DATA AUGMENTATION FOR NEURAL NETWORK OPTIMAL GENERALIZATION

Abdurashitova Muniskhon
Department of Control and Computer Engineering, Turin Polytechnic University in Tashkent, Uzbekistan

https://doi.org/10.5281/zenodo.7739565

Abstract. *To expand the size of a real dataset, data augmentation techniques artificially create various versions of the original dataset. Following their application, many techniques and methods have demonstrated an improvement in the precision of machine learning models. It serves as a regularizer during machine learning model training and aids in lowering overfitting. This article will cover the application of data augmentation techniques based on different noise types that shows improved neural network performance.*

Keywords: *Neural Networks, Data Augmentation techniques, Pink Noise, White Noise.*

I OBJECTIVE

The ability of neural network models to generalize what they have learned to new, unobserved data can be improved with the help of augmentation methods, which can offer variations of dataset. Data augmentation techniques artificially generate different versions of a real dataset to increase its size. Many practices have shown an increase in the accuracy of machine learning models after applying them. According to an experiment, a deep learning model after image augmentation performs better in training loss & accuracy and validation loss & accuracy than a deep learning model without augmentation for the image classification task. In this article, during the training of neural networks, we will employ data augmentation techniques to enhance the robustness of model inference generalization.

II INTRODUCTION

Data augmentation refers to a set of strategies for creating new training samples from existing ones by introducing random variations and disturbances, ensuring that the data is not destroyed. By adding random variations and disturbances, a set of techniques known as "data augmentation" allows for the creation of new training samples from existing ones without erasing the original data. Our goal is to increase the model's capacity for generalization through data supplementation. We cover the creation of augmented sets using the original data along with combined Gaussian white noise and pink noise. Because it is popular in industries like energy management, health tracking, and security, we are investigating human indoor localization in a constrained setting. The goal is to investigate low-cost but effective methods for indoor localization because GPS technology exists for outdoor person tracking and can quickly pinpoint a person's position using a wearable tag or cell phone. People don't always carry phones with them when they are at home or inside a room, so we should look for tagless localization methods. In a 3x3 meter room, four sets of experimental data are gathered for a brief amount of time. An ultrasound-based tag from the Marvelmind Starter Set HW v4.9 is used to gather the person's reference location from a 4x4 pixel Omron D6T-44L-06 thermopile infrared sensor that is mounted on a room's ceiling. There are 18 components in each tuple of testing data. The X and Y coordinates of the individual displaying the label are added to the 16 pixels of the infrared sensor (IR)[1].

III MAIN PART

The model architecture and input data are designed based on the state-of-the-art. Next, we explore the augmentation parameters that improve best the neural network generalization performance. For each combination of parameters, the neural network model is trained 30 times with the same input data because the training process is stochastic. Each time model weights are initialized with random variables resulting in different output values for each training. Out of these 30 independent trainings, the best one is chosen based on the generalization ability. To determine the generalization capability of the neural network model, we use four sets of experimental data, denoted as A, B, C, D. They are collected in different days, for different movement trajectories, and in different environmental conditions. As a metric for the generalization capability is calculated the sum of the mean square error (MSE) for each of the four sets, A, B, C, and D, as follows:

$$OverallPerformance = MSE_A + MSE_B + MSE_C + MSE_D \quad (3.1)$$

We consider only the training with the lowest out of the 30 that we do for each parameter set. First, we generate a baseline against which to track the generalization improvement. The baseline is obtained by training the model with original data without augmented sets. Our aim at generating augmented sets is to acquire improved generalization compared to the model training without augmentation. Noise amplitude that generates augmented sets is related to the signal amplitude, i.e., between the signal when a person is detected and when there is no person. We explore thus the behavior of the model with noise amplitudes at a fraction of the maximum signal, starting from 1 %, and increasing until the effect of the noise is detrimental for the generalization (at around 12 %).

IV EXPERIMENTAL ANALYSIS

In general, all the systems are simultaneously affected by different kinds of noises. We have seen the model behavior that is affected by white and pink noises separately in our previous works [1][2]. To better model real-world settings, we explore the effects of combinations of white and pink noises. We use the results of our previous training. We choose the interval [0; 0.2] for white noise ranges, which is proven as a maximum noise amount improving the model overall generalization. While for the pink noise possible range is proven as [0; 1]. We use white and pink noise amplitudes as two parameters describing the model MSE represented in 3D plots. Let us divide the chosen intervals of white and pink noises as follows: white noise amplitude ranges are selected as {0.001 0.025 0.050 0.075 0.100 0.125 0.150 0.175 0.200} and pink noise amplitudes are considered as {0.01 0.21 0.41 0.61 0.81 1.01 1.21 1.41 1.61}. The model is trained 30 times and out of 30 trainings, the best model which shows the best model generalization is chosen. Overall results are obtained as the sum of MSEs of four sets using equation *(3.1)* and the baseline is calculated as the sum of individual baselines as below:

$$OverallBaseline = Baseline_A + Baseline_B + Baseline_C + Baseline_D \quad (3.2)$$

