Pilot Study For Large-Scale Radiograph Pre-training

Niklas Bühler
Supervisor: Paul Hager

Technische Universität München, Germany
niklas.buehler@tum.de

Abstract. In recent years, medical datasets have expanded significantly, offering great potential for the development of machine learning applications in the medical field. However, manual labeling of such data is costly and poses a significant bottleneck to their utilization.
To address this issue, self-supervised learning (SSL) exploits the data itself to learn embeddings that can be quickly adapted to downstream tasks as needed.
In this work, we show the suitability of self-supervised learning techniques, specifically masked autoencoders (MAE), to generate such embeddings from a large clinical dataset comprising 12,000 radiograph images from various anatomical regions.
By pre-training a MAE model on producing these high-quality embeddings, the need for labeled data in downstream tasks is substantially reduced. This is evidenced by a linear classifier trained on representations from this MAE model achieving 84.58% top-1 accuracy on body-part classification when using only 1% of data, marking a 7% relative improvement over fully supervised training.
This pilot study thus establishes the foundation for applying the MAE strategy to our own large-scale real-world radiograph dataset, comprising 700,000 radiograph images, as well as evaluating on more complex downstream tasks in future work.

Keywords: X-ray · Vision Transformer · Masked autoencoder · Pre-training · Representation learning · Self-supervised learning · Scarce labels.

1 Introduction

The recent increase in large medical datasets presents a significant opportunity for exploitation by machine learning methods. However, this exploitation is hindered by the labor-intensive and costly process of manual labeling, a necessary step for training supervised learning models.

Self-supervised learning (SSL) emerges as a viable solution to this problem, especially within the domain of medical imaging, where the quantity of unlabeled

data typically far exceeds that of labeled subsets. SSL techniques can be further partitioned into contrastive and autoassociative methods.
Although contrastive learning frameworks like SimCLR [1] significantly surpassed previous self-supervised and semi-supervised learning benchmarks on ImageNet, we believe it is not a well-suited approach for learning representations of medical images, due to the core principle of contrastive learning. The contrastive learning technique extracts representations by contrasting positive and negative pairs of instances, which doesn't require models to capture intricate details that are typically present in medical images and oftentimes necessary for more complex downstream tasks.
On the other hand, autoassociative learning techniques, like masked autoencoders (MAEs) [3], are trained to reconstruct their own input data. They lever the inherent structure of the data to learn representations, and seem promising candidates for generating embeddings that extract the fine-grained features of medical images without explicit labels. These embeddings can then serve as a foundation for various medical downstream tasks, potentially reducing their dependency on large amounts of labeled data.

This study presents a pilot exploration of the efficacy of MAEs for pre-training on a large radiograph dataset spanning various anatomical regions. We examine the hypothesis that the MAE SSL strategy can significantly alleviate the labeling bottleneck in medical imaging by facilitating the generation of high-quality embeddings from unlabeled data. By employing MAE pre-training on a corpus of 12,000 radiograph images and subsequently testing the generated embeddings on a downstream body part classification task, we demonstrate the potential of this approach.
Our contributions are twofold: First, we establish the feasibility of MAE pre-training for generating useful embeddings from a substantial radiograph dataset, a critical step towards leveraging SSL in further large-scale medical image analysis. Second, we present preliminary evidence that these embeddings can significantly enhance model performance in downstream tasks with minimal labeled data, thereby paving the way for further exploration into the applicability of MAEs across more complex medical imaging challenges.

2 Related Work

2.1 Transformer Models

Attention Is All You Need [7] introduced the Transformer architecture as a novel alternative to traditional recurrent or convolutional neural networks (RNNs/CNNs). It relies solely on attention mechanisms, offering superior parallelizability and training efficiency compared to existing models. The architecture achieves state-of-the-art results in machine translation tasks and generalizes effectively to various applications.

An Image is Worth 16x16 Words [2] extended the Transformer architecture to computer vision with the Vision Transformer (ViT). This approach reshapes 2D images into sequences of flattened 2D patches, which are then fed into an architecture of alternating multi-headed self-attention and MLP blocks.

