CSE 891 Final Report: Model Explanation for Masked Autoencoders with Finetuning

Zhiying Li
Department of CSE
Michigan State University
East Lansing, MI 48823
xuezhiyu@msu.edu

Zhiyu Xue
Department of CSE
Michigan State University
East Lansing, MI 48823
xuezhiyu@msu.edu

Abstract

Masked autoencoders (MAE) are scalable vision learners [1], suggesting that self-supervised learning (SSL) in vision might embark on a similar trajectory as in NLP. The generative-based SSL model can perform extremely well on downstream tasks, for both vision (e.g. MAE) and language (e.g. BERT[2]). Although it can achieve promising few-shot/zero-shot performance for downstream tasks, few works focus on the interpretability and the inner decision process of large-scale pretrained models.

In this report, we first introduce the overall architecture of MAE, and review the existing methods for model explanation. Then, we apply the model explanation methods for MAE with a finetuned adapter via the downstream datasets. Our experimental results and analysis will be presented in the last section.

1 Introcution

Masked image modeling (MIM) has been recognized as a strong and popular self-supervised pre-training approach in varied vision tasks, due to its promising ability to transfer unbiased representation space. However, the interpretability of the mechanism and properties of the learned representations by such a scheme is so far not well-explored. In our report, we use several model-agnostic model explanation methods to reveal the interpretability of MAE when applying it to the downstream task. By using the model explanation methods, not only we can check the stander performance (e.g. Accuracy) of finetuned MAE, but also the reliability. The reliable ViT-based model [3] must capture the semantic patches, otherwise, it would be attacked easily and caused biased decisions in the real-world application.

Our experimental contributions can be concluded as follows:

- Review the papers of MAE and existing model explanation methods.
- Implement MAE and finetune it on the downstream tasks.
- Implement several model explanation methods (e.g. saliency and deeplift) and use them on MAE, to visualize the informative patches used in the decision process of ViT.
- Apply model explanation methods for MAE on both normal data and medical data, achieving the diversity of downstream tasks. Surprisingly, we find MAE with finetuned adapted can achieve better interpretability on medical images than normal images.

The source codes of our experiments were released at `https://colab.research.google.com/drive/1BZusNV-90M-9utI7WVW3tjsoRDkI_2Vq?usp=sharing` as a Colab template.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

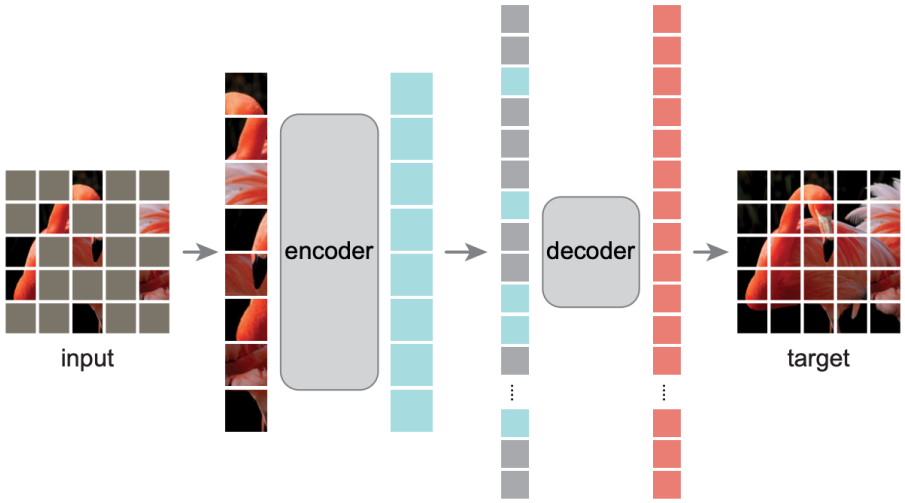

Figure 1: The MAE Architecture (retrieved from [1])

2 Review: Mask Autoencoder (MAE)

By proposing a new schema to enable self-supervised learning in field of computer vision, the Masked Autoencoders (MAE)[1] address the needs for even greater amount of labeled images for training. In the closed related field of Natural Language Processing (NLP), BERT [2] was proposed as a self-supervised learning schema to solve similar issues. By randomly masking out a percentage of input words and asking the BERT model to predict the missing word, it enables a generalizable models for NLP tasks. To borrow such idea for vision autoencoding, MAE address three differences between vision and language tasks.

