From Linearity to Non-Linearity: How Masked Autoencoders Capture Spatial Correlations

Supplementary Material

A. Linear MAEs: Characterizing critical points

We present additional details for Theorem 1 below.

Consider the loss.

$$\ell = ||X - (1-m)XAB||^2 + m(1-m)||GAB||^2, \quad (12)$$

We first set $\partial\ell/\partial A$ to 0 to get

$$-2(1-m)X^\top XB^\top + 2(1-m)^2 X^\top XABB^\top + 2m(1-m)\text{Blkdiag}_p(X^\top X)ABB^\top = 0. \quad (13)$$

Any critical point must satisfy

$$A^* = V^{-1}X^\top XB^\top(BB^\top)^{-1} \quad (14)$$

where $V = (1-m)X^\top X + m\text{Blkdiag}_p(X^\top X)$. Substituting this value of A^* back into the loss, we get

$$\ell = \text{Tr}(X^\top X) - (1-m)\,\text{Tr}[B(X^\top XV^{-1}X^\top X)B^\top(BB^\top)^{-1}]$$

Let $C = X^\top XV^{-1}X^\top X$ and $D = I$. Note that C and D are both symmetric and D is invertible. Using lemma 3, the expression is minimized by the k largest eigenvalues of the generalized eigenvalue problem defined on (C, D). Furthermore, every critical point is a subset of k eigenvectors (from Lemma 2).

B. Non-linear MAEs using linear approximations

Non-linear masked autoencoders under a Taylor series approximation. Consider a nonlinear autoencoder f and the corresponding masked autoencoder loss

$$\ell_m = \mathbb{E}_R||X - f(R \odot X)||^2.$$

Let $X_\mu = (1-m)X$ and $X_r = R \odot X$. Under the Taylor series approximation around 0

$$f(X_R) \approx f(0) + X_R\nabla f(0)^\top = X_R\nabla f(0)^\top,$$

the MAE loss reduces to

$$\ell_m = ||X - (1-m)X\nabla f(0)^\top||^2 + m(1-m)||G\nabla f(0)^\top||^2$$

Note that this approximation holds for small perturbations to the input, and is less likely to hold for large perturbations, i.e., the approximation is only valid for small masking ratio.

We will consider another approximation, but this time calculate the loss for a single sample. We consider a first-order approximation of f for a single sample.

$$f(x_R) \approx f(x_\mu) + \nabla f(x_\mu)(x_R - x_\mu),$$

which when substituted into the above loss gives us

$$\ell_m(x) = ||x - f(x_\mu)||^2 + m(1-m)\,\text{Tr}\left(F_x\text{Blkdiag}(x^\top x)\right)$$

Note that this is the loss for a single sample and F_x is the Fisher information matrix for data point x.

Masked autoencoders: A function space perspective Let us assume that we have access to the true input image signal $x(i, j)$, where $i, j \in [0, 1]$, as opposed to a discretized version of it. The masked autoencoder objective can be posed as an optimization problem over functionals $f \in \mathcal{F}$, i.e.,

$$\ell_m = \int ||f(r \odot x) - x||^2 \,\mathrm{d}r,$$

where r is a mask applied to the image. Assuming that f is linear, then the above objective reduces to

$$\ell_m = ||x - f(\mu)||^2 - ||f(\mu)||^2 + \int ||f(r \odot x)||^2 \,\mathrm{d}r,$$

where $\mu = \int (r \odot x)\,\mathrm{d}r$. In the linear case, the MAE forces f to reconstruct the mean masked image, while minimizing the variance of the predictions made on the masked images.

