


Understanding Deep Learning Requires Rethinking Generalization

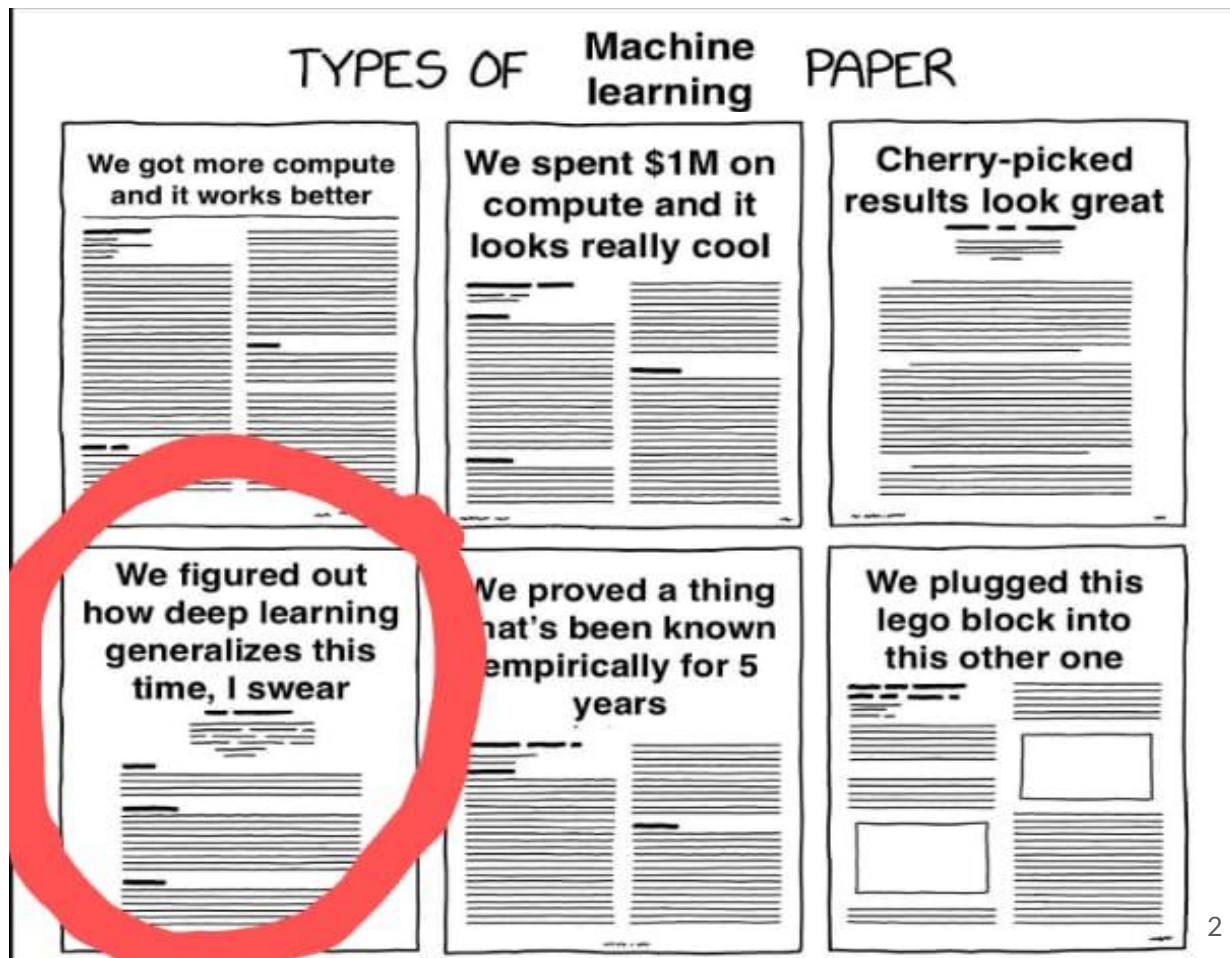
Team: Okay

- Madhav Agarwal (2020900022)
- Siddhant Bansal (2019900091)
- Piyush Singh (2020701031)

Github: <https://github.com/mdv3101/Okay>

Paper: <https://arxiv.org/pdf/1611.03530.pdf>


You think you
know how deep
learning works?



