FOUNDATION MODELS - A REVIEW

Sumeet Shirgure
sshirgur@usc.edu

ABSTRACT

Foundation models such as GPT, LLaMA, CLIP and SAM have gained popularity lately. This paper discusses some of the recent trends pertaining to foundation models, and reviews some of the developments in the past couple of years published in top conferences in AI and machine learning.

1 INTRODUCTION

The term *foundation models*, as coined in Bommasani et al. (2022) refer to large scale models trained on broad modalities of a tremendous amount of data, that are then later specialized to perform specific, related tasks. They're named so because the term ”foundation” captures the fact that such models are incomplete and yet they are important.

These models have been shown to be capable in performing all kinds of tasks arising in various fields like language, vision and robotics. Some of the ingredients for successfully building such models, for example deep neural networks, self supervised learning and transfer learning, were already well known; and still the sheer scale at which foundation models operate gives rise to new *emergent* capabilities. Another central aspect of foundation models is what's termed as *homogenization*, which means that a single model can be specialized to perform multiple downstream tasks.

The objective of this review is to act as an index to some of the developments that have taken place in the past couple of years. The rest of this paper is structured as follows: each section has a collection of papers that are each discussed in some detail. Section 2 goes over recent developments surrounding the theory of foundation models. Section 3 discusses new methodologies for the compression, reuse and adaptation of foundation model weights for various purposes. Section 4 points to papers that reveal new ideas in decentralized training. Section 5 explores results in foundation models being applied to problems in the fields of vision, language and logical reasoning.

2 THEORY SURROUNDING FOUNDATION MODELS

ON THE POWER OF FOUNDATION MODELS

In stark contrast to the use of statistical learning theory, [Yuan (2023)] makes strides in the use of category theory to derive remarkable results applicable to self supervised learning, prompt tuning and foundation models. The paper starts with the question - ”With infinitely many data points, infinite computational power, an infinitely large foundation model with a perfect training algorithm and guaranteed zero generalization error on the pretext task, can the model be used for everything?” In doing so the paper discards most of the assumptions made by existing theory that study sample sizes, computational complexity and scaling laws for example.

Of course the paper makes some assumptions, the most important of which are that (a) different modalities of data like images and texts are modelled as categories (b) for solving a downstream task there are mainly two types of methods : (i) prompt tuning : where no training happens on the downstream task and only a prompt is sent to the model so that it can switch its working mode (ii) fine tuning : refers to supervised training of a small network that is appended to a foundation model whose parameters are frozen. In both of these cases, the foundation model is trained on a pretext task like masked language modelling or image reconstruction. The author then proceeds to use category theory to prove several theorems regarding foundation models.

The first says that the model can solve a downstream task if and only if it is representable (as defined in the paper) in the category defined by the pretext task. In other words, arbitrary pretext tasks are not guaranteed to learn representations that solve a particular downstream task. The paper gives the example of learning image orientation as a pretext task, and proves that it's not able to solve complicated downstream tasks like image segmentation.

The second says that the fine tuning approach has no such restriction: that as long as the foundation model has the minimum required computational capacity for the pretext task, and there is enough training data any downstream task can be solved in the category defined by the pretext task. The role of pretext task is crucial in the sense that if the pretext task fails to extract adequate information from the unlabeled dataset, the power of fine tuning remains restricted.

The last has to do with generalization, in that foundation models are able to generate unseen objects from a category (e.g images) using structural information learned from a source category (e.g texts).

Along the way, the paper also provides a categorical framework for supervised and self - supervised learning which is interesting in itself. As now there is a choice between using the well established, quantitative statistical theory vs. this newly constructed, existential, categorical theory.

3 Playing with foundation model weights

SparseGPT: Massive Language Models Can be Accurately Pruned in One-Shot

This paper [Frantar & Alistarh] shows for the first time that large language models can be pruned to at least 50% sparsity in one-shot without any re-training at minimal loss in accuracy. They call their pruning method ”SparseGPT” as it's specifically designed for the GPT-family of models.

All prior approaches for model compression focused on quantization. While quantization truncates the precision of the model weights to reduce information content, pruning removes network elements like individual weights to entire rows/columns of weight matrices. And even though pruning was successfully applied to vision models before this work, it required extensive retraining to recover a significant portion of the lost accuracy.

