NEURAL INFORMATION PROCESSING SYSTEMS

The Implicit Bias of Gradient Descent toward Collaboration between Layers: A Dynamic Analysis of Multilayer Perceptions

Zheng Wang[1], Geyong Min[1], Wenjie Ruan[2]

[1]University of Exeter [2]University of Science and Technology of China

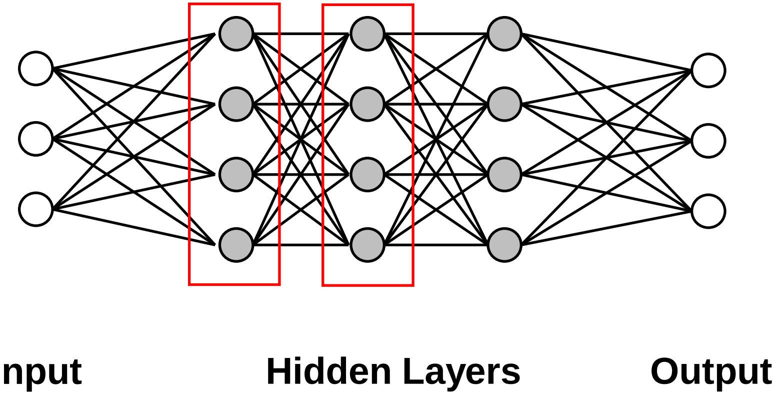
github.com/squarewang2077/co-correlation

Motivation

Whether layers in neural networks collaborate to strengthen adversarial robustness during gradient descent?

Adversarial Example

$x + \delta$



Misled output

$h(x+\delta) \neq h(x) = y_{label}$

Problem Setting

- **Binary classification**
- **Training Set** $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$

Neural Networks

$$f_{linear}(\boldsymbol{x}, W) = \boldsymbol{a}^T (W\boldsymbol{x})$$

$$f_{mlp}(\boldsymbol{x}, W) = \boldsymbol{a}^T (\sigma(W\boldsymbol{x}))$$

Activation function

Prediction

$$y_{pred} = \begin{cases} 1, sig(f_W(\boldsymbol{x})) > 0.5 \\ 0, sig(f_W(\boldsymbol{x})) \leq 0.5 \end{cases}$$

Sigmoid function

$$sig(x) = 1/(1 + e^{-x})$$

Loss function

$$L(f, y) = -\frac{1}{n} \sum_{i=1}^{n} \left[y_i \log(sig(u_i)) + (1 - y_i) \log(1 - sig(u_i)) \right]$$

Outputs

$$u_i = f(\boldsymbol{x}_i, W), i \in [n]$$

Measure Adversarial Risk by Dirichlet Energy

Theorem 4.1. *Given data points $(\boldsymbol{x}, y) \sim P$ and $\boldsymbol{x} \sim P_{\boldsymbol{x}}$, the relationship between adversarial risk and Dirichlet energy for classifier f with differentiable loss function L is shown as*

$$R^{rob}(f, r) \lesssim R(f) + r\mathfrak{S}(L(f)), \tag{8}$$

where $r > 0$ is the largest perturbation budget and $\mathfrak{S}(L(f)) = \sqrt{\mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}\left[\|\nabla_{\boldsymbol{f}} L^T \cdot J_{\boldsymbol{f}}(\boldsymbol{x})\|_2^2\right]}$ indicating the Dirichlet energy of the classifier on loss L.

$$R^{rob}(f, r) = \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim P}\left[\sup_{\boldsymbol{\varepsilon}\in B_r} L(f(\boldsymbol{x}+\boldsymbol{\varepsilon}), y))\right]$$

