Aligning Reconstruction-Based and Supervised Representations Through Masking

A discussion on: How Learning by Reconstruction Produces Uninformative Features For Perception (Balestriero & LeCun, ICML 2024)

Panagiotis Koromilas

University of Athens

CV & Robotics reading group

7 October 2024

Representation Learning Through Reconstruction

Q: Why does reconstruction-based learning generate **highly compelling reconstructed samples**, yet **struggle to produce informative latent representations** for downstream tasks, often **requiring extensive fine-tuning** to be effective?

Linear case

$$\mathcal{L}(\boldsymbol{V}, \boldsymbol{W}, \boldsymbol{Z}) = \left\| \boldsymbol{W}^\top \boldsymbol{V}^\top \boldsymbol{X} - \boldsymbol{Y} \right\|_F^2 + \lambda \left\| \boldsymbol{Z}^\top \boldsymbol{V}^\top \boldsymbol{X} - \boldsymbol{X} \right\|_F^2$$

- encoder $V \in \mathcal{M}_{K,D}(\mathbb{R})$
- decoder $\boldsymbol{Z} \in \mathcal{M}_{D,K}(\mathbb{R})$
- predictor head $\boldsymbol{W} \in \mathcal{M}_{C,K}(\mathbb{R})$, C is the number of classes

Optimal solution of the linear case

Theorem 1. The combined loss function is minimized for

$\boldsymbol{V}^*$ spans $\boldsymbol{P}_{\boldsymbol{X}\boldsymbol{X}^\top}\boldsymbol{D}_{\boldsymbol{X}_{\boldsymbol{X}}}^{-\frac{1}{2}}(\boldsymbol{P}_{\boldsymbol{H}})_{,1:K}$,

$$\boldsymbol{W}^* = \left(\boldsymbol{V}^{*\top}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{V}^*\right)^{-1}\boldsymbol{V}^{*\top}\boldsymbol{X}\boldsymbol{Y}^\top,$$

$$\boldsymbol{Z}^* = \left(\boldsymbol{V}^{*\top}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{V}^*\right)^{-1}\boldsymbol{V}^{*\top}\boldsymbol{X}\boldsymbol{X}^\top,$$

where $\boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{P}_{\boldsymbol{X}\boldsymbol{X}^\top}\boldsymbol{D}_{\boldsymbol{X}\boldsymbol{X}^\top}\boldsymbol{P}_{\boldsymbol{X}\boldsymbol{X}^\top}^\top$, and $\boldsymbol{H} \triangleq \boldsymbol{D}_{\boldsymbol{X}\boldsymbol{X}^\top}^{-\frac{1}{2}}\boldsymbol{P}_{\boldsymbol{X}\boldsymbol{X}^\top}^\top\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{X}\boldsymbol{X}^\top}\boldsymbol{D}_{\boldsymbol{X}\boldsymbol{X}^\top}^{-\frac{1}{2}}$.

Optimal solution of the linear case

Theorem 1. The combined loss function is minimized for

$\boldsymbol{V}^*$ spans $\boldsymbol{P}_{\boldsymbol{X}\boldsymbol{X}^\top} \boldsymbol{D}_{\boldsymbol{X}_{\boldsymbol{X}}}^{-\frac{1}{2}} (\boldsymbol{P}_{\boldsymbol{H}})_{,1:K}$,

$$\boldsymbol{W}^* = \left(\boldsymbol{V}^{*\top}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{V}^*\right)^{-1}\boldsymbol{V}^{*\top}\boldsymbol{X}\boldsymbol{Y}^\top,$$

$$\boldsymbol{Z}^* = \left(\boldsymbol{V}^{*\top}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{V}^*\right)^{-1}\boldsymbol{V}^{*\top}\boldsymbol{X}\boldsymbol{X}^\top,$$

where $\boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{P}_{\boldsymbol{X}\boldsymbol{X}^\top}\boldsymbol{D}_{\boldsymbol{X}\boldsymbol{X}^\top}\boldsymbol{P}_{\boldsymbol{X}\boldsymbol{X}^\top}^\top$, and $\boldsymbol{H} \triangleq \boldsymbol{D}_{\boldsymbol{X}\boldsymbol{X}^\top}^{-\frac{1}{2}}\boldsymbol{P}_{\boldsymbol{X}\boldsymbol{X}^\top}^\top\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{X}\boldsymbol{X}^\top}\boldsymbol{D}_{\boldsymbol{X}\boldsymbol{X}^\top}^{-\frac{1}{2}}$.

