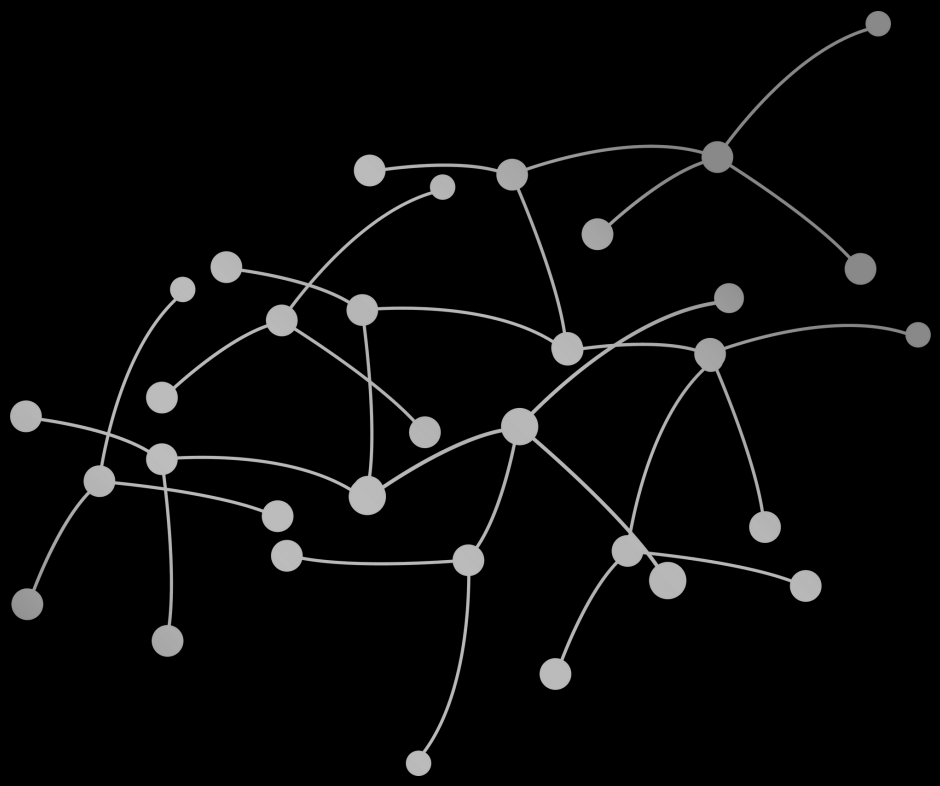
arcee.ai

Case Study:

Innovating Domain Adaptation through Continual Pre-Training and Model Merging



Authors:
Shamane Siri, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Charles Goddard, Mark McQuade

Table of Contents

I. Introduction

In the realm of specialized and secure language models, Arcee stands out with its focus on tailoring solutions that operate within the client's own cloud, leveraging their proprietary data. A cornerstone of our approach is domain adaptation, a critical yet resource-intensive process which maintains a balance between the general language capabilities and the specialized domain expertise of language models. This case study delves into how Arcee harnesses Continual Pre-Training (CPT) and Model Merging for cost-effective domain adaptation, showcasing our cutting-edge strategies in the Medical and Patent domains.

II. The Challenge of Domain Adaptation

Domain adaptation is paramount at Arcee, yet traditional methodologies demand considerable time and resources. In addition, a significant challenge arises with catastrophic forgetting, wherein post-pretraining often results in a deterioration of the model's original general abilities–hindering its fine-tuned performance across various tasks. This underscores the need for a method capable of incorporating domain--specific knowledge while mitigating forgetting and other deterioration. Our breakthrough lies in integrating two key methodologies: Continual Pre-Training (CPT) and Model Merging, designed to enhance efficiency and efficacy in adapting language models to specific domains.

