Mask Models are Token Level Contrastive Learners

Zixuan Pan [1]

Abstract

In recent years, the field of self-supervised learning has seen a surge in the development of mask models, which have been demonstrated to have strong performance on downstream tasks and efficient training. To better understand the underlying mechanism behind these models' success, we propose a theoretical framework for mask modeling. By treating mask modeling as a low-rank recovery task, we demonstrate that it is a parametric version of Spectral Clustering and the reconstruction loss conforms to the form of Spectral Contrastive loss. This means that mask modeling can be understood as a token level Contrastive Learning. Our framework can be used to explain why optimal masking ratios vary among modalities and why there is a large gap between linear probing and finetuning performance for mask models. Additionally, our analysis suggests that the success of mask models depends on the model architecture, where a token mixing layer and layer normalization are crucial for the success of mask models. Our framework has the potential to be a step stone for future algorithm and network architecture design in the field of self-supervised learning.

1. Introduction

With the rapid-growth of deep learning and its increasing demand for data, self-supervised learning arises as a research topic in-demand. Among the successful self-supervised learning models, mask models have received significant attention for their strong downstream performance and efficient training (He et al., 2022; Xie et al., 2021; Devlin et al., 2018; Bao et al., 2022; Liu et al., 2019; Tong et al., 2022; Huang et al., 2022; Bachmann et al., 2022). However, mask-modelling has long been regarded as an engineering trick, and its underlying mechanism remains poorly understood.

[1]Department of Computer Science, University of Michigan, Ann Arbor, the United States. Correspondence to: Zixuan Pan <zxp@umich.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Empirically, we have observed that Mask Image Modelling (MIM) leads to varying performance improvements on different downstream tasks compared to previous baselines (He et al., 2022). Specifically, MIM has been found to perform better on fine-grained tasks such as semantic segmentation and object detection, compared to classification tasks. This phenomenon leads us to hypothesize that the representation learned by MIM models is fundamentally similar to that of image segmentation (clustering).

In this work, we propose a theoretical analysis of mask modelling by treating it as a low-rank recovery (LRR) task. Our analysis further demonstrates that the reconstruction loss can be rewritten as a Contrastive loss (Davenport & Romberg, 2016).

The LRR problem aims to find the low-rank approximation of a given matrix, and has been used as a method for subspace clustering (Patel et al., 2015). Additionally, as the optimal solution of the LRR problem is a combination of leading eigenvectors, we are naturally led to Spectral Clustering, which also utilizes leading eigenvectors (Shi & Malik, 2000; Ng et al., 2001). Our results show that MIM approximates the Spectral Clustering features of an image-related graph, where each node represents a patch of the image.

By viewing the Masked Image Model (MIM) as a parametric version of Spectral Clustering, we can rewrite the reconstruction loss of mask models in the form of Spectral Contrastive loss on the token level (HaoChen et al., 2021).. This allows MIM to be viewed as a token-wise Contrastive Learning method, which attracts similar patches while repelling dissimilar ones, resulting in smaller distances within clusters and larger distances between clusters. However, there are some key differences between mask models and traditional Contrastive Learning methods. Specifically, mask models operate on the token level, whereas traditional Contrastive Learning methods focus on the global feature of the entire input, and in mask models, positive samples are not clearly defined, but are "randomly sampled" based on the similarity between Spectral Clustering features.

Based on the formulation, we could answer several concerning questions about mask models: 1) Why optimal masking ratio vary among modalities? 2) Why is there a large gap between linear probing and finetuning performance for mask

models? 3) Does mask modelling rely on network architectures?

For the first question, we argue that a critical factor that affects the goodness of pretrained features is the number of clusters in Spectral Clustering. For example, if we have an image with a dog on the grass, intuitively we should have two clusters: grass and dog. It could be less representative if we have more clusters and divide one of the existing clusters into different sub-clusters and repel one from each other. The number of clusters is given by the number of leading eigenvectors, which is related to the rank of reconstructed matrix in LRR problem and masking ratio in MIMs. This explains why we need different masking ratios in different modalities (Devlin et al., 2018; He et al., 2022; Tong et al., 2022; Huang et al., 2022).

For the second question, it is due to the nature of token level Contrastive learning. Pretrained mask models learns to divide tokens into clusters, but doesn't always learn which cluster is most related to the class. Therefore, token mixing layers are needed to "select" clusters. Most MIMs apply an extra BatchNorm layer when performing linear probing, otherwise a huge accuracy drop is witnessed (Ioffe & Szegedy, 2015; He et al., 2022). It could be due to the lack of patch selection and a BatchNorm is needed to add non-linearity. In contrast, we found that partially finetuning one linear layer for row mixing with the prediction head could much improve classification accuracy.

