On the Weak Neural Dependence Phenomenon in Deep Learning

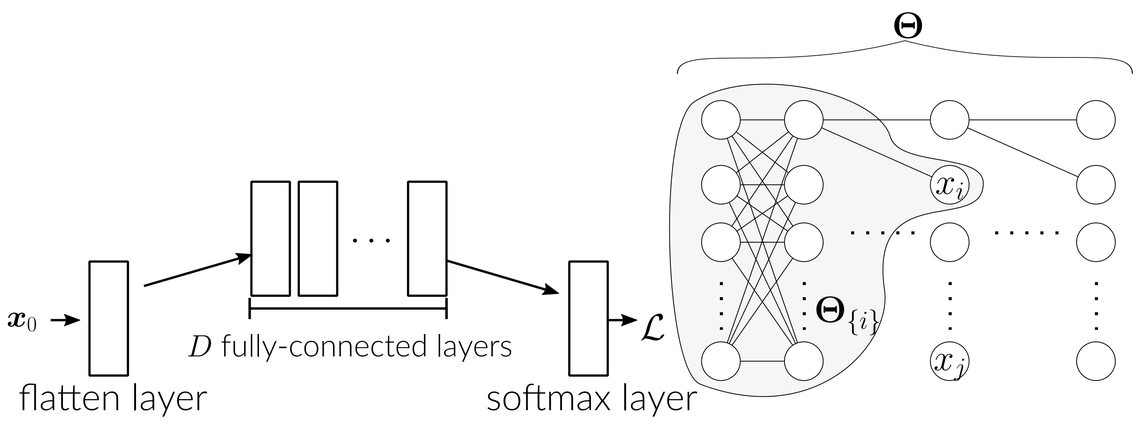
Jiayao Zhang[1] Ruoxi Jia[2] Bo Li[3] Dawn Song[2]

[1]HKU [2]UC Berkeley [3]UIUC

Abstract

We present the weak neural dependence lemma in deep neural networks (DNN) with wide fully-connected layers and arbitrary nonlinear activations. The lemma states that any two activations of the same layer have arbitrarily low dependence regardless of the training stage. To the best of our knowledge, we are the first to give a formal treatment of this phenomenon aided by tools from information theory. Our simulations on MNIST/CIFAR-10 corroborate the theory. The lemma is able to simplify the analysis of deep nets with reasonably wide layers.

Notations



(L) MLP model used. (R) Illustration of $\boldsymbol{\Theta}_S$.

- D, N_l, $\phi(\cdot)$: depth, width, and the activation function of the network.
- $\boldsymbol{z}_l^{(t)}$, $\boldsymbol{x}_l^{(t)}$: pre- and post-activation of layer l at time t.
- $H(x_i)$, $I(x_i; x_j)$: entropy of a neuron x_i and the mutual information (MI) between two neurons x_i, x_j (in the same layer).
- $\boldsymbol{\Theta}_S$: all parameters that contribute to $\{x_k\}_{k\in S}$.

Constant Entropy Assumption

Let z be an arbitrary neuron from any layer but the last, we assume the change of neural entropy $H(z)$, $\Delta H(z)$, after training, approaches to zero when layer width N_l becomes large:

$$\lim_{N_l\to\infty} \Delta H(z) = 0. \tag{1}$$

The Weak Neural Dependence Lemma

Let $I \coloneqq I(x_i; x_j)$ be the MI between any two neurons in the same layer l other than the last, we have

$$\lim_{N_l\to\infty} I = 0. \tag{2}$$

Sketch of the Proof

By definition,

$$I\left(x_i^{(t)}; x_j^{(t)}\right) = H\left(x_i^{(t)}\right) - H\left(x_i^{(t)}|x_j^{(t)}\right). \tag{3}$$

Since conditioning reduces entropy,

$$H\left(x_i^{(t)}|x_j^{(t)}, \boldsymbol{\Theta}_{\{i,j\}}^{(t)} = \boldsymbol{\Theta}_{\{i,j\}}^{(0)}\right) \le H\left(x_i^{(t)}|x_j^{(t)}\right). \tag{4}$$

Since x_i is independent of all parameters that do not contribute to it during the forward pass,

$$H\left(x_i^{(0)}|x_j^{(0)}\right) = H\left(x_i^{(t)}|x_j^{(t)}, \boldsymbol{\Theta}_{\{i,j\}}^{(t)} = \boldsymbol{\Theta}_{\{i,j\}}^{(0)}\right) \ge H\left(x_i^{(0)}|x_j^{(0)}\right). \tag{5}$$

