MedTransformer—Unlocking the Power of Pre-trained Transformers in Medical Image Analysis

Eran Levin * **Gilad Deustch** *

Code: https://github.com/GiilDe/med-hiera

1. Introduction

In recent years, the field of artificial intelligence (AI) has witnessed remarkable progress, particularly in domains that were traditionally challenging to automate. Tasks like classical image recognition, exemplified by ImageNet (Russakovsky et al., 2015), have seen a substantial reduction in error rates, plummeting from a staggering 50% to less than 10%. This transformation has largely been driven by the advent of transformer models, initially designed for natural language processing (NLP) but now dominating the computer vision landscape as well.

However, when it comes to applying these AI breakthroughs to medical tasks, the terrain becomes less accommodating. While datasets for general computer vision tasks have exploded in size, with datasets like ImageNet (Russakovsky et al., 2015) boasting 14 million images and LAION (Schuhmann et al., 2022) stretching to 5.5 billion, medical datasets remain a precious and expensive resource (Alberto et al., 2023). Typically, medical datasets are orders of magnitude smaller than their general task counterparts, making the application of deep learning techniques more challenging in this domain.

This discrepancy in dataset sizes is a critical issue since it has been demonstrated that the performance of deep learning models, including transformers, is highly sensitive to the quantity of training data (Zhai et al., 2022; Chen et al., 2022). These models exhibit a capacity to acquire general domain knowledge during training, a phenomenon often referred to as "transfer learning" (Weiss et al., 2016). This acquired knowledge enables them to generalize effectively to various tasks when subsequently fine-tuned on specific data. Consequently, it has become a common practice to take pre-trained models trained on large general-class datasets, such as ImageNet, and fine-tune them for medical tasks (Parvaiz et al., 2023).

However, our research suggests that this conventional approach may not be optimal for medical image analysis. We propose a more sophisticated scheme that leverages the shared characteristics between medical and general images, even when specific disease-related data is scarce. Our aim is to efficiently adapt pre-trained transformers to medical data, harnessing the models' inherent knowledge of regular images and applying it to medical images. Through this approach, we demonstrate significant improvements in model performance while conserving computational resources.

In this study, we showcase the effectiveness of our scheme on a cutting-edge hierarchical transformer model called "Hiera" (Ryali et al., 2023). Hiera exhibits a hierarchical architecture that offers a range of desirable properties for medical image analysis. By exploring how pre-trained transformers can be tailored to medical data with efficiency and precision, our research contributes to the ongoing effort to bridge the gap between AI advancements and medical applications. We have achieved near state-of-the-art (SOTA) performance, as evidenced by an impressive AUC score of 81.687% (see: appendix .1). It is noteworthy that our accomplishment was attained with significantly fewer computational resources compared to our competitors (by computation time, and number of parameters).

2. Related Work

The subject of X-ray auto-classification has received extensive research. ResNet-based (He et al., 2015) works have been proposed as early as 2017 (Wang et al., 2017). Later, DenseNet (Huang et al., 2018)-based works have been proposed (Yao et al., 2018; Li et al., 2018; Tang et al., 2018; Guendel et al., 2018; Guan and Huang, 2020). All of these works have been published before the major adaption of Vision Transformers (Dosovitskiy et al., 2021) which is believed to be a more robust model for image classification. Sadly, there has been limited success in applying this concept to the field of medicine: (Taslimi et al., 2022) have used Swin-Transformer (Liu et al., 2021) for this task. (Xiao et al., 2022) have used regular transformer with special training regime, claiming to utilize some property of the data to better train the model. They used the same dataset to pre-train (MAE) the model and fine-tune it, in addition they have not utilized any existing weights as we did, and train the model from scratch, requiring much more computation. They also used larger transformer models (more parameters). This work currently holds the SOTA on this task, with our less than a percent apart with the smallest model of theirs. Our work surpasses most of those previous works in performance.

