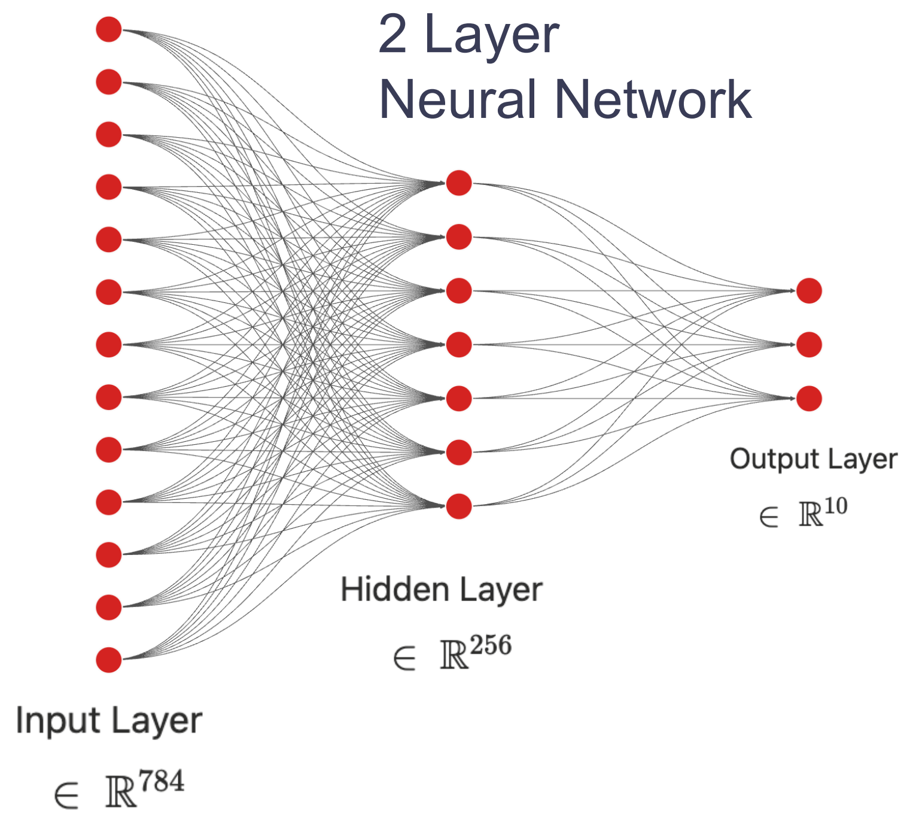
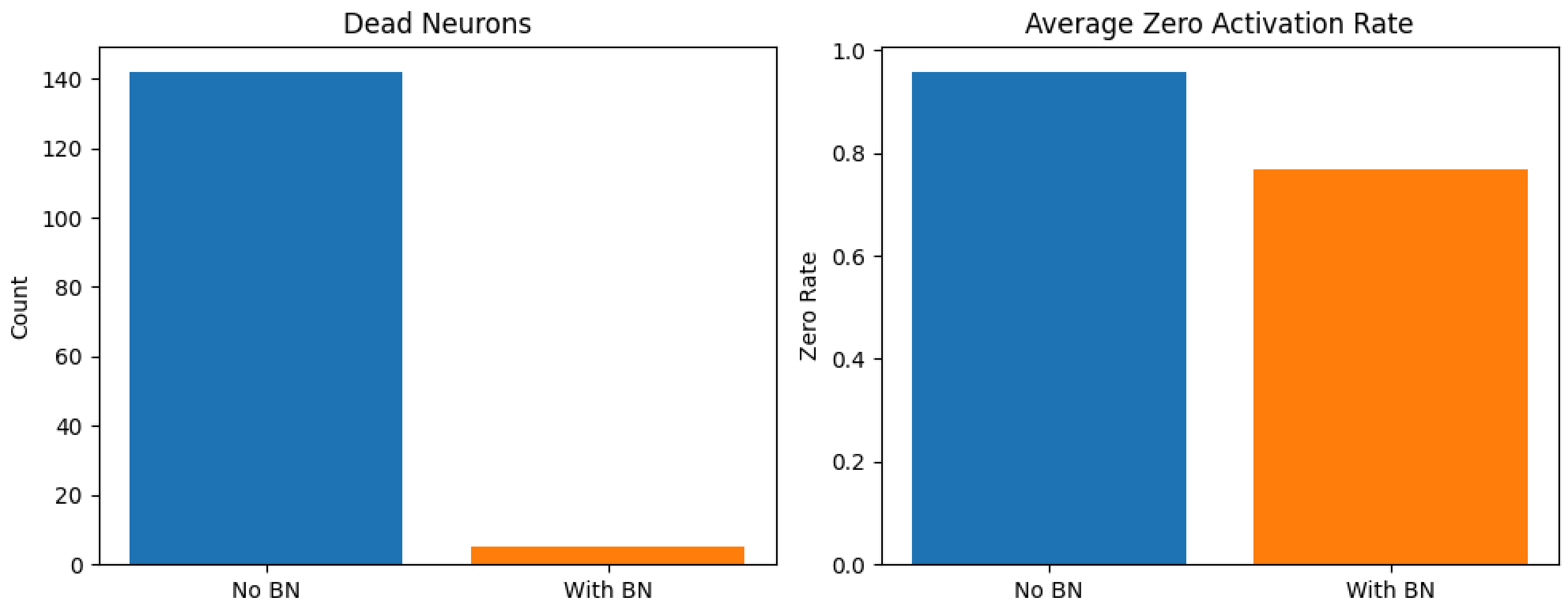
BatchNorm Under Stress Testing - Still Worth it ?

