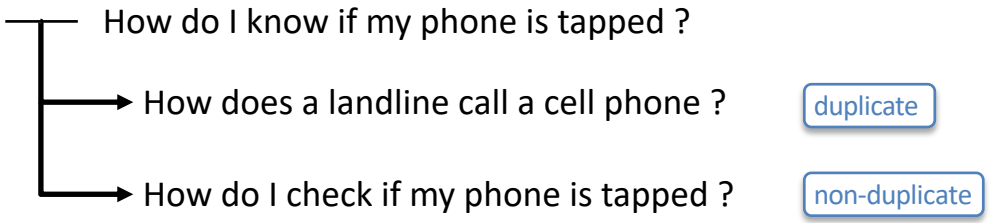
Enhanced Sentence Alignment Network for Efficient Short Text Matching

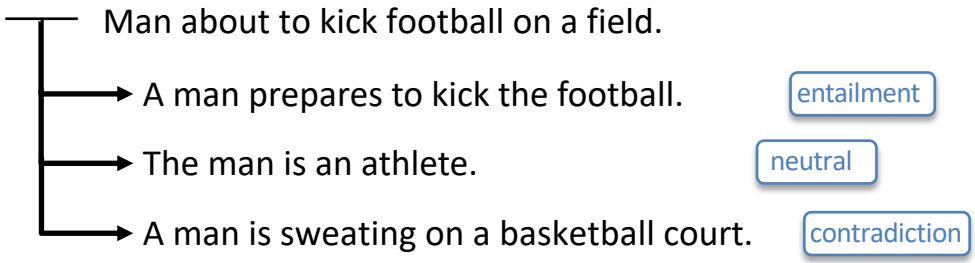
Zhe Hu[1], Zuohui Fu[2], Cheng Peng[1], and Weiwei Wang[1]
[1]Baidu Inc
[2]Rutgers University

Text Matching

- Paraphrase Detection



- Natural Language Inference



Quora Dataset: *https://data.quora.com/First-Quora-Dataset-ReleaseQuestion-Pairs*

SNLI Dataset: *Bowman, Samuel R., et al. "A large annotated corpus for learning natural language inference." arXiv preprint arXiv:1508.05326 (2015).*

Current Methods on Text Matching

- Sentence Encoding Approach

- Sentence Interaction Approach

Current Methods on Text Matching

- Sentence Encoding Approach
- Sentence Interaction Approach
- Pre-trained LM  

Current Methods on Text Matching

- Sentence Encoding Approach
- Sentence Interaction Approach
- Pre-trained LM


Motivations

- Can we design an *efficient* model that can be deployed in the real-world system with a fast inference speed?

Motivations

- Can we design an *efficient* model that can be deployed in the real-world system with a fast inference speed?

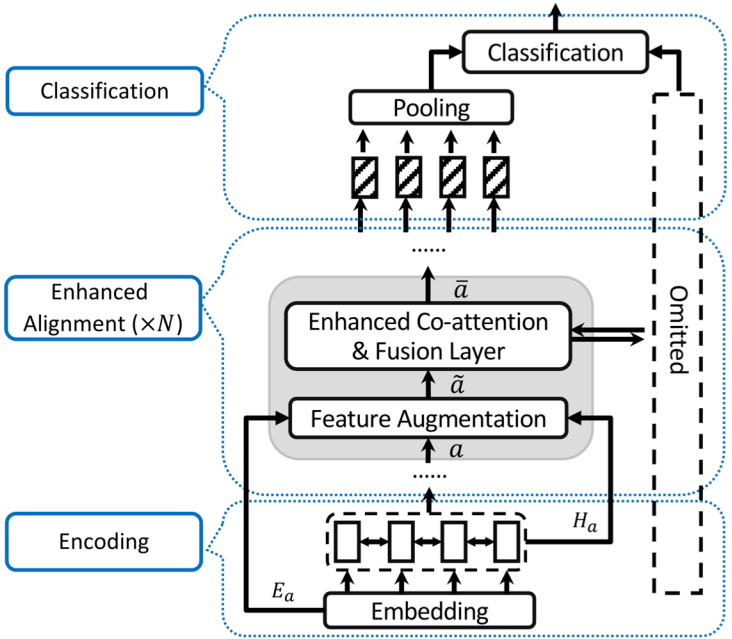
- Can we improve the existing cross-sentence attention to mitigate the *unstable matching* problem between intermediate representations [Liu et al., 2019a]?