Figure 1 describes the model overall generalization ability considering all the sets. The flat blue surface represents the sum of the baselines of individual sets, and the irregular surface is obtained as the sum of all independent sets MSEs. The rough surface has several local minimum points, which are considered points of best generalization because the smallest MSE below the baseline is the best improvement point. In *Figure 1*, all local minimum points are not visible due to overlapping other parts of the surface. We change the representation of the rough surface in another way to enable analyzing the promising areas. We calculate the difference between the baseline and the irregular surface using new formula:

$$NewSurface = Baseline_A + Baseline_B + Baseline_C + Baseline_D - (MSE_A + MSE_B + MSE_C + MSE_D) \quad (3.3)$$

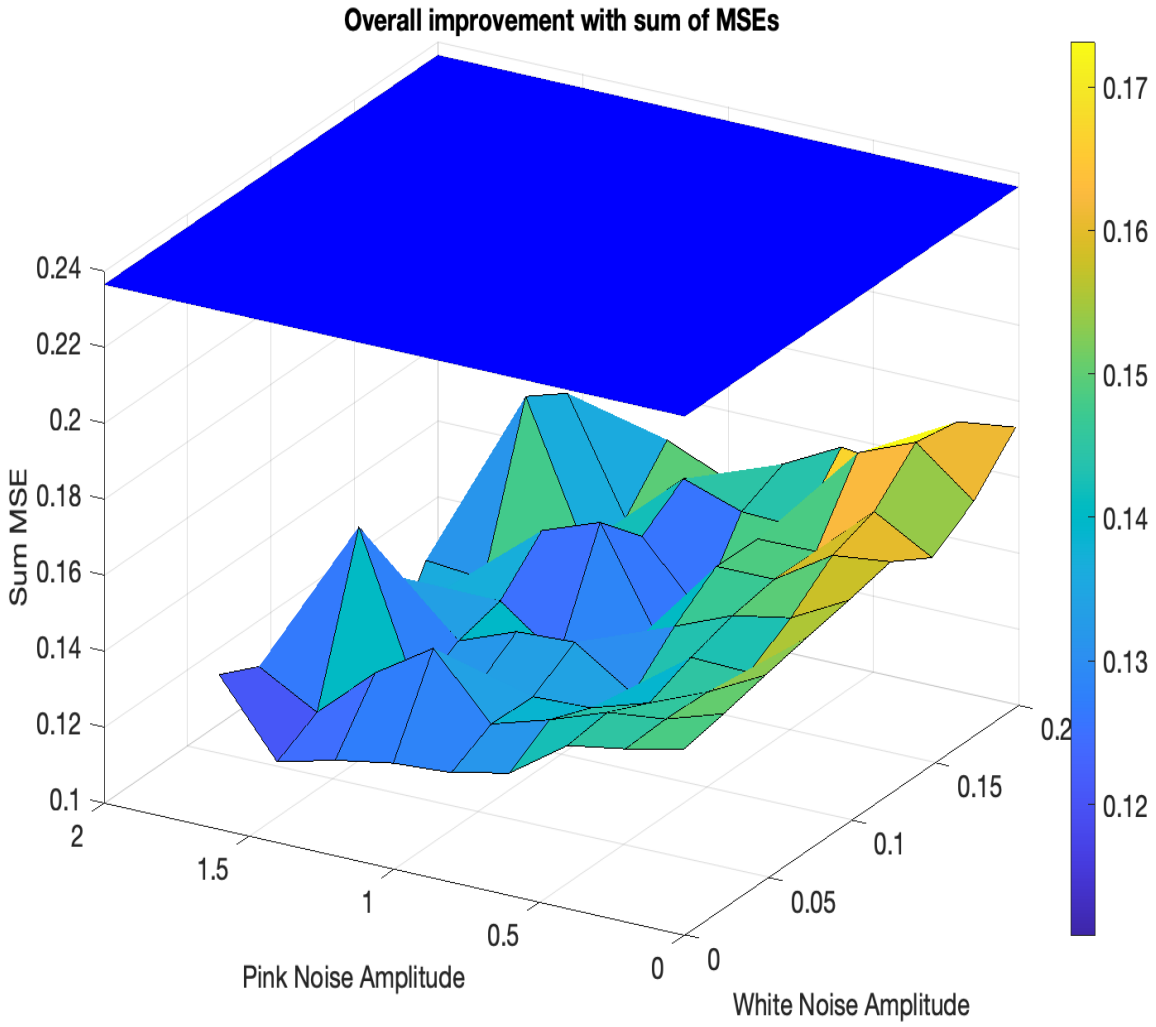

Figure 1: *Comparison of the model overall generalization (obtained by the combination of white and pink noise amplitudes) with the baseline. Flat surface represents overall baseline.*

The longest distance (or the most prominent difference) is the most promising point due to the smallest MSE sum. After calculating the difference for all the surface points, we generate another surface in which calculated deviations are placed on the original coordinate surface of the X and Y axes. As a result, we obtain another surface in *Figure 2* where local maximum points refer to the model best generalization. The next step is to check each local maximum point representing better model generalization and exclude false local maxima in which one or two sets of improvements exceed the terrible results of the other sets. In *Figure 2*, we can see the peaks highlighted with star symbols that are obtained as a sum of individual improvements of all sets.