For the third question, the answer is "Yes". Model architectures containing token mixing layers plays a crucial role in the success of mask models in classification tasks (Vaswani et al., 2017; Dosovitskiy et al., 2021; Liu et al., 2022; Tolstikhin et al., 2021). Finetuning these layers allows the model to learn how to select tokens. Meanwhile, the layer normalization in the decoder might also be important, as it serves as a token level batch normalization, which is commonly used in the projection layer of Contrastive Learning models to improve performance (Ba et al., 2016; Ioffe & Szegedy, 2015). Therefore, we conclude that mask model is dependant of network architecture.

In a summary, our main contributions are:

1. We created a mathematical framework for mask image modeling by viewing it as a low-rank recovery problem.
2. We found that mask model could be viewed as a token level Contrastive Learning, which could account for its good performance on downstream tasks.
3. Our analysis framework could explain several important behaviors of mask models.

We mainly conducted experiments on images, but our findings could be easily generalized to all modalities.

2. Related Works

2.1. Mask Image Modelling

The recent trend in self-supervised learning is to train vision transformers using masked images to reconstruct the original ones. (Dosovitskiy et al., 2021). RDifferent types of reconstruction objectives, such as token-wise, feature-wise, and pixel-wise reconstruction, are being tested. (Zhou et al., 2022; Chen et al., 2022; 2020). These kinds of pretraining tasks are called Masked Image Modeling (MIM) (Bao et al., 2022). There are two main architectures for these models: one that only accesses visible tokens in the encoder and attaches an extra decoder (He et al., 2022; Bao et al., 2022), and another that passes both visible and mask tokens into the encoder and has a single linear layer as a decoder (Xie et al., 2021). Our formulation is based on the first type of architecture. These mask models serve as a pretrain model, and for downstream tasks, we either finetune or perform linear probing. For classification, a linear head is appended, and the parameters are initialized from the pretrain models. The difference between finetuning and linear probing is that the parameters of the pretrained model are frozen in linear probing.

2.2. Theoretical Analysis of Mask Models

Previous works on mask models have provided theoretical frameworks for understanding the attention operation in the encoder (Cao et al., 2022), proposed that MIMs are learning semantics (Pan et al., 2022), and claimed that mask models learn global features that are occlusion invariant (Kong & Zhang, 2022). Our work is distinct from these previous works in that we emphasize the connection between mask modeling and Contrastive Learning. One work also mentioned that the decoder in MIMs is performing low-rank recovery, but the authors did not link this to the success of MIMs (Cao et al., 2022).

2.3. Spectral Contrastive Loss

The Spectral Contrastive loss was proposed as a way to provide a provable guarantee for downstream task performance (HaoChen et al., 2021; Arora et al., 2019). However, some later work has identified issues with the formulation and stronger assumptions are needed to achieve the guarantee (Saunshi et al., 2022). Despite this, the theoretical framework that connects Contrastive Learning and Spectral Clustering is still attractive. Our work is inspired by this analysis framework, but with several differences. In their work, the graph used for Spectral Clustering is inherent, and the authors argue that matrix factorization approximates the node representations of the graph. Our work, instead,

explicitly writes out the adjacency matrix of a graph and shows that it is related to the MIM problem. Additionally, our work highlights the importance of rank, which is often overlooked in previous works.

3. Preliminary and Notations

3.1. Notations of Masked Autoencoder

Our analysis mainly focus on Masked Autoencoder (MAE) style encoder-decoder structure, where the input size of encoder is smaller than that of decoder (He et al., 2022). Denote the encoder in the mask modeling by f, and the decoder by g, the sampled visible subset by X, and the unmasked part of the original image by X_0. We adopt the Transformer architecture as backbone, where f and g don't change the shape of inputs (Vaswani et al., 2017).

Definition 3.1. To train the masked autoencoder and achieve the best performance can be interpreted as solving the minimization problem:

$$\underset{\boldsymbol{f},\boldsymbol{g},\boldsymbol{X}}{\operatorname{argmin}} \|g \circ f(X) - X_0\|_F^2, \tag{1}$$

where $X \in \mathbb{R}^{N \times F}$, $X \in \mathbb{R}^{N_0 \times F}$. We reshape the matrix of image so that N is the number of visible patches and N_0 is the number of masked patches. We also have the loss defined as

$$\mathcal{L}_{MAE}(f, g, X) = \|g \circ f(X) - X_0\|_F^2 \tag{2}$$

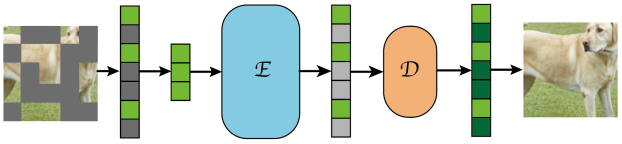

Figure 1. **Overall structure of Mask Autoencoders.** Only visible tokens are passed through the encoder, and a decoder applies encoder features to reconstruct mask tokens.

