Towards Understanding Learning Representations: To What Extent Do Different Neural Networks Learn the Same Representation

Abraham Jose
ID :5068109, CAP6614
abraham@knights.ucf.edu

1 Summary

A study of how similar the representations learned by two networks with identical architecture, trained from different initialization of the weights using the neuron subspace match model. The neuron activation subspace match model algorithm finds the maximum match and simple match between two identical neural networks to find the similarity in the learning representations learned.

2 Strengths of the proposal

1. They have provided a complex analysis for the neuron activation subspace match model for proving its correctness and it provide us with an intuitive way to analyse their methodology.
2. This study can potentially help us to understand how neural network works and to revisit the optimization techniques that we believe will optimize to the global minima. Dissecting the neural networks and the features using these techniques can help us to understand better.

3 Weaknesses of the proposal

1. For the representations/features learned, we should be looking at the similarity in the features irrespective of the skew or the distortion for the same input in the identical neural networks. Subspace model is a point to point model, which cannot accommodate for the slight distortions or the skewness of similar features.
2. L_2 distances may not be ideal difference to find the matches between two layers in identical neural networks with different initialization.
3. Randomness in the input data set or the amount of augmentation or the characteristics of a batch during the testing has to be discussed for the better understanding of the experiment carried out.

4 Results

It was observed that the there is very little or no similarity,using the subspace model in the middle layers of the neural network when experimented with the VGG and Resnet models on the CIFAR10 and ImageNet dataset, collecting output after activation RELU.

5 Discussion

The training profile(including accuracy, loss etc.) for identical networks with different initialization are the similar and it is possible that the same features will have different orientations in the network. Else, how is it possible to have high similarity in the CNN blocks at the end?