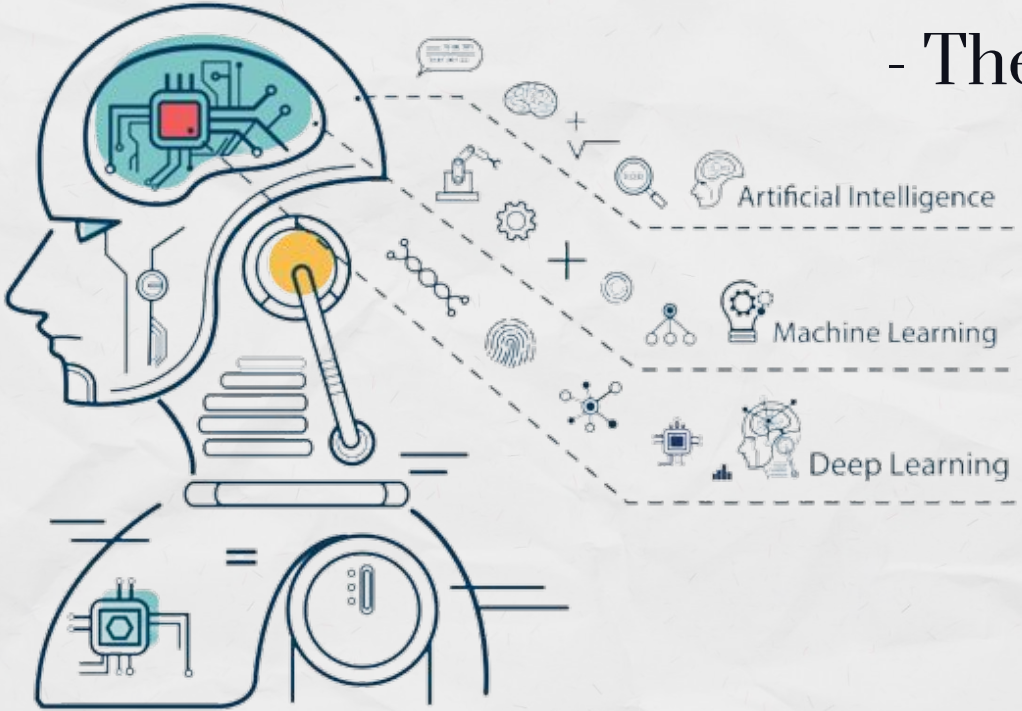
An Experimental Study of What Makes Deep Neural Networks Generalize Well

- The Role of Regularization



Using Fashion MNIST

Neha Thonta (nt446)
Ganesh Raj Kyatham (grk62)
Manisha Gayatri Damera (md1723)

Introduction

- Key feature of deep neural networks is their capacity for memorizing large amounts of data. This capacity for memorization can also lead to overfitting.
- Generalization Error: It is a measure of how well a machine learning model performs on new, unseen data.
- Deep neural networks have been observed to achieve 0 training error even when trained on completely random labels, highlights the importance of controlling generalization error.
- Controlling generalization error can be challenging due to their large number of parameters that can be tuned to fit the training data very closely.

Motivation

What is it that distinguishes neural networks that generalize well from those that don't?

Paper summary

- ➔ Despite their massive size, successful deep neural networks can exhibit a remarkably small difference between training and test performance.
- ➔ Importance of controlling generalization error in deep neural networks, but notes that explicit forms of regularization are not always necessary for good performance.
- ➔ Performed image classification using different deep neural networks to analyze their generalization performance and gain a better understanding.

Dataset - Fashion MNIST

The Fashion MNIST dataset is a collection of 70,000 grayscale images of fashion items, including 10 different categories such as t-shirts, dresses, sneakers, and sandals.

- 0 - T-shirt/top
- 1 - Trouser
- 2 - Pullover
- 3 - Dress
- 4 - Coat
- 5 - Sandal
- 6 - Shirt
- 7 - Sneaker
- 8 - Bag
- 9 - Ankle boot

Deep Neural Network Models

MODELS IMPLEMENTED

01

MLP

Feedforward artificial neural network, used for Regression and Classification.

02

CNN

Deep learning architecture for Image Analysis.

03

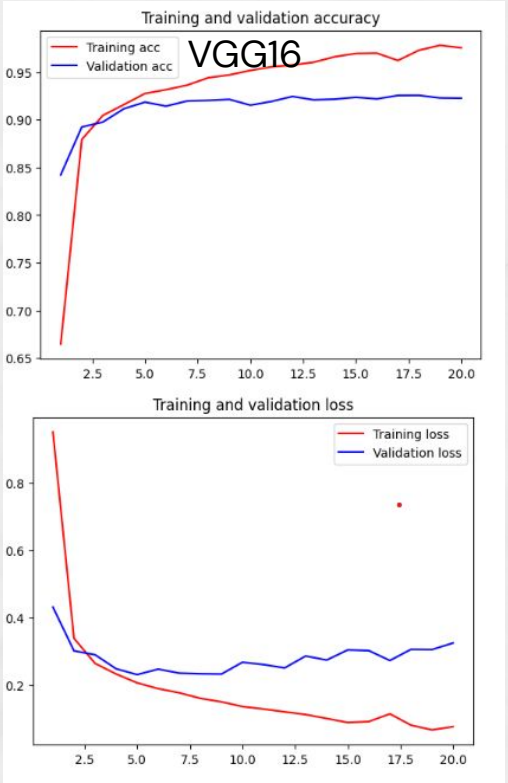
VGG16

Deep convolutional neural network architecture for Image Recognition.

