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

Background & Motivations

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*

哪个输入法好用？
Which input method works well?

→ 输入法哪个好用？ *duplicate*
Which input method is easy to use?

→ 你用过什么输入法？ *non-duplicate*
Which input method have you used?

Background & Motivations

Current Methods for Text Matching

Sentence Encoding Approach

Sentence Interaction Approach

Pre-trained LM

Background & Motivations

Current Methods for Text Matching

Sentence Encoding Approach

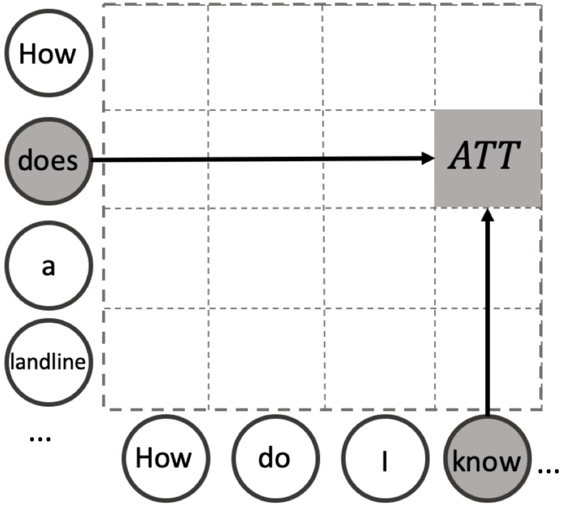
Sentence Interaction Approach

Pre-trained LM

Attention mechanism

Background & Motivations

Cross-attention is widely adopted for text matching



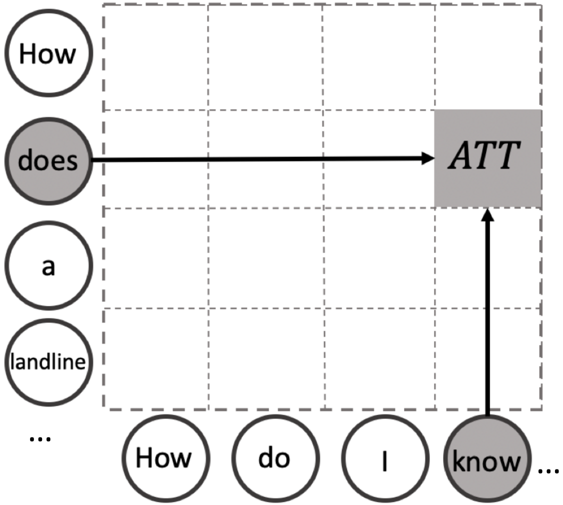
cross-attention

Current word-level cross-attention:

- Computes a word-by-word attention matrix to obtain alignments between two sequences

Background & Motivations

Current **Cross-attention** mostly focuses on word-level links



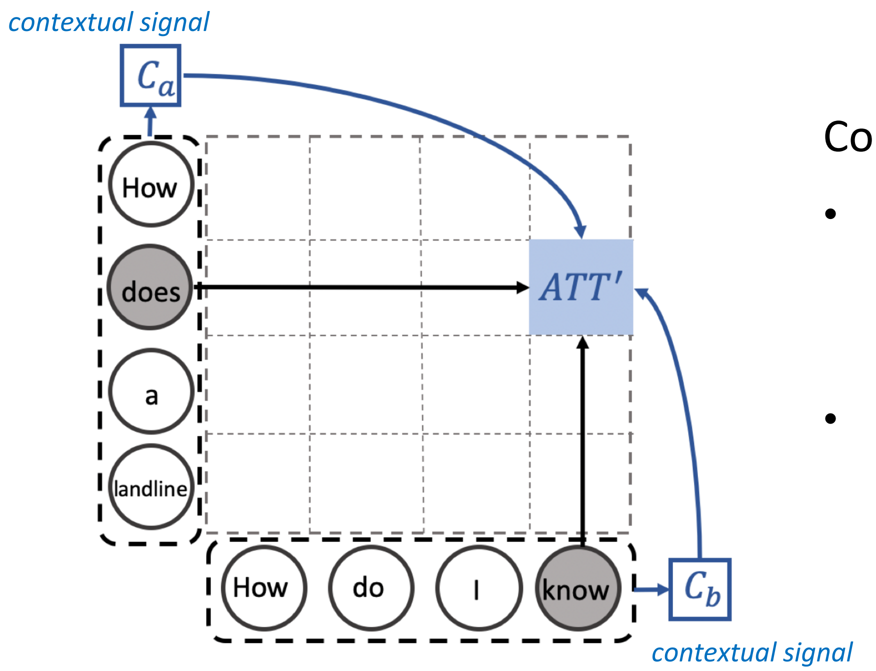
cross-attention

Current word-level 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

Background & Motivations

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



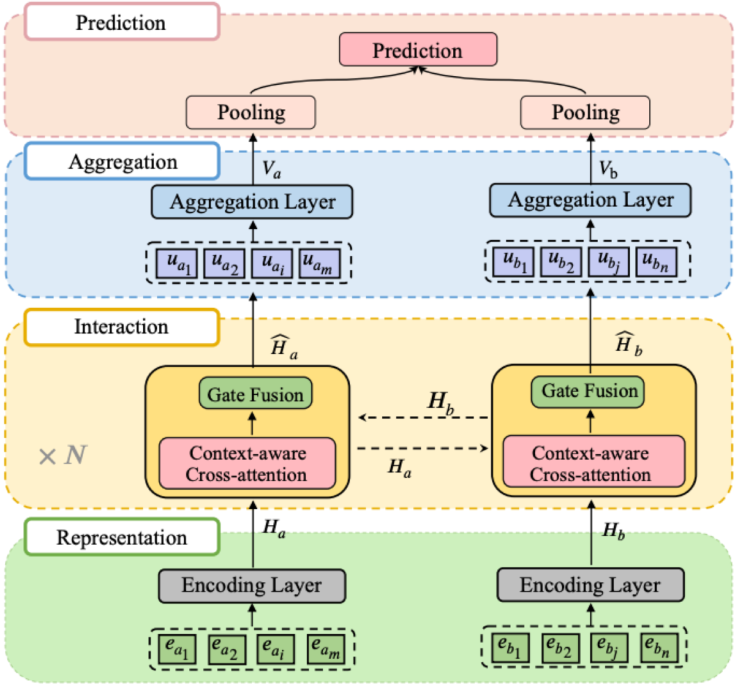
context-aware cross-attention

Context-aware cross-attention:

- Enables the model to consult contextual information while computing the attention matrix to measure the word relevance
- Yields better contextualized alignments for semantic reasoning

Methods

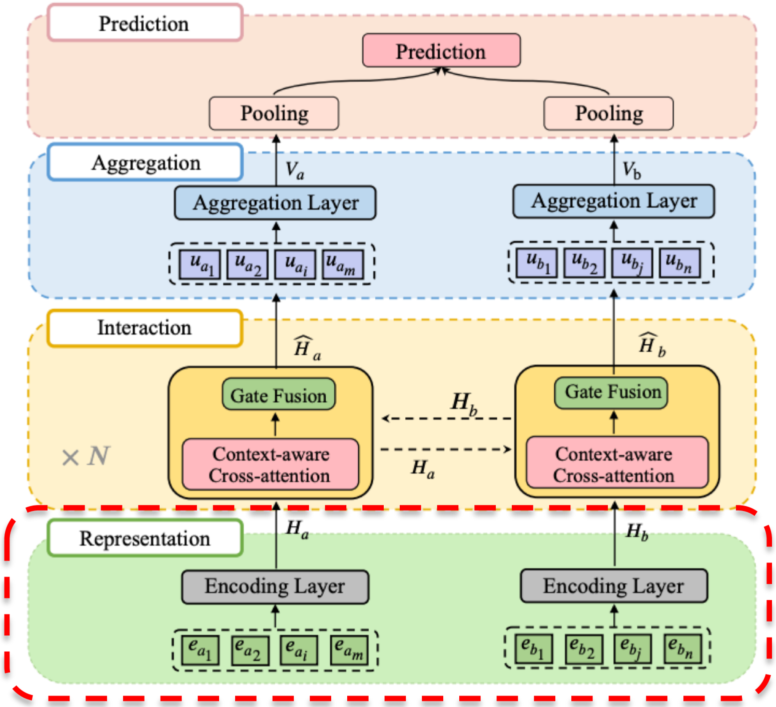
COIN: COntext-aware Interaction Network



- Pooling & Prediction Layer
- Aggregation Layer
- Context-aware interaction Layer
- Input Representation Layer

Methods

COIN: COntext-aware Interaction Network



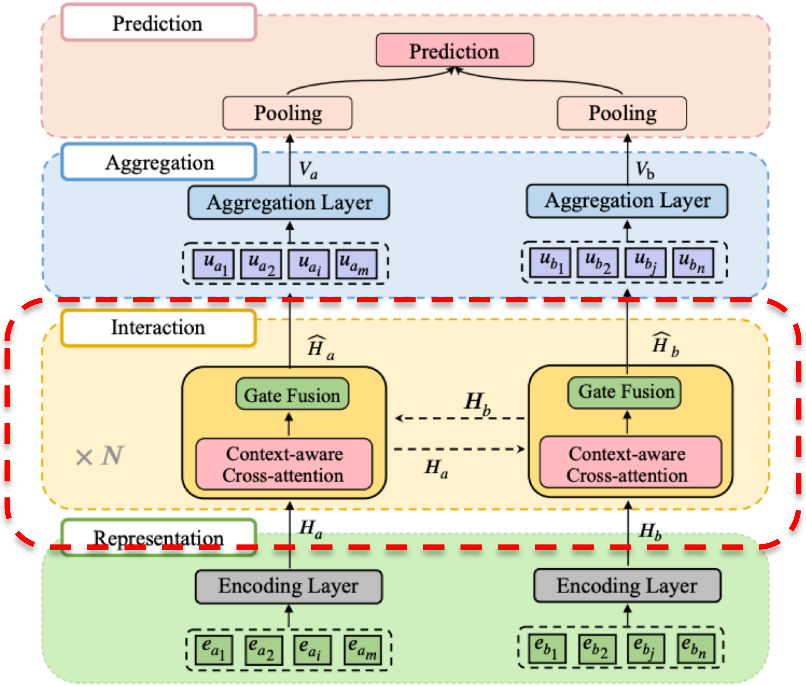
- **Input Representation Layer**: converts sentences into matrix representations with an embedding and encoding layer