V CONCLUSION

he combination of two noises is investigated simultaneously, representing the model characteristics in 3D for white and pink noise amplitudes. From the *Figure 2*, the model improvement by induced noise levels is visible in the area covered by white noise amplitude range [0; 0.2] and pink noise amplitude range [0; 1]. The diagonal part of the diagnosed surface provides the model with the best improvement points.

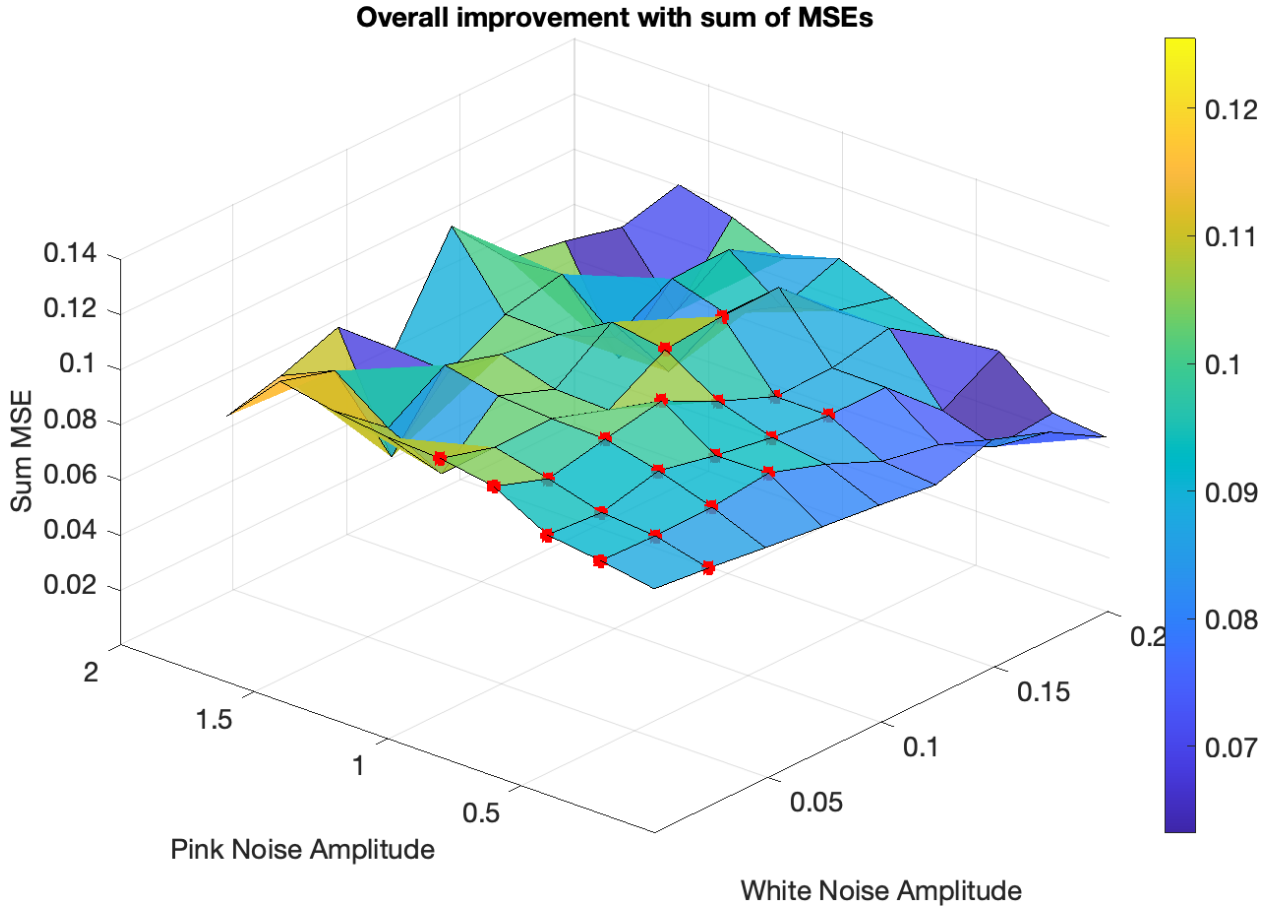

Figure 2: *Comparison of the model overall generalization with the baseline.*

REFERENCES

1. Abdurashitova, M. (2022). The Neural Networks Performance Improvement With Gaussian White Noise Augmentation. *Acta of Turin Polytechnic University in Tashkent*, *12*(1), 24–27. Retrieved from https://acta.polito.uz/index.php/journal/article/view/134
2. Abdurashitova, M. (2022). The Neural Networks Performance Improvement Using Pink Noise Augmentation. *Acta of Turin Polytechnic University in Tashkent*, *12*(2). Retrieved from https://acta.polito.uz/index.php/journal/article/view/139