KNOWLEDGE DISTILLATION FOR EFFICIENT TINY MASKED AUTOENCODERS FOR MEDICAL IMAGING

Dani Dassum, Maximo Librandi, Manuel Paez
Department of Computer Science
Columbia University
New York, NY 10027, USA
{dd3007,ml5014,map2332}@columbia.edu

ABSTRACT

Masked Autoencoders (MAE) have become one of the most widely used self-attention architectures in the medical and biological domain due to their great scalability, computational efficiency, and how pre-trained representations generalize well to various downstream tasks. However, two challenges remain: (1) training can still be prohibitively costly as Vision Transformer (ViT) models require significant amounts of data, and (2) smaller Vision Transformer (ViT) models have poor performance and poor pre-trained representations compared to strong, cumbersome models. Due to the lack of annotated medical image data available, ViT-based models are more unfavorable for downstream tasks compared to their Convolutional Neural Network (CNN) counterparts. In this paper, we study several techniques to improve the efficiency of smaller models for medical image tasks. Firstly, we adopt a masked autoencoder (MAE) to enable efficient and competitive performance in various downstream medical image tasks. Secondly, we perform pre-training along with knowledge distillation to transfer the feature representation knowledge from larger, strong models into smaller ones. Finally, we adopt a recipe for fine-tuning two X-ray medical datasets for Multi-label Image Classification. Our resulting model, the Efficient Tiny Masked Autoencoder (ETMAE), performs similar to other baselines on NIH ChestX-ray14 and Stanford CheXpert datasets but being 15 times smaller. Our code and models are publicly available at https://github.com/dd3007/AdvTopicsDL-Project

1 INTRODUCTION

A fundamental challenge in medical image analysis is quantifying the interdependencies and relationships of different anatomical structures that functionally interact with other structures and regions in the human body. With small and limited medical image datasets, where many images are unlabeled, it is vital for models to learn the representations of these medical images.

Following recent advances in NLP, the transformer architecture shows excellent potential in computer vision, especially when pre-trained with a large amount of unlabeled data and using self-supervised learning methods. Compared with Convolutional Neural Networks (CNNs), Vision Transformers (ViT) better leverage the image data, the long-range spatial context of an image, and are much more computationally efficient.

The pre-training of CNNs has been widely investigated and adopted for medical image analysis due to their powerful ability to learn representations from unlabeled images. One powerful feature pre-train method for transformers is Masked Image Modeling (MIM), which trains models to predict the masked signals of the input image. A reasonably widespread adoption of this work is masked autoencoders (MAE), which have been shown to efficiently and effectively pre-train Vision Transformers with strong feature representations and performance on various downstream tasks.

However, given the limited access to computational resources in the medical domain, developing efficient, faster, smaller, yet powerful models for medical image tasks needs to be investigated. We ask: how can we develop a fast and efficient framework for smaller ViT-based models in medical imaging tasks? We answer this question through several different perspectives.

1. Masked autoencoders (MAE), an asymmetric encoder-decoder architecture, is viable for self-supervised learning and for visual representation learning. Autoencoding is a classical technique

for learning representations. An encoder maps input to a latent representation, and a decoder reconstructs the input. For vision, He et al. (2022) showed that masking random patches for the input and reconstructing the missing patches from the latent representations is effective for vision representation learning tasks. Their decoder design, which features shifting the mask tokens to the decoder, results in a model having a large reduction in computation consumption.

2. Knowledge Distillation is more effective for solid performance on downstream tasks when used in pre-training. For Masked Autoencoders, fine-tuning knowledge distillation methods such as Touvron et al. (2021) were shown to be ineffective due to the inability to transfer the feature representation of the teacher model to student models for downstream performance. With MAE, a pre-training knowledge distillation framework that efficiently transfers feature representations of stronger models is the most viable option.