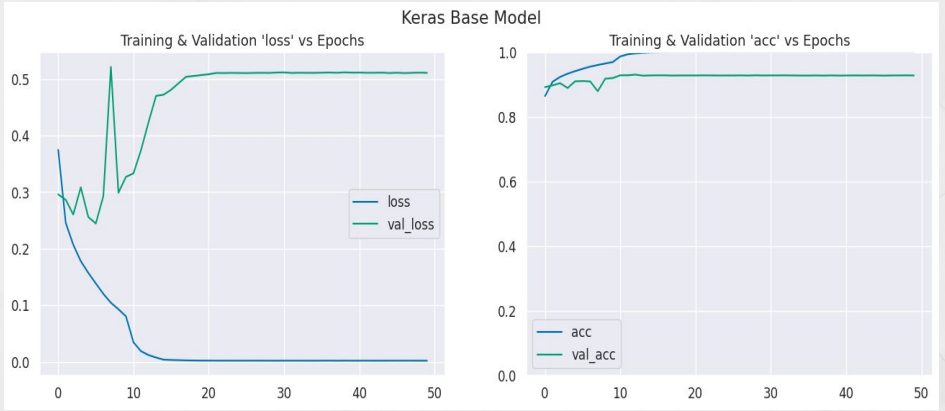
Vapnik-Chervonenkis theory, VC Dimension

- The VC dimension of a model is the maximum number of datapoints that can be arranged so that f shatters them
- $$\mathrm{P}\left(error_{test} \leq error_{training} + \sqrt{\frac{1}{N}\left[D\left(log\left(\frac{2N}{D}\right)+1\right) - log\left(\frac{\eta}{4}\right)\right]}\right) = 1-\eta$$
- This upper bound is valid when D<<N. Fails for neural networks where D >> N

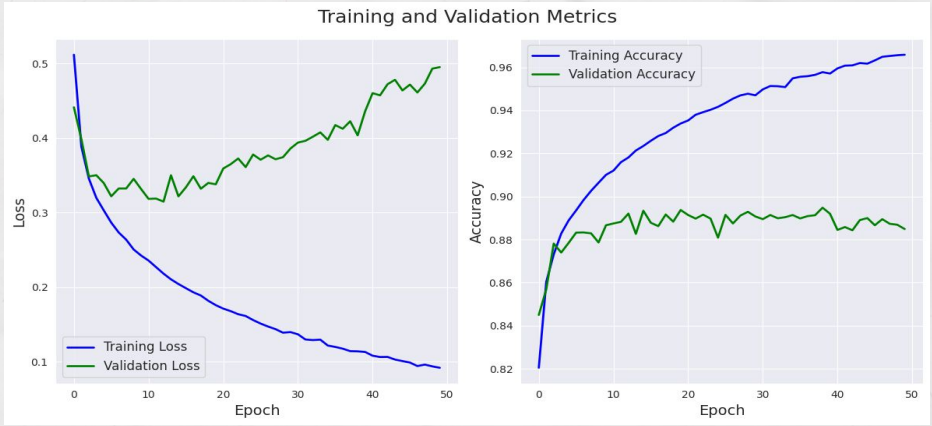
Without Regularization



CNN

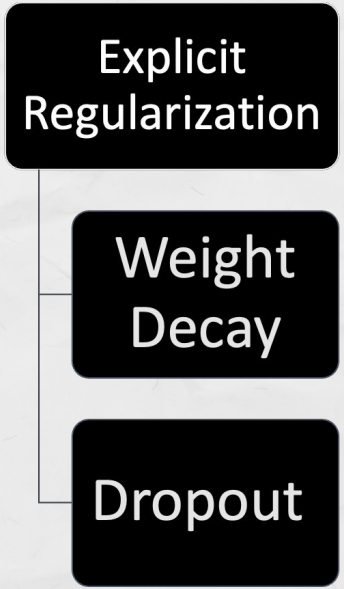


MLP



Regularization

- Regularization refers to a set of different techniques that lower the complexity of a neural network in order to prevent overfitting
- Neural networks easily fit most of the data given to them
- Most common explicit regularization techniques are dropout and weight decay.
- Reduce the effective capacity of the model and typically require the use of deeper and wider architectures to compensate for the reduced capacity



L2 Regularization (Weight Decay)

