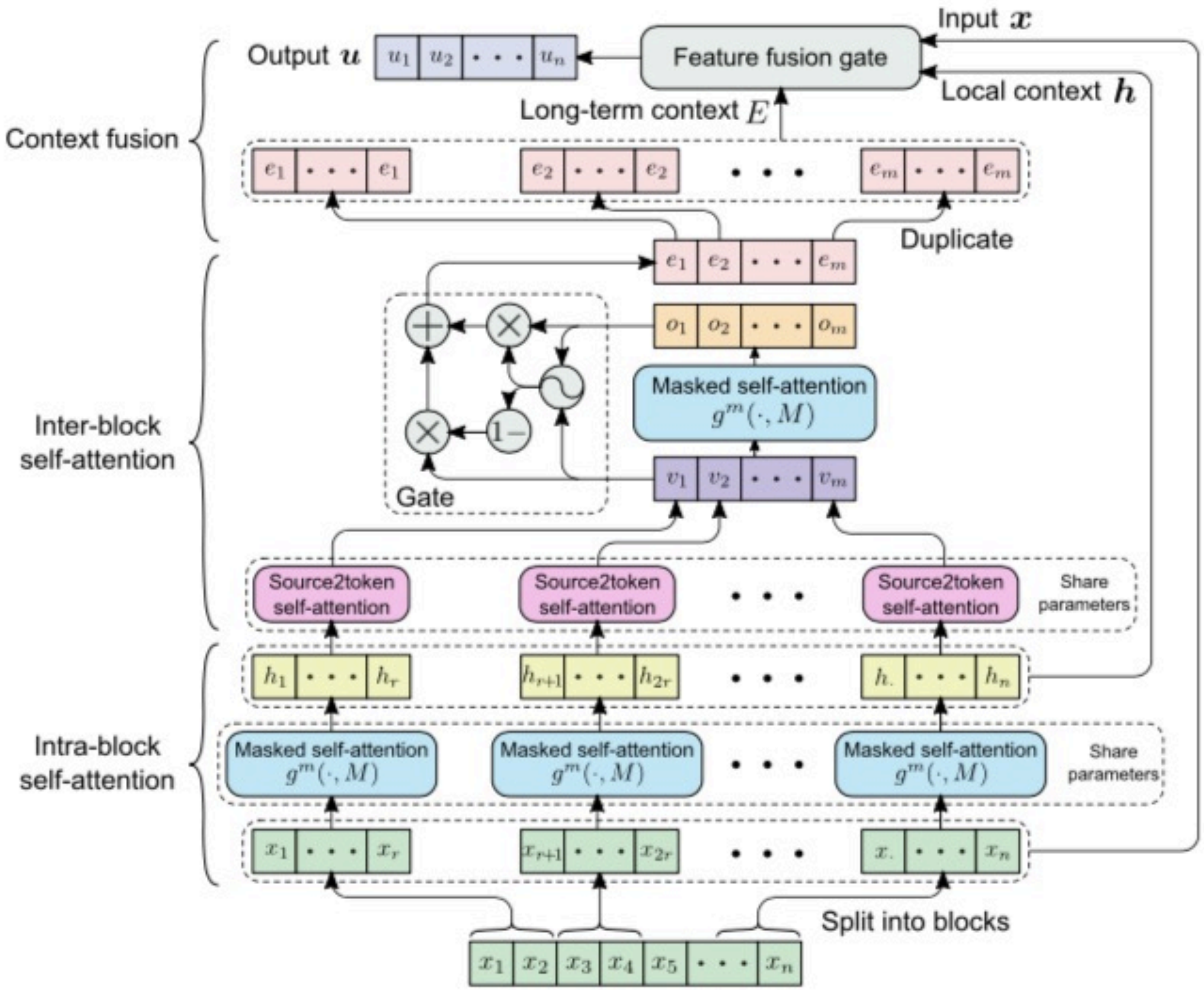
BI-DIRECTIONAL BLOCK SELF-ATTENTION FOR FAST AND MEMORY-EFFICIENT SEQUENCE MODELING

Presenter: Baosong Yang

Motivation

- SAN can model dependencies via highly parallelizable computation, but memory requirement grows rapidly in line with sequence length.
- Solusion: Splits the entire sequence into blocks.
  - intra-block SAN to each block for modeling local context
  - inter-block SAN to the outputs for all blocks to capture long-range dependency.

