Multi-Scale Self-Attention for Text Classification

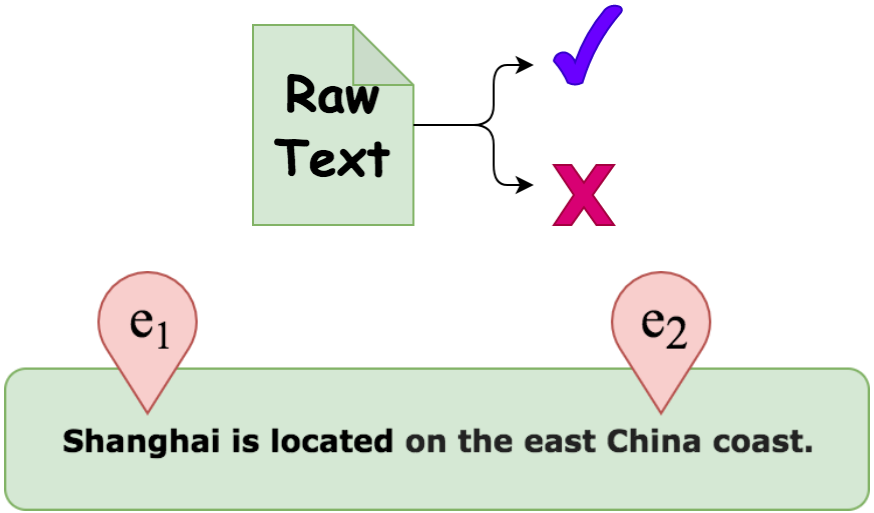
Qipeng Guo* Xipeng Qiu Pengfei Liu Xiangyang Xue Zheng Zhang

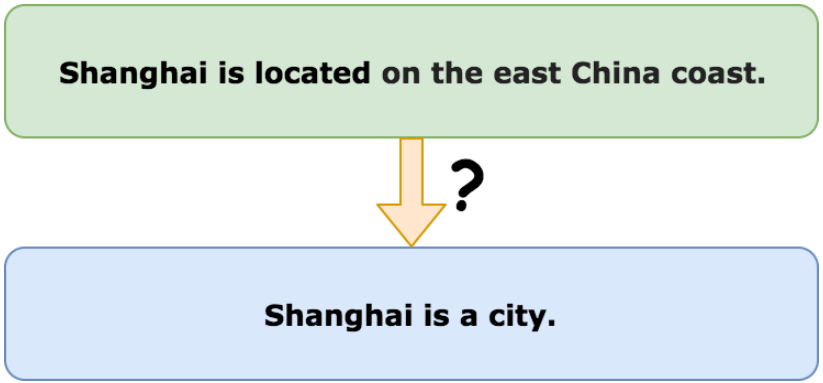
Fudan University
AWS Shanghai AI Lab

*presenter

Classification Tasks

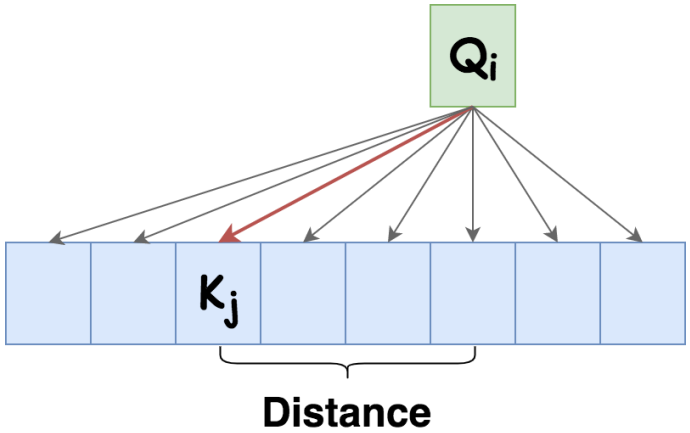
Text Classification, Sequence Labeling, Natural Language Inference





Attention? Scale?

