Neural Collapse in Deep Homogeneous Classifiers and the role of Weight Decay

Akshay Rangamani, Andrzej Banburski

Center for Brains, Minds, Machines, MIT

Paper ID: 4758

Modern Deep Learning Practice and the Terminal Phase of Training

- These days we use networks that are **massively overparameterized**

Modern Deep Learning Practice and the Terminal Phase of Training

- These days we use networks that are **massively overparameterized**
- Networks are usually trained to fit the training set (nearly **zero training error**)

Modern Deep Learning Practice and the Terminal Phase of Training

- These days we use networks that are **massively overparameterized**
- Networks are usually trained to fit the training set (nearly **zero training error**)
- Often, they are trained for longer, beyond zero training error, to go towards **zero training loss**

Modern Deep Learning Practice and the Terminal Phase of Training

- These days we use networks that are **massively overparameterized**
- Networks are usually trained to fit the training set (nearly **zero training error**)
- Often, they are trained for longer, beyond zero training error, to go towards **zero training loss**
- This is what we call the **Terminal Phase of Training**

Modern Deep Learning Practice and the Terminal Phase of Training

- These days we use networks that are **massively overparameterized**
- Networks are usually trained to fit the training set (nearly **zero training error**)
- Often, they are trained for longer, beyond zero training error, to go towards **zero training loss**
- This is what we call the **Terminal Phase of Training**

- What happens at the end of this? (Focus on the last layer)