3. Methodology

3.1. Preliminaries

For this project, we investigate the following assumptions:

3.1.1. Medical data similarity

In the context of medical data, where datasets are often limited in size for specific diseases or conditions, it is noteworthy that common properties exist among diverse medical datasets. Even datasets associated with unrelated diseases can exhibit shared characteristics. For instance, datasets pertaining to various gastric organs tend to exhibit common organ shapes within the abdominal region. This observation suggests the potential to train models that can recognize and extract generic features relevant to multiple medical detection tasks. By leveraging these common properties, we can mitigate the challenge of data scarcity for individual problems, allowing us to amalgamate information from a variety of datasets to construct a more resilient and comprehensive dataset. This approach offers promising prospects for enhancing medical data analysis and diagnosis.

3.1.2. Medical data uniqueness

Medical data possesses unique characteristics distinct from general datasets like ImageNet. These distinctions encompass factors such as a limited color range, with medical data typically presented in grayscale. Moreover, medical data exhibits a higher degree of repetitive patterns, often adhering to a fixed high-level structure within the human body, in contrast to other tasks where objects can display more diverse structural variations. This intrinsic disparity hinders the transferability of models trained on general datasets to the domain of medical datasets. Our empirical investigations substantiate this notion, revealing that transformers trained for ImageNet reconstruction tasks encounter substantial challenges when tasked with reconstructing medical images. These findings underscore the argument that medical images lie outside the distribution of ImageNet data and consequently pose significant challenges when fine-tuning for the classification of medical images.

Our mission is to strike a balance between retaining valuable features from ImageNet and harnessing the rich diversity inherent in medical-oriented datasets.

3.1.3. Architecture

Transformers exhibit remarkable suitability for our objectives, given their capacity to leverage unlabeled data for the acquisition of comprehensive domain knowledge. Additionally, they demonstrate heightened robustness in preserving symmetries, a critical attribute in tasks where holistic context, as seen in the medical field, holds paramount importance. Our choice of employing hierarchical transformers (Liu et al., 2021) aligns with recent findings highlighting their superior efficiency compared to conventional vision transformers, thus enhancing the efficacy of our approach.

3.2. Suggested Solution – New Training Scheme

For exploiting the knowledge that the model has acquired while leveraging existing data, we suggest a new scheme:

3.2.1. cocktail dataset

We propose a novel approach within the domain of medical data analysis, one that leverages diverse datasets encompassing various medical tasks. Our method involves the utilization of pre-existing weights, previously trained on the ImageNet dataset MAE task (He et al., 2021), and subjecting them to training on another Masked Auto Encoder (MAE) task.

In this MAE task, the primary objective is not to predict the specific tabular values within each dataset. Instead, the primary focus is on extracting and learning the underlying structural features inherent to medical images. This approach enables the amalgamation of disparate datasets, each containing distinct medical fields and characteristics, into a cohesive framework.

It is noteworthy that the datasets incorporated into this "cocktail" need not necessarily possess tabular values. This unique feature allows us to delve into the realm of semi-supervised learning paradigms, opening up promising avenues for exploration.

Subsequently, the trained weights, enriched with the knowledge of medical image structures, can be subjected to a fine-tuning process tailored to the specific target task. This final refinement step ensures that our model not only comprehends the shared structural elements across diverse medical datasets but also becomes adept at addressing task-specific nuances. The general scheme is therefore:

1. MAE of ImageNet or other large dataset (for all the transformers those weights are publicly available).

2. MAE of the cocktail dataset. This is the new stage which does not exist in the general scheme

3. Fine-tune of the model with the relevant dataset with its tabular values to the task.

3.2.2. Task Selection

We used ChestX-ray14 (Wang et al., 2017) as an example task for this research. ChestX-ray14 comprises 112,120 frontal-view X-ray images of 30,805 unique patients with fourteen common disease labels. This task has been explored before by many papers, including transformers-based, and contain X-rays data which is abundant in the medical world due to its price.

3.2.3. Cocktail Dataset Selection

For the cocktail dataset, we used the following datasets:

1. BraTS 2020 (Menze et al., 2014; Bakas et al., 2017; et al., 2019) is a collection of pre-operative MRI scans of brain tumors known as gliomas, collected from multiple institutions. These tumors exhibit inherent diversity in their appearance, shape, and histological characteristics.