  - Conducting cross-sentence attention between intermediate representations are uncertain and unstable because semantics are changed at different layers.

  - The intermediate representations tend to be affected by error propagation in multi-layered attentions.

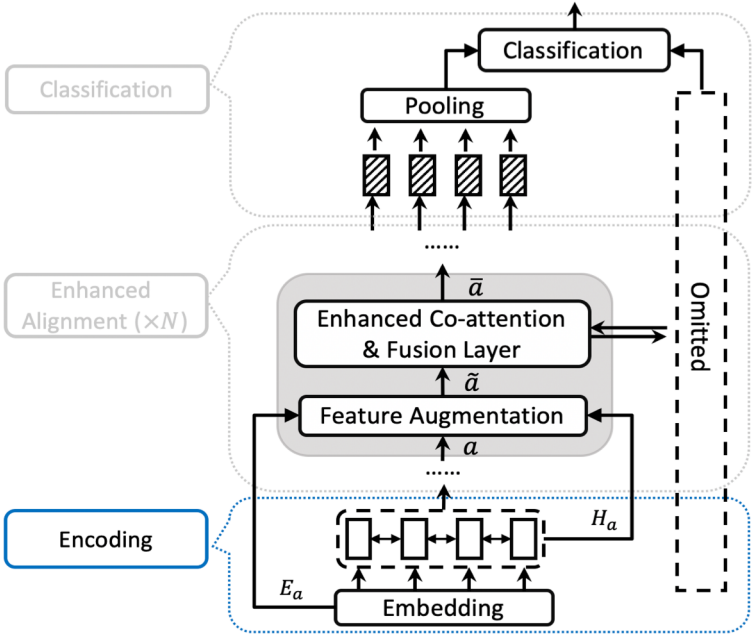
ESAN: enhanced sentence alignment network

- Embedding & Encoding layer
- Enhanced Sentence Alignment layer ($\times N$)
- Pooling and Classification layer



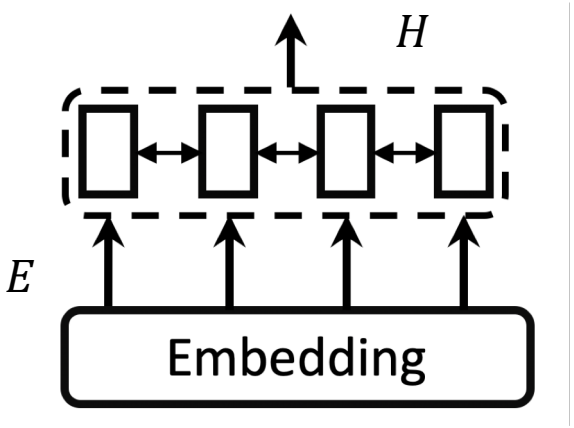
ESAN: enhanced sentence alignment network

- Embedding & Encoding layer
- Enhanced Sentence Alignment layer ($\times N$)
- Pooling and Classification layer