C. Supporting lemmas

Lemma 1 *The loss of the masked autoencoder is*

$$\ell_m = ||X - (1-m)XAB||^2 + m(1-m)||GAB||^2$$

Proof 1 *Expanding the term inside the expectation, we get*

$$\mathbb{E}_R||X-(R\odot X)AB||^2 = \mathbb{E}_R Tr\,[X^\top X - 2X^\top(R\odot X)AB + B^\top A^\top(R\odot X)^\top(R\odot X)AB].$$

We note that $\mathbb{E}[R \odot X] = (1-m)X$ *and*

$$\mathbb{E}[(R\odot X)^\top(R\odot X)] = \begin{cases}(1-m)X_i^\top X_j & X_i, X_j \text{ in the same patch}\\ (1-m)^2X_i^\top X_j & X_i, X_j \text{ not in the same patch.}\end{cases}$$

$$= (1-m)^2X^\top X + m(1-m)\text{Blkdiag}_p(X^\top X).$$

Substituting back into the masked autoencoder loss, we get

$$\begin{aligned}
&\mathbb{E}_R||X-(R\odot X)AB||^2 \\
&= Tr\left[X^\top X - 2(1-m)X^\top XAB + (1-m)^2X^\top X\right. \\
&\qquad \left. + m(1-m)B^\top A^\top \text{Blkdiag}_p(X^\top X)AB\right] \\
&= ||X-(1-m)XAB||^2 + m(1-m)||GAB||^2
\end{aligned}$$

where $G^\top G = \text{Blkdiag}_p(X^\top X)$.

Lemma 2 *For matrices* $C \in \mathbb{R}^{d\times d}$, $X \in \mathbb{R}^{d\times k}$ *and an invertible matrix* $D \in \mathbb{R}^{d\times d}$, *every critical point of*

$$L(X) = \text{Tr}[(X^\top DX)^{-1}X^\top CX]$$

that is full-rank can be expressed as UQ, *where* Q *is an invertible matrix and* U *is any subset of* k *eigenvectors of the generalized eigenvalue problem for* (C, D).

Proof 2 *Taking the derivative of* $L(X)$ *with respect to* X *and setting it to 0, we get*

$$2DX(X^\top DX)^{-1}X^\top CX = 2CX.$$

Let (Λ_D, Φ_D) *be the eigenvectors and eigenvalues of* D. *Let* $(\Lambda_{\tilde{C}}, \Phi_{\tilde{C}})$, *be the eigenvectors and eigenvalues* $\Lambda_D^{-1/2}\Phi_D^\top C\Phi_D\Lambda_D^{-1/2}$. *In addition, we define* $\Phi = \Phi_{\tilde{C}}\Lambda_D^{-1/2}\Phi_{\tilde{D}}$ *and* $\tilde{X} = \Phi X$. *We choose this definition of* Φ, *since it diagonalizes both* C *and* D.

$$\begin{aligned}
&2DX(X^\top DX)^{-1}X^\top CX = 2CX \\
\implies &D\Phi^\top\tilde{X}(\tilde{X}^\top\Phi D\Phi^\top\tilde{X})^{-1}\tilde{X}^\top\Phi C\Phi^\top\tilde{X} = C\Phi^\top\tilde{X} \\
\implies &D\Phi^\top\tilde{X}(\tilde{X}^\top\tilde{X})^{-1}\tilde{X}^\top\Lambda_{\tilde{C}}\tilde{X} = C\Phi^\top\tilde{X} \\
\implies &\Phi D\Phi^\top\tilde{X}(\tilde{X}^\top\tilde{X})^{-1}\tilde{X}^\top\Lambda_{\tilde{C}}\tilde{X} = \Phi C\Phi^\top\tilde{X} \\
\implies &\tilde{X}(\tilde{X}^\top\tilde{X})^{-1}\tilde{X}^\top\Lambda_{\tilde{C}}\tilde{X} = \Lambda_{\tilde{C}}\tilde{X} \\
\implies &P_{\tilde{X}}\Lambda_{\tilde{C}}\tilde{X} = \Lambda_{\tilde{C}}P_{\tilde{X}}\tilde{X}
\end{aligned}$$