2. COVID-19 Radiography Database (Rahman et al., 2021)—a database of chest X-ray images for COVID-19 positive cases, along with Normal and Viral Pneumonia images.

3. COVIDx CXR-3 (Wang et al., 2020b)—a database of chest X-ray images for COVID-19 cases along with healthy images.

4. ChestX-ray14.

4. Experiments

4.1. Overview of experimental settings

Our experimental methodology consists of a two-stage process, namely, 1) Model Pre-training and 2) Classification Fine-tuning. Additionally, it is worth noting that there exists an implicit initial stage, denoted as Stage 0, wherein we leverage pre-trained weights from the ImageNet dataset. Within both the pre-training and classification phases, we employ a data partitioning approach to create distinct training and validation subsets. These validation subsets play a crucial role in our hyperparameter tuning process. It is imperative to highlight that our classification dataset is further partitioned to include a dedicated test set for independent evaluation. Given the substantial scale of our datasets and the computationally intensive nature of the training process, we opt not to employ techniques such as K-fold cross-validation or related methodologies in our hyperparameter tuning endeavors. This decision is rooted in the resource constraints associated with our computational infrastructure.

4.1.1. Preprocessing

In our experimentation, we first calculate the mean (μ) and standard deviation (σ) of the training dataset. During training, we normalize each pixel (p) in the images using the transformation $p' = \frac{p-\mu}{\sigma}$, aligning the dataset with the standard normal distribution $\mathcal{N}(0, 1)$.

A pertinent question arises when dealing with separate pre-training and classification datasets: should we maintain consistency by using the same μ and σ, or recalibrate for the classification stage to ensure strict adherence to the standard normal distribution? In the interest of methodological consistency, we opt for the former approach. However, the exploration of the latter, where distinct preprocessing for the classification stage is implemented, remains a potential avenue for future research.

4.2. Pre-training

This phase, as previously elucidated, revolves around the training of a Masked Autoencoder (MAE) on medical data. Our empirical inquiry commences with three distinct experiments, thoughtfully structured as follows:

Initialization with pre-trained weights from ImageNet followed by pretraining on our comprehensive dataset amalgamation. Commencement with ImageNet pre-trained weights, followed by exclusive pretraining on the ChestX-ray14 dataset. The adoption of a random weight initialization strategy, followed by pretraining on the entire dataset amalgamation. The results of these experiments are graphically depicted in Figure 4. Our findings underscore the paramount significance of ImageNet pre-training (Step 0 in our pipeline), which consistently outperforms random weight initialization. This observation is further substantiated by the occurrence of the "exploding" loss phenomenon when employing excessively high learning rates, as evident in Figure 5.

It is noteworthy that this experiment alone does not definitively establish the efficacy of our dataset amalgamation, often referred to as the "cocktail" as it manifests a slightly higher loss. A comprehensive evaluation of its effectiveness is deferred until the subsequent classification stage, as delineated in Figure 1.

Table 4 presents the configuration parameters of our optimal MAE model following an exhaustive hyperparameter search. Notably, the most effective parameters appear relatively conventional, potentially attributed to our deliberate choice of a compact Transformer architecture.

4.3. Classification

This is the final step in our training pipeline. We experiment with 4 different experiments: 1. Starting from random weights 2. Starting from ImageNet pre-trained weights 3. Starting from MAE pretraining on the ChestX-ray14 dataset. 4. Starting from MAE pretraining on the full cocktail. We show the results in Figure 1. We can see that the pre-training significantly increases the AUC average for the ChestX-ray14 pretrained and the full cocktail pretrained models. We can also see a clear, albeit humble, increase of the AUC average of the model pretrained on the full cocktail over the model pretrained on the ChestX-ray14 only. Note that the MAE pre-training lasted for 20 epochs in this case, more epochs could have shown a bigger increase in performance.

Table 1 shows the parameters of our best run after conducting hyperparameter sweeping. Figure 2 shows our numerous training runs as part of our hyperparameter sweeping. Figure 3 shows the hyperparameter importance analysis we compute.

During our classification training, we evaluate the AUC mean for each epoch on the validation set. We train for 10 epochs and due to over-fitting our best model is after 7 epochs. It reaches an AUC mean of 80.01%. We evaluate this model on the test set and achieve an AUC mean of 81.6%. Table 2 shows the AUC for each label.