Source - Internet

Introduction



- Deep Neural Networks have the ability to generalise well on the given dataset.
- They show very small difference in the training and test performance.
- But the questions to consider here are:
 - What gives them the ability to generalize well?
 - Are they learning the ‘actual’ association?
 - Can we train the network to learn random associations?
 - Does regularization play a significant role while training a deep network?
- Due to the way in which neural networks are formulated, we cannot use traditional methods (VC dimension, uniform stability, etc.) to understand the generalization ability.

Introduction



- This work shows that the traditional view of generalization is *incapable of distinguishing between different neural networks that have different generalization performance*.
- For that purpose we do the following experiments:
 - **Randomized testing**
 - As the name suggests, here the neural networks are provided a different set of inputs and outputs and the performance is checked.
 - **Role played by *explicit* regularization**
 - Here we try to understand how well various regularization techniques like, weight decay, dropout, and data augmentation, are playing a role in generalization.
 - **Role played by *implicit* regularization**
 - Implicit regularization techniques such as batch norm, SGD and early stopping are applied to check the performance of neural network

Randomized Testing

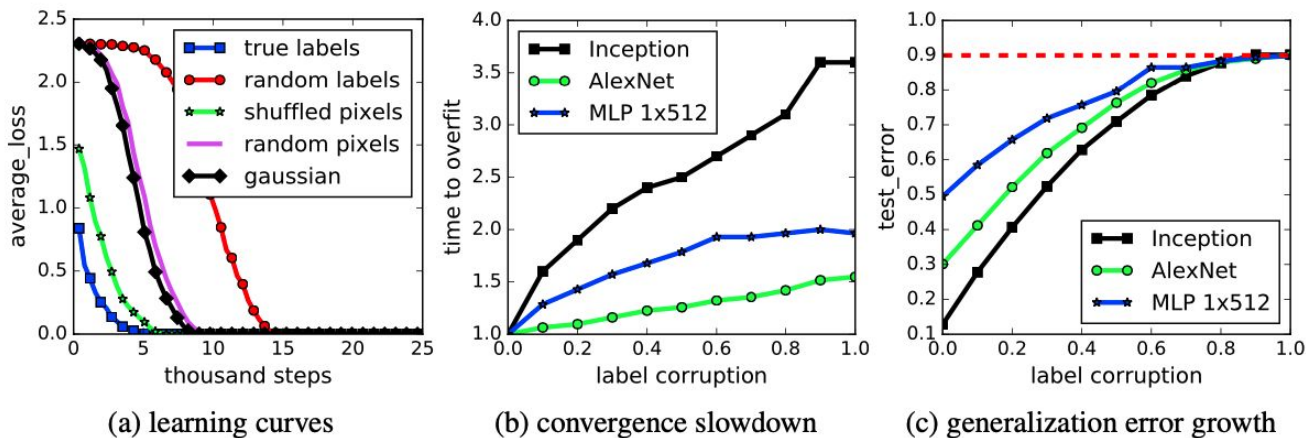


Figure 1: Fitting random labels and random pixels on CIFAR10. (a) shows the training loss of various experiment settings decaying with the training steps. (b) shows the relative convergence time with different label corruption ratio. (c) shows the test error (also the generalization error since training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