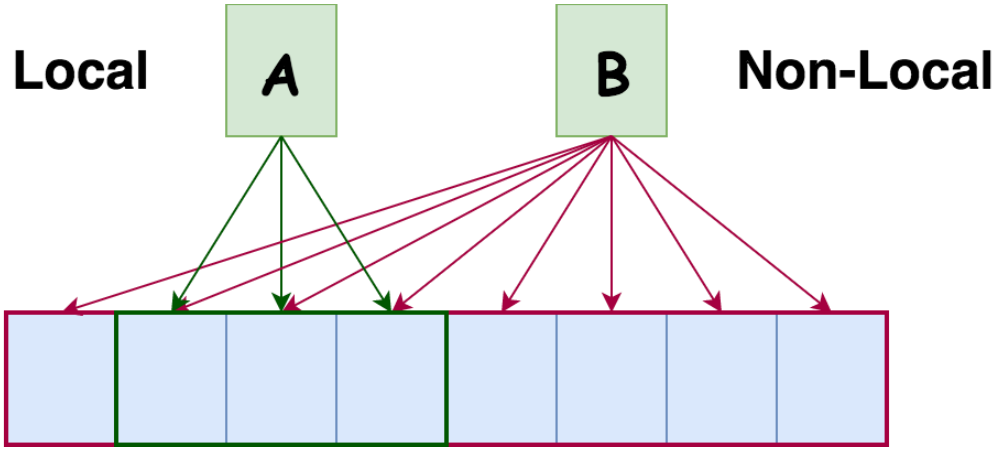
The scale measures the distance between two endpoints of attention edges on average.



$$\text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})$$

$$\mathbf{z}_{ij} = \frac{<\mathbf{Q}_i, \mathbf{K}_j>}{\sqrt{d}}$$

Which scale should be used?

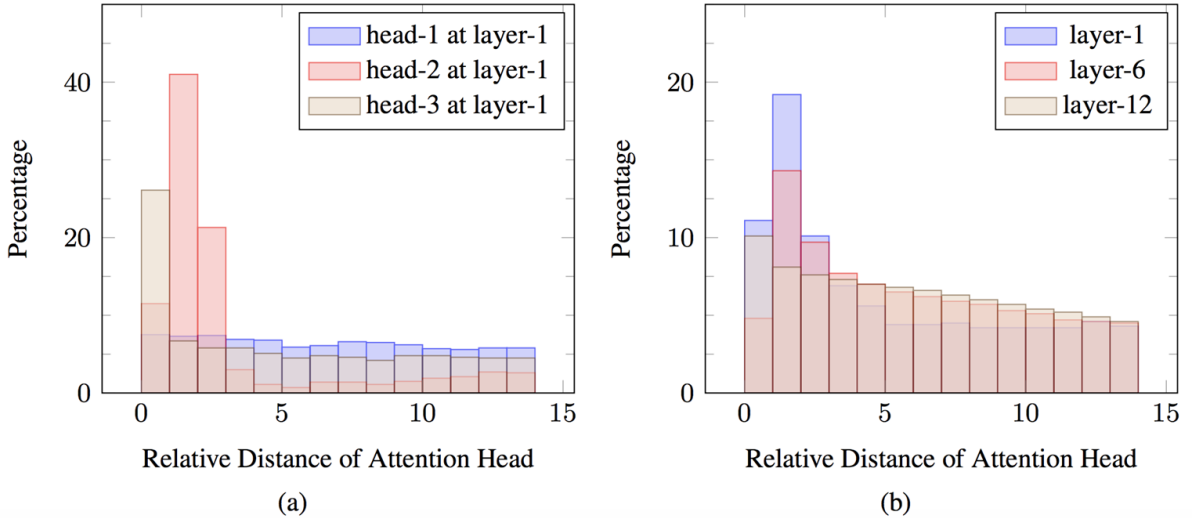


Small-Scale: Local patterns, N-gram

Large-Scale: Non-Local patterns, long-term dependencies

Observations of Pre-trained models

It is hard to make the trade-off. Take a look at what the model learned from data.