3.2. Basic Low-rank Recovery Problem

The basic low-rank recovery problem solves the following optimization problem:

$$\underset{\widehat{\boldsymbol{D}}}{\operatorname{argmin}} \ \|\widehat{D} - D\|_{\mathrm{F}} \ \text{ subject to } \ \operatorname{rank}(\widehat{D}) \leq \operatorname{rank}(D) \tag{3}$$

Based on the Eckart–Young–Mirsky theorem (Horn & Johnson, 1985), the low-rank approximation problem has a solution in terms of singular value decomposition of the original matrix, which is in the form: $\widehat{D} = \sum_{i=1}^{r} \sigma_i u_i v_i^T$, where σ_i is the i^{th} singular and u_i and v_i are its corresponding left and right singular vectors. We could also write it as $\widehat{D} = U_r \Sigma_r V_r^T$, where U_r, V_r contains the first r columns of U and V, and Σ_r is an $r \times r$ matrix with the top r leading singular values as diagonal.

3.3. Spectral Clustering with Normalized Adjacency Matrix

Suppose we have a n-node graph G with the adjacency matrix A:

$$A = \begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nn} \end{bmatrix}$$

The normalized adjacency matrix is defined as $\mathscr{A} = D^{-\frac{1}{2}} \mathrm{A} D^{-\frac{1}{2}}$, where D is the degree matrix of graph G, which is a diagonal matrix such that $D_{ii} = \sum_{j=1}^{n} w_{ij} = w_i$. To get k cluster of the graph, take the leading top k eigenvectors of $\mathscr{A}$ as node embedding and perform k-means algorithm on the node embedding (Ng et al., 2001).

4. Mask Modeling as a Patch-wise Contrastive Learning

4.1. Mask Modeling is Low Rank Recovery

In this section, we formalize MAE as a low-rank recovery task. As X is a smaller portion of the original image and f doesn't change the size of input, $f(X)$ naturally has a lower rank compared to X_0. we assume the following condition is true:

Assumption 4.1. $g \circ f(X)$ has a lower rank than X_0.

Remark 4.2. In practice, even if g is a non-linear function that doesn't guarantee low rank assumption, we find that $g(f(X))$ still has a very low rank. Arguably it is because reconstructing unseen tokens is very hard to optimize and only leading singular vectors can be approximated

Under this assumption, the minimization problem can be rewritten as

$$\underset{\boldsymbol{f},\boldsymbol{g},\boldsymbol{X}}{\operatorname{argmin}} \|g \circ f(X) - X_0\|_F^2 \tag{4}$$
$$\text{subject to } \ \operatorname{rank}(g \circ f) < \operatorname{rank}(X_0).$$

4.2. Mask Modeling is a Parametric Version of Spectral Clustering

In section 3.2, it is showed that the low-rank approximation problem is solved by singular value decomposition of the higher-ranked matrix. Suppose the required rank is k, then the optimal solution is a linear combination of top k eigenvectors of $X_0 X_0^T$ obtained from singular value decomposition.

Consider Spectral Clustering which clusters a graph into k connected components such that there is minimal effect on graph Laplacian. Spectral clustering performs dimensional reduction with k eigenvectors corresponding with the largest k eigenvalues of the normalized adjacency matrix.

Both mask modeling and Spectral Clustering are utilize k eigenvectors, hence, we propose that the behaviors of mask modeling is similar to the behaviors of Spectral Clustering. Consequently, the classifier trained based on mask modeling based f and Spectral Clustering based f gives the same prediction. Formally we have:

Theorem 4.3. *Define weights of adjacency matrix for graph G as $w_{ij} = \langle X_{0r,i}, X_{0r,j} \rangle$, where $X_{0r,i}$ is the low-rank approximation of the representation for ith patch of X_0. Given the corresponding normalized adjacency matrix $\mathscr{A}$, optimizing mask modeling is equivalent to optimize the following loss on classification downstream tasks.*

$$\mathcal{L}_{spec}(f, g, X) = \left\| (g \circ f(X))(g \circ f(X))^\top - \mathscr{A} \right\|_F^2 \tag{5}$$

Proof. The SVD of X_0 gives $X_0 = U\Sigma V^T$, then $A = X_{0r}X_{0r}^T = U_r\Sigma_r^2 U_r^T$.