III. Our Approach

Continual Pre-Training (CPT)

In language, CPT was studied under the name of domain adaptation pre-training where the new dataset comes from a new domain.[1] For instance, PMC-LLaMA[2], an open-source medical-specific large language model, incorporates data-centric knowledge injection with pure CPT and medical-specific instruction tuning. It stands out as the first of its kind, showcasing superior performance on diverse medical benchmarks with significantly fewer parameters compared to both ChatGPT and LLaMA-2. As another example, ChipNeMo investigates the utility of large language models (LLMs) in industrial chip design, employing a domain-adaptive CPT approach in their adaptation process. They assess their model across three specific chip design applications: an engineering assistant chatbot, EDA script generation, and bug summarization and analysis. Their findings demonstrate that their domain adaptation pipeline enhances LLM performance substantially compared to general-purpose models, achieving up to a 5x reduction in model size while maintaining or improving performance across various design tasks.[3] Inspired by prior work, CPT at Arcee involves extending the training of a base model, such as Llama-2-base or Mistral-7B-base, using domain-specific datasets. This process allows us to fine-tune models to the nuances of specialized fields.

FOOTNOTES:

[1] Gupta, Kshitij, et al. "Continual Pre-Training of Large Language Models: How to (re) warm your model?." arXiv preprint arXiv:2308.04014 (2023).

[2] Wu, Chaoyi, et al. "Pmc-llama: Towards building open-source language models for medicine." arXiv preprint arXiv:2305.10415 6 (2023).

[3] Liu, Mingjie, et al. "Chipnemo: Domain-adapted llms for chip design." arXiv preprint arXiv:2311.00176 (2023).

Model Merging

Model Merging involves synthesizing the capabilities of multiple pre-trained models into a single, more versatile checkpoint. This technique enables us to combine domain-specific models with general-purpose chat models, leveraging the strengths of both.[4] [5] [6]

Benefits of the Our Method

- Domain-Specific Data Utilization: By employing CPT, we can incorporate proprietary client data, ensuring models are finely-tuned to specific requirements.
- Efficiency in Model Development: Utilizing existing chat models accelerates development, avoiding the need for complex and expensive model tunings to have chat-like capabilities.
- Cost-Effectiveness: Fine-tuning smaller language models (SLMs) for specific domains yields substantial cost savings, with SLMs requiring only thousands of dollars for training compared to the billions needed for large language models (LLMs). Through Model Merging, our approach combines the specialized expertise of public SLMs with the broad domain-adapted SLMs, ensuring cost-effective and high-performance language model development.

FOOTNOTES:

[4] Yu, Le, et al. "Language models are super mario: Absorbing abilities from homologous models as a free lunch." arXiv preprint arXiv:2311.03099 (2023).

[5] Stoica, George, et al. "ZipIt! Merging Models from Different Tasks without Training." arXiv preprint arXiv:2305.03053 (2023).

[6] Yadav, Prateek, et al. "Ties-merging: Resolving interference when merging models." Advances in Neural Information Processing Systems 36 (2024).

IV. Case Study Highlights

Continual Pre-Training (CPT) Stage

Medical Domain:

- Our project in the medical domain entailed the development of a CPT checkpoint from a vast dataset sourced from medical articles and books, as per the PMC-Llama[2-1] paper protocol. This initiative generated a dataset which is similar to the Meditron[7] dataset, which was then utilized to enhance a Llama-2-7B base model, without employing traditional data cleaning techniques like de-duplication and topic filtering. We stopped the training process after 3500 steps when approximately 27 billion tokens of the dataset were processed.
- The model was trained using a packed strategy, with each example containing 4096 tokens. This approach was implemented with a learning rate of (1.5 \times 10^{-5}) and batch sizes of 2048, utilizing the Trainium architecture. For additional hyperparameters, we used the methodologies outlined in Gupta et al.'s[1-1] work.
- ***Note**: Our strategy did not extend to training beyond the 3500 steps due to the existence of Meditron[7-1], an open-source PMC Llama-2 chat model trained on a curated and well-cleaned 48B token medical dataset, compared to our former dataset. Given Meditron's exemplary performance, we acknowledge it as the pinnacle of CPT achievements in the medical domain and use it in place of the model our CPT efforts would have converged to.*
- Both of the models helped in facilitating our exploration into how the quality of a CPT checkpoint impacts the task performance of a downstream merged model.