$$\mathbf{H}_a = \mathrm{Layer}_{\mathrm{input}}(\mathbf{S}_{\mathrm{a}})$$

$$\mathbf{H}_b = \mathrm{Layer}_{\mathrm{input}}(\mathbf{S}_{\mathrm{b}})$$

Methods

COIN: COntext-aware Interaction Network



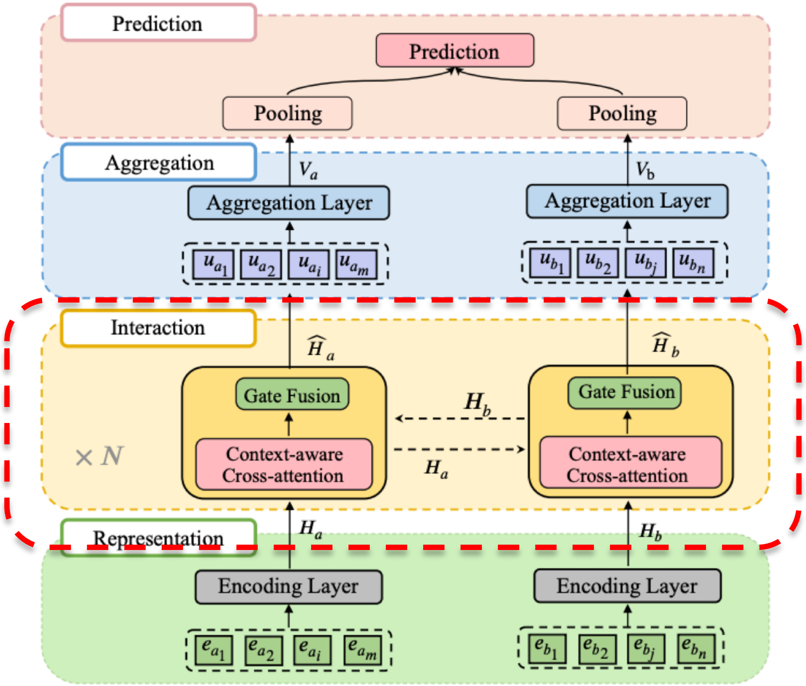
- **Context-aware interaction Layer**

1. Context-Aware Cross-Attention Layer: computes the aligned information of each sequence;

2. Gated Fusion Layer: updates sequence representations by blending the alignments with the original ones;

Methods

COIN: COntext-aware Interaction Network



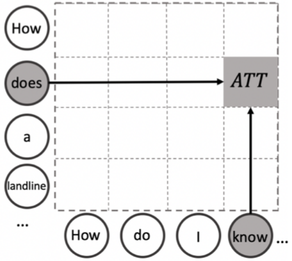
- **Context-aware interaction Layer**

1. Context-Aware Cross-Attention Layer

Original cross-attention

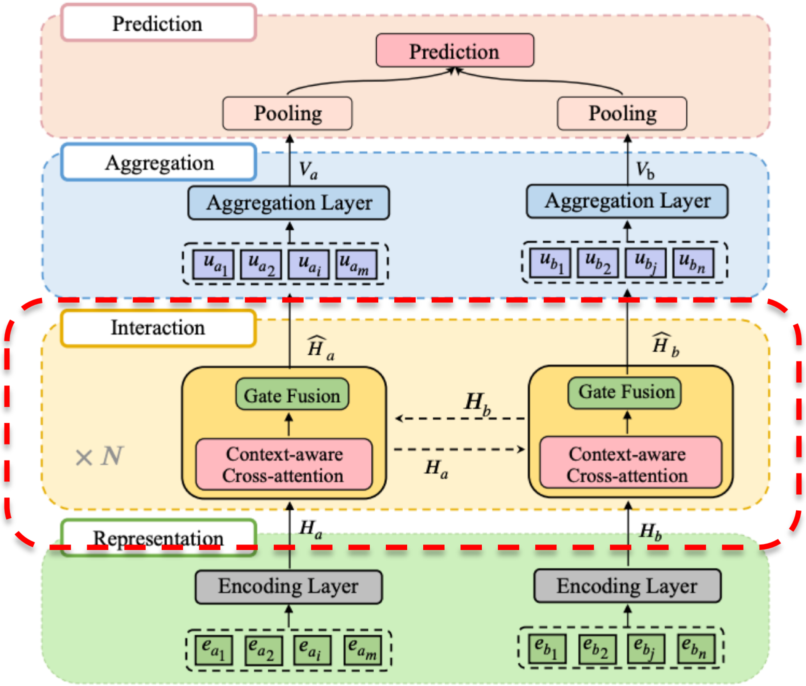
Step 1: compute attention matrix:

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



Methods

COIN: COntext-aware Interaction Network



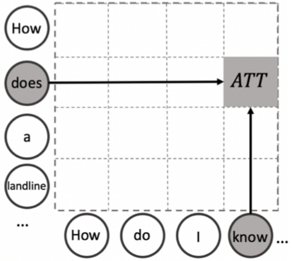
- **Context-aware interaction Layer**

1. Context-Aware Cross-Attention Layer

Original cross-attention

Step 1: compute attention matrix:

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



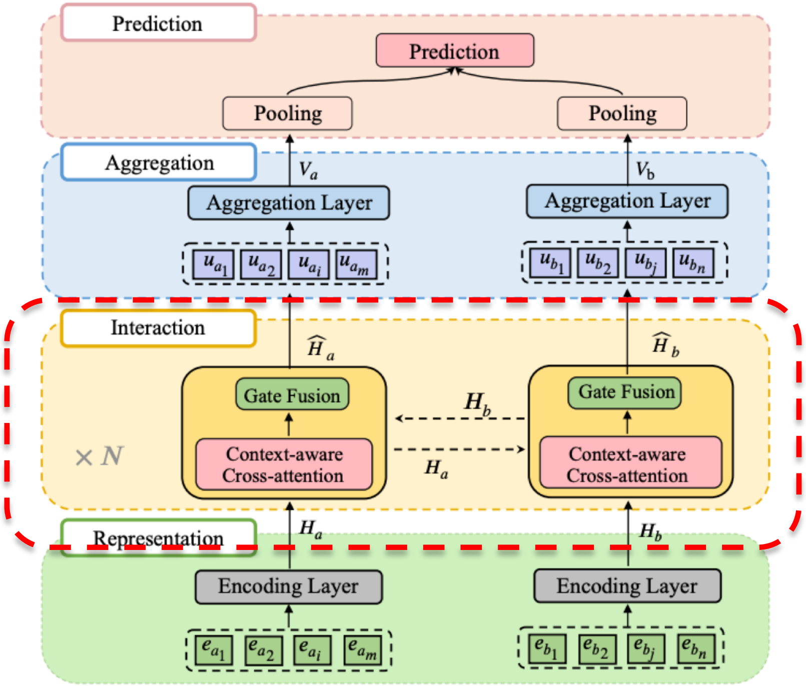
Step 2: compute alignments

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

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

Methods

COIN: COntext-aware Interaction Network



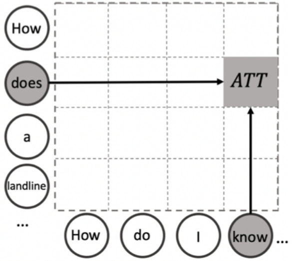
- **Context-aware interaction Layer**

1. Context-Aware Cross-Attention Layer

Original cross-attention

Step 1: compute attention matrix:

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



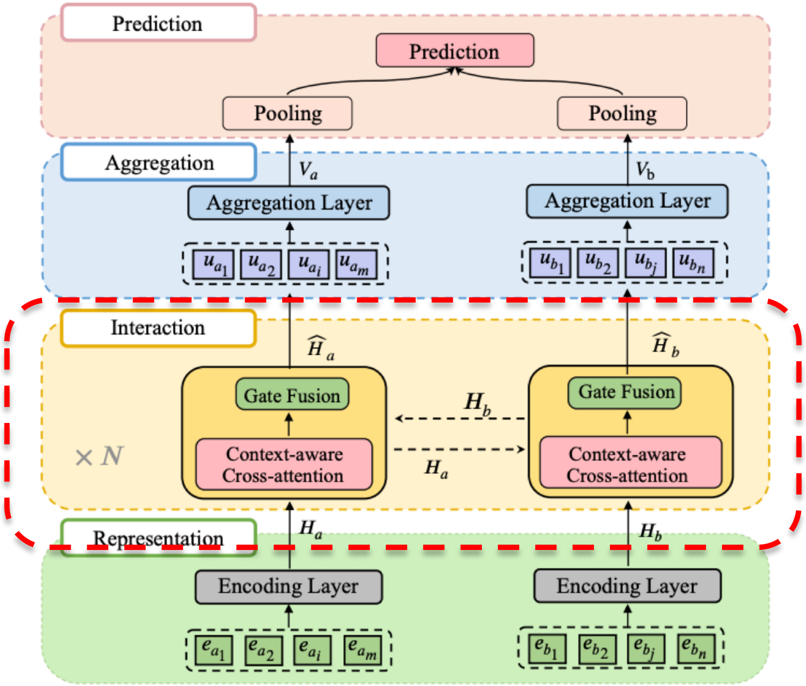
Step 2: compute alignments

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

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

Methods

COIN: COntext-aware Interaction Network

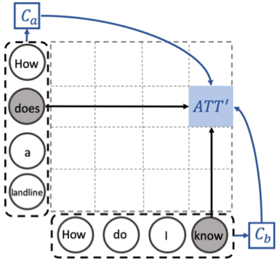


- **Context-aware interaction Layer**

1. Context-Aware Cross-Attention Layer

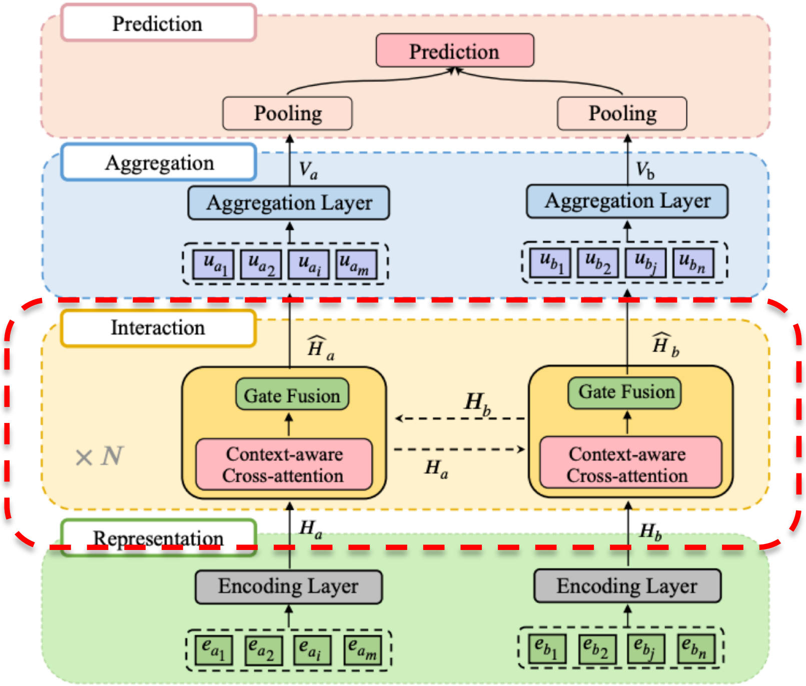
Context-aware cross-attention

- *We want to integrate **contextual features C***



Methods

COIN: COntext-aware Interaction Network

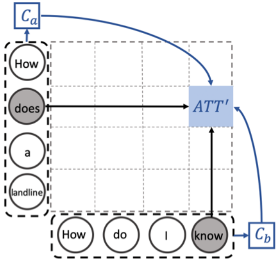


- **Context-aware interaction Layer**

1. Context-Aware Cross-Attention Layer

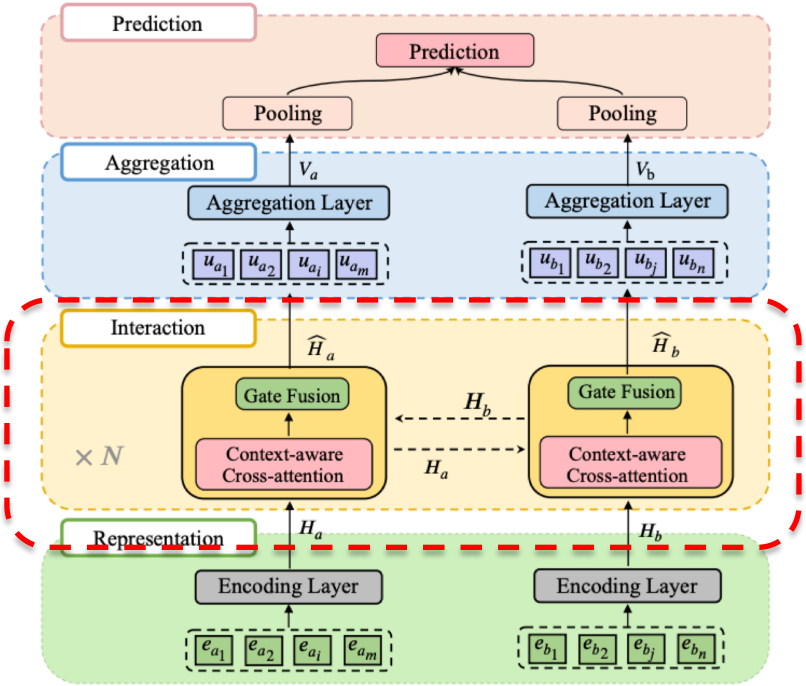
Context-aware cross-attention

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



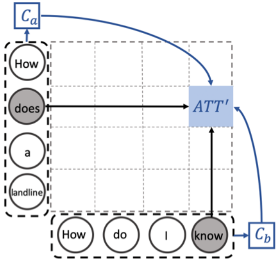
Methods

COIN: COntext-aware Interaction Network



- **Context-aware interaction Layer**

1. Context-Aware Cross-Attention Layer



Context-aware cross-attention

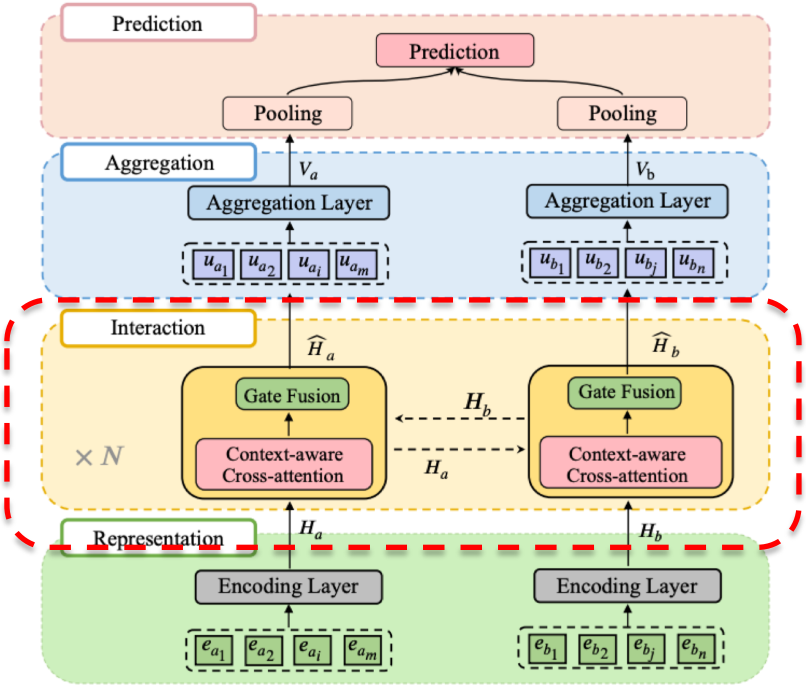
- *We want to integrate **contextual features C***
- *We apply a **self-alignment layer** to aggregate pertinent contextual information*