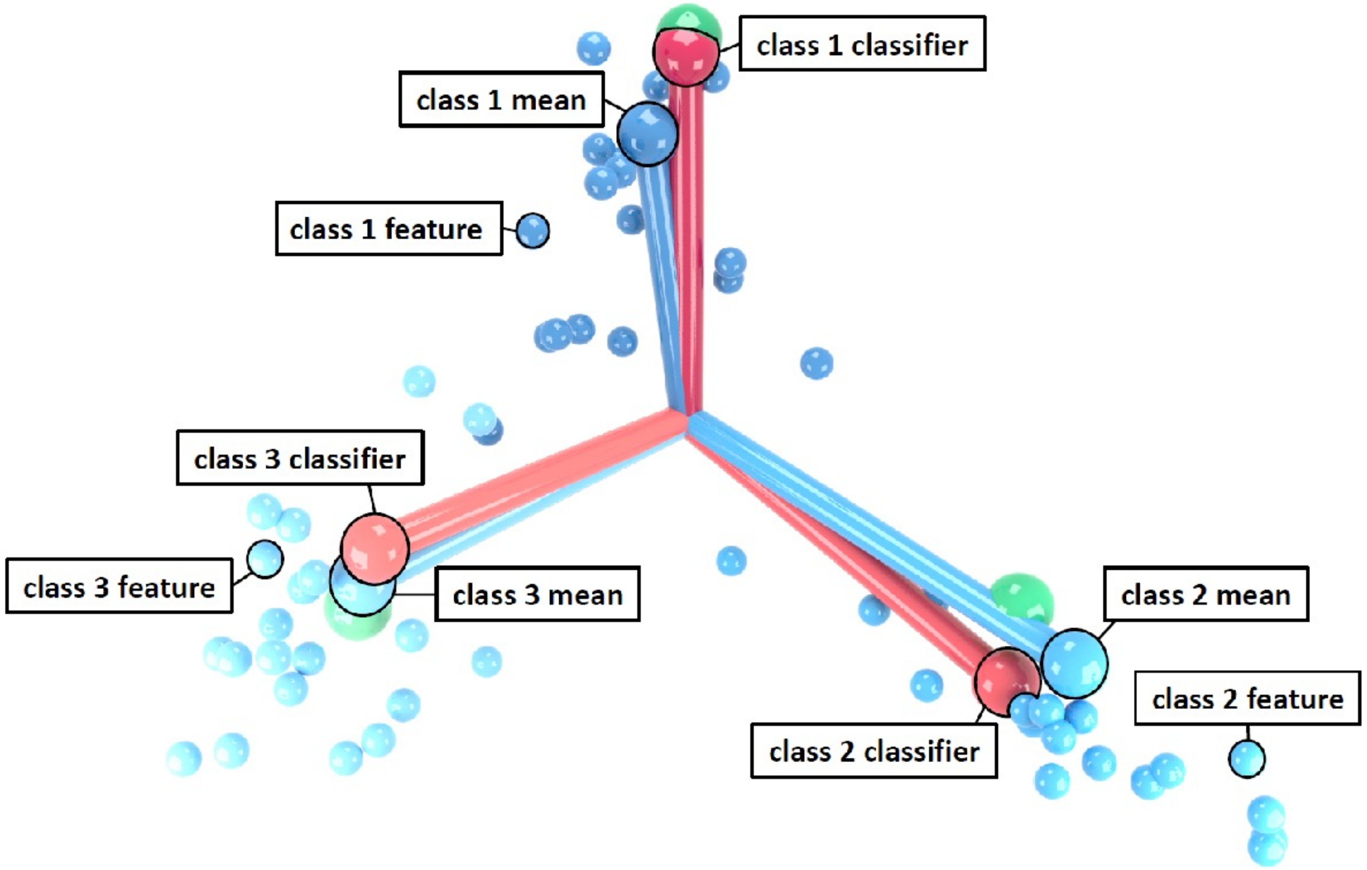
Some Notation

- We have a classification problem with C classes
- Inputs – $\boldsymbol{x}_{n(c)}, n = 1 \dots N, c = 1 \dots C$
- Deep Network divided into features and classifier: $f_{\boldsymbol{W}}(\boldsymbol{x}) = \boldsymbol{W}_L \boldsymbol{h}(\boldsymbol{x})$
- Last layer features: $\boldsymbol{h}_{n(c)}, n = 1 \dots N, c = 1 \dots C$
- Classifiers: $\boldsymbol{W}_L^c, \quad c = 1 \dots C$
- Classification rule: $\mathrm{argmax}_{c'} \left\langle \boldsymbol{W}_L^{c'}, \boldsymbol{h}(\boldsymbol{x}) \right\rangle$

