Supplementary Material to Abstraction Mechanisms Predict Generalization in Deep Neural Networks

Alex Gain [1] **Hava Siegelmann** [2]

1. Relationship Between the Gradients of CNA and Supervised Loss

We give additional exposition and description of the similarity betwen the CNA and supervised loss gradients mentioned in section 4.1:

Consider: Databatch X consisting of n samples and corresponding label batch Y and error terms $\mathcal{E}$, network layers $1, \ldots, L$, let n_ℓ denote the number of neurons in layer ℓ, and let z_ℓ^k denote the activation value of the network for neuron k in layer ℓ. Lastly denote the supervised loss as $C(X, Y)$. Then the supervised loss gradient[1] derives to:

$$\nabla C(X, Y) = \frac{1}{n} \sum_{x \in X, \varepsilon \in \mathcal{E}} \varepsilon \nabla z_L(x) \tag{1}$$

And the CNA gradient[2] derives to:

$$\nabla CNA_\alpha(X) = \frac{1}{n} \sum_{x \in X} \alpha(x) \nabla \beta(x) \tag{2}$$

The slope gradient is defined by

$$\nabla \beta(x) = \sum_{\ell=1}^{L} \frac{C_\ell^\dagger}{n_\ell} \sum_{k=1}^{L} \nabla z_\ell^k(x) \tag{3}$$

where $C^\dagger$ is the row of the pseudoinverse matrix used in the least squares regression corresponding to the slope term $\beta(x)$.

Focusing on the terms ε, $\alpha(x)$, $\nabla z_L(x)$, and $\nabla \beta(x)$, we observe some resemblance between the two gradients. They both have scalars (ε and $\alpha(x)$) multiplied by gradients of network output terms ($\nabla z_L(x_i)$ and $\nabla \beta(x)$). We now show that the terms in both pairs are closely related.

On the similarity of $\nabla z_L(x)$ and $\nabla \beta(x)$: Note that $\nabla \beta(x)$ is a weighted sum[3] of network output terms ∇z_ℓ for $\ell = 1, \ldots, L$. There is similarity between $\nabla z_L(x)$ and $\nabla \beta(x)$ since they are both linear functions of activation output gradients. In practice, when training on the MNIST dataset with an MLP network, we find that, for the vast majority of datapoints, the cosine angle between $\nabla z_L(x)$ and $\nabla \beta(x)$, when using mean-aggregated activations, does not exceed 0.05 radians, meaning their directions are *very* similar.

Given this large degree of similarity between $\nabla z_L(x)$ and $\nabla \beta(x)$, if ε and $\alpha(x)$ correlate, we would expect updates via $\nabla C(X, Y)$ and $\nabla CNA_\alpha(X)$ to take the network along similar optimization paths.

2. Additional Expermental Details

We provide additional experimental details for verification and reproducability.

For the test accuracy and generalization gap experiments seen section 4:

For the MLP architecture, the depth was fixed at 5 hidden layers of size 500 each. Regularization was not used for the MLPs, but batch-normalization was used for the VGG-18 and ResNet architectures. Max-pooling was used after every block of the VGG-18 and ResNet architectures followed by average pooling in the last block.

For VGG and ResNet architectures, standard image augmentation was used for SVHN, CIFAR-10, CIFAR-100, and ImageNet. Otherwise, image augmentation was not used. Across all experiments, ImageNet was downsampled

[1]Department of Computer Science, The Johns Hopkins University, Baltimore, MD 21218, USA [2]School of Computer and Information Sciences, University of Massachusetts Amherst, Amherst, MA 01003, USA. Correspondence to: Alex Gain <again1@jhu.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

[1]For brevity, our expressions correspond to the 1-dimensional output case (derivations straightforwardly generalize to larger output dimensions).

[2]Also for brevity, we consider the mean-aggregated CNA gradient without the standard deviation normalization terms. The direction of the non-normalized CNA gradient is approximately equal to the direction of CNA gradient with these terms included, i.e. they have a cosine similarity of close to 1 in practice. The direction is what is important since our analysis is primarily concerned with the cosine similarity between the supervised loss and CNA gradients (cosine similarity is invariant to scaling).

[3]In fact, given the nature of $C^\dagger$, it can be shown that is $\nabla \beta(x)$ is a weighted *average* of ∇z_ℓ terms.

to resolution 32 x 32 for computational expediency.

For results shown in Figure 6, a quadratic fit of the form $ax^2 + bx + c$ was performed to arrive at the green dotted curve shown. Points were smoothed in bins of 25 networks each, ordered by test accuracy, for cleaner visualization.

