Week 5 Paper Summaries

Joseph Camacho

July 27, 2022

Attention Is All You Need

Earlier methods of using attention combines attention with recurrent neural networks (RNNs), which bottlenecks parallelization (since every timestep requires the previous). This paper, titled "Attention Is All You Need", proposes to speed up training by only using attention (which is massively parallelizable).

The Transformer

The transformer is their basic building block. It consists of an attention block combined with a fully connected feed-forward block, with residual connections in between them. In this paper, the attention block is multi-headed, meaning that they have many copies of identical looking attention blocks that are concatenated together right before the feed-forward block.

Masking

It's important to mask future outputs in the decoder so the model actually has to learn to predict the future. In this paper, that's done by setting to $-\infty$ all those values right before the softmax, effectively zeroing their contribution.

Feed-forward Networks

The feed-forward networks are fully connected *per position*. It's like a 1x1 convolution. They do not transfer information between different positions (i.e. words). This is important in the decoder so that early words can't glimpse the future.

Results

Their results are the next best (though only slightly better at En-Fr) at translating both English to German and English to French. Their training costs are significantly less on the English-French translation, and about half of the next best model for English-German.

Judging:

Weakness 1: It's so inefficient! In a single attention block, only a small fraction of the values actually matter. Instead of using a softmax, why can't they just us a max (or k-max)? This would definitely make backpropagation faster, and probably would also speed up the forward step.

Weakness 2: Multi-headed attention seems like an unnecessary complication. See "Single Headed Attention RNN: Stop Thinking With Your Head", for a rather (informal) example of where single headed attention is comparable to multi-headed attention.

Strength: Attention is useful in tons of applications other than just language processing. I know, for example, that they've been used in image recognition & captioning.

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

Much of the first four papers is the authors' insistence that their paper is unique because they "do not use traditional left-to-right or right-to-left models". Indeed, they almost appear to try to hide evidence of others' use of bidirectional methods in order to make their work seem more innovative, such as by barely citing Mikolov et al.'s paper about word2vec at all in their "Related Work" section, in stark contrast to their paragraphs and paragraphs of text about other, left-to-right models.

The authors then briefly touch on their what their "innovation" actually is: Instead of masking all future words, just mask a random fraction of all the words and predict what the values behind the masks are.

The authors then spend a while discussing their experiments with fine-tuning their model. They do much, much better than the next best model (most often OpenAI GPT) on a wide variety of tasks, and are even superhuman on one.

They conclude by stating, yet again, that their method is so innovative because it is "*bidirectional*" (their emphasis, not mine).

Weaknesses

- Their work is hardly innovative.
- Their choice of replacing the `[Mask]` token with random tokens 10% of the time is arbitrary, and any gains are in large part due to randomness. Their chosen model's results in the table in Appendix C.2 all fall within two standard deviations of the other models' results.
- There are basically no new problems that BERT solves. It just does a better job at solving problems other neural networks have already mostly solved.

Unsolved Problems in ML Safety

Robustness

Black Swans

The unexpected and unusual happen all the time. How can we make sure ML models don't crash when they encounter such? This paper proposes data augmentation, having model-environment feedback loops, and having unusual but useful datasets.

Adverserial Robustness

Right now, most ML systems are super vulnerable to adverserial attacks. How can we train them to be robust to them? Some ideas include blackboxing neural networks (making it much harder to train an adversary against them), looking for discrepancies through the use of multiple sensors, and having evolving defenses.

Monitoring

Anomaly Detection

Few ML models are designed to detect anomalies. However, this could be incredibly useful to determine if an adversary is attacking it or to determine if humans should take a closer look at what is going on.

Past research in anomaly detection includes out-of-distribution detection, open-set detection, and one-class learning. Anomaly detection right now suffers from determining if noise is random or anomalous.

Representative Model Outputs

Communicate uncertainty. Communicate honsetly.

Hidden Functionality

"Backdoors" exist in most ML systems, where particular circumstances can lead to detrimental results. They differ from adversarial examples because they are inserted by bad actors at training time, not test time. One way of making a backdoor is by "poisoning" images/text/data on the internet with your own data. Getting rid of hidden functionality will likely be a constant competition between white-hat and black-hat ML hackers.

Alignment

Okay, I'm done writing so much in a "summary". From now on, only the biggest headings get text.

This section consists of lots of opaque/circular/vague methods to solve problems that may come up in the future regarding alignment. E.g., the authors write, "To get a sense of an agent's values and see how it make tradeoffs between values, researchers could also create diverse environments that capture realistic morally salient scenarios and characterize the choices that agents make when faced with ethical quandaries." Why couldn't the authors just write, "Researchers could give the agents ethical dillemas to test their morals?"

Systemic Safety

Machine learning can eventually be used to hack, and this is dangerous and needs to be countered with machine learning. We also need to make sure our ML models aren't hacked, because (1) they could be dangerous in the hands of the public and (2) they could be subverted to give incorrect answers.

Weaknesses

- This paper doesn't really address any *actual* solutions to any of the problems it presents. I guess that's why it's titled "**Unsolved** Problems in ML Safety".
- There's much less of an emphasis on securing models that could be dangerous in the hands of the public than there should be. E.g., The weights for a model that can create RNA sequences from desired protein folds should probably not be published on Github for any bad actor to download and design superviruses. "Unsolved Problems in ML Safety" makes no mention on restricting such information.