In this work, we present the Efficient Tiny Masked Autoencoder (ETMAE) model for medical image tasks. We adopt the MAE from He et al. (2022), which features an asymmetric encoder-decoder design to mask random patches from the input image and reconstruct the missing patches in the pixel space. During pre-training, we apply a knowledge distillation framework similar to Bai et al. (2023) from a large, strong teacher MAE model to a small student MAE, which is an effective method for transferring feature representations of the distilled model. Distilling pre-trained models in this framework leads to substantial performance improvements of smaller models on downstream tasks. Finally, we developed a fine-tuning recipe similar to Xiao et al. (2023) for medical image classification, which has been shown to be competitively comparable to CNN's on the medical image tasks. By combining these frameworks and techniques, ETMAE is a faster, more efficient, and more scalable student model for medical image classification.

We experiment with our model on two public X-ray datasets for Multi-label Image Classification: NIH ChestX-ray14 (75,312 X-rays) and Stanford CheXpert (191,028) X-rays). We found that for a 75 % masking ratio, the ETMAE was able to achieve similar performance to other baselines on these datasets while using a model that is 15 times smaller.

To summarize, our main contributions include:

- We adopt the MAE for our models for its performance on various downstream medical imaging tasks.
- We employ a specialized pre-training distilling method to learn MAE's feature representations.
- We develop specialized pre-training and fine-tuning receipts for medical image classification on NIH ChestX-ray14 and Stanford CheXpert datasets.

We hope this work can benefit future research on efficiently unleashing the power of small, efficient pre-train models for medical image analysis.

2 Related Work

Vision Transformers

The introduction of the Transformer architecture Vaswani et al. (2017) in Natural Language Processing marked a seminal point in the evolution of models, moving away from recurrent models towards those based on self-attention mechanisms for sequence modeling. This paradigm shift in NLP has led to the development Vision Transformers (ViTs) in computer vision, which optimize data processing by focusing selectively on various segments of input data. The seminal paper Dosovitskiy et al. (2020) demonstrated the potential of ViTs to handle image data by processing images as sequences of fixed-size patches. This method proved particularly beneficial for managing complex, large datasets such as ChestXray14 and CheXpert in medical imaging.

Masked Autoencoders

Autoencoders are classical neural network models that learn to compress input data into a latent-space representation and then reconstruct the output from this representation, serving as a tool for dimensionality reduction and feature identification Tschannen et al. (2018). Masked Autoencoders (MAEs), as introduced by He et al. (2022), advance this concept by using an asymmetric encoder-decoder architecture that processes only visible, unmasked patches of an image. This approach significantly reduces computational demands. Our approach mirrors this architecture by utilizing a Vision Transformer (ViT) based encoder that selectively encodes these visible patches, optimizing computational resources Cao et al. (2022). The corresponding lightweight ViT decoder is then tasked with reconstructing the full image from this sparse encoded representation, focusing on restoring the image with high fidelity from minimal input data. This strategic reduction in computational demands is particularly beneficial for medical imaging where data volumes are large and processing efficiency is crucial.

Knowledge Distillation

Following the seminal paper Hinton et al. (2015), Knowledge Distillation is a framework to transfer knowledge from large models to smaller ones while allowing the small, less-resource-intensive model to maintain similar performance levels as the larger model. In transformers, papers such as DeIT Touvron et al. (2021) have shown ways to distill knowledge in the fine-tune phase. Bai et al. (2023) advances our understanding of using Masked Autoencoders (MAEs) for knowledge distillation. This research is crucial for our approach, particularly because it focuses on the distillation process at the encoder level of the student model. Knowledge distillation, as explored by Bai et al. (2023) Huang et al. (2023), Lao et al. (2023), Wu et al. (2022), involves using masked inputs which serve a dual purpose: they reduce the computational complexity of the distillation process and enhance the student model's ability to selectively absorb critical features from the teacher model. This selective learning enables the student model to efficiently master essential diagnostic features, crucial for the high accuracy required in medical imaging. Implementing this technique streamlines the training process, considerably improving the performance of the student model while ensuring it remains resource-efficient, and enables further developments such as that proposed by Chen et al. (2022) which explores leveraging high-level representations in distillation processes.

