Contrastive learning for unsupervised medical image clustering and reconstruction

Matteo Ferrante*[1], Tommaso Boccato[1], Andrea Duggento[1], Simeon Spasov[2], and Nicola Toschi[1,3]

[1]Department of Biomedicine and Prevention, University of Rome, Tor Vergata, Rome (RM), Italy
[2]Department of Computer Science and Technology, University of Cambridge, Cambridge, UK
[3]Martinos Center for Biomedical Imaging, MGH and Harvard Medical School (USA), Boston, USA

Abstract

The lack of large labeled medical imaging datasets, along with significant interindividual variability compared to clinically established disease classes, poses significant challenges in exploiting medical imaging information in a precision medicine paradigm, where in principle dense patient-specific data can be employed to formulate individual predictions and/ or stratify patients into finer-grained groups that may follow more homogeneous trajectories and therefore empower clinical trials. In order to efficiently explore the effective degrees of freedom underlying variability in medical images in an unsupervised manner, in this work, we propose an unsupervised autoencoder which is augmented with a contrastive loss to encourage high separability in the latent space. The model is validated on (medical) benchmark datasets. As the cluster labels are assigned to each example according to the cluster assignments, we compare performance with a supervised transfer learning baseline. Our methods achieve performance similar to the supervised architecture, indicating that separation in the latent space reproduces expert medical observer-assigned labels. The proposed method could be beneficial for patient stratification, exploring new subdivision of larger classes or pathological continua, or, due to its sampling abilities in a variation setting, data augmentation in medical image processing.

1 Introduction

In current medical practice, the difficulties in specifically targeting any disease process are rooted in recent evidence showing that current diagnostic categories do not actually represent a single disease, but rather heterogeneous clinical syndromes underpinned by different pathogenic mechanisms. Today, we still do not know how heterogeneity in organ function and anatomy is linked to this clinical variability and to the risk of following different 'trajectories'. This represents a drawback in personalizing diagnosis and therapy, which is commonly based on a 'one size fits all' approach. This lack of understanding of the basic mechanisms underlying (multimorbid) syndromes is a major roadblock to the development of the P4[1] medicine paradigm today, as well as in designing cost- and discovery- efficient clinical trials. For these reasons, there is significant interest in developing unsupervised methods that can discover clinical subtypes (or, more generally, more fine-grained patient strata) in patient populations based on patient data [11]. In this paper, we propose an unsupervised framework for image-based patient stratification based on an autoencoder network. We augment the reconstruction loss of the autoencoder with a contrastive learning

*Corresponding Author: matteo.ferrante@uniroma2.it

[1]P4: predictive, preventative, personalized, participatory

https://doi.org/10.7557/18.6819

component inspired by [5, 6] to encourage better separation of the latent space [14, 1, 3, 10]. In the first stage of training (warmup), we focus on learning structured latent representations, while in the second stage we fine-tune the decoder for reconstruction. Using a simple function to map between cluster labels and real labels, we are able to produce classification results close to a ResNet18 [12] supervised baseline, and outperform a feature extraction (last feature layer of ResNet18 combined with KMeans clustering) baseline.

Other recent work addressed the use of constrastive learning on biomedical images for clustering and patient stratification. For example, in [4] the authors propose strategies for extending the contrastive learning framework for the segmentation of volumetric medical images, such as magnetic resonance or computed tomography scans. The proposed method uses domain-specific and problem-specific cues to improve the performance of the contrastive learning framework in a semisupervised setting, where only a limited amount of labeled data are available. One key contribution is the use of domain-specific cues to improve the contrastive learning process, leveraging the inherent structure and similarity in the data to provide more complex similarity cues than data augmentation alone can provide. The method is evaluated on three MRI datasets and yields substantial improvements compared to other self-supervision and semi-supervised learning techniques. Again on segmentation, in [18] the authors propose two federated self-supervised learning frameworks for medical image segmentation with limited annotations. The first framework is suitable for high-performance servers, while the second is more suitable for mobile devices. Both frameworks use self-supervised contrastive learning followed by fine-tuning with limited annotations. Experiments on a cardiac magnetic resonance data set show that the proposed frameworks improve segmentation and generalization performance compared to state-of-the-art techniques. In [17] the authors propose a new method for content-based whole-slide image (WSI) retrieval, called Retrieval with Clustering-guided Contrastive Learning (RetCCL). The RetCCL framework combines a self-supervised feature learning method with a global ranking and aggregation algorithm to improve the performance of WSI-level image retrieval. The feature learning

