Delving into Masked Autoencoders for Multi-Label Thorax Disease Classification

Junfei Xiao  Yutong Bai  Alan Yuille  Zongwei Zhou*

Johns Hopkins University

Code: https://github.com/lambert-x/Medical_MAE

Abstract

Vision Transformer (ViT) has become one of the most popular neural architectures due to its great scalability, computational efficiency, and compelling performance in many vision tasks. However, ViT has shown inferior performance to Convolutional Neural Network (CNN) on medical tasks due to its data-hungry nature and the lack of annotated medical data. In this paper, we pre-train ViTs on 266,340 chest X-rays using Masked Autoencoders (MAE) which reconstruct missing pixels from a small part of each image. For comparison, CNNs are also pre-trained on the same 266,340 X-rays using advanced self-supervised methods (e.g. MoCo v2). The results show that our pre-trained ViT performs comparably (sometimes better) to the state-of-the-art CNN (DenseNet-121) for multi-label thorax disease classification. This performance is attributed to the strong recipes extracted from our empirical studies for pre-training and fine-tuning ViT. The pre-training recipe signifies that medical reconstruction requires a much smaller proportion of an image (10% vs. 25%) and a more moderate random resized crop range (0.5∼1.0 vs. 0.2∼1.0) compared with natural imaging. Furthermore, we remark that in-domain transfer learning is preferred whenever possible. The fine-tuning recipe discloses that layer-wise LR decay, RandAug magnitude, and DropPath rate are significant factors to consider. We hope that this study can direct future research on the application of Transformers to a larger variety of medical imaging tasks.

1. Introduction

There has been great progress in the Vision Transformer (ViT) architecture [26] and its variants [62, 43, 37, 91], showing that Transformers surpass and supersed Convolutional Neural Networks (CNN) in various natural imaging tasks. In comparison with CNNs, Transformers can better leverage the rapidly increasing image data, long-range spatial context of an image [21, 25], and share properties of the human visual system [71, 77, 30, 92]. Training Transformers requires considerably more data than CNNs [83, 89], but medical data are small and labels are hard to obtain. As a result, directly applying Transformers to the medical domain is found to be problematic and challenging. There are several early attempts [70, 99, 67], but their performance is often inferior to state-of-the-art CNNs (elaborated in §2). Recent surveys suggest that a range of successful cases are using a hybrid architecture of Transformers and CNNs [58, 81]. In contrast, the stand-alone and vanilla ViT architecture remains the concentration of this study to strive for simplicity. We ask: What is the full potential of ViT architecture in medical imaging tasks? The answer, based on our study, is that vanilla ViT can achieve a similar or even better performance than state-of-the-art CNNs if equipped with (I) a large-scale pre-training on unlabeled medical data and (II) strong pre-training and fine-tuning recipes, customized by unique characteristics of medical images.

The pre-training of CNNs has been widely investigated in the medical domain [49], resulting in several publicly available Foundation models [113, 17, 99]. Numerous pre-training methods can enable CNNs to learn representation from unlabeled images, including contrastive learning [84], predictive learning [115], restorative learning [15], and their combination [38, 39]. At the time this paper is written, however, neither contrastive nor predictive pre-training is mature for vanilla ViT architectures yet. The most popular pre-training scheme for ViTs is called Masked Autoencoders (MAE) [45]. Its task is to mask random patches of the input image and reconstruct the missing pixels. We adopt MAE in this paper because of its great scalability, computational efficiency, and compelling performance in many vision tasks.

This paper customizes the recipe of pre-training and fine-tuning MAE for the medical domain and verifies its effectiveness on three chest X-ray datasets. We have also made the pre-training and fine-tuning code publicly available and released ViT-Small and ViT-Base that are pre-trained on 510K X-ray images as well as the pre-trained CNNs. The pre-trained ViT encoder can be fine-tuned to improve classification tasks (validated in §5) and detection tasks (see Github). In summary, four contributions are made.

*Corresponding author: Zongwei Zhou (zzhou82@jh.edu)
1. The usefulness of ViT pre-trained on ImageNet (14M data & labels) and chest X-rays (0.3M data) is evaluated, underlining the opportunity of in-domain transfer learning and self-supervised learning (Table 1).