1. **Architecture**. With the advancement of Vision Transformer (ViT)[3], MAE uses ViT as the architecture instead of commonly used convolutional network, with which mask tokens and positioinal enbedding doesn't work well.
2. **Information Density**. To randomly mask a high portion of the image patches, it reduces the spatial redundancy and increases the information density.
3. **Role of decoder**. Unlike the BERT, the decoder design of MAE is important in determine the semantic level of the learned latent representation.

In terms of the overall performance, the original work of MAE [1] reports threefold ($3\times$) acceleration of the pretraining using 75% masking on the original image. And they have also demonstrated that ViT-Huge model achieves 87.8 % accuracy from ImageNet-1K data using their methods. This implies the potential of Mask Autoencoder as a scalable model to train large model more effectively.

2.1 MAE Architecture

As is shown in Figure 1, at high-level, MAE randomly masks a portion of patches from the input image and reconstruct the image's missing patches. They used asymmetric design: the encoder encodes the latent representation from partial patches, and the decoder takes mask token to reconstruct the latent representation into the full image. The loss function for the implementation is mean squared error (MSE). The details of masks, encoder, and decoder is introduced as follow:

- **Mask.** This step randomly selects a high portion of subsets of equally devided patches as mask. The random process is uniformly distributed. The high portion is to guarantee information density to avoid trivial cases when the task can be solved by extrapolation.

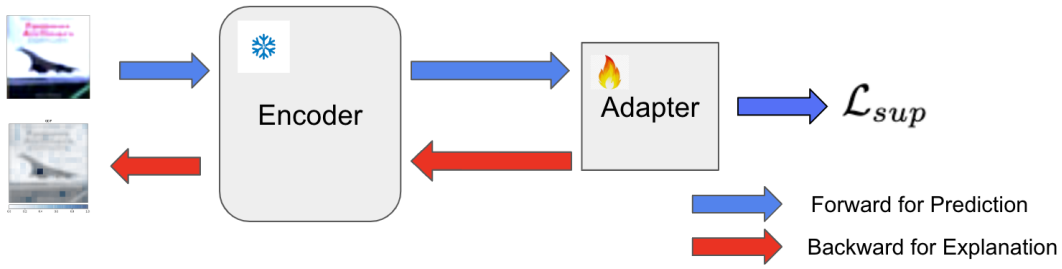

Figure 2: The Architecture for Finetune.

- **Encoder.** This step takes as input the linear projection of non-masked patches and the positioning embedding of each patch. The encoder is ViT. It is because the usage of only non-masked patches that greatly reduce the number of computation and memory usage of larger encoders.
- **Decoder.** The decoder take the encoded latent representation and mask token to reconstruct the image. The mask token can help to determined the position of masked patches with respect to the input patches. One key point is that, given full set of mask tokens, the decoder can be independent with encoder design and can be lightweight as compared with encoder (10% is reported in the original paper).

3 Review: Methods for Model Explanation

For this project, we have selected several model explanation methods: Saliency [4], GradCAM [5], and DeepLift [6, 7].

3.1 Saliency

Saliency is one of the input gradient-based methods to explain input's attribution. It is one of the most fundamental model explination methods. The key idea is to determine the importance of the pixels that matters the most in the ConvNet.

In terms of equation, this can be formalized as directional derivative on a trained network. Suppose that x is input, c is the class, and f is the ConvNet. Salency is characterized by vanilla input gradient:

$$\mathcal{I}(x; c) = |\nabla_x f_c(x)|\,,$$

The difference between saliency and backpropagation is that saliency maps are calculated after a network has finished training, whereas backpropagation is for training. And saliency maps take absolute values of the gradient so as to query the spatial support of the input image.

3.2 GradCAM

GradCAM is one of class-discriminate input attribution map of the explanation method. The prior work CAM (class activation map) is the average of weighted activation maps of the final convolutional layer. It can be used to visualize sensitive region of input patterns as heatmap that is used to make certain prediction.

GradCam is the generalized the version of CAM with less trade-off of model complexity and performance. Compared to CAM, GradCAM specifies importance weights of activation maps by input gradient α_k^c.