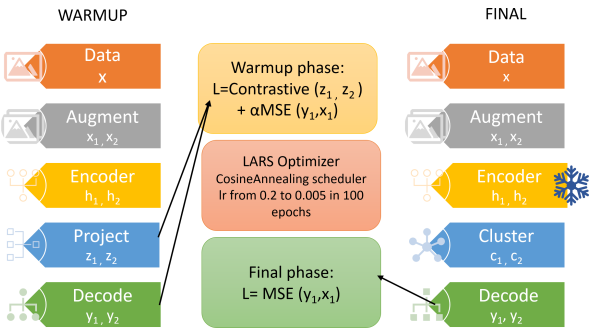

Figure 1: A diagram of the proposed training scheme based for our architecture. The first column, is related to the warmup phase and information flow from top to bottom, so data, augmentation, encoding, projection and decoding. Projected and decoded representation are used in the loss for the warmup phase and all the components are updated. The central column shows the loss function for the first warmup phase (yellow box) and the loss for the last phase of the training (green box). The right column report the same scheme as the left one, but with frozen encoder. This latter configuration is used in the last phase of training, where only the decoder is updated and the loss function is only the reconstruction term.

method uses large-scale unlabeled histopathological image data to learn universal features that can be used directly for WSI retrieval tasks without additional fine-tuning. Finally, in [21] the authors propose ConVIRT, an unsupervised strategy for learning visual representations of medical images using paired descriptive text. ConVIRT uses a bidirectional contrastive objective to pretrain medical image encoders without requiring additional expert input. The authors evaluated ConVIRT on four medical image classification tasks and two zero-shot retrieval tasks, showing that it outperforms strong baselines in most settings and demonstrates superior data efficiency. While contrastive learning can be successfully used to solve many medical problems, from segmentation to patient statification. To solve this latter task, we propose a simple method based on a constastive loss in a two-step learning process.

2 Material and Methods

Our architecture has the dual objective of reconstructing images while generating a latent space whose structure separation is driven by the contrastive loss. The encoder e (a convolutional network with 3 2D layers, kernel size=4, stride=2, fol-

lowed by GELU activations and an average pooling layer (kernel size= 2) and a linear layer that maps the features into a latent space of dimension 128) maps images x to the latent space $h = e(x)$, while the projector p (a multilayer perceptron with 3 layers) projects the images into another space ($z = p(h)$) where the similarity function $\text{sim}(\cdot, \cdot)$ is computed as in [5]). The encoded representations are then passed to a decoder that learns how to reconstruct the images driven by the reconstruction loss. Parameter values were chosen to halve the image dimensions three times while increasing the receptive field, in order to create filters that process at the entire image feature map before the linear layers. During the '*warmup*' (first) phase of training, the contrastive loss term was applied.