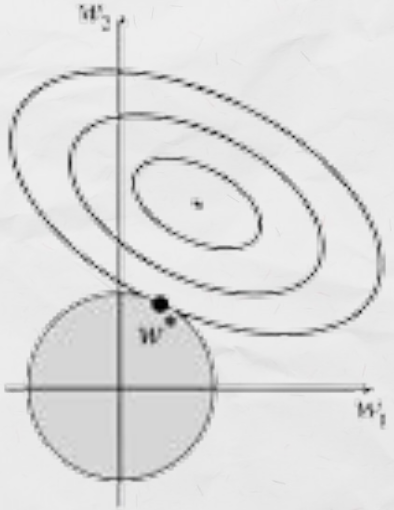
- Works at reducing overfitting
- Standard weight update:

$$W_{t+1} \leftarrow W_t - \eta_t \frac{\delta L(w)}{\delta W t}$$

- $L(W) = L_0(W) + \lambda/2\, w^2$ *
- New weight update:

$$W_{t+1} \leftarrow W_t - \eta_t \frac{\delta L0(w)}{\delta W t} - \eta_t \lambda W_t^*$$

- Forces the weights to become small “decay”

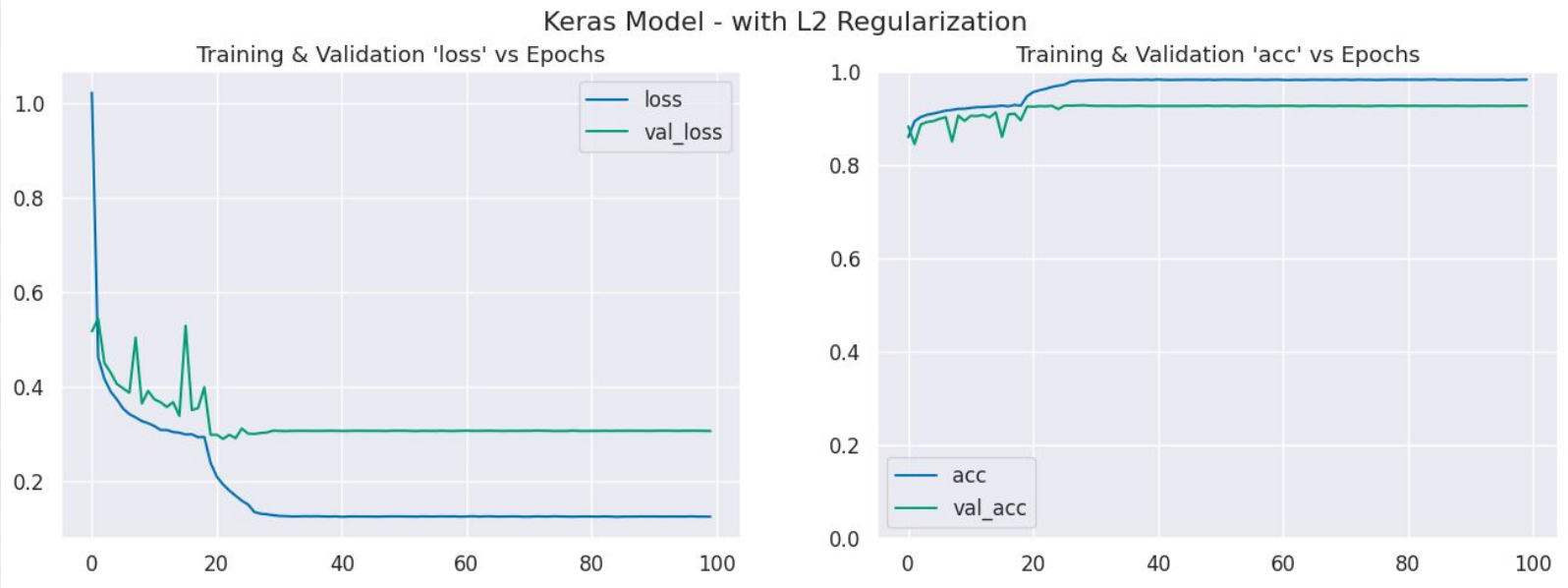


Weight Decay - Results

MLP

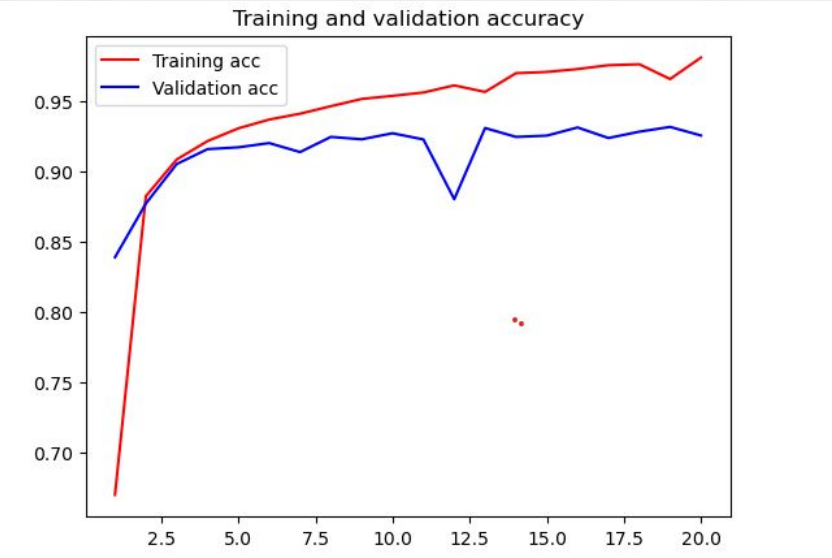
Weight Decay - Results

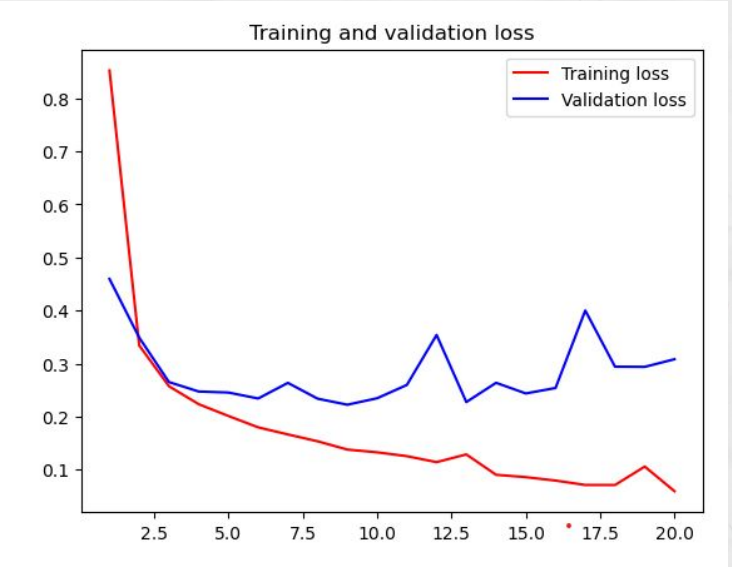
CNN



Weight decay - Results

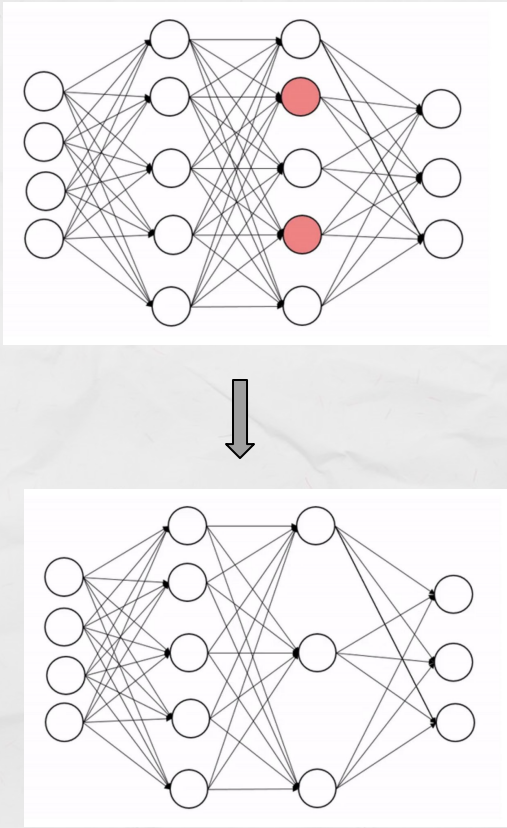