Masked Autoencoders for Medical Image Analysis

Masked Autoencoders (MAEs) in the context of medical imaging have shown considerable innovative approaches like that of Luo et al. (2022), who enhance MAE's encoder capabilities through self-distillation for histopathological image analysis. Our interest particularly lies in Xiao et al. (2023)'s application, where MAEs are adapted for thorax disease classification using Vision Transformers (ViTs) and large-scale datasets such as CheXpert and ChestX-ray14. Their approach customizes MAEs to handle the unique challenges of medical datasets, which feature high variability and specificity in imaging characteristics. This adaptation aligns with our project as we utilize these datasets to train our dual-model system, involving knowledge distillation between a complex teacher model and a student model, and finally demonstrate that a 90% masking ratio during pre-training effectively reduces computational demands while enabling the model to learn essential features from minimal data.

3 Approach

3.1 Masked Autoencoders

Our method is built upon MAE, a powerful autoencoder-based MIM approach. Specifically, the MAE encoder first projects unmasked patches to a latent space, which are then fed into the MAE decoder to help predict pixel values of masked patches. The core elements in MAE include:

Masking: MAE operates on image tokens formed when the image is divided into non-overlapping patches. An arbitrarily small subset of these patches is used for the MAE encoder, and the rest is set as the predicting target of the MAE decoder. We apply a masking ratio of 75% for our model.

MAE encoder: The MAE encoder is a standard ViT architecture that is modified to take unmasked patches on input. This design reduces the computational cost of the encoder.

MAE decoder: The MAE decoder receives masked tokens and the encoded features of unmasked patches as input. The masked token is a learned vector shared across all missing positions. We choose a lightweight decoder to reduce computational cost.

Model: We use MAE ViT-B/16, MAE ViT-S/16, and MAE ViT-T/16 to denote MAE ViT-Base, MAE ViT-Small, and MAE ViT-Tiny with a patch size of 16 x 16.