More Notation

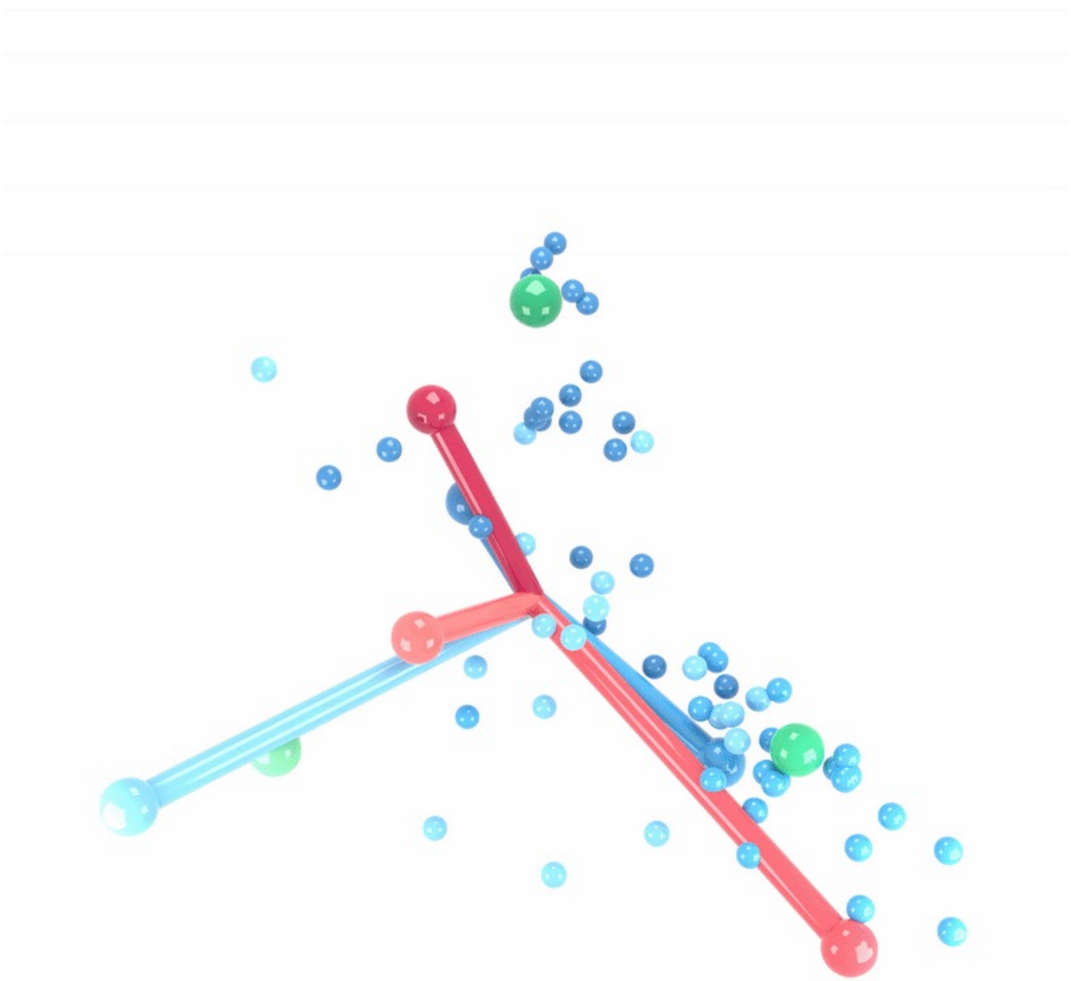
- Statistics of the last layer features:
    - Global mean: $\boldsymbol{\mu}_G = \frac{1}{NC}\sum_{n=1,c=1}^{N,C} \boldsymbol{h}_{n(c)}$
    - Class means: $\boldsymbol{\mu}_c = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{h}_{n(c)}\,, c = 1 \ldots C$
    - Within class covariance: $\Sigma_W = \frac{1}{NC}\sum_{n=1,c=1}^{N,C}\left(\boldsymbol{h}_{n(c)} - \boldsymbol{\mu}_c\right)\left(\boldsymbol{h}_{n(c)} - \boldsymbol{\mu}_c\right)^T$
    - Between class covariance: $\Sigma_B = \frac{1}{C}\sum_{c=1}^{C}(\boldsymbol{\mu}_c - \boldsymbol{\mu}_G)(\boldsymbol{\mu}_c - \boldsymbol{\mu}_G)^T$

Illustration



Papyan, V., Han, X. Y., & Donoho, D. L. (2020). Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, *117*(40), 24652-24663.

Neural Collapse



Papyan, V., Han, X. Y., & Donoho, D. L. (2020). Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, *117*(40), 24652-24663.

Neural Collapse

(NC1) Variability Collapse: All last layer features are the same within the same class $\left(\boldsymbol{h}\left(\boldsymbol{x}_{\boldsymbol{n}(\boldsymbol{c})}\right) = \ \boldsymbol{\mu}_{\boldsymbol{c}}\right)$

(NC2) Equinorm and Equiangularity of the centered class means

$$\left|\left|\boldsymbol{\mu}_{\boldsymbol{c}} - \boldsymbol{\mu}_{\boldsymbol{G}}\right|\right|_{\boldsymbol{2}} = \left|\left|\boldsymbol{\mu}_{\boldsymbol{c}'} - \boldsymbol{\mu}_{\boldsymbol{G}}\right|\right|_{\boldsymbol{2}};\ \langle\boldsymbol{\mu}_{\boldsymbol{c}} - \boldsymbol{\mu}_{\boldsymbol{G}}, \boldsymbol{\mu}_{\boldsymbol{c}'} - \boldsymbol{\mu}_{\boldsymbol{G}}\rangle\ \propto \frac{\boldsymbol{C}}{\boldsymbol{C}-\boldsymbol{1}}\boldsymbol{\delta}_{\boldsymbol{c},\boldsymbol{c}'} - \frac{\boldsymbol{1}}{\boldsymbol{C}-\boldsymbol{1}}$$

(NC3) Self Duality of classifier and last layer features $\frac{\boldsymbol{W}^{\boldsymbol{T}}}{\left|\left|\boldsymbol{W}\right|\right|_{\boldsymbol{F}}} = \frac{\dot{\boldsymbol{M}}}{||\dot{\boldsymbol{M}}||_{\boldsymbol{F}}}$

