Modular Large Language Models

Suchin Gururangan

Paul G. Allen School of Computer Science

Abstract

Despite the diversity of stakeholders and interests in NLP technologies, a small set of corporations drive the development of large language models (LMs). We propose to introduce a new class of LMs that are fundamentally *modular*, where components (or *experts*) of the LM are specialized to distinct data distributions in the training corpus. At inference time, users can mix, add, or remove experts, enabling rapid customization to new use cases. We propose 1) to develop new modular LM architectures and 2) to use modularity for decentralized LM development, where anyone can contribute to and maintain experts at very modest computational cost.

Collaborators: Noah A. Smith (UW/AI2), Luke Zettlemoyer (UW/Meta AI)

1 Introduction

The NLP community relies on a small set of corporations with enough computational resources to select data for, train, and responsibly release large language models (LMs). This centralization increases the risk of unwanted model behavior [Gehman et al., 2020], reinforces biased data curation practices [Gururangan et al., 2022], and necessitates costly model adaptation to new data distributions after training [Gururangan et al., 2020].

We instead propose to re-design LMs with **modularity** [Gururangan et al., 2021] (Figure 1). Under this framework, smaller components of the LM (called **experts**) are specialized to distinct domains of a large training corpus (e.g., scientific or legal text). During training, experts are only updated with respect to the domains they have been assigned to. At inference-time, experts can mixed (to generalize to heterogeneous domains), added (to incorporate new domains), or removed (to forget unwanted domains). This flexibility will allow for cheap and rapid adaptation of a large LM to new use cases after training, as needed for different end tasks.

Modularity will also enable the first **decentralized** LM, where researchers can asynchronously train experts on different domains, contribute experts to a large modular LM, and combine experts in sparse ensembles for end tasks. An end-task developer can also update or disable experts after deployment if, for example, a domain develops as a source for problematic text (e.g., hate speech; Gehman et al. 2020), or the expert degrades in performance over time [Luu et al., 2021]. With modularity, responsibility for data procurement, model training, and model release is distributed across the NLP community.

To build these models, we will need two key advances:

1. **New Modular Architectures (§2.1):** New transformer LM architectures that are modular by design, and new training algorithms to specialize experts.

2. **Decentralized LMs (§2.2):** We also need new algorithms to asynchronously train a modular LM with thousands of experts, and new tools to continuously maintain experts after deployment.

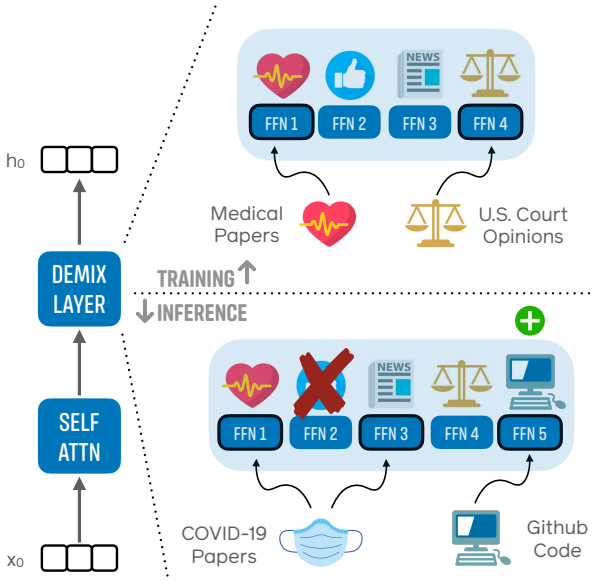

Figure 1: Illustration of a modular LM in a single transformer block, as proposed in [Gururangan et al., 2021]. During training, expert feedforward networks are conditionally activated based on the domain (here, document provenance) of the input sequence. At inference time, the LM has new modular functions; mixing, adding, and removing experts for rapid adaptation. In this proposal, we develop new modular architectures, and use modularity to decentralize LM development.

2 Proposed Work

2.1 New Modular Architectures

We propose to develop new transformer LM architectures that are modular by design, building on our prior work [Gururangan et al., 2021]. We will also develop new training algorithms that enhance expert specialization while improving model generalization to new domains.

