Distance-based Self-Attention Network for Natural Language Inference

Jinbae Im and **Sungzoon Cho**
Department of Industrial Engineering
Seoul National University
Seoul, South Korea
jinbae@dm.snu.ac.kr, zoon@snu.ac.kr

*on arXiv 2017.12.06

Motivation

- Transformer: fully attention-based seq2seq model.
- Self-Attention Network (Shen et al. 2017): fully attention-based sentence encoder for Natural Language Inference, reflecting directional information (directional mask).
- Intuition: positional information of words includes direction and distance.
- This paper: Distance-based Self-Attention Network, introducing a distance mask.

Shen et al.: Disan: Directional self-attention network for rnn/cnn free language understanding. –AAAI'17

Natural Language Inference

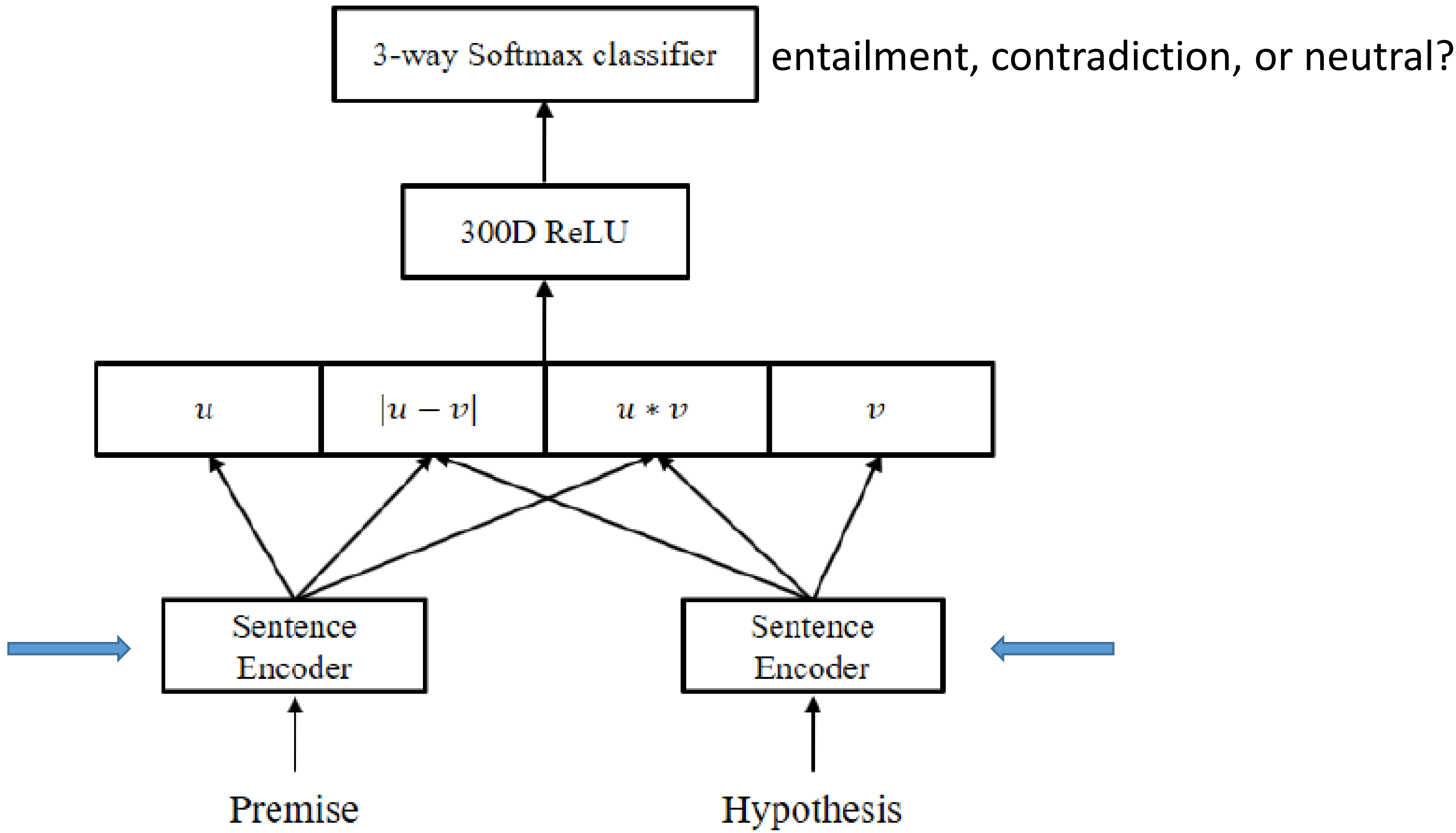


Figure 1: **Overall architecture**

Sentence Encoder