Another notable observation made by the authors is that larger models are more compressible, i.e they drop significantly less accuracy at a fixed sparsity when compared to smaller models. The method also allows sparsification to be compounded with quantization. Lastly, SparseGPT is entirely local in the sense that no global information (e.g gradients) need to be computed.

Contrastive Adapters for Foundation Model Group Robustness

While foundation models can generalize well to various datasets and exhibit impressive robustness to some distribution shifts in the overall data, this [Zhang & Ré (2022)] paper shows that in the zero-shot regime foundation models can have poor *group robustness*. [1] For example, these groups could be sensitive features such as subject's skin tone in an image generation setting. In particular, there could be as large as 80 percentage point gap between the average and the worst group accuracy for certain classification tasks.

Improving robustness of existing models is challenging because all prior methods involved retraining entire models. For foundation models this can be expensive due to their size and scale; not to mention the weights themselves may not even be available, just access to model outputs via APIs like that of OpenAI.

This paper studies effective and efficient solutions for better group robustness in foundation models. As a baseline the authors consider linear probes –small linear layers– and adapters –small bottleneck MLPs – on top of foundation model embeddings. The authors note that poor zero-shot robustness results when FMs embed same-class samples in different groups “far apart” in embedding space.

They then propose *contrastive adapting*, a simple adapter training method that prioritizes bringing those ”far apart” points together, while also applying a supervised contrastive loss over *other sample embeddings*. The method provides a way to ”pull together” far apart sample embeddings in the same

[1] defined in the paper as the difference between the average error and the worst performing group error

class and ”push apart” nearby sample embeddings in different classes. Doing so they achieve near SoTA worst-group accuracy on popular robustness benchmarks with only 1.0% of the trainable parameters.

In summary, while FM zero-shot classification may not be group-robust, we can significantly improve robustness without any fine tuning. This suggests the information to classify groups is present in their pretrained embeddings; and contrastive adapting is a good method to extract it.

Model Ratatouille: Recycling Diverse Models for Out-of-Distribution Generalization

This paper [Rame et al. (2023)] by Meta AI targets the handful of fine tuned foundation models that are available online and proposes a new strategy to recycle them into an ensemble that performs better on out-of-distribution samples. As per the paper, the internet is full of a few foundation models, each of which have been fine tuned by various practictioners. These fine-tunings are isolated in themselves, and don’t benefit from each other. The proposed *model ratatouille* is a novel strategy to recycle multiple fine-tunings of the same foundation model on diverse auxiliary tasks.

The proposed way is to repurpose these auxiliary weights as initializations for multiple parallel fine-tunings on the target task, and finally the fine-tuned weights are averaged to get the ensemble model. The diversity in auxiliary tasks gives rise to a natural diversity in weights and the proposed method harnesses it. The paper goes on to study the empirical improvement in out-of-distribution generalization on a standard benchmark.

Lastly, this paper also contributes to the emerging paradigm of updateable ML. Raffel (2023) An exciting foresight presented in the paper is the possibility of *updateable machine learning*, where akin to open-source software development, the community collaborates to reliably update machine learning models. It’s not entirely obvious how such averaging can be used to update existing foundation models, but this paper is certainly a step in this direction.

4 Decentralizing the training of foundation models

Decentralized Training of Foundation Models in Heterogeneous Environments

The paper [Yuan et al. (2022)] takes the first steps towards heterogeneous decentralized training of foundation models. Some of the terminology that the paper uses begs explanation : simply distributing deep learning compute over multiple GPUs in a single data center is considered a *homogenous* workload. Such parallelization is not considered true *decentralization*, where compute is distributed over geospatially distributed GPUs. It’s such decentralization that is akin to that of the Folding@Home Shirts & Pande (2001) project that’s considered *heterogeneous* in the paper. The fundamental characteristics of a heterogeneous network are high latency and low bandwidth of interconnects.

The advantage of doing so is the huge savings in electricity being used across GPUs, not to mention the wide availability of compute ”clusters”. More importantly, training foundation models is costly; as per one estimate in the paper training a single GPT3-175B instance takes 3.6K petaflops-days, amounting to a total of $4M on AWS instances at the time.

Among the key technical contributions of the paper is a scheduling algorithm that allocates computational ”tasklets” in the training of foundation models to a group of decentralized GPU devices. Each such tasklet comprises of a micro-batch of data and a subset of layers of the foundation model. Note that such parallelization is more fine-grained that just pure data parallelization. The scheduling algorithm is optimized for a cost model that is also introduced in the paper. The cost model is decomposed into two levels, the first level deals with the communication cost of data parallelism, while the second level deals with the cost of compute pipeline parallelism.