Role played by regularization

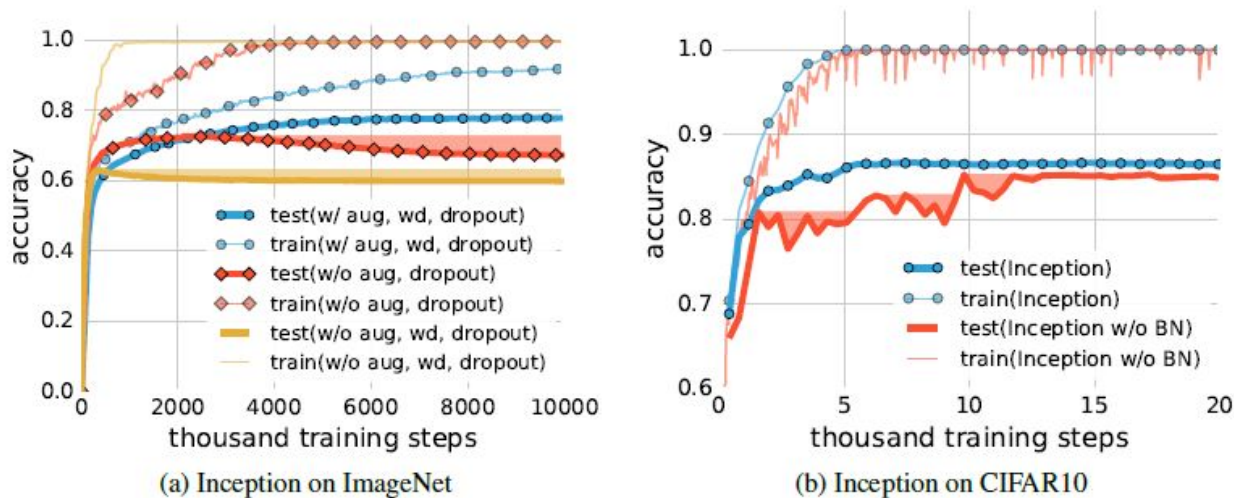


Figure 2: Effects of implicit regularizers on generalization performance. aug is data augmentation, wd is weight decay, BN is batch normalization. The shaded areas are the cumulative best test accuracy, as an indicator of potential performance gain of early stopping. (a) early stopping could potentially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

Datasets

CIFAR-10

- Dataset contains 50,000 training and 10,000 validation images
- Total Number of classes: 10
- Image size is 32x32 with 3 channels (cropped to 28x28)

airplane

automobile

bird

cat

deer

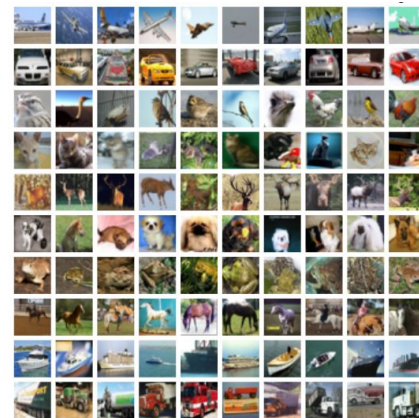
dog

frog

horse

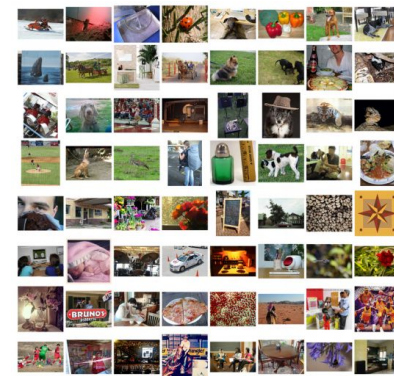
ship

truck



ImageNet

- Dataset contains 1,281,167 training and 50,000 validation images
- Total Number of classes: 1000
- Image resized to 299x299 with 3 channels



Randomized Testing



- Here the goal is to see how well the neural networks generalize to data which is not 'logical' for us humans.
- To make sure that the results are what is being observed, we perform these set of experiments:
 - Using **true labels** as a baseline.
 - **Corrupting the labels** by replacing a label with a random class's label.
 - **Shuffling pixels** of the input image.
 - **Randomizing the pixels** of the input image.
 - Creating **noisy images** using Gaussian noise while using the variance and mean of the dataset.
- Logically, it feels like apart from the first case, the models should not converge but, as we will see in the slides ahead, the results are surprising.