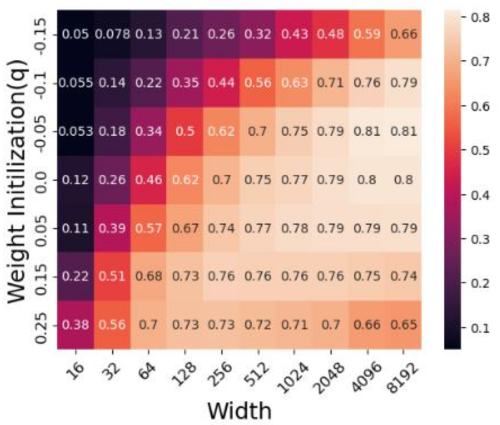
$$R(f) = \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim P}[L(f(\boldsymbol{x}), y))]$$

$$B_r = \{\|\boldsymbol{\varepsilon}\|_2 \leq r\}$$

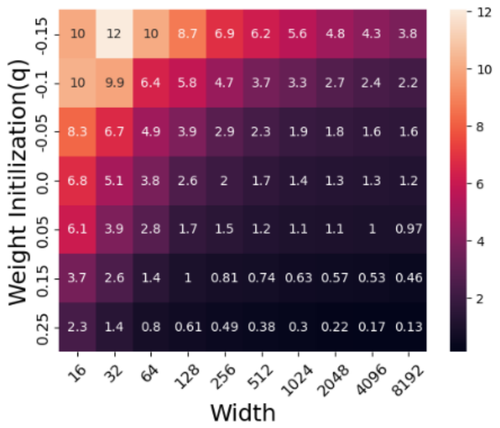
- **The proof is based on 1st order Taylor's expansion**

UNIVERSITY OF EXETER

Measure Adversarial Risk by Dirichlet Energy



(a) Robust Acc.



(b) $\mathfrak{S}(f)$

- **2-Layer MLPs with width from 2^4 to 2^{13}**
- **Initialize the weight matrix $w_{\{i,j\}} \sim N\left(0, \frac{1}{m^{1+2q}}\right)$**
- **Let q change from -0.15 to 0.25**
- **Dirichlet Energy of f can be a good representation**
- **It can measure individual layers therefore the correlations**

On Dynamics of Co-Correlation

Robustness Decomposition

Theorem 4.5 (Robustness Decomposition). *Given the same assumption in Definition 4.2, the measurement for overall adversarial robustness can be decomposed as*

$$\mathfrak{S}(\phi \circ \varphi) = \left(\mathbb{E}_{\boldsymbol{x}\sim P}[\|J_{\phi\circ\varphi}(\boldsymbol{x})\|_2^2]\right)^{\frac{1}{2}} = \varrho_{\phi,\varphi}\left(1 + \frac{var_{\phi,\varphi}}{\mu^2_{\phi,\varphi}}\right)^{\frac{1}{2}} \rho_{\phi,\varphi}\mathfrak{S}(\phi)\mathfrak{S}(\varphi) \qquad (15)$$

$$f_{linear}(\boldsymbol{x}, W) = \boldsymbol{a}^T(W\boldsymbol{x})$$

$$f_{mlp}(\boldsymbol{x}, W) = \boldsymbol{a}^T(\sigma(W\boldsymbol{x}))$$

Co-correlation

$$\varrho_{\phi,\varphi} \triangleq \frac{\left(\mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}\|J_{\phi\circ\varphi}(\boldsymbol{x})\|_2^2\right)^{\frac{1}{2}}}{\left(\mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}[\|J_{\phi}(\varphi)\|_2^2 \cdot \|J_{\varphi}(\boldsymbol{x})\|_2^2]\right)^{\frac{1}{2}}}$$

$$var_{\phi,\varphi} \triangleq Var_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}[\|J_{\phi}(\varphi)\|_2 \cdot \|J_{\varphi}(\boldsymbol{x})\|_2]$$

