On the Opportunities and Risks of Foundation Models

Technological Foundations

Sept 14, 2021

SUT
Speech Processing Lab

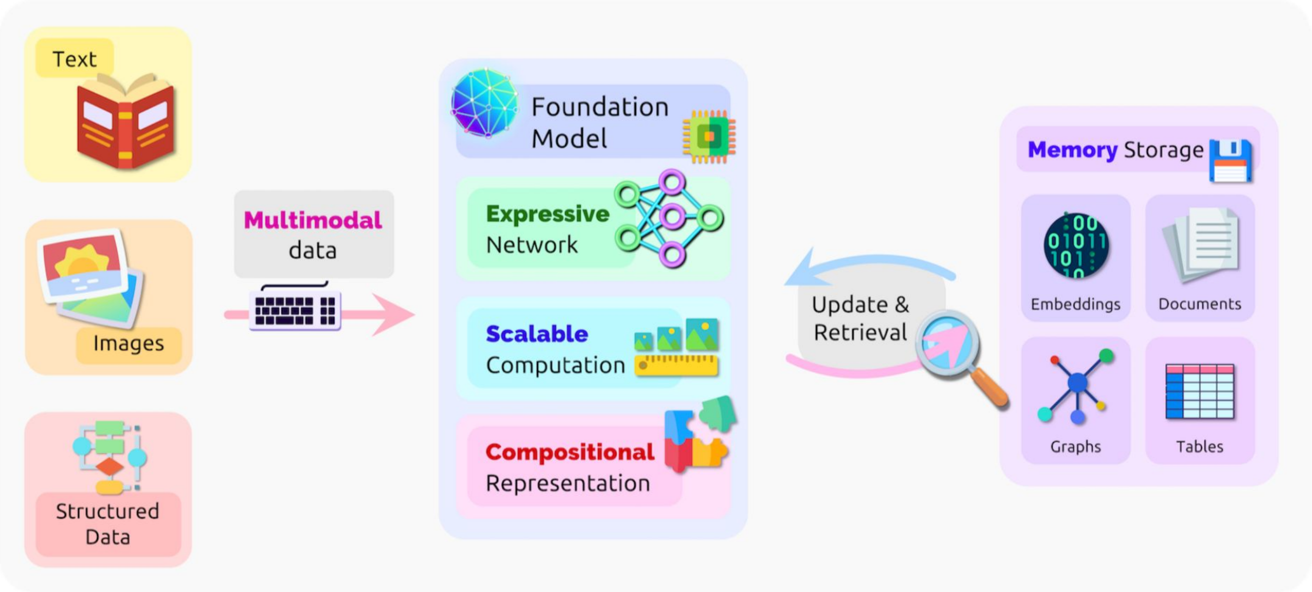
Index

Foundation Models

1. What capabilities do they have?
2. What if we just scale up the size & parameters of model?
3. Is any research possible without extreme-scale computing power?
4. How are foundation models different from the other neural networks?
5. What kind of data is needed?
6. How can we be sure of the quality of the data?

Modeling

Key Properties



The five key properties of a foundation model: expressivity, scalability, multimodality, memory capacity, and compositionality.

Expressivity

Capacity of a network to <u>model</u> the trained data distribution and <u>represent</u> it flexibly.

1. Inductive Biases
2. Transformer Networks & Attention
3. General-Purpose Computation
4. Challenges & Future Directions

Expressivity

1. Strong evidence for the high expressivity of neural networks from breakthroughs in generative models (Brown et al. 2020; Devlin et al. 2019)
2. Effectiveness of attention & gating units in comparison to the mechanisms involved in MLP & CNN networks (better at adapting the computation to the input, e.g. considering context in an NLP task). (Zavrel et al. 1997)
3. Not strongly tied to a particular task or domain (Liu et al. 2019; Dosovitskiy et al. 2020; Hudson and Zitnick 2021)
4. Trade-off between efficiency & expressivity. (Zavrel et al. 1997)
   - The need for ways to create a balance between the two
   - Focusing on other modalities like structural and perceptual

Scalability

The foundation models must keep up with the progress rate of **increasing resources** and **computational power**.

1. Models’ depth & width
2. Training time
3. Number of parameters
4. Amount of data

Scalability

Foundation models should be:

1. Easy-to-train
2. Easy-to-adapt

Multimodality

A key component of intelligence, and crucial factor for the development of comprehension of the world.

