30 May 2022

Frontier Talk

Memo by Marco Almada

Foundation Models in AI: what impact for policies and law?

Speaker: Matthias Gallé (Naver Labs Europe)

This event has been organised by the Technological Change and Society Interdisciplinary Research Cluster

Matthias Gallé *is an AI researcher with a background in Natural Language Processing (NLP). He heads the "AI for our Digital World" group at Naver Labs Europe (Grenoble)—Naver is a leading Korean tech company, which is the leading actor in the search market in South Korea but has activities in other sectors.* ***Gallé*** *is also involved in the BigScience project, a large open-source initiative to train a large, multilingual language model.*

Gallé began his talk by pointing out that he is not an expert on law or policy. Instead, he comes from a computer science background and wants to raise the issues he has seen in his work to an audience of lawyers, social scientists, and policy scholars. For that purpose, he divided his talk into two parts: an introduction of the technology behind current AI systems, followed by a policy discussion.

The presentation covered four main topics. First, Gallé addressed why the machine learning community believe foundational models are a big deal. Then, he argued that foundation models are a natural evolution of the machine learning research agendas from the past decades. The third part of the presentation showed the most recent advances in foundation models. Finally, Gallé covered some of the policy implications of these developments.

To open the discussion of foundation models, Gallé provided an earlier example of machine learning advancement. In 2016, the AlphaGo model managed to defeat Lee Sedol, a leading player of Go. This was one of the breakthrough moments of deep learning, after deep learning systems achieved super-human performance in image recognition and translation tasks. Drawing from his experience working for an Asian company, Gallé pointed out that the victory in Go—a game which is seen as requiring "intuition"—has been a wake-up call for AI research in Asian countries, as shown by the boon in Chinese patents on deep learning technologies in the years since. The US and, to a later extent, Europe have also seen considerable growth in the same period.

These advancements are grounded on large-scale neural networks with billions or even trillions of parameters. These parameters give the model flexibility to adapt to specific tasks, especially when coupled with non-linear functions. What is striking about these recent models is that they provide very surprising outputs. Imagen, an image generation model, achieves a considerable level of compositionality (see image). Coming to language models, Google's Palm can provide explanations of jokes that rely on double entendres, looking at parts of language and situating them in context.

Such results are impressive, unexpected, but quite useless. They are very generic at the moment, but Gallé conjectured that we will likely see their application in the next few years—probably months—towards specific problems. However, these large language models do not have much of a direct market value at this stage.

A EUI Professor then pointed out that some applications, such as chatbots, might benefit directly from the capabilities afforded by foundational models. Gallé replied that these models produce unexpected outputs, which generate an element of surprise that is not desirable in many applications. However, Gallé acknowledged that this creative aspect of current foundation models may already be interesting for applications such as interactive fiction but requires more work for applications such as bureaucracy. Nevertheless, he expects that in the next few years, we will see these models being used in various contexts, with specific adaptation layers tailoring the general capabilities of a foundation model to the particular application one might want to use the model for.

Gallé introduced the history of the term "foundation model", coined in a Stanford paper. "Foundation" does not refer to a revolutionary breakthrough. Instead, these systems are meant as a platform for future development. The term "foundation" has been subject to much discussion in the community, and Gallé suggested that "platform" might be an excellent description of what these models offer. There is a strong belief in the machine learning community that research in such models still has much to offer.

The earlier founding moments of NLP as a discipline came in the 1950s. Not only the usual suspects came into play—Alan Turing's 1950 paper on computing machinery and intelligence, the 1955 Dartmouth seminar in which the term "artificial intelligence" was coined—but some earlier work, such as the Georgetown–IBM machine translation

experiment in 1954. From the 1950s to the late 1960s, AI research relied on single-layer linear neural networks for classification. However, in the late 1960s, it was pointed out that such a model could not learn how to reproduce a relatively simple function—the XOR logical function—and this negative result led researchers away from neural networks for decades.

From the late 1960s to the early 1990s, research in NLP—and in AI more generally—was dominated by symbolic systems. In these systems, the machine does not learn language rules from the data. Instead, programmers codify a formal representation of the knowledge available to a human expert in the application domain. For example, language models would rely on grammatical trees and other structures. The system would then apply this formalized reasoning to the problem at hand, but its rules would stay static. Expert systems like those found some success in specific fields, such as medicine, but they are not very common nowadays. Gallé mentioned that this scarcity of expert systems stems, at least in part, from their difficulty in dealing with noisy data. Applying an expert system requires neatly organized data, but this data is hard to come by in the real world, particularly in language. For example, depending on the context, the word “bank” may refer to a financial institution, a slope or a place of storage. Given this limitation, much research from the mid-1980s to the late 2000s, much effort was dedicated to combining symbolic and machine learning approaches. In particular, *feature engineering* became a critical problem.

After 2012 or so, deep neural networks gained prominence. Instead of using a single layer, like the old-school neural networks, these networks relied on various specialized layers of functions. In addition, each of these layers used non-linear functions, increasing their versatility. These innovations, and the gains in computational power, meant one no longer had to spend much time with feature engineering. Now it would be possible to feed a system with raw data and let the various layers of the neural network find out what is relevant. This has been a boon to natural language processing, as one could rely on much text without much pre-processing. Gallé invoked what has been called “the bitter lesson” of machine learning: generally, a model with more data and computing power will perform better than a fine-tuned model that incorporates human knowledge and feature engineering. However, as a participant pointed out, there is always some implicit feature selection involved in model design and selection of data, though Gallé points out this tendency is mitigated by the kind of comprehensive models discussed later in this talk.