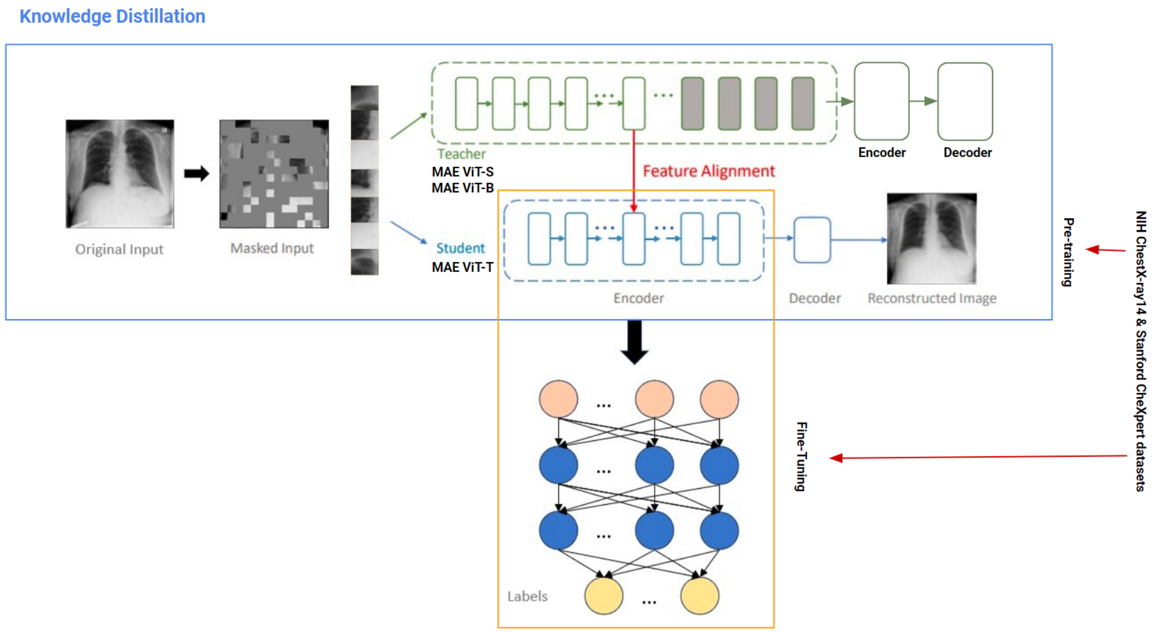


Figure 1: **Our ETMAE Architecture.** We divide the model in a pre-train (Knowledge Distillation) step, which performs MAE teacher-student feature alignment in their respective encoders to extract masked inputs as done in Bai et al. (2023); and a fine-tuning step, in which our student models are fine-tuned for the downstream image classification task. We use both the NIH ChestX-ray14 and Stanford CheXpert datasets for this.

Reconstruction: The MAE reconstructs image pixel values using the Mean Squared Error (MSE) applied to masked tokens to calculate the loss.

Figure 1 shows the proposed architecture. The upcoming subsections provide details of each component.

3.2 Knowledge Distillation

MAE has shown exceptional capabilities in learning high-capacity models efficiently, effectively, and with scalability. This is especially useful for analyzing downstream tasks in the medical image domain, where there are few datasets and fewer computational resources to leverage. In this paper, we seek to bridge knowledge distillation with the MAE framework to develop a small, efficient, and fast model with similar performances as powerful and larger models. This framework will then be used to pre-train and fine-tune on two datasets for medical image classification. Although approaches such as the DeiT Touvron et al. (2021) framework have shown to work for ViT-based architectures, Bai et al. (2023) shows that this framework has little to no performance improvements for MAEs while also failing to leverage its design for computational efficiency. With this, we adopt and modify Bai et al. (2023) approach of directly applying knowledge distillation at the pre-training stage.

For our specific approach, we extract the features from the specific layers of the student model. This is done as MAE pre-training has no categorical labels, which leads to distilling the intermediate features. After feeding these features into a small project head, the outputs are then used to mimic the features from the corresponding layers of the teacher model.

Formally, it is as follows: For MAE reconstruction, consider H as image height and W as image width, let $x \in \mathbb{R}^{3HW \times 1}$ be the input pixel RGB values and let $y \in \mathbb{R}^{3HW \times 1}$ be the predicted pixel values. The MAE reconstruction loss L_{MAE} is defined as follows:

$$L_{MAE} = \frac{1}{\Omega(X_M)} \sum_{i \in M} (y_i - x_i)^2 \tag{1}$$

where M denotes the set of masked pixels, $\Omega(\cdot)$ is the number of elements, and i is the pixel index. For feature alignment distillation, let $z_l^S, z_l^T \in \mathbb{R}^{LC \times 1}$ be the features extracted from the respective l-th layer of the student model and the teacher model, where l denotes the patch numbers and C denotes the channel dimension. We use $\sigma(\cdot)$ to denote the projection function of the network. Our feature alignment distillation loss L_{DAE} is defined as follows:

$$L_{Dist} = \sum_{l} \frac{1}{\Omega(z_l^T)} \sum_{i} \left\| \sigma(z_l^S)_i - z_{l,i}^T \right\|_1 \tag{2}$$

where $\|\cdot\|_1$ is the l_1 norm. For our final loss L in pre-training, a weighted sum of the MAE reconstruction loss and the feature alignment loss multiplied with a parameter α is used:

$$L = L_{MAE} + \alpha \times L_{Dist} \tag{3}$$

The framework of the Knowledge Distillation structure is shown in Figure 1. Following MAE, our method takes masked inputs and performs Masked Image Modeling (MIM). The corresponding features are aligned between the teacher model and student model. Since (1) the Knowledge Distillation method operates on a tiny subset of visible patches; and (2) aligning intermediate features reduces the computational cost of the teacher model, our method is computationally efficient.