For example, in Gururangan et al. 2021 we used *document provenance* to define domains that we would assign experts to. While provenance is convenient and easy to understand, it may not correspond to the best segmentation of a corpus — domains may be organized hierarchically [Chronopoulou et al., 2021] or have fuzzy boundaries between coarse provenance categories [Gururangan

et al., 2020]. *Learning* domains via unsupervised methods (e.g., clustering pretrained representations of documents; Aharoni and Goldberg 2020) may result in expert assignments that improve downstream LM performance. A number of clustering algorithms could be evaluated, each with their own tradeoffs: minibatch k-means clustering [scikit, 2022] is scalable, but agglomerative clustering may be more effective at representing a hierarchy of domains [Chronopoulou et al., 2021]. We can use the domain hierarchy in a training algorithm that initializes new experts with that of their parent domain, reducing the costs of adding new experts to the system.

Moreover, in Gururangan et al. 2021 we modularized the feedforward layers of the transformer into experts. However, sharing other parts of the network (e.g., self-attention layers, vocabularies) across domains may be detrimental to distinct domains with different context lengths and word usage. As such, we propose to develop new architectures for modularity. These architectures may include experts in low-rank parameters of the network (e.g., bias terms, Zaken et al. 2021), the vocabulary, or only in certain layers of the network. The modular architectures could depend on characteristics of the domains. For example, preliminary experiments suggest that heterogeneous domains (e.g., web crawls), which span a variety of styles, genres, and word usages, tend to benefit from more parameter sharing across domains, while distinct domains (e.g. legal text) do not.

Finally, training algorithms could be considered where the transformer is gradually modularized, beginning with a phase of dense training. Increased modularization may improve specialization of experts, while initial phases of dense training may help experts retain cross-domain generalization. Experiments on gradual modularization could assume an overall computational budget and explore how much time to dedicate to each phase of training. We hypothesize a tradeoff between generalization and modularity when sweeping over different budgets for dense and modular training.

2.2 Decentralized LMs

To democratize large LM training and maintenance, we envision a new decentralized system of LM development. Under this system, researchers can asynchronously train experts on their own (carefully documented) language data, and submit them to a larger LM that can incorporate newly trained experts into sparse ensembles for new domains. A key research question is how to handle shared parameters during asynchronous training, since they could go out of sync if updated with each newly submitted expert. We will also investigate how to efficiently return the right set of experts for a new domain, which could be framed as a sparse retrieval problem [Seo et al., 2019]. Once experts are retrieved, we will compare the effects of different sparse ensembling methods, such as output-ensembling or weight averaging [Wortsman et al., 2022], to maximize performance on the new domain. Retrieved experts could be distilled into a much smaller LM for efficient deployment [Bapna et al., 2021].

Modularity opens up new opportunities for decentralized maintenance of experts after model deployment. In Gururangan et al. 2021, we discovered that *expert removal* can dynamically control the LM's data exposure at test time; one can simulate an LM that has not been exposed to the domains the experts were specialized to. If an expert has been exposed to problematic content or is degrading in performance over time [Luu et al., 2021], they may be disabled or removed to prevent misuse, or updated with further adaptation. We propose building evaluation tools to continuously monitor experts for degenerate behavior. A reasonable data source for such a tool is Bloomberg's vast trove of **breaking news**, which represents a rapidly changing domain with new terms and concepts. With an evolving LM benchmark of breaking news, we can measure how well experts fit text related to emerging events, and update them accordingly.

Our vision necessitates a system of community governance to maintain and distribute experts in a decentralized LM. Who has the power to audit and remove degenerate experts? What constitutes fair use of experts in a decentralized LM, especially they have been trained on private data? We propose to draw on literature in content moderation [Gillespie, 2020] and data governance [Janssen et al., 2020] to investigate these questions.

3 Expected Results & Impact

We aim to build new modular architectures and training algorithms in the first year that will be scaled in the later year(s). We have an overall goal of decentralizing LMs, which will have broad impact due to the pervasiveness of LMs in virtually every modern NLP application [Bommasani et al., 2021]. Decentralized LMs will enable broad community participation in the training and deployment of large LMs.