where $P_{\tilde{X}}$ *is the projection operator. Note that* $P_{\tilde{X}}\tilde{X} = \tilde{X}$. *Since* $\Lambda_{\tilde{C}}$ *is diagonal,* $P_{\tilde{X}}$ *must also be diagonal in order for the matrices to commute. Furthermore,* $P_{\tilde{X}}$ *has exactly* k *eigenvalues equal to 1 and the rest set to 0, since* X *has rank* k. *Hence,* $\tilde{X}$ *must be of the form* $I_{S_k}Q$ *where* S_k *selects a subset of* k *dimensions and* $Q \in \mathbb{R}^{k\times k}$ *is an invertible matrix. Hence* $X = \Phi_{S_k}Q$ *where* Φ_{S_k} *is a subset of* k *eigenvectors of the generalized eigenvalue problem.*

Lemma 3 *For matrices* $C \in \mathbb{R}^{d\times d}$, $X \in \mathbb{R}^{d\times k}$ *and an invertible matrix* $D \in \mathbb{R}^{d\times d}$, *the global maximum of*

$$L(X) = \text{Tr}[(X^\top DX)^{-1}X^\top CX]$$

is $\sum_{i=1}^k \Lambda_k$ *where* Λ *are the eigenvalues of the generalized eigenvalue problem* (C, D).

Proof 3 *From lemma 2, we know that any critical point is of the form* $\Phi_{S_k}Q$. *Subtituting this into* $L(X)$, *we get*

$$\begin{aligned}
L(X) &= \text{Tr}[(X^\top DX)^{-1}X^\top CX] \\
&= \text{Tr}[(Q^\top Q)^{-1}Q^\top\Phi_{S_k}^\top C\Phi_{S_k}Q] \\
&= \text{Tr}[\Phi_{S_k}^\top C\Phi_{S_k}] = \sum_{i\in S_k}\Lambda_i.
\end{aligned}$$

The loss is maximized by the largest k *eigenvalues and minimized by the smallest* k *eigenvalues.*

Lemma 4 *Under the Taylor series approximation of* $f(X_R) \approx f(0) + X_R\nabla f(0)^\top = X_R\nabla f(0)^\top$, *the MAE loss for a non-linear function* f *is*

$$\ell_m = ||X-(1-m)X\nabla f(0)^\top||^2 + m(1-m)||G\nabla f(0)^\top||^2.$$

Proof 4

$$\begin{aligned}
\ell_m &= \mathbb{E}_R||X - X_R\nabla f(0)^\top||^2 \\
&= ||X||^2 + \mathbb{E}_R||X_R\nabla f(0)^\top||^2 - 2\mathbb{E}_R\,\text{Tr}(X^\top X_R\nabla f(0)^\top) \\
&= \text{Tr}(X^\top X) + (1-m)^2\,\text{Tr}\left(\nabla f(0)X^\top X\nabla f(0)^\top\right) \\
&\qquad + m(1-m)\,\text{Tr}\left(\nabla f(0)\text{Blkdiag}(X^\top X)\nabla f(0)^\top\right) \\
&\qquad - 2(1-m)\mathbb{E}_R\left[X^\top X\nabla f(0)^\top\right] \\
&= ||X-(1-m)X\nabla f(0)^\top||^2 + m(1-m)||G\nabla f(0)^\top||^2.
\end{aligned}$$

Lemma 5 *Under the Taylor series approximation of* $f(x_R) \approx f(x_\mu) + \nabla f(x_\mu)(x_R - x_\mu)$, *the MAE loss, reduces to*

$$\ell_m(x) = ||x-f(x_\mu)||^2 + m(1-m)\,\text{Tr}\left(F_x\text{Blkdiag}(x^\top x)\right).$$

