Recent Explorations of Self-Attention Networks

Presenter: Baosong Yang

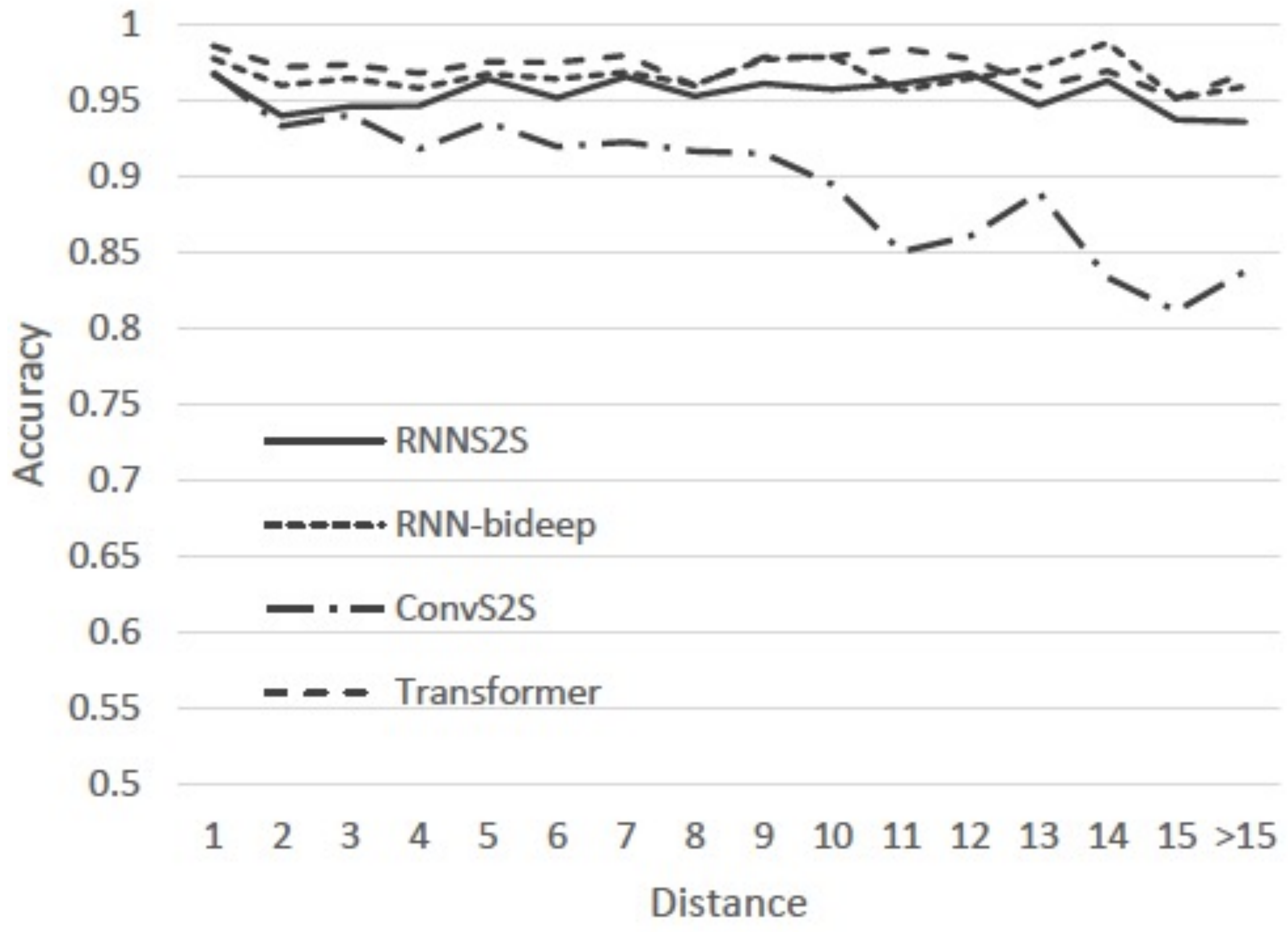
Why Self-Attention? (EMNLP 2018)

- Motivation
  - Modeling long-range dependencies through shorter-path? Theoretic!
- Experiments
  - Subject-verb agreement over long distances
    - Verbs must agree with their subjects in both grammatical number and person
  - Extraction of semantic feature in word sense disambiguation
    - Replace the ambiguous word with other translations of its translation
  - Setting:
  - RNNS2S: encoder(1-bi+6-uni) and decoder (8 uni)
  - RNNbideep: 4-bi
  - Transformer: 8 SAN

Long-distance

- Cannot conclude that Transformer models are stronger than RNN models for long-distances
- CNN: accuracy increases when the local context size becomes larger, but the BLEU score not.
  - BLEU measures only on the n-grams