$$\mathcal{L}_{contrastive} = -\log\left(\frac{e^{\frac{sim(z_i,z_j)}{\tau}}}{\sum_{negatives} e^{\frac{sim(z_i,z_k)}{\tau}}}\right) \quad (1)$$

and a standard mean squared error loss;

$$\mathcal{L} = \mathcal{L}_{contrastive} + \alpha\mathcal{L}_{recon} \quad (2)$$

are combined into the total loss , where $\alpha = 0.1$ is a fixed scalar term that encourages learning those features which are potentially also useful for reconstruction. All network weights are updated with the LARS optimizer [20]. Then, the encoder is frozen and a search for optimal number of clusters is run using KMeans combined with the elbow method. [15, 2]. This optimizer has shown superior performances in training models with large batch sizes, so we choose it instead of a standard Stocastic Gradient Descent or Adam. Cluster centroids are then stored in a matrix whose columns represent cluster *prototypes*. In the second phase, the decoder is turned into a conditional decoder $d(\cdot)$ by adding layers that process information from a soft label assignment computed in the latent space. The images are then encoded and their representation are compared to each of the prototypes using cosine distance as similarity metric. The temperature-scaled softmax of the vector of similarities produces a soft label assignment passed to the decoder that will learn how to decode images including this information. The architecture of $d(\cdot)$ is symmetrical with respect to $e(\cdot)$.

After obtaining proof-of-concept results on the MNIST [8] digit images, we validate our approach

Algorithm 1 Contrastive Learning

for epoch in warmup epochs **do**
  for x, y in dataloader **do**
    x_1, x_2 =augment(x)
    h_1, h_2 =encode(x_1,x_2)
    z_1, z_2 = project(h_1,h_2)
    y_1, y_2=decode(h_1,h_2)
    $\mathcal{L}_{sim}$=Contrastive(z_1,z_2)
    $\mathcal{L}_{recon}$=MSE(x_1,y_1)+MSE(x_2,y_2)
    $\mathcal{L} = \mathcal{L}_{sim} + \alpha\mathcal{L}_{recon}$
    update(e,p,d)
  end for
end for
for epoch in(warmup_epoch,total_epochs) **do**
  for x, y in dataloader **do**
    x_1, x_2=augment(x)
    h_1, h_2=encode(x_1,x_2)
    c_1, c_2=clusters_labels(h_1, h_2)
    y_1, y_2=decode(h_1, h_2, c_1, c_2)

    $\mathcal{L}$=MSE(x_1, y_1)+MSE(x_2, y_2)
    update(d)
  end for
end for

on the *pneumoniaMNIST* dataset (2D chest X-Ray images labelled as healthy or affected by pneumonia), part of MedMNIST (`https://medmnist.com/`) [19]. We generate *positive* pairs through augmentation, that is, we randomly apply (with probability $p = 0.5$), rotations of up to 30 degrees, Gaussian blur, Gaussian noise, horizontal flips, and randomly rescaled crops. For all images in the training set, we compute the encoded latent representations and their cluster labels. We then map each cluster into the mode of the real labels of items belonging to that specific cluster to generate a map between the cluster labels and the real labels. The latter map is used to predict the labels for the validation and the test set and compute the performance in a way that is comparable to what is done in a supervised model. An outline of the algorithm is shown below.

We also trained a supervised baseline by performing transfer learning from a ResNet-18 architecture pre-trained on ImageNet [7] while changing the number of units in the last fully connected layer to two. This model was trained while freezing all parameters before the avgpool layer (number