Proof 5

$$\begin{aligned}
\ell_m(x) &= ||x - f(x_\mu) - \nabla f(x_\mu)(x_R - x_\mu)||^2 \\
&= ||x-f(x_\mu)||^2 + \mathbb{E}_R||\nabla f(x_\mu)(x_R-x_\mu)||^2 \\
&\qquad + 2\mathbb{E}_R(x-f(x_\mu))^\top(\nabla f(x_\mu)(x_R-x_\mu)) \\
&= ||x-f(x_\mu)||^2 + \\
&\qquad \mathbb{E}_R(x_R-x_\mu)^\top\nabla f(x_\mu)^\top\nabla f(x_\mu)(x_R-x_\mu) \\
&= ||x-f(x_\mu)||^2 + \\
&\qquad \mathbb{E}_R\,\text{Tr}\left(\nabla f(x_\mu)^\top\nabla f(x_\mu)(x_R-x_\mu)(x_R-x_\mu)^\top\right) \\
&= ||x-f(x_\mu)||^2 + m(1-m)\,\text{Tr}\left(F_x\text{Blkdiag}(x^\top x)\right).
\end{aligned}$$

Lemma 6 *The masked autoencoder loss, for a input image* x *and linear functional* f *is*

$$\ell_m = ||x-f(\mu)||^2 - ||f(\mu)||^2 + \int ||f(r\odot x)||^2\,\text{d}r,$$

where $\mu = \int(r\odot x)\,\text{d}r$

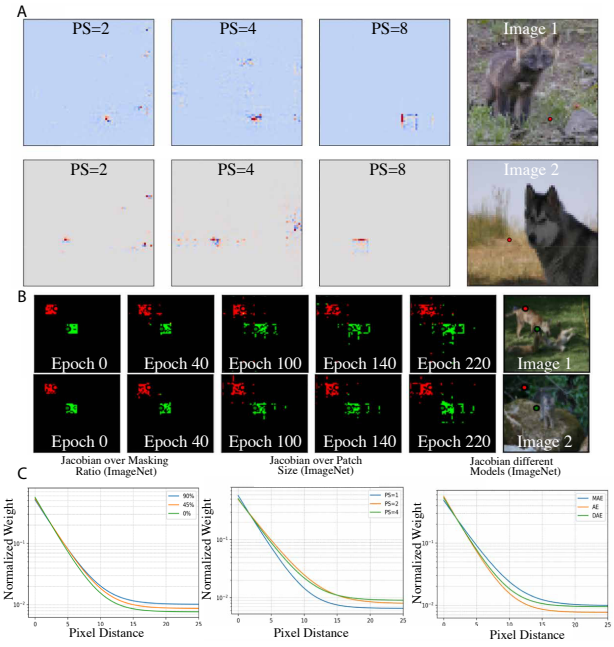

Figure 13. Analogous to Figs. 5, 6, 8, 9, but for the ImageNet-64 dataset. A) Visualization of the Jacobian ($m = 0.8$) for nonlinear MAE. B) Jacobian across different stages of training for a specific output pixel ($ps = 8$, $m = 0.8$) for nonlinear MAE. C) Normalized weight for different hyper-parameters of a linear MAE, AE and DAE.

Proof 6

$$
\begin{aligned}
\ell_m &= \int ||x - f(r \odot x)||^2 \, \mathrm{d}r \\
&= \int ||x - f(\mu) - (f(r \odot x) - f(\mu))||^2 \, \mathrm{d}r \\
&= \int ||f(\mu) - x||^2 + ||f(r \odot x) - f(\mu)||^2 \, \mathrm{d}r \\
&\quad - \int 2\langle x - f(\mu), f(r \odot x) - f(\mu) \rangle \mathrm{d}r \\
&= \int ||f(\mu) - x||^2 + ||f(r \odot x) - f(\mu)||^2 \, \mathrm{d}r \\
&= ||x - f(\mu)||^2 - ||f(\mu)||^2 + \int ||f(r \odot x)||^2 \, \mathrm{d}r.
\end{aligned}
$$

D. Additional experiments

Additional details about MAE pretraining We train MAEs using the architecture in He et al. [20]. We divide the image into patches of size p and randomly mask a fraction m of the patches before feeding it to the MAE. The encoder projects the unmasked patches to a d-dimensional embedding using a linear layer. The sequence of patches are then fed to a series of Transformer blocks. The decoder adds a learnable vector and position encoding for every masked patch and reconstructs the masked patches.

