Towards an Understanding of Neural Networks in Natural-Image Spaces

Yifei Fan Anthony Yezzi
Georgia Institute of Technology
85 5th Street, NW Atlanta, GA 30308 USA
yifei@gatech.edu, anthony.yezzi@ece.gatech.edu

Abstract

Two major uncertainties, dataset bias and adversarial examples, prevail in state-of-the-art AI algorithms with deep neural networks. In this paper, we present an intuitive explanation of these issues as well as an interpretation of the performance of deep networks in a natural-image space. The explanation consists of two parts: the variational-calculus view of machine learning and a hypothetical model of natural-image spaces. Following the explanation, we (1) demonstrate that the values of training samples differ, (2) provide incremental boosts to the accuracy of a CIFAR-10 classifier by introducing an additional "random-noise" category during training, and (3) alleviate over-fitting thereby enhancing the robustness of a classifier against adversarial examples by detecting and excluding illusive training samples that are consistently misclassified. Our overall contribution is therefore twofold. First, while most existing algorithms treat data equally and have a strong appetite for more data, we demonstrate in contrast that an individual datum can sometimes have disproportionate and counterproductive influence, and that it is not always better to train neural networks with more data. Next, we consider more thoughtful strategies by taking into account the geometric and topological properties of natural-image spaces to which deep networks are applied.

1. Introduction

Recent years have witnessed the rapid development of artificial intelligence (AI) and deep learning. However, two major issues remain and prevent us from establishing robust real-world applications with current algorithms. One is dataset bias [1], meaning that a machine-learning algorithm that performs well on one dataset may fail on another. The other is adversarial examples [2], which shows that tiny modifications on inputs may lead to incorrect outputs by deep networks, even though the perturbations are almost imperceptible by humans. In this paper, we present our understanding of the behavior of neural networks and explain the uncertainties with a hypothetical model of natural-image spaces. Our contributions are as follows:

- We provide a unified explanation for dataset bias and adversarial examples from the perspective of variational calculus and properties of natural-image spaces.
- We illustrate that the values of training samples differ. Training with more samples does not guarantee higher accuracy, and even random noise can sometimes help improve the performance of neural-network classifiers.
- We present a hypothetical model for natural-image spaces, which can potentially guide a network to alleviate over-fitting, enhance robustness against adversarial examples, and improve accuracy.

2. Variational-Calculus View of Learning

The goal of the learning scheme is to obtain an approximation of the underlying target fucntion $f(\cdot)$ by optimizing an objective loss $L(\cdot)$. According to the geometric view taken in functional analysis, the target function $f(\cdot)$ is an extremum point, the values of which may be explained as components of an infinite-dimensional vector indexed by its domain (i.e., $\{f_x\}_{x\in X}$). Meanwhile, the objective function $L(\cdot)$ corresponds to a functional $L(f(\cdot))$ in the calculus of variations. One can discretize a functional and obtain an objective function for learning algorithms.

$$L(f) = \int_0^1 (f(x) - y(x))^2\, dx \Longrightarrow \frac{1}{n}\sum_{i=1}^{n}(h(x_i) - y_i)^2 = \text{MSE}$$

The variational-calculus view shows that the training process in machine learning is heavily data-dependent because the output of the objective loss function (i.e., a functional) depends not only the *summand* f, but also the *interval* (i.e., the data x). Therefore, deficiencies in data themselves and data-usage will impact learning-based algorithms. Such deficiencies, however, cannot be mitigated by improving optimizers or network structures for f.

We then introduce the *point-function duality* of the underlying target function f to emphasize that the training

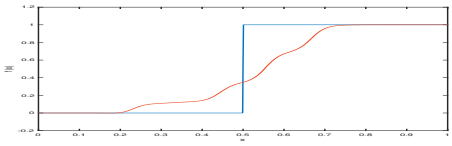

Figure 1: Training samples affect the geometry of approximation of $f(x) = \text{sign}(x - 0.5)$. Blue: 40,000 from [0.4, 0.6); orange: 20,000 from [0, 0.1) and 20,000 from [0.9, 1).

process is under-constrained. As an optimal point of the loss functional $L(f)$, the target function f is unaware of its actual domain and range. As a function, the topological properties of its domain and range do matter, especially if we apply $f : X \to Y$ to test inputs. Different from admissible functions in variational calculus, the approximated target function sets no restrictions on the value of its inputs and outputs, and might be biased and locally valid around training data. As Figure 1 shows, the shape of the learned curves and boundaries is determined by training data and sometimes cannot be dictated directly by learning algorithms.