3.3 Medical Image Fine-tuning

Fine-tuning is a crucial step in adapting the pre-training model to downstream tasks. Once the student model has been pre-trained using the Knowledge Distillation framework, we fine-tune the model for the multi-label image classification task. To achieve this, we take the encoder layers of the student model and add a classifier on top, as shown in Figure 1. The classifier is a fully connected network with the number of classes as the output dimension. We fine-tune the model separately on the training split of each dataset and evaluate it on each validation split.

3.4 Data

As described in Xiao et al. (2023), the teacher ViT-B and ViT-S models were pre-trained on data from three public X-ray datasets: 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.

Because of our limited resources, for the Knowledge Distillation process, we pre-train our student models on NIH ChestX-ray14 and Stanford CheXpert. These student models are later fine-tuned for the multi-label image classification task. The NIH ChestX-ray14 dataset contains 14 different disease labels, including Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening, and Hernia. The Stanford CheXpert dataset includes five different disease labels - Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion.

We use the train split of the datasets for the training and the validation splits for evaluation. Lastly, we evaluate the models fine-tuned on NIH ChestX-ray14 on the official test data, while the models fine-tuned on Stanford CheXpert are evaluated on the official validation data. This is because the test set for Stanford CheXpert is hidden, and the competition is already closed.

3.5 Evaluation Metrics

Both NIH ChestX-ray14 and Stanford CheXpert datasets require the multi-label classification task to be solved. Thus, we report the Mean Area Under Curve (mAUC) metric for comparison with existing benchmarks. The mAUC measures the average area under the ROC curve for each label. The ROC curve represents the true positive rate against the false positive rate for different thresholds. The mAUC is the average of the area under the ROC curve for each label. The mAUC is a good metric for evaluating the performance of multi-label classification models, as it considers the trade-off between true positive and false positive rates for each label.

	Teacher		Student	
	Model	# Parameters	Model	# Parameters
Experiment 1	ViT-Small	47.6M	ViT-Tiny	12.5M
Experiment 2	ViT-Base	111.9M	ViT-Tiny	12.6M

Table 1: Configurations for the teacher and student models on each pre-training experiment.

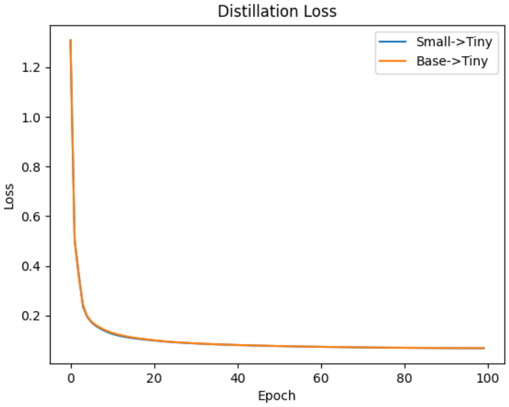


Figure 2: Training loss curves for each experiment during the pre-training phase.

4 Results and Analysis

Below, we discuss the results obtained during both the knowledge distillation and fine-tuning steps of our pipeline. We lastly evaluate our models on the corresponding datasets and benchmark their results.

4.1 Knowledge Distillation

For feature alignments, we follow Bai et al. (2023) and choose to align the features from the $\frac{3}{4}$ depth of both the student model and the teacher model. From all the models we tested, we noticed exciting results.