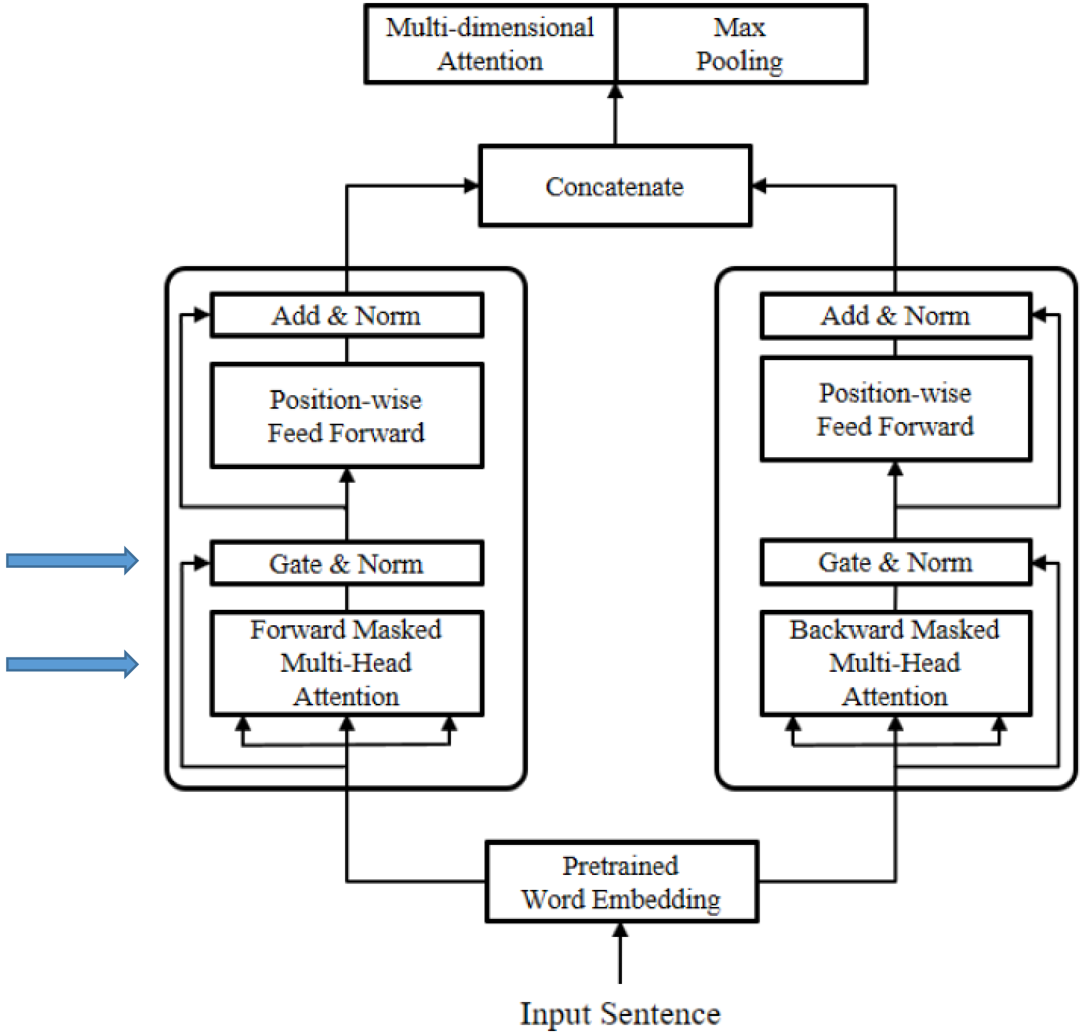


Figure 2: **Sentence encoder**

Masked Multi-Head Attention

- Transformer:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (6)$$

- This work:

$$\begin{aligned} &\text{Masked}(Q, K, V) \\ &= \text{softmax}(\frac{QK^T}{\sqrt{d_k}} + M_{dir} + \alpha M_{dis})V \end{aligned} \quad (7)$$

Concentrate on the local words around the reference word.

	1	2	···	n
1	-∞	-∞	-∞	-∞
2	0	-∞	-∞	-∞
⋮	0	0	⋱	-∞
n	0	0	0	-∞

(a) Forward mask

	1	2	···	n
1	-∞	0	0	0
2	-∞	-∞	0	0
⋮	-∞	-∞	⋱	0
n	-∞	-∞	-∞	-∞

(b) Backward mask

Figure 3: **Directional mask**

	1	2	···	n
1	0	-1	⋱	-(n-1)
2	-1	0	⋱	-(n-2)
⋮	⋱	⋱	⋱	⋱
n	-(n-1)	-(n-2)	⋱	0

Figure 4: **Distance mask**

Positional Information

- Transformer: positional encoding, considering absolute position of the words.
- This paper: distance mask, considering relative position of words, which is more important in sentence modeling.

