DiSAN: Directional Self-Attention Network for RNN/CNN-free Language Understanding

Xiao i - Chen Lu
School of Computer and Software Engineering - ICA

DiSAN: Directional Self-Attention Network for RNN/CNN-free Language Understanding

Xiao i - Chen Lu
School of Computer and Software Engineering - ICA

DiSAN: Directional Self-Attention Network for RNN/CNN-free Language Understanding

Xiao i - Chen Lu
School of Computer and Software Engineering - ICA

DiSAN: Directional Self-Attention Network for RNN/CNN-Free Language Understanding

Tao Shen† **Tianyi Zhou‡** **Guodong Long†**
Jing Jiang† **Shirui Pan†** **Chengqi Zhang†**

†Centre of Artificial Intelligence, FEIT, University of Technology Sydney
‡Paul G. Allen School of Computer Science & Engineering, University of Washington
tao.shen@student.uts.edu.au, tianyizh@uw.edu
{guodong.long, jing.jiang, shirui.pan, chengqi.zhang}@uts.edu.au

Outline

- Background
- Two Proposed Attention Mechanisms
- Directional Self-Attention Network
- Experiments
- Conclusions

Outline

- Background
- Two Proposed Attention Mechanisms
- Directional Self-Attention Network
- Experiments
- Conclusions

Background

- Task: Language understanding

  Natural language understanding (NLU) is a branch of artificial intelligence (AI) that uses computer software to understand input made in the form of sentences in text or speech format.

Background

- Context dependency

  e.g.

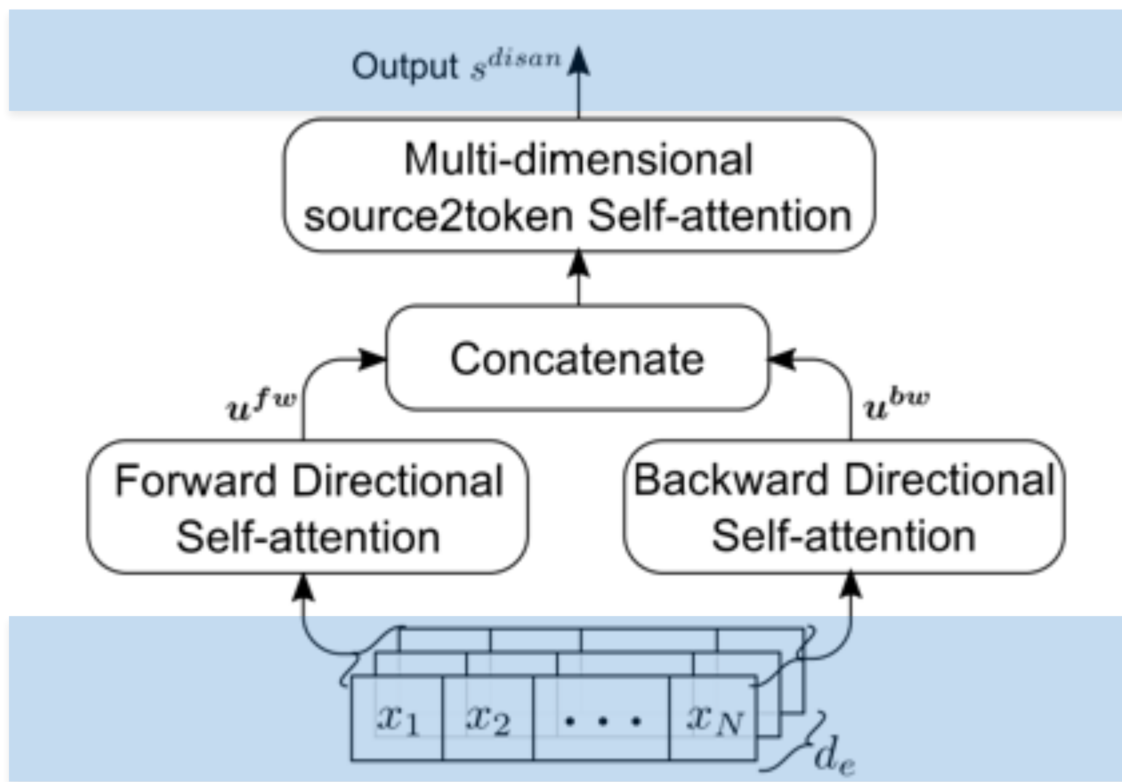
  - RNN with sequential architecture: Capturing long-range dependencies

  - CNN with hierarchical architecture: Capturing local or position-invariant dependencies

Background

- Sentence encoding

  e.g.



