[On the Opportunities and Risks of Foundation Models]
(Até a seção 1, página 12)
Author: Marcus Vinícius Borela de Castro

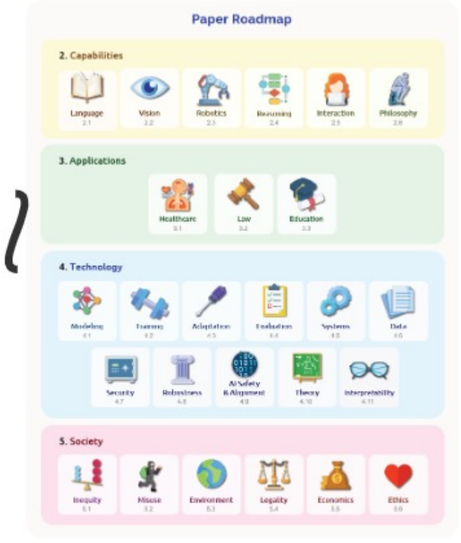
Contextualization

AI is undergoing a paradigm shift with the rise of models (e.g., BERT, DALL-E, GPT-3) trained on broad data (generally using self-supervision at scale) that can be adapted to a wide range of downstream tasks

We call these models **foundation models** to underscore their critically central yet incomplete character.

This report provides a thorough account of the opportunities and risks of foundation models, ranging from their capabilities and technical principles to their applications and societal impact

Paper Roadmap



- Definitions
  - Emergence means that the behavior of a system is implicitly induced rather than explicitly constructed; it is both the source of scientific excitement and anxiety about unanticipated consequences
    - how a task is performed emerges (is inferred automatically) from examples; with deep learning, the high-level features used for prediction emerge; and with foundation models, even advanced functionalities such as in-context learning emerge
  - Homogenization indicates the consolidation of methodologies: learning algorithms (machine learning aspect), architectures (deep learning) and models (foundation model).

AI History

Machine learning era

AI started in the 1990s: rather than specifying how to solve a task, a learning algorithm would induce it based on data — i.e., the how emerges from the dynamics of learning. It still required domain experts to perform “feature engineering”

Deep learning era

Around 2010: Deep neural networks would be trained on the raw inputs (e.g., pixels), and higher-level features would emerge through training (a process dubbed “representation learning”). Rather than having bespoke feature engineering pipelines for each application, the same deep neural network architecture could be used for many applications

Around 2013 (word embeddings): begigning of self-supervised learning technique: tasks depending on unlabeled data designed to force the model to predict parts of the inputs, making them richer and potentially more useful than models trained on a more limited label space.

Foundations models era
By the end of 2018: a model is trained on a surrogate task (often just as a means to an end) and then adapted to the downstream task of interest via fine-tuning (transfer learning).

Transfer learning is what makes foundation models possible, but scale is what makes them powerful. Scale required three ingredients:
(i) improvements in computer hardware — e.g., GPU throughput and memory have increased 10× over the last four years
(ii) the development of the Transformer model architecture [Vaswani et al. 2017] that leverages the parallelism of the hardware to train much more expressive models than before; and
(iii) the availability of much more training data

After 2019, self-supervised learning with language models became more of a substrate of NLP. The acceptance that a single model could be useful for such a wide range of tasks marks the beginning of the era of foundation models.

While the deep learning revolution has fostered open science, the rise of foundation models is reversing this trend, with some models and datasets not being released, and the actual training of these models remaining inaccessible to most AI researchers due to high computational costs and complex engineering requirements.

It difficults research as some functionalities like in-context learning have only been demonstrated in models of sufficient size, so scale is needed to even ask the right questions in models of sufficient size

Risks

The scale results of foundation models in new emergent capabilities, and their effectiveness across so many tasks incentivizes homogenization

The widespread use of a few foundational models, such as BERT, RoBERTa, BART, T5, in Natural Language Processing (NLP) not only amplifies improvements across the field but also risks propagating the same potential biases of these models throughout all AI systems.

The inherent complexity and unpredictability of foundation models (our current limited comprehension of their functionality, failure points, and potential capabilities due to their emergent properties), make their broad adoption risky, thus making risk mitigation a key challenge in their further development, especially from an ethical and AI safety perspective.

The multifaceted social impact of foundation models encompasses potential social inequities, economic and environmental effects, risks of disinformation, legal implications, ethical issues from homogenization, underscoring the challenge of responsibly anticipating and addressing these concerns given the models' versatile nature and potential for adaptation to unforeseen systems.

Wrong incentives can also lead to market failures and underinvestment in areas where the value of innovation is not directly profitable. These incentives may overlook social externalities such as technological displacement of labor, the health of an informational ecosystem necessary for democracy, the environmental cost of computing resources, and the sale of technologies to non-democratic regimes for profit.

What to do?

Academia and industry need to collaborate. The unique, broad-ranging expertise and altruistic motivations of educational institutions can offer valuable, ethically and technically sound direction in this process.

To establish the professional norms that will enable the responsible research (to promote their social benefit and mitigate their social harms) and deployment of foundation models. These norms, among other things, will define:

(for each new downstream use case):
(i) surrogate metrics for a representative set of potential downstream evaluation, and
(ii) a commitment to documenting these metrics

monitoring and intervention that is needed at every stage of foundation model sevelopment

when models are “safe” to release

how the community should react in response to methodological misconduct

To estabilish market-driven commercial that considers social benefit

Reduce the gap between private models that the industry can train and those who are open to the community (combating its fundamental centralization nature)

EleutherAI and Hugging Face's BigScience project are attempting to train large foundation models (jul2022)

In the US, the nascent National Research Cloud initiative is a step in this direction