3. Models of Natural-image Spaces

As the learning process is heavily data-dependent, we deem that the characteristics of the input image space are worth studying. Another motivation for studying the input space is that neural networks cannot classify even and odd numbers, which implies that the topological properties matter. In this work, we address natural-image spaces in a discrete manner. Let $\mathbb{Z}_{[l,u]}$ denote the set of integers from l to u. A natural image I with resolution $w \times h$ and d channels is considered as a point of the $\mathbb{Z}_{[0,255]}^{w \times h \times d}$ -based natural-image space $\mathbb{I}^{w \times h \times d}$. The essence of the discrete set-up is to consider a space for each resolution as a "chart" of the manifold, and these spaces are not necessarily dense. We then study the properties of natural-image spaces and reveal that the properties are closely related to data-usage in learning.

3.1. Sparsity and the scale space effect

Natural-image spaces are sparse since the probability that a random sample in $\mathbb{Z}_{[0,255]}^{w \times h \times d}$ belonging to $\mathbb{I}^{w \times h \times d}$ is small. Moreover, they are denser in lower dimensions because the quantity of valid natural images increases more slowly than that of possible cases as resolution increases. These properties are consistent with results that show algorithms for image generation and translation are better at producing images with lower resolutions. In image classification, however, such properties are not addressed properly, especially when, as is typically the case, a limited number of categories are provided. Under current settings, classifiers assume that the underlying target function is well-defined everywhere and learn to paint the entire input space with a fixed number of colors. When classifiers eliminate all possible categories, they have to pick the remaining one as the final output. Unfortunately, no category for exception exists. To classify the input space continuously and smoothly, the decision boundaries are often deformed. As a first step for handling exceptions, we augment training samples with an extra random-noise category with the hope that it would fill in the invalid regions among categories and push the decision boundaries towards the centroid of each cluster. Suppose there are N training samples ($\mathbf{x}$) and M classes. The improved cross-entropy loss becomes:

$$L(p(x), \mathbf{x}, y) = -\sum_{i=1}^{N} w(\mathbf{x}^i) \sum_{c=1}^{M+1} y_c^i \log p_c(\mathbf{x}^i) \qquad (1)$$

in which y_c^i is the binary indicator (0, 1) if class label c is the correct classification, $p_c(\mathbf{x}^i)$ is the predicted probability class c, and $w(\mathbf{x}^i)$ is the weight, all for $\mathbf{x}^i$.

3.2. Connectivity

Based on the sparsity property, we infer that two images belonging to the same category are not always "path-connected" through adjacent grid points in $\mathbb{Z}_{[0,255]}^{w \times h \times d}$. At the micro level, a natural-image space can be regarded as a *quotient space* $Y = X/\sim$ consisting of numerous equivalence classes of "path-connected" images that can be derived from each other via non-destructive operations (e.g., using filters). These equivalence classes, however, are not necessarily connected. One gap exists between a labeled natural-image space and the one approximated by a neural network, because the regions classified by the network tend to be connected [3]. We claim that such a gap is one of the key reasons for uncertainties in neural networks. In addition, the gap is exacerbated by a setting of a fixed yet insufficient number of categories. The setting reflects a subconscious assumption in machine learning: algorithms should always learn and follow the exact concepts (e.g., object classes) from humans. In fact, they may require multiple classes to fully comprehend a human-level concept. It is possible that instances in one dataset are "essentially" different from instances in another, even though they correspond to the same "biased" concept according to humans. Therefore, to break the constraints from the limit on number of categories, we allow more equivalence classes that are linked with a table to represent a category. Intuitively, these classes slice the input space into discontinuous pieces and glue them together with the table. Similar to equation (1), the improved cross-entropy function can be expressed as:

$$L_j(p(x), \mathbf{x}, y) = -\sum_{i=1}^{N} w(\mathbf{x}^i) \sum_{c=1}^{M_j(p,\mathbf{x},y)} y_c^i \log p_c(\mathbf{x}^i) \qquad (2)$$

in which the number of sub-networks j and equivalence classes M, as well as weights on training data $w(\cdot)$ will be determined by algorithms.

