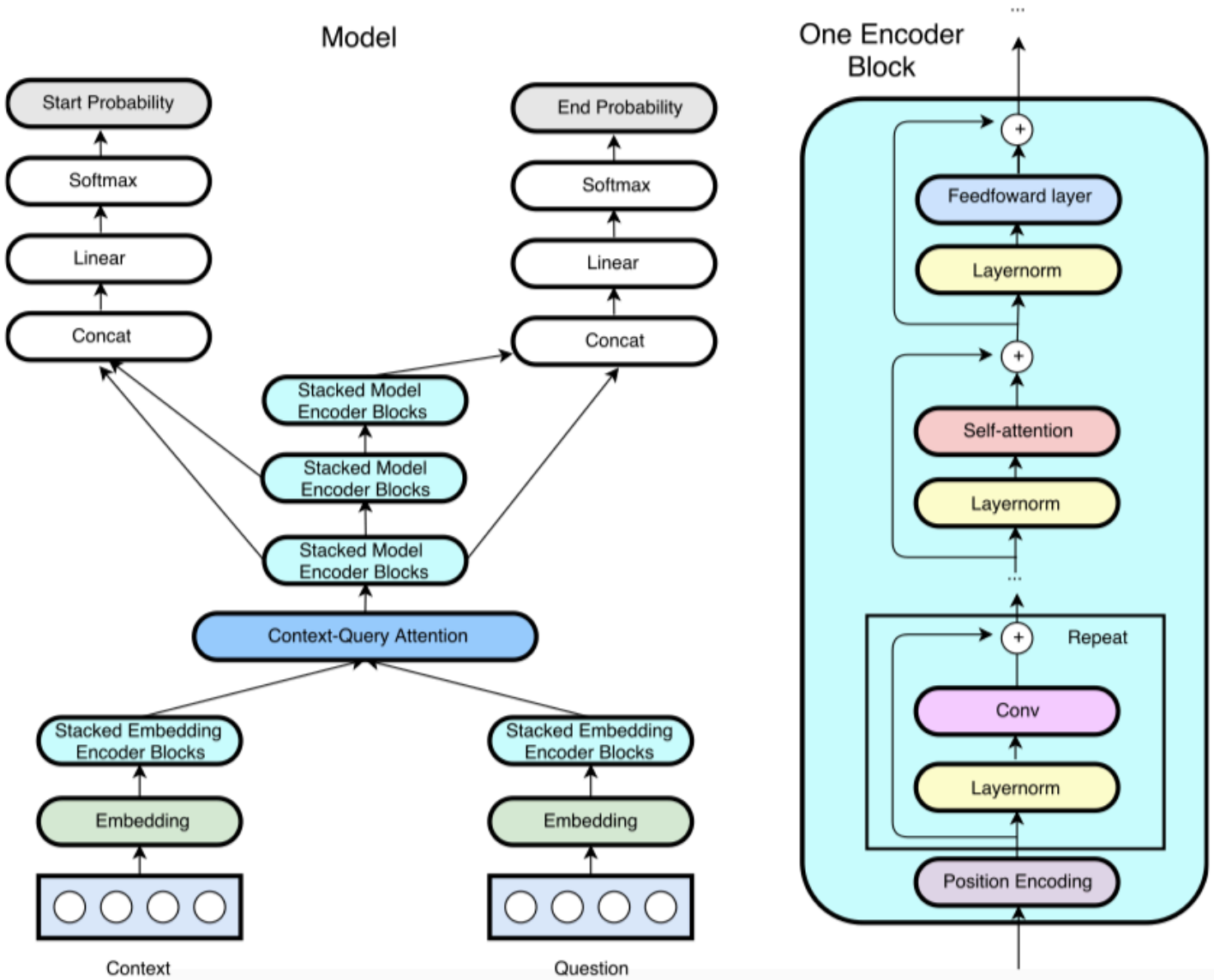
QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension

Baosong Yang

Motivation

- RNN slow
- Contribution: An Reading comprehension model
  - CNN learns the local information
  - SAN model the global dependencies
- Reading Comprehension:
  - Given a context paragraph with n words C = {c1, c2, ..., cn} and the query sentence with m words Q = {q1, q2, ..., qm}, output a span S = {ci,ci+1,...,ci+j} from the original paragraph C.

Architecture



Details

- Embeddings: Word + Characters
  - Word: Pretrained by GloVe
  - Character: 16*d represent a word, then, take maximum value of each row
- Encoder Block: CNN+SAN
  - CNN: modeling the local information with kernel size 7 and 4 layers
  - SAN: multi-head self-attention

Experiments

- Data：SQuAD with 87.5K for training, 10.1K for validation, 10.1K for testing.
- Speed up 5x.

	Published[12]	LeaderBoard[13]
Single Model	EM / F1	EM / F1
LR Baseline (Rajpurkar et al., 2016)	40.4 / 51.0	40.4 / 51.0
Dynamic Chunk Reader (Yu et al., 2016)	62.5 / 71.0	62.5 / 71.0
Match-LSTM with Ans-Ptr (Wang & Jiang, 2016)	64.7 / 73.7	64.7 / 73.7
Multi-Perspective Matching (Wang et al., 2016)	65.5 / 75.1	70.4 / 78.8
Dynamic Coattention Networks (Xiong et al., 2016)	66.2 / 75.9	66.2 / 75.9
FastQA (Weissenborn et al., 2017)	68.4 / 77.1	68.4 / 77.1
BiDAF (Seo et al., 2016)	68.0 / 77.3	68.0 / 77.3
SEDT (Liu et al., 2017a)	68.1 / 77.5	68.5 / 78.0
RaSoR (Lee et al., 2016)	70.8 / 78.7	69.6 / 77.7
FastQAExt (Weissenborn et al., 2017)	70.8 / 78.9	70.8 / 78.9
ReasoNet (Shen et al., 2017b)	69.1 / 78.9	70.6 / 79.4
Document Reader (Chen et al., 2017)	70.0 / 79.0	70.7 / 79.4
Ruminating Reader (Gong & Bowman, 2017)	70.6 / 79.5	70.6 / 79.5
jNet (Zhang et al., 2017)	70.6 / 79.8	70.6 / 79.8
Conductor-net	N/A	72.6 / 81.4
Interactive AoA Reader (Cui et al., 2017)	N/A	73.6 / 81.9
Reg-RaSoR	N/A	75.8 / 83.3
DCN+	N/A	74.9 / 82.8
AIR-FusionNet	N/A	76.0 / 83.9
R-Net (Wang et al., 2017)	72.3 / 80.7	76.5 /84.3
BiDAF + Self Attention + ELMo	N/A	**77.9/ 85.3**
Reinforced Mnemonic Reader (Hu et al., 2017)	73.2 / 81.8	73.2 / 81.8
Dev set: QANet	**73.6 / 82.7**	N/A
Dev set: QANet + data augmentation ×2	**74.5 / 83.2**	N/A
Dev set: QANet + data augmentation ×3	**75.1 / 83.8**	N/A
Test set: QANet + data augmentation ×3	**76.2 / 84.6**	76.2 / 84.6

Conclusion

- New architecture for local-global modeling.
- SAN is still valid for long sequence.
- Localness modeling replaced by SAN？ Maybe more flexible.