BLOCK SELF-ATTENTION



- split the input into m blocks of equal length r.
- In each block, do masked self attention.
- Generate a vector representation v of each block using a s2t self-attention
- Applies a masked self-attention to capture global dependency.
- To combine the local and global context features, a gate is used:

$$G = \mathrm{sigmoid}\left(W^{(g1)}\boldsymbol{o} + W^{(g2)}\boldsymbol{v} + b^{(g)}\right)$$
$$\boldsymbol{e} = G \odot \boldsymbol{o} + (1 - G) \odot \boldsymbol{v}$$

- Combine x,h,e:

$$F = \sigma\left(W^{(f1)}[\boldsymbol{x}; \boldsymbol{h}; E] + b^{(f1)}\right),$$
$$G = \mathrm{sigmoid}\left(W^{(f2)}[\boldsymbol{x}; \boldsymbol{h}; E] + b^{(f2)}\right),$$
$$\boldsymbol{u} = G \odot F + (1 - G) \odot \boldsymbol{x},$$

Appendix: Block Length

- Fixed sentence length n:

$$\xi \propto r^2 \cdot m + m^2 \cdot 1 = r^2 \cdot \frac{n}{r} + (\frac{n}{r})^2.$$

$$r = \sqrt[3]{2n}$$

- The sentence lengths that follow a normal distribution:
  - The upper bound of the expectation of random variable Z + μ

$$Z = \max_i X_i, \text{ for } i = 1, 2, \ldots, B.$$

$$\mathbb{E}[Z] \leq \sigma\sqrt{2 \ln B}$$

$$r = \sqrt[3]{2n} = \sqrt[3]{2(\sigma\sqrt{2 \ln B} + \mu)}$$

Experiments: Natural language inference

- training/dev/test split of 549,367/9,842/9,824 samples.

Table 2: Time cost and memory consumption of the different methods on SNLI. **Time(s)/epoch**: average training time (second) per epoch. **Memory(MB)**: Training GPU memory consumption (Megabyte). **Inference Time(s)**: average inference time (second) for all dev data on SNLI with test batch size of 100.

Model	Time(s)/epoch	Memory(MB)	Inference Time(s)	Test Accuracy
Bi-LSTM (Graves et al., 2013)	2080	1245	9.2	85.0
Bi-GRU (Chung et al., 2014)	1728	1259	9.3	84.9
Bi-SRU (Lei & Zhang, 2017)	1630	731	8.2	84.8
Multi-CNN (Kim, 2014)	284	529	2.4	83.2
Hrchy-CNN (Gehring et al., 2017)	343	2341	2.9	83.9
Multi-head (Vaswani et al., 2017)	345	1245	3.0	84.2
DiSAN (Shen et al., 2017)	587	2267	7.0	85.6
480D Bi-BloSAN	508	1243	3.4	85.7

Table 3: An ablation study of Bi-BloSAN. "Local" denotes the local context representations $\boldsymbol{h}$ and "Global" denotes the global context representations E. "Bi-BloSAN w/o mBloSA" equals to word embeddings directly followed by a source2token attention and "Bi-BloSAN w/o mBloSA & source2token self-attn." equals to word embeddings plus a vanilla attention without q.

Model	$\|\theta\|$	Test Accuracy
Bi-BloSAN	2.8m	85.7
Bi-BloSAN w/o Local	2.5m	85.2
Bi-BloSAN w/o Global	1.8m	85.3
Bi-BloSAN w/o mBloSA	0.54m	83.1
Bi-BloSAN w/o mBloSA & source2token self-attn.	0.45m	79.8

Experiments

- Reading Comprehension

Table 4: Experimental results for different methods on modified SQuAD task.

Context Fusion Method	$\|\theta\|$	Time(s)/Epoch	Dev Accuracy
Bi-LSTM (Graves et al., 2013)	0.71m	857	68.01
Bi-GRU (Chung et al., 2014)	0.57m	782	67.98
Bi-SRU (Lei & Zhang, 2017)	0.32m	737	67.32
Multi-CNN (Kim, 2014)	0.60m	114	63.58
Multi-head (Vaswani et al., 2017)	0.45m	140	64.82
Bi-BloSAN	0.82m	293	**68.38**

- Semantic Relatedness (4,500/500/4,927 instances for training/dev/test sets.)