4. Experiments

We present experimental results that support our statements on the impact of training samples and the benefits of utilizing the topological and geometric properties of natural-image spaces. As classification is the basis for advanced tasks, we set up controlled image-classification experiments on CIFAR-10 [4]. To measure the location of samples in the learned space, we assume that the output probability of a category is negatively correlated with the distance from the sample to the centroid of that category.

4.1. Impact of Training Samples

A regular training process is terminated at a premature stage before over-fitting occurs. After training, we categorized the training samples into two super-classes: those that were correctly classified and those that were misclassified. Within each super-class, we sorted the samples according to their maximum probability score, which is denoted by "confidence" for the correctly classified or "illusiveness" for the misclassified. From the correctly classified samples, we selected two subgroups with relatively higher (S_{hc}) and lower (S_{lc}) confidences. Similarly, two subgroups with higher (S_{hi}) and lower illusiveness (S_{li}) were selected from the misclassified. With all subgroups the same size, we then retrained the network with the selected subgroups to illustrate the impact of different training samples. The left half of Table 1 shows the performance of the classifier after retraining with subgroups of the original data.

We observe the following: More training samples do not guarantee a higher accuracy. High-confidence images are required for higher test accuracy, especially when a limited number of training samples are provided. Highly illusive images are misleading when the size of the training set is small; as the training set expands, however, such adverse images become valuable and lead to even higher accuracy. In this sense, the highly illusive images contain higher entropy (more information) than low-illusiveness images after a certain number of iterations. Classifiers trained with $S_{\mathrm{hc}} \cup S_{\mathrm{li}}$ are determined (smaller σ_A) and confident (higher P_c, P_i) regardless of whether they are right or wrong. By contrast, classifiers trained with $S_{\mathrm{lc}} \cup S_{\mathrm{hi}}$ are relatively hesitant (larger σ_A) in that the average probabilities for the output category (P_c, P_i) are lower; moreover, even if the prediction is wrong, they still assign a certain probability on the ground-truth category (P_g). The results demonstrate that bias can occurs within the same dataset because it is an intrinsic property of the learning scheme.

4.2. Training with a Random-Noise Class

Adding random-noise samples to training data is the easiest way to change the topology of the input space. To the best of our knowledge, researchers have never treated random noise as a collection of independent training samples containing information that can be directly employed for training neural networks. We repeated the experiments in the previous section with an extra category of random noise as negative training samples. Surprisingly, as shown in the right half of Table 1, the noise category improves the test accuracy. The risk of "misclassification as random noise" vanishes as the number of natural training images increases. In general, classifiers trained with random noise tend to be more determinant (higher P_c, P_s, lower σ_A); surprisingly, for misclassified samples, the probability of the ground-truth category also increases (higher P_g). Besides the enhancement, the existence of the unusual samples is more inspiring, and the samples may not be restricted to random noise. Such samples used to be considered as "off-the-manifold" and completely irrelevant to classification. Our experiment results leave open the possibility of seeking more hidden samples that may complement data augmentation, especially when training samples are insufficient.

4.3. Training in the Natural-image Spaces

We can also improve a classifier by slicing the learned connected regions to match the quotient-space model. Let us first recognize that a given network has limited, finite capability of learning the quotient space. Thus, there will often be training samples that are consistently misclassified during training, sometimes even with very high confidence. Henceforth, we will refer to these samples as illusive samples. Higher-order statistics of the training process, including the number of times that a training sample has been correctly classified, indicate crucial characteristics of the learned space. The following experiments were conducted based on the higher-order training statistics.

We retrained the network without illusive samples. As Figure 2 shows, despite a slight drop in test accuracy during early epochs, over-fitting is alleviated. If we further exclude illusive test samples, over-fitting is even more weakened. A correlation seems to exist between over-fitting and the illusive samples that cannot be mapped to correct equivalence classes. Moreover, the rationale behind dataset bias and over-fitting may be similar; over-fitting occurs at equivalence classes that are already observed whereas dataset bias happens at unseen locations in the input space. Another consequence of removing illusive training samples is that fewer uncertain regions are required to connect separated equivalence classes, which could potentially enhance the robustness against adversarial examples. We retrained the CIFAR-10 tutorial in CleverHans [5] and obtained results in Table 2 as expected.

