Measurements of Three-Level Hierarchical Structure in the Outliers in the Spectrum of Deepnet Hessians

Vardan Papyan

Department of Statistics
Stanford University

June 13, 2019

Setting

- C-class classification problem

Setting

- C-class classification problem
- Loss:

$$\mathcal{L}(\theta) = \text{Ave}_{i,c}\{\ell(f(x_{i,c};\theta), y_c)\}$$

Setting

- C-class classification problem
- Loss:

$$\mathcal{L}(\theta) = \text{Ave}_{i,c}\{\ell(f(x_{i,c};\theta), y_c)\}$$

- Hessian:

$$\text{Hess}(\theta) = \text{Ave}_{i,c}\left\{\frac{\partial^2 \ell(f(x_{i,c};\theta), y_c)}{\partial \theta^2}\right\}$$

Setting

- C-class classification problem
- Loss:

$$\mathcal{L}(\theta) = \text{Ave}_{i,c}\{\ell(f(x_{i,c};\theta), y_c)\}$$

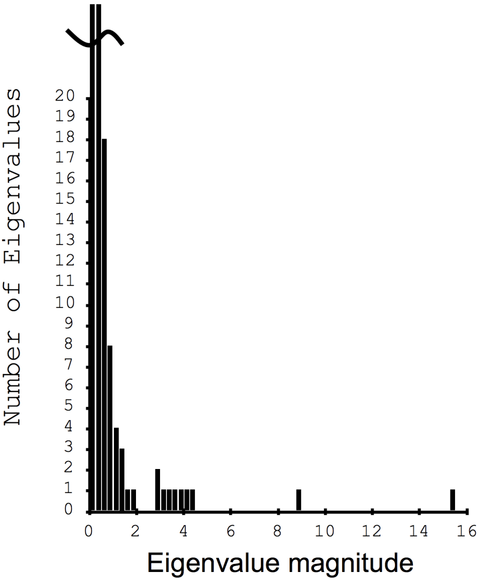
- Hessian:

$$\text{Hess}(\theta) = \text{Ave}_{i,c}\left\{\frac{\partial^2 \ell(f(x_{i,c};\theta), y_c)}{\partial \theta^2}\right\}$$

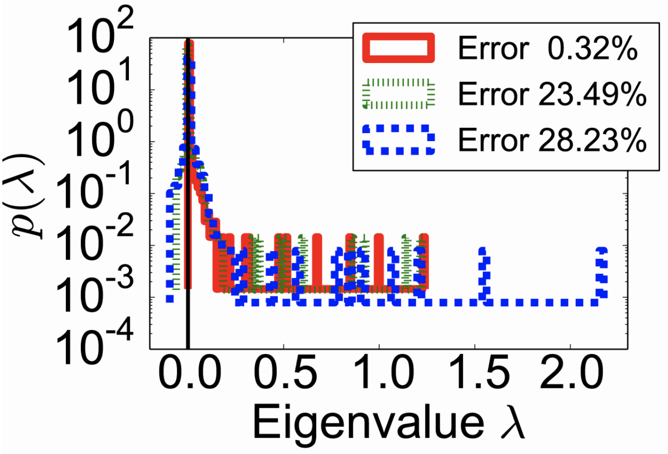
- Gauss-Newton decomposition:

$$\text{Hess} = G + H$$

Previous work: LeCun et al. (1998)



Previous work: Dauphin et al. (2014)



Previous work: Sagun et al. (2017)

- Noticed that the spectrum can be decomposed into:

Previous work: Sagun et al. (2017)

- Noticed that the spectrum can be decomposed into:
  - Bulk+outliers

Previous work: Sagun et al. (2017)

- Noticed that the spectrum can be decomposed into:
  - Bulk+outliers
  - Number of outliers $\approx$ number of classes

What is causing the outliers in the spectrum?

