NEURAL INFORMATION PROCESSING SYSTEMS

Shortcut Features as Top Eigenfunctions of NTK: A Linear Neural Network Case and More

Jinwoo Lim[1], Suhyun Kim[2], Soo – Mook Moon[1]

[1]Seoul National University

[2]Kyung Hee University

Virtual Machine & Optimization Laboratory

Dept. of Electrical and Computer Engineering

Seoul National University

Motivation

In shortcut learning, interesting phenomena occur such as …

- Shortcut features are **learned faster** than other features
- Shortcut features contribute more to the network output than other features after convergence
- While data samples with no shortcut features can also be learned with near-zero loss during training, such data outside the training set are still not well-predicted

We analyze those phenomena upon the framework of Neural Tangent Kernel (NTK) theory

- **Case study**: training a linear neural network on a dataset of an almost-separable Gaussian mixture model, of which the majority of the samples are clustered by a certain feature

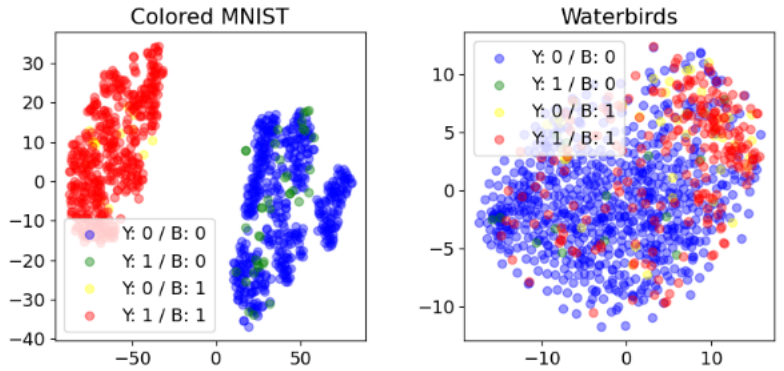
Problem formulation

We define a simple setting of training data distribution:

- All samples x are labelled as $y \in \{-1, 1\}$
- Data samples are clustered as a simple Gaussian mixture model based on the attributes of samples
- Biased cluster: $B_{y,i} \sim N(\mu_{B_{y,i}}, \sigma^2_{B_{y,i}} \mathrm{I})$,
- Bias-conflicting cluster: $C_{y,i} \sim N(\mu_{C_{y,i}}, \sigma^2_{C_{y,i}} \mathrm{I})$

$$p(x) = \sum_{y \in \{-1,1\}} [\sum_i \pi_{B_{y,i}} \mathcal{N}(\mu_{B_{y,i}}, \sigma^2_{B_{y,i}}) + \sum_j \pi_{C_{y,j}} \mathcal{N}(\mu_{C_{y,j}}, \sigma^2_{C_{y,j}})]$$

This setting resembles real-world scenarios



(a) Colored-MNIST (b) Waterbirds

Figure 2: t-SNE visualization of 1000 input data samples in datasets. Each input is marked in color corresponding to its label - Y: ground-truth label, B: label from a shortcut feature.

Spectral bias of NTK on biased dataset

Proposition 3.1. *Assume data* $x \in \mathbb{R}^d$ *in a Gaussian Mixture Model of* $p(x) = \sum_{k=1}^{K} \pi_k N(\mu_k, \sigma_k^2 I)$. *The kernel* $k(x, y) = \langle x, y \rangle$ *has eigenfunctions* ϕ_i *and corresponding eigenvalues* λ_i *as follows:*

$$\phi_i(x) = \begin{cases} x^\top v_i / c_i & \text{if } i = 1, \ldots, m \\ x^\top v_i^\perp / c_i & \text{otherwise} \end{cases}$$

$$\lambda_i = \begin{cases} \sum_{k=1}^{K} \pi_k \sigma_k^2 + a_i & \text{if } i = 1, \ldots, m \\ \sum_{k=1}^{K} \pi_k \sigma_k^2 & \text{otherwise} \end{cases}$$

when $(\sum_{k=1}^{K} \pi_k \mu_k \mu_k^\top) v_i = a_i v_i$, $v_i^\perp$ *is a vector perpendicular to* μ_k *for* $k \in \{1, \ldots, K\}$, $m = rank\left(\sum_{k=1}^{K} \pi_k \mu_k \mu_k^\top\right)$, $|v_i| = 1$, $|v_i^\perp| = 1$ *and* $c_i = \sqrt{\lambda_i}$.

$\pi_{B_{y,\cdot}}$ is much larger than $\pi_{C_{y,\cdot}}$, so we can expect that the inner products with data from these "shortcut clusters" will converge faster than others.