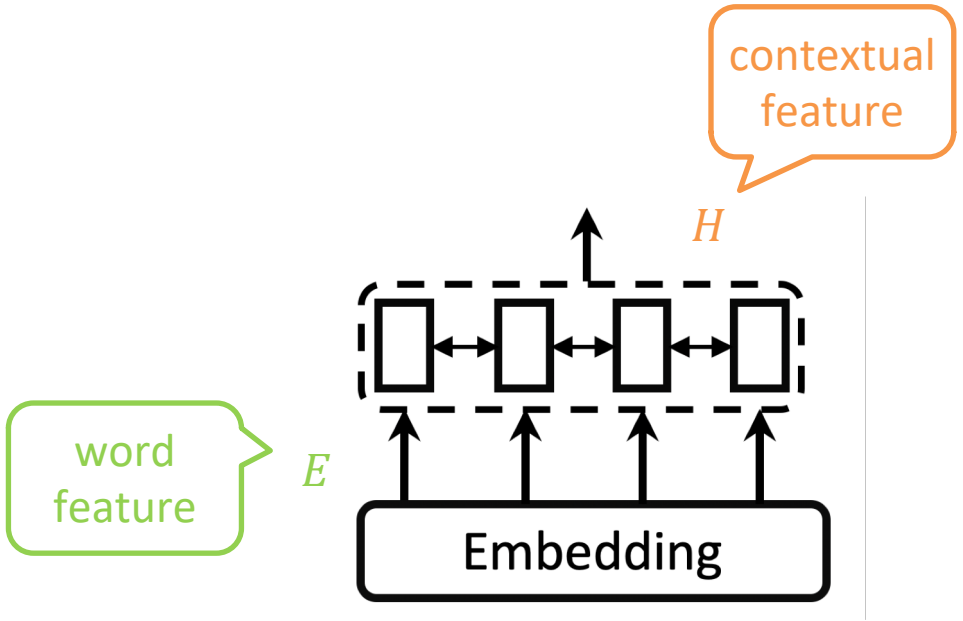
ESAN: enhanced sentence alignment network

- Embedding
  - static word embedding, fixed
  - char embedding
  - lexical features
- Encoding
  - Bi-LSTM Encoder



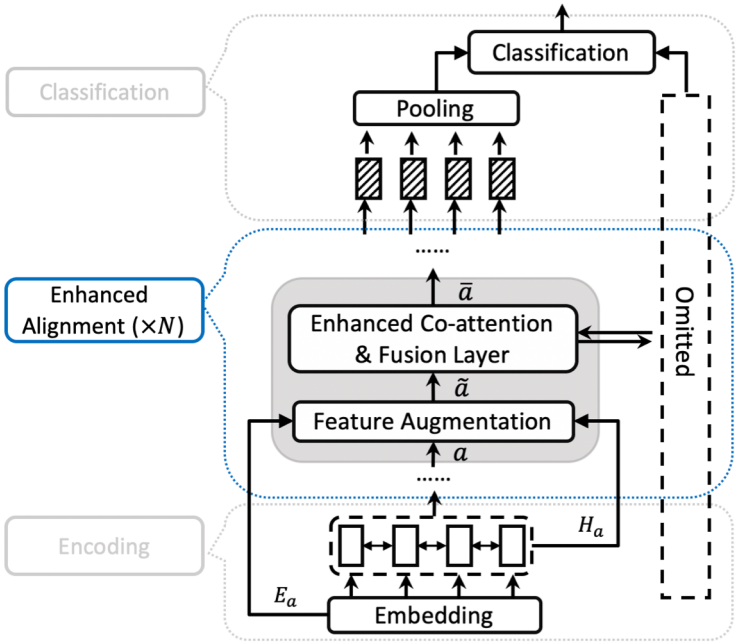
ESAN: enhanced sentence alignment network

- Embedding
  - static word embedding, fixed
  - char embedding
  - lexical features
- Encoding
  - Bi-LSTM Encoder