5. Discussion

5.1. Datasets similarity

In this work, we have shown that pre-training a model using Masked Auto Encoding (He et al., 2021) on a collection, e.g. “cocktail” of datasets is effective in improving a classification model, even when the datasets diverge from the classification dataset in some properties, e.g. include different diseases. Yet, admittedly, most of our cocktail consists of the same medical image type—chest X-ray. There are only 14,911 images in our cocktail that are not chest x-ray which makes it challenging to study how much of the increase in performance can be attributed to them. Whether pre-training with datasets that are dissimilar not only in labels, but in overall structure helps to the classification, and quantifying such dissimilarity may be a fascinating next step. Another interesting question is whether the chest x-ray datasets that we add to the cocktail aids our training because of the healthy patients’ images, or are images of diseases that we do not try to classify also helped?

5.2. Pre-training vs. classification performance

In this work, we have seen that there is a correlation between the performance in the MAE pre-training and the performance in the classification fine-tuning. Yet, this relationship is not absolute—for example, when we pre-train a model on the ChestX-ray14 dataset only it has better slightly better loss in the pre-training stage, but is less performant in the classification stage with respect to both AUC and evaluation loss. This adds another element of complexity to the hyperparameter tuning, even if we sweep and find the best model in the pre-training, is it also the best for classification?

6. Conclusion

In this work, we leverage recent progress in training of hierarchical transformers for the classification of chest X-ray diseases. We leverage a cocktail of medical datasets for Masked Auto Encoding (He et al., 2021) pre-training. Our cocktail is composed of 4 datasets with varying degrees of similarity to the dataset of interest (i.e. the one we perform classification on). We show that the combination of the Hiera architecture (Ryali et al., 2023) and MAE pre-training on the cocktail we built allows us with relatively few resources to be close to the state of the art on the ChestX-ray14 dataset. We achieve 81.6% mean AUC where the state of the art is 83% (Xiao et al., 2022). We also conduct ablation experiments and show each step of our training pipeline increases our model’s performance — ImageNet pre-training, cocktail pre-training, and of course the classification training.

7. Limitation

1. The complete source code of the Hiera model remains unreleased to date. It is understood that certain techniques employed by the authors have not been fully disclosed or implemented. It is widely acknowledged that these undisclosed strategies have a cumulative effect, potentially bridging the existing performance gap between our approach and the current SOTA, which frequently leverages similar techniques, such as augmentations.

2. We opted for the employment of the smallest available transformer model in our experimentation due to resource constraints. As illustrated in the comparative table, this choice may account for a portion, if not the entirety, of the observed performance gap between our approach and the current SOTA.

.1. Appendix

References

Isabelle Alberto, Nicole Rose Alberto, Arnab Ghosh, Bhav Jain, Shruti Jayakumar, Nicole Martínez, Ned McCague, Dana Moukheiber, Lama Moukheiber, Mira Moukheiber, Sulaiman Moukheiber, Antonio Yaghy, Andrew Zhang, and Leo Celi. 2023. The impact of commercial health datasets on medical research and health-care algorithms. *The Lancet. Digital health* 5 (05 2023), e288–e294. https://doi.org/10.1016/S2589-7500(23)00025-0