$$\mathbf{C}_a = \text{Layer}_{\text{self-align}}(\mathbf{H}_\text{a}) \qquad \mathbf{C}_b = \text{Layer}_{\text{self-align}}(\mathbf{H}_\text{b})$$

$$\begin{aligned}\mathbf{E}^{\text{c}}_{ij} &= \text{Att}_{\text{context}}(\mathbf{h}_{a_i}, \mathbf{h}_{b_j}, \mathbf{c}_{a_i}, \mathbf{c}_{b_j}) \\ &= FFN(\mathbf{h}_{a_i} + \mathbf{c}_{a_i})^{\text{T}} FFN(\mathbf{h}_{b_j} + \mathbf{c}_{b_j})\end{aligned}$$

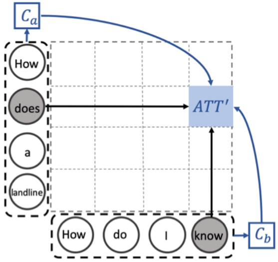
Methods

COIN: COntext-aware Interaction Network



- **Context-aware interaction Layer**

1. Context-Aware Cross-Attention Layer



Context-aware cross-attention

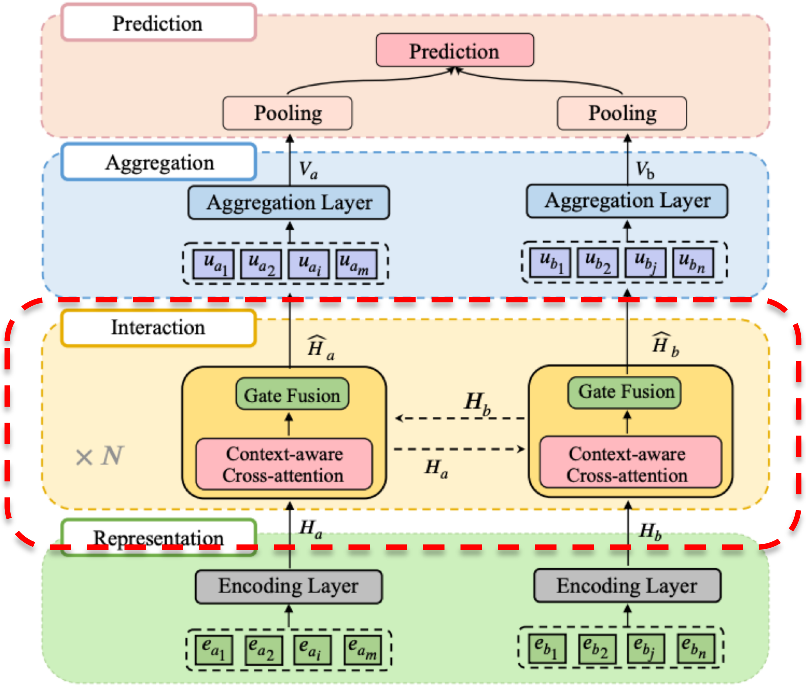
- *We want to integrate **contextual features C***
- *We apply a **self-alignment layer** to aggregate pertinent contextual information*

$$\mathbf{C}_a = \text{Layer}_{\text{self-align}}(\mathbf{H}_\text{a}) \qquad \mathbf{C}_b = \text{Layer}_{\text{self-align}}(\mathbf{H}_\text{b})$$

$$\begin{aligned}\mathbf{E}^{c}_{ij} &= \text{Att}_{\text{context}}(\mathbf{h}_{a_i}, \mathbf{h}_{b_j}, \mathbf{c}_{a_i}, \mathbf{c}_{b_j}) \\ &= FFN(\mathbf{h}_{a_i} + \mathbf{c}_{a_i})^{\mathrm{T}} FFN(\mathbf{h}_{b_j} + \mathbf{c}_{b_j})\end{aligned}$$

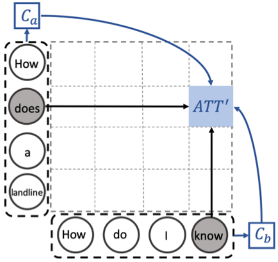
Methods

COIN: COntext-aware Interaction Network



- **Context-aware interaction Layer**

1. Context-Aware Cross-Attention Layer



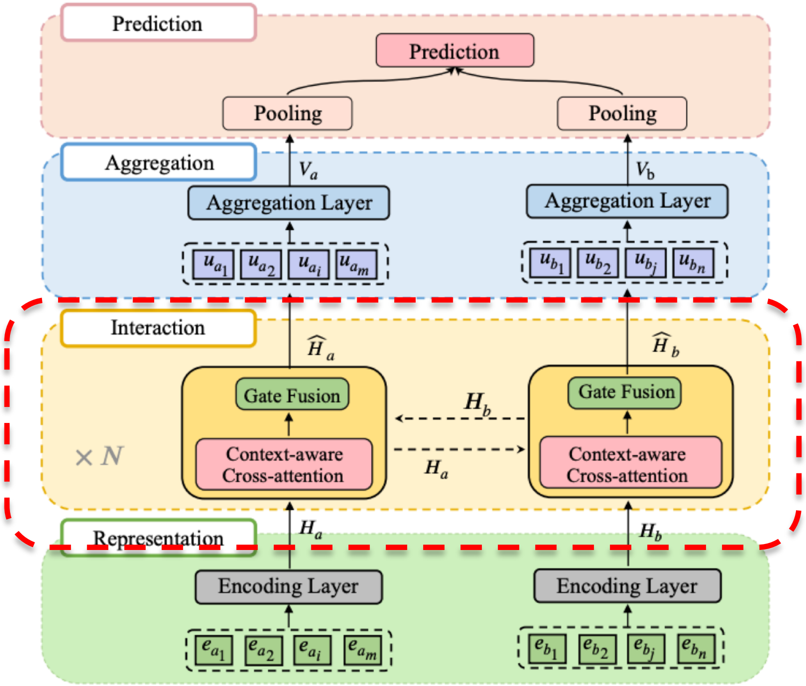
Context-aware cross-attention

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

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.

Methods

COIN: COntext-aware Interaction Network



- **Context-aware interaction Layer**

2. Gated Fusion Layer

$$\boldsymbol{f}_i = \sigma(\mathbf{W}_1 \mathbf{h}_{a_i} + \mathbf{W}_2 \widetilde{\mathbf{h}}_{a_i} + \mathbf{b}_g)$$

gate

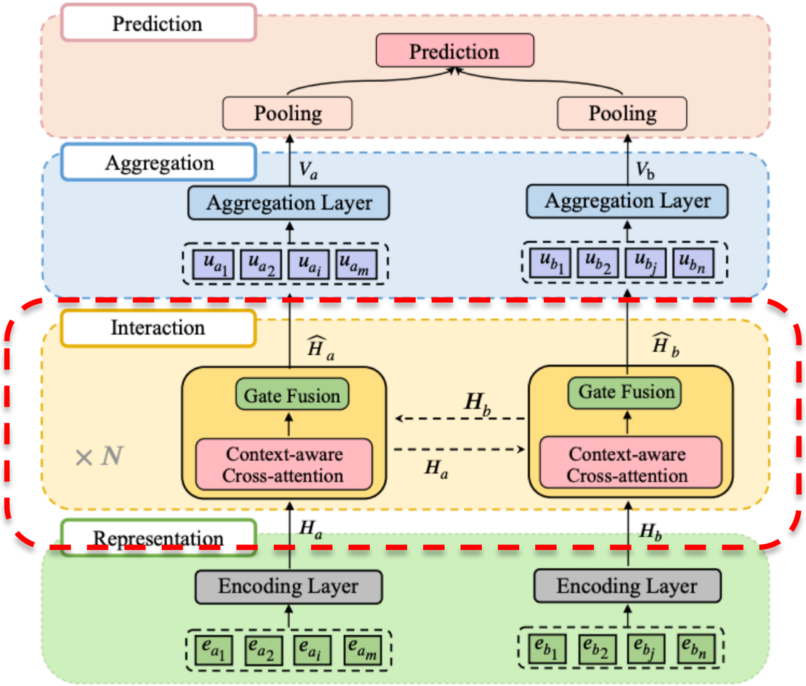
updated representation

$$\widehat{\mathbf{h}}_{a_i} = \boldsymbol{f}_i \odot \mathbf{h}_{a_i} + (\mathbf{1} - \boldsymbol{f}_i) \odot \widetilde{\mathbf{h}}_{a_i}$$

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

Methods

COIN: COntext-aware Interaction Network



- **Context-aware interaction Layer**

2. Gated Fusion Layer

$$\boldsymbol{f}_i = \sigma(\mathbf{W}_1 \mathbf{h}_{a_i} + \mathbf{W}_2 \widetilde{\mathbf{h}}_{a_i} + \mathbf{b}_g)$$ *gate*

