Paper summary, week 5

David McSharry

1 Attention is all you need

This paper describes a new paradigm in the field of NLP, namely by removing sequential processing of data (like would be present in RNN and LSTM architecture) and instead more heavily relying on attention. One key issue with RNNs and LSTMs is that they process language in a fundamentally sequential way, and so parallelization is not feasible and so these architectures have limitations. Encoders and decoders work by translating natural language to a series of representations and then back again. Attention allows a model to make connections between separate parts of a text, which the transformer will use to make connections between it's particular abstractions in any given text. The innovation of transformers is most clearly seen in the fact that it becomes more complex linearly with the input text length rather than quadratically (like RNN or LSTM). This architecture also reaches LLM benchmarks in a faster training time than previous architectures. A non trivial strength of this paper is that it employs the "bitter lesson", by removing the sequential constraint and instead focusing on allowing models to come up with their own abstractions of language seems like a key innovation of this paper to me.

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

This paper introduces a new LLM architecture called BERT, which in a nutshell allows the aforementioned representations of language to flow in both directions, as opposed to strictly from left to right as in the case of previous LLMs like GPT. This for example, improves fine tuning compared to GPT which has to update all pretrained parameters. BERT instead masks some of the pretraining data set and tries to predict the masked section. This allows BERT to combine the left and the right flow of representations and thus "allows BERT to read in both directions" to anthropomorphise. BERT takes advantage of transfer learning by first being trained on unlabeled data, and then being fine tuned while taking advantage of it's bidirectional data flow. BERT performs better than other transformers and LLM on several benchmark NLP data sets. Because BERT must contextually determine words as part of it's training process, this extra contextual processing leads to comparatively higher computation costs compared to other LLM, and this should be taken into account when exploring the applications and potential of BERT.

3 Unsolved Problems in ML Safety

This paper explores ML safety and defines some key concepts that provide a good framework for thinking about some unsolved AI safety problems, namely robustness, monitoring, alignment, and systemic safety. Robustness refers to an AIs ability to act predictably and reliably in tail risk events. A case is made that a considerable amount of expected negative impacts of AI could be contained in such unlikely events, so we should be taking them seriously. Stress testing and testing on a wider variety of scenarios is an important direction to go in to explore further the robustness of AI systems. Monitoring refers to the fact that deploying large AI models will have to be an extremely careful process to avoid unexpected consequences. Organisations will have to prove their reliability and deployment will have to be managed by such highly reliable organisations. Making models be able to assess their domain of competence could be a good way to improve the reliability of those tasked with monitoring models as it will not purely be left up one person whether or not deploying an AI to an application will lead to unexpected consequences. Backdoors could also allow those monitoring to have deeper insights into the AI, and have some more control over the AI once it is deployed. Alignment refers to the goals of the AI, and how we have not yet solved how to instill out goals and values in AI. This is extremely important as a misaligned AI that wants something different to us will take actions that could impede our pursuit of our goals and even actively perform actions that run completely contrary to our goals. Some specific problems include the fact that goals are difficult to define properly, the can be difficult to optimize for, and proxy goals can often turn out to be misaligned in out of training distribution events. Systemic safety refers to the problem that AI systems could potentially be altered or tampered with in a manner that would lead to undesirable or unexpected consequences, for example by cyber attacks. AI cybersecurity is therefor an important field that we have thus far not solved. The fact that AI are "in good hands" now does not guarantee that this will continue to be the case forever, so this must be accounted for. A combination of solutions to all of these four problems are needed to ensure the safety of AI systems. One nontrivial problem with this paper is also it's best feature in my opinion, it provides a very good framework for thinking about currently known problems in AI safety. This is good for spreading awareness, but could also induce a myopia to a reader and would pen them into thinking about problems in AI in this nice framework so they might miss other problems that fit less nicely into this framework.