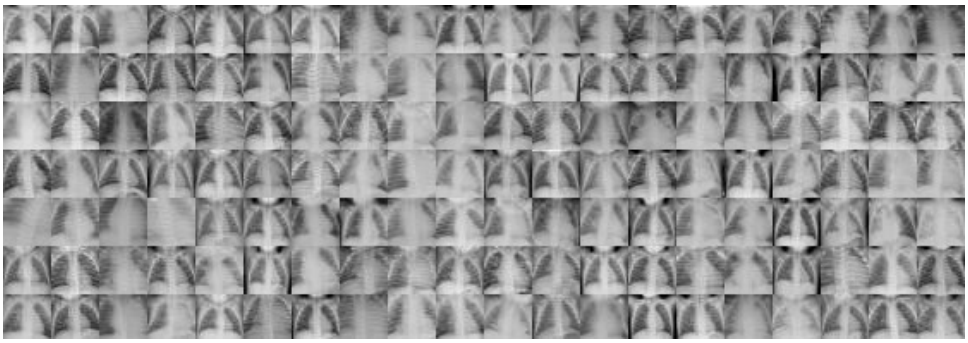

Figure 2: Examples of data instances from the PneumoniaMNIST dataset

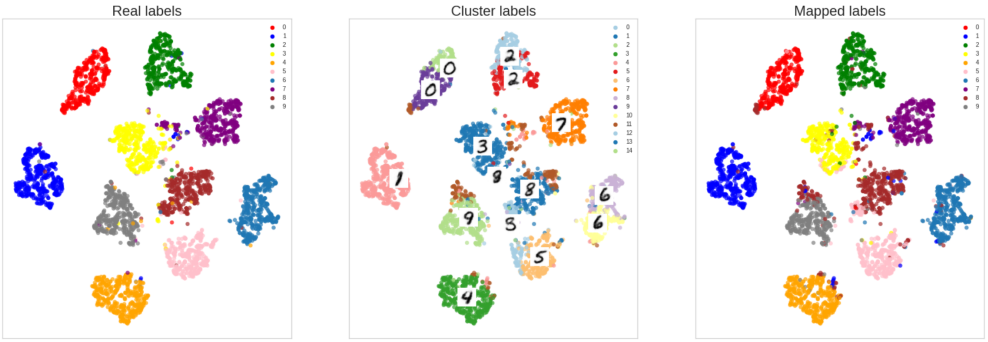

Figure 3: Results on MNIST test dataset.**A**: real labels, **B**: cluster labels, **C**: cluster labels mapped on real labels

of epochs: 100, Adam optimizer, learning rate=1e-4). Additionally, to explore whether embedding the separation in the training results in performance increases compared to simply clustering the latent space, we extracted features from the last layer of ResNet18 and performed a KMeans search for the optimal number of clusters, followed by a mapping between cluster labels and real labels. The last two experiments served as baselines for comparison for the architecture proposed in this paper. We choose ResNet18 as the benchmark architecture because it is used in the original [19] publications and the benchmark data are already available and comparable. All experiments were run using Pytorch, 50 warm-up epochs, and 100 total epochs. The LARS optimizer was used with a StepLR scheduler that linearly increases the learning rate from lr=0.01 to 0.25 (first 10 epochs), after which a cosine annealing scheduler reduces the learning rate to 0.05 in 90 epochs.

3 Evaluation

For visual evaluation, plot a 2D representation of the latent space clustered using the t-SNE algorithm [16]. Successively, after training the model in an self-supervised manner, the evaluation phase employs part of the available labels that are used for performance evaluation in three different approaches, all based on the downstream classification task: a) a statistical mapping approach between the cluster labels and the real labels, b) the kNN algorithm (k-Nearest Neighbors) [9], and c) the training of a simple linear layer with categorical cross-entropy as a loss function.

In the statistical approach (a), samples drawn from a subportion of the training set are associated with a cluster label which is calculated as the index of the closest prototype based on the distance between cosines. Then, the distribution of real labels across each cluster label is computed, and each cluster label is associated with the most frequent real label that occurs in the samples associated with the

