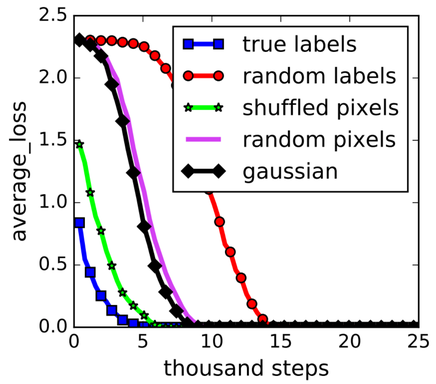
Bad Global Minima Exist and SGD Can Reach Them

Shengchao Liu — Quebec Artificial Intelligence Institute (Mila), Université de Montréal
Dimitris Papailiopoulos — University of Wisconsin-Madison
Dimitris Achlioptas — University of Athens

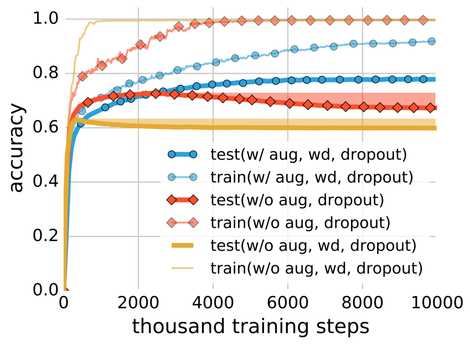
Rethinking Generalization [Zhang et al. ICLR'17]

Overparameterized, SGD-traind models:

- Can fit even completely random labels (i.e., huge capacity)



- Yet, generalize well



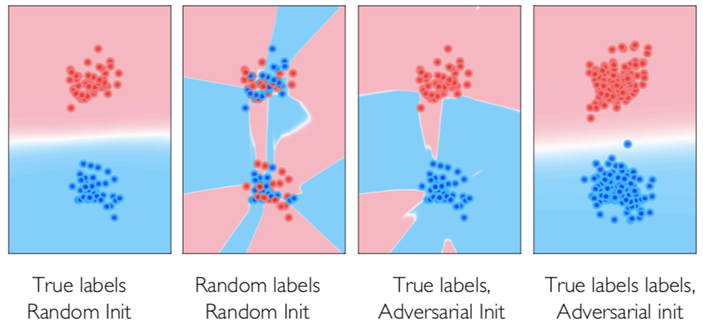
Possible Explanations

- Every model that fits the training data generalizes well (No bad global minima)
- SGD "avoids" bad global minima?

Our Results

- **Bad Global Minima Exist**
  - BAD Global Minimum: Permit fit on training data, but poor performance on test.
- **SGD Can Reach Them**
  - Very easy to construct initial conditions using only unlabeled data such that SGD converges to bad global minimum.
- **Explicit Regularization Affects Search Dynamics**
  - Regularization helps beyond telling apart good from bad global minima.
  - A combination of l2 and data augmentation allow SGD to escape bad initializers.

A Toy Example



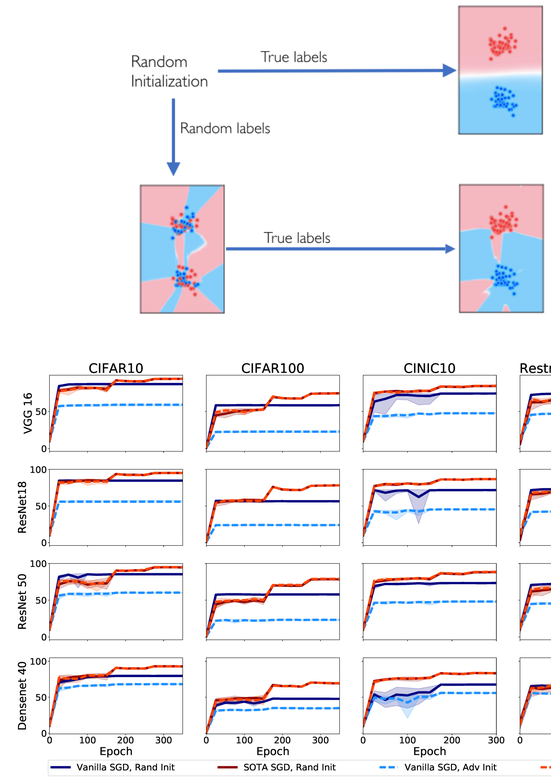
Confusion Data Generation

Algorithm 1 Creating the Confusing (Random) Data Set

Input: Original training dataset S; Replication factor R; Noise factor N
$C = \emptyset$
for every image $x \in S$ **do**
  for i from 1 to R **do**
    $x_i \leftarrow$ zero-out a random subset comprising $N\%$ of the pixels in x
    $y_i \leftarrow$ Uniformly random label
    Add (x_i, y_i) to C
Train the architecture to 100% accuracy on C from a random initialization using vanilla SGD
Output: The weight vector of the architecture when training ends

Experiments

- Datasets: CIFAR10/100, CINIC10 and Restricted ImageNet
- Models: VGG16, ResNet18/50 and DenseNet40



Findings and Conclusions

- Adversarial initialization causes SGD up to 40% drop in the test accuracy.
- The model found is close to the adversarial initialization.
- Data augmentation (DA), momentum (M), and l2 regularization all contribute to SGD escaping adversarial initialization.
- And two of {DA, M, l2} are enough.

Codes are available at https://github.com/chao1224/BadGlobalMinima

Email: liusheng@mila.quebec