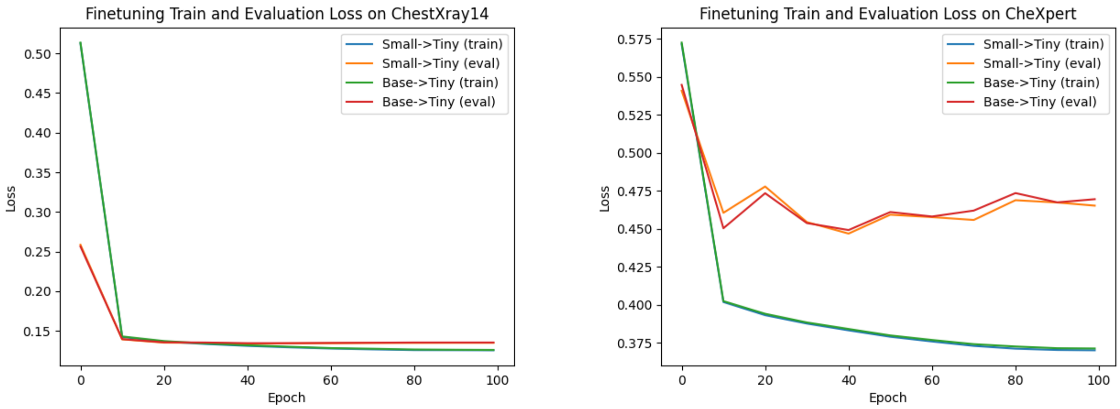
As a first experiment, we apply knowledge distillation by considering a ViT-Small as the teacher model and a ViT-Tiny as the student model. The second experiment involves a ViT-Base as the teacher model and a ViT-Tiny as the student again. In both cases, the pre-training was run for 100 epochs using both NIH ChestX-ray14 and Stanford CheXpert datasets, a batch size of 32, a learning rate of 1.5e-4, and a weight decay of 0.05. The optimizer used was AdamW. The configurations for the teacher and student models on each experiment are shown in Table 1. Figure 2 shows the training loss curves for each experiment during the pre-training phase.

4.2 Fine-tuning

Once the MAE student models are pre-trained on the medical data, we fine-tune them on each dataset for the multi-label image classification task. We use the training split of each dataset for the training phase and the validation split for evaluation.

First, we fine-tune the models on the NIH ChestX-ray14 dataset, which involves predicting 14 different disease labels. Next, we fine-tune the models on the Stanford CheXpert dataset, which includes five disease types. In both cases, we fine-tune for 100 epochs using the AdamW optimizer, a batch size 32, and a base learning rate of 2.5e-4. We use a learning rate scheduler with a layer decay of 0.55 and a weight decay 0.05. The loss function used is the Binary Cross-Entropy loss, and the models are evaluated with the Mean Area Under Curve (mAUC) metric.

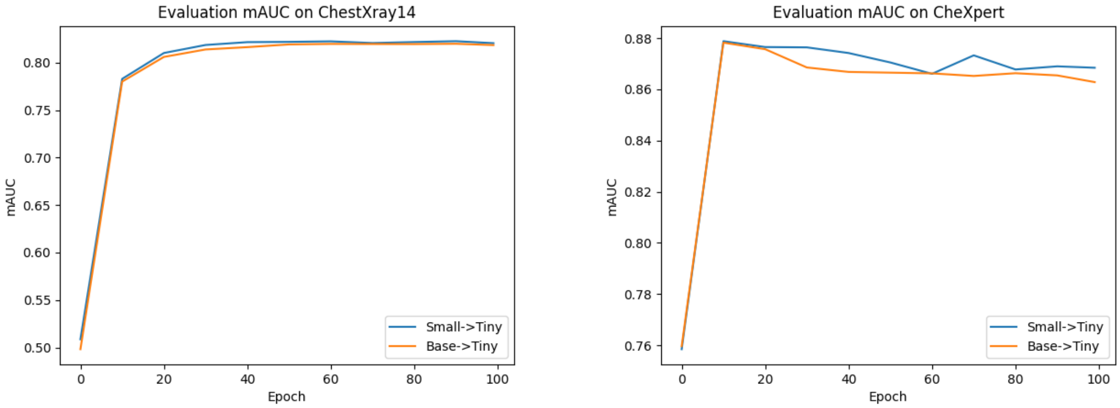
Figure 3 shows both the training and validation loss curves for the fine-tuning phase of the models on each dataset. The training loss decreases as the model learns to predict the disease labels, while the validation loss monitors overfitting. The mAUC scores for each model on the validation partition of each dataset are shown in Figure 4.