In our last proof-of-concept, we computed the cumulative confusion matrix (CCM) during training. If an illusive training sample is consistently misclassified and appears to be top confusion (i.e., large values in the CCM), we will

Table 1: 50-time average performance of the classifier retrained with subgroups of training data. S_{c}^{p}: the top p of training samples selected with criteria c. The number of uniformly distributed noise samples is 5× the number of legitimate samples. Evaluation metrics are average test accuracy (A), standard deviation of test accuracy (σ_A), average confidence of correctly classified samples (P_c), average illusiveness of misclassified samples (P_i), average probability of the ground-truth category for misclassified samples (P_g), and average number of test samples that are misclassified as "noise" (N).

Training set	A	σ_A	P_c	P_i	P_g	Training set	A	σ_A	P_c	P_i	P_g	N
$S_{\mathrm{lc}}^{.25} \cup S_{\mathrm{li}}^{.25}$	27.12%	2.40%	26.56%	22.91%	14.65%	$S_{\mathrm{lc}}^{.25} \cup S_{\mathrm{li}}^{.25} \cup$ noise	30.82%	2.11%	28.31%	24.14%	15.43%	2.1
$S_{\mathrm{lc}}^{.25} \cup S_{\mathrm{hi}}^{.25}$	19.26%	0.92%	24.68%	22.86%	13.67%	$S_{\mathrm{lc}}^{.25} \cup S_{\mathrm{hi}}^{.25} \cup$ noise	21.32%	0.91%	25.27%	23.40%	14.30%	1.08
$S_{\mathrm{hc}}^{.25} \cup S_{\mathrm{li}}^{.25}$	53.61%	1.23%	79.67%	58.00%	12.73%	$S_{\mathrm{hc}}^{.25} \cup S_{\mathrm{li}}^{.25} \cup$ noise	55.27%	0.86%	79.24%	58.60%	12.86%	5.36
$S_{\mathrm{hc}}^{.25} \cup S_{\mathrm{hi}}^{.25}$	51.49%	1.29%	66.57%	44.77%	14.57%	$S_{\mathrm{hc}}^{.25} \cup S_{\mathrm{hi}}^{.25} \cup$ noise	53.26%	1.06%	67.22%	44.81%	14.87%	2.78
$S_{\mathrm{lc}}^{.50} \cup S_{\mathrm{li}}^{.50}$	52.16%	1.58%	45.73%	36.39%	18.61%	$S_{\mathrm{lc}}^{.50} \cup S_{\mathrm{li}}^{.50} \cup$ noise	53.11%	1.36%	46.70%	37.26%	19.23%	0.9
$S_{\mathrm{lc}}^{.50} \cup S_{\mathrm{hi}}^{.50}$	41.07%	1.18%	35.11%	30.86%	17.99%	$S_{\mathrm{lc}}^{.50} \cup S_{\mathrm{hi}}^{.50} \cup$ noise	40.82%	1.10%	35.05%	31.23%	18.55%	0.68
$S_{\mathrm{hc}}^{.50} \cup S_{\mathrm{li}}^{.50}$	65.78%	0.58%	88.08%	70.23%	11.41%	$S_{\mathrm{hc}}^{.50} \cup S_{\mathrm{li}}^{.50} \cup$ noise	66.37%	0.53%	88.56%	70.03%	11.61%	1.12
$S_{\mathrm{hc}}^{.50} \cup S_{\mathrm{hi}}^{.50}$	60.48%	0.67%	75.62%	50.50%	15.92%	$S_{\mathrm{hc}}^{.50} \cup S_{\mathrm{hi}}^{.50} \cup$ noise	65.13%	0.65%	76.10%	50.47%	16.16%	0.72
$S_{\mathrm{lc}}^{.75} \cup S_{\mathrm{li}}^{.75}$	68.76%	1.16%	74.25%	52.34%	18.36%	$S_{\mathrm{lc}}^{.75} \cup S_{\mathrm{li}}^{.75} \cup$ noise	70.03%	0.81%	73.88%	53.26%	18.42%	0.14
$S_{\mathrm{lc}}^{.75} \cup S_{\mathrm{hi}}^{.75}$	60.55%	1.31%	58.33%	43.00%	20.33%	$S_{\mathrm{lc}}^{.75} \cup S_{\mathrm{hi}}^{.75} \cup$ noise	61.95%	1.38%	54.95%	43.80%	20.69%	0.50
$S_{\mathrm{hc}}^{.75} \cup S_{\mathrm{ls}}^{.75}$	70.81%	0.31%	92.67%	76.09%	10.28%	$S_{\mathrm{hc}}^{.75} \cup S_{\mathrm{li}}^{.75} \cup$ noise	71.05%	0.54%	92.43%	75.64%	10.54%	0.22
$S_{\mathrm{hc}}^{.75} \cup S_{\mathrm{hi}}^{.75}$	70.85%	0.47%	80.45%	55.76%	16.24%	$S_{\mathrm{hc}}^{.75} \cup S_{\mathrm{hi}}^{.75} \cup$ noise	71.42%	0.55%	80.94%	55.75%	16.48%	0.34
S^1 (all)	74.67%	0.46%	83.22%	59.39%	16.49%	S^1 (all) ∪ noise	75.29%	0.36%	83.53%	58.61%	16.85%	0.14