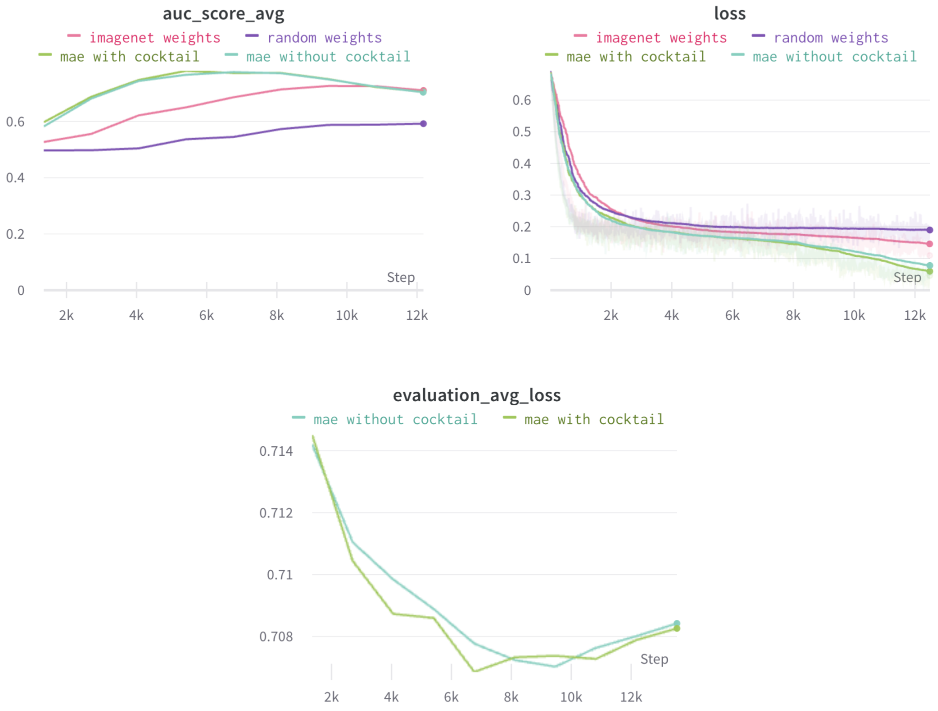

Figure 1. Classification Training & Evaluation Loss. We run classification training using 4 different models: 1. model initialized with random weights 2. model with MAE pre-training on ImageNet 3. model with MAE pre-training on ChestX-ray14 dataset 4. model with MAE pre-training on our full cocktail.

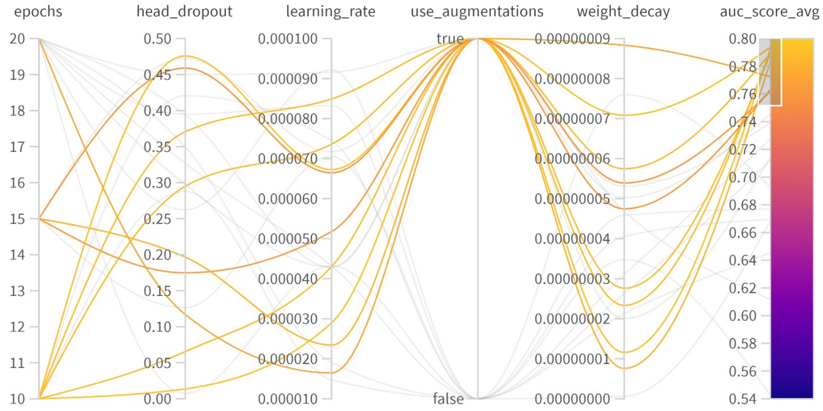

Figure 2. configs and AUC mean for our different sweeping runs. Runs with poor scoring have been outlined in grey.

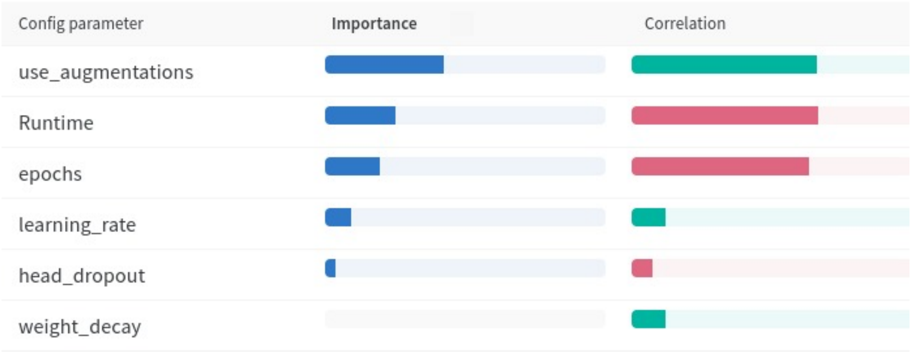

Figure 3. Hyperparameter sweep analysis. We compute the correlation and importance of hyperparameters. Importance is based on training a random forest on hyperparameters values as input and the mean – AUC as the target. Then, for each hyperparameter, its importance is defined as its feature importance in the training.

