

- We first compare the original representations and the aligned ones from difference perspectives
- We use gate operation to enable the model to flexibly incorporate aligned features by **controlling gates**;
- Gate operation is similar to a **skip connection** in mitigating the additional model complexity coming from the deeper structure

MOTIVATION

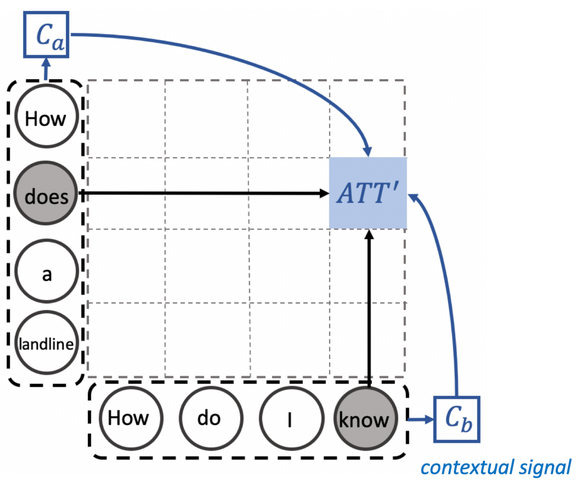
- **Cross-attention** is widely adopted for text matching



cross-attention

- Computes a **word-by-word attention matrix** to obtain alignments between two sequences
- Each value of the attention matrix is based on just two **individual tokens** from the sequences
- Mostly focus on **word-level local matching** and fail to fully account for **the overall semantics**

- We aim to contextualize **cross-attention** for better interaction



- Accurate matching requires a deeper understanding of the two sentences along with **pertinent linguistic patterns and constructions**
- We enables the model to consult contextual information while computing the attention matrix to measure the word relevance
- Yields better contextualized alignments for semantic reasoning

EXPERIMENTS

- **Datasets**: Quora Question Pairs & LCQMC
- **Results** :

Model	Acc (%)	F1 (%)
Lattice-CNN	82.1	82.4
ESIM	82.0	84.0
BiMPM	83.3	84.9
GMN	84.6	86.0
COIN (Ours)	**85.6**	**86.5**
BERT	85.7	86.8
Sentence-BERT	85.4	86.6
COIN (5-run ensemble)	**86.2**	**87.0**

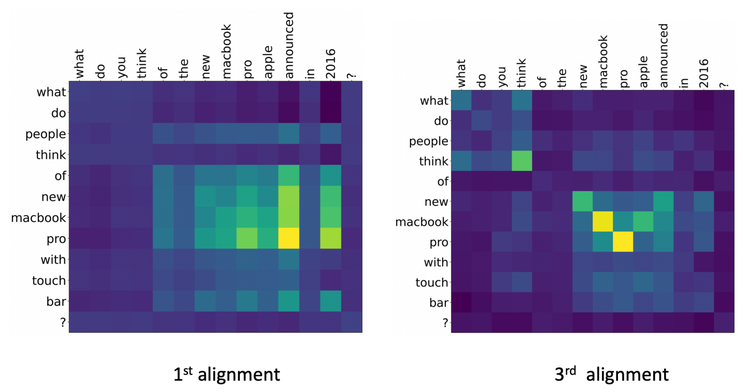
Table 1: Experimental results on LCQMC.

Model	Acc. (%)	Params
BiMPM	88.2	1.6M
DIIN	89.0	4.4M
CAFE	88.7	4.7M
OSOA-DFN	89.0	10.0M
RE2	89.2	2.8M
COIN (ours)	**89.4**	6.5M
BERT	90.1	109.5M
Sentence-BERT	90.6	109.5M
COIN (ensemble)	**90.7**	32.5M

Table 2: Experimental results on Quora dataset.

- Better results than non-pretrained methods
- Comparable results with pre-trained Methods with fewer parameters
- 5-run ensemble model outperforms BERT and Sentence-BERT

- **Visualization of Attention**:



1st alignment　　3rd alignment

- model learns to refine the alignments from low-level to high-level;
- the structured phrase such as " what do you think of " is also connected.

- **Ablation**:

Model	Quora	LCQMC
original	89.6	85.4
w/o context	89.1	84.8
simple fusion	88.8	85.2
w/o aggregat.	89.2	84.9
simple pool	89.4	85.2

- Ablation results confirms the effectiveness of each module;
- Removing context-aware alignments brings perform decrease on both datasets.