**An Empirical Analysis of Boosting Deep Networks**

Sai Saketh Ramabhatla1, Michael Jones2, Rama Chellappa1,3
1University of Maryland, College Park 2Mitsubishi Electric Research Laboratories 3Johns Hopkins University

**Goal:** Compare accuracy of a *boosted ensemble* of Deep Neural Networks with the accuracy of a *single* large Deep Neural Network with same number of parameters.

**Introduction**

- Boosting is a method for finding a highly accurate hypothesis by linearly combining many “weak” hypotheses, each of which may be only moderately accurate.
- Boosting can be applied to any classifier and AdaBoost has been proven to reduce the training error as more weak classifiers are added to the ensemble.
- Boosting was studied extensively with decision trees, and a large ensemble of decision trees has better performance than a single decision tree on the test set (“win”).
- **Missing in current literature:** Analysis on whether an ensemble of MLPs or CNNs is a “win” in terms of decreasing the testing error below what is achievable with a single network with the same number of total parameters as in an ensemble.

**Key Takeaway:** Better off training a *single* large network than a *boosted ensemble* of small networks.

### AdaBoost

- AdaBoost maintains a set of weights per training example.
- On each round of boosting, the weight on each example is updated with a specific equation that gives less weight to examples the weak classifier got right and more weight to examples it got wrong.
- The next weak classifier will be forced to classify more of the incorrect examples correctly.
- For AdaBoost, at round \( t \), the equation to update weights is

\[
 w_{i,t+1} = w_{i,t} e^{-\alpha_i m_i / Z_t}
\]

**Base Architectures**

- **CNN**
  - LeNet style
  - 5954 trainable parameters
- **MLP**
  - Two hidden layers
  - 41088 trainable parameters
- Boosting a base classifier \( N \) rounds makes the total number of parameters \( N \times \) the number of parameters of the base classifier.
- For single model, only width (# of filters per layer) is increased to increase parameters (not depth).
- CNN experiments are run five times and the results are averaged
- Two different optimizers and three boosting algorithms were used – SGD and Adam

### Datasets and Boosting algorithms

**Datasets**

- MNIST
- CIFAR-10
- CIFAR-100
- SVHN

**Boosting algorithms**

- AdaBoost
- SAMME
- LogitBoost

**Task:** Classification  **Metric:** Accuracy

### Experiments

**Decision Trees**

- Single large Decision Trees overfit while the boosted ensemble does better on all three datasets
- With the same number of leaves, ensemble is better than a **single large tree**

**MLP**

- Comparison of Boosted and Single MLP on CIFAR100 dataset
- Number of Parameters (x10^9)

**CNN**

- Comparison of Boosted and Single CNN on CIFAR10 dataset
- Test Accuracy

**VGG-8**

- Number of Parameters (x10^9)

---

**Visit the paper for more information:**

- Boosting is a method for finding a highly accurate hypothesis by linearly combining many “weak” hypotheses, each of which may be only moderately accurate.
- Boosting can be applied to any classifier and AdaBoost has been proven to reduce the training error as more weak classifiers are added to the ensemble.
- Boosting was studied extensively with decision trees, and a large ensemble of decision trees has better performance than a single decision tree on the test set (“win”).
- **Missing in current literature:** Analysis on whether an ensemble of MLPs or CNNs is a “win” in terms of decreasing the testing error below what is achievable with a single network with the same number of total parameters as in an ensemble.

**Key Takeaways**

- Single large Decision Trees overfit while the boosted ensemble does better on all three datasets
- With the same number of leaves, ensemble is better than a **single large tree**