Across all experiments, training continued until approximately 0 training loss was achieved. For all standard datasets, we trained MLPs for 100 epochs with a learning rate of 0.05 and momentum of 0.8 via SGD. For the VGG-18 and ResNet experiments, a learning rate of 0.01 and momentum of 0.8 were used. The VGG and ResNet architectures were trained for 100 epochs on CIFAR-10 and CIFAR-100, 40 epochs on MNIST, Fashion-MNIST, and SVHN, and 50 epochs on ImageNet. On ImageNet, the MLP architecture was excluded from analysis since it failed to converge passed 32% training accuracy after 1000 epochs. For the Gaussian noise dataset, the same optimization settings were used except all architectures were trained for 40 epochs (since this was a sufficient number for memorization of the training set).

The Gaussian noise dataset is of training size 50,000 and test size 10,000, where datapoints are of shape (3,32,32), are drawn from the standard normal distribution, and are then normalized between 0 and 1. A total of 10 classes are randomly assigned to each datapoint.

For the shuffled label datasets (MNIST, Fashion-MNIST, SVHN, CIFAR-10, CIFAR-100), for each experiment, a percentage of the training labels was shuffled and trained on with a new network. The percentages considered were 10%, 20%, 30%, 40%, and 50%, with metrics of the network measured at the end of training once approximately 0 training error had been achieved, for a total of 75 additional trained networks considered. All training settings were consistent with the original datasets, except the number of epochs were doubled, since shuffled label datasets require more training iterations to memorize the training set(Zhang et al., 2016). The networks at the end of training on shuffled label datasets were incorporated into analysis of the networks trained every 20 epochs on the corresponding non-shuffled datasets. Lastly, for the MLP architecture, other metrics (L2, L2-path, Spectral norm, and the 2018 bound) *negatively* correlated with generalization gap, performing particularly bad in comparison to other settings, thus we show them at 0 correlation for ease of visualization.

3. Additional Neuroscience Details

We give a slightly expanded description of the neuroscience study from (Taylor et al., 2015), with more descriptive and technical terms for those who want to become more familiar with the neuroscience science details of the study:

The neuroscience study employed large-scale analyses of fMRI data, spanning twenty years and tens of thousands of studies to determine the relationships between a variety of tasks having different levels of abstraction and the associated neuronal firing patterns across the whole brain. The study revealed a strong correlation between a given cognitive behavior's "firing slope" (a geometric characterization of global neuronal firing patterns) and its level of abstraction. Firing slope utilizes a distance metric – combining fMRI latency and accessibility (DTI) that measures the "connectome distance" (CD) of each neuronal Region Of Interest (ROI) from the brain's inputs (sensory cortices). Regions of Interest were binned by their CD to create a layered "connectome depth network" (CDN), similar in structure to a deep neural network. All fMRI experiments were projected on the CDN, and experiments that measured the same cognitive behavior (typically about 1000) were analyzed together.

While brain activity was present at all connectome depths for each cognitive behavior, findings demonstrated that deep neurons (those farther from brain inputs on the CDM) showed higher activation values than shallow neurons when the brain was engaged in reasoning and other abstract behaviors. When graphed against CD, this neuronal activity showed a positive slope. Conversely, shallow neurons had higher activation values than deep neurons and neuronal activity showed a negative slope when less abstract (shorter CD) tasks were performed. Each of the recorded behaviors was identifiable by a specific geometric slope on the CDN that correlated with the behavior's level of abstraction. Perhaps the most novel aspect of the research was the almost perfect correlation between firing slope and level of abstraction.

4. Additional Generalization Gap Results

Here, we show the correlation of the generalization metrics conditional on datasets in the following table. Conditional on dataset only, all metrics appear to be inconsistent when considering all architectures in aggregate. We also include the CNA-Area, which performs particulary well for this task in comparison.

References

P. Taylor, J. Hobbs, J. Burroni, and H. Siegelmann. The global landscape of cognition: hierarchical aggregation as an organizational principle of human cortical networks and functions. *Scientific reports*, 5(1):1–18, 2015.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

	CNA-Area	CNA-Margin	L2	L2-path	Spectral norm	2018 Bound
MNIST	**0.98**	0.39	0.13	0.05	0.17	0.17
F-MNIST	**0.72**	0.11	-0.24	-0.45	-0.22	-0.22
SVHN	**0.28**	0.04	-0.15	-0.16	-0.12	-0.12
CIFAR-10	**0.45**	0.02	-0.24	0.11	-0.2	-0.2
CIFAR-100	-0.23	0.18	-0.29	**0.25**	-0.26	-0.26
ImageNet-32	**0.82**	-0.54	-0.5	-0.28	-0.5	-0.5
Gaussian	-0.37	**0.37**	0.22	-0.29	0.21	0.21