Fusion Gate

$$S = \begin{bmatrix} - & w_1 & - \\ - & w_2 & - \\ & \vdots & \\ - & w_n & - \end{bmatrix} H = \begin{bmatrix} - & h_1 & - \\ - & h_2 & - \\ & \vdots & \\ - & h_n & - \end{bmatrix} \quad (10)$$

Embedding

Multi-head attention

First, we generate S^F, H^F by projecting S, H using $W^S, W^H \in R^{d_e \times d_e}$. Mathematically:

$$\begin{aligned} S^F &= SW^S \\ H^F &= HW^H \end{aligned} \quad (11)$$

Then create gate F as shown in equation 12 where $b^F \in R^{d_e}$.

$$\begin{aligned} \mathrm{Gate}(S, H) &= F \odot S^F + (1 - F) \odot H^F \\ \text{where } F &= \mathrm{sigmoid}(S^F + H^F + b^F) \end{aligned} \quad (12)$$

SNLI Results

Model Name	$\|\theta\|$	T(s)/epoch	Train Acc(%)	Test Acc(%)
Feature-based models				
Unlexicalized features (Bowman et al., 2015)			49.4	50.4
+Unigram and bigram features (Bowman et al., 2015)			99.7	78.2
Sentence encoding-based models				
100D LSTM encoders (Bowman et al., 2015)	220k		84.8	77.6
300D LSTM encoders (Bowman et al., 2016)	3.0m		83.9	80.6
1024D GRU encoders (Vendrov et al., 2015)	15m		98.8	81.4
300D Tree-based CNN encoders (Mou et al., 2015)	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, 2016b)	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, 2016a)	3.0m		86.2	84.6
600D Deep Gated Attn. BiLSTM encoders (Chen et al., 2017)	11.6m		90.5	85.5
600D Directional Self-Attention Network (Shen et al., 2017)	2.4m	587	91.1	85.6
Our self-attention network (without distance mask)	4.7m	687	88.1	86.0
Our Distance-based Self-Attention Network	4.7m	693	89.6	**86.3**

Table 1: **Experimental results of different models on SNLI data.** $|\theta|$: number of parameters (excluding word embedding part). T(s)/epoch : average training time (second) per epoch.

MultiNLI Results

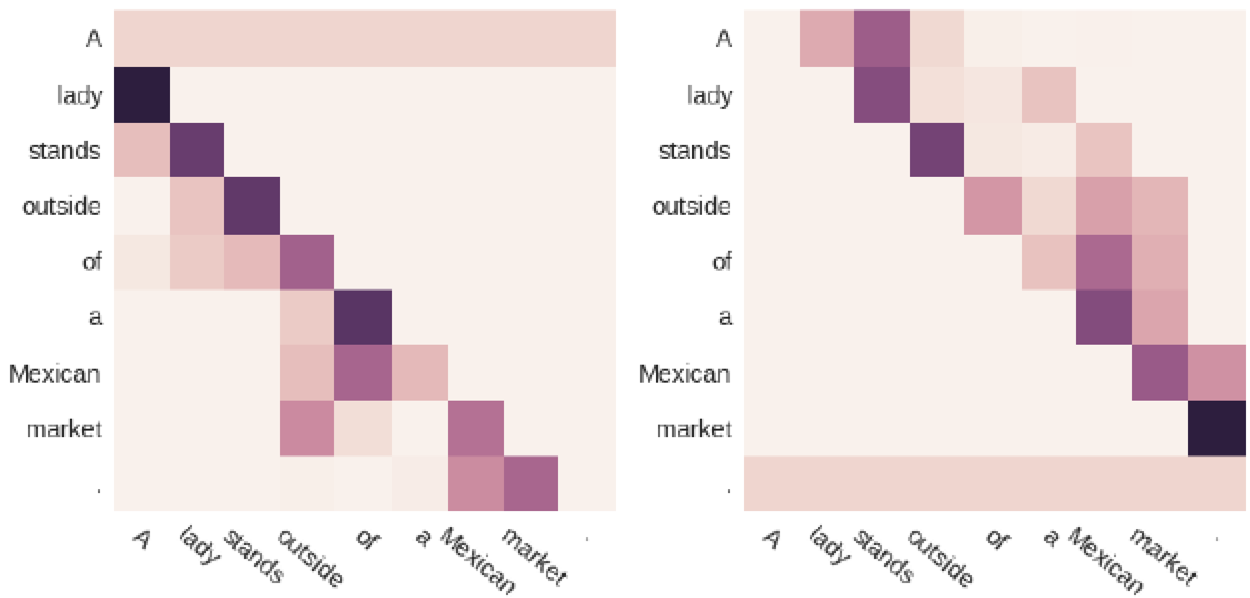
- Longer sentences.