2. A strong pre-training recipe, consisting of more unlabeled data (266,340), a higher masking ratio (90%), and a modest random cropping scale (0.5~1.0), is developed for MAE to learn image representation from chest X-rays efficiently (§4.2–4.4).

3. Three of the most important hyper-parameters are determined to fine-tune ViT in multi-label thorax disease classification: layer-wise LR decay, RandAug magnitude, and DropPath rate (Table 3).

4. This is among the first efforts to approach vanilla ViT’s performance to the state-of-the-art CNNs on three predominant chest X-ray benchmarks, yielding mAUC of 82.3%, 89.2%, 99.3% on NIH ChestX-ray14, Stanford CheXpert, and COVIDx, respectively (§5.1–5.3).

By intention or non-intention, the empirical comparisons between old and new techniques (e.g., CNN vs. ViT) are often biased to the newer one [63, 4]. In this paper, we try our best not to over-sell or under-analyze the ViTs’ potential in the medical domain. To provide a fair and comprehensive benchmark, the performance of CNNs is truly state-of-the-art in each dataset based on our extensive literature review.

2. Related Works

Preliminary. Radiography images possess unique characteristics compared with photographic images, resulting in considerable difficulties when switching computer vision advancements to medical imaging [111, 112, 58, 81]. Photographic images, particularly those on ImageNet [23], contain large, apparent objects in the center of the images, residing in varying backgrounds. Learning discriminative features (e.g., color, texture, and shape) primarily from the foreground objects is important in computer vision. In contrast, radiography images are generated from pre-defined imaging protocols, so the background exhibits anatomical consistency across images (see examples of the chest anatomy in Figure 4). Clinically relevant information is dispersed throughout the image, whereas the diseased region (as the foreground) often encloses much more local, subtle, and fine-grained variations than photographic images. As a result, the model must be able to extract both global and local features to identify various diseases from the normal anatomy. In the following sections, we describe the difference between computer vision and medical imaging in the choice of model architectures and self-supervised methods, followed by a review of current state-of-the-art solutions for multi-label thorax disease classification.

ViTs or CNNs for medical imaging? Transformers have gained prevalence in numerous AI applications (e.g. AlphaFold2 [53], Google Translate [8]). In computer vision, there is a heated debate between the adoption of ViTs and CNNs, in terms of performance [63, 110, 6, 97, 91, 25], robustness [4, 69, 105, 109], data requirement [26, 83, 89], computational efficiency [74]. This discussion has finally been converging to an agreement that ViTs could serve as alternatives to CNNs in a variety of tasks [56, 42]. ViT has substantial potential for radiography imaging tasks, but currently, the superior performance of ViT has not been translated to radiography imaging, where CNN is still the dominant architecture. (1) ViTs’ performance has lagged behind that of SOTA CNNs [72, 88], in which we believe the poorly configured training recipe is one of the major causes; (2) most existing studies report ViTs’ performance on medical tasks without comparing with CNN under a similar experimental setting [70, 58]; and (3) multiple works focus on designing hybrid architectures by integrating benefits of ViTs and CNNs to claim the superior performance to CNNs [14, 107, 98, 86]. Conducting a fair comparison between ViTs and CNNs should take into account the number of parameters, volumes of computations, usages of GPUs, and suitable pre-training schemes. So far, there is no broad benchmark to fairly compare ViTs and CNNs in medical tasks, leaving us wondering whether we could trivially switch to ViTs in medical tasks. Unlike the aforementioned studies, our objective is to faithfully benchmark between ViTs and SOTA CNNs in radiography imaging tasks; to improve the recipe of existing ViTs with respect to data, model, optimization aspects; and to visualize how ViTs and CNNs interpret radiography images (§6).