Masked Autoencoders are Scalable Vision Learners [3] proposed Masked Autoencoders (MAE) for image reconstruction, demonstrating their effectiveness in achieving state-of-the-art accuracy in image classification and transfer learning tasks. The asymmetric encoder-decoder architecture involves an encoder that operates only on unmasked patches and a lightweight decoder responsible for image reconstruction.

Due to the proven effectiveness of this class of model architectures across multiple domains, the Vision Transformer serves as one of our baseline models, and the Masked Autoencoder serves as the base for our pre-trained models.

2.2 Self-supervised Learning in Medical Imaging

A Simple Framework for Contrastive Learning of Visual Representations [1] introduced SimCLR, another approach to learning visual representations. In contrast to the MAE presented in [3], SimCLR uses a contrastive learning approach.

Self Pre-training with Masked Autoencoders for Medical Image Classification and Segmentation [9] applied Masked Autoencoders to the pre-training of Vision Transformers for medical image analysis. The MAE aggregates contextual information to infer masked image regions, enhancing the understanding of interdependencies among anatomical structures crucial in the medical image domain. The method involves pre-training a ViT on the same dataset as the downstream task and fine-tuning with task-specific heads. Experimental results demonstrate significant enhancements in medical image segmentation and classification performance compared to random initialization and traditional ImageNet pre-training methods. Notably, MAE self-pretraining shows promising performance even on small-scale medical datasets, surpassing existing approaches, including ImageNet-transfer learning.

Self-Supervised Learning Application on COVID-19 Chest X-ray Image Classification Using Masked Autoencoder [8] applied MAE for COVID-19 chest X-ray image classification, showcasing superior performance with a self-supervised approach. This method demonstrated remarkable efficiency even with limited labeled data, highlighting its potential in medical image analysis.

This work builds upon these previous works, leveraging insights from self-supervised learning, Transformer-based architectures, and their applications in medical imaging to develop an efficient and effective approach for large-scale radiograph analysis.

3 Method

3.1 Dataset and Preprocessing

This paper focuses on the IRMA radiograph dataset [5] and in this way serves as a pilot study for later employing and extending the presented methods on the larger MRI dataset (our own).

The IRMA Dataset is a compilation of anonymous radiographs, sourced from routine procedures at the Department of Diagnostic Radiology at RWTH Aachen University. The contained radiographs capture a diverse range of patient demographics, imaging views, and pathological conditions, resulting in significant variability in image quality. To facilitate uniform processing and analysis, all images within the dataset have been standardized to a resolution of 224×224. The dataset is distinguished by its extensive classification schema known as the IRMA code, which categorizes images into 193 distinct classes, spanning nine different body regions. The labeled subset of the dataset comprises 12,677 radiographs, each annotated with its corresponding IRMA code.

To prepare the IRMA data for subsequent deep learning tasks, we normalize pixel intensities using the 0-1 min-max scaling technique, ensuring consistency across all images.

The class distribution of the IRMA dataset is shown in figure 1. Due to very limited available data in the *whole body* class, we drop it from the dataset, leaving eight remaining classes.