Model Name	SNLI Mix	$\|\theta\|$	Matched Test Acc(%)	Mismatched Test Acc(%)
Baseline				
CBOW (Williams et al., 2017)	O		66.2	64.6
BiLSTM (Williams et al., 2017)	O		67.5	67.1
RepEval 2017 (Nangia et al., 2017)				
Cha-level Intra-attention BiLSTM encoders (Yang et al., 2017)	O		67.9	68.2
BiLSTM + enhanced embedding + max pooling (Vu et al., 2017)	X		70.7	70.8
BiLSTM + Inner-attention (Balazs et al., 2017)	O		72.1	72.1
Deep Gated Attn. BiLSTM encoders (Chen et al., 2017)	X	11.6m	73.5	**73.6**
Shortcut-Stacked BiLSTM encoders (Ni and Bansal, 2017)	O	140.2m	**74.5**	73.5
Fully attention-based models				
Directional Self-Attention Network (Shen et al., 2017)	X	2.4m	71.0	71.4
Our Distance-based Self-Attention Network	X	4.7m	74.1	72.9

Table 2: **Experimental results of different models on MultiNLI data.** SNLI Mix : use of SNLI training dataset. $|\theta|$: number of parameters (excluding word embedding part).

Case Study: Self-Attention

- Sentence: a lady stands outside of a Mexican market.



(a) Forward (b) Backward

Figure 7: **Masked multi-head average attention weights**

Effect of Distance Mask

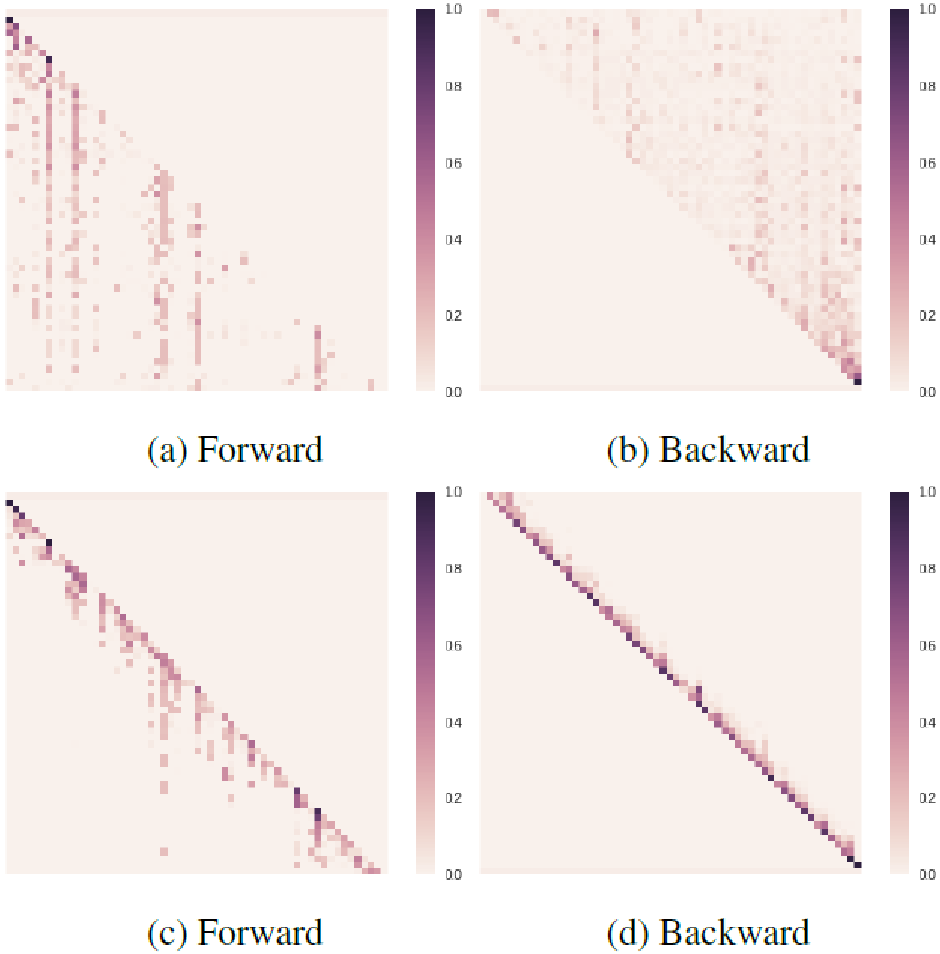
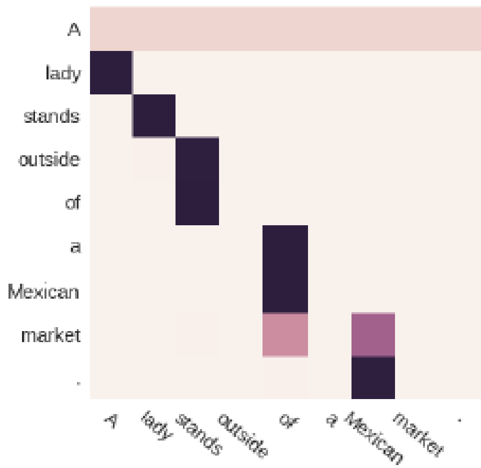