The collaboraters conduct extensive experiments that simulate real working conditions of such decentralized training, and show that it is merely $1.7 - 3.5\times$ slower than Megatron-Deepspeed in data centers, *even though the internetworks can be* $100\times$ *slower*. The paper doesn’t go into further engineering details like checkpointing and fault tolerance and leaves that to future work.

CocktailSGD: Fine-tuning Foundation Models over 500Mbps Networks

In a continuation of [Yuan et al. (2022)] the paper [Jue et al. (2023)] introduces ”CocktailSGD”, which is a training framework that combines three distinct compression techniques. The authors make a point that communication is indeed the key bottleneck in scaling the training of foundation models. To relieve this bottleneck in a decentralized setting, CocktailSGD combines three distinct well known techniques to compress gradients to lower the overhead, and achieve a much greater compression than each individual technique alone.

The three methods bring complementary benefits and advantages : (i) *Top-K sparsification* : improves gradient sparsity but leaves overhead to encode the sparse pattern (ii) *quantization* : decreases the overhead but has limited compression ratio (iii) *random sparsification*: improves sparsity and provides a way to encode the sparse vector, but might miss important directions

The key challenge, as noted in the paper, is to balance the communication compression ratio with the convergence of SGD. And the key contributions of the paper therefore are (i) CocktailSGD : an asynchronous training framework that superimposes communication and local gradient computation. (ii) Theory : proof that CocktailSGD ensures local convergence at a rate $O(1/\sqrt{NT})$ where N is the number of workers and T is the number of iterations. (iii) Experiments : large scale experiments are conducted that fine tune open foundation models up to 20B parameters. As a result CocktailSGD is shown to converge to comparable training loss with a comparable number of iterations while requiring $117\times$ *less data transmission*.

5 Foundation models in vision and language paradigms

Visual Classification via Description from Large Language Models

This paper [Menon & Vondrick (2023)] gives a way to classify images by descriptive words generated using a large language model. Their thesis is that vision-language models (VLMs) like CLIP [Radford et al. (2021)] only use the category name for performing classification, and in doing so ”neglect to make use of the rich context of additional information that language affords”. The procedure also gives no interpretable reasoning of why the category is chosen.

The authors thus present an alternative framework for classification with VLMs. They ask VLMs to check for *descriptive features* rather than just the broad category. E.g: to find a tiger, look for its stripes, its claws and its tail and so on. This process naturally generates interpretable data on why the model classifies the image as a tiger. The descriptive features, as you might have guessed, come from language models like GPT. ”Large language models (LLMs) such as GPT-3 can be thought of as implicit knowledge bases, noisily condensing the collective knowledge of the Internet in a way that can be easily queried with natural language” - Petroni et al. (2019). The authors simply ask an LLM, much like a child, ”what does a tiger look like?” to get its description. As noted, this method requires no additional training, and the authors report a 4-5% improvement in zero-shot top-1 ImageNet accuracy, which is quite impressive.

Visual Clues: Bridging Vision and Language Foundations for Image Paragraph Captioning

The paper [Xie et al. (2022)] proposes a new computer vision system called BEST (Bridging with Explicit Structured Textual clues) for image paragraph captioning. The authors argue that textual representation like image paragraph captions make a good description for the very idea of human vision. And their case is strong - it’s more holistic than mere categorical labels. Moreover the way captions are generated in the paper also allows machines to interpret visual signals.

The proposed method starts by constructing a semantic representation of the image, referred to as visual clues. It comprises of rich semantic components like object and attribute tags, and is powered by recent advances in vision foundation models like Florence [Yuan et al. (2021)] and CLIP [Radford et al. (2021)]. These visual clues are then ingested by language foundation models like GPT along with an engineered prompt, which in turn produces crisp language descriptions that are sensible while also uncluttered with irrelevant information.

The contributions are thus twofold (a) proposal of a general framework for semantic visual representation and its application to image paragraph captioning (b) benchmarking the effectiveness of the proposed framework and setting state of the art results

OmniVL: One Foundation Model for Image-Language and Video-Language Tasks

This paper [Wang et al. (2022)] presents OmniVL, a new foundation model supporting both image-language and video-language tasks using a single universal transformer-based encoder - decoder architecture. The novelty of this work is in showing that joint image-language and video-language pretraining benefits both image and video tasks, as opposed to the conventional one-directional transfer from image to video for example.