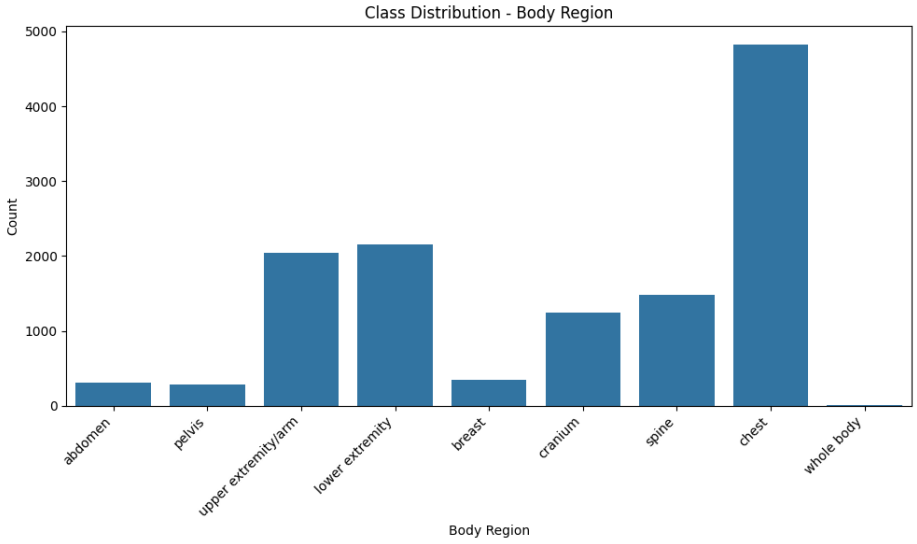
3.2 Models

We evaluate multiple deep learning models on the classification and pre-training tasks. For all base models, we use the standard implementations from their introducing papers:

- **ResNet50 Convolutional Neural Networks [4]:** Utilized for body part classification tasks due to their proven effectiveness in image classification.
- **Vision Transformers (ViT) [2]:** Investigated for both body part classification and masked autoencoder (MAE) pre-training tasks.
- **Masked Autoencoder (MAE) [3]:** Employed for self-supervised pre-training, enabling feature learning without explicit supervision.

For the linear probing of the pre-trained ViT MAE encoder, we explore different aggregation methods for the encoded output sequence, including calculating the mean values across the sequence, flattening it, and probing a learned CLS token.

Fig. 1. The class distribution of the IRMA dataset clearly shows the lack of available data in the *whole body* class, as well as a surplus of data for the *chest* class.



3.3 Training

We determine the learning rate for all model architectures by running a hyper-parameter search across ten different learning rates and optimizing for the best validation accuracy. The MAE model is tuned using the best validation loss. All models are trained until convergence, by employing early stopping with a patience of 10 epochs without improvement.

The weights of all base models (ResNet50, ViT and MAE) are initialized with their default pre-trained ImageNet weights.

We use a train-validation-test split of $(8000, 2000, 2600)$.

4 Results

We present the results of our experiments on the IRMA dataset, focusing on masked autoencoder pre-training and bodypart classification as an example downstream task.

4.1 Masked Autoencoder Pre-training

For the MAE pre-training task on the IRMA dataset, we evaluate the MSE loss of the MAE reconstructions compared to the input radiographs. However, the results of the pre-training should be judged by the quality of the resulting

embeddings. Therefore, the results of the produced embeddings in the ensuing downstream tasks are more informative on the success of the MAE pre-training than the direct MSE loss.

The MAE model trained for 36 epochs before stopping due to no further improvement on the validation loss. It achieved a final test MSE loss of 0.002475.

4.2 Full Data Regime: All Models Achieve Similarly High Accuracy

For the body part classification task, we evaluate the top-1 accuracy of our models on the IRMA dataset. Table 1 summarizes the results obtained from the different model and pre-training configurations.

When using the total available data for training the models, they all achieve an accuracy upwards of 96%, the ResNet50 even achieves 99.12%.

Table 1. Body Part Classification Results on the full IRMA dataset. Although all models achieve high accuracies, the ResNet50 model performed best.

Model	**IRMA Top-1 Accuracy**
ResNet50	**0.9912**
ViT	0.9685
ViT MAE Linear Probing (Flattened)	0.9604
ViT MAE Linear Probing (Mean)	0.9688
ViT MAE Linear Probing (CLS)	0.9692

4.3 Low Data Regime: MAE Outperforms Supervised ViT

In the low data regime of the body part classification task, we evaluate the top-1 accuracy of our models on the IRMA dataset, with a training set that is limited to only 1% of its original size, comprising only 80 images. Table 2 summarizes the results obtained from the different model and pre-training configurations.