Since A is symmetric and D is diagonal, $\mathscr{A}$ is symmetric and SVD of $\mathscr{A}$ has the form of $U_{\mathscr{A}}\Sigma_{\mathscr{A}}U_{\mathscr{A}}^T$. Plug in A gives

$$\begin{aligned}\mathscr{A} &= D^{-\frac{1}{2}}AD^{-\frac{1}{2}} \\ &= D^{-\frac{1}{2}}U_r\Sigma_r^2U_r^\top D^{-\frac{1}{2}} \\ &= \left(D^{-\frac{1}{2}}U_rD^{\frac{1}{2}}\right)\left(D^{-\frac{1}{2}}\Sigma_r^2D^{-\frac{1}{2}}\right)\left(D^{-\frac{1}{2}}U_rD^{\frac{1}{2}}\right)^T.\end{aligned}$$

Therefore, $U_{\mathscr{A}} = D^{-\frac{1}{2}}U_rD^{\frac{1}{2}}$ and $\Sigma_{\mathscr{A}} = D^{-\frac{1}{2}}\Sigma_r^2D^{-\frac{1}{2}}$. The SVD of X_0 can be rewritten as $X_{0r} = D^{\frac{1}{2}}U_{\mathscr{A}}D^{-\frac{1}{2}}\Sigma_r V^T$. With Eckart–Young–Mirsky Theorem, we rewrite the minimization problem of mask modeling as

$$\underset{\boldsymbol{f},\boldsymbol{g},\boldsymbol{X}}{\operatorname{argmin}} \left\| g \circ f(X) - D^{\frac{1}{2}}U_{\mathscr{A}}D^{-\frac{1}{2}}\Sigma V^T \right\|_F^2,$$

whose optimal solution is $D^{\frac{1}{2}}U_{\mathscr{A}}D^{-\frac{1}{2}}\Sigma V^T$. Note that D is a diagonal matrix, so we could find B, such that $D^{\frac{1}{2}}U_{\mathscr{A}}D^{-\frac{1}{2}}\Sigma V^T B = U_{\mathscr{A}}D^{-\frac{1}{2}}\Sigma V^T$. Therefore, we can discard $D^{\frac{1}{2}}$, making the optimization problem into:

$$\underset{\boldsymbol{f},\boldsymbol{g},\boldsymbol{X}}{\operatorname{argmin}} \left\| g \circ f(X) - U_{\mathscr{A}}D^{-\frac{1}{2}}\Sigma V^T \right\|_F^2,$$

Multiply $g \circ f(X)$ and $U_{\mathscr{A}}D^{-\frac{1}{2}}\Sigma V^T$ with their transpose and make an optimization problem, we finish the proof. □

Therefore, MAE learns to approximate the Spectral Clustering features. We further discuss the importance of having appropriate k in Section 5.1

4.3. L_{spec} is a Spectral Contrastive Loss

Rewrite L_{spec}, mask modeling can be viewed as a token level Contrastive Learning. We define the i^{th} row of $g \circ f(X)$ as $\sqrt{w_i}u_i$, the predicted patch representation.

We could rewrite L_{spec} into,

$$\begin{aligned}\mathcal{L}_{\text{spec}} &= \left\| (g \circ f(X))(g \circ f(X))^\top - \mathscr{A} \right\|_F^2 \\ &= \left\| (g \circ f(X))(g \circ f(X))^\top - D^{-\frac{1}{2}}AD^{-\frac{1}{2}} \right\|_F^2 \\ &= \sum_{i,j} \left(\frac{w_{ij}}{\sqrt{w_iw_j}} - (\sqrt{w_i}u_i)^\top(\sqrt{w_j}u_j) \right)^2 \\ &= \sum_{i,j} \left(\frac{w_{ij}^2}{w_iw_j} - 2w_{ij}u_i^\top u_j + w_iw_j \cdot \left(u_i^\top u_j\right)^2 \right)\end{aligned} \tag{6}$$

Apply the kernel trick, changing w_{ij} into W_{ij}, such that $W_{ij} = \exp(\frac{w_{ij}}{2\sigma^2})$ (Hofmann et al., 2008). With a choice of σ, we have W_{ij} defined as (or approximates) the probability of u_i and u_j to be a positive pair. Following the notations of Spectral Contrastive Loss, we make Equation (6) into the form of a Contrastive loss (HaoChen et al., 2021).

$$\mathcal{L}_{\text{spec}} = \mathcal{L}_{\text{cont}} + \text{ const}, \tag{7}$$

where $\mathcal{L}_{\text{cont}} = -2 \cdot \mathbb{E}_{u,u^+}\left[u^\top u^+\right] + \mathbb{E}_{u,u^-}\left[\left(u^\top u^-\right)^2\right]$

The above shows that MAE loss is equivalent to a Contrastive loss on masked tokens. We further show that it inherently perform Contrastive Learning on visible tokens.