FOOTNOTES:

[2-1] Wu, Chaoyi, et al. "Pmc-llama: Towards building open-source language models for medicine." arXiv preprint arXiv:2305.10415 6 (2023).

[1-1] Gupta, Kshitij, et al. "Continual Pre-Training of Large Language Models: How to (re) warm your model?." arXiv preprint arXiv:2308.04014 (2023).

[7, 7-1] Chen, Zeming, et al. "Meditron-70b: Scaling medical pretraining for large language models." arXiv preprint arXiv:2311.16079 (2023).

Patent Domain:

- A similar approach was taken in the patent domain, adapting the methodology to the unique content and requirements of the United States Patent and Trademark Office (USPTO) dataset.[8] We took 10B patent tokens, as well as general tokens to reduce catastrophic forgetting, and did continual pre-training runs using Llama-2-7B as a base model. This resulted in a domain-adapted 7B patent model that performed exceptionally well on patent QA (synthetically generated), which was synthetically generated by us using new patents (held out patents), much better than a closed-source model with the same query.
- The model training was conducted in accordance with the DOREMI[9] settings, blending domain-specific data with a broad dataset of general red pajama data, totaling 30 billion tokens.
- We also created an instruction-tuned version of our domain-adapted patent base model with the use of a synthetically-generated instruction dataset.

Merging Stage

Leveraging Mergekit, we explored various merging techniques, such as Linear[10], SLERP[11], TIES[6–1], and DARE[4–1] to integrate our CPT checkpoints with general-purpose chat models. Model Merging maintains a balance between general and domain-specific knowledge while mitigating the risk of catastrophic forgetting, as the weights in the foundational general model can remain frozen. This stage was crucial for enhancing the model's adaptability and performance in specific domains.

FOOTNOTES:

[8] Marco, Alan C., et al. "The USPTO patent assignment dataset: Descriptions and analysis." (2015).

[9] Xie, Sang Michael, et al. "Doremi: Optimizing data mixtures speeds up language model pretraining." Advances in Neural Information Processing Systems 36 (2024).

[10] Wortsman, Mitchell, et al. "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time." International Conference on Machine Learning. PMLR, 2022.

[11] https://github.com/Digitous/LLM-SLERP-Merge

[6:1] Yadav, Prateek, et al. "Ties-merging: Resolving interference when merging models." Advances in Neural Information Processing Systems 36 (2024).

[4:1] Yu, Le, et al. "Language models are super mario: Absorbing abilities from homologous models as a free lunch." arXiv preprint arXiv:2311.03099 (2023).

IV. Experiments and Results

Our research assessed the effectiveness of Continual Pre-Training (CPT) models and model merging strategies in the medical and patent domains. Performance of our final merged models on medical and patent benchmarks, showcasing our pipeline's' ability to adapt to a certain domain.

Relationship Between Medical and General Benchmarks and Checkpoint Steps

To assess the quality of our CPT efforts, we focused on the medical domain, recognizing the Meditron-7B[7-2] checkpoint for its superior refinement and domain-specific performance. This checkpoint served as a benchmark for evaluating the effectiveness of our CPT process. Our analysis spanned medical and general benchmarks[12]: USMLE, MedMCQA, PubMedQA, Arc Challenge, HellaSwag, MMLU.

	Medical Benchmarks				General Benchmarks		
CPT Checkpoint	**USMLE**	**MedMCQA**	**PubMedQA**	**Perplexity**	**Arc Challenge**	**HellaSwag**	**MMLU**
27B tokens chkpt	37.6	28.31	73.6	5.35	41.7	55.4	40.75
Final (Meditron 1.5 epochs)	38.96	30.93	76.2	4.33	43.94	57.5	45.03

FOOTNOTES:

[7:2] Chen, Zeming, et al. "Meditron-70b: Scaling medical pretraining for large language models." arXiv preprint arXiv:2311.16079 (2023).

[12] Gilson, Aidan, et al. "How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment." JMIR Medical Education 9.1 (2023): e45312.

Observations:

- Better CPT checkpoints can improve the final results after the merging stage.
- Final evaluation with the Meditron checkpoint emphasized the importance of carefully selected CPT settings and high-quality datasets.
- Comparative results revealed that the quality of CPT checkpoints is vital for superior model performance after merging.

Which Merge Methods Work Well for the Medical Domain?

With a selection of refined checkpoints at hand, our next goal was to determine the most effective Model Merging techniques for the medical domain. We experimented with various methods, including SLERP, TIES, and Linear, to merge the Meditron-7B[7–3] checkpoint with Llama2-chat models, the base model of both being the Llama2 base model.

	Medical Benchmarks			General Benchmarks		
Method	**USMLE**	**MedMCQA**	**PubMedQA**	**Arc Challenge**	**HellaSwag**	**MMLU**
Llama2 7B Chat [11]	35.9	35.45	73.4	44.2	55.4	46.37
Meditron -7B [9]	38.4	24.07	71.4	40.2	54.5	33.06
Linear	39.1	36.65	75.6	46.76	58.66	48.44
Slerp	39.2	36.91	75.6	46.84	58.67	47.97
Dare-Ties	36.37	27.56	72.2	42.92	54.79	41.17
Ties	38.73	32.27	75.6	45.05	58.23	45.03

FOOTNOTES:
[7:3] Chen, Zeming, et al. "Meditron-70b: Scaling medical pretraining for large language models." arXiv preprint arXiv:2311.16079 (2023).

Observations

- The region between and around the models seems to be filled with low loss models as verified by evaluating various exploratory configurations.[5:1]
- Linear interpolation (Lerp) merge and Spherical Interpolation (Slerp) merge end up doing the same weight interpolation because the weights represented as flattened vectors are pretty much collinear (at which point Slerp transforms to Lerp).
- As per observations, linear merge works the best in this scenario.
- Ties merge method doesn't have any edge here as the task vectors for the two models (Meditron and Llama-7B-chat) relative to the Llama2 base model are mostly orthogonal and as a consequence there is very little interference to resolve. And in the scenario of conflict resolution, Meditron larger magnitude most likely wins statistically maintaining Meditron as the larger contributor as is the case with the Lerp scenario
- Through various experiments with different Ties-merging configurations, reducing the threshold value which is a redundant weights filter seems to add to performance making for better evals (especially when it comes to the thresholding Meditron associated task vector). This could be explained by the product space distance between Meditron and Llama2-chat-hf is vast enough that: 1) The weight distributions are different, and 2) that missing portions of either model induced by the Ties mechanism (trimming) cannot be made up by fractional values by the corresponding model's delta weights especially in the case when the induced values are those of Meditron's.

FOOTNOTES:

[5:1] Stoica, George, et al. "ZipIt! Merging Models from Different Tasks without Training." arXiv preprint arXiv:2305.03053 (2023).

Comparing the Patent Domain Checkpoints

As part of our exploration, we also worked in the patent domain. This area, unlike the medical domain, lacks public benchmarks–leading us to develop our own evaluations for this domain.

We introduced two benchmarks: 1) generating a patent abstract summary, and 2) a closed-book Question Answering where the questions were synthetically generated using a powerful LLM and manually-curated prompts. The former involves feeding the system the main content of a patent and requesting it to produce an abstract. The latter benchmark entails generating synthetic question-answer pairs and assessing the model's capability to provide accurate answers.