Spectral bias of NTK on biased dataset

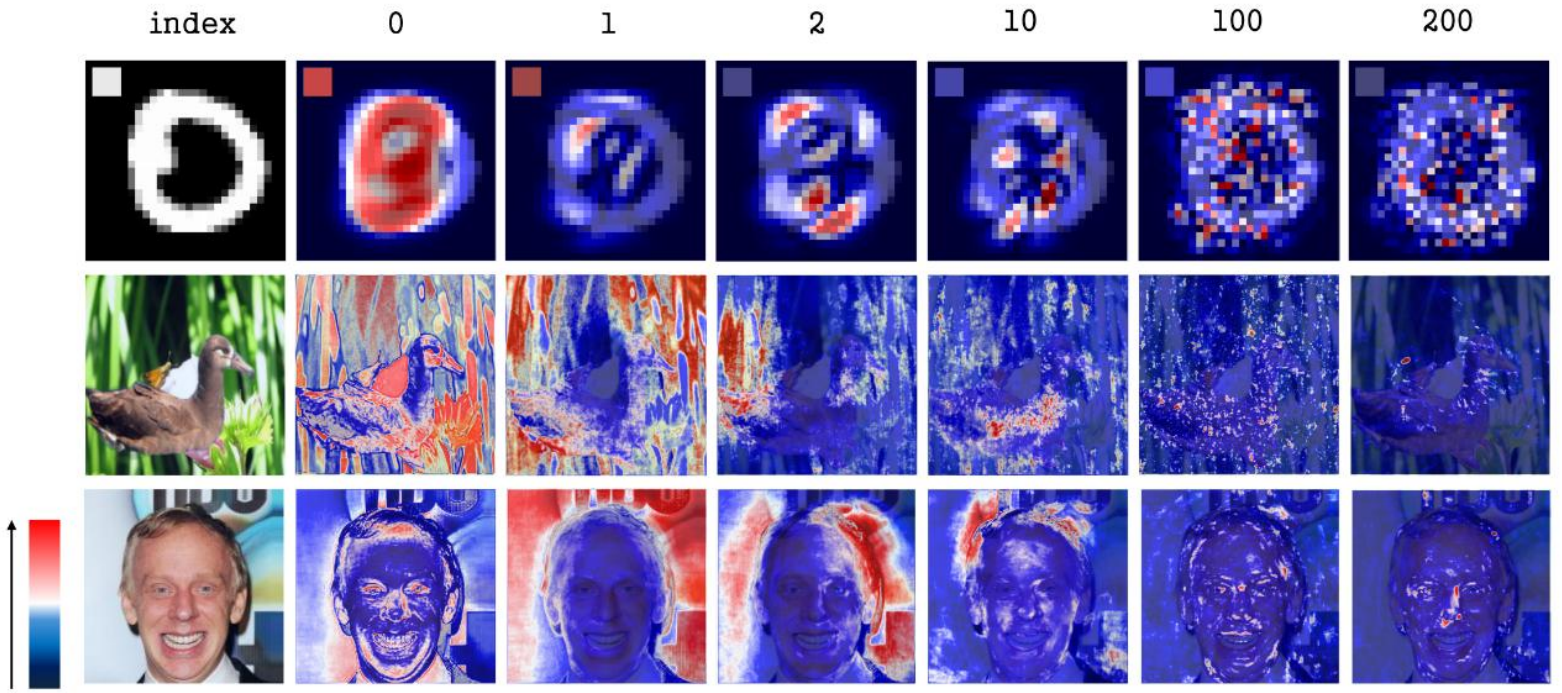


Figure 3: Original images and saliency maps from each feature of outputs from two-layer ReLU CNN networks. Saliency maps on the left side shows the spatial support of features with large eigenvalues, while saliency maps on the right side shows the spatial support of features with smaller eigenvalues. Indices indicate the ranks of the eigenvalues in terms of magnitude. A saliency map from the i-th index indicates the saliency map from a feature with the $(i+1)$-th largest eigenvalue. Features with larger eigenvalues focus on biased attributes of samples, i.e., in CelebA, features with larger eigenvalues focus on the edges of the face, or the background of an image rather than the hair itself.

Features after convergence

Proposition 3.2. *Assume data* $x \in R^d$ *in a Gaussian Mixture Model of* $p(x) = \sum_{k=1}^{K} \pi_k N(\mu_k, \sigma_k^2 I)$. *A binary label function* $y(x) \in \{-1,1\}$ *nearly separates the mixture model that data from clusters with mean* $\mu_c (c \in \mathcal{C})$ *are labelled as 1 and otherwise -1. When the linear neural network is optimized for y(x) with the MSE loss and dissected into* $f(x) = \sum_{k=1}^{m} w_k f_k(x)$ ($f_k(x) = x^\top v_k \propto \phi_k(x)$, $|v_k| = 1$), *and if* $\mu_i \perp \mu_j$ *for* $i \neq j$ *and* $v_k = \mu_k / |\mu_k|$, *then*

$$w_k = \begin{cases} \frac{\pi_k \|\mu_k\|_2}{\sum_{i=1}^{K} \pi_i \sigma_i^2 + \pi_k \|\mu_k\|_2^2} & if\ k \in \mathcal{C} \\ -\frac{\pi_k \|\mu_k\|_2}{\sum_{i=1}^{K} \pi_i \sigma_i^2 + \pi_k \|\mu_k\|_2^2} & if\ k \in \mathcal{C}^c \end{cases}$$

- Data samples with biased attributes, $B_{y,\cdot}$, belong to a larger cluster ($\pi_{B_{y,\cdot}}$), so biased features have large eigenvalues and a large influence on the output at the same time.
- Such a dependency of w_k on cluster weights π_k originates from the existence of the variances, σ_i, among data in the clusters.