Self-supervised methods in medical imaging. Self-supervised learning has shown enormous potential in medical imaging due to the sparsity of high-quality annotation [111]. Two major trends are based on contrastive and restorative pre-training. In computer vision, contrastive pre-training [19, 18, 34, 11] holds state-of-the-art performance, surpassing supervised ImageNet pre-training in some tasks; while in medical imaging, restorative pre-training [113, 86, 29] presently reaches a new height in performance. We attribute this popularity asymmetry to the marked difference between photographic and radiography images. Since radiography imaging protocols assess patients in a fairly consistent orientation, the generated images have great similarity across various patients [96, 39]. The inherent consistency eases the analysis of numerous critical problems but also causes a significant problem for contrastive pre-training. Contrastive pre-training (e.g., MoCo [46, 19]) treats each image as a distinct class and minimizes the simi-
larity of representations derived from different images. This concept, in theory, might not work properly for radiography imaging because the negative pairs appear too similar (empirically evidenced in our Table 1). In contrast, restorative pre-training is good at conserving fine-grained textures embedded in image context, so it has been widely adopted in medical pre-training. Restorative pre-training is formulated as the task of pixel-wise image reconstruction [2, 15, 114, 115, 16, 100]. Following this spirit, we take masked autoencoders (MAE) [45] as pre-training task for its simplicity, efficiency, scalability, and compelling performance. We are among the first to configure a strong recipe for both ViT pre-training and fine-tuning on enormous chest X-rays. Besides, we extend MAE to pre-train CNNs on the same scale medical data, establish the first direct benchmark between ViTs and SOTA CNNs on public radiography imaging datasets, and extract reusable insights to the medical vision community.

3. Method

Data. Data from three public X-ray datasets are used to pre-train ViTs (and CNNs as comparison): NIH ChestX-ray14 (75,312 X-rays), Stanford CheXpert (191,028 X-rays), and MIMIC-CXR (243,334 X-rays). All data are in the posteroanterior (PA) or anteroposterior (AP) view, and resized to 256×256 as input. All the X-rays are standardized by mean and standard deviation computed from ImageNet. We perform random resized cropping with a scale range of (0.5~1.0) and random horizontal flipping. No other data augmentation is applied unless noted. The pre-training does not require any annotations shipped with the datasets.

Task. The ViT pre-training² is analogous to the image reconstruction task proposed in MAE [45]: to reconstruct the masked image patches from visible ones. Mean squared error is computed between the reconstructed and original images in the pixel space, averaged over masked patches [24]. An image is divided into regular non-overlapping patches as a sequence of embeddings. We randomly sample patches to be masked. The optimal masking ratio we observe is 90%, which substantially accelerates the pre-training by 2.5× compared with the original MAE [45]¹ and enables us to scale up ViTs with greater model capability (Figure 1a).

Model. The vanilla ViT [26] is used as encoder and applied only on the visible image patches. This design reduces time and memory complexity [45]: a masking ratio of 90% (used in our paper) can reduce the encoder complexity to <1/10. The decoder is another ViT and only used during pre-training to reconstruct the masked patches. Therefore, the decoder is made to be more lightweight than the encoder (depth=2, width=512). As a result, although the decoder processes both visible and masked image patches, its complexity is much smaller than the encoder. Positional embeddings are added to visible and masked patches in this full set to preserve information about their original location in the image. We use ViT-S/16 and ViT-B/16 to denote ViT-Small and ViT-Base with a patch size of 16×16 for simplicity.

4. Pre-training: Recipe and Results

Implementation details. We adopt AdamW [65] optimizer with β₁ = 0.9, β₂ = 0.95 and set the weight decay to 0.05. Transformer blocks are initialized with xavier_uniform [32]. We set learning rate (lr) and batch size to 1.5e-4 and 2048. lr is warmed up for 20 epochs [33] and scheduled with cosine annealing strategy [64]. The pre-training stage takes 800 epochs in total. Random resized crop and horizontal flip are used as data augmentation.

4.1. On the importance of in-domain transfer

Table 1 provides a comprehensive comparison on three sets of model initialization: random, ImageNet pre-training, and X-ray pre-training. Unlike CNNs, ViTs trained from scratch show very poor performance even with a strong training recipe and 2.7× larger number of training epochs than fine-tuning. On the contrary, ViTs achieve comparable and sometimes superior performance with the help of pre-training on large-scale datasets (e.g. ImageNet and X-rays). Specifically, after supervised pre-trained on ImageNet (following [91]), ViT-S/16 shows acceptable performance on the three datasets but is still distanced to the state-of-the-art CNN-based methods. In-domain transfer

---

²The MAE-style pre-training for CNNs (for comparison in Table 1) is similar to the image in-painting task proposed in Models Genesis [113].