VGG16 - ImageNet





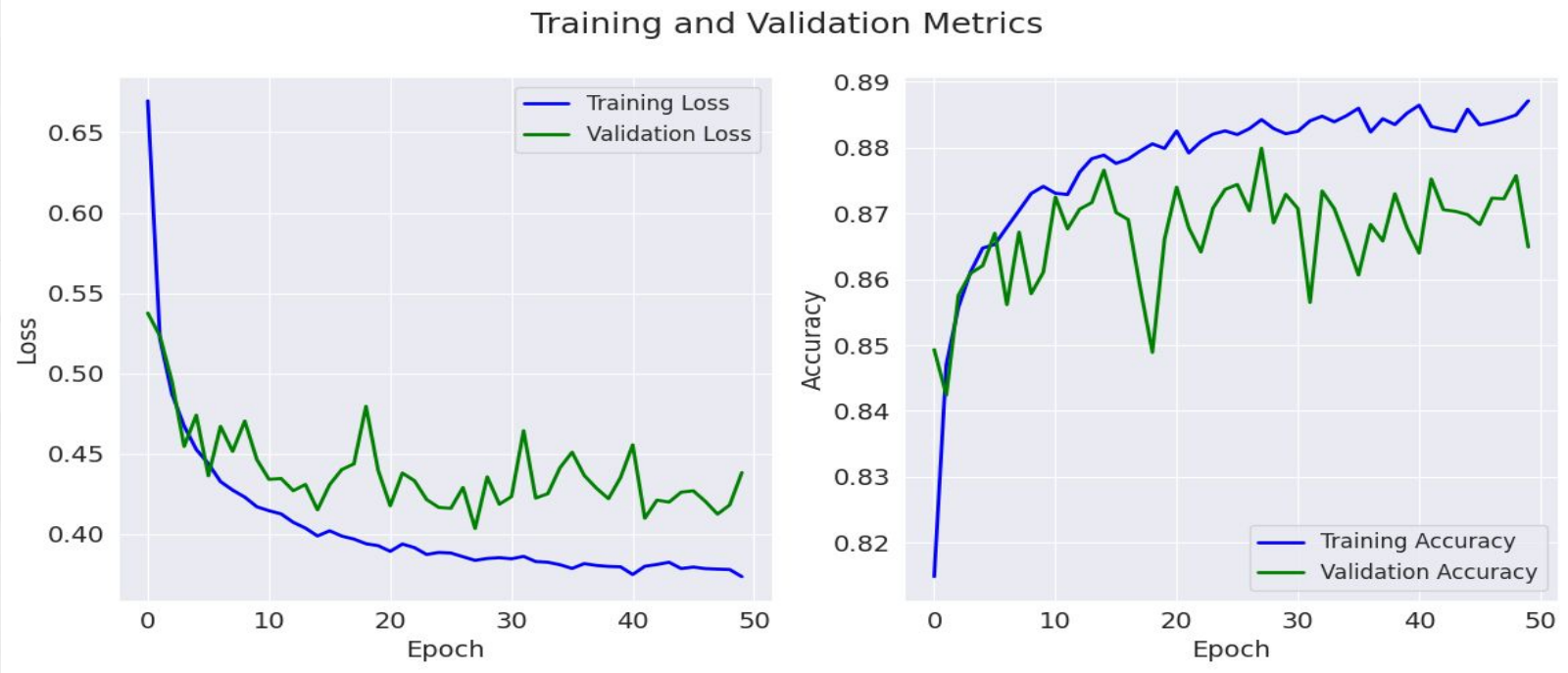
Dropout

- Randomly drop neurons from layers in the network
- Removes reliance on individual neurons
- Learn Redundancies
- Learn more nuanced set of feature detectors
- Mask out each element of a layer output randomly with a given dropout probability.