Features after convergence: max-margin bias

The result on features after convergence is not due to max-margin bias.

In fact, max-margin bias alone cannot explain shortcut learning.

$$l_{SD}(f(x), y) = \log(1 + \exp(-yf(x))) + \frac{\lambda}{2}|f(x)|^2$$

- When NN is trained with SD (loss above) with a very strong regularization and margin is minimized ($\lambda \to \infty$), the decision boundary converges to the one of MSE loss, which is still affected by shortcut features. ($f(x) = \sum_{k=1}^{m} w_k f_k(x)$)

$$\lim_{\lambda \to \infty} \frac{(w_i)_{SD}}{(w_j)_{SD}} = \frac{(w_i)_{MSE}}{(w_j)_{MSE}}$$

Predictability and availability

We empirically extend theoretical results to the case of more complex networks.

- **Predictability** measures how well a feature g is aligned with the ground-truth.
  - Predictability is $\mathbf{y}^\top \mathbf{g}/|\mathrm{X}|$. $\mathbf{y} = [y(x) \text{ for } x \in \mathrm{X}]^\top$ is a vector of ground-truth labels, and $\mathbf{g} = [g(x) \text{ for } x \in \mathrm{X}]^\top$ is a vector of labels assigned by a certain feature $g(x)$.
- **Availability** measures how "easy" it is to learn a label for a neural network. For a discrete case, availability is an alignment of label $\mathbf{g}$ to empirical NTK (eNTK) of $K(X, X)$

$$A(K, \mathbf{g}) := \frac{\mathbf{g}^\top K(\mathcal{X}, \mathcal{X})\mathbf{g}}{\|\mathbf{g}\|_2^2 \|K(\mathcal{X}, \mathcal{X})\|_F} = \sum_i \frac{\lambda_i}{\sqrt{\sum_j \lambda_j^2}} \langle v_i, \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \rangle^2$$

Predictability and availability

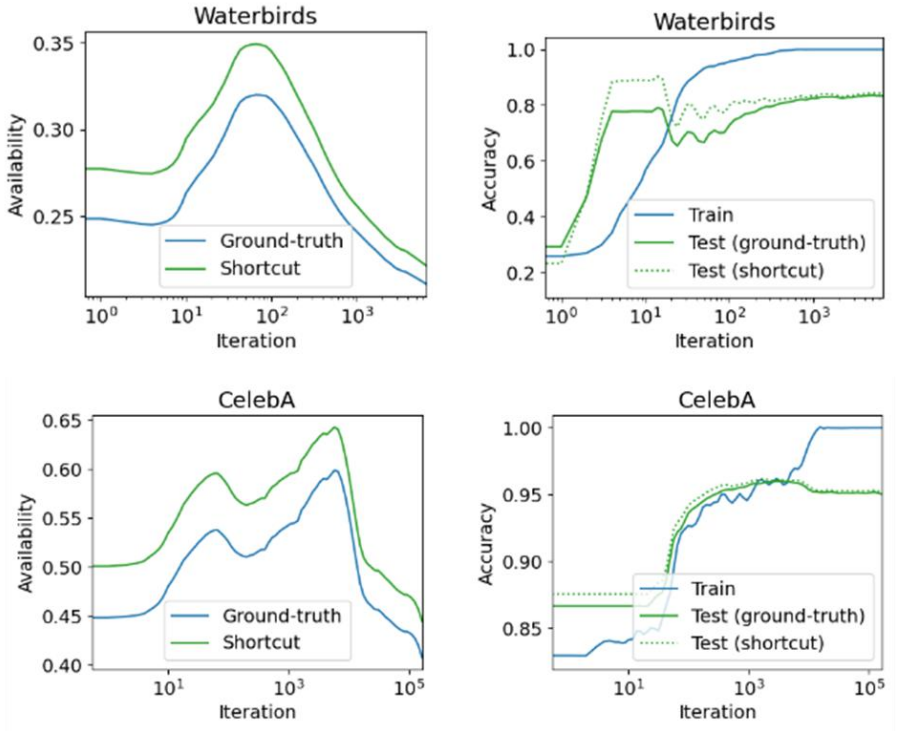


Figure 4. (a) Availability and test accuracy of ground-truth labels and shortcut labels in two realistic datasets: Waterbirds and CelebA. The tested model was a pretrained ResNet-18.

- The availability of shortcut labels was larger than the availability of the ground-truth labels.
- Since the availability measures the alignment of a label to top eigenvectors of eNTK, it is possible to say that the shortcut labels were more aligned with top eigenfunctions of eNTK.

THANK YOU!