Hence

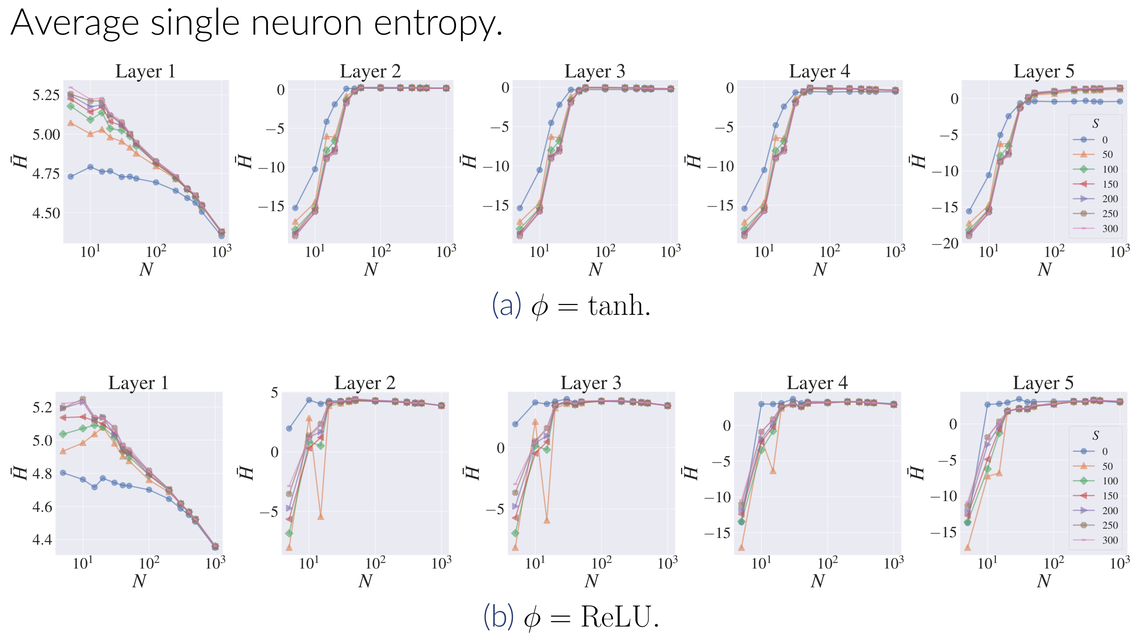
$$I\left(x_i^{(t)}; x_j^{(t)}\right) = H\left(x_i^{(t)}\right) - H\left(x_i^{(t)}|x_j^{(t)}\right) \le H\left(x_i^{(t)}\right) - H\left(x_i^{(0)}|x_j^{(0)}\right), \tag{6}$$

and the lemma follows from the assumption and the initial condition $I\left(x_i^{(t)}; x_j^{(t)}\right) = 0$.

Simulation Results

We tested on a five-layer ϕ-activated MLP, trained on MNIST and CIFAR-10 (not shown). We observe the convergence behaviour of the (average) single neuron entropy as layer width becomes larger; and the convergence of the (average) pairwise MI to the vicinity of zero, indicating low dependence.

1. Average single neuron entropy.

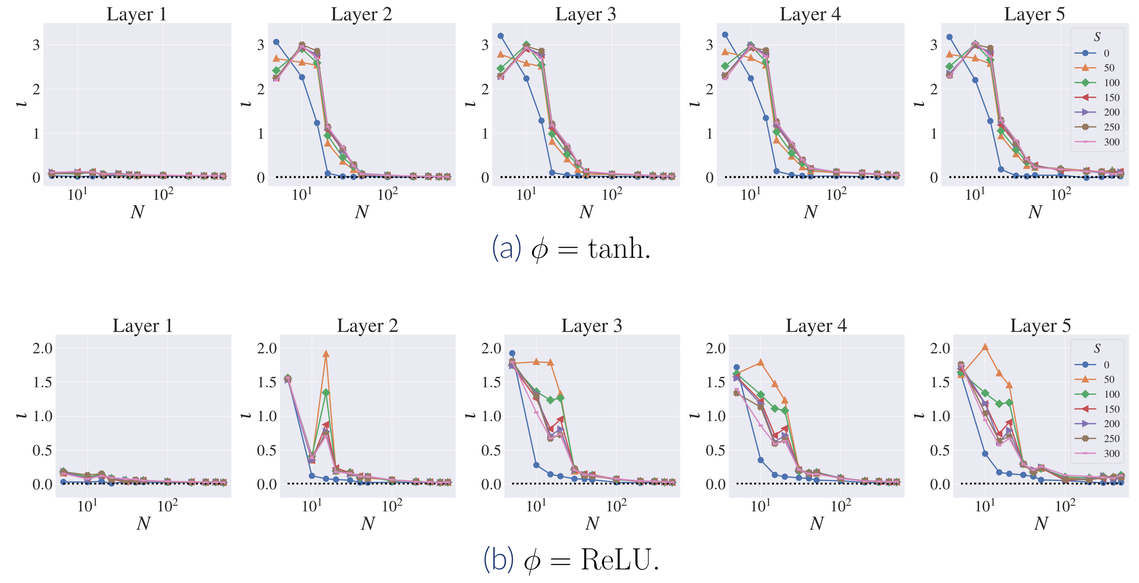


(a) $\phi = \tanh$.

(b) $\phi = \mathrm{ReLU}$.

Average neuron entropy $\bar{H}$ against different network width.

2. Average pairwise MI, which we call the dependence score ι.



(a) $\phi = \tanh$.

(b) $\phi = \mathrm{ReLU}$.

Dependence score ι of each layer against network width.

Consequences of the Lemma

- **Applicable to both pre/post-activations.** The extension of the lemma to pre-activations can be made with ease by redoing the proof. In $\tanh$-activated nets, the extension is even simpler since MI is invariant to reparameterization given by homeomorphisms (smooth invertible maps), in particular $\tanh^{-1}$.
- **Neural activities as multi-variate Gaussians.** When one looks at a particular a network at some training stage, by central limit theorem, the activations are approximately Gaussians. The Gaussian approximation has been used in multiple previous work for untrained nets; our results further suggest that the same holds for trained MLPs given reasonably wide layers.
- **Layers do not have to be too wide.** The average individual neuron entropy exhibited a convergence behaviour, and is correlated with the training accuracy (not shown). Information-theoretically, the average entropy is able to characterize the uncertainty in neural activities, and relate to the capacity in terms of the the uncertainty/diversity of hidden features. Increasing layer width after convergence at a higher cost of computation, therefore, may not induce further gain.