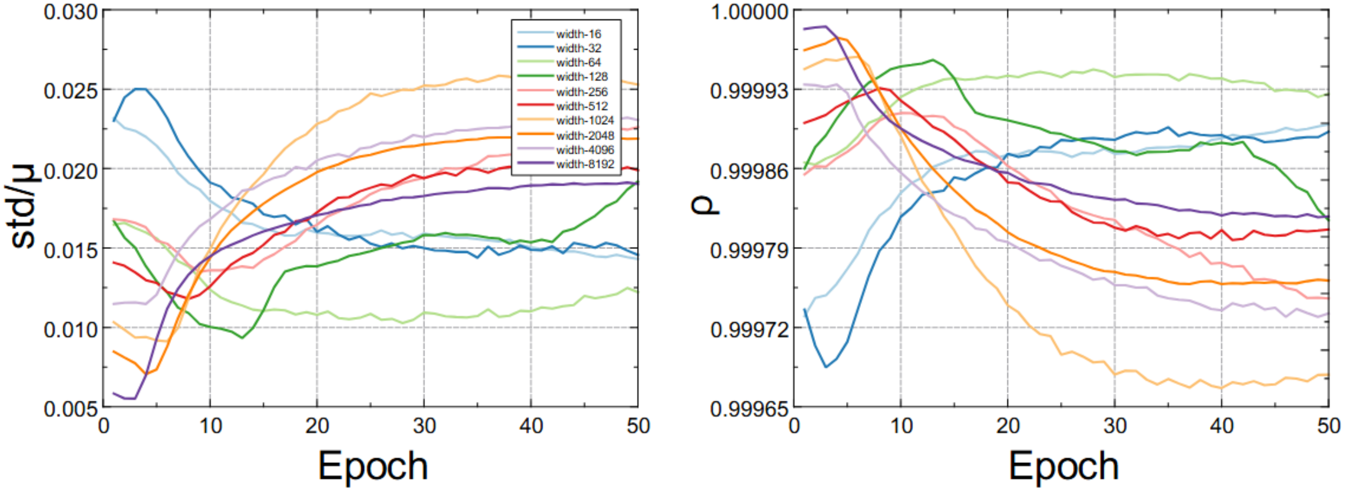
$$\rho_{\phi,\varphi} \triangleq \frac{\mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}[\|J_{\phi}(\varphi)\|_2 \cdot \|J_{\varphi}(\boldsymbol{x})\|_2]}{\left(\mathbb{E}_{\varphi\sim\varphi(\boldsymbol{x})}\|J_{\phi}(\varphi)\|_2^2\right)^{\frac{1}{2}}\left(\mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}\|J_{\varphi}(\boldsymbol{x})\|_2^2\right)^{\frac{1}{2}}}$$

$$\mu_{\phi,\varphi} \triangleq \mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}}[\|J_{\phi}(\varphi)\|_2 \cdot \|J_{\varphi}(\boldsymbol{x})\|_2]$$

- **Regard the *neural network* as *function composition***
- **The proof is straight forward**

On Dynamics of Co-Correlation

Robustness Decomposition



- **Empirically, co-correlation is more influential**
- **We focus on the co-correlation**

UNIVERSITY OF EXETER

On Dynamics of Co-Correlation

Linear Model

Assumption 5.1. We assume that each element $w_{i,j}$ in the weight matrix $W(0) \in \mathbb{R}^{m \times d}$ at initialization follows the Gaussian distribution $N(0, \frac{1}{m^{1+2q}})$, with $q > 0$. Additionally, each element $a_r, r \in [m]$ in $\boldsymbol{a}$ is randomly selected from the set $\{-\frac{1}{\sqrt{m}}, \frac{1}{\sqrt{m}}\}$, and fixed during training.

Assumption 5.2. We assume that for each $(\boldsymbol{x}_i, y_i) \in D, i \in [n]$, $\boldsymbol{x}_i$ is L_2 norm bounded such that $\|\boldsymbol{x}_i\|_2 = 1$ for all $i \in [n]$.

