Understanding Deep Learning Requires Rethinking Generalization

Wenhao Yang

School of Mathematical Sciences
Peking University

April 28, 2017

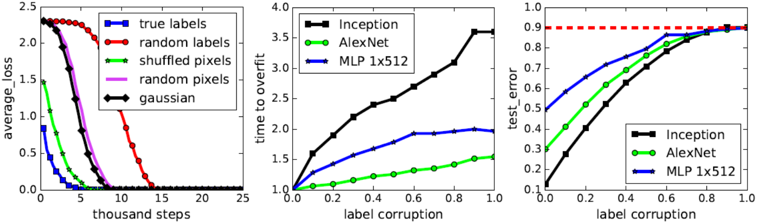
Outline

1. Deep neural networks easily fit random labels
2. Explicit Regularization
3. Finite Sample Expressivity
4. Implicit Regularization
5. Reference

Fitting Random Labels and Pixels

- **True labels:** the original dataset without modification.
- **Partially corrupted labels:** independently with probability p, the label of each image is corrupted as a uniform random class.
- **Random labels:** all the labels are replaced with random ones.
- **Shuffled pixels:** a random permutation of the pixels is chosen and then the same permutation is applied to all the images in both training and test section
- **Random pixels:** a different random permutation is applied to each image independently.
- **Gaussian:** A Gaussian distribution(with matching mean and variance to the original image dataset) is used to generate random pixels for each image.

Fitting Random easily fit random labels



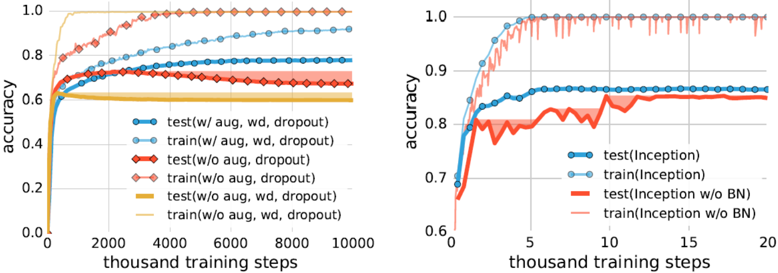
(a) learning curves (b) convergence slowdown (c) generalization error growth

Figure 1: Fitting random labels and random pixels on CIFAR10. (a) shows the training loss of various experiment settings decaying with the training steps. (b) shows the relative convergence time with different label corruption ratio. (c) shows the test error (also the generalization error since training error is 0) under different label corruptions.

The Role of Explicit Regularization

- Explicit regularization may improve generalization performance, but is neither necessary nor by itself sufficient for controlling generalization error.
- **Data augmentation:** augment the training set via domain-specific transformations. For image data, commonly used transformations include random cropping, random perturbation of brightness, saturation, hue and contrast.
- **Weight decay:** equivalent to a l_2 regularizer on the weights; also equivalent to a hard constrain of the weights to an Euclidean ball, with the radius decided by the amout of weight decay.
- **Dropout:** Mask out each element of a layer output randomly with a given dropout probability. Only the Inception V3 for ImageNet uses dropout in our experiments.

The Role of Explicit Regularization



(a) Inception on ImageNet

(b) Inception on CIFAR10

Figure 2: Effects of implicit regularizers on generalization performance. aug is data augmentation, wd is weight decay, BN is batch normalization. The shaded areas are the cumulative best test accuracy, as an indicator of potential performance gain of early stopping. (a) early stopping could potentially improve generalization when other regularizers are absent. (b) early stopping is not necessarily helpful on CIFAR10, but batch normalization stablize the training process and improves generalization.

Finite Sample Expressivity

- Generically large neural networks can express any labeling of the training data.
- E.g. A simple 2-layer ReLU network with $p = 2n + d$ parameters can express any labeling of any sample of size n in d dimensions: For weight $w, b \in R^n$ and $a \in R^d$consider the function:

$$c(x) = \sum_{j=1}^{n} w_j \max\{< a, x > -b_j, 0\} \tag{1}$$

Fix sample $z_1, .., z_n$and target $y_1, .., y_n$find a, b that satisfies $x_i =< a, z_i >$ and $b_1 < x_1 < b_2 < ... < b_n < x_n$. And find w that satisfies $y = Aw$ as A is a full rank matrix, where $A_{ij} = \max\{x_i - b_j, 0\}$

The Role of Implicit Regularization

- Blackboard

Reference

Reference:
[1] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, Oriol Vinyals; Understanding deep learning requires rethinking generalization