MoE Meets Reparameterization: Reparameterizable Mixture-of-Experts Model Enhances Dermatology Diagnosis via Dense-to-Experts Distillation

Anonymous authors
Paper under double-blind review

1 Related Work

Mixture-of-Experts. Early sparsely gated MoE layers demonstrated that conditional parameter utilisation can enlarge model capacity without proportional compute (Shazeer et al., 2017). GShard scaled this idea to ~ 600 B parameters in multilingual translation by integrating automatic tensor sharding with expert routing (Lepikhin et al., 2020). Switch Transformer further simplified routing to one-expert selection and trained a trillion-parameter language model with reduced communication overhead (Fedus et al., 2022). GLaM extended sparsity to multilingual web corpora and showed strong zero-shot generalisation (Du et al., 2022). Subsequent open-source efforts improved stability and transfer: ST-MoE introduced noise-based gating to equalise expert load (Zoph et al., 2022), OpenMoE provided modular implementations for continual pre-training (Xue et al., 2024), and MegaScale-MoE proposed pipeline parallelism for production-scale deployment (Jin et al., 2025). Auxiliary regularisers alleviate expert imbalance and collapse, for example expert-choice routing encourages uniform utilisation (Zhou et al., 2022), gating dropout reduces communication cost (Liu et al., 2022), and residual MoE blends dense and sparse paths (Wu et al., 2022a). Vision transformers later adopted sparse experts: V-MoE matched the accuracy of large dense ViT-G while saving FLOPs through token-level routing (Riquelme et al., 2021), M^{3}ViT co-designed accelerator-aware MoE blocks for multi-task learning (Fan et al., 2022), ViMoE explored depth-wise expert placement (Han et al., 2024), and Mobile V-MoE delivered mobile-scale latency by weight pruning and sparsity (Daxberger et al., 2023).

Reparameterisation. Structural reparameterisation trains over-parameterised multi-branch blocks then algebraically merges them into single-branch kernels for inference efficiency. ACNet first showed asymmetric convolutions can be fused into 3×3 filters for robustness (Ding et al., 2019). RepVGG stacked 3×3 convolutions during deployment and achieved $> 80\%$ ImageNet top-1 with VGG-style speed (Ding et al., 2021c). Diverse Branch Block generalised fusion to heterogeneous branches for improved generalisation (Ding et al., 2021b). ExpandNets applied linear over-parameterisation to compress student CNNs after training (Guo et al., 2020). ResRep decoupled "remembering" and "forgetting" filters to prune channels losslessly (Ding et al., 2021a). FastViT combined RepMixer token mixing with ViT blocks to reduce latency and improve accuracy (Vasu et al., 2023). RepFormer extended the idea to facial landmark transformers by fusing pyramid heads (Li et al., 2022).

Dermatology Vision Models. Esteva *et al.* trained an end-to-end CNN on $\sim 130{,}000$ clinical and dermoscopic images and reached dermatologist-level melanoma detection (Esteva et al., 2017). Subsequent reader studies confirmed AI superiority over clinicians in large cohort benchmarks (Brinker et al., 2019). Human–computer collaboration improved overall sensitivity by combining visual saliency with expert opinion (Tschandl et al., 2020). Reinforcement learning further aligned predictions with clinical risk preferences (Barata et al., 2023). Self-supervised pre-training on 75 M radiology and dermatology images advanced few-shot transfer (Azizi et al., 2021). Federated self-supervised contrastive learning preserved privacy across mobile dermatology assistants (Wu et al., 2022b). SwAVDerm mined unannotated community images via contrastive clustering and improved long-tail disease coverage (Shen et al., 2024). PanDerm combined slide-level and pixel-level supervision over > 2 M images plus 45 K pathology reports to set new state of the art across multiple benchmarks (Yan et al., 2025).

REFERENCES

Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3478–3488, 2021.

Catarina Barata, Veronica Rotemberg, Noel CF Codella, Philipp Tschandl, Christoph Rinner, Bengu Nisa Akay, Zoe Apalla, Giuseppe Argenziano, Allan Halpern, Aimilios Lallas, et al. A reinforcement learning model for ai-based decision support in skin cancer. *Nature Medicine*, 29 (8):1941–1946, 2023.

Titus J Brinker, Achim Hekler, Alexander H Enk, Joachim Klode, Axel Hauschild, Carola Berking, Bastian Schilling, Sebastian Haferkamp, Dirk Schadendorf, Tim Holland-Letz, et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer*, 113:47–54, 2019.

