Paper review for "Bender, Gebru et al., On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?"

EECS 598 Paper Review – Week 13 - Changyuan Qiu

Large-scale pre-trained language models like BERT, GPT-2/3 and their variants have dominated NLP by extending the state of the art in a wide series of downstream tasks measured by canonical benchmarks like GLUE, SQuAD, and SWAG. However, while the enormous amounts of the uncurated, Internet-based data available on the web has enabled these large-scale language models to achieve high accuracy on specific benchmarks, the data could encode and reinforce hegemonic stereotypical and derogatory bias along all aspects of social ethics. And this paper asked the question of "Can language models be too big?" and looking into whether enough consideration has been put into the potential risks associated with the scaling of language models and strategies to mitigate these risks.

Here the author identified a wide variety of costs and risks associated with the rush for ever larger language models (LMs), including: (1) environmental costs (training a big Transformer with neural architecture search as one in the paper *"Attention is All You Need"* emitted 284t of CO_2 while the average human is responsible for an estimated 5t CO_2 per year, and training a single BERT base model without hyperparameter tuning on GPUs was estimated to require as much energy as a trans-American flight); (2) financial costs, which in turn erect barriers to entry, limiting who can contribute to this research area and which languages can benefit from the most advanced techniques (an increase in 0.1 BLEU score using neural architecture search for English to German translation results in an increase of $150,000 compute cost); (3) opportunity cost, as researchers pour effort away from directions requiring less resources; and (4) the risk of substantial harms due to the biased uncurated Internet-based dataset (authors justified this with 4 aspects: 1 - unevenly distributed Internet access, 2 - moderation practices that limit underrepresented populations to add data and share thoughts on use-generated

content sites like Reddit, Twitter and Wikipedia, 3 - data in minorities' fora is less likely to be included in the process of crawling training data, 4 – current practice of filtering datasets tend to attenuate the voices of people from marginalized identities), including stereotyping, denigration, increases in extremist ideology, and wrongful arrest, should humans encounter seemingly coherent LM output and take it for the words of some person or organization who has accountability for what is said looked into.

In order to mitigate these potential risks, authors urged researchers to carefully weigh these risks before starting to build either datasets or large-scale systems trained on datasets. More specifically, they advocate for making time in the research process for considering environmental impacts, for doing careful data curation and documentation, for engaging with stakeholders early in the design process, for exploring multiple possible paths towards long-term goals, for keeping alert to dual-use scenarios, and finally for allocating research effort to harm mitigation in such cases. In addition, they call on the field to recognize that applications that aim to believably mimic humans bring risk of extreme harms, where downstream effects need to be understood and modeled in order to block foreseeable harm to society and different social groups.

Compare with prior works about AI ethics, this paper come with many solid examples as supporting evidence (like they covered details about how the filtering process of GPT-3's training data leads to bias), and it presented a very in-depth and structured analysis of potential risks about scaling of LMs. Regarding limitations of this paper, I thought that the authors could give more comprehensive comparisons of large-scale LMs and previous small-scale counterparts in Section 6: Stochastic Parrots to give a better sense of the risks from deploying large-scale LMs.