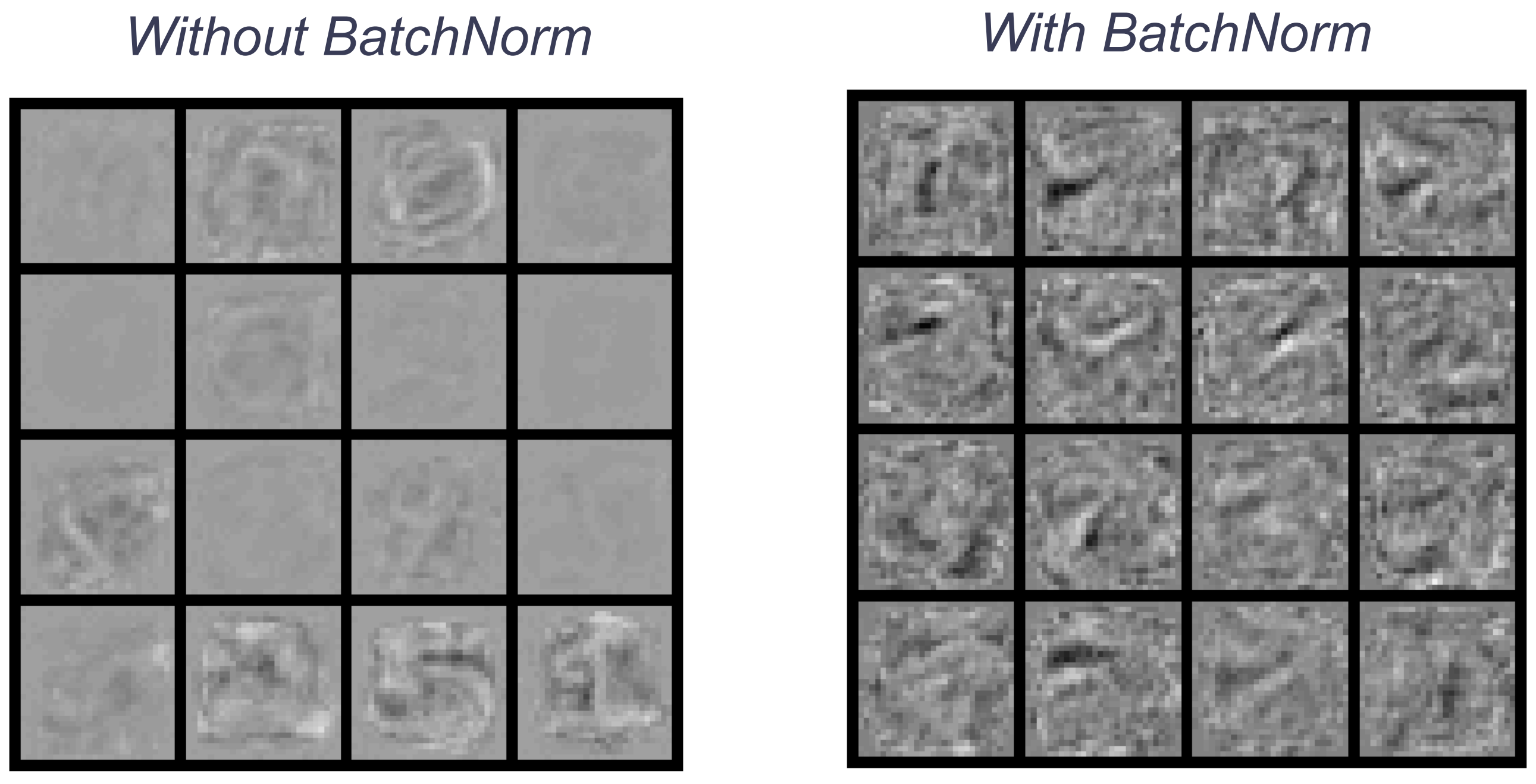
MNIST Data



Bar Chart of Dead Neurons Ratio and the Average Zero Activation Rate



First Layer's Weights on a 2D grid.

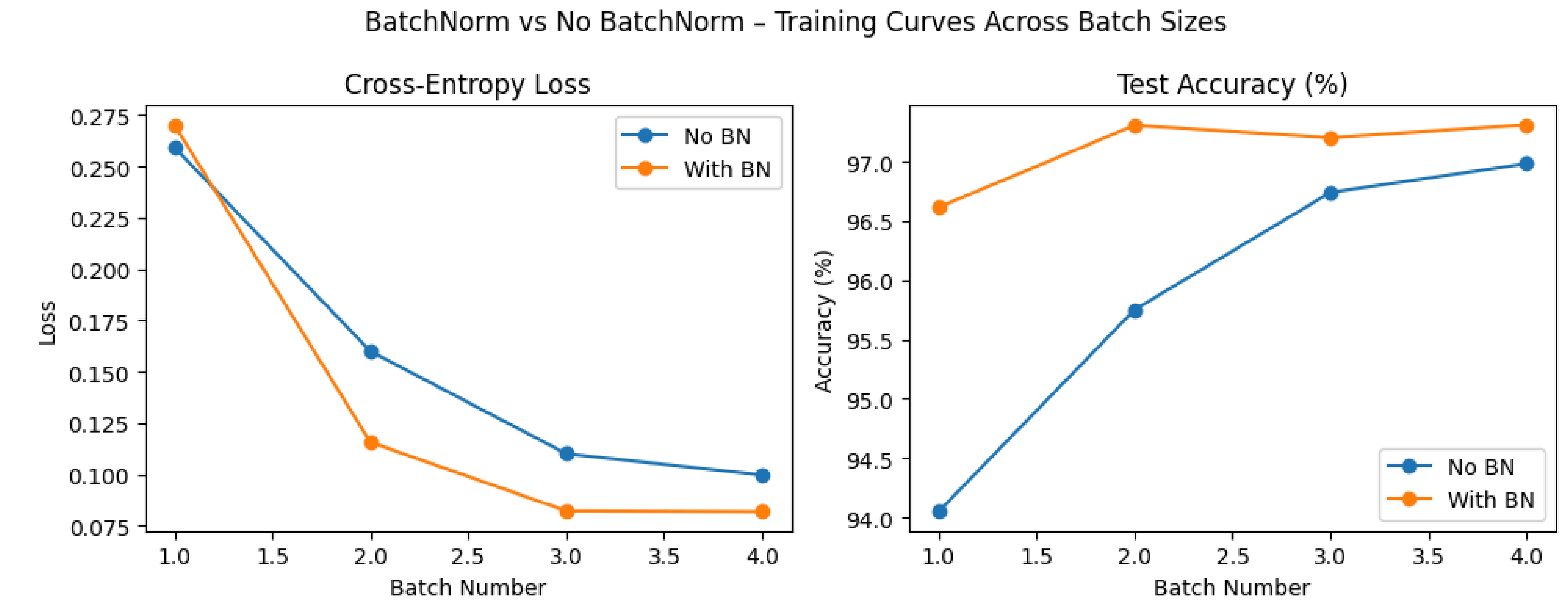


Results

The non-normalized architecture shows a substantially higher proportion of dead neurons and a higher average zero activation rate across the network. This behavior can also be visually seen by examining the first 16 neurons in the first layer of both models as an example.

Fixed Settings Across each single Dataset:
Learning rate, Train/Test size, Loss Type = Cross Entropy, Optimizer = Adam, ...