G is a second moment of gradients with structure on indices

- Define the gradient:

$$\delta_{i,c,c'}{}^T = \sqrt{p_{c'}(x_{i,c};\theta)}(y_{c'} - p(x_{i,c};\theta))^T \frac{\partial f(x_{i,c};\theta)}{\partial \theta}$$

G is a second moment of gradients with structure on indices

- Define the gradient:

$$\delta_{i,c,c'}{}^T = \sqrt{p_{c'}(x_{i,c};\theta)}(y_{c'} - p(x_{i,c};\theta))^T \frac{\partial f(x_{i,c};\theta)}{\partial \theta}$$

  - $\delta_{i,c,c}$: gradient of i-th example in c-th class (up to a scalar)

G is a second moment of gradients with structure on indices

- Define the gradient:

$$\delta_{i,c,c'}{}^T = \sqrt{p_{c'}(x_{i,c};\theta)}(y_{c'} - p(x_{i,c};\theta))^T \frac{\partial f(x_{i,c};\theta)}{\partial \theta}$$

  - $\delta_{i,c,c}$: gradient of i-th example in c-th class (up to a scalar)
  - $\delta_{i,c,c'}$: gradient of i-th example in c-th class, if it belonged to class c' instead (up to a scalar)

G is a second moment of gradients with structure on indices

- Define the gradient:

$$\delta_{i,c,c'}{}^T = \sqrt{p_{c'}(x_{i,c};\theta)}(y_{c'} - p(x_{i,c};\theta))^T \frac{\partial f(x_{i,c};\theta)}{\partial\theta}$$

  - $\delta_{i,c,c}$: gradient of i-th example in c-th class (up to a scalar)
  - $\delta_{i,c,c'}$: gradient of i-th example in c-th class, if it belonged to class c' instead (up to a scalar)
- These gradients can be indexed by three numbers:

G is a second moment of gradients with structure on indices

- Define the gradient:

$$\delta_{i,c,c'}{}^{T} = \sqrt{p_{c'}(x_{i,c};\theta)}(y_{c'} - p(x_{i,c};\theta))^{T}\frac{\partial f(x_{i,c};\theta)}{\partial\theta}$$

  - $\delta_{i,c,c}$: gradient of i-th example in c-th class (up to a scalar)
  - $\delta_{i,c,c'}$: gradient of i-th example in c-th class, if it belonged to class c' instead (up to a scalar)
- These gradients can be indexed by three numbers:
  - i: observation

G is a second moment of gradients with structure on indices

- Define the gradient:

$$\delta_{i,c,c'}{}^T = \sqrt{p_{c'}(x_{i,c};\theta)}(y_{c'} - p(x_{i,c};\theta))^T \frac{\partial f(x_{i,c};\theta)}{\partial \theta}$$

  - $\delta_{i,c,c}$: gradient of i-th example in c-th class (up to a scalar)
  - $\delta_{i,c,c'}$: gradient of i-th example in c-th class, if it belonged to class c' instead (up to a scalar)
- These gradients can be indexed by three numbers:
  - i: observation
  - c: true class

G is a second moment of gradients with structure on indices

- Define the gradient:

$$\delta_{i,c,c'}{}^T = \sqrt{p_{c'}(x_{i,c};\theta)}(y_{c'} - p(x_{i,c};\theta))^T \frac{\partial f(x_{i,c};\theta)}{\partial \theta}$$

  - $\delta_{i,c,c}$: gradient of i-th example in c-th class (up to a scalar)
  - $\delta_{i,c,c'}$: gradient of i-th example in c-th class, if it belonged to class c' instead (up to a scalar)
- These gradients can be indexed by three numbers:
  - i: observation
  - c: true class
  - c': potential class

G is a second moment of gradients with structure on indices

- Define the gradient:

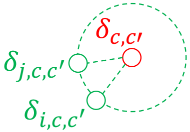
$$\delta_{i,c,c'}{}^T = \sqrt{p_{c'}(x_{i,c};\theta)}(y_{c'} - p(x_{i,c};\theta))^T \frac{\partial f(x_{i,c};\theta)}{\partial \theta}$$