Theorem 5.3 (Dynamics of the Co-correlation for Linear Model). *Given the linear model defined in Equation (3a) and training dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. Assume that assumptions 5.1 and 5.2 hold for W and $\boldsymbol{a}$. The gradient descent applied to the weights results in the dynamics of the co-correlation being expressed as:*

$$\dot{\varrho}_{\boldsymbol{a},W}(t) = \eta C(t) \varrho_{\boldsymbol{a},W}, \tag{18}$$

and with high probability,

$$C(t) \geq \frac{\sum_{\tau=1}^{t} \widetilde{\boldsymbol{x}}(\tau)^T \widetilde{\boldsymbol{x}}(t)}{\|W(t)\|_2^2} \cdot \left(1 - \left(\boldsymbol{v}(t)^T \boldsymbol{a}\right)^2\right) + \mathcal{O}\left(\frac{1}{m^q}\right) \tag{19}$$

where the $\boldsymbol{v}(\boldsymbol{t})$ is the dominate eigenvector for $W(t)W(t)^T$.

When m is sufficiently large, and during the initial steps of the optimization process, $\widetilde{\boldsymbol{x}}(\tau), \tau \in [t]$ are quite similar to each other in terms of cosine similarity, implying an acute angle to each other, which leads to $\sum_{\tau=1}^{t} \widetilde{\boldsymbol{x}}(\tau)^T \widetilde{\boldsymbol{x}}(t) \geq 0$. As a result, we can conclude that $C(t) \geq 0$.

- **$\varrho_{a,W}$ increase during the initial stages and become saturated to its later stages.**
- ***The speed of the accumulation of $\varrho_{a,W}$ is inversely related to $\|W(t)\|_2$***

$$\widetilde{\boldsymbol{x}}(t) = \frac{1}{n} \sum_{i=1}^{n} \left[y_i - sig\left(u_i(t)\right)\right] \boldsymbol{x}_i$$

On Dynamics of Co-Correlation

2-Layer MLP

Assumption 5.4. The derivative of the activation function $\sigma'(x)$ in non-linear neural networks is bounded by M. In other words, we have $|\sigma'(x)| \leq M$.

Theorem 5.5. *(Dynamics of the Co-correlation for MLP) Given the MLP defined in Equation (3) with training dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, $\boldsymbol{x} \in \mathcal{X}$ such that $\boldsymbol{x} \sim P_{\boldsymbol{x}}$. Assume that Assumption 5.1 and 5.2 hold for W and $\boldsymbol{a}$, and Assumption 5.4 holds for the activation function. we have*

$$\dot{\varrho}_{\boldsymbol{a},\sigma\circ W}(t) = \eta C(t) \varrho_{\boldsymbol{a},\sigma\circ W}(t).$$

With high probability,

$$C(t) \geq \frac{\sum_{\tau=1}^{t} \left(1 - \boldsymbol{a}^T \boldsymbol{v}(\tau) \boldsymbol{a}^T \boldsymbol{v}(t)\right) \mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}} \left[\widetilde{\boldsymbol{x}}_*^T(\tau) \widetilde{\boldsymbol{x}}_*(t)\right]}{\mathbb{E}_{\boldsymbol{x}\sim P_{\boldsymbol{x}}} \|D(t)W(t)\|_2^2} + \max\left\{\mathcal{O}\left(\frac{1}{\sqrt{m}}\right), \mathcal{O}\left(\frac{1}{m^q}\right)\right\},$$

where

$$D(t) = diag(\sigma'(\boldsymbol{w}_1(t)^T \boldsymbol{x}), \cdots, \sigma'(\boldsymbol{w}_m(t)^T \boldsymbol{x})),$$

and $\boldsymbol{v}(t)$ denotes the dominant eigenvector for $W(t)W(t)^T$, with $\widetilde{\boldsymbol{x}}_*^T$ is defined in Equation (21). Similar to the Theorem 5.3, when m is sufficiently large, and during the initial steps of the optimization where the error-weighted inputs $\widetilde{\boldsymbol{x}}_*^T(\tau), \tau \in [t]$ do not significantly fluctuate, we have that $C(t) \geq 0$.