Performance of NN using BatchNorm vs Non BatchNorm Across different Batch Sizes

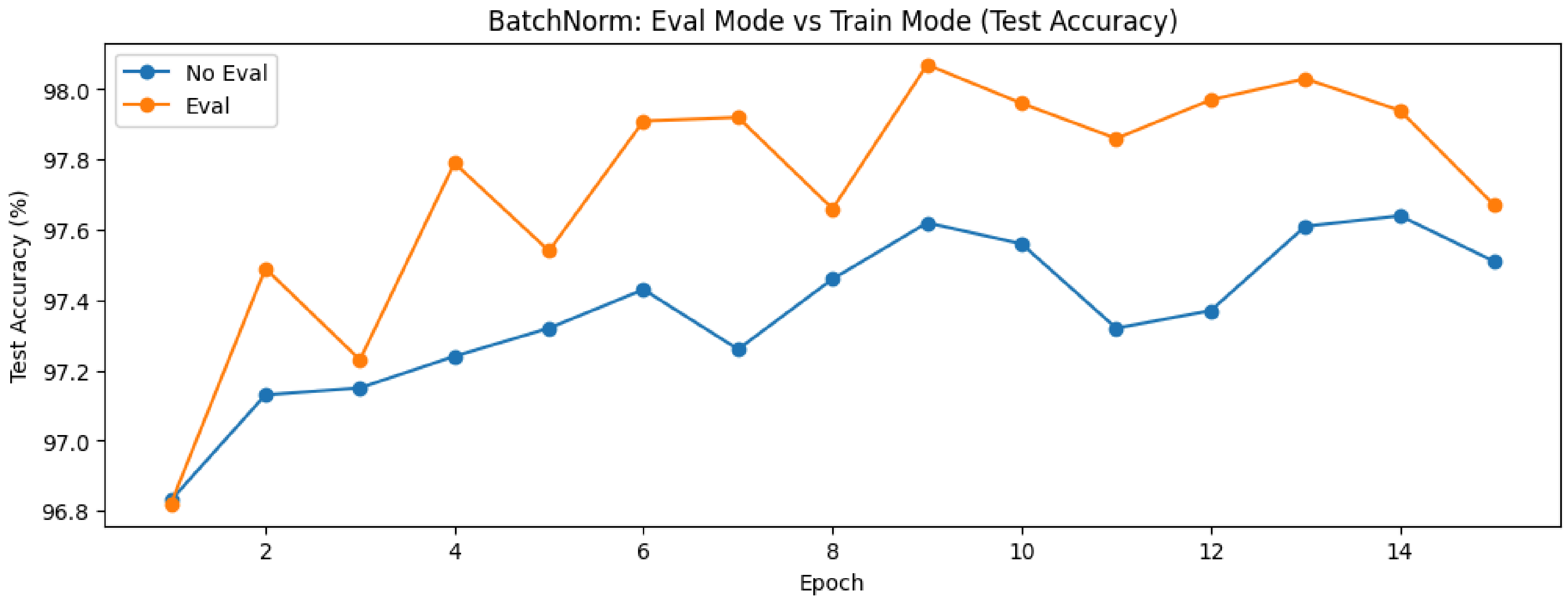


The Batch Number from 1 to 4 represent batch sizes respectively of [8, 32, 128, 256].

The performance difference between normalized and non-normalized models is big for small batch sizes and decreases as the batch size increases. This behavior is not seen in the training loss -> Better Generalization from Batch Normalization.

We could explore smaller batch sizes but that would require more computing power due to more frequent updates.