Suppose there are k feature map, and one of them is denoted as $A^k \in \mathbb{R}^{u \times v}$. The goal is to determine how the pixel (i, j) of the initial image contribute into the final feature map A^k. GradCAM first computes α_k^c the average gradient score of the class c with respect to the k-th feature map. It's pixel-wise average. i.e.

$$\alpha_k^c = \frac{1}{uv} \sum_i \sum_j \frac{\partial y^c}{\partial A_{i,j}^k}$$

Therefore, $\mathcal{I}(x; c)$ given input x and class c in terms of GradCAM is given as:

$$\mathcal{I}(x; c) = \sum_k \alpha_k^c A^k$$

3.3 DeepLift

DeepLIFT solves the relative feature importance problem of other other explanation method by comparing the activation of each neuron to its 'reference activation'. For DeepLIFT, it characterizes the change in terms of difference with the baseline, instead of the gradient, since using the gradient may not work well for nonlinear models, such as ReLU.

For each layer is calculated in terms of slope rather than gradient. Scores can be computed efficiently in this way. And the feature importance of x_i in $\mathbf{x}$ with respect to the output Y is given as:

$$\left(x_i - x_i^{\text{baseline}}\right) \frac{\Delta_i Y}{\Delta_i x}$$

4 MAE Fine-tuning

Since MAE is for pre-training of ViTs, the encoder has already produced image representations for decoder to reconstruct target image. Therefore, one can use the encoder for fine-tuning. In the original paper [1], by plugging the weights into the various downstream tasks (COCO object detection and segmentation, ADE20K semantic segmentation, and image classification), we can all see different level of improvement as compared to the baseline methods. For COCO object detection and segmentation, MAE improves by 4% than supervised pretraining. For ADE20K semantic segmentation, MAE improves by 3.7% over supervised pretraining. For classification, MAE improves by 0.8% 8% for iNaturalists and Places dataset.

For this report, we select the downstream tasks of MedMNIST on the dataset of CIFAR10 We will apply model explanation to show the model interpretability and the inner decision process. As Fig 2, we fixed the encoder of pretrained MAE as the backbone of the feature extractor, and only finetune a simple adapter. During the explanation process, we backward the gradients of whole model combing the encoder and adapter.

During finetuning, we compute the representation of the inputted image as q^0 as the class token of $q = f_{enc}(z)$, and the logit of classification as $p = Softmax(Wq^0)$. Note that p_c denotes the output probability for class c. Details of classification loss $\mathcal{L}_{sup}$ can be shown in

$$\mathcal{L}_{sup} = -\frac{1}{|X_{lab}|} \sum_{x \in X_{lab}} y_c log(p_c) \tag{1}$$

As illustrated in Fig 2, during finetuning, we fixed the encoder of MAE as the feature extractor, since MAE is a transferable learner constructing liner separable feature space for downstream tasks [8]. More details will be provided in the experimental results section.

5 Experiments

5.1 Experimental Settings

Datasets. We finetune pretrained MAE on normal image classification task and medical image classification task, respectively. We choose CIFAR10 as the normal image classification task, while MedMNIST as the medical image classification task. These two datasets are introduced as follows:

- CIFAR10 [9]. The CIFAR-10 dataset consists of 60000 color images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images. In our experiment, considering the limitation of computing resources, we only use 1000 images for training data.
- MedMNIST [10, 11]. A large-scale MNIST-like collection of standardized biomedical images, including 12 datasets for 2D. We selected Dermanist and more results of other benchmarks of MedMNIST will be reported in the Appendix. As we mentioned above, we only use 1000 images for training data considering the limitation of computing resources.