Foundation models should:

1. Connect together the different modalities
2. Distill information into a shared multifaceted representation
3. Capture the full range of inter-connections and relations among them

Multimodality

Memory

Separate out computation from memory

1. Separation of explicit facts & implicit knowledge:
    a. Alleviates models' size and number of parameters
    b. Improves models' trust and reliability
    c. Key for memory update, manipulation or adaptation

Compositionality

The principle according to which the meaning of the whole is derived from the meaning of its constituent parts, and the rules applied to combine them [Janssen and Partee 1997; Bottou 2014]

1. Model
2. Computation
3. Training & Data
4. Representation

Training

Design Trade-offs

1. Level of abstraction
2. Generative vs discriminative models
3. Capturing multimodal relationships

Future Path

1. Out-of-box self-supervised models
   a. Not so easy to understand the underlying principles
   b. Highly domain-specific
2. Goal-directed training of foundation models

Adaptation

Adaptation Procedure

1. Inclusion of new data
2. Prompt in input data
3. Updating some or all of the parameters

Use Cases

1. Task specialization
2. Temporal adaptation
3. Domain specialization
4. Local model editing
5. Applying constraints

Evaluation

Intrinsic Evaluation

1. Approaches:
    a. Meta-benchmarks
    b. Evaluation of intrinsic properties
2. Design principles:
    a. Inspiration from evaluation of humans
    b. Human-in-the-loop evaluation
    c. Validity of intrinsic measures

Extrinsic Evaluation

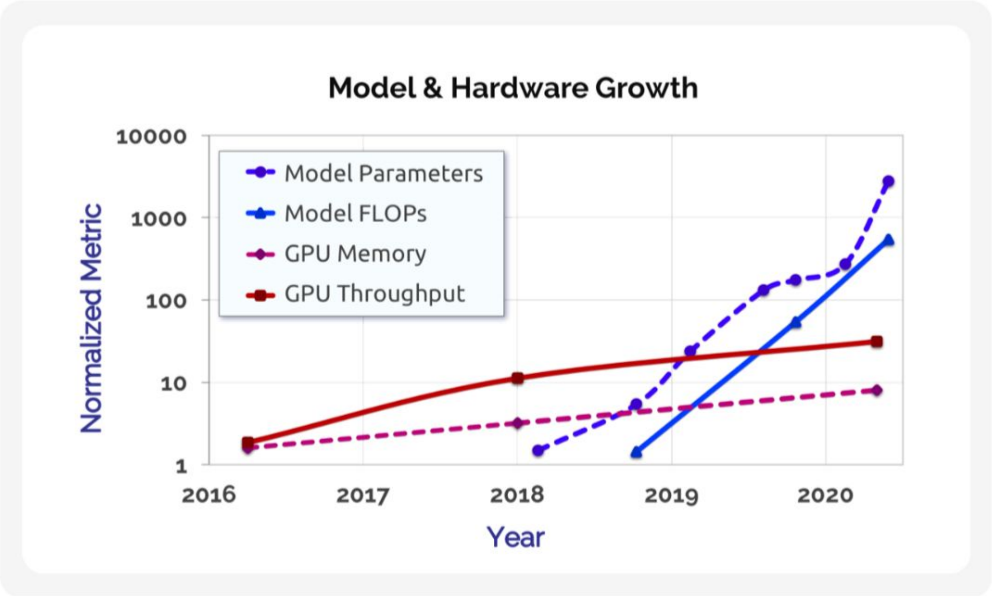
1. Accounting for the process & resources

Evaluation Design

1. Traditional models: Large training set for learning, an optional validation set to optimize hyperparameters, and a test set to for evaluation
2. Foundation models:
    a. Much smaller and far more diverse benchmarks for individual tasks
    b. Nature of foundation models may cause a shift in nature of benchmarks, de-emphasizing quantity as opposed to quality and diversity
    c. Measurements across diverse fronts and more than just accuracy (e.g., robustness, fairness, efficiency and environmental impact)

Systems

Model & Hardware Growth



Designing Systems

1. Training Phase:
   a. Automatic discoveries and optimizations
   b. Sharing pretrained model between two models
   c. Leveraging volunteer computing (like Learning@Home)
2. Production:
   a. Model compression techniques (distillation, quantization, pruning, and sparsity)
   b. Parallelization techniques
   c. Automated dataset curation (behavioral testing)
   d. Model quality assurance (model assertions)

Data

Challenges

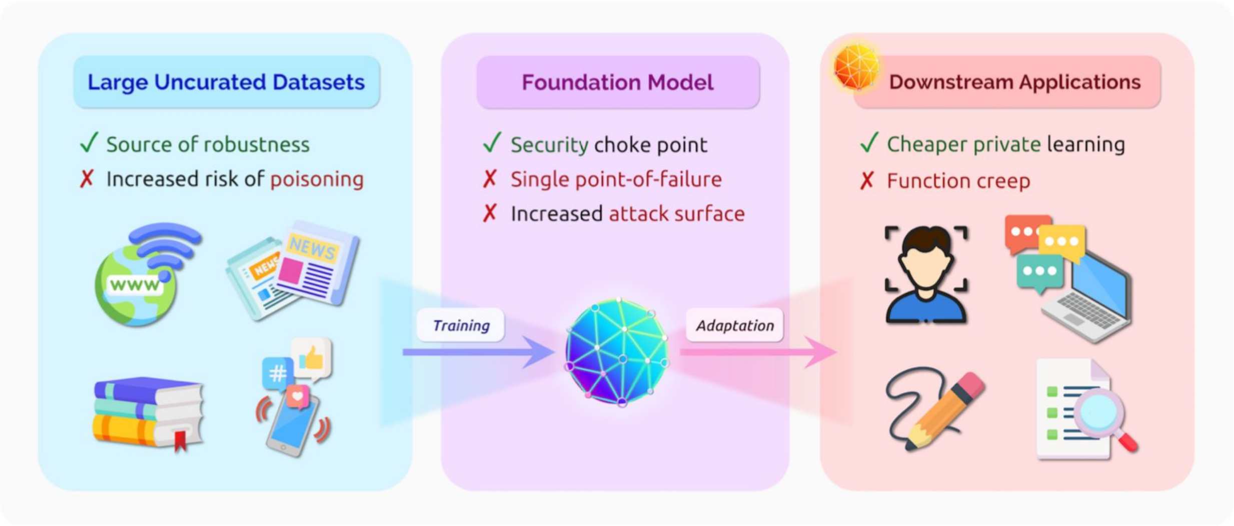
1. Scalability
2. Data integration
3. Privacy and governance controls
4. Understanding data quality

