Analysis of the Relation between Shannon Entropy and Generalization Gap

Mentor: Tolga Dimlioglu **School:** NYU Tandon School of Engineering
Name: Sumit Dhar (The Dalton School), Su Pyae Sone Win @ Samantha Sue (West End Secondary School)

INTRODUCTION

The field of Deep Learning (DL) has been advancing rapidly in different directions such as Computer Vision, Natural Language Processing, etc. Despite these advancements, the geometry and the properties of the loss landscape of the DL models are still not well explored and understood.

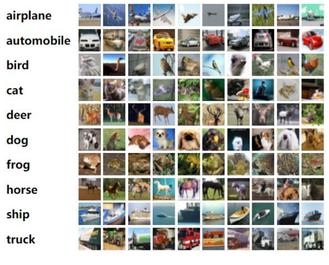
This project's aim is to investigate the correlation between the generalization capability of the model's converged local minimum with the Shannon Entropy metric that is calculated on the DL model's output neurons. We will be using Kendall's Tau Correlation Coefficient to calculate the correlation. Our hope is to better understand the characteristics of the generalization gap, and to find an effective way to reduce it.

Definitions:
- Shannon Entropy: A measure of uncertainty of our model's predictions; calculated using output probabilities
- Generalization Gap: The difference between test set error and train set error.

**Loss/Error function: Cross Entropy

DATASETS

- **CIFAR-10 dataset:** a collection of 60,000 images sorted into 10 classes (airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks)



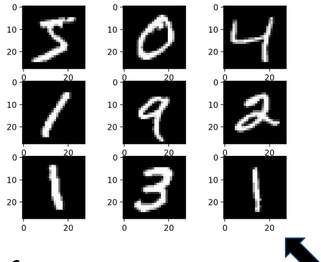


Figure 1

- **MNIST dataset** - a collection of 70,000 images of handwritten digits (0-9)

Models

- Residual Neural Network (ResNet): a convolutional neural network with multiple layers
- Multilayer perceptron (MLP): a simpler neural network model

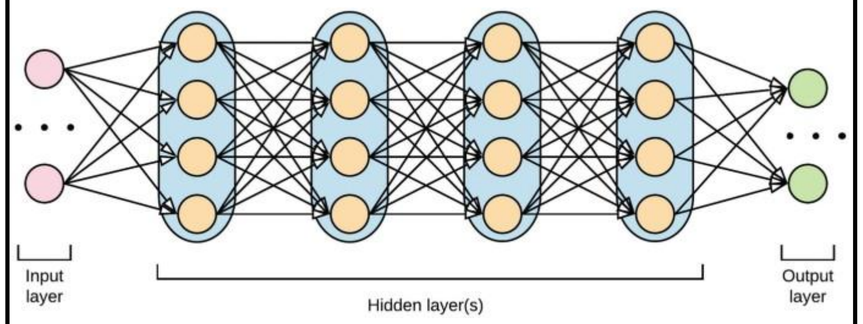


Figure 2

Experiment Settings & Training Procedure

Tuned parameters

- Learning Rate: 0.01, 0.005
- Momentum: 0.1, 0.5, 0.9
- Weight Decay: 0, 0.0001
- Batch size: 32,128, 512
- Resnet width: 4, 6, 8
- Skip connection: True, False } ResNet
- MLP width: 4, 8
- MLP hidden neurons: 256, 1024 } MLP

**We tested every permutation of the varying parameters (MLP: 144 experiments; ResNet: 216 experiments) above on 4 separate GPUs.

**We trained our two models with 200 epochs each using Stochastic Gradient Descent as our optimizer. We also decreased the learning rate by 10-fold after 100 and 150 epochs.

**Train errors greater than 1 were disregarded (focusing on good performance results)

Potential limitation(s):
Different seeds weren't used in the experiment settings. It is good practice to use 3 - 4 different seeds.

RESULTS

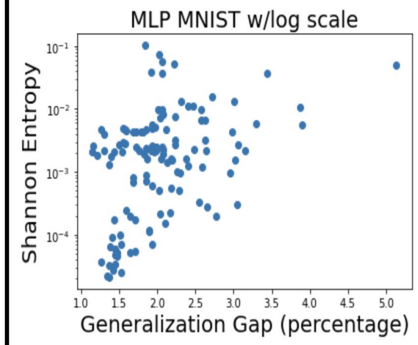


Figure 3.Kendall's τ correlation coeff: 0.28

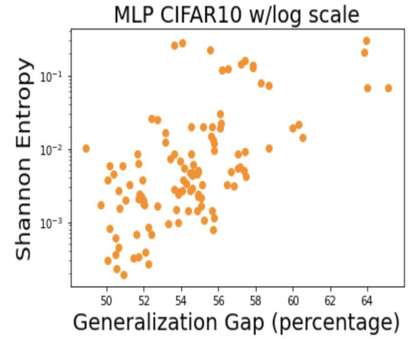


Figure 4.Kendall's τ correlation coeff: 0.41

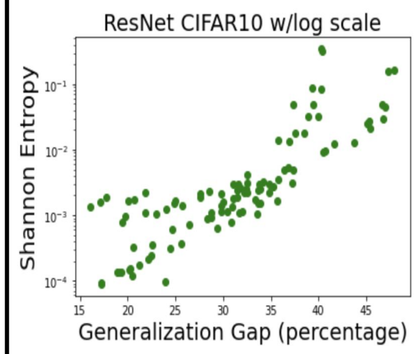


Figure 5. Kendall's τ correlation coeff: 0.70

Kendall's τ correlation coefficient	Strength of correlation
0 - 0.1	very weak
0.1 - 0.2	weak
0.2 - 0.3	moderate
0.3 - 1	strong

Figure 6. Kendall's τ correlation coeff table

CONCLUSION

There is a moderate to strong correlation between the Generalization Gap & Shannon Entropy. As Shannon Entropy increases (more uncertainty of the model's predictions), the Generalization Gap increases.

To apply our new understanding of the relation between Shannon Entropy and the Generalization Gap, our *next steps* would be to incorporate Shannon Entropy into the loss function to see if it's an effective method to train the model.

CONTACTS

Tolga Dimlioglu: td2249@nyu.edu
Sumit Dhar: sd5319@nyu.edu
Su Pyae Sone Win (Samantha Sue): sw5496@nyu.edu