Are Natural Domain Foundation Models Useful For Medical Image Classification?

Joana Palés Huix[1,2,3,†], Adithya Raju Ganeshan[1,2,†], Johan Fredin Haslum[1,2,4], Magnus Söderberg[3], Christos Matsoukas[1,2,3], Kevin Smith[1,2]

[1] KTH Royal Institute of Technology, Stockholm, Sweden [2] Science for Life Laboratory, Stockholm, Sweden [3] Cardiovascular, Renal and Metabolism Pathology, Clinical Pharmacology & Safety Sciences, R&D AstraZeneca, Gothenburg, Sweden [4] Cell & Molecular Pharmacology, Research and Early Development, Cardiovascular, Renal and Metabolism, BioPharmaceuticals R&D, AstraZeneca, Gothenburg, [†] Equal contribution

Deep learning is converging towards the use of foundation models. We attempt to investigate the transferability of these to medical image classification tasks. We explore different training settings to fully harness their potential and gain insights on their applicability in this domain.

1 MODELS

Foundation models:

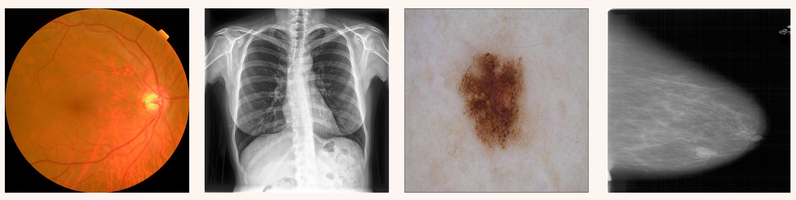
- SAM (88.8M parameters) [1]
- SEEM (29.9M parameters) [2]
- DINOv2 (86.5M parameters) [3]
- OpenCLIP (86.2M parameters) [4]
- BLIP (85.7M parameters) [5]

Baselines:

- ImNet-1k-Sup (DeiT) [6]
- DINOv1 [7]
- ResNet152 [8]

2 METHODS

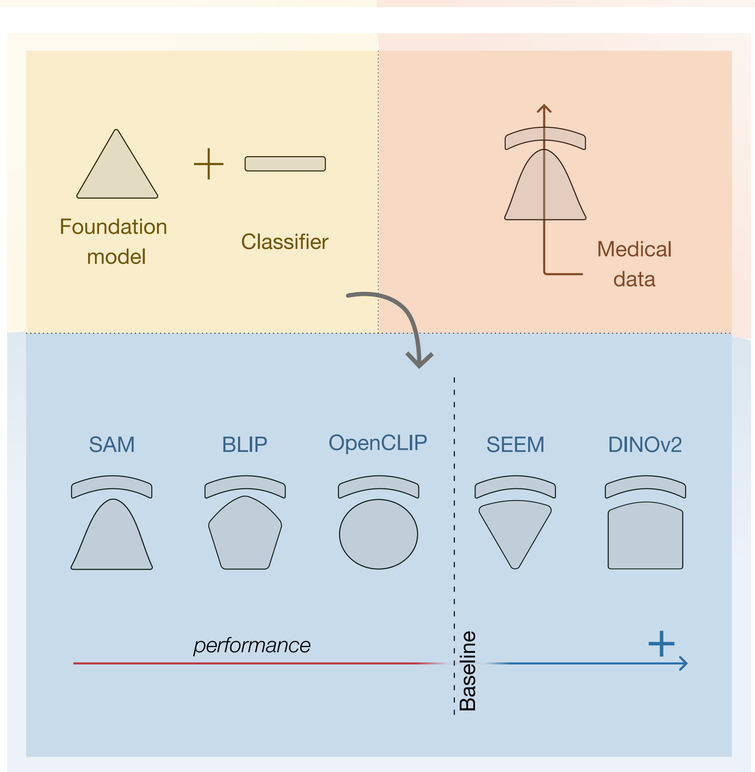
Medical datasets:



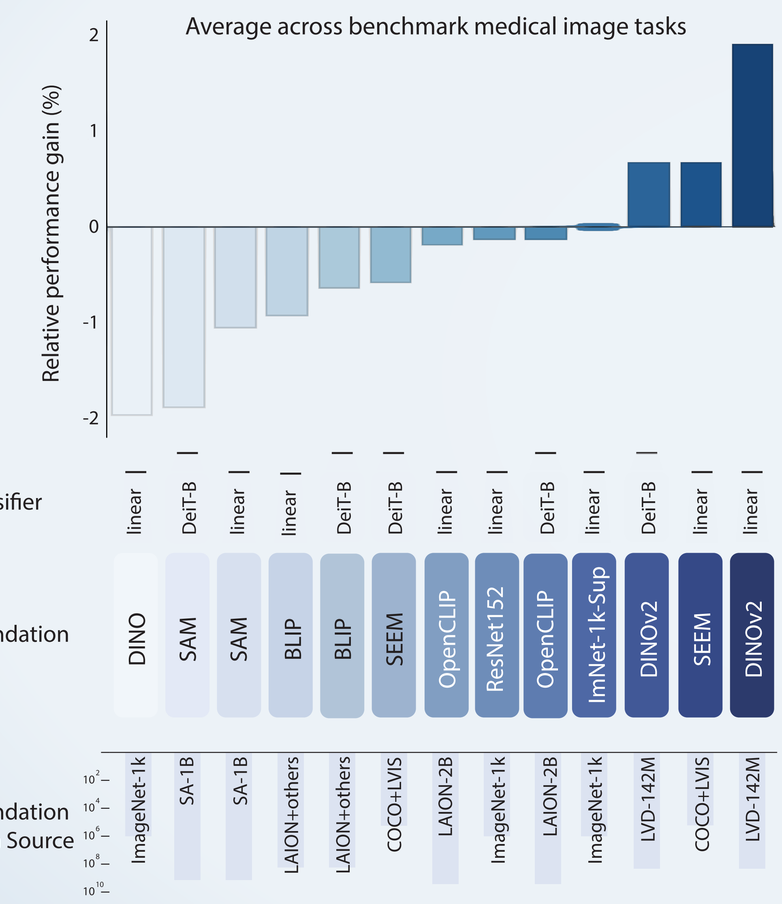
APTOS2019 CheXpert ISIC2019 DDSM

Evaluation scenarios:

- Frozen foundation with linear head
- Unfrozen foundation with linear head
- Frozen foundation with appended DeiT classifier
- Unfrozen foundation with appended DeiT classifier



3 RESULTS



KEY FINDINGS:

- DINOv2 serves as a solid base for transfer learning and SEEM competes effectively while being a smaller model. Other foundations fail to outperform baselines effectively.
- DINOv2 converges faster compared to the other models.
- Adapting the foundation models to downstream tasks, especially their deeper layers, is key.
- Appending a DeiT classifier to the foundation models results in marginal performance gains.
- Both architectural differences and pre-training objectives significantly influence transfer learning success.

REFERENCES

[1] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., ... & Girshick, R. (2023). Segment anything. arXiv:2304.02643.
[2] Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Gao, J., & Lee, Y. J. (2023). Segment everything everywhere all at once. arXiv:2304.06718.
[3] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ... & Bojanowski, P. (2023). Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193.
[4] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PMLR.
[5] Li, J., Li, D., Xiong, C., & Hoi, S. (2022, June). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International Conference on Machine Learning (pp. 12888-12900). PMLR.
[6] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021, July). Training data-efficient image transformers &distillation through attention. In International conference on machine learning (pp. 10347-10357). PMLR.
[7] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., ... & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 9650-9660).
[8] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference oncomputer vision and pattern recognition (pp. 770-778).

READ ME!

AstraZeneca

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM