Paper Summaries

Paper Summaries

By James Camacho

Attention *Is* All You Need

The authors use attention to vastly improve language models using far less compute. The idea of attention is certain words in a sentence are much more related to certain other words, and it would be useful for a neural net to be able to pick up on that. Their attention mechanism is almost like a dictionary for arbitrary keys. They have a bunch of (key, value) pairs and query the dictionary with some query. It then adds together the values multiplied by how similar the query and corresponding key is (which is determined by dot product or another neural net). Of course, you need more than one attention "head," maybe one to keep track of the subject, another the verb, etc., so they use "multi-headed" attention. One benefit of these attention heads is they might make interpretability much easier.

It works extremely well, gaining several BLEU points against previous state-of-the-art models. However, it does have a few weeknesses. First of all, you cannot have attention for entire books. The model would either be way too large to store/train, or restricted to looking at a few paragraphs at a time, but the latter would mean the neural net has completely forgotten about the beginning of the book by the end. Second, language isn't necessarily forward thinking. They restrict the neural net from seeing "future" tokens in the decoder, but writers skip words/sentences/paragraphs all the time and come back later to fill in the gaps. Finally, attention is extremely inefficient. Trying to rediscover the relevant key/values is much more computationally complex than keeping track of them the entire time.

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

The major difference in the model is it is bidirectional. This solves one of my issues with attention above. The rest of the paper talks about how they train the bidirectional model, as it's almost exactly same structure as the transformer model from "Attention Is All You Need." They train the big model on two things:

”fill in the blank” and ”what is the next word?” For the former, they mask 15% of the ”words” (sentence tokens) with either a random token or a [MASK] token. All the model trains on is finding the correct word, as it should trivially know the rest of the words in the sentence. For the sentence pair task, they take pairs of sentences and 50% of the time replace the second sentence with a random one. The goal is to construct the correct second sentence.

The structure of BERT makes it easy to fine-tune it. Once they trained the base model, all they had to do was add a few layers and fine-tune it on a bunch of different tasks, and it achieved top scores in practically every language model metric. Feature extraction worked similarly well (the difference between the two is you don’t modify parameters in BERT if you’re only extracting features).

I think the largest weakness in this paper is how uninnovative it is. It was quite a slog to read through all the details of training and fine-tuning when the whole new concept could be summed up in two words (bidirectional transformers). With the actual model, I think a weakness is training inefficiency. It seemed they spent a large time talking about training because of how difficult it was to get it to train correctly. Also, the masking/sentence randomization seems like it would ”confuse” the neural net during training, making the gradients change unpredictably.

Unsolved Problems in ML Safety

This paper summarizes many issues in ML safety and provides ideas for how to resolve them. Some of them, like anomaly detection and adversarial robustness, already are being figured out. The issue right now is adding anomaly detection or robustness is pretty bad (<70%) and makes models perform worse, so there is little motivation to do so.

Many of the issues are relatively small. E.g. backdoors in ML could let a thief escape security footage, but really isn’t much worse than any other kind of backdoor. Some are a little more worrying, but being worked on: How confident is the neural net in its output? How truthful is it in its output? Interpretability is a very young science, and the more advanced the models the more important it is. I do feel like most researchers have very good reason for working on interpretability—after all it would help them make more accurate models if they can understand the big black box—so I don’t think this is too large an issue.

Probably the most dangerous issues are on capabilities and alignment. It’s practically impossible know an AI is aligned with human morals/objectives, which to be fair is also true of other humans. The only issue with large AI models is they can be much more capable than any human. If a doctor prescribes the wrong medication to make an extra buck, they can be caught eventually. A capable enough AI could hide their tracks well enough to never get caught.

One thing this paper didn’t mention was keeping NN models proprietary. Some motivations could be: it takes a lot of money to train a NN, you want to beat out competitors, or it makes it harder to perform adversarial attacks if you don't have the model. Unfortunately, it's not so simple. Even without the model you can recreate a very similar one by just looking at how your inputs affect its outputs. In fact, researchers currently use this to downsize large models, decreasing overfitting and making them run faster. One way to fix this is limiting access to the model, but then it can't be nearly as useful.