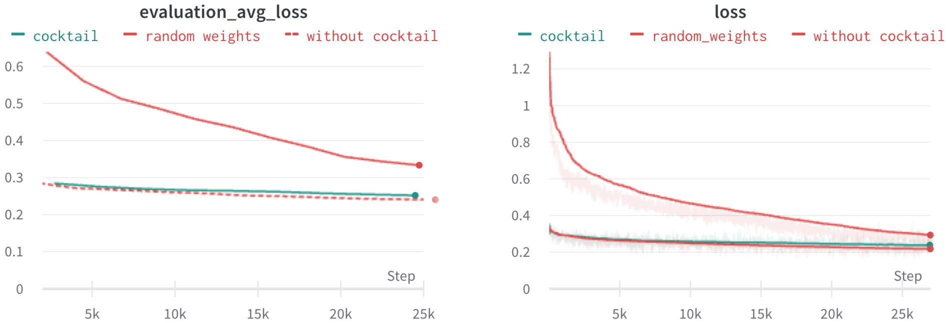

Figure 4. Mae Training & Evaluation Loss. We tested starting MAE training from random weights or from ImageNet weights. For the ImageNet weights case we test using the full cocktail or only the ChestX-ray14 dataset. We can see that the full cocktail has slightly higher loss, despite this fact when we apply classification fine-tuning we see that the full cocktail pre-trained model has a higher AUC. Additionally, it can be observed that it is much more efficient to start from the ImageNet pre-trained model than random weights

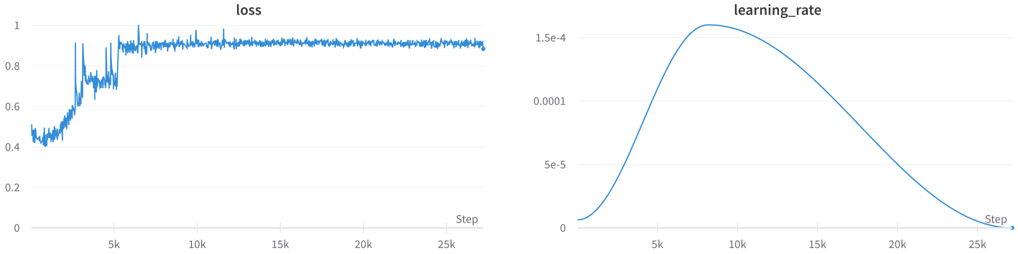

Figure 5. The phenomenon of an "exploding loss" during MAE training is illustrated in the figures above. On the left, we observe the loss curve itself, while the right figure displays the corresponding learning rate. It is important to note that an excessively high learning rate can lead to a situation where the model effectively "forgets" the valuable ImageNet pre-training it underwent, causing it to lose a significant portion of its acquired general-world knowledge. This "forgetting" process can be detrimental, as the model struggles to recover, and the loss remains persistently higher than the initial loss. This observation serves as compelling evidence that the model indeed leverages the knowledge it gained in the initial stage of training.

hyperparameter	Value
pretraining	full-cocktail MAE, Figure 4
maximal learning rate	0.0001
batch size	64
epochs	10
learning rate schedule	OneCycleLR (Smith and Topin, 2018)
optimizer	Adam
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$
weight decay	1e-8
rotation augmentation angle	30

Table 1. Our classification hyperparameters.

Diagnosis	AUC
Edema	87.24
Atelectasis	77.55
Cardiomegaly	85.03
Consolidation	77.28
Effusion	85.83
Emphysema	90.39
Fibrosis	81.83
Hernia	89.15
Infiltration	70.28
Mass	81.17
No Finding	79.14
Nodule	76.99
Pleural Thickening	78.24
Pneumonia	79.86
Pneumothorax	85.24

Table 2. AUC per Diagnosis with our suggested model

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. 2017. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific data* 4, 1 (2017), 1–13.

