Neural Network Interpretation Using Descrambler Groups: Do All Roads Lead to the SVD?

Shashank Sule, Richard Spencer, Wojciech Czaja

University of Maryland, College Park National Institute on Aging, Baltimore MD

Introduction

Attributing human-understandable explanations for the workings of a neural network is an active problem in deep learning. In [1] this problem is addressed through **Descrambling Transformations**. Descrambling proceeds as follows: fix input data X and *wiretap* the network output F at the kth layer with a matrix $P = \rho(g)$ where ρ is representation of the **descrambler** group G:

$$F = \sigma W_m \cdots \sigma P^{-1} P \underbrace{W_k \cdots W_1 X}_{f_k(X)} \qquad (1)$$

Then P **is computed such that** PW_k **is an interpretation of** W_k. *Smoothness criterion descrambling* is defined as computing P to promote smoothness of the intermediate data over $G = SO(d)$. Fixing D as a differentiation stencil, we write P as

$$P = \text{argmin}\, \|DQf_k(X)\|_F^2 \qquad \text{w.r.t}\, Q^\top Q = I \quad (2)$$

We characterize the properties of the minimizers in (2) and give three applications to deep learning in inverse problems.

Main Result

1. Let X be a random $d \times N$ matrix of N samples drawn from

$$x = s(z) + \alpha y \qquad (3)$$

where s is a signal of a parameter z distributed independently from a zero mean noise vector y with an isotropic density.

2. Let U_J be the left SVD of $Jf_k(\overline{X})$ where $\overline{X} := \mathbb{E}[X]$ and J is the Jacobian of f_k defined in (1).

3. Let T be the $m \times m$ matrix such that every kth column has lth entry $t_k(l)$ given by

$$t_k(l) = \begin{cases} \cos\frac{\pi l k}{m} & k \text{ even} \\ \sin\frac{\pi l (k+1)}{m} & k \text{ odd} \end{cases} \qquad (4)$$

4. Let T_r the submatrix of T given by picking the first r columns.

Let P be defined as (2). Then we have that

$$\|PU_J - T_r\|_F \leq \frac{K(d)}{\sqrt{N}} + C(\alpha, d)\|U_J\|_F \qquad (5)$$

Here $|C(\alpha, d)| \to 0$ as $\alpha \to \infty$. Furthermore, we obtain a number of simplifications when $k = 1$:

1. If $s \equiv K$ then $C(\alpha, d) = 0$. In this case

$$\lim_{N\to\infty} P = T_r U_1^\top \qquad (6)$$

2. In general for $k = 1$, $P \to$

$$\underset{Q^\top Q = I}{\text{argmin}}\, \mathbb{E}\left[\|DQW_1 s(z)\|_2^2 + \alpha^2 \|DPW_1 y\|_2^2\right] \qquad (7)$$

Application 1: Explaining DEERNet

DEERNet [2] is a fully-connected NN for solving the inverse problem of recovering p from the noisy measurements of Γ in deep electron-electron resonance (DEER):

$$\Gamma(t) = \int_\Omega p(r)\gamma(r,t)\,dx + \xi, \;\; \gamma(r,t) := \sqrt{\frac{\pi r^3}{6t}}\left[\cos[tr^{-3}]C\left[\sqrt{\frac{6t}{\pi r^3}}\right] + \sin[tr^{-3}]S\left[\sqrt{\frac{6t}{\pi r^3}}\right]\right] \qquad (8)$$

Here S and C are Fresnel Integrals. We present a descrambling analysis of DEERNet mirroring [1]. A potential advantage of descrambling over the SVD is the signal term in (7); however, this may render the descrambler analysis dependent upon specific data rather than just the NN itself. However, the SVD analysis shows that this is not the case since the network also learns copies of the data.

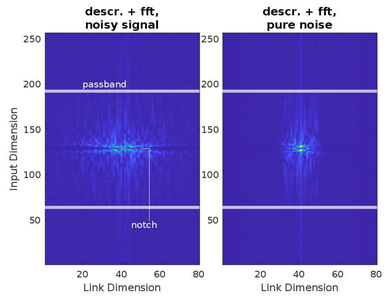

Figure 1: We visualize the 2-D FFT's of PW_1 where W_1 is the first weight matrix of the DEERNet [2] and P is computed from (2). Left: Amey et al. show that the descrambled first weight matrix DEERNet contains a notch filter and a bandpass filter on the Fourier Domain. Right Visualizing the 2-D DFT of the first weight matrix descrambled with noise recovers the bandpass filter but not the notch.

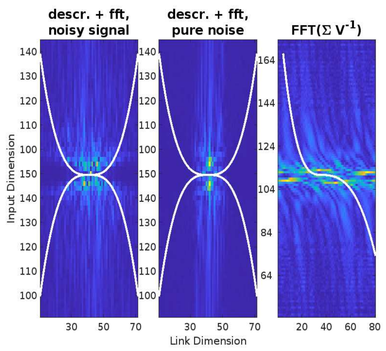

Figure 2: Left: The notch filter from Figure 1 vaguely corresponds to the cubic time-distance conversion in (8). Right: We show that this is much better inferred by directly looking at Fourier Transform of the right singular vectors of the weights.