4 Data, Software and Ethics Policy

This research proposal is a multi-year agenda, with the aim to submit to top-tier conferences, open-access preprints, open-source software, and presentations throughout. In all experiments, we will only use data associated with friendly licenses that support reproducibility. We will make our best effort to anonymize personally identifiable information in any scraped corpus. The LMs we train will be partially exposed to web corpora, which contains undesirable content [Bender et al., 2021]. The ethical and legal norms around using public-facing web data are still in flux [Fiesler et al., 2020], and may not align with user perceptions of what constitutes fair use of online communications [Williams et al., 2017]. For model and data release, we will clearly acknowledge these risks, leveraging model cards [Mitchell et al., 2019] and datasheets [Gebru et al., 2018] for documentation.

Bloomberg employee consultants: No Bloomberg employee was consulted in the preparation of this application.

References

Roee Aharoni and Yoav Goldberg. Unsupervised domain clusters in pretrained language models, 2020.

Ankur Bapna, Dmitry (Dima) Lepikhin, Maxim Krikun, Orhan Firat, Sneha Reddy Kudugunta, Thang Luong, and Yanping Huang. Beyond distillation: Task-level mixture-of-experts for efficient inference. In *Beyond Distillation: Task-level Mixture-of-Experts for Efficient Inference*, 2021.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of FAccT*, 2021. doi: 10.1145/3442188.3445922. URL `https://doi.org/10.1145/3442188.3445922`.

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, 2021. URL `https://arxiv.org/abs/2108.07258`.

Alexandra Chronopoulou, Matthew E. Peters, and Jesse Dodge. Efficient hierarchical domain adaptation for pretrained language models, 2021. URL `https://arxiv.org/abs/2112.08786`.

Casey Fiesler, Nathan Beard, and Brian Keegan. No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods in social media terms of service. In *Proceedings of ICWSM*, 2020. URL `https://ojs.aaai.org//index.php/ICWSM/article/view/7290/7144`.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, and Kate Crawford. Datasheets for datasets, 2018. URL `https://arxiv.org/abs/1803.09010`.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models, 2020.

Tarleton Gillespie. Content moderation, ai, and the question of scale. *Big Data & Society*, 7(2):2053951720943234, 2020. doi: 10.1177/2053951720943234. URL `https://doi.org/10.1177/2053951720943234`.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.740. URL `https://www.aclweb.org/anthology/2020.acl-main.740`.

Suchin Gururangan, Michael Lewis, Ari Holtzman, Noah A. Smith, and Luke Zettlemoyer. Demix layers: Disentangling domains for modular language modeling. *ArXiv*, abs/2108.05036, 2021.

Suchin Gururangan, Dallas Card, Sarah K. Drier, Emily Kalah Gade, Leroy Z. Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. Whose language counts as high quality? measuring language ideologies in text data selection. *ArXiv*, abs/2201.10474, 2022.

Marijn Janssen, Paul Brous, Elsa Estevez, Luis S. Barbosa, and Tomasz Janowski. Data governance: Organizing data for trustworthy artificial intelligence. *Government Information Quarterly*, 37(3):101493, 2020. ISSN 0740-624X. doi: https://doi.org/10.1016/j.giq.2020.101493. URL `https://www.sciencedirect.com/science/article/pii/S0740624X20302719`.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith. Time waits for no one! analysis and challenges of temporal misalignment. *ArXiv*, abs/2111.07408, 2021.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, jan 2019. doi: 10.1145/3287560.3287596. URL https://doi.org/10.1145%2F3287560.3287596.

scikit. Sklearn.cluster.minibatchkmeans, 2022. URL https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MiniBatchKMeans.html.

Minjoon Seo, Jinhyuk Lee, Tom Kwiatkowski, Ankur P. Parikh, Ali Farhadi, and Hannaneh Hajishirzi. Real-time open-domain question answering with dense-sparse phrase index, 2019. URL https://arxiv.org/abs/1906.05807.

Matthew L Williams, Pete Burnap, and Luke Sloan. Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology*, 51(6): 1149–1168, 2017. doi: 10.1177/0038038517708140. URL https://doi.org/10.1177/0038038517708140.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time, 2022. URL https://arxiv.org/abs/2203.05482.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models, 2021. URL https://arxiv.org/abs/2106.10199.