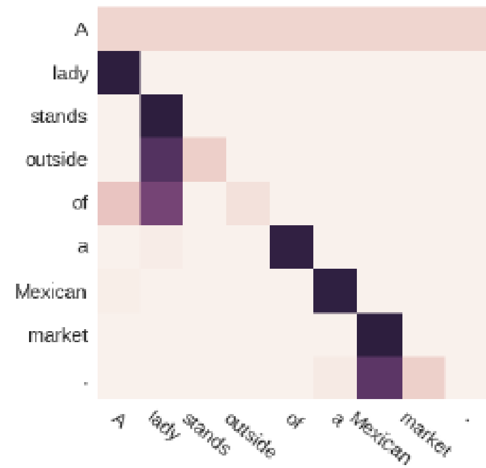
Figure 8: **Masked multi-head average attention weights : without/with distance mask.** (a), (b) : without distance mask. (c), (d) : with distance mask

Neighboring words are seen more intensively, capturing local dependency.

Effect of Multi-Head Attention

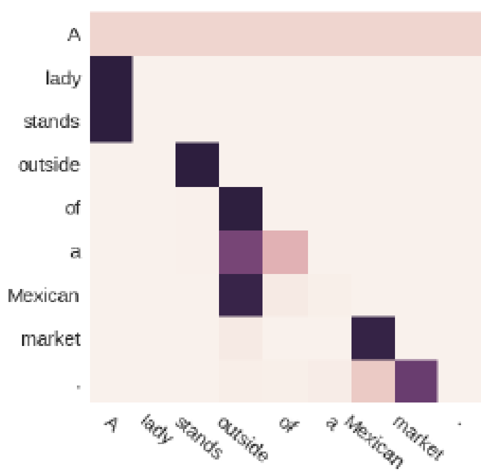


(a) head0

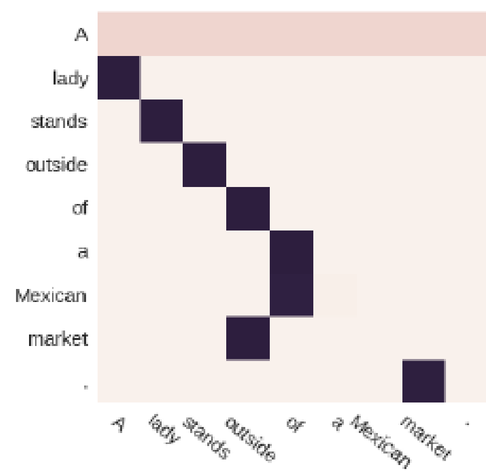


(b) head1

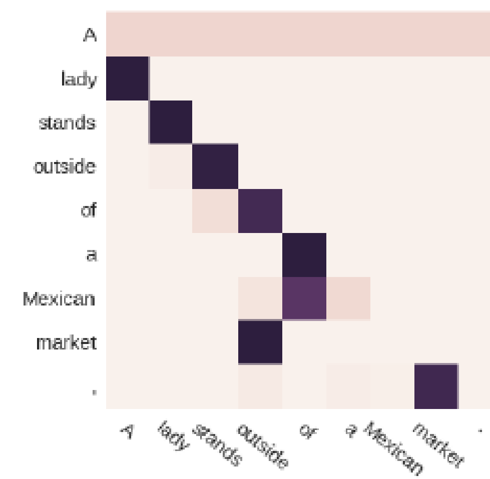
Attention weights are different for each head.



(c) head2



(d) head3



(e) head4