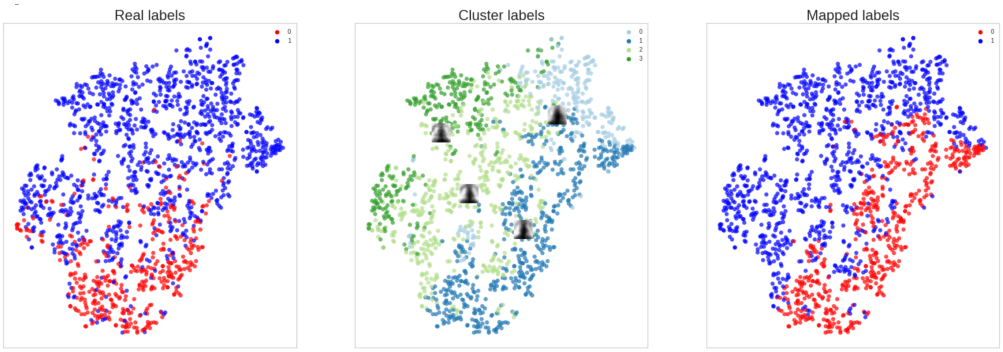

Figure 4: Results on test (PT) and validation (PV) sets of the PneumoniaMNIST dataset. **A**: real labels, **B**: cluster labels, **C**: cluster labels mapped on real labels

prototype currently in use. This results in associating each prototype to the mode of the real labels which are most similar to itself and can be seen as a way to measure how well each prototype captures the key elements that intrinsically define a class. In the inference phase, each sample of the test set is associated to one cluster and a predicted label is generated using this pre-computed map between clusters and real labels.

In the second approach (b), a subset of the training set is used to compute the "memory bank" for the kNN algorithm. In the inference phase, each sample of the test set is compared with every element stored in the memory bank, computing the pairwise distance for every pair. The predicted label is then defined as the mode of the labels of the k closest samples (in our case $k = 5$).

In the third approach (c), a linear layer maps from z to the number of possible classes. This linear layer is trained for 200 epochs using the Adam [13] optimizer with a learning rate of $3 * 10^{-4}$ over a subportion (20% of the training set, chosen to mimic a semi-supervised approach in a situation where the number of available labels is low). of the training set, with categorical cross-entropy as loss function.

4 Results

Results are summarized in Table I. Since the validation data sets were not used for training or model selection, we tested our framework on the validation and test data sets provided, which have different proportions of classes and produce different results. Our approach systematically outperforms feature extraction + KMeans in both the test set and the validation set. Importantly, it also performs close to the supervised baseline, even though it is optimized for reconstructing images in an unsupervised manner. When combined with kNN or a linear classifier, our approach generates features, which result in even higher performance compared to the above performance evaluation approach. Figure 3 shows a 2D representation of the latent space after encoding for both real and cluster labels. In the latter case, the *prototypes* are also superimposed over the image to demonstrate the reconstruction of mean representatives of each cluster.

5 Conclusions

This proof of concept study demonstrates the potential of using contrastive learning to encourage latent space separability in autoencoders and possibly other generative frameworks. In benchmark datasets, including medical imaging, our unsupervised stratification method delivers nearly equal results (in terms of performance) to supervised baselines and outperforms the alternative strategy of clustering features extracted from the penultimate layer of the supervised baseline. The autoencoder framework allows for simultaneous image reconstruction and sampling from a specific cluster if used in a variational setting. The method is also apt to refine/redefine existing classes or to completely re-stratify a disease continuum. Current limitations include the need of choosing the num-

Model	Dataset	Accuracy	Precision	Recall
Resnet18	Pneumonia Test	0.74	0.77	0.74
Feature+KM	Pneumonia Test	0.69	0.74	0.69
Our Model (stat)	Pneumonia Test	0.73	0.73	0.73
Our Model (kNN)	Pneumonia Test	0.84	0.85	0.83
Our Model (lin)	Pneumonia Test	0.84	0.80	0.88
Resnet18	Pneumonia Val	0.86	0.86	0.86
Feature+KM	Pneumonia Val	0.78	0.77	0.78
Our Model (stat)	Pneumonia Val	0.80	0.79	0.80
Our Model (kNN)	Pneumonia Val	0.84	0.86	0.81
Our Model (lin)	Pneumonia Val	0.84	0.80	0.87
Resnet18	MNIST	0.995	0.995	0.995
Feature+KM	MNIST	0.69	0.68	0.69
Our Model (stat)	MNIST	0.90	0.91	0.90
Our Model (kNN)	MNIST	0.98	0.98	0.97
Our Model (lin)	MNIST	0.98	0.97	0.98