We use perplexity as the base primary metric in the absence of human expert labels. Using classical long text generation task metrics like ROUGE/BLUE are known to provide inconsistent correlations with actual quality for complex domains such as legal texts.

The results, detailed in the table below, compare the performance of domain-adapted patent models and instruction-tuned versions against the baseline performances of the LLama2-7B model and Llama2-7B chat. This comparison highlights the tailored models' effectiveness in navigating the complex and nuanced patent domain.

Model	Abstract Generation (pplg)	Q&A (ppl)
LLama2 7b	13.37	41.7
LLama2 7B Chat	8.5467	39.1
Patent-Base-7b (fp16)	11.6	33.8
Patent-Instruct -7b (fp16)	10.5	26.5

Test Set: U.S. Patents from 8/2023-12/2023

Which Merge Methods Work Well for the Patent Domain?

Similar to the medical domain, we also explore different merging methods with our CPT and Instruction checkpoints.

Model 1	Model 2	Merge Type	Abstract Gen (ppl)	Q&A (ppl)
LLama2 7B Chat	Patent - Instruct -7b (fp16)	Linear	11.6	23.5
LLama2 7B Chat	Patent - Instruct -7b (fp16)	Slerp	11.62	23.56
LLama2 7B Base +LLama2 7B chat	Patent - Instruct -7b (fp16)	Ties	10.52	22.56

Patent Merged Model

Observations:

- As with the medical domain, the immediate region between the models seems to be filled with low loss models as verified by evaluating different configurations.
- As with the previous scenario (medical domain), Linear interpolation (Lerp) merge and Spherical Interpolation (Slerp) merge end up doing the weight interpolation because the weights represented as flattened vectors are collinear.
- In contrast to the previous domain where Ties-merging did not seem to make a significant positive difference, here it does make a significant difference. This is because the distance between Llama2-chat and patent-instruct is somewhat the same relative to the Llama2-base model.

VI. Conclusion

The integration of Continual Pre-Training and Model Merging at Arcee.ai represents a significant leap forward in domain adaptation. Our case studies in the Medical and Patent domains demonstrate the potential of these methodologies to enhance the relevance and performance of language models across specialized fields. By leveraging domain-specific data, existing open-source checkpoints, and innovative Model Merging techniques, we are delivering cost-effective, high-quality models tailored to our clients' unique needs.

References

1. Gupta, Kshitij, et al. "Continual Pre-Training of Large Language Models: How to (re) warm your model?." arXiv preprint arXiv:2308.04014 (2023).
2. Wu, Chaoyi, et al. "Pmc-llama: Towards building open-source language models for medicine." arXiv preprint arXiv:2305.10415 6 (2023).
3. Liu, Mingjie, et al. "Chipnemo: Domain-adapted llms for chip design." arXiv preprint arXiv:2311.00176 (2023).
4. Yu, Le, et al. "Language models are super mario: Absorbing abilities from homologous models as a free lunch." arXiv preprint arXiv:2311.03099 (2023).
5. Stoica, George, et al. "ZipIt! Merging Models from Different Tasks without Training." arXiv preprint arXiv:2305.03053 (2023).
6. Yadav, Prateek, et al. "Ties-merging: Resolving interference when merging models." Advances in Neural Information Processing Systems 36 (2024).
7. Chen, Zeming, et al. "Meditron-70b: Scaling medical pretraining for large language models." arXiv preprint arXiv:2311.16079 (2023).
8. Marco, Alan C., et al. "The USPTO patent assignment dataset: Descriptions and analysis." (2015).
9. Xie, Sang Michael, et al. "Doremi: Optimizing data mixtures speeds up language model pretraining." Advances in Neural Information Processing Systems 36 (2024).
10. Wortsman, Mitchell, et al. "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time." International Conference on Machine Learning. PMLR, 2022.
11. https://github.com/Digitous/LLM-SLERP-Merge
12. Gilson, Aidan, et al. "How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment." JMIR Medical Education 9.1 (2023): e45312.


arcee.ai