BatchNorm in Evaluation Mode VS in Non-Evaluation Mode

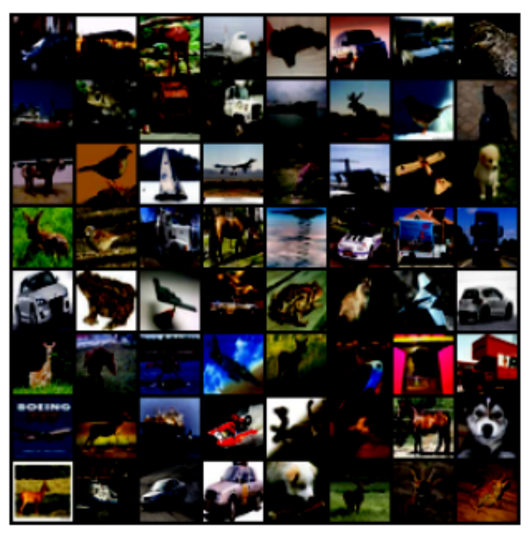


Results

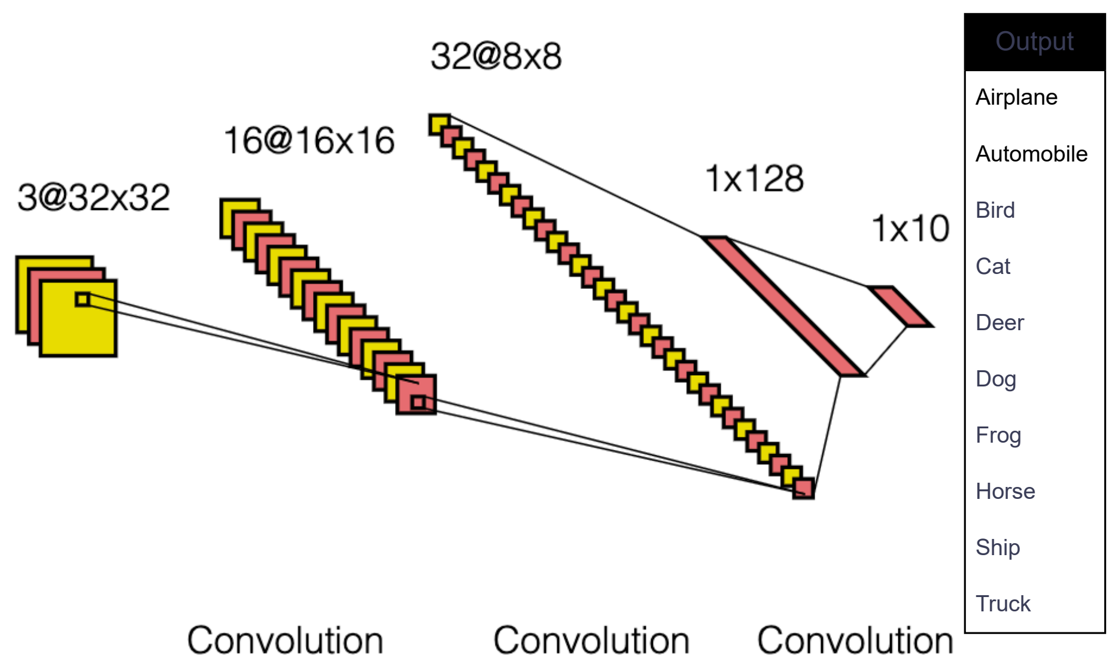
The accuracy is lower for the model in which model.eval() was not called before testing. Calling model.eval() activates the use of the running mean and variance in Batch Normalization layers and prevents these statistics from being updated. On the MNIST dataset, allowing the running statistics to be updated during testing degrades generalization performance.

We Proceed with a Second Experiment of Another More Complicated Dataset.

