Paper 3: BERT

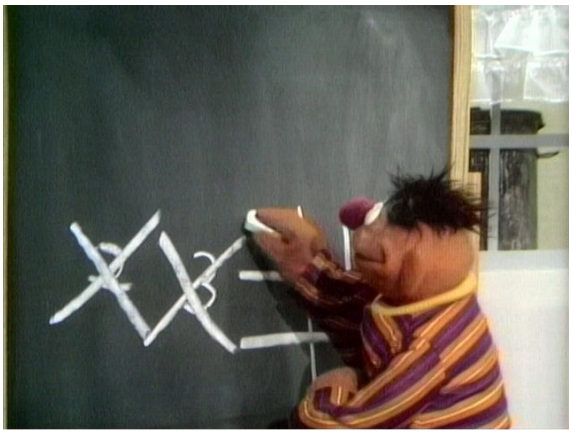
tl;dr: we do the calculations, you simply use them

paper: https://arxiv.org/pdf/1810.04805.pdf
easier explanation: https://jalammar.github.io/illustrated-bert/
Kaggle code: https://www.kaggle.com/mateiionita/taming-the-bert-a-baseline

What does it do?
BERT transforms text in a meaningful way. In other words, BERT delivers us features.
How does it work?
What happens is that BERT takes in a token sentence (e.g. „play“, „##ing“, „[SEP]“). This input is represented internally as a sum of token, segment and position embeddings (consider embeddings as vector representations). It then pushes these inputs through a number of bidirectional transformers (transformers basically translate some input into some output, rather than make a decision - treat them like translators) and returns a vector for each input. It also returns a special classification output vector „[CLS]“ that can be used for, you guessed it - classification. We can then supply it to another ML algorithm in order to teach it and get results.
How does it learn?
Two ways. First: guessing game. Say, you have a sentence „my dog is hairy“. The task is for BERT to figure out 15% of the words in a sentence - in this case, „hairy“. How do you hide that word?

- 80% of the time, you replace it with „[MASK]“, e.g. „my dog is [MASK]“
- 10% of the time, you replace it with a random word, e.g. „my dog is apple“
- 10% of the time, you do nothing - that is so that BERT tries to fill in masks with „hairy“

Second: do sentences follow each other? BERT can actually accept two sentences, concatenated with „[SEP]“ tag. We supply BERT with a set of pairs - 50% of them follow one another, 50% do not (these ones are picked at random). This apparently makes BERT work way better.
Does it work?
On average, 4.5% better than other methods, across multiple NLP tasks.
How can we use it?
To get useful features we do not understand, yet work very well.
Why is this paper so short?
Because for us, BERT is basically a huge, black box. We don’t know how it works or what it returns, but when you put it to use with a neural network, the results tend to be great.
Can I get a Bert strip here?
There you go.



Ernie tries to prove that
mathematics is a hoax