Erik Daxberger, Floris Weers, Bowen Zhang, Tom Gunter, Ruoming Pang, Marcin Eichner, Michael Emmersberger, Yinfei Yang, Alexander Toshev, and Xianzhi Du. Mobile v-moes: Scaling down vision transformers via sparse mixture-of-experts. *arXiv preprint arXiv:2309.04354*, 2023.

Xiaohan Ding, Yuchen Guo, Guiguang Ding, and Jungong Han. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1911–1920, 2019.

Xiaohan Ding, Tianxiang Hao, Jianchao Tan, Ji Liu, Jungong Han, Yuchen Guo, and Guiguang Ding. Resrep: Lossless cnn pruning via decoupling remembering and forgetting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4510–4520, 2021a.

Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Diverse branch block: Building a convolution as an inception-like unit. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10886–10895, 2021b.

Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13733–13742, 2021c.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International conference on machine learning*, pp. 5547–5569. PMLR, 2022.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

Zhiwen Fan, Rishov Sarkar, Ziyu Jiang, Tianlong Chen, Kai Zou, Yu Cheng, Cong Hao, Zhangyang Wang, et al. M^3vit: Mixture-of-experts vision transformer for efficient multi-task learning with model-accelerator co-design. *Advances in Neural Information Processing Systems*, 35:28441–28457, 2022.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Shuxuan Guo, Jose M Alvarez, and Mathieu Salzmann. Expandnets: Linear over-parameterization to train compact convolutional networks. *Advances in Neural Information Processing Systems*, 33:1298–1310, 2020.

Xumeng Han, Longhui Wei, Zhiyang Dou, Zipeng Wang, Chenhui Qiang, Xin He, Yingfei Sun, Zhenjun Han, and Qi Tian. Vimoe: An empirical study of designing vision mixture-of-experts. *arXiv preprint arXiv:2410.15732*, 2024.

Chao Jin, Ziheng Jiang, Zhihao Bai, Zheng Zhong, Juncai Liu, Xiang Li, Ningxin Zheng, Xi Wang, Cong Xie, Wen Heng, et al. Megascale-moe: Large-scale communication-efficient training of mixture-of-experts models in production. *arXiv preprint arXiv:2505.11432*, 2025.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

Jinpeng Li, Haibo Jin, Shengcai Liao, Ling Shao, and Pheng-Ann Heng. Repformer: Refinement pyramid transformer for robust facial landmark detection. *arXiv preprint arXiv:2207.03917*, 2022.

Rui Liu, Young Jin Kim, Alexandre Muzio, and Hany Hassan. Gating dropout: Communication-efficient regularization for sparsely activated transformers. In *International Conference on Machine Learning*, pp. 13782–13792. PMLR, 2022.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Yue Shen, Huanyu Li, Can Sun, Hongtao Ji, Daojun Zhang, Kun Hu, Yiqi Tang, Yu Chen, Zikun Wei, and Junwei Lv. Optimizing skin disease diagnosis: harnessing online community data with contrastive learning and clustering techniques. *NPJ Digital Medicine*, 7(1):28, 2024.

Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. Human–computer collaboration for skin cancer recognition. *Nature medicine*, 26(8):1229–1234, 2020.

Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. Fastvit: A fast hybrid vision transformer using structural reparameterization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5785–5795, 2023.

Lemeng Wu, Mengchen Liu, Yinpeng Chen, Dongdong Chen, Xiyang Dai, and Lu Yuan. Residual mixture of experts. *arXiv preprint arXiv:2204.09636*, 2022a.

Yawen Wu, Dewen Zeng, Zhepeng Wang, Yi Sheng, Lei Yang, Alaina J James, Yiyu Shi, and Jingtong Hu. Federated self-supervised contrastive learning and masked autoencoder for dermatological disease diagnosis. *arXiv preprint arXiv:2208.11278*, 2022b.

Fuzhao Xue, Zian Zheng, Yao Fu, Jinjie Ni, Zangwei Zheng, Wangchunshu Zhou, and Yang You. Openmoe: An early effort on open mixture-of-experts language models. *arXiv preprint arXiv:2402.01739*, 2024.

Siyuan Yan, Zhen Yu, Clare Primiero, Cristina Vico-Alonso, Zhonghua Wang, Litao Yang, Philipp Tschandl, Ming Hu, Lie Ju, Gin Tan, et al. A multimodal vision foundation model for clinical dermatology. *Nature Medicine*, pp. 1–12, 2025.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114, 2022.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*, 2022.