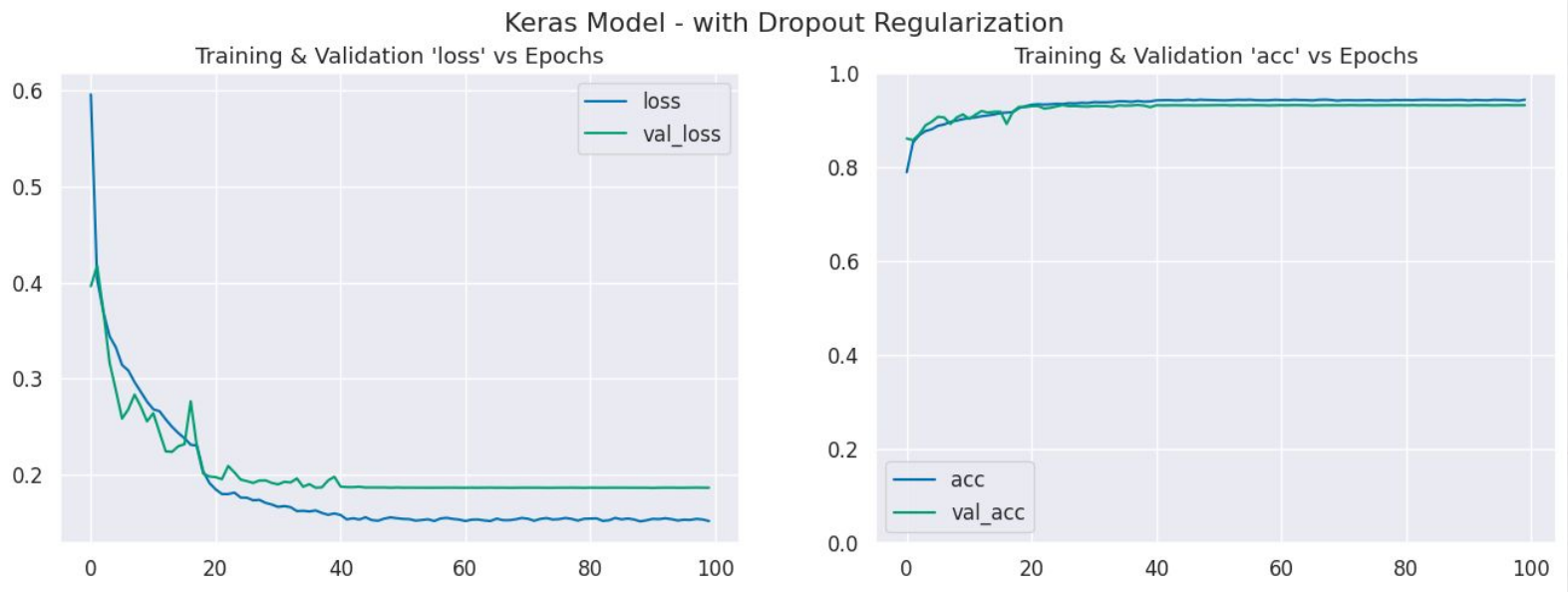
Dropout Regularization - Results

MLP



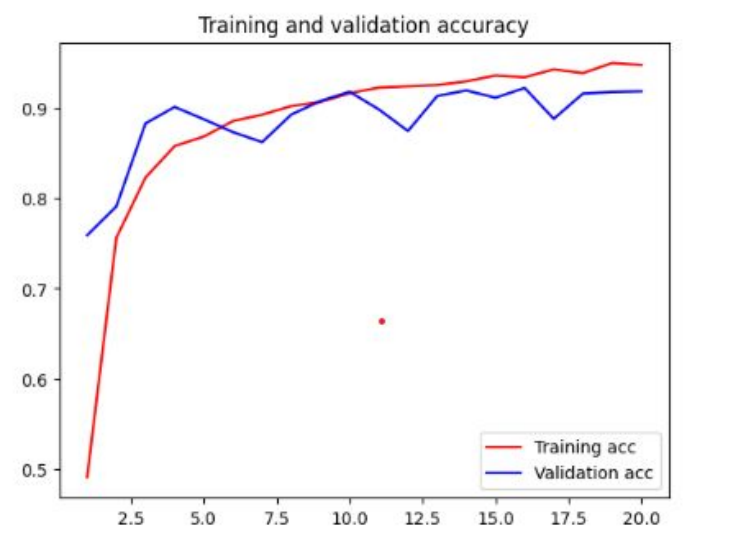
Dropout Regularization - Results

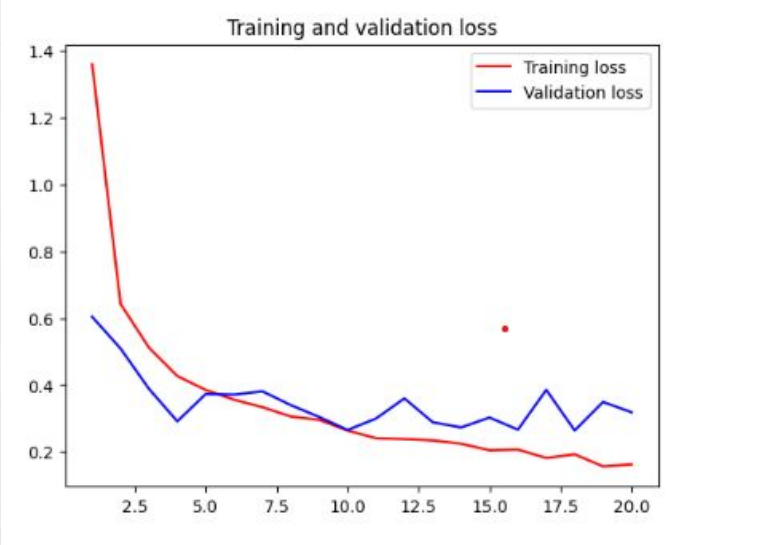