¹It should have taken ~16.7 GPU days for the original MAE to pre-train ViT-S/16 (the smallest ViT) on 510K X-rays.
Table 1: Pre-training on ImageNet vs. X-rays. A direct comparison is performed between ViT and three groups of CNNs on three public datasets, considering the number of parameters, volumes of computations, usages of GPUs, and suitable pre-training schemes. The results suggest that ViT (I) consistently exceeds the CNNs that are pre-trained by state-of-the-art pre-training schemes on ImageNet, underlining the importance of in-domain transfer learning (§4.1); (II) surpasses the CNNs that are pre-trained by MAE and MoCo v2 on the same number of medical data (0.3M X-rays); (III) performs comparably (or even better) than state-of-the-art CNNs reported in the literature (detailed in Tables 4–6). Additionally, several important observations are obtained: (i) training from scratch takes longer epochs to converge than fine-tuning pre-trained weights (200 vs. 75 epochs); (ii) ViT shows inferior performance to CNN when training from scratch on X-ray images or fine-tuning from ImageNet; (iii) restorative pre-training (MAE) outperforms contrastive pre-training (MoCo v2) in radiography imaging.

Table 2: A modest random cropping scale is preferred for medical pre-training because the diseased regions (as foreground) is more local than photographic images, and pathological disorder could disperse over the entire X-rays [39] (rather than the center of the image). 

4.2. Learning from 266,340 unlabeled X-rays

Training ViTs from scratch is harder than CNNs because ViTs lack inductive bias in modeling local visual representation and generally require more data to figure out the image content on their own [83, 63, 70]. As shown in Table 1, supervised pre-training on ImageNet brings performance gain from 67.9% to 78.6% for ViTs, and from 80.4% to 82.1% for CNNs on ChestX-ray14. We ask: How many X-rays are needed for ViT pre-training? Figure 1a shows that ViT-S/16 pre-trained on 75K, 191K, 266K, and 510K X-rays achieve a mAUC of 79.3%, 81.9%, 82.3%, and 82.3% on ChestX-ray14. The improvement from 75K to 266K is statistically significant (p-value=1.2×10^{-127}), but the performance gain is negligible from 266K to 510K—a bottleneck for ViT-S/16 (with 22M parameters). Although larger ViTs (e.g. ViT-B/16 with 86M parameters) can produce higher performance, considering computational cost and the fairness of the ViT vs. CNN, we end up pre-training ViT-S/16 using 266,340 unlabeled X-rays for benchmarking.

4.3. Masking out 90% X-ray content

The optimal masking ratio is related to the information redundancy in the data: BERT [55] uses a masking ratio of 15% for language and MAE [45] uses a ratio of 75% for images. Most recent study suggests that videos, due to its greater redundancy in the temporal dimension, can apply only 90% masking ratio for pre-training [28]. Given the great similarity in the chest anatomy, naturally, we hypothesize that chest X-rays require even larger masking ratio for pre-training. This is in line with the assumption that chest X-rays are more information-redundant than photographic images. We experimented with masking ratios ranging from 75% to 95%, incremented by intervals of 5%. Figure 1b indicates that 90% is the optimal masking ratio for MAE pre-training on chest X-rays. The larger masking ratio results in a more efficient pre-training, which is 2.5× faster than the original MAE. The efficient pre-training, in turn, enables us to scale up to larger ViT architectures and more diverse datasets.
Layer-wise LR decay | mAUC (%) | RandAug magnitude | mAUC (%) | DropPath rate | mAUC (%)
---|---|---|---|---|---
45 | 82.1 | 4 | 82.0 | 0.1 | 81.5
55 | 82.3 | 6 | 82.2 | 0.2 | 82.3
65 | 82 | 8 | 82.1 | 0.3 | 82.1

(a) **Layer-wise LR decay.** Learning rate decay in layer-wise needs to be tuned closely.