- **The dynamics for $\varrho_{a,\sigma\circ W}$ is the same to $\varrho_{a,W}$**
- ***The speed of the accumulation of $\varrho_{a,\sigma\circ W}$ is inversely related to $\|D(t)W(t)\|_2$***

Serve as similar purpose of $\widetilde{x}(t)$

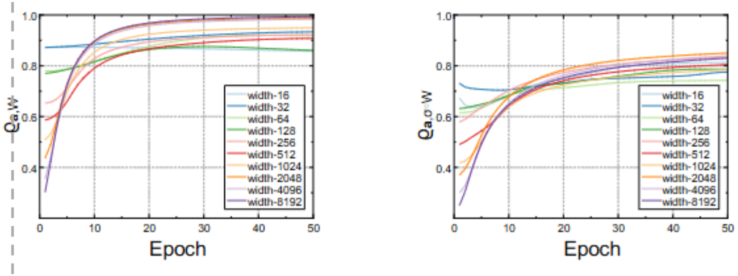
$$\alpha_i(t, \boldsymbol{x}) \triangleq \mathbb{E}_{W(0)}\left[\sigma'(\boldsymbol{w}(t)^T \boldsymbol{x}) \sigma'(\boldsymbol{w}(t)^T \boldsymbol{x}_i)\right]$$

$$\widetilde{\boldsymbol{x}}_*(t) \triangleq \frac{1}{n} \sum_{i=1}^{n} \alpha_i(t, \boldsymbol{x}) (y_i - sig(u_i(t))) \boldsymbol{x}_i$$

Experiments

2-Layer MLP

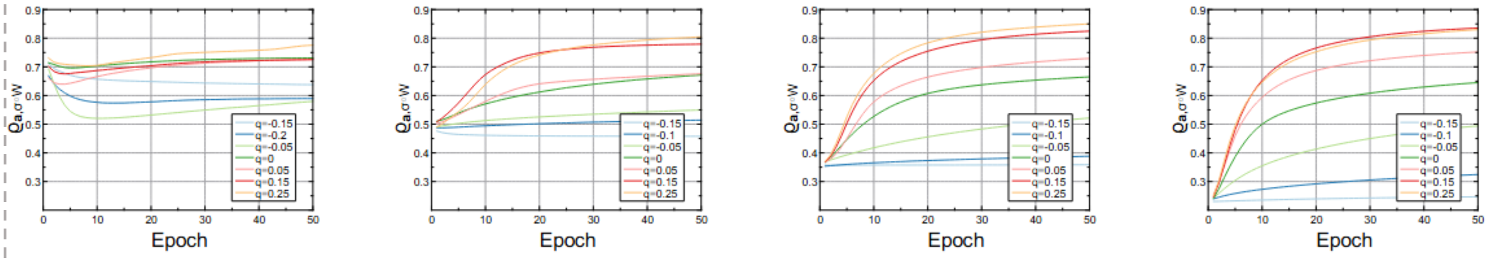
Different width



(c) Linear Model (d) MLP

- ***The upward trend is true for all 2-Layer MLPs***
- ***The theorem is quite tight on*** q