Background

- Attention

  The attention is proposed to compute an alignment score between elements from two sources. That is, large score means one contributes important information to another.

$$a = [f(x_i, q)]_{i=1}^{n}$$

$$p(z|\boldsymbol{x}, q) = \mathrm{softmax}(a)$$

specifically, $$p(z = i|\boldsymbol{x}, q) = \frac{\exp(f(x_i, q))}{\sum_{i=1}^{n} \exp(f(x_i, q))}$$

$$s = \sum_{i=1}^{n} p(z = i|\boldsymbol{x}, q) x_i = \mathbb{E}_{i \sim p(z|\boldsymbol{x}, q)}(x_i)$$

Background

- Additive attention & Multiplicative attention

$$f(x_i, q) = w^T \sigma(W^{(1)} x_i + W^{(2)} q)$$

$$f(x_i, q) = \left\langle W^{(1)} x_i, W^{(2)} q \right\rangle$$

Background

- Self-Attention

  It is a special case of the attention mechanism introduced above. It replaces q with a token embedding x_j from the source input itself.

$$f(x_i, q) = W^T \sigma \left(W^{(1)} x_i + W^{(2)} q + b^{(1)} \right) + b$$

$$f(x_i, x_j) = W^T \sigma \left(W^{(1)} x_i + W^{(2)} x_j + b^{(1)} \right) + b$$

Outline

- Background
- Two Proposed Attention Mechanisms
- Directional Self-Attention Network
- Experiments
- Conclusions

Two Proposed Attention Mechanisms

- Multi-dimensional Attention
- Two Types of Multi-dimensional Self-attention
- Directional Self-Attention

Two Proposed Attention Mechanisms

- Multi-dimensional Attention
- Two Types of Multi-dimensional Self-attention
- Directional Self-Attention

Two Proposed Attention Mechanisms

- Multi-dimensional Attention

It is a natural extension of additive attention at feature level. Multi-dimensional attention computes a feature-wise score vector for x_i.