(a) ChestX-ray14 dataset

(b) CheXpert dataset

Figure 3: Training and validation loss curves for the fine-tuning phase of the models on each dataset.



(a) ChestX-ray14 dataset

(b) CheXpert dataset

Figure 4: mAUC scores for each model on the validation data of each dataset.

4.3 Evaluation

The last step is to evaluate the models on the official test data of the NIH ChestX-ray14 dataset and the official validation data of the Stanford CheXpert dataset, since the test data is hidden. Table 2 shows the mAUC scores for each model on their corresponding fine-tuned dataset. We also include the mAUC scores of both models presented in Xiao et al. (2023) for comparison.

As shown in Table 2, our ETMAE models does not perform as well as the models from Xiao et al. (2023) on the NIH ChestX-ray14 dataset. Our ViT-Tiny model distilled from the ViT-Small model has a difference of 4.2 and 4.9 mAUC points compared to the ViT-Small and ViT-Base models, respectively. The ViT-Tiny model distilled using the ViT-Base model differs by 4.1 and 4.8 mAUC points from the ViT-Small and ViT-Base models, respectively.

For the Stanford CheXpert dataset, our models again perform worse than the models from Xiao et al. (2023), but the difference is smaller. Our ViT-Tiny model distilled from the ViT-Small model has a difference of 2.3 and 2.4 mAUC points compared to the ViT-Small and ViT-Base models, respectively, while the ViT-Tiny model distilled from the ViT-Base model differs by 1.4 from the ViT-Small and 1.5 mAUC points from the ViT-Base.

Despite the small differences in performance, our goal to build efficient, smaller models for medical image tasks is achieved. With no significant loss in performance, our ETMAE models are up to 15 times smaller

System	Model	# Params	ChestX-ray14	CheXpert
Xiao et al. (2023)	ViT-S	21.7M	82.3	89.2
Xiao et al. (2023)	ViT-B	85.8M	**83.0**	**89.3**
ETMAE	ViT-Tiny (from Small)	5.5M	78.1	86.9
ETMAE	ViT-Tiny (from Base)	5.5M	78.2	87.8

Table 2: mAUC scores for each model on the test data of the NIH ChestX-ray14 dataset and the validation data of the Stanford CheXpert dataset.

than the models from Xiao et al. (2023), which makes them more computationally efficient. Our ViT-Tiny models have 5.5M parameters, while the ViT-Small model from Xiao et al. (2023) has 21.7M parameters, and the ViT-Base model has 85.8M parameters.

By showing these results, we conclude that smaller models benefit more from knowledge distillation when distilled from larger models rather than smaller models. This is consistent with the results from Bai et al. (2023) and Xiao et al. (2023). Moreover, we show that pre-training on medical data is crucial for the performance of the models on medical image tasks.

5 Conclusion

Self-supervised pre-training models have demonstrated great success in the medical and biological domains. However, because of the limited resources available in these domains, models with a balanced trade-off between performance and computational efficiency in real-world applications are often preferred. Envisioned to help medical institutions and practitioners with limited computational resources, our ETMAE is a small yet efficient model that unleashes the potential of Knowledge Distillation and Masked Autoencoders.

This work presents an architecture combining the MAE framework with Knowledge Distillation to develop a small, efficient, and fast model for medical image tasks. We pre-train our models on the NIH ChestX-ray14 and Stanford CheXpert datasets by performing feature alignment distillation between the teacher and student models. These models are then fine-tuned for the multi-label image classification task on the same datasets. Lastly, our models are evaluated on the official data partitions to benchmark their performance. Extensive experiments on multiple model scales demonstrate the effectiveness of our strategy.

