Tech Review | Chronology of the state of the art models in NLP

In this tech review I aim to go through the most popular NLP state of the art technologies that have made the biggest impact in the recent years and give a brief overview of their characteristics. It starts with the Recurrent Neural Networks and in specific the Long Short-Term Memory model, which has been used in all kinds of problems that involve sequential data, like text analysis. Second, the attention mechanism with the Transformer, that since it was published in 2017, has proved to have an important impact in NLP and has been the key in further developments. Finally, the BERT model is described to finish the review. There have been more improvements in the most recent times but they depart from the ideas described in these three technologies.

RNN and LSTM

Recurrent Neural Networks (RNNs) are more flexible than other types of Artificial Neural Networks at using the context information. We need to use the context to complete our understanding of each word of a sentence and each sentence of a document. RNNs achieve this by adding loop, for example, by making the output of the network at time t-1 part of the input at time t, allowing the information to persist. This is inspired by the cyclical connectivity of neurons on the brain. But they have one problem for problems in NLP which is not being able to store information for long periods of time, limiting the context they can access.

Long Short-Term Memory networks (LSTMs) are a redesigned RNN capable of learning long-term dependencies. They are based on units called memory cells, capable of storing information for long timespans. These memory cells process information with one input gate, one output gate and a third 'forget' gate. Also LSTM networks can be unidirectional and bidirectional in terms of where they learn the context from.
They have been applied to a large variety of problems like speech recognition, anomaly detection in sequenced data and sentiment analysis. LSTM have many variants and can be combined with other models like hidden Markov models.

Attention Mechanism and Transformer

There are some cases where the LSTM have a weakness on handling long sentences even though they have shown to be better at using context than RNNs. More importantly, LSTMs can't give more importance to some of the input words over others, which is highly desired in specific situations when the context has several syntactic groups and some of them much more relevant than others. The attention mechanism works specifically on fixing this issue.

The attention mechanism was applied first on a Sequence to sequence architecture for machine translation. On a high level description, the attention model computes a set of attention weights that evaluates the contexts words and assigns them weights depending on how much value they provide. To compute these weights we use a small neural network that has as input the state of the RNN in the previous time step and

the features computed from the words in the current time step. The output of this small NN is the weights for each of the features. This weighted features are then used to produce the output of the complete model, which in this case is the translated input.

The small NN that computes the weights is also train along with the model and it learns to tell you how much attention your model should pay to the features of the input. One downside of the attention mechanism is that has a quadratic time complexity because of the training of the NN that computes the attention weights.

The attention mechanism achieve its revolutionary status in the NLP community with the paper Attention is all you need published in 2017 by researchers at Google and the University of Toronto. Transformer is a deep learning model for sequential data. One of its advantages is that it can process the data in any order and it is not constrained by processing the beginning before the end, which allows parallelization and faster training times.

BERT

Bidirectional Encoder Representations from Transformers (BERT) was proposed in 2018 achieving state of the art accuracy in NLP:

- General Language Understanding Evaluation score: 80.5% (7.7% improvement)
- MultiNLI accuracy: 86.7% (4.6% improvement)
- SQuAD v1.1: 93.2% (1.5% improvement)
- SQuAD v1.2: 83.1% (5.1% improvement)

The BERT uses a semi-supervised learning and it can be described as the encoder part of the transformer architecture. Basically the BERT model is trained on massive datasets acquiring language-processing abilities. The trained BERT model can then be used to empowered other models and be built and trained in a supervised way. BERT would take a sequence of words as input. The model then contains a set of encoder layers to apply self-attention and then process the results on a neural network. The output of BERT then can be used as the input of a second model, for example a classifier.

Two versions of the BERT model were released to the public and pre-trained:

- $BERT_{BASE}$: 12 layers in the encoder task. Useful to compare performances with other models.
- $BERT_{LARGE}$: 24 layers in the encoder task. Trained on a extremely large dataset and the one used to achieve the state of the art results shown in the paper.

Conclusion

Those are the main ideas for these three major NLP steps. To conclude, BERT is a huge improvement in the NLP world and the fact that pre-trained model were released together with the original paper has been very useful for the community to make it available for everyone to use as an encoder for many different models and understand the model from first hand. BERT has proved a real impact in NLP and is currently being used in technologies like Google Search.

References

https://medium.com/dataseries/the-current-state-of-the-art-in-natural-language-processing-nlp-5c440f889e15
https://arxiv.org/pdf/1503.04069.pdf and
http://proceedings.mlr.press/v37/jozefowicz15.pdf
https://arxiv.org/abs/1706.03762
https://arxiv.org/abs/1810.04805
http://jalammar.github.io/illustrated-bert/
https://github.com/google-research/bert