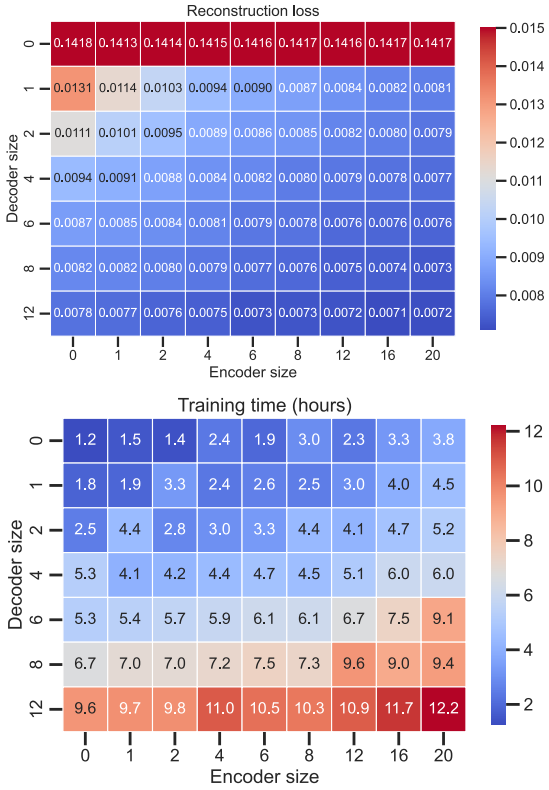

Figure 14. We vary the number of encoder and decoder layers and record the reconstruction loss at the end of training and the time required to train the MAEs. Note that **MAEs are slow to train** with training time growing faster the the size of the decoder. The reconstruction loss becomes smaller with increasing size of both the decoder and the encoder. However, we find that the **training loss is not a good proxy for downstream task performance**.

Number of Encoder and decoder layers We train MAEs on CIFAR10 for different encoder and decoder sizes. We find that the reconstruction loss decreases with increasing size of both the encoder and the decoder (Fig. 14). However, the training time grows faster than the size of the decoder, making it computationally expensive to train large decoders. We also find that training loss is not a good proxy for downstream task performance. We evaluate the performance of the trained encoder using linear probing and find that the accuracy improves as we increase the size of the encoder. However, the optimal decoder size is 2-4 layers (Fig. 15).

If the model are trained only using the supervised loss, i.e., we do no MAE pretraining, then the accuracy on CIFAR plateaus around 83-84%. In fact the accuracy for a 20-layer network is worse than the accuracy for a 12-layer

network which differs from the trend for masked autoencoders.