$$f(x_i, q) = \underline{w}^T \sigma(W^{(1)} x_i + W^{(2)} q)$$

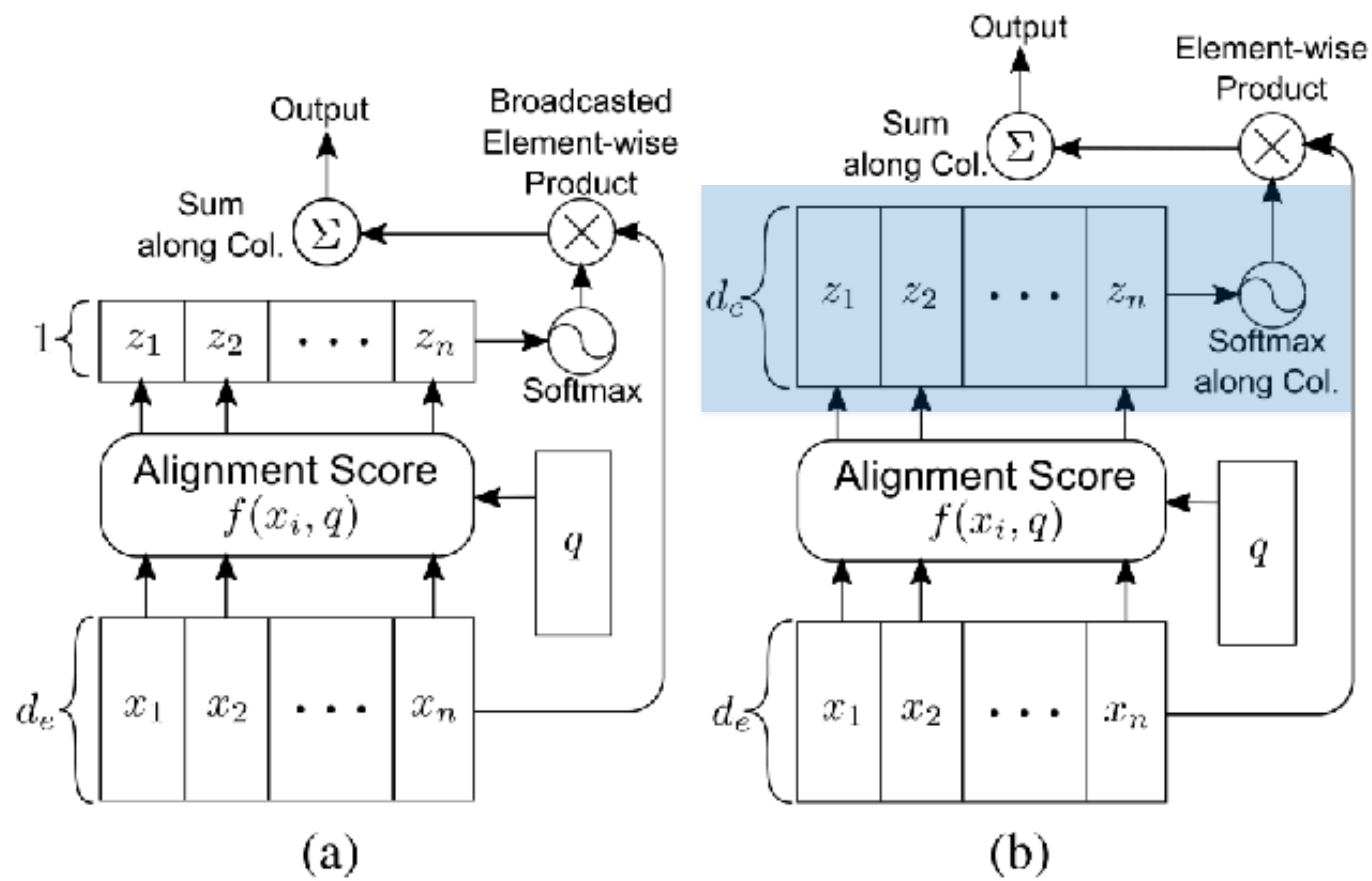
$$f(x_i, q) = \underline{W}^T \sigma\left(W^{(1)} x_i + W^{(2)} q\right)$$

$$f(x_i, q) = W^T \sigma\left(W^{(1)} x_i + W^{(2)} q + b^{(1)}\right) + b.$$

$$s = \left[\sum\nolimits_{i=1}^{n} P_{ki} \boldsymbol{x}_{ki}\right]_{k=1}^{d_e} = \left[\mathbb{E}_{i \sim p(z_k | \boldsymbol{x}, q)}(\boldsymbol{x}_{ki})\right]_{k=1}^{d_e}$$

Two Proposed Attention Mechanisms

- Multi-dimensional Attention



Two Proposed Attention Mechanisms

- Remark: Multi-dimensional Attention

  - The word embedding usually suffers from the polysemy in natural language. Since traditional attention cannot distinguish the meaning of the same word in different contexts.

  - Multi-dimensional attention computes a score for each feature of each word, so it can select the features that can best describe the word specific meaning in any given context.