ESAN: enhanced sentence alignment network

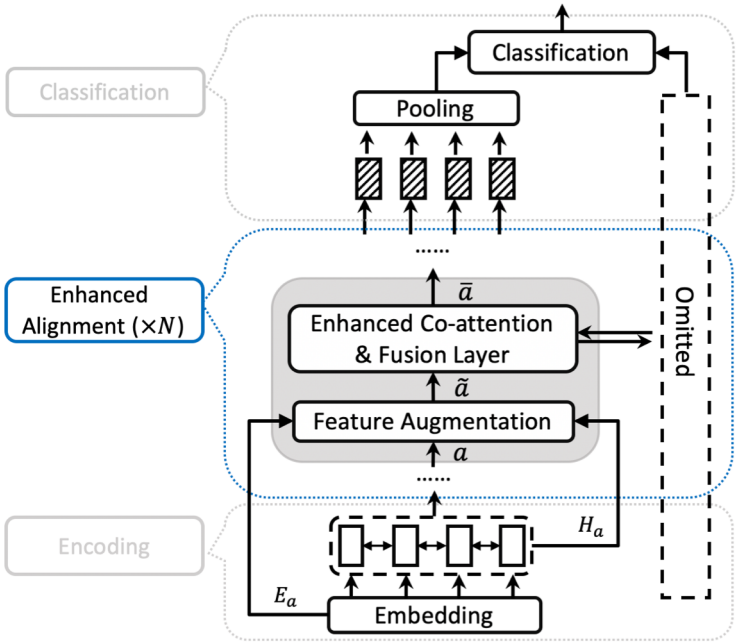
- Embedding & Encoding layer
- Enhanced Sentence Alignment layer ($\times N$)
- Pooling and Classification layer



ESAN: enhanced sentence alignment network

Enhanced Sentence Alignment layer

We integrate the word and contextual features to each *intermediate representations* to improve the cross-sentence attention and mitigate the unstable matching problem.



ESAN: enhanced sentence alignment network

1. Gated Feature Augmentation:

For the intermediate representation a = $[a_1, a_2, \ldots, a_m]$ in each cross-sentence attention

gate 1:

$$g_{e_i} = \sigma(W_g a_i + W_e e_{a_i} + b_e)$$

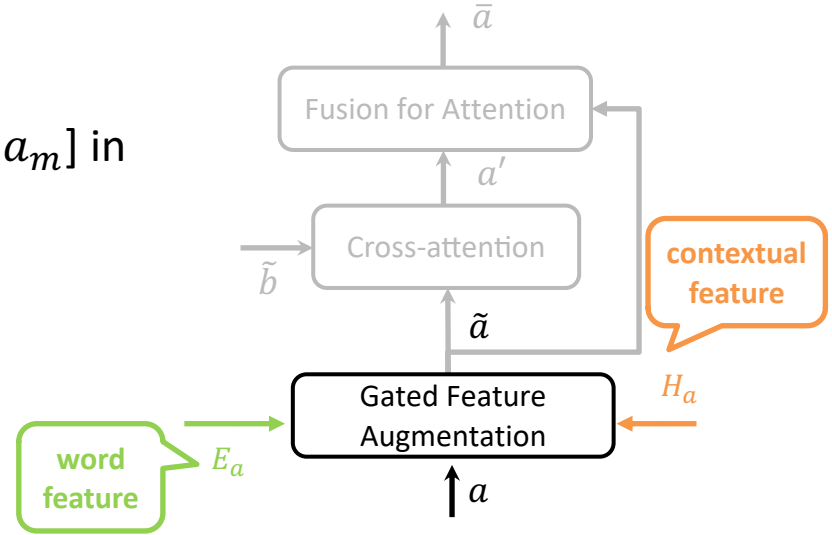
gate 2:

$$g_{h_i} = \sigma(W_g a_i + W_h h_{a_i} + b_h)$$

Output:

$$\tilde{a}_i = a_i + g_{e_i} \cdot e_i + g_{h_i} \cdot h_i$$

⇐ Enhanced Representation



ESAN: enhanced sentence alignment network

2. Enhanced Cross-sentence Attention & Fusion

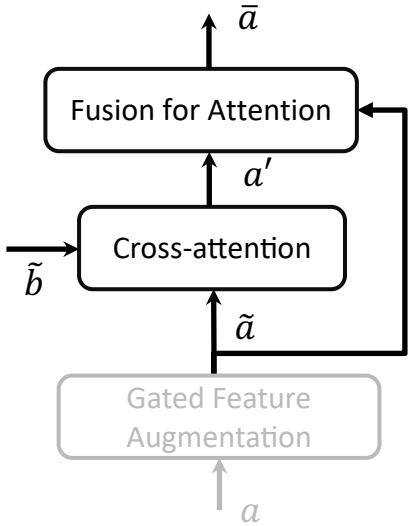
- Cross-sentence Attention

$$e_{ij} = ReLU(W_c\widetilde{a_i})^T\ ReLU(W_c\widetilde{b_j})$$

$$a' = \sum_{j=1}^{n} \frac{\exp(e_{ij})}{\sum_{k=1}^{n}\exp(e_{ik})}\widetilde{b_j}, \qquad \bar{b} = \sum_{j=1}^{m} \frac{\exp(e_{ij})}{\sum_{k=1}^{m}\exp(e_{kj})}\widetilde{a_i}$$

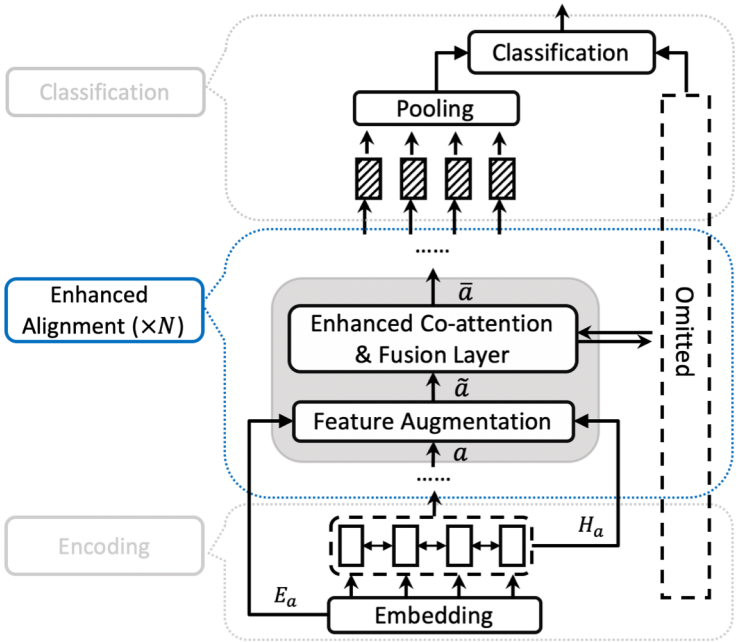
- Fusion for Attention

$$\bar{a} = F([\widetilde{a_i}; a'; \widetilde{a_i} - a'; \widetilde{a_i} \cdot a'])$$



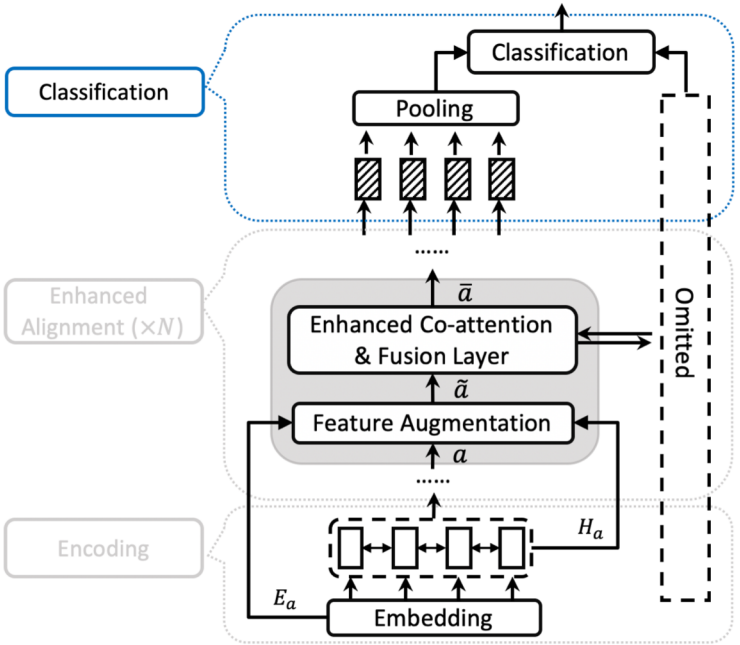
ESAN: enhanced sentence alignment network

Multiple alignment layers are stacked



ESAN: enhanced sentence alignment network

- Embedding & Encoding layer
- Enhanced Sentence Alignment layer ($\times N$)
- Pooling and Classification layer



Experiments: datasets

- ❑ Paraphrase Identification
  - Quora Question Pairs

- ❑ Natural Language Inference
  - SNLI
  - MultiNLI

Experiment: results & analysis

- Experiment Results

Model	Quora (acc)	SNLI (acc)	MNLI-m (acc)	MNLI-mm (acc)
DIIN (Gong et al., 2017)	89.1	88.0	78.8	77.8
MwAN (Tan et al., 2018)	88.2	86.9	78.5	77.7
CAFE (Tay et al., 2018)	88.5	88.7	78.7	77.9
ADIN (Liang et al., 2019)	-	88.8	78.8	77.9
Ours	**89.3**	**89.0**	**79.3**	**78.4**

Experiment: results & analysis

- Experiment Results

Model	Quora (acc)	SNLI (acc)	MNLI-m (acc)	MNLI-mm (acc)
DIIN (Gong et al., 2017)	89.1	88.0	78.8	77.8
MwAN (Tan et al., 2018)	88.2	86.9	78.5	77.7
CAFE (Tay et al., 2018)	88.5	88.7	78.7	77.9
ADIN (Liang et al., 2019)	-	88.8	78.8	77.9
Ours	**89.3**	**89.0**	**79.3**	**78.4**
BERT (Devlin et al., 2019)	**90.1**	**90.8**	**84.6**	**83.4**

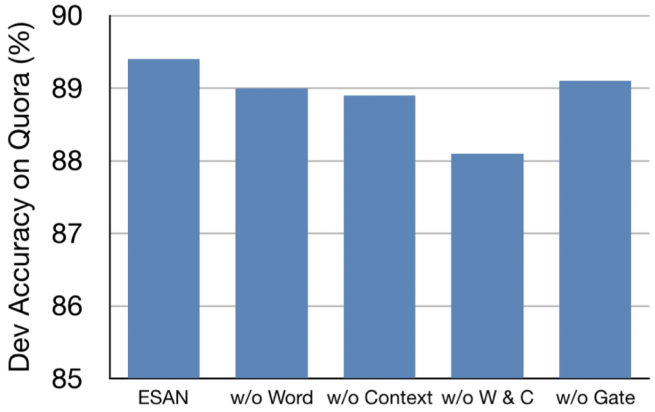
Experiment: results & analysis

- Model Efficiency

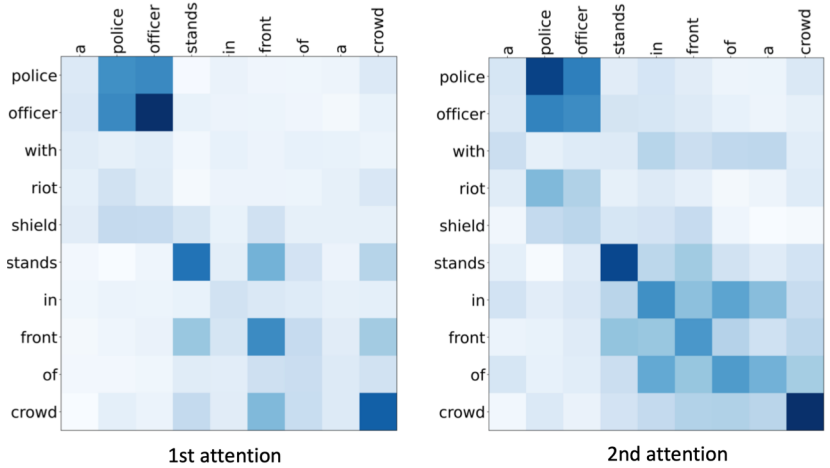
Model on Quora	# Params	CPU Inference Time (s/batch) *
BERT	109.5 M	0.88 ± 0.06
Ours	3.9 M	0.04 ± 0.01

Experiment: results & analysis

- Ablation Results



- Attention Visualization



THANKS