Paper Review CSE 573,

Dissecting Contextual Word Embeddings:
Architecture and Representation

Alexander Van Roijen

February 26, 2020

I Introduction

Bidirectional language models (biLMs) have shown superior performance in various NLP tasks with their contextual representations versus classic word embeddings. However, the various choices for biLMs (CNN, LSTM, Transformer) has not been well studied. Which one is better for what situations? This paper demonstrates that LSTMs show the best results and then dig deeper into what the layers of the LSTM are learning when creating the embeddings.

II Contributions

Highlighted in table 1, they identified that CNNs and Transformer architectures for deep biLMs are superior in inference time and in perplexity than their LSTM counter parts (Peters et al. 3). However, in table 2, they highlight that when using an ELMo presentation of their context vectors, LSTM based biLMs allow for better performance on all 4 NLP tasks benchmarked (Peters et al. 4).

Lastly I would highlight their description of these context vectors for a particular example.

> In addition, we can see how the biLM is implicitly learning other linguistic information in the up-per layer. For example, all of the verbs (“says”,“can”, “afford”, “maintain”, “meet”) have high similarity suggesting the biLM is capturing part-of-speech information. We can also see some hints that the model is implicitly learning to perform coreference resolution by considering the high contextual similarity of “it” to “government”,the head of “it”s antecedent span (Peters et al. 5).

Both of these contributions are what I would claim to be the most significant contributions of this paper compared to the related works highlighted and discussed in the related works section.

III Significance

Overall, I think this paper fills in the blanks and more rigorously supports the use of LSTM based biLMs for context vectors for NLP tasks. There are plenty of examples of "taken for granted" processes and actions. Having data to support those intuitions can lead to future work improving previous systems[2], and give a better basis to defend their usage than simply a "that is whats been done before" explanation.

IV Strengths

- **Background**
  I appreciate the amount of background given throughout the paper, particularly the discussion of biLMs and how they work. This educates the reader on the subject as well as motivates intuitive reasoning of why they should provide better performance on NLP tasks.
- **Concreteness**
  The use of trained examples to demonstrate the representations or contexts learned by the biLMs, seen in figure 1, provide concrete proof to their claims.
- **Storytelling**
  The paper has a great flow, not stopping too much to focus on the numbers for the results and transitioning to each subject in an expected manner. Leaving details of model implementation and hyperparameter tuning to the appendix certainly aids in this endeavor.

V Critiques

- **Spelling/Grammatical errors**
  There were a fair number of mistakes in their grammar throughout the paper. For example

  > "...Transformer accuracies 0.2% / 0.6% (matched/mismatched) less **then** the 2-layer LSTM" (Peters et al. 5).

  > "... between all pairs **of words in single sentence** using the 4-layer LSTM" (Peters et al. 6).

"As our results **have show that** computationally efficient architectures also learn high quality representations "(Peters et al. 8).

VI References

[1] Peters, M. E., Neumann, M., Zettlemoyer, L.,& Yih, W. T. (2018). Dissecting contextual word embeddings: Architecture and representation. arXiv preprint arXiv:1808.08949.
[2] Li, G., Zhang, P., & Jia, C. (2018). Attention Boosted Sequential Inference Model. arXiv preprint arXiv:1812.01840.