Understanding controlling parameters and reasons for generalization performance of neural networks is crucial to making neural networks more interpretable and developing reliable architectures. The paper challenges traditional beliefs involving role of parameters and complexity measures- VC dimension, Rademacher complexity, uniform stability in controlling generalization error.

Experiments by the authors show that state of the art CNNs fit randomly labelled data with 0 training error. However, the test error was very high. It is observed that with no changes in the model and by simply randomizing the labels the generalization error can increase by large extent.
Similarly, data with noisy images obtained by replacing true images by random pixels was found to fit with 0 training error.

Training loss for data with true labels, random labels, shuffled pixels, random pixels and Gaussian distribution (generate random pixels for each image) were studied using CIFAR10 dataset. For all cases, once fitting starts, it converges quickly. Random pixels and Gaussian converge faster than random labels. Random pixels are more distinct from other images in same category making it easier to train for arbitrary labels.

Data was trained on Inception v3, AlexNet, MLP and the convergence time was observed to slow down on increasing label noise for each of these architectures. Generalization errors on varying the label noise was also studied for the 3 architectures. Significant difference in test error was observed between Inception and MLP for lower label noise level. However, generalization error converges to 0.9 for all 3 architectures for noise level 1.

Role of regularization in generalization performance was studied using Inception v3 with variation of using explicit regularizers like data augmentation, weight decay and dropout to train on ImageNet dataset. For random labels, top1 accuracy 95.2% was obtained without explicit regularization and 91% accuracy was reached with regularization.

Augmenting data is observed to improve test accuracy better than weight decay and dropout. For true labels, top5 test accuracy of AlexNet was much better than Inception v3 suggesting that using different architecture could be better choice for improving generalization than use of regularizers.

Role of batch normalization as implicit regularization measure was studied. It improved generalization performance by 3-4%. It was seen that networks perform well even without regularization. Thus, regularization is a tuning parameter and does not govern generalization performance.

The authors emphasize that expressivity of neural networks of finite sample is more relevant than expressivity over entire domain by suggesting that 2 layer network with ReLu activation can represent any function of input sample.

Generalization of linear models was studied to help study parallel insights for better understanding of neural networks. The solution to which Stochastic Gradient Descent (SGD) converges was studied since the choice of model is an output of running SGD thus acts as implicit regularizer. It was found that we can fit any set of labels by forming kernel matrix and solving linear system K(alpha)=y. SGD often converges to minimum norm solution but minimum norm does not always give good generalization performance.