To simplify the model, we make the following assumption:

Assumption 4.4. Denote one of the original patch of the i^{th} masked token as $X_{0,i}$, the predicted feature u_i is a linear combination of features of visible tokens, such that $u_i = \sum_j a_j u'_j$. Assume this transformation is made by decoder g.

Lemma 4.5. *Optimal a_j is positively correlated with patch similarity $X_{0,i}^T X_{0,j}$.*

Proof. Consider the gradient flow from MAE loss to a_k passing u_i.

$$\begin{aligned}\frac{\partial \mathcal{L}_{MAE}}{\partial a_k} &= \frac{\partial \mathcal{L}_{MAE}}{\partial u_i}\frac{\partial u_i}{\partial a_k} \\ &= 2(u_i - X_{0,i})^\top u'_k \\ &= 2(\sum_j a_j u'_j - X_{0,i})^\top u'_k\end{aligned}$$

$\frac{\partial \mathcal{L}_{MAE}}{\partial a_k} = 0$ when $a_k u_k'^\top u'_k = (X_{0,i} - \sum_{j \neq k} a_j u'_j)^\top u'_k$.

We assume u' is uniformly distributed, then

$$\sum_{j \neq k} a_j u'_j \approx C(1 - a_k),$$

where C is a constant that is the same for all patches. Hereby we obtain the optimal $\hat{a_k}$

$$\hat{a_k} = \frac{X_{0,i}^\top u'_k - C}{C + \|u'_k\|^2}$$

Meanwhile, as there's only one transformation between u'_k and $X_{0,k}$, we view u'_k as an approximation to $X_{0,k}$:

$$\hat{a_k} \approx \frac{X_{0,i}^\top X_{0,k} - C}{C + \|X_{0,k}\|^2}.$$

It indicates that $\hat{a_k}$ is higher regarding u_i when the original patches are similar to each other.

□

As the representation of masked tokens is mainly composed of similar tokens, performing Contrastive Learning on masked tokens inherently performs Contrastive Learning on visible tokens.

Remark 4.6. We could view the layer normalization layer in MAE's decoders as a token level batch normalization, and the entire decoder as a non-linear projection layer in Contrastive Learning methods (Ba et al., 2016).

Remark 4.7. Though reconstructing masked tokens make it more complicated and less explainable, it is required as reconstruction visible tokens could lead to a shortcut solution of identity mapping.

5. Patch-wise Contrastive Learning Explains Mask Model Behaviors

Based on the theoretical framework proposed, we could explain several parameter choice and architecture design for mask models.

5.1. Mask Ratio for Different Modalities

In Section 4.2, we demonstrate that mask models are a parametric version of Spectral Clustering, and they learn to decrease intra-cluster distances while increasing distances between different clusters through Contrastive Learning. Therefore, an appropriate number of clusters is a crucial factor that affects the quality of the features learned. When we consider each cluster has a pseudo-class label, too few or too many classes can both be indistinct when trying to separate the classes. Therefore, we define the following:

Definition 5.1. Let s be the ratio of appropriate cluster numbers to total number of tokens. We have

$$s = \frac{num\_cluster}{num\_tokens} = \frac{k}{N_0}, \tag{8}$$

where k is the number of leading eigenvectors in Spectral Clustering, and N_0 is the number of tokens for reconstruction.

In mask models, k is subjected to $rank(g \circ f(X))$, which is determined by the number of visible tokens N. If we assume $rank(g \circ f(X))$ is proportional to N, we have:

$$s \propto \frac{N}{N_0} \tag{9}$$

As $\frac{N}{N_0} = 1 - mask\_ratio$, s is thus determined by the masking ratio.

Intuitively we know that s is smaller for modalities with lower information density, such as video, vice versa. Therefore, we need a higher masking ratio for lower-density modalities and a smaller one for higher-information-density modalities (Huang et al., 2022; Wettig et al., 2022; He et al., 2022; Tong et al., 2022).