Two Proposed Attention Mechanisms

- Multi-dimensional Attention
- Two Types of Multi-dimensional Self-attention
- Directional Self-Attention

Two Proposed Attention Mechanisms

- Two Types of Multi-dimensional Self-attention

  - token2token

$$f(x_i, x_j) = W^T \sigma \left(W^{(1)} x_i + W^{(2)} x_j + b^{(1)} \right) + b$$

$$s_j = \sum_{i=1}^{n} P_{\cdot i}^{j} \odot x_i$$

  - source2token

$$f(x_i) = W^T \sigma \left(W^{(1)} x_i + b^{(1)} \right) + b$$

$$s = \sum_{i=1}^{n} P_{\cdot i} \odot x_i$$

Two Proposed Attention Mechanisms

- Multi-dimensional Attention
- Two Types of Multi-dimensional Self-attention
- Directional Self-Attention

Two Proposed Attention Mechanisms

- Directional Self-Attention

  - A “masked” multi-dimensional token2token self-attention block to explore the dependency and temporal order, and a fusion gate to combine the output and input of attention block.

Two Proposed Attention Mechanisms

- Directional Self-Attention

  - A “masked” multi-dimensional token2token self-attention block to explore the dependency and temporal order.

$$f(h_i, h_j) =$$
$$c \cdot \tanh\left([W^{(1)} h_i + W^{(2)} h_j + b^{(1)}]/c\right) + M_{ij}\mathbf{1}$$

$$M_{ij}^{diag} = \begin{cases} 0, & i \neq j \\ -\infty, & i = j \end{cases}$$

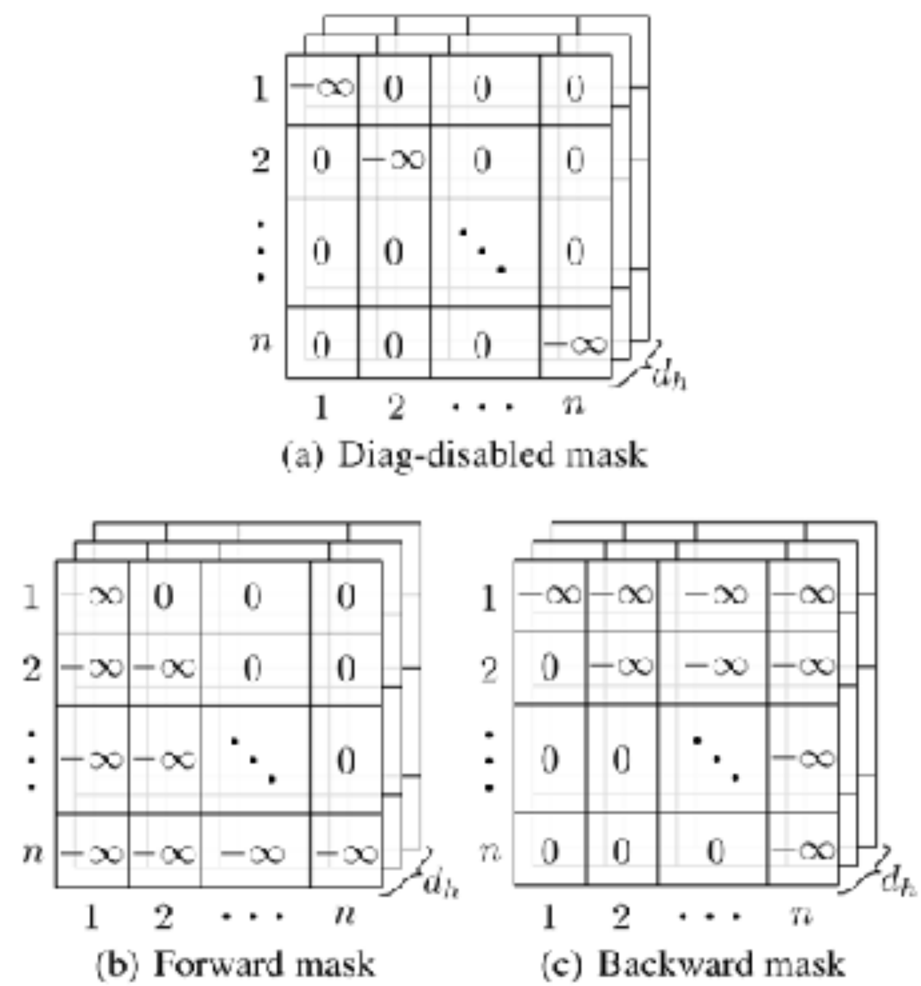
$$M_{ij}^{fw} = \begin{cases} 0, & i < j \\ -\infty, & otherwise \end{cases}$$

$$M_{ij}^{bw} = \begin{cases} 0, & i > j \\ -\infty, & otherwise \end{cases}$$

Two Proposed Attention Mechanisms

- Directional Self-Attention

  - A “masked” multi-dimensional token2token self-attention block to explore the dependency and temporal order.