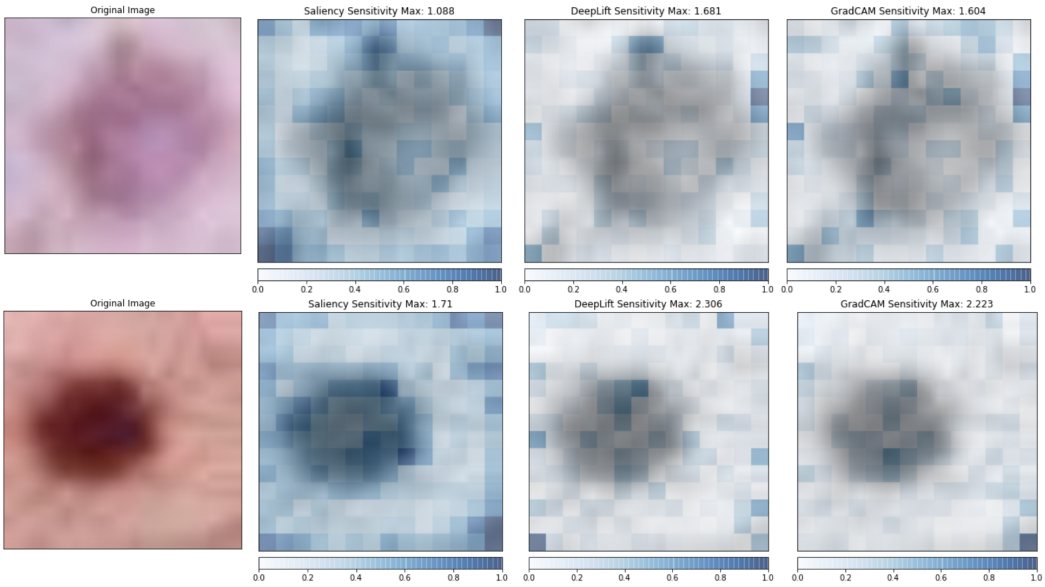

Figure 3: Visualization results for DermaMNIST. The max explanation sensitivity scores were shown on the title of each explanation figure.

Metric. We use the explanation sensitivity based on adversarial perturbations [12] as the metric of our explanation methods. Explanation sensitivity measures the extent of explanation change when the input is slightly perturbed. It has been shown that the models that have high explanation sensitivity are prone to adversarial attacks.

Finetune. We use a two-layer MLP as the adapter and finetune it over the training set of downstream tasks for 20 epochs. We use Adam [13] as the optimizer with learning rate equal to 0.1. The training curves of these two different will be shown in the Appendix.

5.2 Results & Analysis

The experimental results of DermaMNIST and CIFAR 10 are shown in Fig 3 and Fig 4 with the explanation sensitivity scores reported on them.

We found that, for both medical image dataset and normal image dataset. Our structure, where a generative-based pretrained ViT with a finetuned adapter can achieve promising results of explanation. Also, we find in most cases, DeepLift is more likely to locate at the sematic patches compared to the GradCAM and Saliency.

The pretrained MAE we used in our experimented was released by Meta, where it pretrained on the large amount of unlabelled images collected from the social media apps. Intuitively, the finetuned model should achieve better interpretability on normal images (CIFAR10) rather than medical images (MedMNIST), since the domain gap between pretrained dataset and CIFAR10 is smaller than the one between MedMNIST. However, the results is contrary to our assumption. We conclude the potential reasons:

- The medical images are much cleaner than the normal images. Medical images do not contain noises in the background, while normal images do (e.g. A dog image without the head).
- The main objectives of medical images are usually more significant compared to normal images. A normal image may contain several different objects (e.g. A image contain a dog and a cat).

References

[1] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*

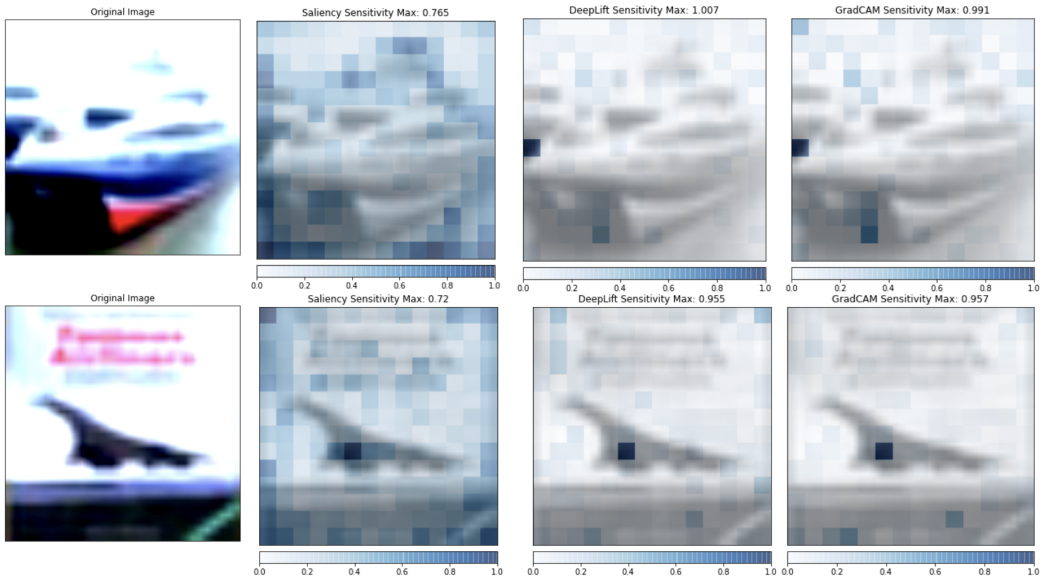

Figure 4: Visualization results for CIFAR10. The max explanation sensitivity scores were shown on the title of each explanation figure.