Moreover, the paper introduces a novel unified vision-language contrastive loss function to leverage the different modalities of data together. Without the use of additional adapters, OmniVL can simultaneously support visual tasks like image classification, cross-modal alignment tasks like image retrieval, and multimodal understanding like visual question answering (VQA). OmniVL is motivated by the unified contrastive learning used in Florence Yuan et al. (2021) and extends its scope to cover video-text and image-label data.

Selection-Inference: Exploiting Large Language Models for Interpretable Logical Reasoning

Large language models (LLMs) have impressive few/zero-shot capabilities, but still perform poorly on multi-step logical reasoning problems. The authors of this paper [Creswell et al.] propose a new framework named the Selection-Inference (SI) framework that alternates between selection and inference to generate a series of causal reasoning steps leading to the final answer. It's shown that a 7 billion parameter LLM used within the SI framework in a 5-shot setting outperforms the larger 280 billion parameter LLM when used in a few-shot manner. Not to mention that the answers produced by the SI framework are accompanied by a causal NLP reasoning trace which has implications for interpretability, safety and trustworthiness of the system.

The SI framework decomposes logical reasoning into two stages : (i) *selection* : which involves choosing a subset of relevant information emitted by the LLM for a single step of inference (ii) *inference* : which only sees the limited information provided by the selection module and uses it to infer a new intermediate piece of evidence on the way to the final answer.

The reasoning trace that is thus generated is *causal*, in that each step follows from the previous step. Each inference step is made in isolation, which is in contrast with the more common *post-hoc rationalization*. The SI framework is also shown to be better at formulating explanations than other approaches like Chain-of-Thought prompting (COT) [Wei et al. (2022)] where the reasoning traces are often wrong even though the final answer is correct.

6 Conclusion

We looked at several notable papers pertaining to foundation models published in the years 2022-2023. The author hopes that this paper serves as a good starting point to explore these recent developments in detail. Thank you for reading.

References

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya

Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022.

Antonia Creswell, Murray Shanahan, and Irina Higgins. Selection-inference: Exploiting large language models for interpretable logical reasoning, 2022. *URL https://arxiv. org/abs/2205.09712*.

Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot, 2023. *URL https://arxiv. org/abs/2301.00774*.

WANG Jue, Yucheng Lu, Binhang Yuan, Beidi Chen, Percy Liang, Christopher De Sa, Christopher Re, and Ce Zhang. Cocktailsgd: Fine-tuning foundation models over 500mbps networks. 2023.

Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=jlAjNL8z5cs.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL https://aclanthology.org/D19-1250.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

Colin Raffel. Building machine learning models like open source software. *Commun. ACM*, 66(2): 38–40, jan 2023. ISSN 0001-0782. doi: 10.1145/3545111. URL https://doi.org/10.1145/3545111.

Alexandre Rame, Kartik Ahuja, Jianyu Zhang, Matthieu Cord, Léon Bottou, and David Lopez-Paz. Model ratatouille: Recycling diverse models for out-of-distribution generalization. 2023.

Michael Shirts and Vijay Pande. Screen savers of the world unite. *Science (New York, N.Y.)*, 290: 1903–4, 01 2001. doi: 10.1126/science.290.5498.1903.

Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 5696–5710. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/259a5df46308d60f8454bd4adcc3b462-Paper-Conference.pdf.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Yujia Xie, Luowei Zhou, Xiyang Dai, Lu Yuan, Nguyen Bach, Ce Liu, and Michael Zeng. Visual clues: Bridging vision and language foundations for image paragraph captioning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17287–17300. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/6e4df3406bcf04443ea26d5695454355-Paper-Conference.pdf`.

Binhang Yuan, Yongjun He, Jared Davis, Tianyi Zhang, Tri Dao, Beidi Chen, Percy S Liang, Christopher Re, and Ce Zhang. Decentralized training of foundation models in heterogeneous environments. *Advances in Neural Information Processing Systems*, 35:25464–25477, 2022.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

Yang Yuan. On the power of foundation models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 40519–40530. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/yuan23b.html`.

Michael Zhang and Christopher Ré. Contrastive adapters for foundation model group robustness. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 21682–21697. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/8829f586a1ac0e6c41143f5d57b63c4b-Paper-Conference.pdf`.