Challenges: Data Hub Solution

1. **Scalability:** Standard data management solutions, scalable interfaces, heterogeneous compute, and cloud infrastructure to support scalable solutions in different environments
2. **Privacy and governance controls**
3. **Data quality tooling:** Automatic & manual data correcting tools and analyzing tools regarding model errors

Security and Privacy

Security Risks & Opportunities



Risks and opportunities raised by foundation models for security and privacy of ML systems

Computer Security Threats in ML Systems

1. Confidentiality of user data
   a. Data inference and reconstruction attacks
   b. Model stealing attacks
2. The Integrity of ML systems
   a. Adversarial examples
   b. Data poisoning attacks
3. Availability of ML systems
   a. Resource-depletion attacks

Risks (and Opportunities)

1. **Single points of failure**
    a. Data poisoning attacks
    b. Adversarial examples
    c. Data privacy
    d. Data Centralization
    e. Model stealing attacks
    f. Denial-of-service attacks
2. **Function creep & dual use**
    a. Overlearning
        i. Attributes that are not part of the learning objective
        ii. Attributes that are sensitive from a privacy or bias perspective
    b. Adversarial reprogramming
        i. Reprogramming CLIP for facial recognition
3. **Multimodal inconsistencies**

AI Safety and Alignment

AI Safety and Alignment

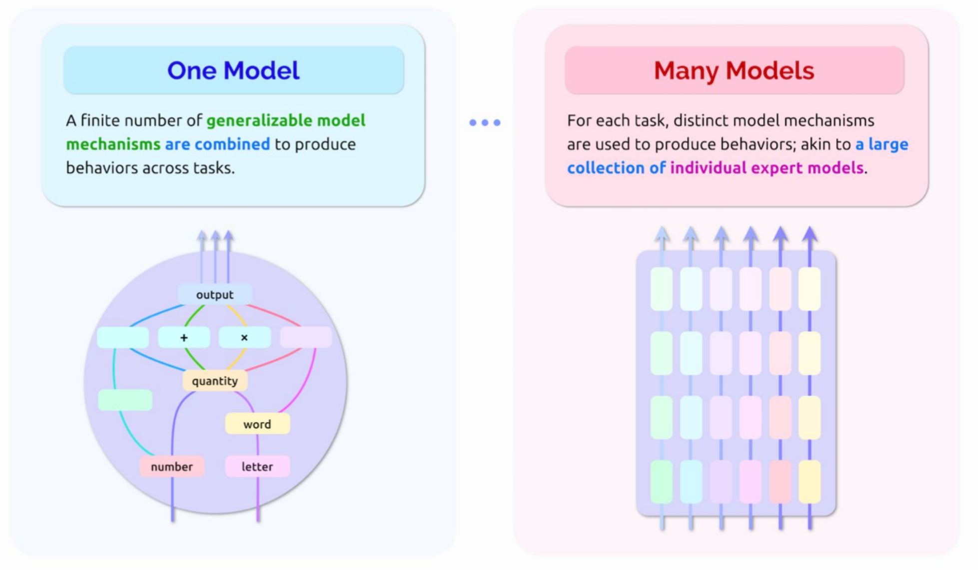
Challenges & Potential Risks

1. Generality of foundation models
2. Unexpected changes (e.g., “Prompting” in GPT-3)
3. Complex characterization of capabilities

1. Catastrophic robustness failures
2. Misspecified goals

Interpretability

Security Risks & Opportunities



Should we consider foundation models as one huge model or some separate, task-specific models?

Understanding Foundation Models

1. Characterizing behavior
2. Explaining behavior
3. Characterizing model mechanisms

References

Bommasani, Rishi, et al. “On the Opportunities and Risks of Foundation Models.” ArXiv:2108.07258 [Cs], Aug. 2021. arXiv.org, http://arxiv.org/abs/2108.07258.

Disclaimer: Not all of the references have been provided. To see all of the references, please refer to the original paper.