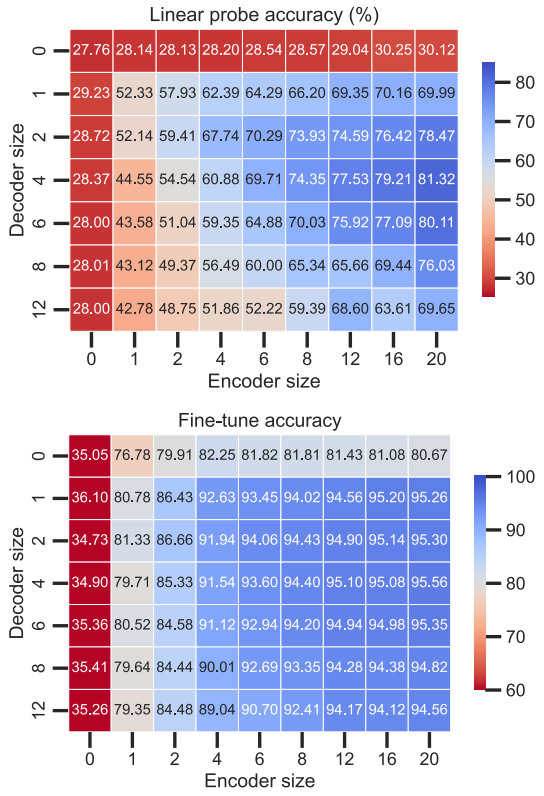

Figure 15. We vary the number of encoder and decoder layers (transformer blocks) and plot the linear probe accuracy (left) and the accuracy after fine-tuning for 100 epochs. The **accuracy of the trained encoder continues to improve as we increase its size**. Linear probe accuracies are usually indicative of performance after fine-tuning.

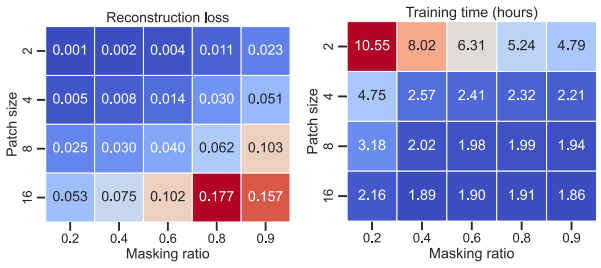

Figure 16. We vary the masking ratio and patch-size of the masked autoencoder. While larger masking ratio lead to smaller training times, even smaller masking ratios work quite well. Smaller patch-sizes are a lot slower to train but usually perform better than larger patch-sizes.

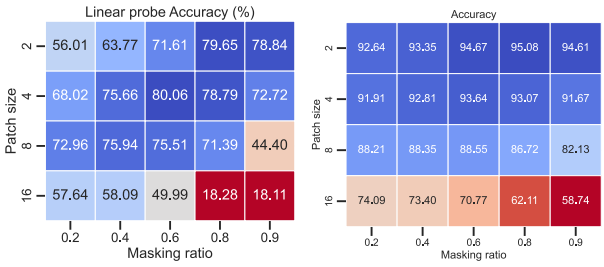

Figure 17. We plot the linear probe (Left) and accuracy after fine-tuning (right) for different patch-size and masking ratio. Patch-sizes of 2 and 4

Patch-size vs. Masking ratio The masking ratio and patch-size control the basis learned by the MAE. We surprisingly find that many different parameters work surprisingly well for downstream task accuracy. We also note that reconstruction loss is not indicative of downstream task performance.

Are long training times even necessary? MAEs are typically trained for a large number of epochs and the reconstruction error continues to decrease over the course of training. However, the reconstruction error is not predictive of both the linear-probe and fine-tuning accuracies. In Fig. 10, we consider multiple checkpoints over the course of pretraining and plot the number of pretraining epochs against the linear probe accuracy of that checkpoint. We find that the **linear probe accuracy continues to increase even after 1500 epochs of training** particularly for larger models, justifying the need to pretrain for a large number of epochs (see Fig. 15).

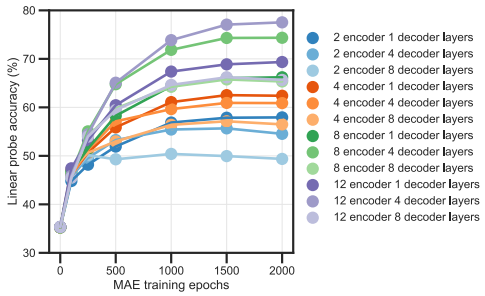

Figure 18. We plot the linear probe accuracies of different masked autoencoders over the course of training. They accuracy continue to increase even after 1000 epochs of training, justifying the need for long training times. Larger models tend to require longer training times.