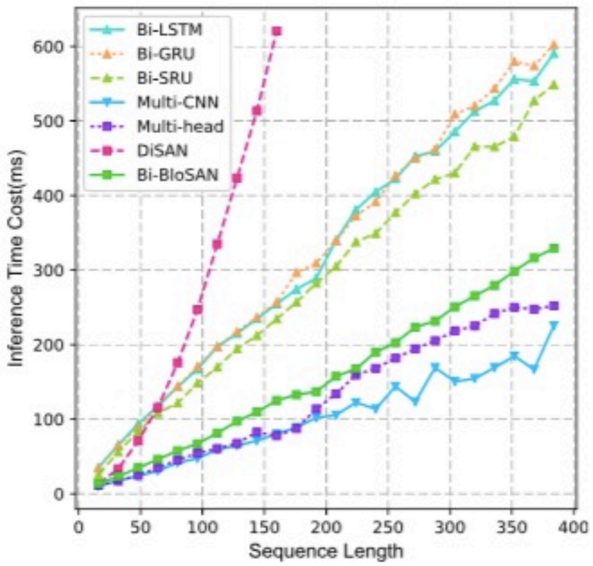
Model	Pearson's r	Spearman's ρ	MSE
Meaning Factory (Bjerva et al., 2014)	0.8268	0.7721	0.3224
ECNU (Zhao et al., 2014)	0.8414	/	/
DT-RNN (Socher et al., 2014)	0.7923 (0.0070)	0.7319 (0.0071)	0.3822 (0.0137)
SDT-RNN (Socher et al., 2014)	0.7900 (0.0042)	0.7304 (0.0042)	0.3848 (0.0042)
Constituency Tree-LSTM (Tai et al., 2015)	0.8582 (0.0038)	0.7966 (0.0053)	0.2734 (0.0108)
Dependency Tree-LSTM (Tai et al., 2015)	0.8676 (0.0030)	0.8083 (0.0042)	**0.2532 (0.0052)**
Bi-LSTM (Graves et al., 2013)	0.8473 (0.0013)	0.7913 (0.0019)	0.3276 (0.0087)
Multi-CNN (Kim, 2014)	0.8374 (0.0021)	0.7793 (0.0028)	0.3395 (0.0086)
Hrchy-CNN (Gehring et al., 2017)	0.8436 (0.0014)	0.7874 (0.0022)	0.3162 (0.0058)
Multi-head (Vaswani et al., 2017)	0.8521 (0.0013)	0.7942 (0.0050)	0.3258 (0.0149)
DiSAN (Shen et al., 2017)	**0.8695 (0.0012)**	**0.8139 (0.0012)**	0.2879 (0.0036)
Bi-BloSAN	0.8616 (0.0012)	0.8038 (0.0012)	0.3008 (0.0091)

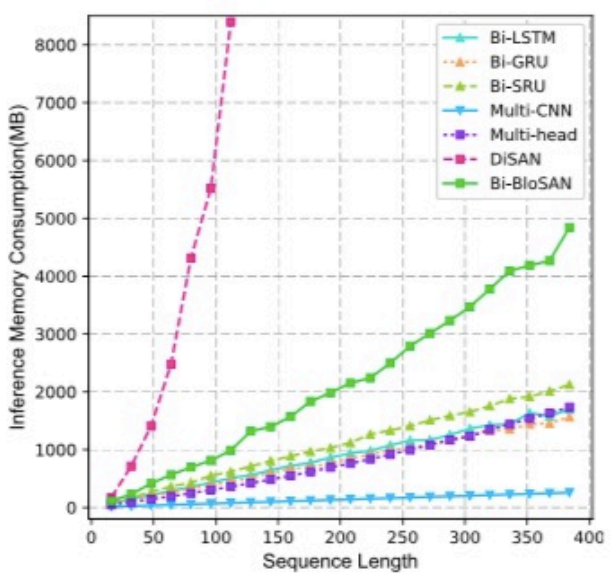
Experiments

- Sentence Classifications

Model	CR	MPQA	SUBJ	TREC	SST-1	SST-2
cBoW (Mikolov et al., 2013a)	79.9	86.4	91.3	87.3	/	/
Skip-thought (Kiros et al., 2015)	81.3	87.5	93.6	92.2	/	/
DCNN (Kalchbrenner et al., 2014)	/	/	/	93.0	86.8	48.5
AdaSent (Zhao et al., 2015)	83.6 (1.6)	**90.4 (0.7)**	92.2 (1.2)	91.1 (1.0)	/	/
SRU (Lei & Zhang, 2017)	**84.8 (1.3)**	89.7 (1.1)	93.4 (0.8)	93.9 (0.6)	**89.1 (0.3)**	/
Wide CNNs (Lei & Zhang, 2017)	82.2 (2.2)	88.8 (1.2)	92.9 (0.7)	93.2 (0.5)	85.3 (0.4)	/
Bi-LSTM (Graves et al., 2013)	84.6 (1.6)	90.2 (0.9)	**94.7 (0.7)**	94.4 (0.3)	87.7 (0.6)	49.9 (0.8)
Multi-head (Vaswani et al., 2017)	82.6 (1.9)	89.8 (1.2)	94.0 (0.8)	93.4 (0.4)	83.9 (0.4)	48.2 (0.6)
DiSAN (Shen et al., 2017)	**84.8 (2.0)**	90.1 (0.4)	94.2 (0.6)	94.2 (0.1)	87.8 (0.3)	**51.0 (0.7)**
Bi-BloSAN	**84.8 (0.9)**	**90.4 (0.8)**	94.5 (0.5)	**94.8 (0.2)**	87.4 (0.2)	50.6 (0.5)

- Consuming





Conclusion

- Split long sentence into blocks to model the local context and global context.

- Strange: the block size is calculated based on the memory requirement with out any linguistic prior.
- Only 1 layer? Is it fair for conventional self-attention model?
- Without “bidirection” and “multi-dimension”?