ELMo

Deep contextualized word representations

-Elmo is a function providing contextualized embedding , it takes sentence (context) as an input and outputs different vectors (embeddings) for the same word

-The magic behind ELMo is that it uses all internal states of biLM (concatenated forward LM and backward LM)

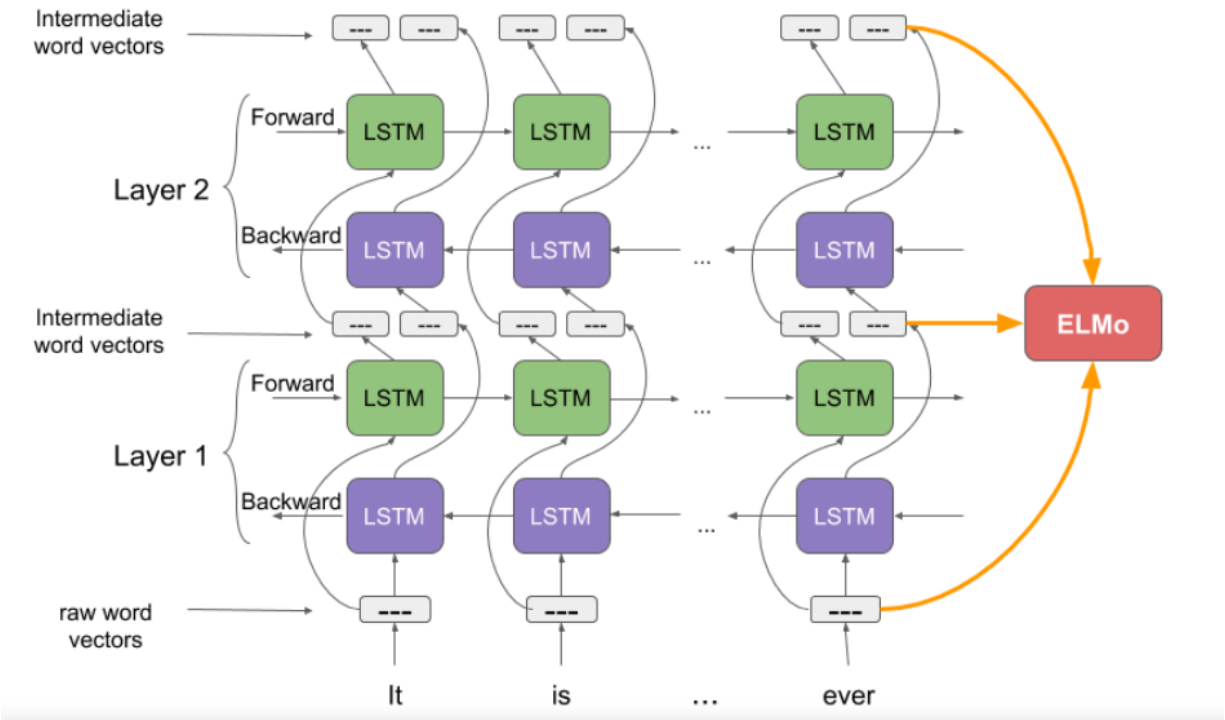
-One of the most benefits of using LM for NLP model is: you don't need additional labels on data for train, but you just need large data to predict next word or previous word in the given data

- Forward LM trained to predict next word given previous words
- Backward LM trained topredict previous word given future words

-ELMo uses all internal states while SOTA(State-of-art)uses only last layer's output for contextualized embedding vector

- ELMo word vectors are computed on top of a two-layer bidirectional language model (biLM). This biLM model has two layers stacked together. Each layer has 2 passes — forward pass and backward pass:

- given a sentence, then the first word is converted into character embedding then the character embedding goes to first LSTM cell
- The architecture uses a character-level convolutional neural network (CNN) to represent words of a text string into raw word vectors
- These raw word vectors act as inputs to the first layer of biLM
- The forward pass contains information about a certain word and the context (other words) before that word
- The backward pass contains information about the word and the context after it
- This pair of information, from the forward and backward pass, forms the intermediate word vectors
- These intermediate word vectors are fed into the next layer of biLM
- The final representation (ELMo) is the weighted sum of the raw word vectors and the 2 intermediate word vectors



While the initial layer outputs context independent embeddings, the later hidden layers outputs context dependent embeddings

The first LSTM output has residual connection with character embedding, they didn't say why but residual connection helps in 2 ways:

1) The later layers can learn from initial layer's features well.
2) While doing training, the residual connection can prevent gradient vanishing issue during back propagation.

2 Main reasons for using character embedding from the cnn

1) The initial layer supposed to be context independent
2) They wanted to compare ELMo score to the model using pretrained word embeddings so,
   They didn't use the pretrained word embedding in ELMo