Proof steps:

- find optimal $\mathbb{W}^*$ and $\mathbb{Z}^*$ as a function of $\mathbb{V}$
- find optimal V as the solution of a generalized eigenvalue problem

Optimal solution of the linear case

Theorem 1. The combined loss function is minimized for

$\boldsymbol{V}^*$ spans $\boldsymbol{P}_{\boldsymbol{X}\boldsymbol{X}^\top} \boldsymbol{D}_{\boldsymbol{X}\boldsymbol{X}}^{-\frac{1}{2}} (\boldsymbol{P}_{\boldsymbol{H}})_{,1:K}$,

$$\boldsymbol{W}^* = \left(\boldsymbol{V}^{*\top}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{V}^*\right)^{-1}\boldsymbol{V}^{*\top}\boldsymbol{X}\boldsymbol{Y}^\top,$$

$$\boldsymbol{Z}^* = \left(\boldsymbol{V}^{*\top}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{V}^*\right)^{-1}\boldsymbol{V}^{*\top}\boldsymbol{X}\boldsymbol{X}^\top,$$

where $\boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{P}_{\boldsymbol{X}\boldsymbol{X}^\top}\boldsymbol{D}_{\boldsymbol{X}\boldsymbol{X}^\top}\boldsymbol{P}_{\boldsymbol{X}\boldsymbol{X}^\top}^\top$, and $\boldsymbol{H} \triangleq \boldsymbol{D}_{\boldsymbol{X}\boldsymbol{X}^\top}^{-\frac{1}{2}}\boldsymbol{P}_{\boldsymbol{X}\boldsymbol{X}^\top}^\top\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{X}\boldsymbol{X}^\top}\boldsymbol{D}_{\boldsymbol{X}\boldsymbol{X}^\top}^{-\frac{1}{2}}$.

Corrolary: The solution from Theorem 1 recovers the OLS solution for $\boldsymbol{W}^{*\top}\boldsymbol{V}^{*\top}$ as $\lambda \to 0$, and the PCA solution for $\boldsymbol{Z}^{*\top}\boldsymbol{V}^{*\top}$ as $\lambda \to \infty$.

Measuring alignment

Q: under which condition $\mathbb{V}^*$ is not impacted by λ?

$$\text{alignment}(k) \triangleq \frac{\left\| \boldsymbol{Y}^\top \boldsymbol{Y} \left(\boldsymbol{P}_{\boldsymbol{X}\boldsymbol{X}^\top}\right)_{1:k} \right\|_F^2}{\left\| \boldsymbol{Y}^\top \boldsymbol{Y} \boldsymbol{P}_{\boldsymbol{X}\boldsymbol{X}^\top} \right\|_F^2}$$

is the minimum supervised error that can be achieved given the $\left(\boldsymbol{V}^\top \boldsymbol{X}\right)$ minimizes reconstruction which is measured by how much of the matrix $\boldsymbol{Y}^\top \boldsymbol{Y}$ can be reconstructed from the top- k subspa of $\boldsymbol{X}^\top \boldsymbol{X}$

Corollary 1.2. alignment(k) increases with k, has value 0 iff the two losses are misaligned, and has value 1 iff the two losses are aligned.

Measuring alignment

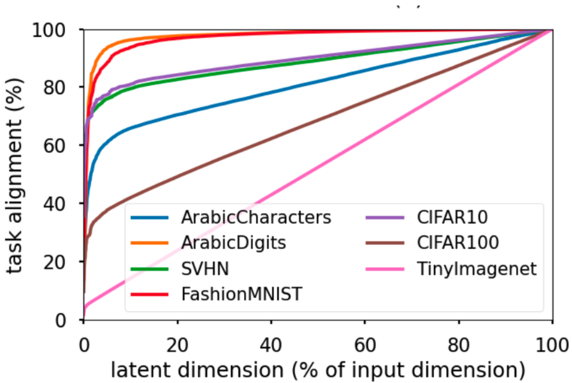


Figure: Figure from Balestriero & LeCun (ICML 2024)

- alignment for images without background
- alignment decreases for more classes (CIFAR10 vs CIFAR100)
- alignment decreases for higher resolution

Non-linear case

$$\mathcal{L}(\boldsymbol{W},\theta,\gamma) = \left\| \boldsymbol{W}^\top f_\theta(\boldsymbol{X}) - \boldsymbol{Y} \right\|_F^2 + \lambda \left\| g_\gamma\left(f_\theta(\boldsymbol{X})\right) - \boldsymbol{X} \right\|_F^2$$

Theorem 2. For any high-capacity encoder f_θ, studying the linear and the non-linear equation is equivalent at initialization for any decoder, and is always equivalent when the decoder is linear.