Three factors underpin the current boon in deep learning technologies. The first one is the availability of compute, that is, software and hardware resources that can be used for making the computations involved in these networks. The second factor is the availability of large data sets. Finally, open-source libraries are also available that provide high-level

frameworks for creating deep neural networks. By relying on those frameworks, software developers do not have to worry about implementation details and can specify neural networks at a very abstract level. Reliance on these libraries reduces considerably the work involved in implementation, allowing for the spread of deep learning solutions.

Foundation models represent the next step in this development. In the case of natural language processing—the core of our discussion—a foundation model is a very large model trained on a large set of unlabelled data. For example, one can simply feed a model with an unlabelled corpus of English-language texts and train it to find patterns in the data without teaching the model to classify anything. The result is a large language model, which developers can direct towards specific texts by doing a second training stage. When they train a pre-trained model, developers are fine-tuning it toward a specific task, and they can do so with a relatively small set of annotated data. But, after this final training, the model will also benefit from the statistical patterns it gathered from the larger dataset at a fraction of the cost and time needed for the pre-training.

Over the last few years, AI has moved toward homogenization. In terms of architecture, most top-end AI systems are now deep learning systems, particularly relying on the Transformer architecture. The weights of these models are learned through variations of the same general technique: Stochastic Gradient Descent. And even the learning goal of the training process has been made uniform: these systems nowadays rely on some sort of self-supervised learning, in which they learn patterns from the data without the need for human annotation. This homogenization, which is not limited to NLP, opens up the possibility of building multi-modal systems that can solve problems of various types: for example, the same system may be able to read images and interact with text.
Gallé highlighted five challenges for future ML research:

1. How do we operate these large models to allow people to make the best use of them?
2. How do we understand the operation of these models to ensure transparency and accountability in models with an absurdly large number of parameters?
3. How do we develop models that are efficient in terms of energy usage and drawing inferences?
4. How do we incorporate new information into these models?
5. How do we control these models to ensure they match social mores and legal requirements?

He also pointed out three pressing issues he is often asked about that require perspectives

beyond technical debates. The first one is copyright, intellectual property, and fair use. A strong, recent example is GitHub's Copilot, a system that has been produced by fine-tuning GPT-3, a large language model, on code available at various public software repositories. What is the copyright status of the code produced by such a tool? For common-law jurisdictions, is the output of Copilot, or something like it, considered fair use of the original code? Is the output necessarily open source if it draws upon code licensed with "viral" licenses such as the GPL?

An attendee asks about the financial incentives for actors such as OpenAI to produce large language models. How can they profit from releasing these models to the public? Gallé gives two answers. First, some of these models are developed by people interested in open science, such as the BigScience project initiated by Hugging Face. This project was created because there is no model as big as GPT-3 available to the public, and a coalition of corporations and scientists wanted to reduce the concentration of NLP resources in the hands of the few actors currently able to train such models. The second incentive would be that companies make these models available because they don't know how to profit directly from their large models. So, actors such as OpenAI provide these models as a platform with the idea of letting other actors find a killer application, which would then license the large language model for its continued use.

Gallé pointed out questions associated with the right to be forgotten. Large language models are absurdly good at memorizing data, and recent research suggests that generalization is, in fact, impossible without memorization. There is no surefire way to erase one given data item from the model, especially as it is not known how exactly a specific data item is embedded in the model. So there remains the possibility of extracting this data item. Gallé provided an example: if somebody adds their email address to the comments of a program, can Copilot use this program as part of its training dataset? Another EUI Professor responds that this use might be possible if there is a legitimate interest under the GDPR, especially if the training set and not embedded in the model itself. Gallé stated that he does not know of any methods for removing personal data from a model other than removing the personal data from the training set itself, effectively retraining the model.

In the ethics and transparency part, Gallé pointed out the shortage of work on formal verification of machine learning models, which could offer guarantees about the output of AI systems. There is some work on documentation, in which techniques such as data sheets and model cards provide information about the data and processes used to train a given machine learning model. However, AI is increasingly a topic of attention for regulators in various countries, so we are likely to see more formal requirements, such as those proposed

in the AI Act.

After this initial presentation, a participant asked how far machine learning models can go without incorporating some sort of theoretical reasoning to move beyond correlation. Gallé replies that this incorporation of causal thought is already underway; for example, in debates about explainable decision-making, but that it comes at the cost of system performance.

A PhD researcher asked whether models are likely to continue growing in scale or if we are going to find limits to growth. Gallé points out that growth is unlikely to stop any time soon, but data might soon be a stronger bottleneck than compute. In some languages, all available texts are already used for training, raising the question of how to take training further.

A participant asked about the social implications of the use of models by a broader public, especially in the case of biased models. Gallé acknowledged the issue, mentioning that science has long relied on trust relationships, which are now put into question with the popularization of these models. Other sciences, such as physics after the Manhattan project, have had to deal with similar issues. However, the solutions adopted by these disciplines might not be easily translated to computer science, as language models are much easier to repurpose than research that requires expensive lab equipment. Gallé concluded the discussion by framing this as an open problem.