updated representation

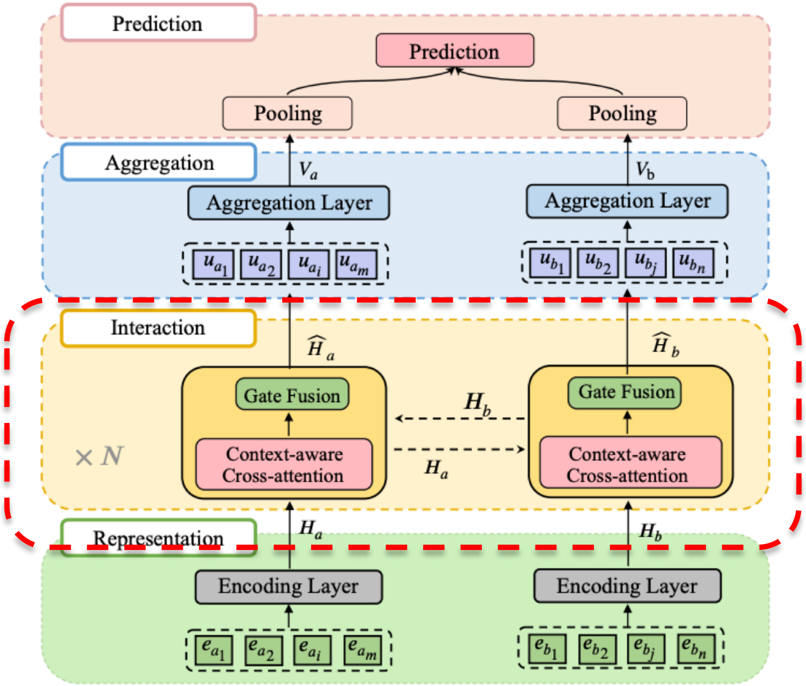
$$\widehat{\mathbf{h}}_{a_i} = \boldsymbol{f}_i \odot \mathbf{h}_{a_i} + (\mathbf{1} - \boldsymbol{f}_i) \odot \widetilde{\mathbf{h}}_{a_i}$$

Gated Fusion

- *Enable model to flexibly incorporate aligned features by* ***controlling gates****;*
- *Similar to a* ***skip connection*** *in mitigating the additional model complexity coming from the deeper structure (multiple interactions)*

Methods

COIN: COntext-aware Interaction Network



- **Context-aware interaction Layer**

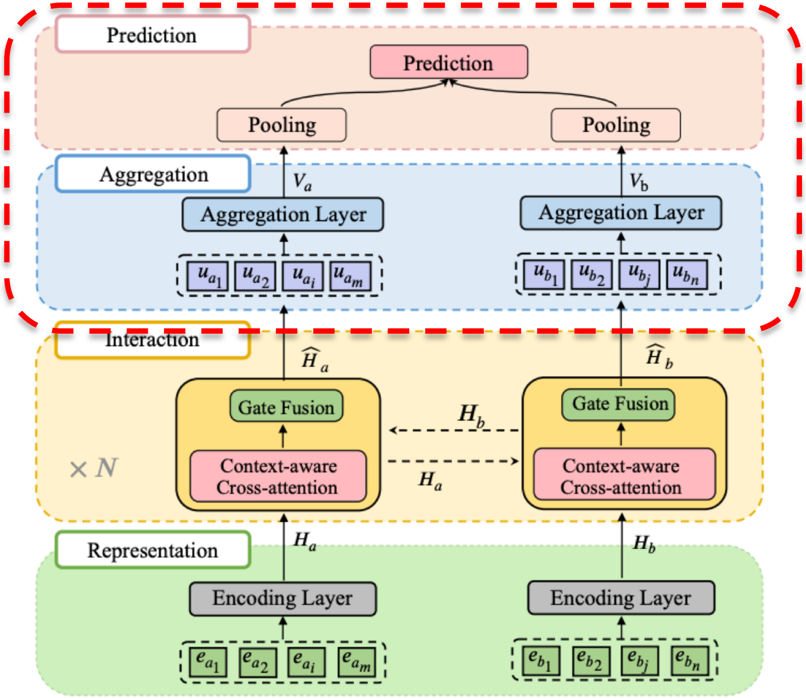
2. Gated Fusion Layer

  - *We compare the original representations and the aligned ones from difference perspectives*
  - *encourage model to better learn the semantic relationship and update the original representations*

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

Methods

COIN: COntext-aware Interaction Network



- **Aggregation & Prediction Layer**

$$\mathbf{P} = \mathrm{FFN}([\mathbf{V}'_{\mathrm{a}}; \mathbf{V}'_{\mathrm{b}}; \mathbf{V}'_{\mathrm{a}} - \mathbf{V}'_{\mathrm{b}}; \mathbf{V}'_{\mathrm{a}} \odot \mathbf{V}'_{\mathrm{b}}])$$

Experimental Setup

Datasets

- Quora Question Pairs
  - English question pairs from Quora.com
  - Use the splits from (Wang et al. 2017)
- LCQMC Corpus (Liu et al. 2018)
  - Open-domain question matching corpus from Baidu Knows
  - Use the original splits

Dataset	Train	Dev	Test	# Classes
QUORA	384K	10K	10K	2
LCQMC	239K	9K	13K	2

Evaluation Metrics

- Accuracy
- F1-score

Experiment Results

Better results than non-pretrained methods

Model	Acc (%)	F1 (%)
Lattice-CNN	82.1	82.4
ESIM (Chen et al., 2017)	82.0	84.0
BiMPM (Wang et al., 2017)	83.3	84.9
GMN (Chen et al., 2020)	84.6	86.0
COIN (Ours)	**85.6**	**86.5**
BERT (Devlin et al., 2019)	85.7	86.8
SBERT (Reimers and Gurevych, 2019)	85.4	86.6
COIN (ensemble)	**86.2**	**87.0**

Results on LCQMC

Model	Acc. (%)	Params
BiMPM (Wang et al., 2017)	88.2	1.6M
DIIN (Gong et al., 2017)	89.0	4.4M
CAFE (Tay et al., 2018)	88.7	4.7M
OSOA-DFN (Liu et al., 2019)	89.0	10.0M
RE2 (Yang et al., 2019b)	89.2	2.8M
ESAN (Hu et al., 2020)	89.3	3.9M
Enhanced-RCNN (Peng et al., 2020)	89.3	7.7M
COIN (ours)	**89.4**	6.5M
BERT (Devlin et al., 2019)	90.1	109.5M
SBERT (Reimers and Gurevych, 2019)	90.6	109.5M
COIN (ensemble)	**90.7**	32.5M

Results on Quora

Experiment Results

Comparable results with pre-trained methods

Model	Acc (%)	F1 (%)
Lattice-CNN	82.1	82.4
ESIM (Chen et al., 2017)	82.0	84.0
BiMPM (Wang et al., 2017)	83.3	84.9
GMN (Chen et al., 2020)	84.6	86.0
COIN (Ours)	**85.6**	**86.5**
BERT (Devlin et al., 2019)	85.7	86.8
SBERT (Reimers and Gurevych, 2019)	85.4	86.6
COIN (ensemble)	**86.2**	**87.0**