That is, linear and non-linear encoders
learn the principal subspace early in the training

Emperical non-linear case

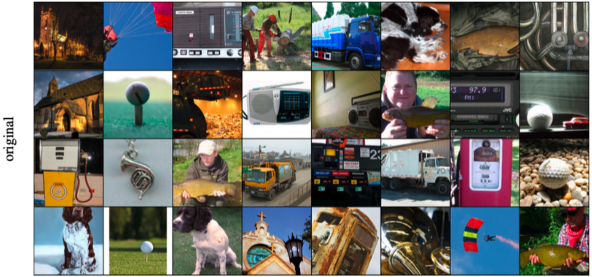
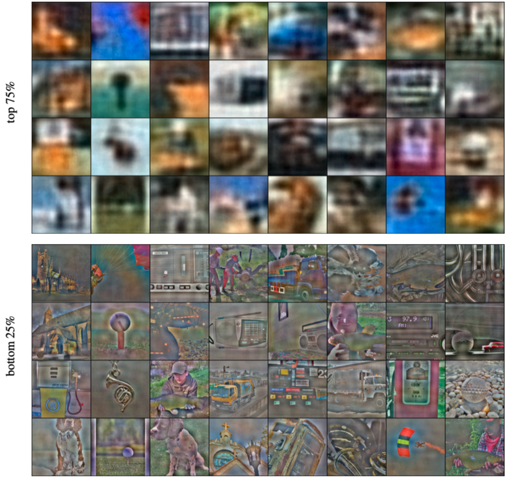


Figure: Figure from Balestriero & LeCun (ICML 2024)



Emperical non-linear case

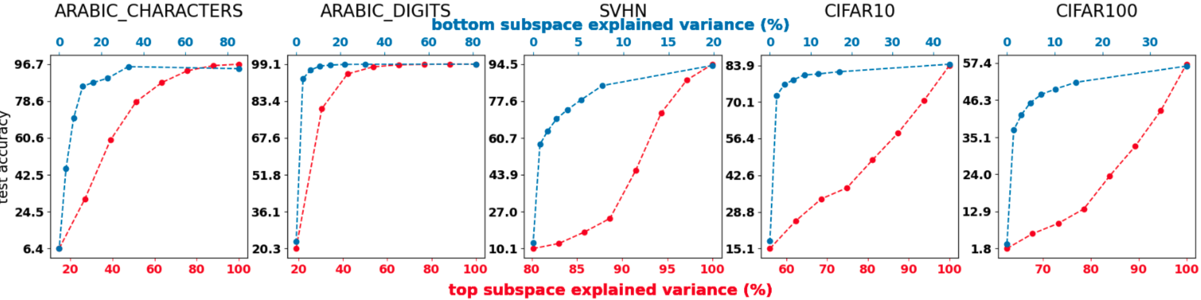
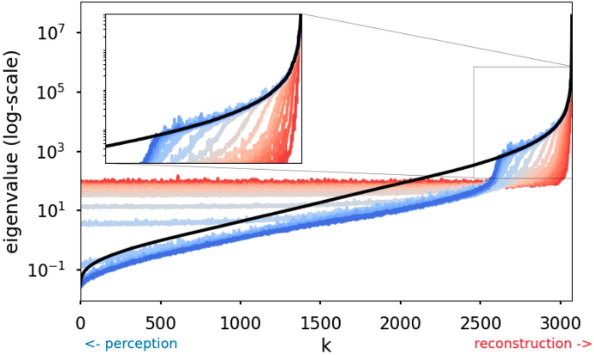


Figure: Figure from Balestriero & LeCun (ICML 2024)

Emperical non-linear case



bottom subspace is learned
exponentially slower than the top subspace

Figure: Figure from Balestriero & LeCun (ICML 2024)

F-principle in reconstuction

F-principle: DNNs fit target functions from low to high frequencies

Is this paper just a specific case of the F-principle?

How do other SSL methods perform well on downstream tasks while following the F-principle?