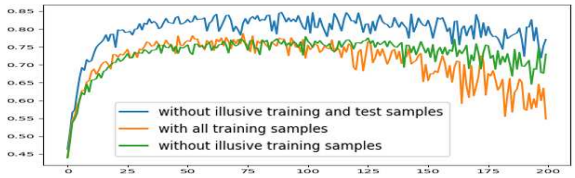

Figure 2: Training without consistently misclassified samples help prevent over-fitting.

Table 2: Test accuracy (in %) from CleverHans. n: number of training epochs; ϵ: max-norm eps; $A_{\mathrm{leg}}/A_{\mathrm{adv}}$: average test accuracy (20 runs) on legitimate/adversarial samples; $A^*_{\mathrm{leg}}/A^*_{\mathrm{adv}}$: average accuracy w/o illusive training samples.

n	ϵ	A_{leg}	A^*_{leg}	A_{adv}	A^*_{adv}	n	ϵ	A_{leg}	A^*_{leg}	A_{adv}	A^*_{adv}
6	.3	78.90	78.88	10.84	11.74	50	.3	90.39	88.81	10.37	10.40
6	.1	78.73	78.22	9.34	11.31	50	.1	90.51	88.77	10.75	14.35
6	.05	78.98	78.21	11.16	13.87	50	.05	90.53	88.91	13.28	21.88
6	.01	78.84	78.73	45.47	48.39	50	.01	90.33	88.90	39.03	48.50

relabel the sample with a new category. According to the quotient-space model, the network successfully learns that the illusive sample belongs to a particular equivalence class, but fails to connect the equivalence class with others in the same category due to limited capability of approximating discontinuous functions. Thus, we sliced the equivalence class with a new label, trained a separate classifier on those illusive samples, and glued the new equivalence class to existing ones with a look-up table. When illusiveness of test samples is available, we can select which classifier to apply at test time. According to test results, the proposed strategy can potentially increase test accuracy by 3% within current budget. In addition, we can train networks with larger learning rate because over-fitting is weakened by following the input space learned by the network.

5. Discussion

We hope the paper will stimulate discussion in the community regarding the intrinsic properties of the input space to which deep networks are applied. Open problems include the entropy of training samples, features of the decision boundaries, equivalence relation among images, and better representation of the image space. Understanding the topological and geometric properties of natural-image spaces with a more rigorous model will help us interpret the performance of state-of-the-art deep neural networks. Moreover, it may provide a more comprehensive understanding of the theoretical basis for deep neural networks. In practice, we may enhance the performance of neural networks by improving the quality of training samples or altering how we use data.

References

[1] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 1521–1528, IEEE, 2011.

[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[3] A. Fawzi, S.-M. Moosavi-Dezfooli, P. Frossard, and S. Soatto, "Empirical study of the topology and geometry of deep networks," in *IEEE CVPR*, no. CONF, 2018.

[4] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.

[5] N. Papernot, I. Goodfellow, R. Sheatsley, R. Feinman, and P. McDaniel, "cleverhans v1.0.0: an adversarial machine learning library," *arXiv preprint arXiv:1610.00768*, vol. 10, 2016.