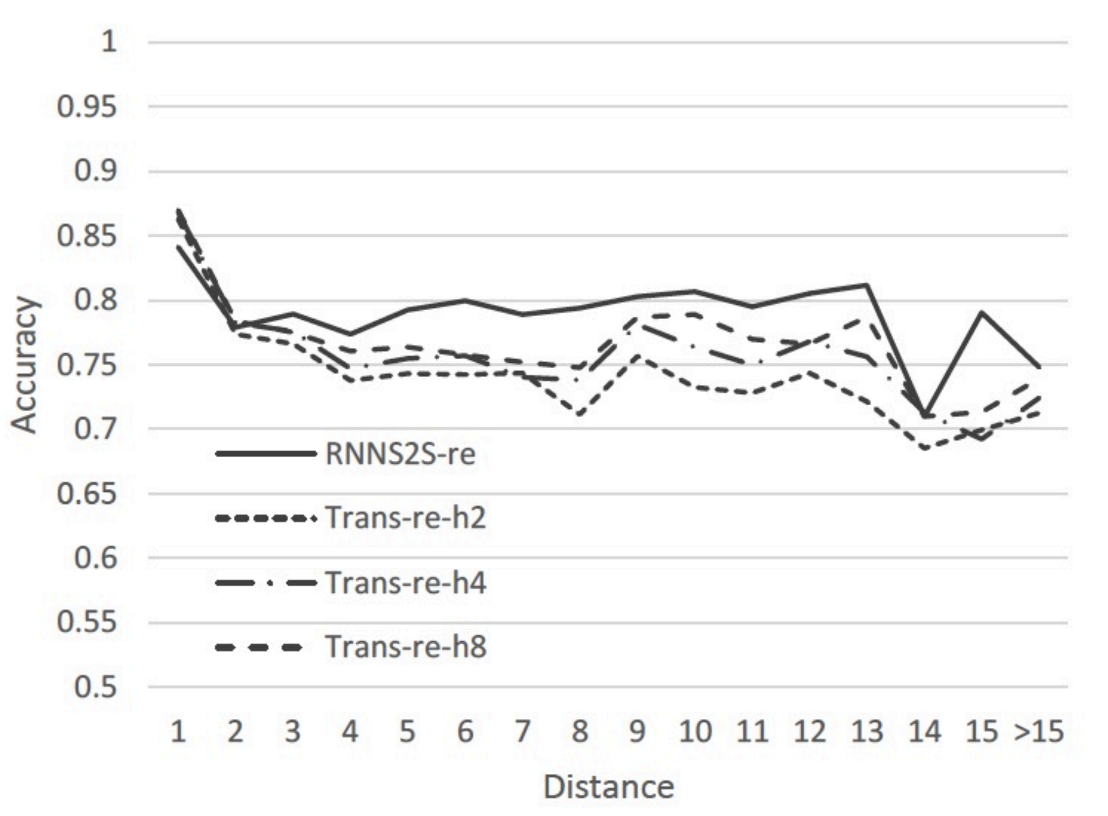
Layer	K	Ctx	2014	2017	Acc(%)
4	3	4	22.9	24.2	81.1
6	3	6	23.6	25.0	82.5
8	3	8	**23.9**	**25.2**	84.9
8	5	16	23.5	24.7	89.7
8	7	24	23.3	24.6	**91.3**



Multi-Head Benefits More

- The accuracy over long-distances can be improved substantially, by increasing the number of heads.

Model	2014	2017	PPL	Acc(%)
RNNS2S-s	7.3	7.8	47.8	77.3
Trans-s	7.2	8.0	44.6	74.6
RNNS2S-re	9.2	10.5	39.2	77.7
Trans-re-h2	9.6	10.7	36.9	71.9
Trans-re-h4	9.5	11.9	35.8	73.8
Trans-re-h8	9.4	10.4	36.0	75.3



WSD

- Transformers distinctly outperform RNNS2S and ConvS2S models on WSD tasks.
- Chen et al., 2018 (ACL) encoder(SAN) and decoder (RNN) is worse than Transformer.

Model	DE→EN				DE→FR		
	PPL	2014	2017	Acc(%)	PPL	2012	Acc(%)
RNNS2S	5.7	29.1	30.1	84.0	7.06	16.4	72.2
ConvS2S	6.3	29.1	30.4	82.3	7.93	16.8	72.7
Transformer	**4.3**	**32.7**	33.7	**90.3**	**4.9**	**18.7**	**76.7**
uedin-wmt17	–	–	**35.1**	87.9	–	–	–
TransRNN	5.2	30.5	31.9	86.1	6.3	17.6	74.2

Conclusion

- CNN and SAN which have shorter paths through networks are not empirically superior to RNNs in modeling long distance dependencies.
- Multi-head affects the ability mentioned above.
- Transformer outperform other models on WSD.
- Simply computing BLEU score may be insufficient to learn the performance of a model.

