EUROPE'S TECHNICAL DEBT: WHY WE NEED WEB SEARCH IN THE AGE OF GENERATIVE AI

Malte Ostendorff, Pedro Ortiz Suarez, Julian Moreno-Schneider, Georg Rehm, DFKI GmbH, Germany

INTRODUCTION

Generative AI has completely changed the landscape of machine learning, allowing researchers and practitioners to tackle tasks thought to be impossible. The proliferation of generative AI models revolutionises the way we process and use information. However, this age of rapid technological innovations is dominated by US-based enterprises. Europe is lagging behind in developing large AI models and is expected to have a hard time catching up. One of the reasons is the *technical debt* that Europe has been accumulating since it lost and stopped competing in the race of the previous technological revolution – Web search.

Progress in generative AI is mainly driven by two factors: computational power and data (neglecting algorithmic improvements). Despite the fact that US-based cloud providers currently possess the lion's share of computational power, this does not pose a major hindrance for European AI developments. Europe is actively investing in its compute infrastructure, reallocating resources, and making them accessible for AI research, such as through initiatives like the EuroHPC Joint Undertaking. The real issue lies in the deficiency of the second key ingredient – Web data and its retrieval, which can be attributed to the absence of strong European Web search initiatives. EU projects such as Open Web Search are promising, though.

WEB CRAWLS FOR PRETRAINING

Large language models (LLMs) and other generative models are statistical models based on training data. This training data is crucial for the success of any AI model, e. g., affecting the language capabilities, biases, and cultural representations. Given the increasing size of these models, larger training datasets are required. The most prominent source that provides data at the scale required is the Web, accounting for a significant portion of the training data for recent LLMs, often more than 80%. Especially Web data from Common Crawl, or processed versions such as OSCAR, is widely-used for LLM training. This reliance on Web data introduces several limitations, especially in the European context. Web crawls from Common Crawl are only a sample of the whole Web, i. e., important European websites might be omitted. Also, since the Common Crawl crawler operates with a US user-agent and an IP number located in the US, the crawler appears to websites as a user from the US. As a result, English language content represents the largest share of Common Crawl data by far (30%). Web data is generally unbalanced in terms of included languages, which further increases the technological gap between English and other, more multilingual regions. To address these limitations, a European Web crawl is needed to collect a training dataset that adequately covers Europe's diversity including its languages, countries, and cultures. While there are already ongoing projects and initiatives working on this or related problems, we need to significantly strengthen them to obtain valid extensions to Common Crawl. Something as crucial as the training data of AI models should not solely depend on a single Californian non-profit organisation that operates on AWS-donated infrastructure. From one day to another, AWS may throttle Common Crawl's bandwidth and hence delay European AI research projects, as happened recently.

WEB-BASED LLM AUGMENTATION

Generative models have severe shortcomings. Among others, the enormous training costs prohibit frequent retraining on new data. Thus, the "knowledge" encoded in LLMs can become outdated quickly. For instance, ChatGPT's knowledge cut-off date is September 2021. Moreover, LLM output may contain factually incorrect information ("hallucinations"). One promising approach to address this issue is augmenting LLMs with retrieval systems. The idea is to retrieve factual and updated information from trustworthy sources and then let the LLM generate output based on the retrieved information. One example is Microsoft's Bing chatbot that retrieves information from the Web. As with pre-training, the Web represents the most extensive resource from which information can be retrieved. However, building a retrieval-augmented LLM obviously requires Web search, making it, once again, quite difficult for Europe to compete since no European Web search exists. Relying on Web search APIs from one of the big technology enterprises is no valid option either since it would introduce a strong dependency, hampering technological sovereignty. In fact, Microsoft tripled the prices of the Bing Search API briefly after the introduction of their own retrieval-augmented LLM. Thus, there is a pressing need for European Web search APIs to build the next generation of retrieval-augmented AI models.

CONCLUSIONS

Europe's lack of investment in Web search infrastructure, crawling and retrieval, has led to a significant technical debt. To be able to compete in the next technological revolution, generative AI, this debt needs to be paid off. Although catching up is feasible, it requires a collective effort and substantial investment from industry and academia.

ACKNOWLEDGEMENTS

The work presented in this paper has received funding from the German Federal Ministry for Economic Affairs and Climate Action (BMWK) through the project OpenGPT-X (project no. 68GX21007D).