(a) Diag-disabled mask

(b) Forward mask

(c) Backward mask

Two Proposed Attention Mechanisms

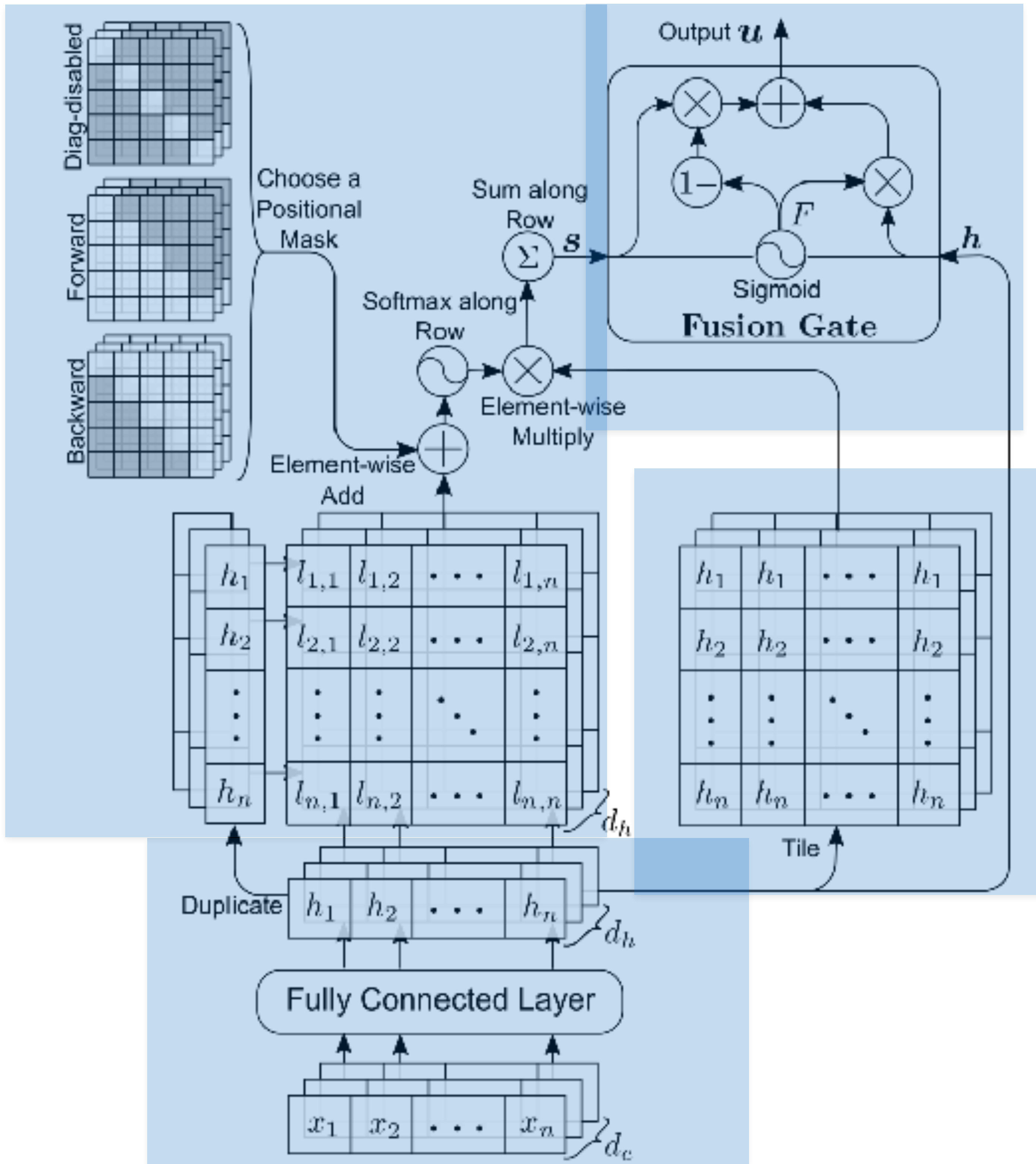
- Directional Self-Attention

  - A fusion gate to combine the output and input of attention block.

$$F = \text{sigmoid}\left(W^{(f1)}\boldsymbol{s} + W^{(f2)}\boldsymbol{h} + b^{(f)}\right)$$