Results on LCQMC

Model	Acc. (%)	Params
BiMPM (Wang et al., 2017)	88.2	1.6M
DIIN (Gong et al., 2017)	89.0	4.4M
CAFE (Tay et al., 2018)	88.7	4.7M
OSOA-DFN (Liu et al., 2019)	89.0	10.0M
RE2 (Yang et al., 2019b)	89.2	2.8M
ESAN (Hu et al., 2020)	89.3	3.9M
Enhanced-RCNN (Peng et al., 2020)	89.3	7.7M
COIN (ours)	**89.4**	6.5M
BERT (Devlin et al., 2019)	90.1	109.5M
SBERT (Reimers and Gurevych, 2019)	90.6	109.5M
COIN (ensemble)	**90.7**	32.5M

Results on Quora

Experiment Results

fewer parameters than many SOTA methods

Model	Acc (%)	F1 (%)
Lattice-CNN	82.1	82.4
ESIM (Chen et al., 2017)	82.0	84.0
BiMPM (Wang et al., 2017)	83.3	84.9
GMN (Chen et al., 2020)	84.6	86.0
COIN (Ours)	**85.6**	**86.5**
BERT (Devlin et al., 2019)	85.7	86.8
SBERT (Reimers and Gurevych, 2019)	85.4	86.6
COIN (ensemble)	**86.2**	**87.0**

Results on LCQMC

Model	Acc. (%)	Params
BiMPM (Wang et al., 2017)	88.2	1.6M
DIIN (Gong et al., 2017)	89.0	4.4M
CAFE (Tay et al., 2018)	88.7	4.7M
OSOA-DFN (Liu et al., 2019)	89.0	10.0M
RE2 (Yang et al., 2019b)	89.2	2.8M
ESAN (Hu et al., 2020)	89.3	3.9M
Enhanced-RCNN (Peng et al., 2020)	89.3	7.7M
COIN (ours)	**89.4**	6.5M
BERT (Devlin et al., 2019)	90.1	109.5M
SBERT (Reimers and Gurevych, 2019)	90.6	109.5M
COIN (ensemble)	**90.7**	32.5M

Results on Quora

Experiment Results

5-run Ensemble results even outperform pre-trained methods

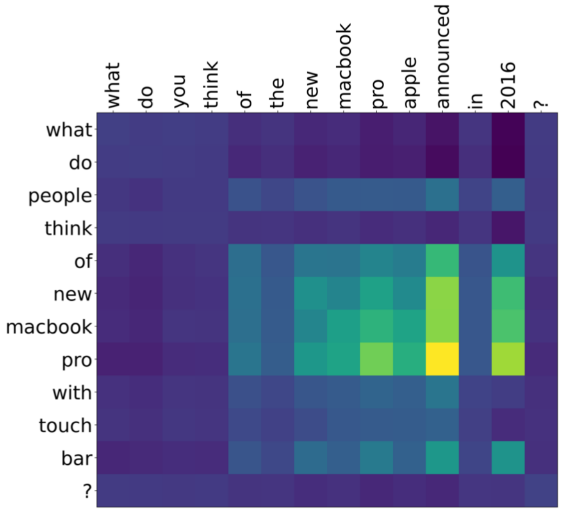
Model	Acc (%)	F1 (%)
Lattice-CNN	82.1	82.4
ESIM (Chen et al., 2017)	82.0	84.0
BiMPM (Wang et al., 2017)	83.3	84.9
GMN (Chen et al., 2020)	84.6	86.0
COIN (Ours)	**85.6**	**86.5**
BERT (Devlin et al., 2019)	85.7	86.8
SBERT (Reimers and Gurevych, 2019)	85.4	86.6
COIN (ensemble)	**86.2**	**87.0**

Results on LCQMC

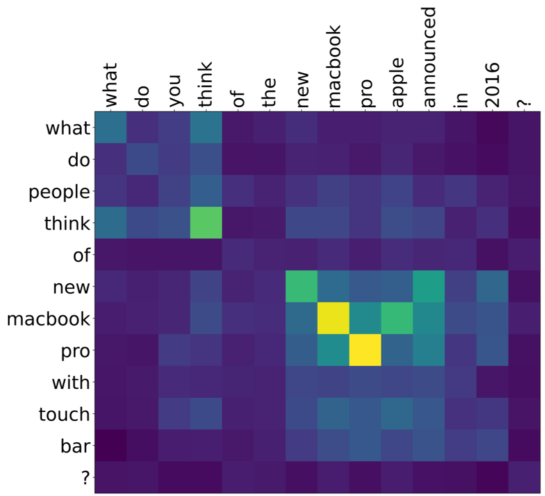
Model	Acc. (%)	Params
BiMPM (Wang et al., 2017)	88.2	1.6M
DIIN (Gong et al., 2017)	89.0	4.4M
CAFE (Tay et al., 2018)	88.7	4.7M
OSOA-DFN (Liu et al., 2019)	89.0	10.0M
RE2 (Yang et al., 2019b)	89.2	2.8M
ESAN (Hu et al., 2020)	89.3	3.9M
Enhanced-RCNN (Peng et al., 2020)	89.3	7.7M
COIN (ours)	**89.4**	6.5M
BERT (Devlin et al., 2019)	90.1	109.5M
SBERT (Reimers and Gurevych, 2019)	90.6	109.5M
COIN (ensemble)	**90.7**	32.5M

Results on Quora

Visualization of Alignments



1st alignment



3rd alignment

- Aided by the context, the model learns to **correctly align the salient words & phrases**;
- Model **refines alignments** from low level to high level;
- Structed phrase *"what do you think of"* is also connected, which is an important feature for question matching

Conclusion

- Improve cross-attention by incorporating contextual cues to better align two sequences
- Leverage a gate fusion layer to flexibly integrate the aligned features
- Achieve better results on two question matching datasets.

Thank You