Recognition (CVPR), pp. 16000–16009, June 2022.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[3] A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner, and X. Zhai, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[4] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps.," *CoRR*, vol. abs/1312.6034, 2013.

[5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.

[6] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, p. 3145–3153, JMLR.org, 2017.

[7] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," in *International Conference on Learning Representations*, 2018.

[8] K. Zhang and Z. Shen, "i-mae: Are latent representations in masked autoencoders linearly separable?," *arXiv preprint arXiv:2210.11470*, 2022.

[9] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.

[10] J. Yang, R. Shi, and B. Ni, "Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis," in *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 191–195, 2021.

[11] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, "Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification," *arXiv preprint arXiv:2110.14795*, 2021.

[12] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 3681–3688, 2019.

[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

A Appendix

A.1 Visualization Case of MAE

As Fig 5 illustrated, we provide a visualization case of our pertained MAE, to show it can achieve promising zero-shot reconstruction results on real-world images.

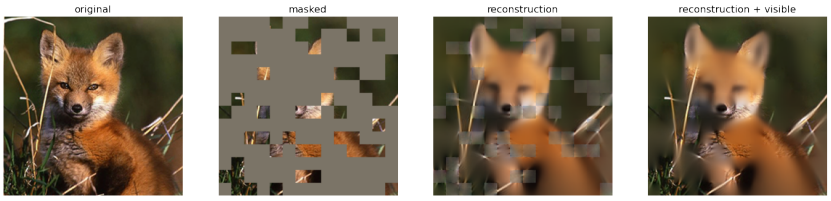

Figure 5: Visualization Case of MAE for the reconstruction performance.

B Learning Curves for Finetuning

The learning curves of finetuning on CIFAR10 and MedMNIST (DermaMNIST, BooldMNIST and RetinaMNIST) were shown in Fig 6.

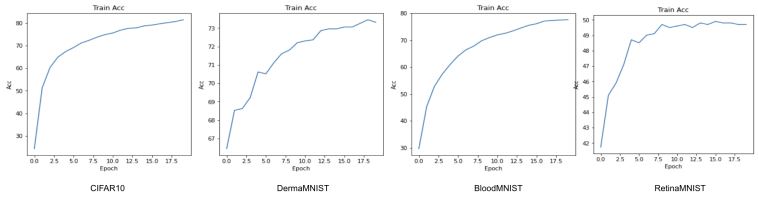

Figure 6: Learning curves of finetuning

B.1 More Visualization Results

We provide more visualization results for other datasets in MedMNIST shown in Fig 7.

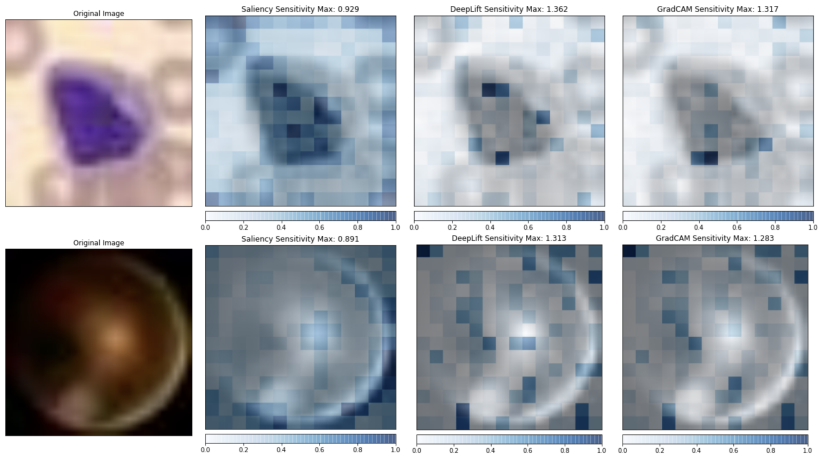

Figure 7: Visulization results for MedMNIST. BooldMNIST (Top) and RetinaMNIST (Bottom)