(NC4) Nearest Class Center Classification

$$\text{argmax}_{c'}\left\langle \boldsymbol{W}_L^{c'}, \boldsymbol{h}(\boldsymbol{x})\right\rangle = \text{argmin}_{c'}\left|\left|\boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{\mu}_{\boldsymbol{c}}\right|\right|$$

Neural Collapse in Deep Homogeneous Models

Deep Homogeneous Networks

In this work we study Deep Homogeneous ReLU (rectified linear unit) networks trained with the square loss

$$f_{\boldsymbol{W}}(\boldsymbol{x}) = \boldsymbol{W}_L\sigma\big(\boldsymbol{W}_{L-1}\dots\boldsymbol{W}_2\sigma(\boldsymbol{W}_1\boldsymbol{x})\big)$$

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{2NC}\sum_{n,c,i=1}^{N,C,C}\left(\boldsymbol{y}_{n(c)}^{(i)} - f_{\boldsymbol{W}}^{(i)}\left(\boldsymbol{x}_{n(c)}\right)\right)^2 + \frac{\lambda}{2}\sum_k ||\boldsymbol{W}_k||_F^2$$

Neural Collapse Solutions Cannot Interpolate

Lemma 1: For Deep Homogeneous networks trained on the unregularized square loss ($\lambda = 0$), solutions that exhibit Neural Collapse do not interpolate the training data and hence are not global minima

Proof sketch: Let $\boldsymbol{H} \in \mathbb{R}^{p\times NC}$ and $\boldsymbol{Y} \in \mathbb{R}^{C\times NC}$ be the matrices of last layer features and one-hot labels respectively

(NC1) means that $\boldsymbol{H}$ can be factorized as $\boldsymbol{H} = \boldsymbol{MY}$, where $\boldsymbol{M} \in \mathbb{R}^{p\times C}$ is a matrix whose columns are the class means (assume that the global mean $\boldsymbol{\mu_G} = \boldsymbol{0}$)

(NC2, NC3) means that $\boldsymbol{WM} = \frac{\alpha C}{C-1}(\boldsymbol{I} - \frac{1}{C}\boldsymbol{1}\boldsymbol{1}^T)$, which means

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{2NC}||\boldsymbol{WMY} - \boldsymbol{Y}||_F^2 = \frac{1}{2}\left((1-\alpha)^2 + \frac{\alpha^2}{C-1}\right) > 0$$

Weight Decay Prevents Interpolation

Lemma 2: The global minima of the regularized square loss of a deep homogeneous ReLU networks do not interpolate the training data

Proof Sketch: We can rewrite the loss using gradient flow on the parameters of the network

$$\mathcal{L}(\boldsymbol{W}) \\ = -\frac{1}{4}\frac{\partial \left| |\boldsymbol{W}_l| \right|^2}{\partial t} + \frac{\lambda}{2}\sum_{k,k\neq l}\left| |\boldsymbol{W}_k| \right|_F^2 + \frac{1}{2NC}\sum_{n=1,c=1}^{N,C} \left| |\boldsymbol{y}_{n(c)}| \right|_2^2 - \left\langle \boldsymbol{y}_{n(c)}, f_{\boldsymbol{W}}(\boldsymbol{x}_{n(c)}) \right\rangle$$

We arrive at a contradiction for interpolating critical points by evaluating the loss using the above expression as well as the direct evaluation

$$\mathcal{L} = \frac{\lambda}{2}\sum\nolimits_k \left| |\boldsymbol{W}_k| \right|_F^2 \qquad \mathcal{L} = \frac{\lambda}{2}\sum\nolimits_{k,k\neq l} \left| |\boldsymbol{W}_k| \right|_F^2$$

Symmetric Quasi-Interpolation

Since Deep Homogeneous networks trained with the regularized square loss cannot interpolate the training data, we make the following assumption about the interpolation errors (this assumption is similar to label-smoothing)