While the ResNet50 model achieves a leading accuracy of 99.12% when trained on the full data, it produces less accurate predictions than all other models (44.08% vs. 79.08% and above), when trained on only 1% of the data. In contrast to this, the ViT is very data efficient, still achieving 79.08% with only 1% of the training data. However, the pre-trained MAE models with a linear classification layer on top outperform these two baseline models, with accuracies of 83%, 83.12% and 84.58% respectively. This five percentage point increase in accuracy shows the advantage of learning in an unsupervised manner over large datasets when minimal labels are available.

Table 2. Body Part Classification Results in Low Data Regime. The pre-trained models outperform the supervised models by five percentage points.

Model	IRMA Top-1 Accuracy (Low Data)
ResNet50	0.4408
ViT	0.7908
ViT MAE Linear Probing (Flattened)	0.8300
ViT MAE Linear Probing (Mean)	0.8312
ViT MAE Linear Probing (CLS)	**0.8458**

5 Further Research

While the IRMA dataset allowed for comparatively easy data handling and fast training runs, the MRI dataset is almost two magnitudes larger. Several new challenges arise due to the extensive nature of the MRI dataset, offering several new directions for further research to explore.

One such challenge revolves around processing images with high variability in sizes using the same model. While Transformer architectures are inherently well-suited for handling inputs of different sizes, training in batches requires intra-batch images to be of the same dimensions. Addressing the resizing or grouping of images into batches thus poses a significant research challenge.
To effectively utilize the variable-sized MRI dataset for deep learning tasks, custom batching strategies can be implemented and explored. Inspired by the VariViT paper [6], one approach is to extend their custom batching strategy (called *strict binning* here) by a *smart binning* strategy. With the original *strict binning* strategy, batches are sampled from all images of the same size. However, the exceptionally high variability of image sizes present in the MRI dataset forces this strategy to inevitably discard a lot of images as there aren't enough images of the same size to build a whole batch from, even with small batch sizes. The *smart binning* strategy thus defines a few image sizes as bins and assigns each image to the bin that matches its original size the closest. In order to sample batches, all images are resized to match their bin's shape, thus having compatible dimensions.
Another challenge is the high computational cost associated with training on the entire dataset. There are several ways to tackle this challenge. For instance, computational cost can be cut down by minimizing the amount of compute spent on uninformative black borders resulting from resizing images to a common resolution and aspect ratio. This optimization can be implemented by further optimizing the previously described binning strategy for batch processing with varying image resolutions or by employing a sophisticated masking strategy within the MAE to automatically mask or discard tokens corresponding to black borders.
A more basic but nonetheless important challenge is the speed at which data processing and loading are performed, considering the impracticality of storing the entire dataset locally. Exploring advanced offline data preprocessing as well

as loading and caching techniques can be further directions of research.

The scarce labels setting can also be a focus of further research, coming up with new ways to use or augment the few existing labels, especially in the context of medical imaging datasets. One possible approach is employing NLP methods for generating pseudolabels from medical reports typically accompanying the radiographs.
Finally, more complex downstream tasks are required in order to better evaluate the suitability of MAEs in the field of medical imaging.

References

1. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
3. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
5. Lehmann, T.M., Schubert, H., Keysers, D., Kohnen, M., Wein, B.B.: The irma code for unique classification of medical images. In: Medical Imaging 2003: PACS and Integrated Medical Information Systems: Design and Evaluation. vol. 5033, pp. 440–451. SPIE (2003)
6. Varma, A., Shit, S., Prabhakar, C., Scholz, D., Li, H.B., Rueckert, D., Wiestler, B., et al.: Varivit: A vision transformer for variable image sizes. In: Medical Imaging with Deep Learning (2024)
7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
8. Xing, X., Liang, G., Wang, C., Jacobs, N., Lin, A.L.: Self-supervised learning application on covid-19 chest x-ray image classification using masked autoencoder. Bioengineering **10**(8), 901 (2023)
9. Zhou, L., Liu, H., Bae, J., He, J., Samaras, D., Prasanna, P.: Self pre-training with masked autoencoders for medical image classification and segmentation. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–6. IEEE (2023)