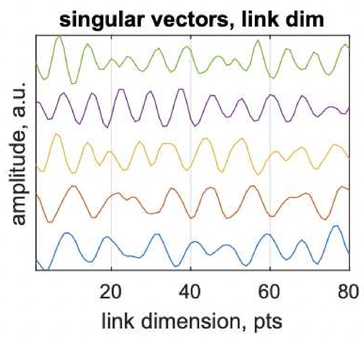

Figure 3: Figure 4 from [1]: Right singular vectors of the second weight matrix after descrambling are approximately sinusoidal. This is predicted by (5), which shows that the singular vectors on the descrambled side are close to an oscillating basis.

Application 2: When learning happens in the SVD

Figure 2 suggests that the singular vectors of the weights of a network can hold interpretable properties. We test and confirm this hypothesis on two classes of networks from magnetic resonance relaxometry [3] that solve a non-linear problem of recovering the rate parameters $(T_{2,1}, T_{2,2})$ from noisy samples of a biexponential curve $0.6\exp(-t/T_{2,1}) + 0.4\exp(-t/T_{2,2})$. In this case, the (ND, Reg) class is trained on a concatenation of noisy and smooth data (top rightmost column, Figure 4) and the (ND, ND) class is trained on a concatenation of two copies of noisy data (bottom rightmost column, Figure 4).

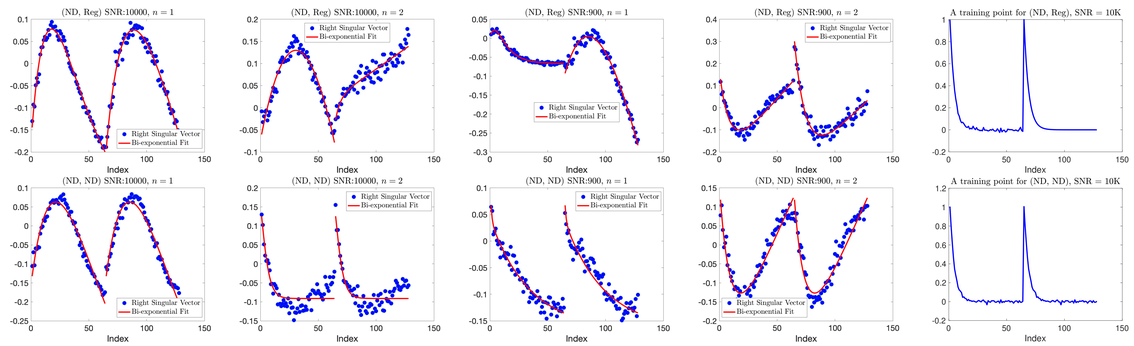

Figure 4: We find that the first layers weights of networks trained to process biexponential data exhibit biexponential singular vectors. Here n is the nth singular vector and SNR is α^{-2} from (7).

Application 3: Limitations of descrambling

We highlight two applications where the descrambling interpretation reveals information already available to us:

Corollary (CNN's)**.** In the main result, let x admit a zero-mean isotropic density and set $k = 1$. Suppose f is a 1-D convolutional neural network with stride 1. Then $PW \to W$ so descrambling does nothing to the weight matrix.

Corollary (oscillatory data analysis)**.** Let $z \sim \text{Unif}[-uN, vN]$ for $u, v \in \mathbb{Z}$ and $s(z) = (\exp 2\pi i k T/N)_{k=0}^{N-1}$. Then $P \to T_r U^\top$ where T_r is the trigonometric basis from the main result and U is the left SVD of the weight W. In this case, descrambling reveals information already available via the SVD.

Conclusions and Further Work

- Descrambling transformations interact with the SVD of the Jacobian of the network when trained on noisy samples of a smooth signal.
- Descramblers computed on highly noisy data reveal exactly the information available in the SVD. Moreover, descramblers P can represent artefacts of the data which may have nothing to do with the network.
- However, we find that this is not the case for two networks trained to solve the inverse problem of signal recovery. In fact, the networks themselves learn forms of the data in the right singular vectors.
- SVD's have been used for network compression and deep interpretation [4]. But singular vector learning of training data and its relationship to generalization is an unexplained phenomenon and is a current research direction.

References

[1] Jake L Amey, Jake Keeley, Tajwar Choudhury, and Ilya Kuprov. "Neural network interpretation using descrambler groups". In: *Proceedings of the National Academy of Sciences* 118.5 (2021).

[2] Steven G Worswick, James A Spencer, Gunnar Jeschke, and Ilya Kuprov. "Deep neural network processing of DEER data". In: *Science advances* 4.8 (2018), eaat5218.

[3] Michael Rozowski, Jay Bisen, Chuan Bi, Wojciech Czaja, and Richard Spencer. "Neural Network Analysis of Biexponential Decay Curves Using Regularized Input Data". In: *Bulletin of the American Physical Society* 66 (2021).

[4] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. "Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability". In: *Advances in neural information processing systems* 30 (2017).