$$\boldsymbol{u} = F \odot \boldsymbol{h} + (1 - F) \odot \boldsymbol{s}$$

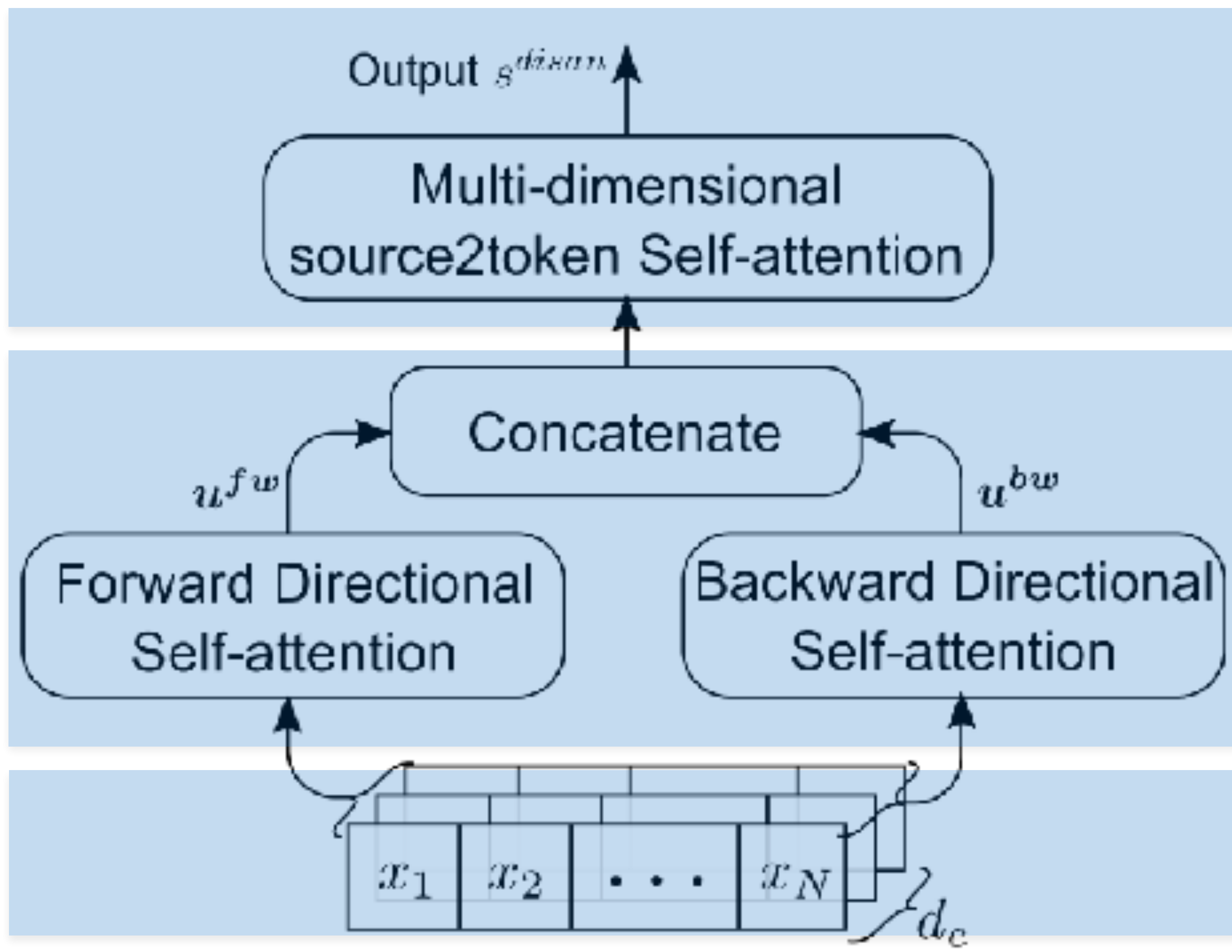
Directional Self-Attention



Outline

- Background
- Two Proposed Attention Mechanisms
- Directional Self-Attention Network
- Experiments
- Conclusions

Directional Self-Attention Network



Directional Self-Attention Network

- Remark: DiSAN

  - Forward/backward DiSA blocks work as context fusion layers.

  - Multi-dimensional source2token self-attention compresses the sequence into a single vector.

Outline

- Background
- Two Proposed Attention Mechanisms
- Directional Self-Attention Network
- Experiments
- Conclusions

Experiments

- Natural Language Inference

Model Name	$\|\theta\|$	T(s)/epoch	Train Accu(%)	Test Accu(%)
Unlexicalized features (Bowman et al. 2015)			49.4	50.4
+ Unigram and bigram features (Bowman et al. 2015)			99.7	78.2