Overview of Recent Studies on SAN

- Dependencies
  - Short-term
  - Long-term
  - Phrasal Pattern
- Multi-heads
  - Multiple features
  - Interaction
- Temporal Information
  - Position
  - Recurrence
- Others
  - Multi-layer
  - Training
  - Components
- Applications

Dependency

- Short term:
  - Local SAN (Sperber et al., Interspeech2018)
  - Block (Shen et al., ICLR 2018)
  - CNN (Yu et al., ACL 2018)
  - Block Review (Hao Jie)
- Long term:
  - Global Context (Yang et al., AAAI 2019; Hao et al., AAAI 2019)
  - Not good at subject-verb agreement (Tang et al., EMNLP 2018; Trans et al., EMNLP 2018)
- Phrasal pattern:
  - Local Attention (Luong et al., EMNLP 2015)
  - Localness (Yang et al., EMNLP 2018)
- Hard attention:
  - Hard attention (Shen et al., IJCAI 2018; Xinwei)
  - others (e.g. syntax-aware or dependency enhanced SAN in AAAI 2019)
- The weighted average operation may be insufficient to fully capture dependencies which may be alleviated by multi-head mechanism.
- The description “build direct revelance between words benefits performance (Gehring et al., ICML 2017; Vaswani et al., NIPS 2017) ” may be not strict.

Multi-head

- Multiple features
  - Multi-dim (Lin et al., ICLR 2017; Shen et al., AAAI 2018)
  - Different subspace of representation (Vaswani et al., NIPS 2017)
  - Disagreement (Lin et al., ICLR 2017; Li et al., EMNLP 2018)
  - Different local scopes (Yang et al., EMNLP 2018)
  - Different distance dependencies (Tang et al., EMNLP 2018)
- Interaction
  - Weighted (Ahmed et al., 2018)
  - CNN and Bilinear Pooling (Li et al., AAAI 2019)
- Multi-Head benefits more for Transformer!
- How to extract and interact features?
  - e.g. more linguistic information (Cohn et al., NAACL 2016; Li Jian)

Temporal Information

- Position
  - Absolute (Gehring et al., ICML 2017; Vaswani et al., NIPS 2017)
  - Relative (Shaw et al., NAACL 2018)
- Recurrence
  - Bi-directional (Shen et al., AAAI 2018)
  - RNN+ (Chen et al., ACL 2018)
  - Target mean and nearest attention (Zhang et al., ACL 2018)
    - May drop the BLEU score
    - Mean fail to fully capture the context (Yang et al., AAAI 2019)
  - RNN better than SAN and CNN in language modeling (Trans et al., EMNLP 2018)
- Temporal Information is important. However, in practice, there is marginal improvement by incorporating recurrence.
  - Translation (Hao et al., AAAI 2019)
  - Word Sense Disambigation (Tang et al., 2018)
  - Encoder SAN Decoder RNN performance better in (Domhan et al., ACL 2018), but worse in (Trans et al., EMNLP 2018)
  - Speed!
- Block review? (Hao Jie)

Others

- Multi-layer
  - Semantic and syntactic information (Peter et al., NAACL 2018)
  - Coarse to fine and revise (Domhan et al., ACL 2018; Dehghani et al., Google 2018)
  - Different Local Scope (Yang et al., EMNLP 2018)
  - Aggregation (Dou et al., EMNLP 2018)
- Large Scale
  - Large Batch size and larger learning rate (Facebook 2018)
- Components are important (Domhan et al., ACL 2018)
  - Residual FFN
  - Layer normalization

Applications

- Author profiling (Use twitters to predict age) (Lin et al., ICLR 2017)
  - Author Profiling dataset: http://pan.webis.de/clef16/pan16-web/author-profiling.html
- Sentiment analysis
  - Yelp dataset: https://www.yelp.com/dataset_challenge (Lin et al., ICLR 2017)
  - Stanford SST: https://nlp.stanford.edu/sentiment/ (Shen et al., AAAI 2018)
- Natural language inference (Lin et al., ICLR 2017; Shen et al., IJCAI 2018)
  - SNLI corpus: https://nlp.stanford.edu/projects/snli/
- Semantic Relatedness (Shen et al., 2018)
  - SICK datasets: http://alt.qcri.org/semeval2014/task1/index.php?id=data-and-tools
- Probing Tasks (10 classification tasks) (Conneau et al. ACL 2018) Evaluating Representations!!!
- NMT
- QA (Yu et al., ACL 2018)
  - SQuAD: https://rajpurkar.github.io/SQuAD-explorer/
- Speech (Sperate et al., Interspeech 2018)
- Universal Transformer (Not release)