CNN



Dropout Regularization - Results

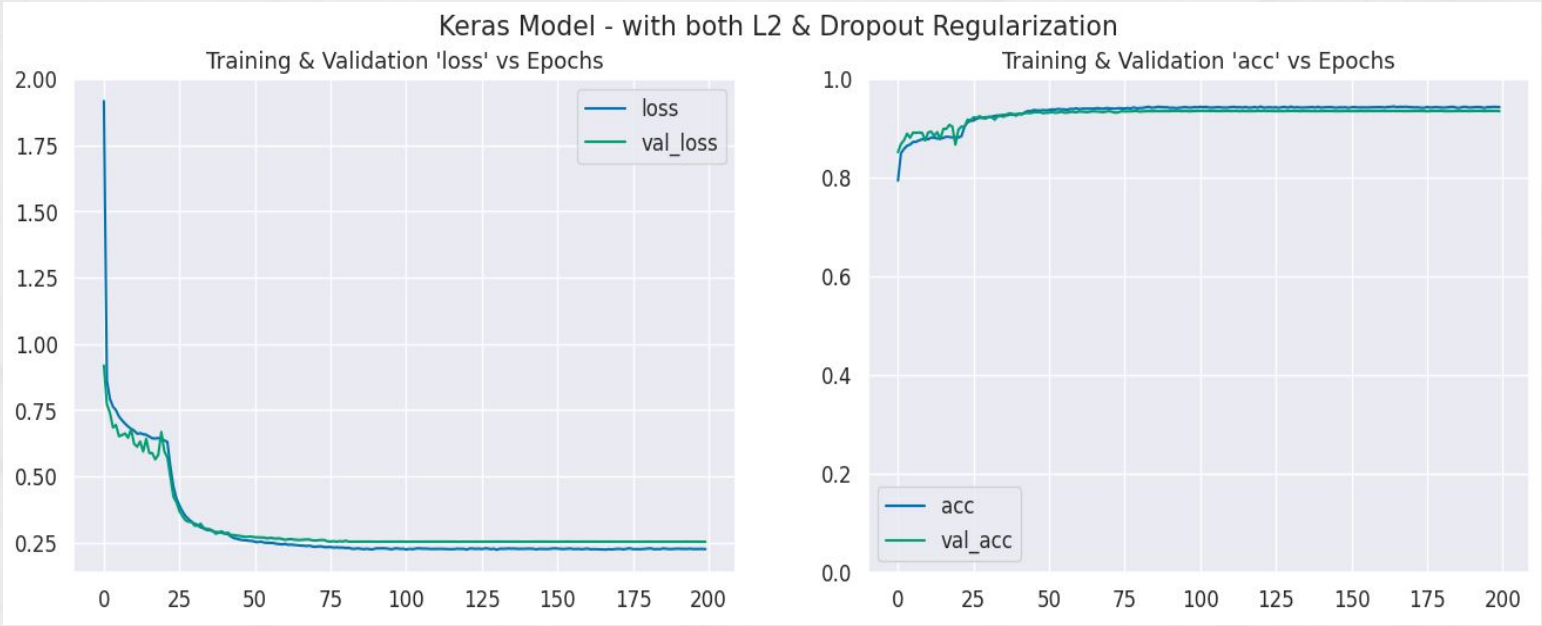
VGG16 - ImageNet





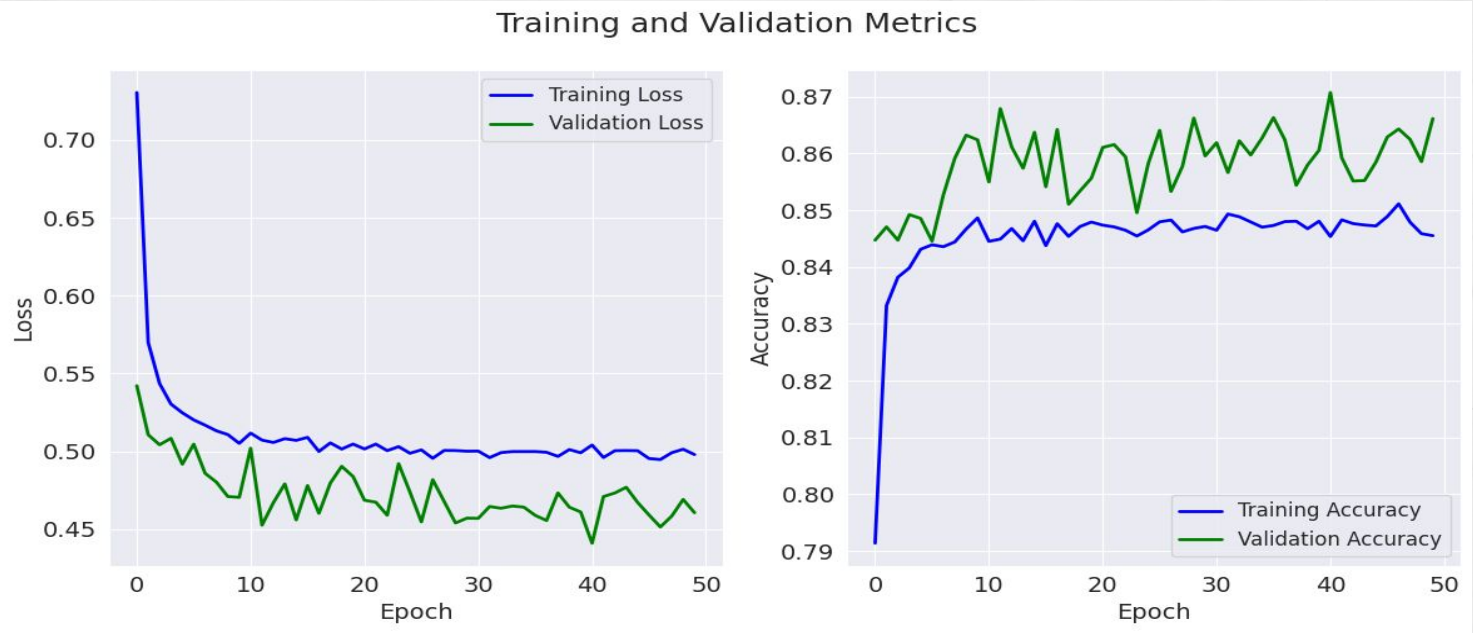
Weight Decay & Dropout Regularization - Results

