R15/Sarthak Mittal/200050129

April 5, 2023

This paper discusses the risks and consequences of developing ever larger **Large Language Models (LLMs)**. The authors analyse the environmental and financial costs and suggest using resources for datasets, carrying out pre-development exercises and encouraging research beyond LLMs.

Among recent works, there seems to have been some competition in producing larger LMs. Though the **risks** associated with these developments have not been prioritized. There are communities that **do not benefit** from LMs but are affected by the **adverse impacts**. The datasets also introduce difficulty in training, which can be reduced by considering data that can be **documented**. The focus should be shifted from state-of-the-art results to a **deeper understanding** of the tasks. LMs are also susceptible to **derogatory language**.

The authors define language modelling as predicting the likelihood of a token, such as a character or a word, given its surrounding or preceding context. In recent years, LMs have progressed from **n-grams** to **pre-trained representations** (using word embeddings) and more recently **transformer** models such as GPT-3. There are also **multi-lingual LMs**, though there is a need for more inclusion of under-represented languages.

The authors discuss that training LLMs incurs significant environmental and financial costs. Estimates from a recent study indicated that training of a single large transformer model emitted **284t of $\mathbf{CO_2}$**, which is equivalent to the energy used in a trans-American flight. There is a need for **energy-efficient** architectures and the use of efficiency as a metric, and even more so because some parts of the population experience more environmental impact than deployment benefit.

Significant resources must be allocated to **data curation** and **documentation**. Due to the **abundance of data** available online, deep learning (DL), natural language processing (NLP) and computer vision (CV) models have achieved high accuracy in specific benchmarks. Though the data also encodes **stereotypical** and **derogatory** language, resulting in biases. Large datasets such as Common Crawl also do not broadly represent views of the world. The model is more likely to retain some characteristics that are **over-represented** such as younger users and developed country citizens. The **filtering** of datasets further reduces the lower-represented identities due to the removal of harsh words.

There is a distinction between tasks for LMs and tasks for **Natural Language Understanding (NLUs)**. The benchmarks lead to efforts being directed to them instead of using **meaning-capturing** approaches or exploring **more effective** ways with carefully curated data. The abilities of LMs need to be characterized, and the commitment to performance could lead us further from the goal of **General Language Understanding (GLU)** systems.

The authors then discuss the harms of deploying LLMs like GPT-3 in the real world. Such LMs may encode **"hegemonic views"**, amplify and cause automation biases and misuse the views. They explore why humans consider the text of an LM to be **meaningful** and point out the risks of using embeddings in other tasks. The **"coherence"** of the output is only apparent to the reader, and it lacks **communicative intent**. The comprehension of the output is "illusory" because of our singular understanding of the language ourselves. LMs are labelled as **"stochastic parrots"** because they **stitch** sequences of commonly observed forms according to training and probabilistic data but do not use the underlying **meaning**.

The authors suggest that the risk posed by LLMs can be controlled through careful **planning** and **evaluation** before setting up datasets and systems. They suggest that researchers evaluate **energy efficiency** and assemble datasets that suit the **overall direction** of their field and the socio-technical impact of their tasks. **Negative effects** of the model and its misuse should also be identified, such as for a "value-sensitive design", both **support and harm** should be discussed, and LLMs explicitly catering to the "marginalized populations" should also be considered.

In conclusion, the authors discuss the rush towards large LLMs in NLP and examine the costs and risks associated with the direction of current research. Though there has been significant progress, the side impacts on the environment, financial issues, opportunity costs as well as substantial harm (stereotyping, derogation, extremism) also need to be considered. NLP researchers are requested to use their resources for techniques that are **effective** instead of being **data-hungry**. They think that a **scholarship** for benefits, harms, and risks of replicating human behaviour, and **thoughtful design** of sufficiently concrete target tasks, could help to motivate researchers.