Ivo Baltruschat, Hannes Nickisch, Michael Grass, Tobias Knopp, and Axel Saalbach. 2019. Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification. *Scientific Reports* (04 2019). https://doi.org/10.1038/s41598-019-42294-8

Wuyang Chen, Xianzhi Du, Fan Yang, Lucas Beyer, Xiaohua Zhai, Tsung-Yi Lin, Huizhong Chen, Jing Li, Xiaodan Song, Zhangyang Wang, and Denny Zhou. 2022. A Simple Single-Scale Vision Transformer for Object Localization and Instance Segmentation. arXiv:2112.09747 [cs.CV]

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs.CV]

Spyridon Bakas et al. 2019. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. arXiv:1811.02629 [cs.CV]

Qingji Guan and Yaping Huang. 2020. Multi-label chest X-ray image classification via category-wise residual attention learning. *Pattern Recognition Letters* 130 (2020), 259–266. https://doi.org/10.1016/j.patrec.2018.10.027 Image/Video Understanding and Analysis (IUVA).

Sebastian Guendel, Sasa Grbic, Bogdan Georgescu, Kevin Zhou, Ludwig Ritschl, Andreas Meier, and Dorin Comaniciu. 2018. Learning to recognize Abnormalities in Chest X-Rays with Location-Aware Dense Networks. arXiv:1803.04565 [cs.CV]

Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Michael B. Gotway, and Jianming Liang. 2022.

Method	Type	Architecture	Pre-training	mAUC
(Wang et al., 2017)	ConvNet	ResNet-50	ImageNet	74.5
(Yao et al., 2018)	ConvNet	ResNet-&DenseNet	-	76.1
(Li et al., 2018)	ConvNet	ResNet-50	-	75.5
(Tang et al., 2018)	ConvNet	ResNet-50	-	80.3
(Guendel et al., 2018)	ConvNet	DenseNet-121	-	80.7
(Guan and Huang, 2020)	ConvNet	DenseNet-121	-	81.6
(Wang et al., 2020a)	ConvNet	ResNet-152	-	78.8
(Ma et al., 2019)	ConvNet	ResNet-101	-	79.4
(Baltruschat et al., 2019)	ConvNet	ResNet-50	-	80.6
(Seyyed-Kalantari et al., 2020)	ConvNet	DenseNet-121	-	81.2
(Ma et al., 2020)	ConvNet	DenseNet-121	-	81.7
(Hermoza et al., 2020)	ConvNet	DenseNet-121	-	82.1
(Kim et al., 2021)	ConvNet	DenseNet-121	-	82.2
(Haghighi et al., 2022)	ConvNet	DenseNet-121	-	81.7
(Liu et al., 2022)	ConvNet	DenseNet-121	-	81.8
(Taslimi et al., 2022)	Tranformer	SwinT	-	81.0
(Xiao et al., 2022) (MoCo v2)	ConvNet	DenseNet-121	X-rays dataset	80.6
(Xiao et al., 2022) (MAE)	Transformer	ViT-S/16	-	82.3
(Xiao et al., 2022) (MAE)	Transformer	ViT-B/16	X-rays dataset	83.0
Ours	Transformer	ViT-T/16	ImageNet	81.687

Table 3. Method Comparison with Type, Architecture, Pre-training, and mAUC

Config	Value
dataset	full-cocktail
pre-training	ImageNet
maximum learning-rate	0.0001
batch size	64
epochs	80
mask ratio	0.6
learning rate schedule	OneCycleLR (Smith and Topin, 2018)
Optimizer	Adam
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$

Table 4. Our MAE pre-training config. No augmentations, weight decay or other methods are used in our best performing run.