Classification accuracy variance

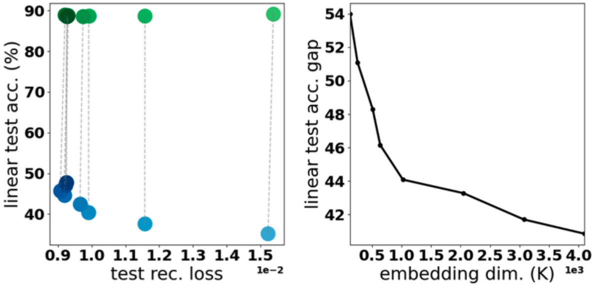


Figure: Figure from Balestriero & LeCun (ICML 2024)

Enforcing a DN to use the "informative" subspace has minimal impact on the reconstruction loss

ARCHIMEDES

Guiding reconstruction

$$
\begin{aligned}
\text{alignment}(k) &\triangleq \min_{\boldsymbol{W}} \left\| \boldsymbol{W}^\top \boldsymbol{V}^{*\top} \boldsymbol{X} - \boldsymbol{Y} \right\|_F^2, \\
\boldsymbol{V}^* &= \arg\min_{\boldsymbol{V}} \min_{\boldsymbol{Z}} \mathbb{E}_{\boldsymbol{X}'|\boldsymbol{X}} \left[\left\| \boldsymbol{Z}^\top \boldsymbol{V}^\top \boldsymbol{X}' - \boldsymbol{X} \right\|_F^2 \right],
\end{aligned}
$$

$$\text{alignment}(k) \triangleq \min_{\boldsymbol{W}} \left\| \boldsymbol{W}^\top \boldsymbol{V}^{*\top} \boldsymbol{X} - \boldsymbol{Y} \right\|_F^2,$$
$$\boldsymbol{V}^* = \arg\min_{\boldsymbol{V}} \min_{\boldsymbol{Z}} \mathbb{E}_{\boldsymbol{X}'|\boldsymbol{X}} \left[\left\| \boldsymbol{Z}^\top \boldsymbol{V}^\top \boldsymbol{X}' - \boldsymbol{X} \right\|_F^2 \right],$$

Theorem 3. The closed form solution for V^* for this problem is given by

$\boldsymbol{V}^*$ spans $\boldsymbol{P}_{\boldsymbol{G}} \boldsymbol{D}_{\boldsymbol{G}}^{-\frac{1}{2}} (\boldsymbol{P}_{\boldsymbol{H}})_{.,1:K}$,

where $\boldsymbol{H} \triangleq \boldsymbol{D}_{\boldsymbol{G}}^{-\frac{1}{2}} \boldsymbol{P}_{\boldsymbol{G}}^\top \boldsymbol{S} \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{S}^\top \boldsymbol{P}_{\boldsymbol{G}} \boldsymbol{D}_{\boldsymbol{G}}^{-\frac{1}{2}}$ and $G \triangleq \mathbb{E}_{\boldsymbol{X}'|\boldsymbol{X}} \left[\boldsymbol{X}' \boldsymbol{X}'^\top \right]$ and $\boldsymbol{S} \triangleq \mathbb{E}_{\boldsymbol{X}'|\boldsymbol{X}} \left[\boldsymbol{X}' \right]$

Guiding reconstruction

$$\text{alignment}(k) \triangleq \min_{\boldsymbol{W}} \left\| \boldsymbol{W}^\top \boldsymbol{V}^{*\top} \boldsymbol{X} - \boldsymbol{Y} \right\|_F^2,$$
$$\boldsymbol{V}^* = \arg\min_{\boldsymbol{V}} \min_{\boldsymbol{Z}} \mathbb{E}_{\boldsymbol{X}'|\boldsymbol{X}} \left[\left\| \boldsymbol{Z}^\top \boldsymbol{V}^\top \boldsymbol{X}' - \boldsymbol{X} \right\|_F^2 \right],$$