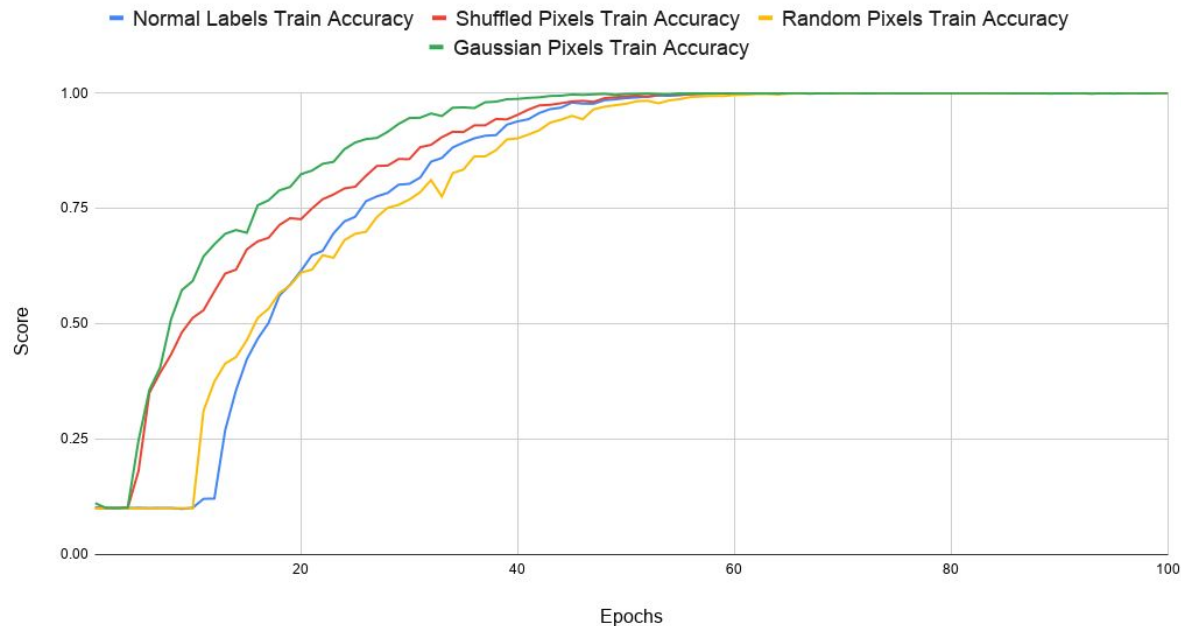
Randomized Testing



- The following trends are observed in the graphs in the original paper
 - With zero input and zero output corruption, the training converges in very few number of iterations
 - The time to converge for different corruption scenarios has the order:
No corruption < Pixel shuffling < Gaussian pixels < Random pixels < Random labels
 - For label corruption experiment, the time to converge increases with increase in the percentage of corrupted labels
 - For label corruption experiment, the generalization error increases with increase in the percentage of corrupted label
- In all randomization experiments, the model is able to “shatter” the training data ie achieve 100% accuracy with sufficient number of iterations.