CNN



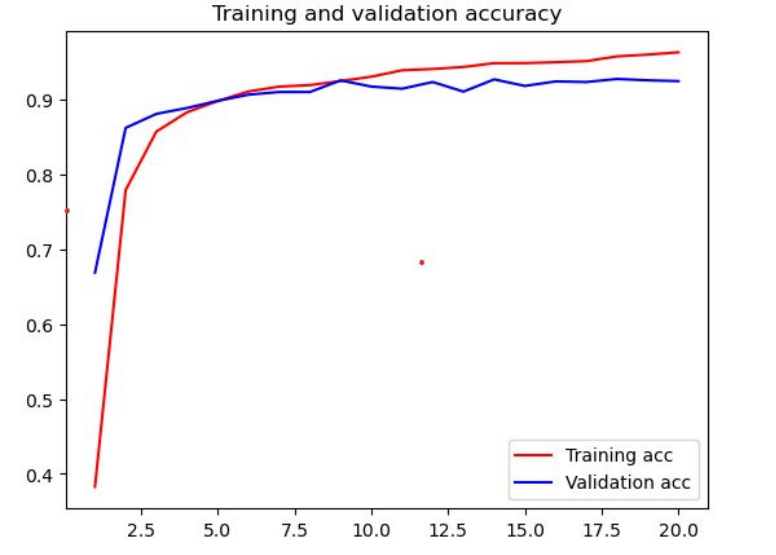
Weight Decay & Dropout Regularization - Results

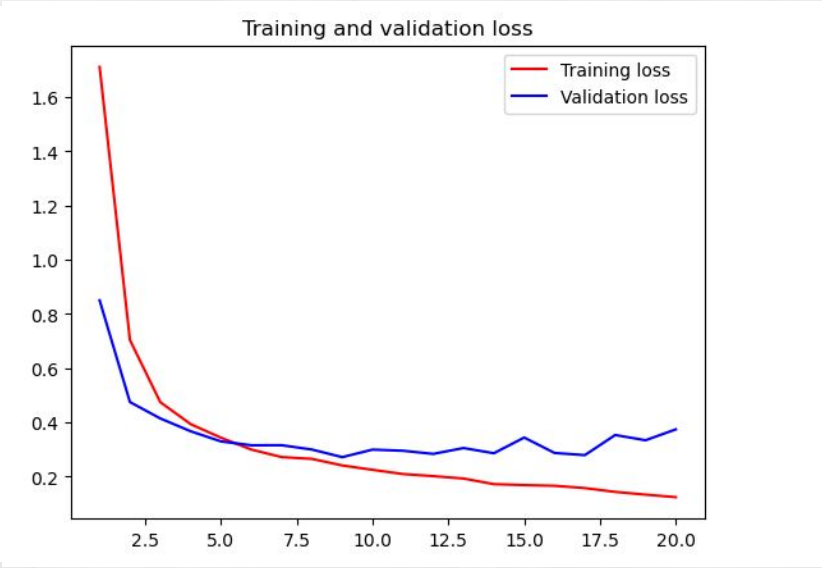
MLP



Weight Decay & Dropout Regularization - Results

VGG16 – ImageNet





Model	L2 Regularization	Dro-pout	Train Accuracy	Test Accuracy
MLP	No	No	96.6	88.5
	Yes	No	88.7	85.8
	No	Yes	84.5	85.3
	yes	Yes	82.4	84.5
CNN	No	No	100	92.8
	Yes	No	98.4	92.6
	No	Yes	96.2	93.2
	Yes	Yes	96.3	93.4
VGG16	No	No	97.5	92.2
	Yes	No	98.1	92.5
	No	Yes	94.7	91.8
	Yes	Yes	96.3	92.5

Conclusion & Implications

- The study demonstrates the effect of explicit regularization methods like weight decay, and dropout performed on the MNIST datasets.
- Bigger gains by changing the model architecture.
- The effective capacity of neural networks is sufficient for memorizing the entire dataset
- Explicit forms of regularization, such as weight decay and dropout, may improve generalization performance, they are not always necessary for good performance.

Implicit Regularization - SGD

Implicit Regularization: When the algorithm itself is implicitly regularizing the solution to converge to certain minima while avoiding others.

Stochastic Gradient Descent

- ➔ Given SGD : $W_{t+1} = W_t - \eta_t e_t x_i$ and w_0 , then $w = \sum_1^n \alpha i x i$ for some coefficients α.
- ➔ $w = X^T \alpha$ Lies in the span of the data points.
- ➔ To get a perfect fit on labels, using kernel matrix $K = XX^T$ and solving to get the minimum L2 norm.
- ➔ Hence SGD converges to the solution with minimum norm

Final Conclusion

- Neural networks have effective capacity and are large enough to shatter training data.
- Traditional measures of model complexity struggle to explain the generalization of large neural network.
- Optimization continues to be easy even when generalization is poor.
- Both explicit and implicit regularizers could help to improve the generalization performance.
- However, it is unlikely that the regularizers are the fundamental reason for generalization.

Thank You