5.2. Linear Probing Mask Image Models

When tuning MIMs on image classification tasks, there is a significant gap between linear probing and finetuning (He et al., 2022). A trick that is often used to improve linear probing performance is to append a batch normalization layer before the linear head (Chen et al., 2021). Without the BN layer, and with an appropriate batch size, the classification accuracy can drop significantly (Wu & Mo, 2022).

We argue that this is due to the nature of token-level Contrastive Learning. MIMs only learn to create and separate several clusters, but do not learn which cluster is indicative of the class label. It is often the case that the class token from a pretrained MIM does not learn the correct cluster. Therefore, partially finetuning a token mixing layer can greatly boost accuracy (He et al., 2022). We also argue that the BN layer adds non-linearity that partly serves as a token mixing layer. Therefore, we may need to rethink whether linear probing is a "fair" method to evaluate MIMs.

5.3. Network Architecture Matters

As discussed in Section 5.2, MIMs do not know how to select important tokens without finetuning. Therefore, a network architecture with token mixing layers is crucial for the success of mask models on classification tasks. Finetuning these layers allows MIMs to understand what are important tokens.

Another factor that may affect token selection is the number of attention heads in the Transformer architecture (Dosovitskiy et al., 2021). If token selection is a random process, more attention heads could increase the chances that tokens in the desired object are chosen. However, it may also reduce the representation capacity as the number of output logits become smaller. In practice, we found that, at least for some existing models, using more attention heads can improve both linear probing and finetuning performance on classification tasks.

6. Experiments

We have verified several of our assumptions and mathematical formulations with MAE models pretrained on Cifar10 and ImageNet-1K (IN-1K) datasets (Krizhevsky et al., 2009; Deng et al., 2009). The model backbones used for Cifar10 and ImageNet are ViT-Tiny and ViT-Base, respectively (Dosovitskiy et al., 2021). For ViT-Base on IN-1K, we followed the settings of MAE (He et al., 2022). The parameters of ViT-Tiny on Cifar-10 are given in Table 4, 5. We have used the same parameters for linear probing and finetuning, as we found that changing the optimizer of linear probing to AdamW gives better performance (Loshchilov & Hutter, 2017).

6.1. Low-rank Approximation of Different Mask Ratio

To verify Assumption 4.1, and our claim in Section 5.1, we computed the average distance of left singular vectors u_i between the reconstructed image and the original image on the Cifar10 dataset. Note that left singular vectors are the eigenvectors of the token similarity matrix ($X_0 X_0^\top$ for the original image), which is correlated with Spectral Clustering. We have visualized the leading 5 vectors in Figure 2.

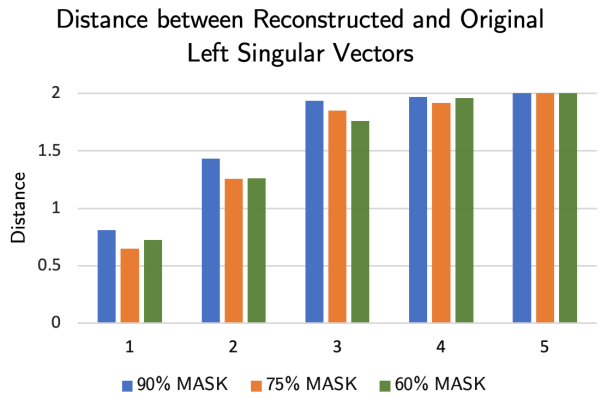

Figure 2. **Distances between left singular vectors of reconstructed image and original image.** We demonstrate the top-5 distances for models pretrained on three different masking ratios.

Our results demonstrate that for all of the three masking ratios, only 2 to 3 singular vectors of the token similarity matrix are notably approximated, which verifies our assumption of a low-rank structure in the token representations.

Additionally, we found that a higher masking ratio leads to the model approaching fewer singular vectors, resulting in fewer clusters in Spectral Clustering. This trend is further supported by our analysis of the distance of singular vectors, as shown in Table 1. We define d_i as the distance of the i^{th} singular vector, with $2 - d_i$ approximating how far it is from random initialization. We observe that when $\frac{2-d_i}{2-d_{i+1}}$ is larger, the model under this masking ratio tends to approach fewer singular vectors, thus verifying our claim in Section 5.1.

Table 1. **Distance ratio between one singular vector and the next.** d_i is the distance of i^{th} singular vectors, we calculate $\frac{2-d_i}{2-d_{i+1}}$ as a measurement of how fast distances increase.

MASKING RATIO	$\frac{2-d_1}{2-d_2}$	$\frac{2-d_2}{2-d_3}$
60 %	1.73	3.01
75 %	1.81	5.07
90 %	2.09	8.89

6.2. Visualizing CLS Token of MAE

In Section 5.2, we discussed about MAE's potential failure on cluster selection, hence, we visualized the CLS tokens's attention map on the last encoder block. This experiment is carried out on Cifar10 with ViT tiny, so we have 3 attention heads and thus 3 maps for each image.