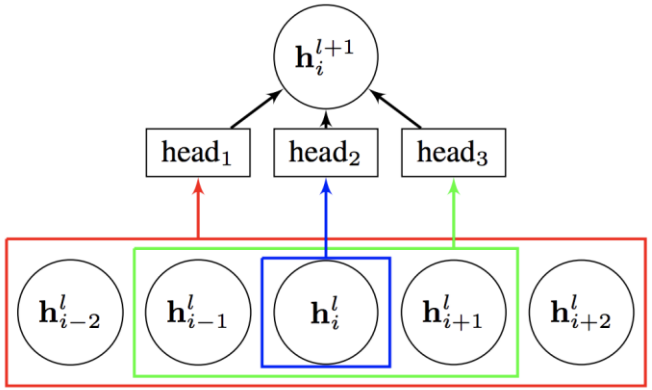


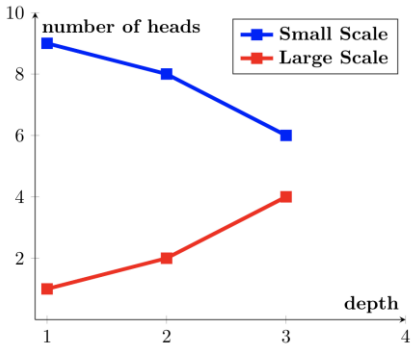
Results from BERT

Impact on model design

- **Different attention heads may work on different scales.**

- **There is a trend of scale over multiple layers.**





Experiments

Model	Acc
BiLSTM (Liu et al. 2016)	83.3
BiLSTM + self-att (Liu et al. 2016)	84.2
4096D BiLSTM-max (Conneau et al. 2017)	84.5
300D DiSAN (Shen et al. 2018a)	85.6
Residual encoders (Nie and Bansal 2017)	86.0
Gumbel TreeLSTM (Choi, Yoo, and Lee 2018)	86.0
Reinforced self-att (Shen et al. 2018b)	86.3
2400D Multiple DSA (Yoon, Lee, and Lee 2018)	87.4
Transformer	82.2
Multi-Scale Transformer	**85.9**

Dataset	Acc (%)			
	MS-Trans	Transformer	BiLSTM	SLSTM
Apparel	87.25	82.25	86.05	85.75
Baby	85.50	84.50	84.51	86.25
Books	85.25	81.50	82.12	83.44
Camera	89.00	86.00	87.05	90.02
DVD	86.25	77.75	83.71	85.52
Electronics	86.50	81.50	82.51	83.25
Health	87.50	83.50	85.52	86.50
IMDB	84.25	82.50	86.02	87.15
Kitchen	85.50	83.00	82.22	84.54
Magazines	91.50	89.50	92.52	93.75
MR	79.25	77.25	75.73	76.20
Music	82.00	79.00	78.74	82.04
Software	88.50	85.25	86.73	87.75
Sports	85.75	84.75	84.04	85.75
Toys	87.50	82.00	85.72	85.25
Video	90.00	84.25	84.73	86.75
Average	**86.34**	82.78	84.01	85.38

Ablation Study and Hyper-parameters

- **Single-Scale vs. Multi-Scale**
- **Ascending vs. Descending**

multi-scale	α	N'	L	Acc
A	1.0	5	3	85.5
B	0.5	5	3	**85.9**
C	0.0	5	3	84.9
D	-0.5	5	3	84.7
E	-1.0	5	3	84.3
single-scale	ω		L	Acc
F	3		3	84.3
G	$N/16$		3	83.9
H	$N/8$		3	81.7
I	$N/4$		3	80.7

Conclusion

- **Both small- and large-scale are important to language understanding**
- **The observation of BERT provides some guidelines of the model design**
- **The trade-off of scale is non-trivial, our experiments show that multi-scale models outperform than the single-scale**

Thank you for listening

Fudan University
AWS Shanghai AI Lab