A Quick Summary:
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Original Paper: https://arxiv.org/pdf/1810.04805.pdf

28 Feb 2019

1 Ideas:

(a) The architecture consists of multiple Transformer encoders stacked on top of each other.

(b) The task is to predict masked words, e.g. "Babies drink [mask] for nutrients", where [mask] here is obviously the word "milk".

2 Explanations:

(a) Refer to the summary of "Attention is all you need" for an explanation of a Transformer.

(b) Word tokens were masked randomly accordingly to the following procedure:

  i 80% of the time replace the word with the [mask] token.

  ii 10% of the time replace the word with a random word.

  iii 10% of the time keep the word unchanged (to bias the representation towards the actual observed word)

  As the transformer does not know which words it will be asked to predict, it is forced to keep a distributional contextual representation of every input token.

3 Model:

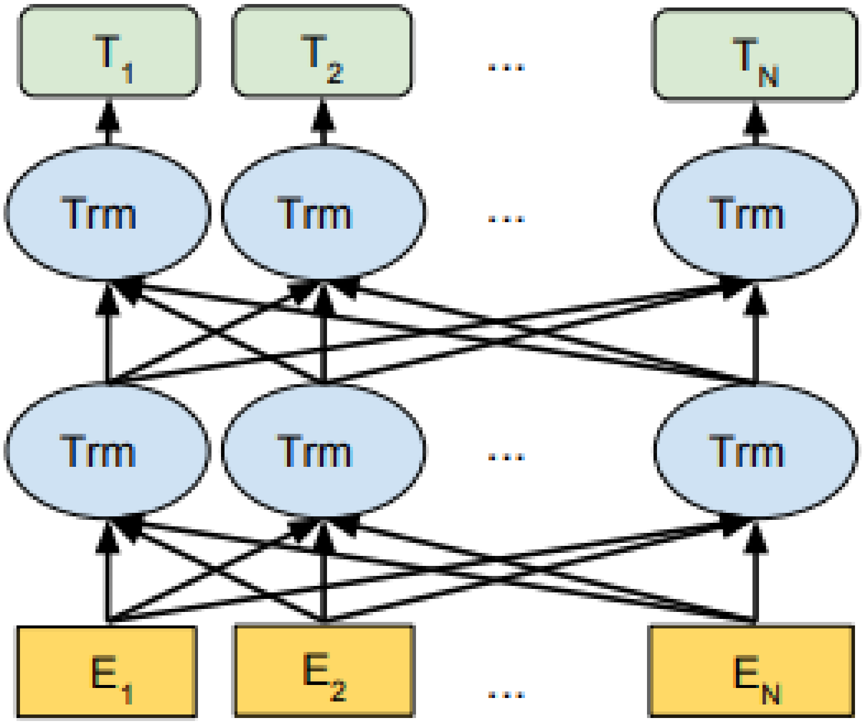

Figure 1: BERT (From original paper)

4 Results:

(a) SOTA for many benchmark tasks

5 Notes:

(a) This seems to be quite interesting. I think it's a springboard to further contextual understanding.

(b) What would happen if we tried to use this with states of stories? For example, if I gave a sequence of words:

`Tiger` → `Rabbit` → `[mask]` → `Tiger eating Rabbit`

will the model be able to recognize the missing state (`Tiger Hunting Rabbit`) via greater contextual understanding?