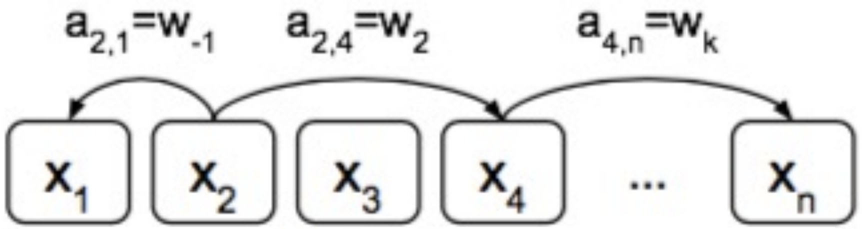
Improvements on Self-attention

Baosong Yang

Contents

Relative Position Representations

- Relative positions:



$$a_{ij}^K = w_{\mathrm{clip}(j-i,k)}^K$$

$$a_{ij}^V = w_{\mathrm{clip}(j-i,k)}^V$$

$$\mathrm{clip}(x,k) = \max(-k, \min(k,x))$$

- Incorporating into self-attention:
  - K: $$e_{ij} = \frac{1}{\sqrt{d_z}} x_i W^Q (x_j W^K + a_{ij}^K)^T$$
  - V: $$z_i = \sum_{j=1}^{n} \alpha_{ij}(x_j W^V + a_{ij}^V)$$

Experiments

- Evaluation on WMT14 En-De (4.5M) and En-Fr (36M)

Model	Position Information	EN-DE BLEU	EN-FR BLEU
Transformer (base)	Absolute Position Representations	26.5	38.2
Transformer (base)	Relative Position Representations	**26.8**	**38.7**
Transformer (big)	Absolute Position Representations	27.9	41.2
Transformer (big)	Relative Position Representations	**29.2**	**41.5**

- Cliping distance k:

k	EN-DE BLEU
0	12.5
1	25.5
2	25.8
4	25.9
16	25.8
64	25.9
256	25.8

- Keys and Values

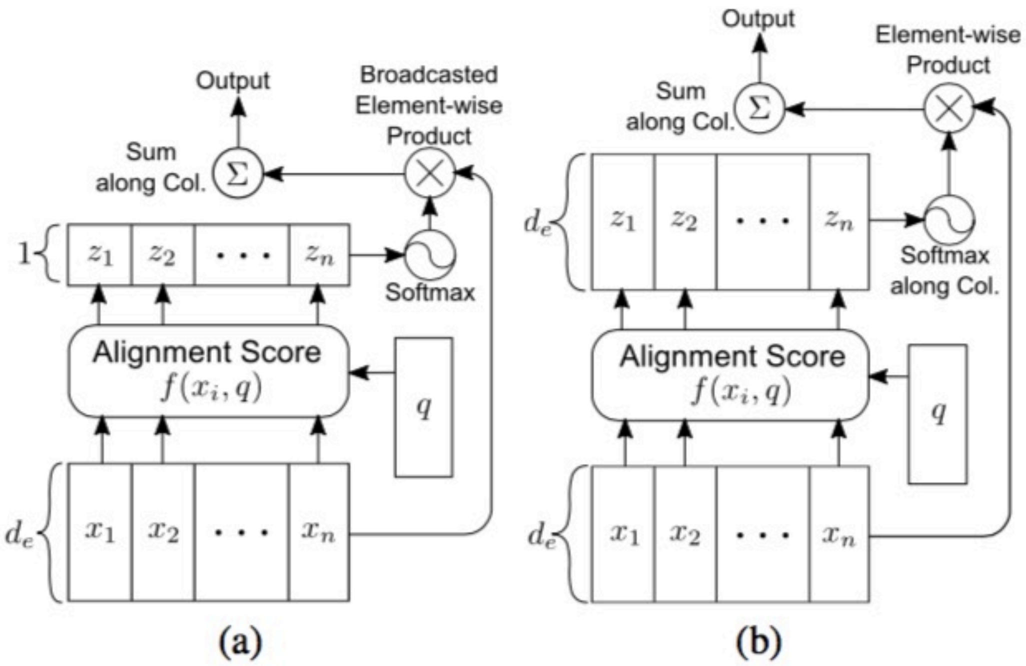
a^V	a^K	EN-DE BLEU
Yes	Yes	25.8
No	Yes	25.8
Yes	No	25.3
No	No	12.5

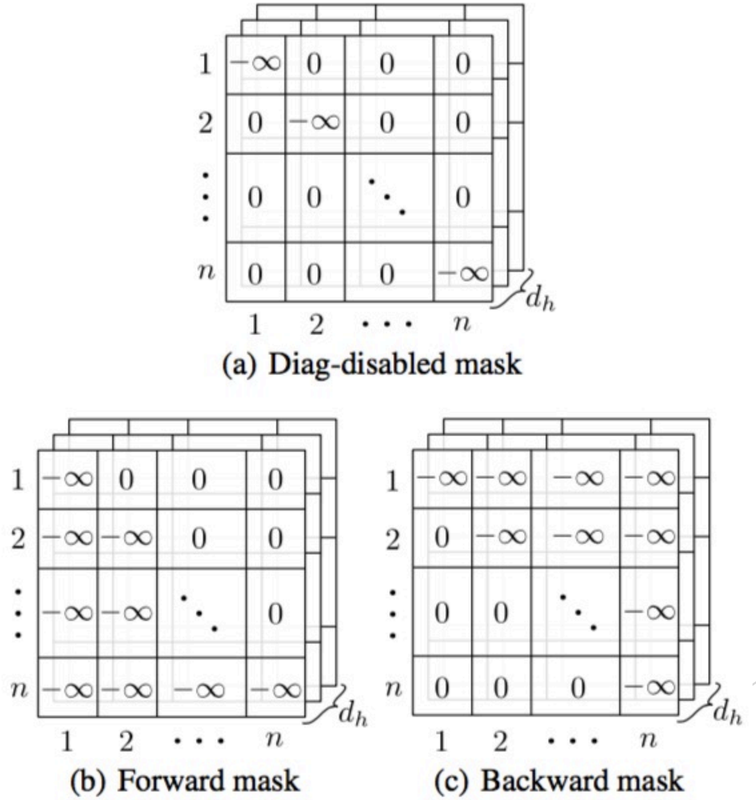
Discussion

- The parameters are shared across different layers and different heads.
- Automatically model the cliping distance?
- Cliping distance conditioned on q? Maybe another choice for local attention.

Directional Self-Attention

- Multi-dimentional: Feature-wise.
- Directional: Temporal order information is lost.





Directional Self-Attention

- Nature Language Inference. (549,367/9,842/9,824)

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**

Directional Self-Attention

- Sentiment Analysis. (8,544/1,101/2,210)

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**

Conclusion

- Multi-dimentional and Multi-head？
- Directional Attention， position embedding and RNN？
- An test: Multi-head with different masks + low-dimentional attetnion.