Our results show that our ETMAE models perform similarly to the models from Xiao et al. (2023) on the NIH ChestX-ray14 and Stanford CheXpert datasets. Despite insignificant decreases in performance, our models are up to 15 times smaller than the baseline models from Xiao et al. (2023), which makes them more computationally efficient. Our results also show that smaller models benefit more from knowledge distillation when distilled from larger models rather than smaller models. We hope future works can study the relationship of distilled larger MAE models with smaller models and how to leverage efficient techniques and small models for self-supervised medical image tasks.

6 Contributions

Manuel researched on Autoencoders, Masked Image Modeling frameworks, Vision Transformers, and Knowledge Distillation. He also researched efficient Attention mechanisms with no coding success. Dani researched on Self-Supervision in the medical domain, especially for image classification. She also researched on Masked Autoencoders and another Knowledge Distillation variant called TinyViT. Maximo researched on Vision Transformers and Knowledge Distillation methods. Before our topic of discovery, Maximo also researched on LLMs and to leveraged them for specific semantic information tasks. After brainstorming on potential ideas, we start with Efficient-Tiny MAE development. Manuel implemented the code required to perform Knowledge Distillation, which takes the features from the teacher model and aligns them with the student model. Maximo helped implement the code for Knowledge Distillation, including essential parameters and design changes for effective distillation. Maximo also implemented the code required to fine-tune the models on the medical image datasets. He also implemented the code to access the datasets. Dani implemented the code required to evaluate the models, including the computation of the mAUC scores. Dani and Maximo ran the experiments on the ETMAE models using Google Cloud Virtual Machines with 4 NVIDIA T4 GPUs, while Manuel replicated the experiments provided in Xiao et al. (2023) for comparison using SLURM 8 NVIDIA A100 GPUs. Due to an unexpected technical problem, Manuel could not provide data for experiments on a variant of the ETMAE model we presented. The three of us collaborated on writing the report and designing the presentation slides, and we all reviewed and edited them. Dani created and uploaded the video presentation of our slides.

7 ACKNOWLEDGEMENT

This work is supported by Google Cloud Research Credits program. We also want to thank Professor Peter Belheumer and TA staff for continuous advice on the project.

REFERENCES

Yutong Bai, Zeyu Wang, Junfei Xiao, Chen Wei, Huiyu Wang, Alan L Yuille, Yuyin Zhou, and Cihang Xie. Masked autoencoders enable efficient knowledge distillers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24256–24265, 2023.

Shuhao Cao, Peng Xu, and David A Clifton. How to understand masked autoencoders. *arXiv preprint arXiv:2202.03670*, 2022.

Yabo Chen, Yuchen Liu, Dongsheng Jiang, Xiaopeng Zhang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Sdae: Self-distillated masked autoencoder. In *European conference on computer vision*, pp. 108–124. Springer, 2022.

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.

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*, pp. 16000–16009, 2022.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Wei Huang, Zhiliang Peng, Li Dong, Furu Wei, Jianbin Jiao, and Qixiang Ye. Generic-to-specific distillation of masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15996–16005, 2023.

Shanshan Lao, Guanglu Song, Boxiao Liu, Yu Liu, and Yujiu Yang. Masked autoencoders are stronger knowledge distillers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6384–6393, 2023.

Yang Luo, Zhineng Chen, Shengtian Zhou, and Xieping Gao. Self-distillation augmented masked autoencoders for histopathological image classification. *arXiv preprint arXiv:2203.16983*, 2022.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Michael Tschannen, Olivier Bachem, and Mario Lucic. Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast pretraining distillation for small vision transformers. In *European Conference on Computer Vision*, pp. 68–85. Springer, 2022.

Junfei Xiao, Yutong Bai, Alan Yuille, and Zongwei Zhou. Delving into masked autoencoders for multi-label thorax disease classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3588–3600, January 2023.