We randomly picked two images from Cifar10 and visualized the attention map mentioned above in Figure 3. In the first image, we can see that Head 0 gives an outlier of the horse, which is a useful cluster that represents the object in the image. However, in the second image, all of the heads fail to capture the object in the image.

6.3. Different Probing Methods

In this experiment, we compare three different probing methods mentioned in Section 5.2. We conducted our experiments on both Cifar10 and IN-1K datasets with 1) linear probing with a linear head (LP), non-learnable batch normalization + linear probing (BN + LP), and partial finetuning (Partial FT). For partial finetuning, we tune a linear head and the last qkv projection layer in the encoder, which is also linear.

From our results in Table 2, it is apparent that there is a significant difference in performance between LP and BN+LP. This gap is larger than what is typically seen in other Contrastive Learning models (Chen et al., 2021). Another observation is that giving the model a simple learnable token

(a) Input (b) Head 0 (c) Head 1 (d) Head 2

Figure 3. **Visualization of CLS token's attention map on the final layer.** Row one demonstrates a successful example where one of the attention heads captures the desired object. Row two is a failed example where none of the attention heads captures the desired object. (a) is the input image, (b) (c) (d) are the attention maps of three different heads.

Table 2. **Classification accuracy with different probing methods.** We tested on three probing methods: linear probing with only a linear head (LP), batch normalization + linear probing (BN + LP), and partial finetuning (Partial FT). For partial finetuning, we tune a linear head and another linear layer from the original model (last qkv projection layer).

DATASET	LP	BN+LP	PARTIAL FT
CIFAR10	64.4	76.6	**83.3**
IN-1K	48.0	68.0	**69.3**

mixing layer could much improve classification accuracy.

6.4. Ablation on Number of Attention Heads

This experiment supports our assertion in Section 5.3 that the number of attention heads can impact the performance of MIMs. By increasing the number of heads in the ViT-Tiny model from 3 to 6 during pretraining, we observed a significant improvement in accuracy for both linear probing and finetuning on the classification task. The results are detailed in Table 3.

Table 3. **Classification accuracy with different attention head numbers.** We pretrained a ViT-Tiny on Cifar10 with 3 and 6 attention heads respectively. This table shows classification accuracy with linear probing (without BN) and finetuning.

HEADS	LP	FT
3	64.4	89.7
6	**67.5 (+3.1)**	**90.2 (+0.5)**

6.5. Finetuning Mask Model on Contrastive Loss

To further demonstrate the similarity of MIMs and Contrastive Learning, we finetuned a pretrained MAE model on a Contrastive loss task using Moco loss (Chen et al., 2021). We used the IN-1K pretrained MAE model and finetuned it for 30 epochs using the Moco loss, with a randomly initialized projection layer. We then compared the loss curve of this finetuned model with a Moco model trained from scratch. The results, shown in Figure 4, indicate that the MAE pretrained model quickly adapts to the Moco loss, suggesting that MIMs and Contrastive Learning share some similarities.

However, it should be noted that this is not a rigorous validation of their equivalence, but rather serves as supportive evidence for our claim.

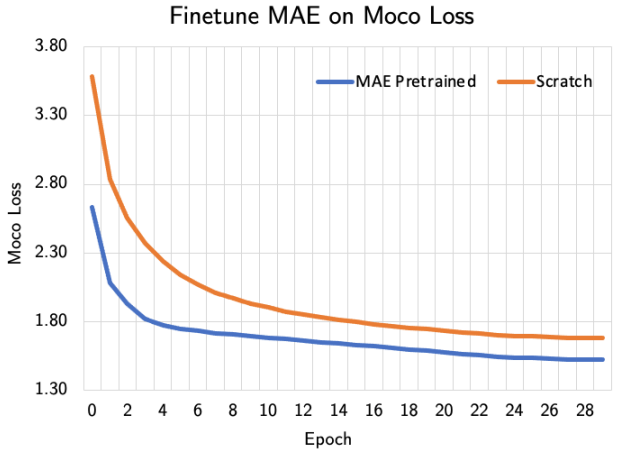

Figure 4. **Finetune MAE on Moco loss.** The blue line starts from a pretrained MAE, while the orange line starts from scratch. We finetuned for 30 epochs.