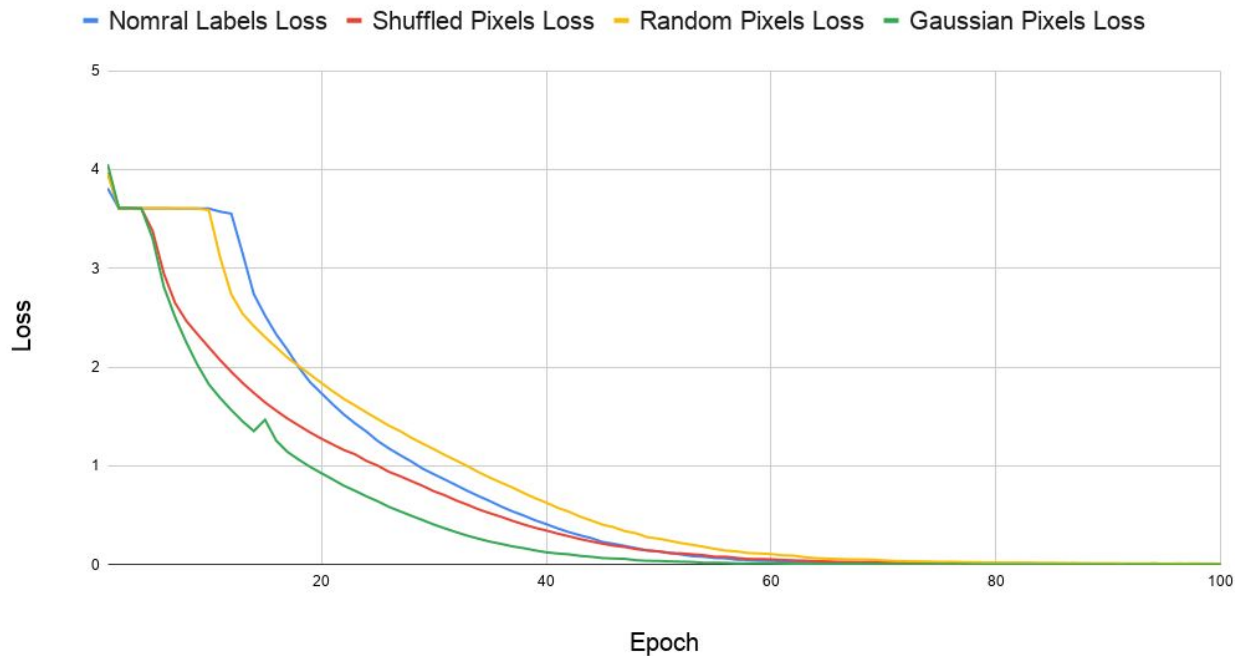
Effect of different inputs

Effect of different inputs on training the Neural Network



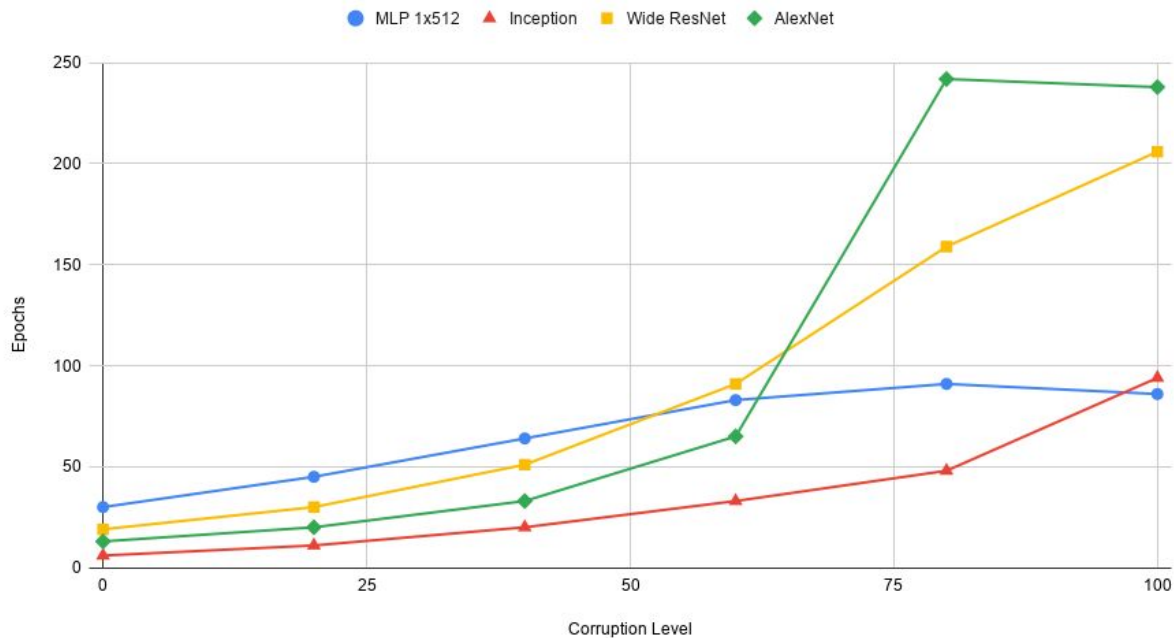
Effect of different inputs

Effect of different inputs on convergence



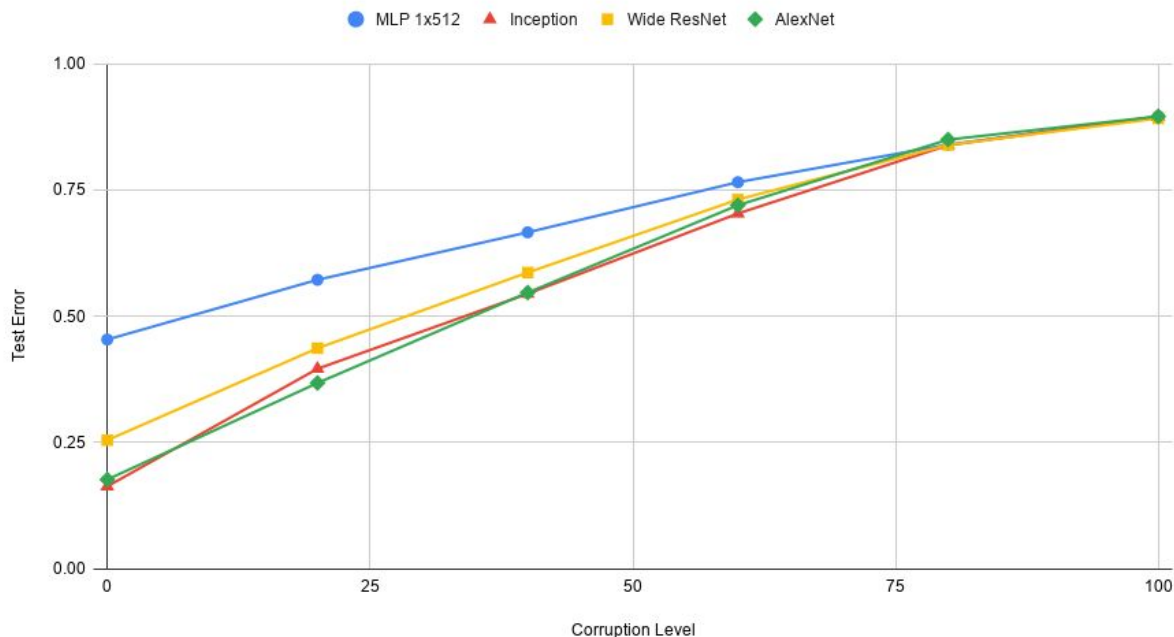
Label corruption: convergence

Epochs to converge vs label corruption level



Label corruption: Generalization Error