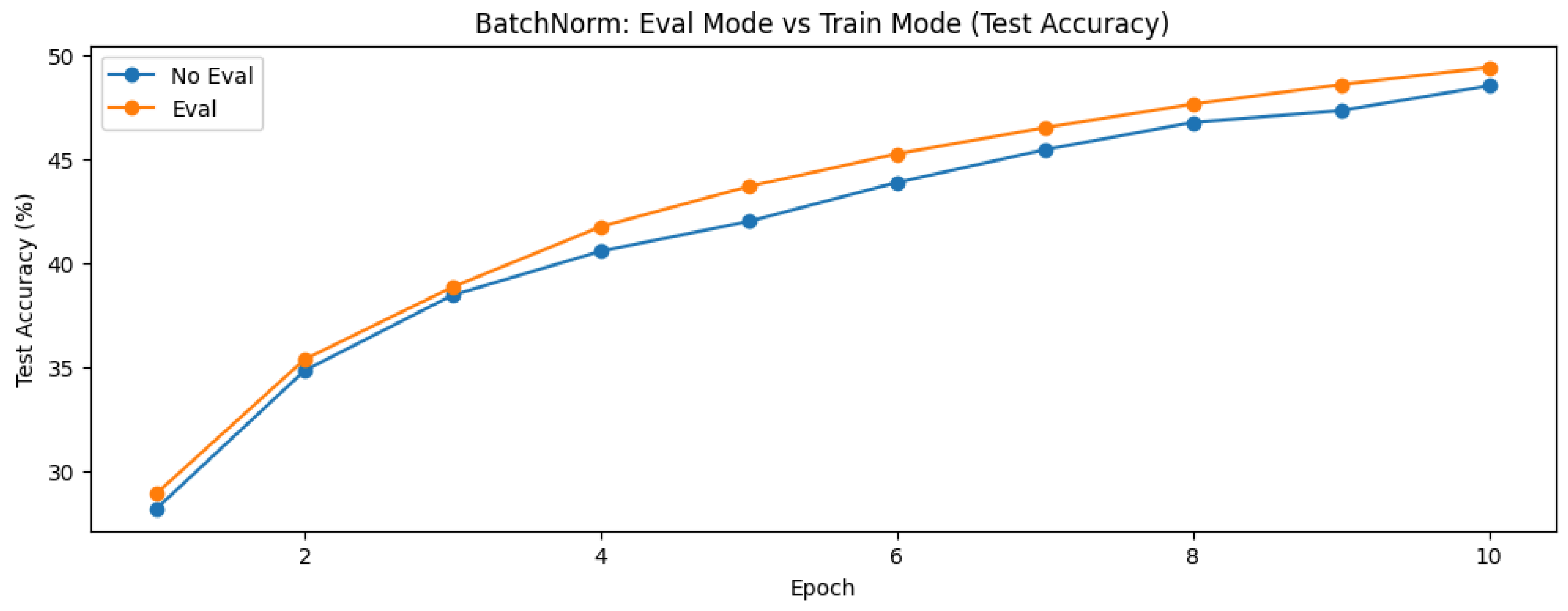
CIFAR10 Data



NN Architecture



BatchNorm in Evaluation Mode VS in Non-Evaluation Mode

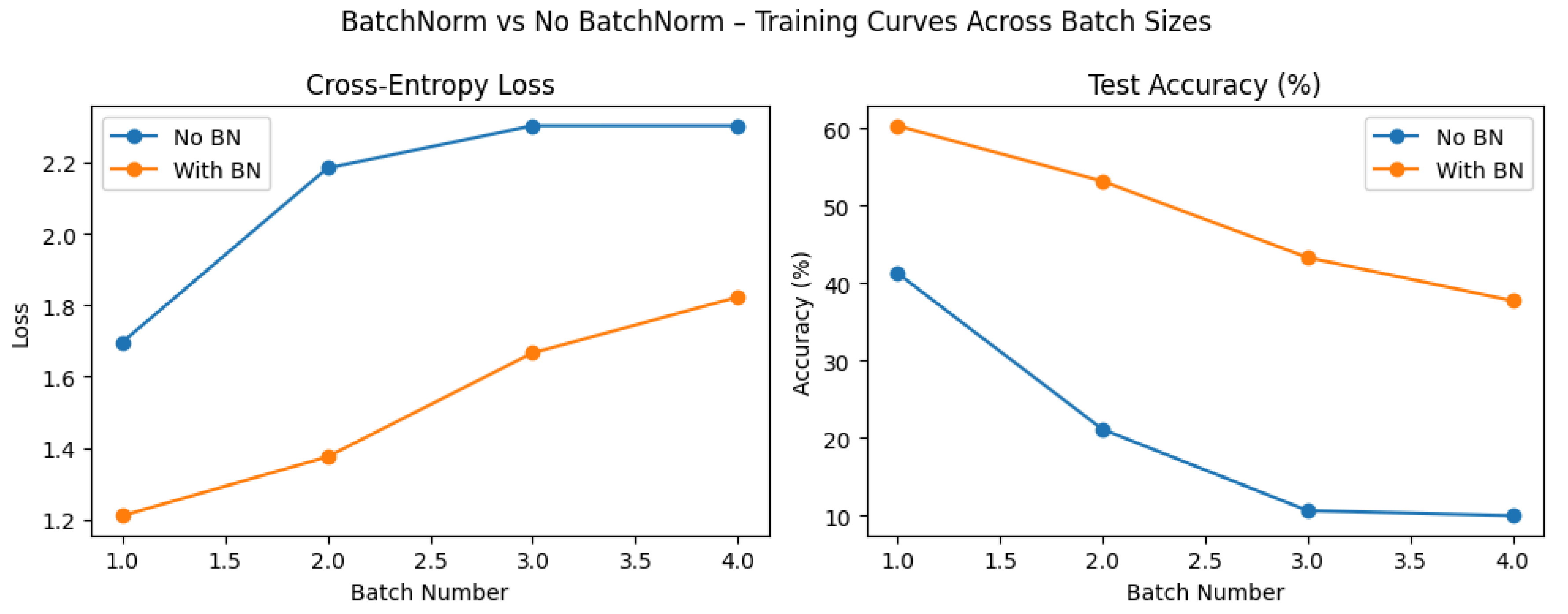


Results

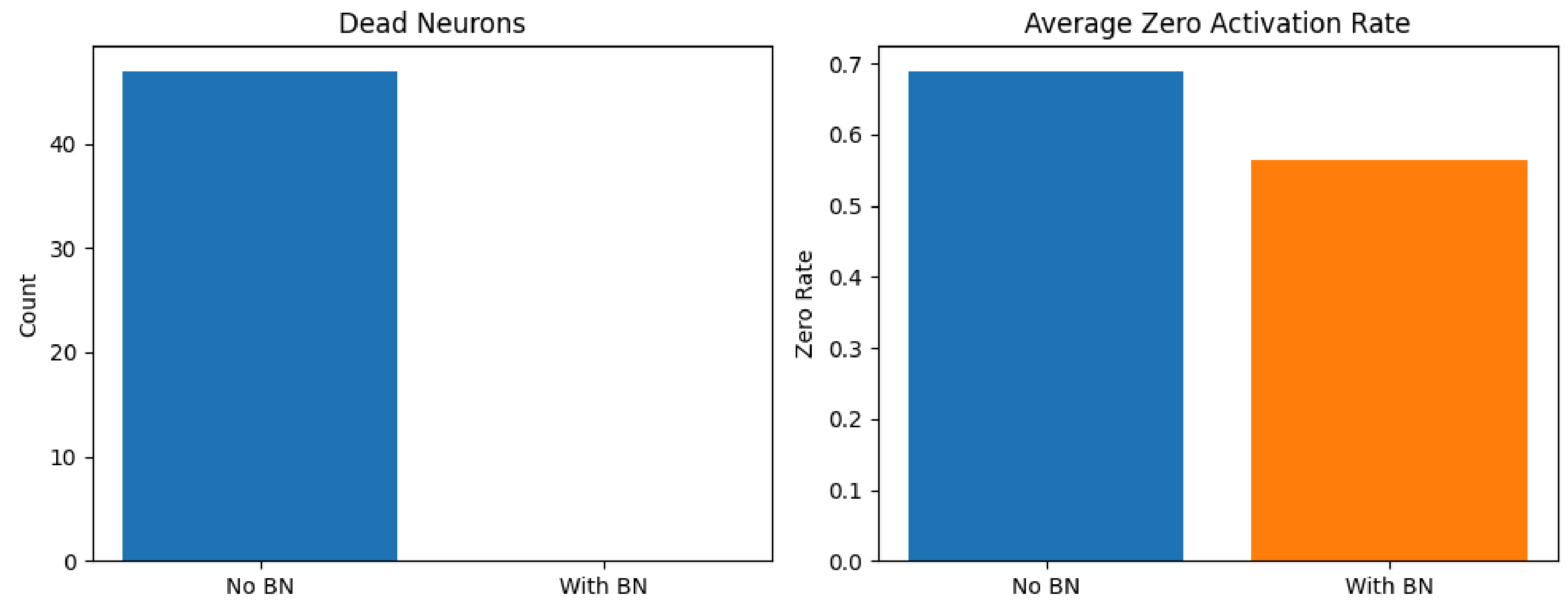
In this architecture, invoking model.eval() continues to yield improved generalization performance. Nevertheless, the magnitude of the improvement is modest, and without a statistical analysis, it is unclear whether the observed difference is statistically significant.

Author: Fedi Ghanmi

Performance of NN using BatchNorm vs Non BatchNorm Across different Batch Sizes



Bar Chart of Dead Neurons Ratio and the Average Zero Activation Rate



Results

For a more complex dataset and a deeper architecture, the model exhibits better generalization when trained with smaller batch sizes, with accuracy consistently higher for the normalized model compared to the non-normalized one. A similar pattern is observed in neuron activity. While the non-normalized model contains approximately 45 dead neurons across the architecture, the batch-normalized model exhibits none. On average, the batch-normalized network has about 50% zero activations, whereas this proportion rises to nearly 70% in the non-normalized model. Since Batch Normalization have zero dead neurons, these zero activations are concluded to be distributed across neurons rather than being concentrated in specific units.