Centered kernel alignment or CKA [30] measures the similarity between representations of two different net-

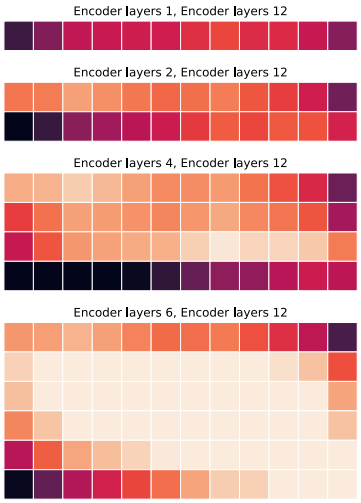

Figure 19. CKA between MAEs trained with different number of encoder layers. Each row and column corresponds to the similarity between a k-layer encoder and the 12-layer encoder (hence 12 columns). The darker shades indicate that the representations for those two layers are not similar.

works. We use CKA to measure the similarity between the representations of MAEs trained with different number of encoder layers and with 4 decoder layers. White indicates that the similarity is high and black/red indicates that the similarity is low. We find that the larger networks are more similar to the 12-layer encoder while while the smaller networks are less similar to the 12-layer network, particularly at the last layer.

E. More experiments with the Ising model

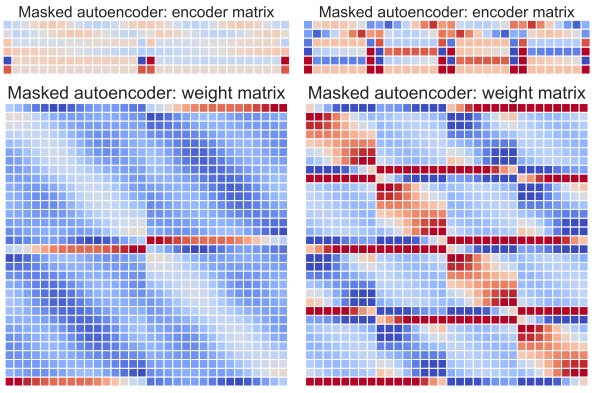

Figure 20. We plot the weight matrix AB and the encoder matrix A for **(left)** patch-size 16 and masking ratio of 0.5 and **(right)** patch-size 8 and masking ratio 0.5. Increasing the patch-size while keeping the masking ratio fixed biases the encoder towards features that capture long-range correlations.

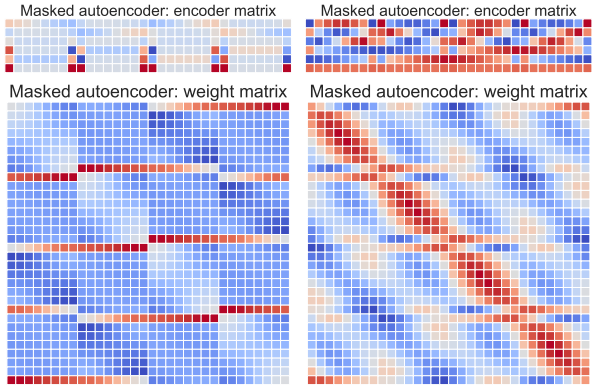

Figure 21. We plot the weight matrix AB and the encoder matrix A for **(left)** patch-size 16 and masking ratio of 0.99 and **(right)** patch-size 8 and masking ratio (0.01). Reducing the masking ratio biases the encoder towards features based on local correlations while increasing the masking ratio prioritizes features that capture long range correlations.

F. MAEs in the Frequency Domain

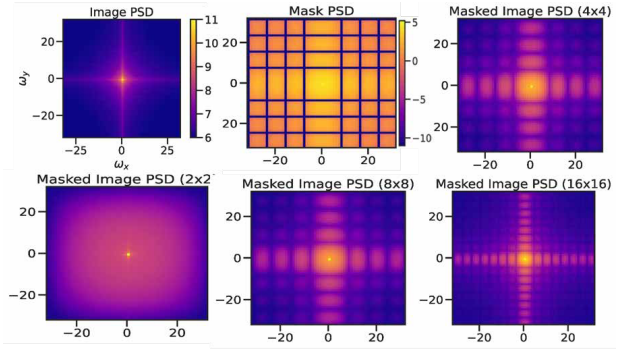

Figure 22. Masking using square patches in a MAE zeros out frequencies in a grating-like pattern in Fourier space (color bar in image log power spectral density used for all plots, but mask).