DiRA: Discriminative, Restorative, and Adversarial Learning for Self-supervised Medical Image Analysis. arXiv:2204.10437 [cs.CV]

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2021. Masked Autoencoders Are Scalable Vision Learners. arXiv:2111.06377 [cs.CV]

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV]

Renato Hermoza, Gabriel Maicas, Jacinto C. Nascimento, and Gustavo Carneiro. 2020. Region Proposals for Saliency Map Refinement for Weakly-supervised Disease Localisation and Classification. arXiv:2005.10550 [cs.CV]

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2018. Densely Connected Convolutional Networks. arXiv:1608.06993 [cs.CV]

Eunji Kim, Siwon Kim, Minji Seo, and Sungroh Yoon. 2021. XProtoNet: Diagnosis in Chest Radiography with Global and Local Explanations. arXiv:2103.10663 [cs.CV]

Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei. 2018. Thoracic Disease Identification and Localization with Limited Supervision. arXiv:1711.06373 [cs.CV]

Fengbei Liu, Yu Tian, Yuanhong Chen, Yuyuan Liu, Vasileios Belagiannis, and Gustavo Carneiro. 2022. ACPL: Anti-curriculum Pseudo-labelling for Semi-supervised Medical Image Classification. arXiv:2111.12918 [cs.CV]

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. arXiv:2103.14030 [cs.CV]

Congbo Ma, Hu Wang, and Steven C. H. Hoi. 2020. Multilabel Thoracic Disease Image Classification with Cross-Attention Networks. arXiv:2007.10859 [cs.CV]

Yanbo Ma, Qiuhao Zhou, Xuesong Chen, Haihua Lu, and Yong Zhao. 2019. Multi-attention Network for Thoracic Disease Classification and Localization. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1378–1382. https://doi.org/10.1109/ICASSP.2019.8682952

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. 2014. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE transactions on medical imaging* 34, 10 (2014), 1993–2024.

Arshi Parvaiz, Muhammad Anwaar Khalid, Rukhsana Zafar, Huma Ameer, Muhammad Ali, and Muhammad Moazam Fraz. 2023. Vision Transformers in medical computer vision—A contemplative retrospection. *Engineering Applications of Artificial Intelligence* 122 (2023), 106126.

Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M. Zughaier, Muhammad Salman Khan, and Muhammad E.H. Chowdhury. 2021. Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. *Computers in Biology and Medicine* 132 (2021), 104319. https://doi.org/10.1016/j.compbiomed.2021.104319

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. 2023. Hiera: A Hierarchical Vision Transformer without the Bells-and-Whistles. arXiv:2306.00989 [cs.CV]

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. arXiv:2210.08402 [cs.CV]

Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y. Chen, and Marzyeh Ghassemi. 2020. CheXclusion: Fairness gaps in deep chest X-ray classifiers. arXiv:2003.00827 [cs.CV]

Leslie N. Smith and Nicholay Topin. 2018. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates. arXiv:1708.07120 [cs.LG]

Yuxing Tang, Xiaosong Wang, Adam P. Harrison, Le Lu, Jing Xiao, and Ronald M. Summers. 2018. Attention-Guided Curriculum Learning for Weakly Supervised Classification and Localization of Thoracic Diseases on Chest Radiographs. arXiv:1807.07532 [cs.CV]

Sina Taslimi, Soroush Taslimi, Nima Fathi, Mohammadreza Salehi, and Mohammad Hossein Rohban. 2022. SwinCheX: Multi-label classification on chest X-ray images with transformers. arXiv:2206.04246 [cs.CV]

Hongyu Wang, Haozhe Jia, Le Lu, and Yong Xia. 2020a. Thorax-Net: An Attention Regularized Deep Neural Network for Classification of Thoracic Diseases on Chest Radiography. *IEEE Journal of Biomedical and Health Informatics* 24 (2020), 475–485. https://api.semanticscholar.org/CorpusID:198171986

Linda Wang, Zhong Qiu Lin, and Alexander Wong. 2020b. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Scientific Reports* 10, 1 (11 Nov 2020), 19549. https://doi.org/10.1038/s41598-020-76550-z

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. https://doi.org/10.1109/cvpr.2017.369

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big data* 3, 1 (2016), 1–40.

Junfei Xiao, Yutong Bai, Alan Yuille, and Zongwei Zhou. 2022. Delving into Masked Autoencoders for Multi-Label Thorax Disease Classification. arXiv:2210.12843 [cs.CV]

Li Yao, Jordan Prosky, Eric Poblenz, Ben Covington, and Kevin Lyman. 2018. Weakly Supervised Medical Diagnosis and Localization from Multiple Resolutions. arXiv:1803.07703 [cs.CV]

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. 2022. Scaling Vision Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12104–12113.