**Deep Contextualized Word Representations (ELMo)**

**Technical Contribution**

In this paper, a new type of deep contextualized word representation is introduced that models both complex characteristics of word use (eg., syntax and semantics), and how these uses vary across linguistic contexts (i.e, to model polysemy). ELMo word representations are functions of the entire input sentence. Vectors used are derived from a bi-directional LSTM (Long short-term memory) that is trained with a couple language model (LM) objective on a large text corpus. ELMo representations are a function of all the internal layers of the biLM. Combining the internal states allows for very rich word representations. The biLM model has two layers stacked together where each layer has two passes - forward pass and backward pass. The forward pass contains information about a certain word and captures the context before that word. In the backward pass, the information about the word is captured along with the context after that word. This pair of information forms the intermediate word vectors. ELMo calculates the weighted sum of the raw word vectors and the intermediate word vectors. This allows for semi-supervised learning, where the biLM is pre-trained at a large scale and easily incorporated into other neural architectures.

Adding ELMo to different models has resulted in increased accuracy in tasks such as in the Stanford Question Answering Dataset (SQuAD) task where adding ELMo to BiDAF model resulted in an accuracy increase of 4.7%. In the sentiment analysis task (in the Stanford Sentiment Tree bank) the state of the art biattentive classification network (BCN) augmented with CoVe when replaced with ELMo (in place of CoVe) yielded in 1.0% absolute accuracy improvement.

**Strengths**

* Word representations in ELMo are rich. Higher-level LSTM states capture context-dependent aspects of word meaning while low level states model aspects of syntax.
* ELMo representations work extremely well in practice. They are easy to add to existing models and the addition of ELMo representations improves the performance by increasing accuracy and reducing relative error by up to 20%.
* In cases of Polysemy where a word could have multiple meanings, ELMo word vectors are able to successfully detect this issue since they take the entire input sentence into equation for calculation. Hence, it generates vectors based on the context they appear in.

**Weaknesses**

* ELMo uses the RTL (right to left) and LTR (left to right) model by training words separately and concatenating the two models, which is twice as computationally expensive as a single bidirectional model.[1]

**Improvements**

* [1] improves on ELMo by using a single bidirectional model that uses one pass on a sentence instead of the two passes (LTR and RTL). This results in half the computation complexity.

**References**

[1] J.Devlin, M.Chang, K.Lee, K.Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019