Generalization error vs label corruption level



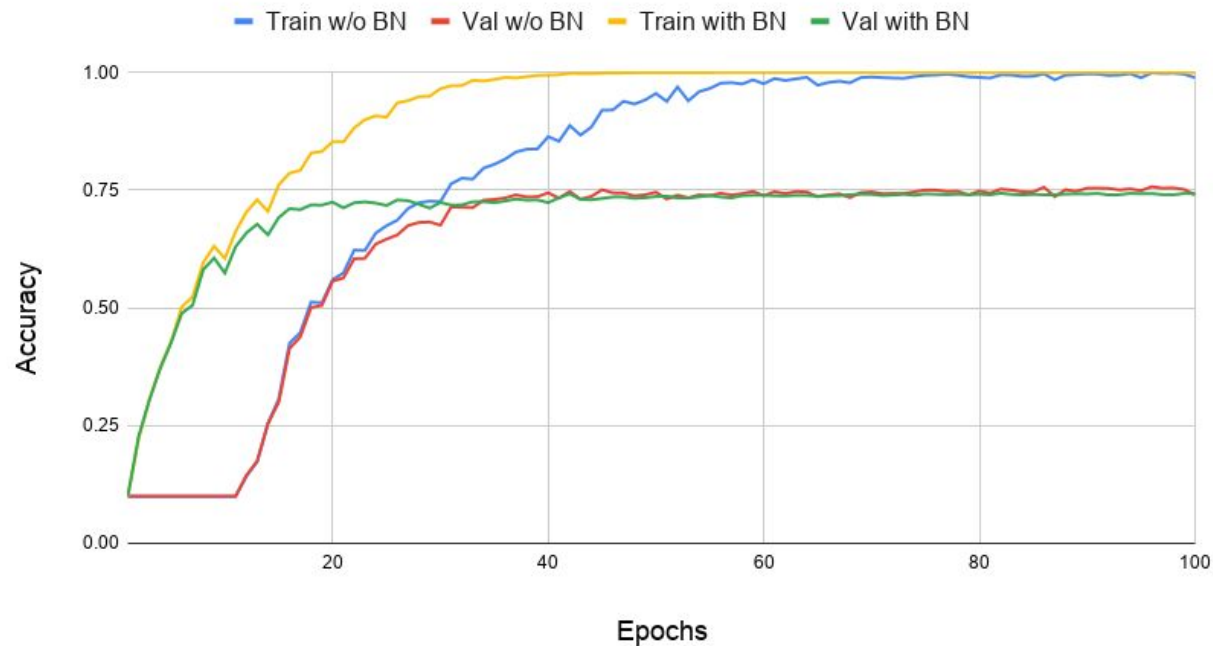
Role played by regularization



- We study the role of applying regularization on model training and its ability to generalize.
- There are some techniques like Batch Normalization which are not meant for regularization but indeed they gave regularization effect (Implicit Regularization).
- Other commonly used regularization techniques like weight decay, dropout and data augmentation are explicitly used to induce regularization (Explicit Regularization).
- The author argues that “Explicit regularization may improve generalization performance, but is neither necessary nor by itself sufficient for controlling generalization error.”
- The experiments suggests that tuning parameters help in improving generalization. The effect of regularization is minimal on generalization.