  - $\delta_{i,c,c}$: gradient of i-th example in c-th class (up to a scalar)
  - $\delta_{i,c,c'}$: gradient of i-th example in c-th class, if it belonged to class c' instead (up to a scalar)
- These gradients can be indexed by three numbers:
  - i: observation
  - c: true class
  - c': potential class
- G is a second moment (not Covariance) of these gradients:

$$G = \text{Ave}_{i,c,c'}\left\{\delta_{i,c,c'}\delta_{i,c,c'}^T\right\}$$

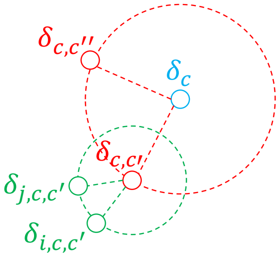
Three-level hierarchical structure in gradients

- Averaging over the index i



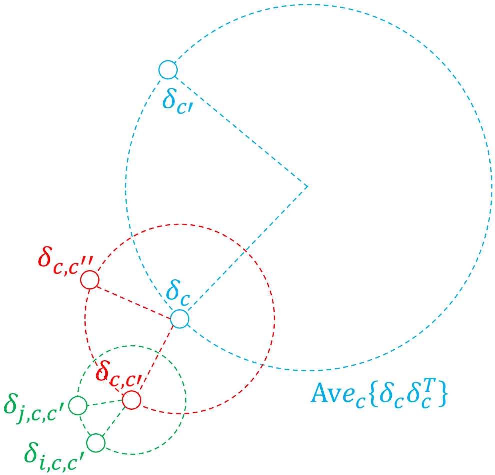
Three-level hierarchical structure in gradients

- Averaging over the index i
- Averaging over the index c'



Three-level hierarchical structure in gradients

- Averaging over the index i
- Averaging over the index c'
- Averaging over the index c



Visualization of three-level hierarchical structure in gradients

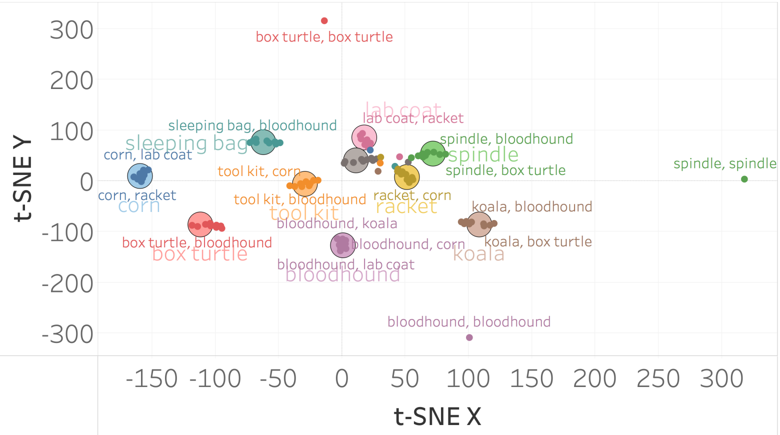
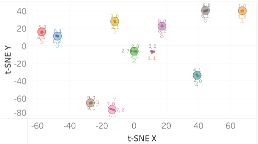
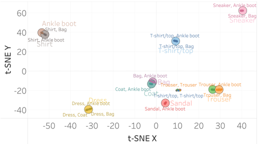


Figure: ResNet50 trained on ImageNet. Large circles: δ_c. Small circles: $\delta_{c,c'}$.

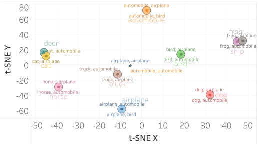
Visualization of three-level hierarchical structure in gradients