(b) **RandAug magnitude.** A modest level of augmentation is preferred for fine-tuning.

(c) **DropPath rate.** Fine-tuning on Chest X-ray images needs stronger regularization than natural images.

Table 3: **The fine-tuning recipe.** The ablation studies are conducted using ViT-S/16 on NIH Chest X-ray14. We report the 14-class average AUC (%). Except for (b) using Layer-wise LR decay 0.65, all the experiments adopt the optimal value for the hyper-parameters (Layer-wise LR decay 0.55, RandAug magnitude 6, and DropPath rate 0.2).

4.4. Cropping patches scaled of (0.5~1.0)

As the spatial consistency of medical images is much higher than photographic images, there is a need to analyze the effectiveness of spatial data augmentation (e.g., random resized crop). Figure 1b suggests that RandomResizedCrop operation has consistent and noticeable benefit to MAE pre-training on chest X-ray imaging under different masking ratios. It enables ViTs to learn multiscale features from X-rays and to avoid the over-fitting problem due to the lack of training examples. More importantly, a relatively smaller cropping ratio than natural imaging is preferred (Table 2). Cropping patches scaled of (0.5~1.0) yields 1.0% higher mAUC than those of (0.2~1.0) (as suggested in [45]). It is intuitive that strong spatial augmentation is harmful since the informative lesions or organs could be cropped and biased and models will be learned with noisy annotations.

4.5. Quality assessment of image reconstruction

We assess reconstruction quality using validation images in Figure 2 for both ViT and CNN. The models are pre-trained on ChestX-ray14 and evaluated on the inputs with varying masking ratios, spanning from 75% to 90%, incremented by intervals of 5%. Both ViTs and CNNs can predict the overall anatomical structures in X-rays, but fail to reconstruct detailed texture such as shoulder bones. This is expected because the ViTs/CNNs only see 10% of the input image and attempt to reconstruct the rest 90% during the training—it is difficult even for expert radiologists. There is no clear evidence showing that the reconstruction capability is positively correlated to the transfer learning performance. On the contrary, the original autoencoders [48] (with a masking ratio of 0%) can certainly reconstruct images better than masked autoencoders, but their resulting representation is not as effective as the masked counterparts [113]. Moreover, studies in both CNNs [87, 113] and ViTs [45, 101] indicate that alternative loss functions (e.g., $l_1$, smooth-$l_1$, SSIM, and adversarial losses) for reconstruction would not contribute to the transfer learning performance. Therefore, we used $l_2$ loss as default. Finally, we should remark that our ultimate goal is not the task of image reconstruction per se. While reconstructing patches is advocated and investigated as a pre-training scheme for ViTs/CNNs, the usefulness of the learned representation must be assessed objectively based on its generalizability and transferability to various downstream tasks (presented in §5).

5. Fine-tuning: Recipe and Results

**Fine-tuning.** The optimizer and $lr$ scheduler are the same as pre-training. The choices of layer-wise LR decay, RandAug [22] magnitude, and DropPath [50] rate are crucial.

*No improvement is obtained by more aggressive augmentation strategies (i.e., mixup [106] and cutmix [104]) since they could produce noisy...*
to fine-tune the pre-trained ViTs. The optimal settings are given by extensive studies in Table 3 and we reuse them for all three radiography imaging tasks in §5.3–5.2. Models are fine-tuned with 75 epochs on all three datasets.

Linear-probing. LARS [103] optimizer is used with momentum=0.9. We set learning rate (lr) and batch size to 0.1 and 16,384. lr is warmed up [33] for 10 epochs and scheduled with cosine annealing strategy [64]. The ViT is trained with 100 epochs. Linear-probing is used in Figure 1b.

5.1. Stanford CheXpert

Experimental setup. CheXpert is a large scale dataset containing 191,028 frontal-view chest X-rays. 14 diseases in radiology reports exist in the dataset and five common diseases are for benchmarking. We resized the images into 224×224 and the test is done on the official validation set. Mean Area Under the Curve (AUC) on five classes is reported for comparison.