Assumption: *For a C-class classification problem, a classifier $f: \mathbb{R}^d \to \mathbb{R}^C$ symmetrically quasi-interpolates a training dataset if for all training examples $\boldsymbol{x}_{n(c)}$ in class c,* $f^{(c)}\left(\boldsymbol{x}_{n(c)}\right) = 1 - \epsilon, f^{(c')}\left(\boldsymbol{x}_{n(c)}\right) = \frac{\epsilon}{C-1}$

Main Result

Theorem: *For a ReLU deep homogeneous network trained on a balanced dataset with the regularized square loss ($\lambda \neq 0$), critical points of gradient flow that satisfy symmetric quasi-interpolation also satisfy the conditions* ***(NC1-4)*** *for Neural Collapse*

Proof Sketch

(NC1) follows from the fact that $f_{\boldsymbol{W}}(\boldsymbol{x}_{n(c)})$ does not depend on n.

(NC3) can be shown by simple algebra from the gradient flow equilibrium condition (setting $\frac{\partial \boldsymbol{W}_L^c}{\partial t} = 0$) . We obtain $\boldsymbol{W}_L^c = \frac{\epsilon}{\lambda(C-1)} \times (\boldsymbol{\mu_c} - \boldsymbol{\mu_G})$

(NC2 Equinorm) can be shown by computing $\left\langle \boldsymbol{W}_L^c, \frac{\partial \boldsymbol{W}_L^c}{\partial t} \right\rangle$, and setting $\frac{\partial \boldsymbol{W}_L^c}{\partial t} = 0$ to obtain $\left|\left|\boldsymbol{W}_L^c\right|\right|^2 = \frac{1}{C\lambda}\left(\epsilon - \frac{C}{C-1}\epsilon^2\right)$

(NC2 Equiangularity) can be shown from the symmetric quasi-interpolation condition, and the duality between $\boldsymbol{W}_L^c$ and $\boldsymbol{\mu_c} - \boldsymbol{\mu_G}$.

We obtain $\frac{\left\langle \boldsymbol{W}_L^c, \boldsymbol{W}_L^{c'} \right\rangle}{\left|\left|\boldsymbol{W}_L^c\right|\right| \left|\left|\boldsymbol{W}_L^{c'}\right|\right|} = -\frac{1}{C-1}$

Experiments

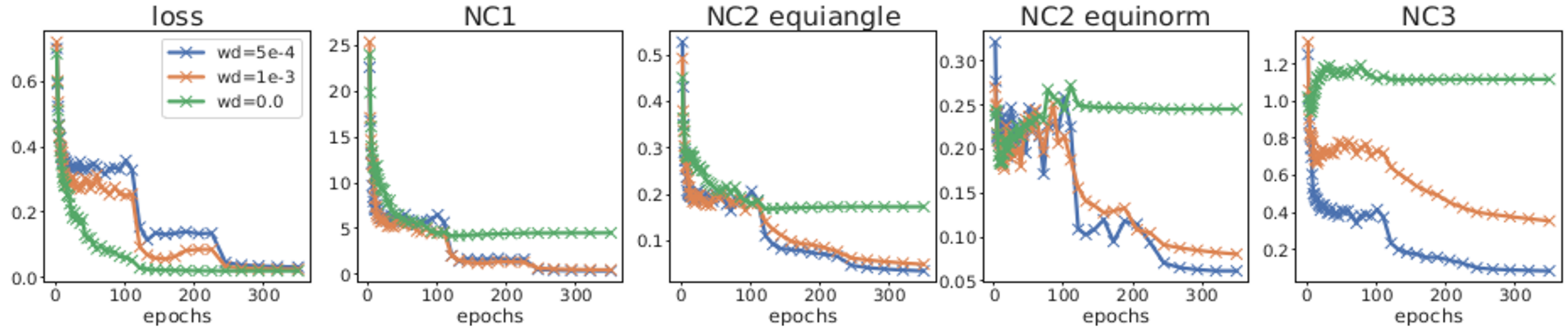


Fig. 1: Measurement of Neural Collapse (NC) on CIFAR10 indicates that it emerges in the presence of weight decay. Exact description of quantities measured can be found in Section 3. Green lines (no weight decay) do not exhibit NC while orange and blue (weight decay) do.

