Making the Most of Limited Data with Self-Supervised Learning for Breast Cancer Screening

Christopher Clark[1], Scott Kinder[1], Syed Rakin Ahmed[2], Giacomo Nebbia[1], Yoga Advaith Veturi[1], Praveer Singh[1], and Jayashree Kalpathy-Cramer[1]

[1] University of Colorado School of Medicine, USA
{christopher.w.clark, scott.kinder, giacomo.nebbia, yogaadvaith.veturi praveer.singh, jayashree.kalpathy-cramer}@cuanschutz.edu
[2] Harvard University, USA
syedrakin_ahmed@fas.harvard.edu

Keywords: Conformal Prediction · Computer Vision · Breast Cancer

1 Introduction

The collection of high-quality, *labeled* images for medical classification remains a difficult and costly endeavor [6], making standard *Supervised Learning* (SL) medical classification tasks difficult. With *Self-Supervised Learning* (SSL), features can be learned from unlabeled data and fine-tuned on limited labeled samples. In this work, we compare SL methods with SSL methods at fractionations of 1, 50, and 100% of the labeled data. Further, we compare pre-training SSL methods on our unlabeled data with using off-the-shelf ImageNet pre-trained weights only. We run experiments on the *Digital Mammographic Imaging Screening* (DMIST) dataset [5]. Through these fractionations, we determine that SSL models trained with a large amount of unlabeled is most beneficial to the downstream classification task in a low-labeled-data regime, but remains a top performer even with a large number of labeled images.

2 Methods

We are using two SSL training methods, *knowledge DIstillation with NO labels* (DINOv1) [1] and *Masked AutoEncoder* (MAE) [3], a contrastive and generative method respectively. The DINOv1 method has as a backbone a CNN, *ResNet50* [4], and a *Vision Transformer* (ViT) [2]. The MAE uses a ViT, as well. We have 83,039 unlabeled images for the SSL pretext finetuning and 72,606 images for the downstream classification training task.

3 Results

In Table 1, we show the 1, 50, and 100% fractionations of the labeled data classification results using the linearly weighted κ metric for four-class classification.

We see the top models by average of five runs at the 1% fractionation is the DINOv1 ResNet50 with in-domain pretext finetuning and the DINOv1 ViTb16 pretext finetuned the same way, followed by the DINOv1 ResNet50 without in-domain pretext finetuning. However, by the 50 and 100% fractionations, the large lead is not present.

Table 1: Comparison of Models at 1%, 50%, and 100% Fractionation Levels. Each cell represents the average linearly weighted κ score with the lower and upper bounds in parentheses. Each model is initialized with ImageNet weights.

Model	1%	50%	100%
SL ResNet50	0.32 (0.19, 0.46)	0.54 (0.54, 0.55)	0.56 (0.55, 0.57)
SL VIT	0.45 (0.43, 0.47)	0.58 (0.57, 0.58)	0.59 (0.59, 0.60)
SSL ViTMAE	0.47 (0.46, 0.49)	0.58 (0.58, 0.59)	0.59 (0.59, 0.59)
SSL ViTMAE Domain	0.53 (0.52, 0.54)	0.59 (0.58, 0.59)	0.60 (0.60, 0.60)
SSL DINO ResNet50	0.54 (0.53, 0.55)	0.57 (0.56, 0.58)	0.59 (0.58, 0.59)
SSL DINO ResNet50 Domain	0.55 (0.54, 0.56)	0.58 (0.58, 0.59)	0.59 (0.58, 0.60)
SSL DINO ViT	0.48 (0.46, 0.49)	0.56 (0.55, 0.57)	0.58 (0.58, 0.59)
SSL DINO ViT Domain	0.55 (0.54, 0.55)	0.58 (0.57, 0.58)	0.59 (0.59, 0.59)

4 Discussion

Table 1 demonstrates the value of SSL, both with and without in-domain pretext finetuning, for overcoming a lack of labeled data needed for pure SL approaches to classification on the DMIST dataset. It also argues that even with a large amount of labeled data, these models are at least as competitive, if not more performative, than standard SL models.

References

1. Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.
2. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
3. Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.
4. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

5. Etta D Pisano, Constantine Gatsonis, Edward Hendrick, Martin Yaffe, Janet K Baum, Suddhasatta Acharyya, Emily F Conant, Laurie L Fajardo, Lawrence Bassett, Carl D'Orsi, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. *New England Journal of Medicine*, 353(17):1773–1783, 2005.
6. Luciano M Prevedello, Safwan S Halabi, George Shih, Carol C Wu, Marc D Kohli, Falgun H Chokshi, Bradley J Erickson, Jayashree Kalpathy-Cramer, Katherine P Andriole, and Adam E Flanders. Challenges related to artificial intelligence research in medical imaging and the importance of image analysis competitions. *Radiology: Artificial Intelligence*, 1(1):e180031, 2019.