A brief write-up about research overview in Deep Learning

Partha Pratim Saha

To share key aspects of some recent research techniques in Deep Learning and Natural Language Processing (NLP).

1 Overview with exploring effectiveness

Deep Natural Language Processing (D-NLP) focuses on how deep learning techniques help us to tackle different NLP tasks. The main goal is the analysis of text morphology, phonology, contextual understanding of the different comments, and semantics to extract relevant knowledge. Examples of such applications are machine translation (e.g. English to German), sentiment analysis, identifying mood & intention of the speaker, polarity detection in social network comments (e.g. twitter or reddit).

2 Research direction

I will review various trends and techniques of Deep learning and NLP in general.

2.1 Transformer[1]

A recent trend is a multilingual machine translation system with varying sequence to sequence modeling using an attention mechanism that takes key, value and query metrics to generate a probability score. Given all the words in a sequence, it finds if some words are semantically closer than others. The model architecture uses a stack of encoders and decoders where the bottom most encoder takes input sequence, top most encoder outputs a context vector of that input to all the decoders which gives the final translated sequence as shown in Figure 1.

Each encoder component is a combination of a self attention layer, followed by a feed forward neural network whereas each decoder is a collection of self attention layer, encoder-decoder attention layer followed by feed forward neural network. The base transformer applies 8 parallel heads to implement a multi-headed attention mechanism. Its effectiveness is evaluated using the BLUE score. One issue of the transformer is that it does not know which words will be asked to predict or which have been replaced by random words.

2.2 Bidirectional Encoder Representations from Transformers (BERT)[2]

This technique takes an encoder stack from the transformer and applies pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers to implement masked language modeling to predict the next token of the sequence to overcome the drawback of the transformer. During masking, 80% words will be predicted by the model, 10% tokens will be replaced by random words whereas the remaining 10% masking will be kept as it is. It

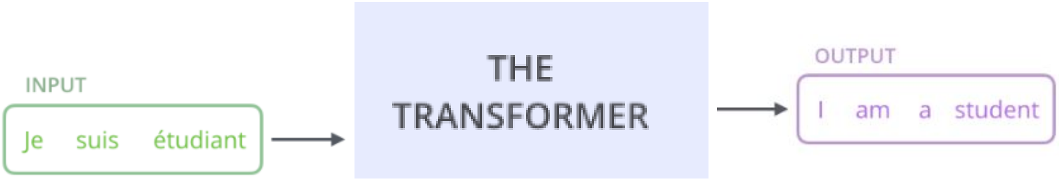

Figure 1: Transformer visualization example

essentially uses a single output layer based on representations from the last layer to compute token level probabilities. Because of the bidirectional system, the prediction becomes context sensitive which increases model accuracy. BERT uses classification [CLS] token to start a sequence and a separator [SEP] token to end one sequence so that between two subsequent sequences there is one separator token for tasks question answering and language inference.

2.3 Clinical-BERT[3]

We can implementation BERT in the healthcare domain. The corpus must have a large volume of clinical notes containing discharge summaries and medication reports of treatment for a patient. This technique predicts the probability of a patient taking readmission after discharge within the next 30 days using laplace smoothing. It captures a final classification for this task. The evaluation is done on Precision, Recall and F1 score.

2.4 Bio-BERT[4]

Another extension of BERT towards text mining for three biomedical tasks: Named Entity Recognition, Biomed Relation Extraction, and Question-Answering. Using word piece vocabulary. Biobert uses wordpiece tokenization which mitigates out-of-vocabulary issues as without it the model might not work well for domain specific Named Entity Recognition (NER). They have used subword embedding, so that when any new terms appear, the model will drill down into multiple existing entities preceded by 2 hash(X). The evaluation is done on Precision, Recall and F1 score.

2.5 ELMo[5] and GPT[6]

Though BERT, ELMo, OpenAI GPT have a lot of real world impact, however, these are not sufficient for real-life NLP applications, e.g., multilingual sarcasm detection. Because BERT is majorly trained on BookCorpus and Wikipedia Data, GPT is focused on Book corpus where ELMo is pre-trained on text corpus. None of them are trained on speech data. In addition to that, BERT predicts [Mask]-ed token bi-directionally to advance on language modeling domain, GPT focuses on left-to-right like transformer, whereas ELMo generates contextual embedding using Bi-LSTM network.

3 Conclusion

For applications like detecting true sarcasm, we may need to build models aligned towards two or more meanings of each sequence: one of for literal meaning, other one is for true meaning and then if the probability of any of these is higher signifies that sequence is of that category. Adding such natural language nuances in intended interpretations is one of of my doctoral research interests.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018.

[3] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. B. A. McDermott, "Publicly available clinical bert embeddings," 2019.

[4] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, sep 2019.

[5] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018.

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, and Sigler, “Language models are few-shot learners,” 2020.