Results and analysis. As shown in Table 4, vanilla ViT-S achieves 89.2% mAUC which is very competitive to the best performance of 89.4%. Moreover, ViT-S yields the best performance on diseases of Atele (83.5%) and Edem (94.0%).

5.2. COVIDx

Experimental setup. COVIDx (version 9A) provides over 30,000 images containing 16,490 positive COVID-19 images. The dataset is annotated with 4 different classes for the training set of 30,130 images while the testing set only has 400 images of 3 classes. To ensure a fair comparison with previous methods, Accuracy and COVID-19 sensitivity on the testing set (3 classes) are reported.

Results and analysis. We compare our vanilla ViT-S model with the state-of-the-art models provided on the official github repository5. Our method beats the two other models when input resolution is 224×224 while achieving a very high accuracy of 95.2% and COVID-19 sensitivity of

Table 4: CheXpert benchmark. ViT achieves comparable performance to the state-of-the-art CNNs on CheXpert (official val) over all five thorax diseases and the best o “Atelectasis” and “Edema” diseases.

Table 5: COVIDx benchmark. ViTs show comparable performance to state-of-the-art CNNs on COVIDx (official val).
5.3. NIH ChestX-ray14

Experimental setup. ChestX-ray14 has 112,120 frontal-view X-rays of 30,805 unique patients with the text-mined fourteen disease labels (where each image can have multiple labels). We follow the official data split which assigns 75,312 images for training and 25,596 images for testing. We rezone the original images from size 1024×1024 into 224×224. Mean AUC on 14 classes is reported and 17 most popular and compelling baseline methods are compared.

Results and analysis. Table 6 provides a systematic comparison with state-of-the-art CNN and Transformer models on NIH ChestX-ray14 over years. The previous best CNN performance was obtained by DenseNet-121 [94] with a mean AUC of 82.6%. The previous best performance of Transformers was 81.0% [88], which was distanced to CNN’s performance. Our vanilla ViT-S shows a very competitive result of 82.3% mean AUC over 14 diseases with the best classification performance on 6 out of 14 thorax diseases. It is worth noting that the research community took four years to improve the AUC score from 74.5 to 82.2 for CNN-type architectures, largely due to the difficulty of the training recipe.

<table>
<thead>
<tr>
<th>Method</th>
<th>Architecture</th>
<th>Pre-training</th>
<th>mAUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wang et al. [95]</td>
<td>RN50</td>
<td>0.5M X-rays</td>
<td>74.5</td>
</tr>
<tr>
<td>Yao et al. [102]</td>
<td>RN6DN</td>
<td>0.5M X-rays</td>
<td>76.1</td>
</tr>
<tr>
<td>Li et al. [59]</td>
<td>RN50</td>
<td>0.5M X-rays</td>
<td>75.5</td>
</tr>
<tr>
<td>Tang et al. [65]</td>
<td>RN50</td>
<td>0.5M X-rays</td>
<td>80.3</td>
</tr>
<tr>
<td>Guendel et al. [56]</td>
<td>DN121</td>
<td>X-rays (0.3M)</td>
<td>80.7</td>
</tr>
<tr>
<td>Guan et al. [35]</td>
<td>DN121</td>
<td>X-rays (0.3M)</td>
<td>81.6</td>
</tr>
<tr>
<td>Wang et al. [93]</td>
<td>R152</td>
<td>X-rays</td>
<td>78.8</td>
</tr>
<tr>
<td>Ma et al. [68]</td>
<td>R101</td>
<td>X-rays (14M)</td>
<td>79.4</td>
</tr>
<tr>
<td>Baltruschat et al. [5]</td>
<td>RN50</td>
<td>X-rays (0.3M)</td>
<td>80.6</td>
</tr>
<tr>
<td>Seyyed et al. [80]</td>
<td>DN121</td>
<td>X-rays (0.3M)</td>
<td>81.2</td>
</tr>
<tr>
<td>Ma et al. [65]</td>
<td>DN121×(2)</td>
<td>X-rays (0.3M)</td>
<td>81.7</td>
</tr>
<tr>
<td>Hermota et al. [47]</td>
<td>DN121</td>
<td>X-rays (0.3M)</td>
<td>82.1</td>
</tr>
<tr>
<td>Kim et al. [57]</td>
<td>DN121</td>
<td>X-rays (0.3M)</td>
<td>82.2</td>
</tr>
<tr>
<td>Haghighi et al. [39]</td>
<td>DN121</td>
<td>X-rays (0.3M)</td>
<td>81.7</td>
</tr>
<tr>
<td>Liu et al. [61]</td>
<td>DN121</td>
<td>X-rays (0.3M)</td>
<td>81.8</td>
</tr>
<tr>
<td>Tadimeti et al. [88]</td>
<td>SwinT</td>
<td>X-rays (0.3M)</td>
<td>81.0</td>
</tr>
<tr>
<td>Baltruschat et al. [5]</td>
<td>MoCo v2</td>
<td>X-rays (0.3M)</td>
<td>80.6</td>
</tr>
<tr>
<td>Ours</td>
<td>MAE</td>
<td>DN121</td>
<td>X-rays (0.3M)</td>
</tr>
<tr>
<td>Ours</td>
<td>MAE</td>
<td>ViT-S/16</td>
<td>82.3</td>
</tr>
<tr>
<td>Ours</td>
<td>MAE</td>
<td>ViT-B/16</td>
<td>X-rays (0.3M)</td>
</tr>
</tbody>
</table>