Table 1: Results of baseline models and of approach. We evaluated our architecture using three different methods: Statistical (stat), where labels are associated to clusters based on the most frequent label present in a cluster, Nearest Neighbor (kNN), where the class is chosen by majority voting amongst the 5 nearest neighbors in the latent space, and a Linear classifier (lin), where a linear layer was trained on top of the latent features. **PT**: PneumoniaMNIST test set, **PV**: PneumoniaMNIST validation set, **MNIST** MNIST test set

ber of clusters with KMeans (which could be substituted with, e.g. a neural network with learnable weights and a clustering loss) and the need for large batch sizes, which could be foregone through e.g. sampling approaches.

References

[1] R. Aggarwal, V. Sounderajah, G. Martin, D. S. W. Ting, A. Karthikesalingam, D. King, H. Ashrafian, and A. Darzi. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *npj Digital Medicine*, 4(1):65, Apr 2021. ISSN 2398-6352. doi: 10.1038/s41746-021-00438-z. URL `https://doi.org/10.1038/s41746-021-00438-z`.

[2] B. Bengfort, R. Bilbro, N. Danielsen, L. Gray, K. McIntyre, P. Roman, Z. Poh, et al. Yellowbrick, 2018. URL `http://www.scikit-yb.org/en/latest/`.

[3] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers, 2021. URL `https://arxiv.org/abs/2104.14294`.

[4] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12546–12558. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/949686ecef4ee20a62d16b4a2d7ccca3-Abstract.html`.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations, 2020. URL `https://arxiv.org/abs/2002.05709`.

[6] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierar-

chical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[8] L. Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[9] E. Fix and J. L. Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review / Revue Internationale de Statistique*, 57(3):238–247, 1989. ISSN 03067734, 17515823. URL http://www.jstor.org/stable/1403797.

[10] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020. URL https://arxiv.org/abs/2006.07733.

[11] H. Hampel, F. Caraci, A. C. Cuello, G. Caruso, R. Nisticò, M. Corbo, F. Baldacci, N. Toschi, F. Garaci, P. A. Chiesa, S. R. Verdooner, L. Akman-Anderson, F. Hernández, J. Ávila, E. Emanuele, P. L. Valenzuela, A. Lucía, M. Watling, B. P. Imbimbo, A. Vergallo, and S. Lista. A path toward precision medicine for neuroinflammatory mechanisms in alzheimer's disease. *Front. Immunol.*, 11:456, Mar. 2020.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

[13] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2014. URL https://arxiv.org/abs/1412.6980.

[14] P. H. Le-Khac, G. Healy, and A. F. Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934, 2020. doi: 10.1109/access.2020.3031549. URL https://doi.org/10.11092Faccess.2020.3031549.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[16] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

[17] X. Wang, Y. Du, S. Yang, J. Zhang, M. Wang, J. Zhang, W. Yang, J. Huang, and X. Han. RetCCL: Clustering-guided contrastive learning for whole-slide image retrieval. *Medical Image Analysis*, 83:102645, Jan. 2023. ISSN 1361-8415. doi: 10.1016/j.media.2022.102645. URL https://www.sciencedirect.com/science/article/pii/S1361841522002730.

[18] Y. Wu, D. Zeng, Z. Wang, Y. Shi, and J. Hu. Distributed Contrastive Learning for Medical Image Segmentation, Aug. 2022. URL http://arxiv.org/abs/2208.03808. arXiv:2208.03808 [cs, eess].

[19] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni. Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification, 2021. URL https://arxiv.org/abs/2110.14795.

[20] Y. You, I. Gitman, and B. Ginsburg. Large batch training of convolutional networks, 2017. URL https://arxiv.org/abs/1708.03888.

[21] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz. Contrastive Learning of Medical Visual Representations from Paired Images and Text, Sept. 2022. URL http://arxiv.org/abs/2010.00747. arXiv:2010.00747 [cs].