100D LSTM encoders (Bowman et al. 2015)	0.2m		84.8	77.6
300D LSTM encoders (Bowman et al. 2016)	3.0m		83.9	80.6
1024D GRU encoders (Vendrov et al. 2016)	15m		98.8	81.4
300D Tree-based CNN encoders (Mou et al. 2016)	3.5m		83.3	82.1
300D SPINN-PI encoders (Bowman et al. 2016)	3.7m		89.2	83.2
600D Bi-LSTM encoders (Liu et al. 2016)	2.0m		86.4	83.3
300D NTI-SLSTM-LSTM encoders (Munkhdalai and Yu 2017b)	4.0m		82.5	83.4
600D Bi-LSTM encoders+intra-attention (Liu et al. 2016)	2.8m		84.5	84.2
300D NSE encoders (Munkhdalai and Yu 2017a)	3.0m		86.2	84.6
Word Embedding with additive attention	0.45m	216	82.39	79.81
Word Embedding with s2t self-attention	0.54m	261	86.22	83.12
Multi-head with s2t self-attention	1.98m	345	89.58	84.17
Bi-LSTM with s2t self-attention	2.88m	2080	90.39	84.98
DiSAN without directions	2.35m	592	90.18	84.66
Directional self-attention network (DiSAN)	2.35m	587	91.08	**85.62**