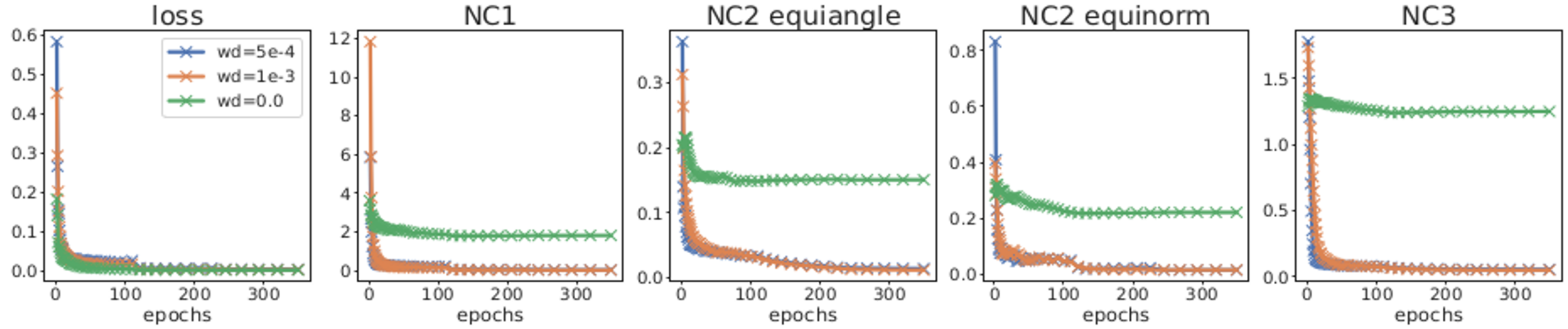


Fig. 2: Measurement of Neural Collapse on MNIST. Neural Collapse is even more clear in the presence of weight decay.

Summary

- Training deep networks beyond data separation leads to Neural Collapse which dramatically simplifies the model of a deep network
- This suggests an intriguing model for a deep classifier – features become templates at the last layer, and we do template matching.
- Weight Decay is necessary for neural collapse to occur in homogeneous models
- Future work:
  - Moving beyond our assumption of symmetric quasi-interpolation
  - Bounding NC measurements in terms of deviations from the ETF structure rather than exactly matching the ETF structure

References

[1] V. Papyan, X. Han, and D. L. Donoho, "Prevalence of neural collapse during the terminal phase of deep learning training,"Proceedings of the National Academy of Sciences, vol. 117, no. 40, pp. 24 652–24 663, 2020

[2] A. Rangamani, M. Xu, A. Banburski, Q. Liao, T. Poggio, "Dynamics and Neural Collapse in Deep Classifiers with the Square Loss," CBMM Memo 117, 2021

[3] T. Poggio and Q. Liao, "Generalization in deep network classifiers trained with the square loss," Center for Brains, Minds and Machines (CBMM) Memo No. 112, 2021

[4] Like Hui and Mikhail Belkin, "Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks," arXiv preprint arXiv:2006.07322, 2020.

[5] Cong Fang, Hangfeng He, Qi Long, and Weijie J. Su, "Layer-peeled model: Toward understanding well-trained deep neural networks," CoRR, vol. abs/2101.12699, 2021.

[6] Dustin G. Mixon, Hans Parshall, and Jianzong Pi, "Neural collapse with unconstrained features," CoRR, vol. abs/2011.11619, 2020.

[7] X. Y. Han, Vardan Papyan, and David L. Donoho, "Neural collapse under MSE loss: Proximity to and dynamics on the central path," CoRR, vol. abs/2106.02073, 2021.

[8] Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu, "A geometric analysis of neural collapse with unconstrained features," 2021.

[9] Jianfeng Lu and Stefan Steinerberger, "Neural collapse with cross-entropy loss," CoRR, vol. abs/2012.08465, 2020.

[10] Stephan Wojtowytsch et al., "On the emergence of tetrahedral symmetry in the final and penultimate layers of neural network classifiers," arXiv preprint arXiv:2012.05420, 2020.

[11] Tolga Ergen and Mert Pilanci, "Revealing the structure of deep neural networks via convex duality," arXiv preprint arXiv:2002.09773, 2020.

Thank You!