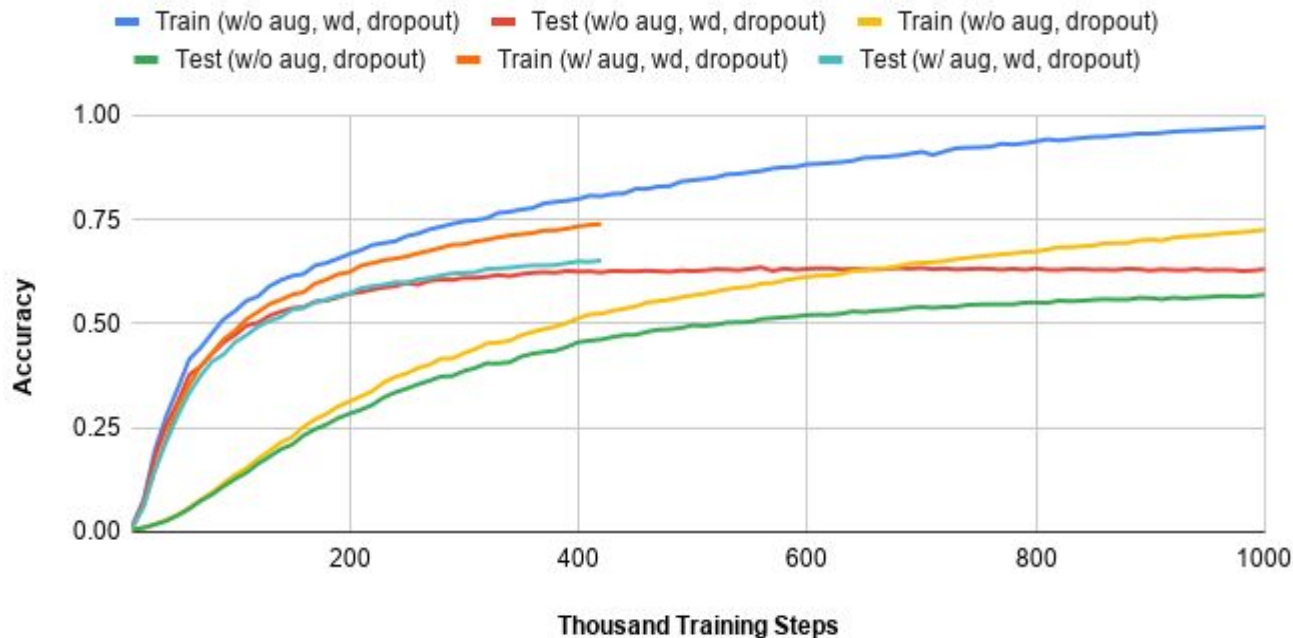
Our experiments on Regularization

Effect of Implicit Regularisation on Generalization Performance



Our experiments on Regularization

Effect of Explicit Regularization on Generalization Performance



Project Summary



- **Aim:** To understand what differentiates neural networks that generalize well from those that do not
- **Datasets:** CIFAR10, ImageNet
- **Models:** MLP-512, Inception (tiny), Wide ResNet, AlexNet, Inception_v3
- **Experiments done:**
 - Effect of explicit regularization like augmentation, weight decay, dropout
 - Effect of implicit regularization like BatchNorm
 - Input data corruption: Pixel shuffle, Gaussian pixels, Random pixels
 - Label corruption with different corruption levels from 1 to 100 %

Conclusion



- Neural networks have the capacity to memorize massive training datasets and achieve near 100% training accuracy with enough training iterations. This is irrespective of both corruption in input data and corruption in output labels
- With input or output corruption, the convergence time goes up only by a small factor
- The progressive label corruption experiment shows that NNs capture remaining signal in the uncorrupted portion of training data and overfit on the corrupted portion
- Explicit regularization may improve generalization performance, but is neither necessary nor by itself sufficient for controlling generalization error
- Implicit regularization like SGD and BatchNorm are also not the primary reasons for generalization capability. They only helps in making training process smooth.
- We have yet to discover a precise formal measure which we can use to construct small models with good generalization capacity. As of now it is based on experimentation

Individual Contribution

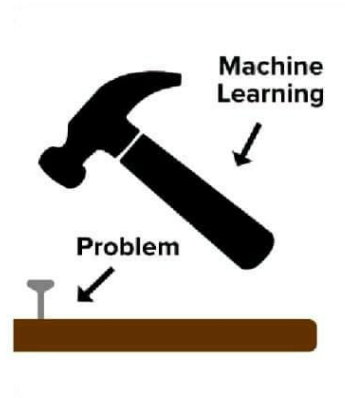


Name	Task
Madhav Agarwal (2020900022)	Implemented the small Inception model, trained InceptionV3 on ImageNet, studied the effect of implicit regularization on CIFAR-10 dataset and, the effect of explicit regularization on the ImageNet dataset.
Siddhant Bansal (2019900091)	Studied the effect of data manipulation on the CIFAR-10 dataset using Inception (small), experiments on normal labels, random pixels, Gaussian noise, and random labels.
Piyush Singh (2020701031)	Studied the effect of label corruption ratio on model convergence and test error using AlexNet, MLP-512, Inception(small) and WideResnet on the CIFAR-10 dataset.

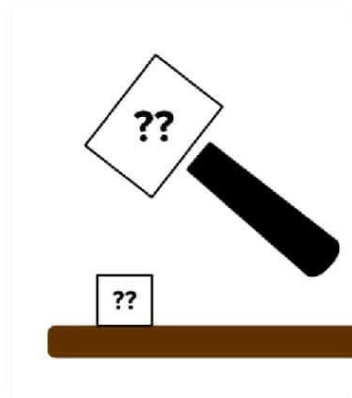


Thank You

how it started



how it's going



You know Nothing, Jon Snow!