MAEs mask a part of the input image (20% of the patches), usually multiple square patches, and train an encoder-decoder pair to reconstruct the intensity at the masked patches. Fig. 22 (top) shows the power spectral density (PSD), i.e., squared amplitude at different spatial frequencies, of the original image (left), the mask itself (top, middle) and masked images using different patch sizes. The Discrete Fourier Transform (DFT) magnitude of these masks is a sinc function modulated by a sum of complex exponentials Eq. (21). This can be shown in a 1D setting by considering a discrete signal $x \in \mathbb{R}^D$. For MAEs with patch size p, we parameterize each individual mask using a rectangular pulse defined by function r as

$$r[n] = u[n] - u[n-p], \tag{15}$$

where p is the patch size, n is the discrete time index and $u[n]$ is the Heaviside step function. A MAE mask $m \in \mathbb{R}^D$ is constructed as a sum of such rectangular pulses:

$$m[n] = \sum_{i=1}^{N} r[n - a_i], \tag{16}$$

where a_i denotes the starting index of the ith mask. The masked signal is then given by elementwise multiplication of the signal x and the mask m:

$$y[n] = x[n] \cdot m[n]. \tag{17}$$

In the frequency domain, by the multiplication property of the DFT, this becomes:

$$Y[k] = \frac{1}{D} X[k] * M[k], \tag{18}$$

where $X[k]$ and $M[k]$ are the DFTs of $x[n]$ and $m[n]$, respectively and $k \in [0, 1, ..., N-1]$ is the frequency index.

Using the time-shift and linearity properties of the DFT, the transform of the mask $m[n]$ can be written as:

$$M[k] = \sum_{i=1}^{N} e^{-j2\pi k a_i} \cdot R[k], \tag{19}$$

where the DFT of $r[n]$ is given by:

$$R[k] = e^{-j\pi k(p-1)/D} \frac{\sin(\pi p k/D)}{\sin(\pi k/D)}. \tag{20}$$

Hence, applying MAE-style masking results in a spectral smoothing effect analogous to sinc filtering, as illustrated in Fig. 22.

$$|R[k]| = \left| \frac{\sin(\pi p k/D)}{\sin(\pi k/D)} \right| \left| \sum_{i=1}^{N} e^{-j2\pi k a_i} \right|. \tag{21}$$

A square patch therefore corresponds to masking frequencies in a grating-like pattern. Masking in an MAE therefore corresponds to zeroing out certain frequencies. The goal of an MAE is to interpolate the value of the masked frequencies from the unmasked version of the spectrum.

G. Related work

Masked Autoencoders. The goal of visual representation learning is to develop representations that go beyond the identity function [44] and instead are predictive of downstream visual tasks. Early methods in this field sought to prevent trivial identity mappings by incorporating various noise distributions, such as blankout noise in Dropout [43] or Marginalized Autoencoders [9], and Gaussian noise in Denoising Autoencoders [47]. A key limitation of these earlier approaches is that they were developed before deep networks could be easily trained end-to-end, resulting in reliance on greedy layer-wise training, which does not generalize as well as full end-to-end optimization. Modern methods, such as Data2Vec [2] and BERT [12], overcome this limitation by training Vision Transformers (ViTs) end-to-end using blackout noise, achieving state-of-the-art representation performance. The core innovation of modern MAEs [20, 54] lies in their architectural design, which encodes only masked image patches, significantly reducing computational costs for masked prediction strategies. Recent advancements in MAEs include cross-attention mechanisms for efficient computation [18], frequency-based loss functions to mitigate oversmoothing [33, 52], and intermediate perception reconstruction instead of direct output reconstruction [42, 50]. Additional innovations involve task-guided masking strategies [15, 31, 53], distillation-based training [3], and numerous other techniques [22].