CNNs are the most widely used models in Deep Neural Networks (DNNs) [1]. They have demonstrated excellent performance in several fields, such as Image Recognition [2], Semantic Segmentation [3], Speech Recognition [4], Object Detection [5] and Natural Language Processing [6]. Since CNN models require great amount of data and computational skills, their success was attributed to dramatic growth in the amount of data as well as an increase in the computational speed [7], [8], [9]. Based on these factors, researchers and practitioners were able to train deeper networks ([10], [11], [12]), which resulted in a higher increase in the number of the CNN models.

However, as the networks get larger, it becomes much more complicated to conduct smooth training. This results in the activation values of each layer being exponentially small or large. These phenomena are called activations gradient vanishing or exploding respectively ([13], [14], [15]). Therefore, in designing the model architecture, it is advisable to take proper precaution since the deeper model does not always mean better performance.

CNN comprises forward and backward passes. Both of them greatly depend on matrix multiplication. In the forward pass, each hidden layer executes multiplication between a specific part of the input and kernel weight matrix. It creates feature maps, which act as inputs for the next layer. Regarding back-propagation [16], the same matrix multiplication operations are executed in the opposite direction. A typical sort of CNN has a large number of layers, thus there are numerous matrix multiplication operations in it. The only available parameter that requires modification is the weight matrix.

Suppose we initialize the weight values from the same standard normal distribution as the inputs, then the model will produce exploding activation values. However, if the initial weight values are small, then we obtain a vanishing activation problem. It will also decrease gradient, which in turn will hinder weight updates and consequently lead to poor learning process [17].

To understand the best strategy for weight initialization that prevents the activations from either exploding or vanishing, we conduct theoretical and practical research. The main contributions of this work are as follows:

- We provide mathematical background to identify the most optimal weight initialization strategy and show its practical implementation in graphs.
- We propose the WIB-ReLU activation function, which performs better for normally distributed values.
- We demonstrate the positive effect of the proposed method on the training process of a CNN model in two different databases.

In the research, we obtain that the most optimal method is to make the activations have standard normal values with mean 0 and standard deviation 1. However, we obtain that the most widely used activation function, which is ReLU, increases the mean value with certain units. In this paper, we address this issue by proposing a method that keeps the activations mean and standard deviation close to 0 and 1 respectively. From the experiments, we obtain that the proposed Weight Initialization Based ReLU (WIB-ReLU) function performs better than regular ReLU in several metrics.

The paper is organized as follows. In Section 2, we present related research works. Section 3 contains the description of the proposed methodology. Section 4 contains the information concerning the experiments and their results. Finally, Section 5 summarizes the work with a conclusion. Directions for further study are suggested as well.