7. Discussion

Although mask modelling is a variant of Contrastive Learning at the token level, mask models have a distinct advantage in terms of pretraining efficiency and scalability as compared to other methods. This raises an interesting question: What is the most representative aspect of an input? Recent research has shown that certain patches within an image can be more informative than the whole image. This suggests that computing on the token level can lead to faster and more efficient training. However, the optimal patch size might be different for different tasks. Previous studies have shown that the optimal patch size for image classification is between 16 to 32 (Xie et al., 2021; Dosovitskiy et al., 2021), while a transformer-based model with a patch size of 8 performs better on semantic segmentation tasks (Strudel et al., 2021). This highlights the need for further research to

determine how to most effectively use unlabeled inputs for different tasks.

8. Conclusion

In this paper, we propose a theoretical framework for analyzing mask models. We discover that mask modeling is a form of Contrastive Learning at the token level, which may account for its success. Our framework also addresses important questions regarding the behavior of mask models. We hope that our study will offer valuable insights into designing self-supervised learning algorithms and model architectures.

Table 4. **Pretraining parameters of ViT-T on Cifar10**

config	value
optimizer	AdamW
base learning rate	1.5e-4
weight decay	0.05
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.95$
batch size	512
learning rate schedule	cosine decay
warmup epochs (Goyal et al., 2017)	200
total epochs	2000
augmentation	None
patch size	2×2

Table 5. **Finetuning and Linear probing parameters of ViT-T on Cifar10**

config	value
optimizer	AdamW
base learning rate	1e-3
weight decay	0.05
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$
batch size	128
learning rate schedule	cosine decay
warmup epochs	5
training epochs	100
augmentation	None

References

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., and Saunshi, N. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Bachmann, R., Mizrahi, D., Atanov, A., and Zamir, A. MultiMAE: Multi-modal multi-task masked autoencoders. 2022.

Bao, H., Dong, L., Piao, S., and Wei, F. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=p-BhZSz59o4.

Cao, S., Xu, P., and Clifton, D. A. How to understand masked autoencoders. *arXiv preprint arXiv:2202.03670*, 2022.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., and Sutskever, I. Generative pretraining from pixels. In *International Conference on Machine Learning*, pp. 1691–1703. PMLR, 2020.

Chen, X., Xie, S., and He, K. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9640–9649, 2021.

Chen, X., Ding, M., Wang, X., Xin, Y., Mo, S., Wang, Y., Han, S., Luo, P., Zeng, G., and Wang, J. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*, 2022.

Davenport, M. A. and Romberg, J. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4): 608–622, 2016.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and 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.

Hofmann, T., Schölkopf, B., and Smola, A. J. Kernel methods in machine learning. *The annals of statistics*, 36(3): 1171–1220, 2008.

Horn, R. A. and Johnson, C. R. *Matrix Analysis*. Cambridge University Press, 1985. doi: 10.1017/CBO9780511810817.

Huang, P.-Y., Xu, H., Li, J., Baevski, A., Auli, M., Galuba, W., Metze, F., and Feichtenhofer, C. Masked autoencoders that listen. In *NeurIPS*, 2022.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

Kong, X. and Zhang, X. Understanding masked image modeling via learning occlusion invariant feature. *arXiv preprint arXiv:2208.04164*, 2022.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Ng, A., Jordan, M., and Weiss, Y. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.

Pan, J., Zhou, P., and Yan, S. Towards understanding why mask-reconstruction pretraining helps in downstream tasks. *arXiv preprint arXiv:2206.03826*, 2022.

Patel, V. M., Van Nguyen, H., and Vidal, R. Latent space sparse and low-rank subspace clustering. *IEEE Journal of Selected Topics in Signal Processing*, 9(4):691–701, 2015.

Saunshi, N., Ash, J., Goel, S., Misra, D., Zhang, C., Arora, S., Kakade, S., and Krishnamurthy, A. Understanding contrastive learning requires incorporating inductive biases. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 19250–19286. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/saunshi22a.html.

Shi, J. and Malik, J. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

Strudel, R., Garcia, R., Laptev, I., and Schmid, C. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7262–7272, 2021.

Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34:24261–24272, 2021.

Tong, Z., Song, Y., Wang, J., and Wang, L. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. 2017.

Wettig, A., Gao, T., Zhong, Z., and Chen, D. Should you mask 15% in masked language modeling? *arXiv preprint arXiv:2202.08005*, 2022.

Wu, J. and Mo, S. Object-wise masked autoencoders for fast pre-training. *arXiv preprint arXiv:2205.14338*, 2022.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., and Hu, H. Simmim: A simple framework for masked image modeling. 2021.

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., and Kong, T. Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=ydopy-e6Dg.