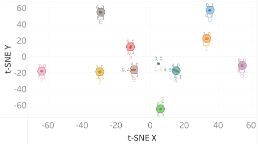
MNIST, 13 examples per class



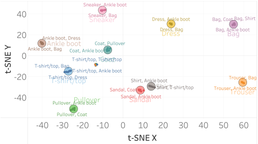
Fashion, 13 examples per class



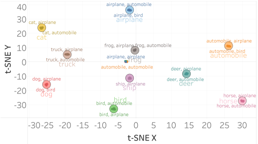
CIFAR10, 13 examples per class



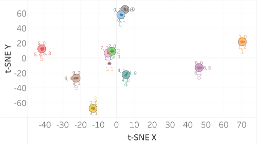
MNIST, 702 examples per class



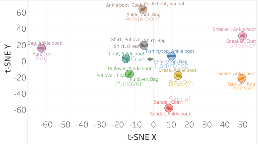
Fashion, 702 examples per class



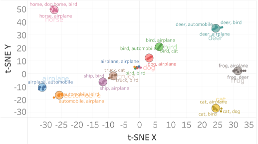
CIFAR10, 702 examples per class



MNIST, 5000 examples per class



Fashion, 5000 examples per class



CIFAR10, 5000 examples per class

$\text{Ave}_c \left\{ \delta_c \delta_c^T \right\}$ causes outliers in G

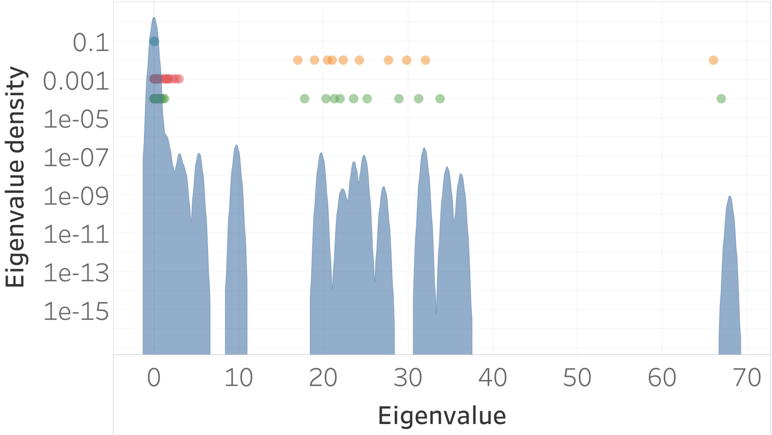
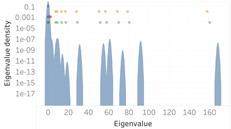
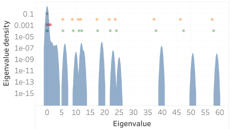


Figure: ResNet18 trained on CIFAR10, 1351 examples per class. Orange: eigenvalues of $\text{Ave}_c \left\{ \delta_c \delta_c^T \right\}$.

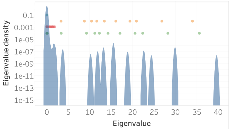
$\text{Ave}_c \left\{ \delta_c \delta_c^T \right\}$ causes outliers in G



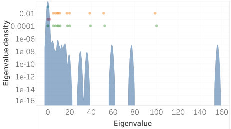
MNIST, 136 examples per class



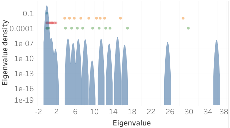
Fashion, 136 examples per class



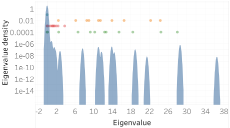
CIFAR10, 136 examples per class



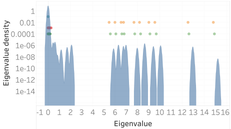
MNIST, 365 examples per class



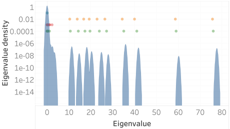
Fashion, 365 examples per class



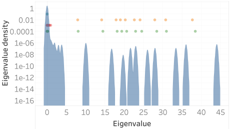
CIFAR10, 365 examples per class



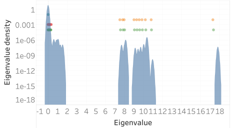
MNIST, 702 examples per class



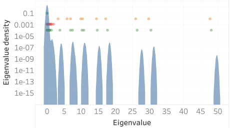
Fashion, 702 examples per class



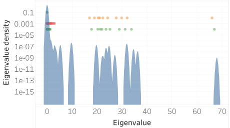
CIFAR10, 702 examples per class



MNIST, 2599 examples per class



Fashion, 2599 examples per class



CIFAR10, 1351 examples per class