Table 6: ChestX-ray14 benchmark. ViT-S/16 achieves comparable performance to previous state-of-the-art CNN-based and Transformer-based methods on ChestX-ray14 (official split) reported in the literature. With the same pre-training scheme (MAE) on 0.3M X-rays, ViT significantly outperforms its CNN counterparts. In addition, ViT-B/16, pre-trained on 0.5M X-rays, hits a new record of 83.0 mAUC. RN, DN, and SwinT denote ResNet, DenseNet, and Swin Transformer.

94.5%. ViT-S shows a great balance between the model size, computation cost, and the performance.

6. Discussion

Can MAE detect anomalies from an image? Anomalies are something that appear differently from the normal X-rays—can be diseases, medical devices, and clinical notations (e.g. arrows, numbers, letters). Since MAE is trained using the original X-rays as ground truth and the majority pixels in an X-ray are normal, the MAE should be able to overfit the normal anatomical patterns. Now, if an anomaly is masked out, can MAE reconstruct a normal pattern? If so, by subtracting the reconstructed output and the original X-ray, the anomaly can be detected and localized. A similar point has been discussed in Zhou et al. [113], but was illustrated using CNNs. Specifically, we input an original X-ray to the trained MAE and plot the difference map between reconstructed output and the original image. As shown in Figure 3, MAE happens to “heal” those anomalies and reconstruct with normal patterns. This behavior can be thought of as an attempt to detect and localize anomalies. More importantly, unlike weakly-supervised detection strategies [108, 7, 9, 82, 96], neither image-level nor pixel-level annotation is required for this approach, making it an attractive and challenging direction to explore [76, 90].

Weakly-supervised disease localization by ViT and CNN. With the help of Grad-CAM⁶ [31], we are able to check which part of the X-ray image is responsible for the model prediction (the diseased region). We use the last dense-block (4th) of DenseNet-121 and the LayerNorm layer in the last transformer block (12th) of ViT-S/16 as the “target layers” for Grad-CAM. The experiments are done in a small subset of ChestX-ray14, which offers 787 cases with bounding-box of a total of eight thorax diseases. The final predicted bounding-box of the diseased region is generated with the thresholded Grad-CAM heatmap, largest connected component, and box regression. The results are evaluated by IoU between ground truth bounding box and the bounding box of the largest connected component in

---

⁶github.com/jacobgil/pytorch-grad-cam

Figure 3: MAE could reveal anomalies. We input a chest X-ray (with anomalies) to the trained MAE and plot the difference map of the reconstructed output and original X-ray. Interestingly, we observe that the MAE happens to “heal” those anomalies by replacing them with normal patterns.
Table 7: Weakly-supervised disease localization. We report the average precision (AP) on 25% and 50% IoUs. The IoU is calculated between the ground truth bounding box and bounding box of the largest connected component in the Grad-CAM heatmap. We also present the statistics of disease sizes, measured by the number of pixels within the bounding box, showing that CNN can detect large diseases (e.g. Cardiomegaly, Pneumonia) better than ViT, while ViT can capture smaller diseases (e.g. nodule).