Theorem 3. The closed form solution for V^* for this problem is given by

$\boldsymbol{V}^*$ spans $\boldsymbol{P}_{\boldsymbol{G}} \boldsymbol{D}_{\boldsymbol{G}}^{-\frac{1}{2}} (\boldsymbol{P}_{\boldsymbol{H}})_{.,1:K}$,

where $\boldsymbol{H} \triangleq \boldsymbol{D}_{\boldsymbol{G}}^{-\frac{1}{2}} \boldsymbol{P}_{\boldsymbol{G}}^\top \boldsymbol{S} \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{S}^\top \boldsymbol{P}_{\boldsymbol{G}} \boldsymbol{D}_{\boldsymbol{G}}^{-\frac{1}{2}}$ and $G \triangleq \mathbb{E}_{\boldsymbol{X}'|\boldsymbol{X}} \left[\boldsymbol{X}' \boldsymbol{X}'^\top \right]$ and $\boldsymbol{S} \triangleq \mathbb{E}_{\boldsymbol{X}'|\boldsymbol{X}} \left[\boldsymbol{X}' \right]$

Corollary. Under these settings, **additive Gaussian noise has no impact in the downsteam task performance** as $\boldsymbol{W}^{*\top} \boldsymbol{V}^*(\sigma)^\top = \boldsymbol{W}^{*\top} \boldsymbol{V}^*(0)^\top, \forall \sigma \geq 0$, regardless of the supervised task.

Masking helps

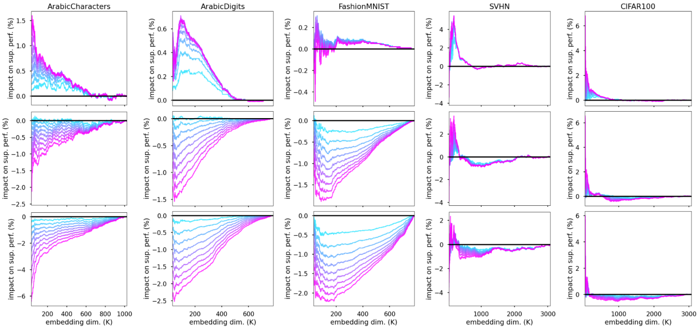


Figure 7. Depiction of the relative alignment difference when employing denoising tasks (recall Eq. (9)) with masking noise, with probability of dropping ranging from 0% to 99% (**cyan** to **pink**) for patch size of $(1, 1)$ recovering multiplicative dropout (**top**), $(2, 2)$ (**middle**), and $(4, 4)$ (**bottom**) on various datasets. A positive number indicates a beneficial impact of using the denoising loss on the supervised performance of the learned representation. We observe that for datasets such as ArabicDigits that already have a strong alignment between the two tasks (recall Fig. 2), the use of any form of masking is detrimental except with shape $(1, 1)$. However for datasets such as CIFAR100 (**right column**) with originally poor alignment, masking is beneficial and increases the alignment between the two tasks. As the original alignment increases with K, as the benefit of masking reduces.

Figure: Figure from Balestriero & LeCun (ICML 2024)

Conclusion

- **Misalignment** between learning by reconstruction and learning for downstream tasks

Conclusion

- **Misalignment** between learning by reconstruction and learning for downstream tasks
- **Ill-conditioned**: Perception features are learned last, requiring long training time

Conclusion

- **Misalignment** between learning by reconstruction and learning for downstream tasks
- **Ill-conditioned**: Perception features are learned last, requiring long training time
- **Ill-posed**: different model parameters can produce same reconstruction error but vastly different perception performance

ARCHIMEDES

Conclusion

- **Misalignment** between learning by reconstruction and learning for downstream tasks
- **Ill-conditioned**: Perception features are learned last, requiring long training time
- **Ill-posed**: different model parameters can produce same reconstruction error but vastly different perception performance
- **Masking helps**

Conclusion

- **Misalignment** between learning by reconstruction and learning for downstream tasks
- **Ill-conditioned**: Perception features are learned last, requiring long training time
- **Ill-posed**: different model parameters can produce same reconstruction error but vastly different perception performance
- **Masking helps**
- MAEs need:
  - large training times and architectures depending on image resolution, background etc
  - fine tuning

References

[1] Balestriero, R., & LeCun, Y. How Learning by Reconstruction Produces Uninformative Features For Perception. In Forty-first International Conference on Machine Learning.
[2] Xu, Z. Q. J., Zhang, Y., Luo, T., Xiao, Y., & Ma, Z. Frequency Principle: Fourier Analysis Sheds Light on Deep Neural Networks. Communications in Computational Physics
[3] Cao, Y., Fang, Z., Wu, Y., Zhou, D.-X., & Gu, Q.. Towards Understanding the Spectral Bias of Deep Learning. In Z.-H. Zhou, Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence.

Congrats Niko Bellic