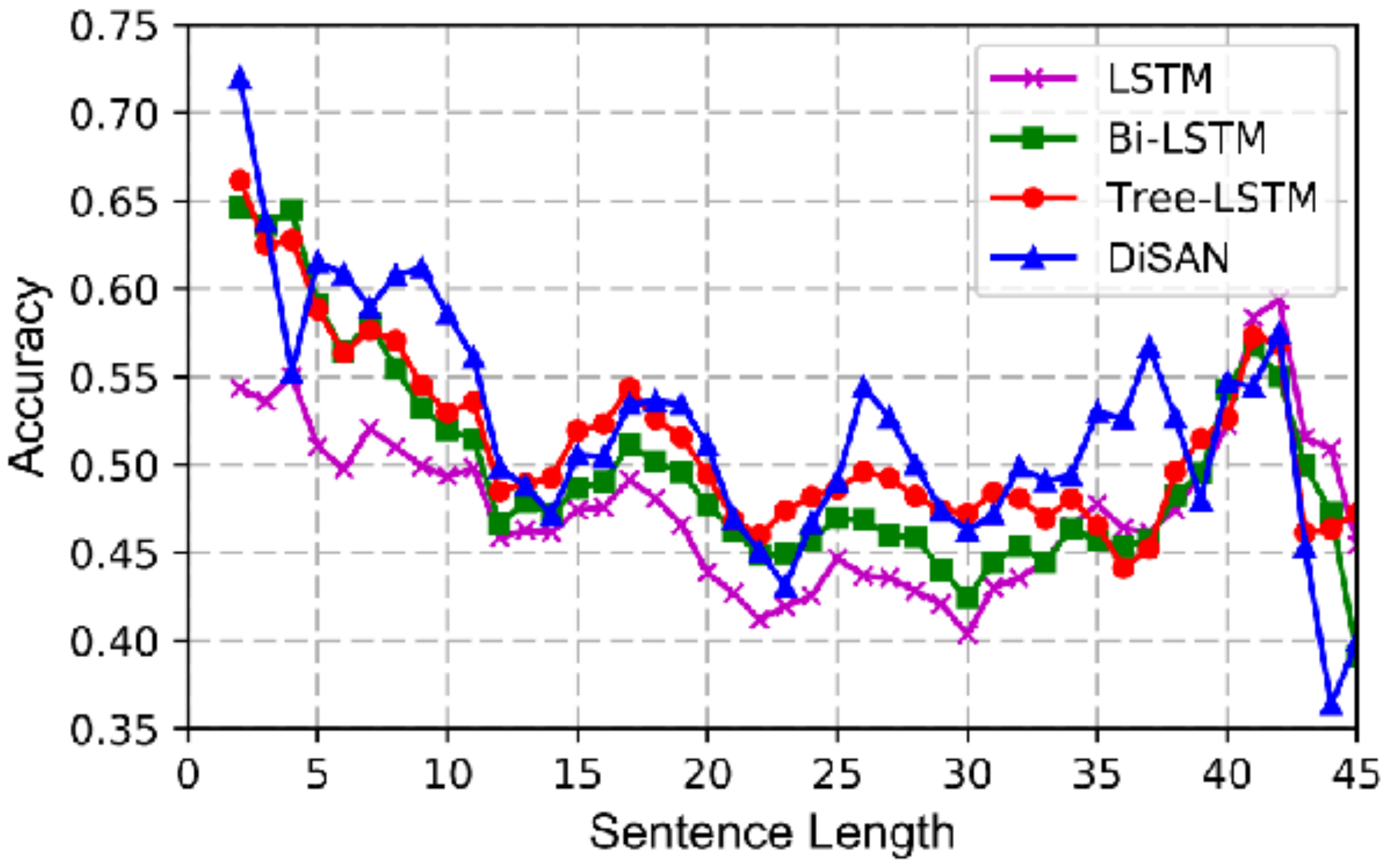
Experiments

- Sentiment Analysis

Model	Test Accu
MV-RNN (Socher et al. 2013)	44.4
RNTN (Socher et al. 2013)	45.7
Bi-LSTM (Li et al. 2015)	49.8
Tree-LSTM (Tai, Socher, and Manning 2015)	51.0
CNN-non-static (Kim 2014)	48.0
CNN-Tensor (Lei, Barzilay, and Jaakkola 2015)	51.2
NCSL (Teng, Vo, and Zhang 2016)	51.1
LR-Bi-LSTM (Qian, Huang, and Zhu 2017)	50.6
Word Embedding with additive attention	47.47
Word Embedding with s2t self-attention	48.87
Multi-head with s2t self-attention	49.14
Bi-LSTM with s2t self-attention	49.95
DiSAN without directions	49.41
DiSAN	**51.72**

Experiments

- Sentiment Analysis



Outline

- Background
- Two Proposed Attention Mechanisms
- Directional Self-Attention Network
- Experiments
- Conclusions

Conclusions

- Multi-dimensional attention
- Directional self-attention
- RNN/CNN-free language understanding network
- Fewer parameters and higher time efficiency
- One more paper: <Attention is all you need>

Thanks

Xiao i - Chen Lu
School of Computer and Software Engineering - ICA