Different weight initialization parameter q

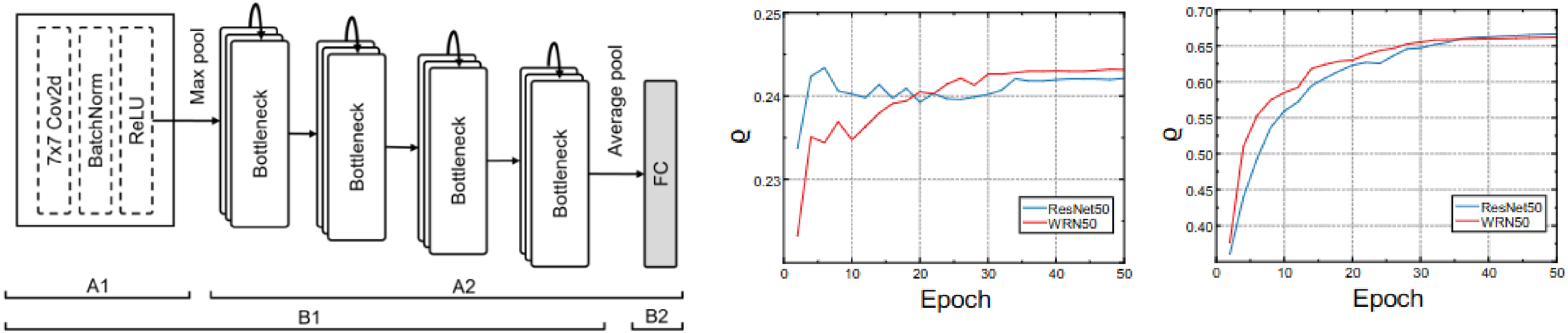


(a) width-32 (b) width-512 (c) width-2048 (d) width-8192

UNIVERSITY OF EXETER

Experiments

ResNets

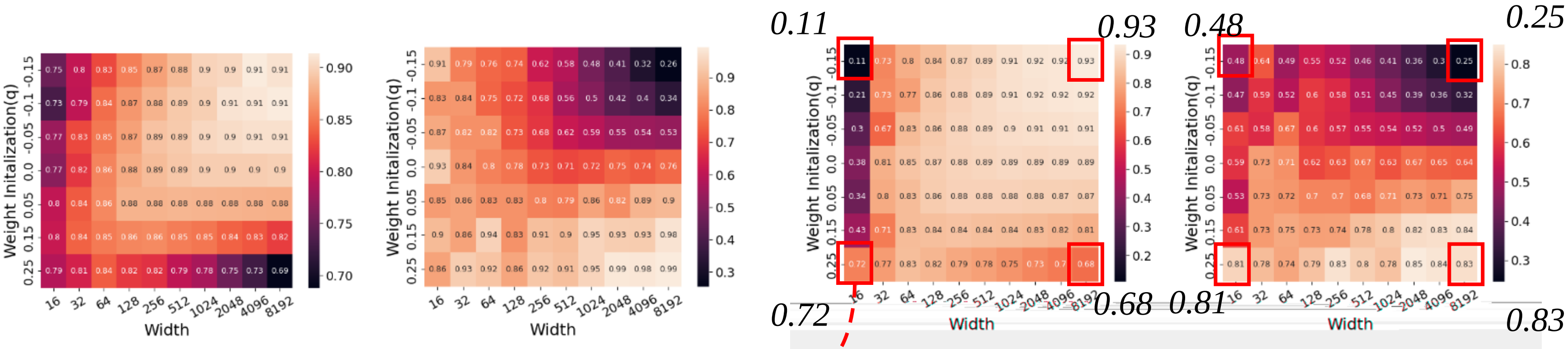


(a) ResNets (b) A1-A2 (c) B1-B2

- **Divide the ResNet50 and WRN in 2 ways**
- **w/o specific weight initialization**
- **On CIFAR10 with Adam optimizer**

Experiments

Different behavior for wide and narrow MLPs



(a) Linear Model/Acc. (b) Linear Model/ϱ (c) MLP ReLU/Acc. (d) MLP ReLU/ϱ

- ***Difference in behavior between wider and narrower neural networks***

UNIVERSITY OF EXETER

Conclusion

- By quantifying the interactions between layers, we found that it not only fails to collaborate against adversarial perturbations but may even hinder resistance to them during gradient descent.

- Wider MLPs exhibit more resistance to increased co-correlation and, therefore, are more adversarial robust.

- Future research can expand upon this by examining the effects of increased network depth and more sophisticated structures on the observed phenomena.