<table>
<thead>
<tr>
<th>Disease</th>
<th>Size (# of px)</th>
<th>DenseNet-121</th>
<th>ViT-S/16</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>AP25</td>
<td>AP50</td>
</tr>
<tr>
<td>Nodule</td>
<td>224</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Mass</td>
<td>756</td>
<td>25.4</td>
<td>1.6</td>
</tr>
<tr>
<td>Atelectasis</td>
<td>924</td>
<td>10.1</td>
<td>2.0</td>
</tr>
<tr>
<td>Pneumothorax</td>
<td>1899</td>
<td>11.6</td>
<td>2.3</td>
</tr>
<tr>
<td>Infiltrate</td>
<td>2754</td>
<td>32.9</td>
<td>12.7</td>
</tr>
<tr>
<td>Effusion</td>
<td>2925</td>
<td>24.5</td>
<td>2.9</td>
</tr>
<tr>
<td>Pneumonia</td>
<td>2944</td>
<td>32.0</td>
<td>6.2</td>
</tr>
<tr>
<td>Cardiomegaly</td>
<td>8670</td>
<td>89.6</td>
<td>53.3</td>
</tr>
<tr>
<td>All</td>
<td>2300</td>
<td>31.0</td>
<td>12.3</td>
</tr>
</tbody>
</table>

7. Conclusion and Future Work

This paper has unleashed the potential of stand-alone, vanilla ViT by devising strong pre-training and fine-tuning recipes. We overcome several technical barriers and bring reusable insights for the medical vision community. Specifically, we (i) improve computational efficiency; (ii) customize data augmentation; (iii) explore larger data scale; and (iv) optimize learning parameters. As a result, the vanilla ViT achieves a comparable (sometimes better) performance to state-of-the-art CNNs. Code and pre-trained models are available.

This paper has also presented an up-to-date benchmark on three predominant chest X-ray datasets. Taking into account the number of parameters, volumes of computations, usages of GPUs, and suitable pre-training schemes, we have performed a fair and comprehensive comparison between vanilla ViT and (i) state-of-the-art CNNs reported in the literature, (ii) CNNs that are pre-trained by advanced pre-training schemes on ImageNet, (iii) CNNs that are pre-trained on the same number of medical data. We hope this study can direct future research on the application of Transformers to a larger variety of medical imaging tasks.

As future work, we will consider three extensions to our current study. First, assembling more publicly available X-ray datasets for pre-training (which account for a total of ∼1M images [10]). Scaling up the data is perhaps the most straightforward way to enhance larger ViTs (e.g. ViT-Large, ViT/Huge) in terms of performance and generalizability based on Figure 1a. Second, extending ViT to its 3D form for higher dimensional medical modalities (e.g. CT, MRI), which is expected to take a considerable computational resource [44] and larger data for pre-training [86], therefore requiring a more efficient method. Third, exploiting paired information of radiology reports and image data for pre-training. We acknowledge the unique ability of Transformers in processing multi-modality data [78, 20].

Acknowledgements. This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research. We thank Y. Zhang for providing data loader of the COVIDx dataset; A. Delaney for improving the writing of this paper.
**References**


[45] 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. 1, 3, 4, 5


European conference on computer vision, pages 646–661. Springer, 2016. 5


[59] Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei. Thoracic disease identification and localization with limited supervision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8290–8299, 2020. 7


[75] Hieu H Pham, Tung T Le, Dat Q Tran, Dat T Ngo, and Ha Q Nguyen. Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels. *Neurocomputing*, 437:186–194, 2021. 6


[97] Junfei Xiao, Yutong Bai, Alan Yuille, and Zongwei Zhou. Transforming radiograph imaging with transformers: Com-
paring vision transformers with convolutional neural networks in multi-label thorax disease classification. In Radiological Society of North America (RSNA), 2022. 2


[104] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoo Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In ICCV, 2019. 5


