# SMAI Project Theory Report

**Team - SaMURai**

**Sarthak Bansal - 2021101134**

**Krishna Singh - 2021112005**

## Motivation

What is it that distinguishes neural networks that generalize well from those that don't? A satisfying answer to this question would not only help to make neural networks more interpretable, but it might also lead to more principled and reliable model architecture design

Generalization capabilities of massive size successful deep artificial neural networks has been a topic of study for some time. Traditional wisdom leads us to believe that over parameterization causes over-fitting of models while these large parameterized neural networks show marginal difference in train and test performance.

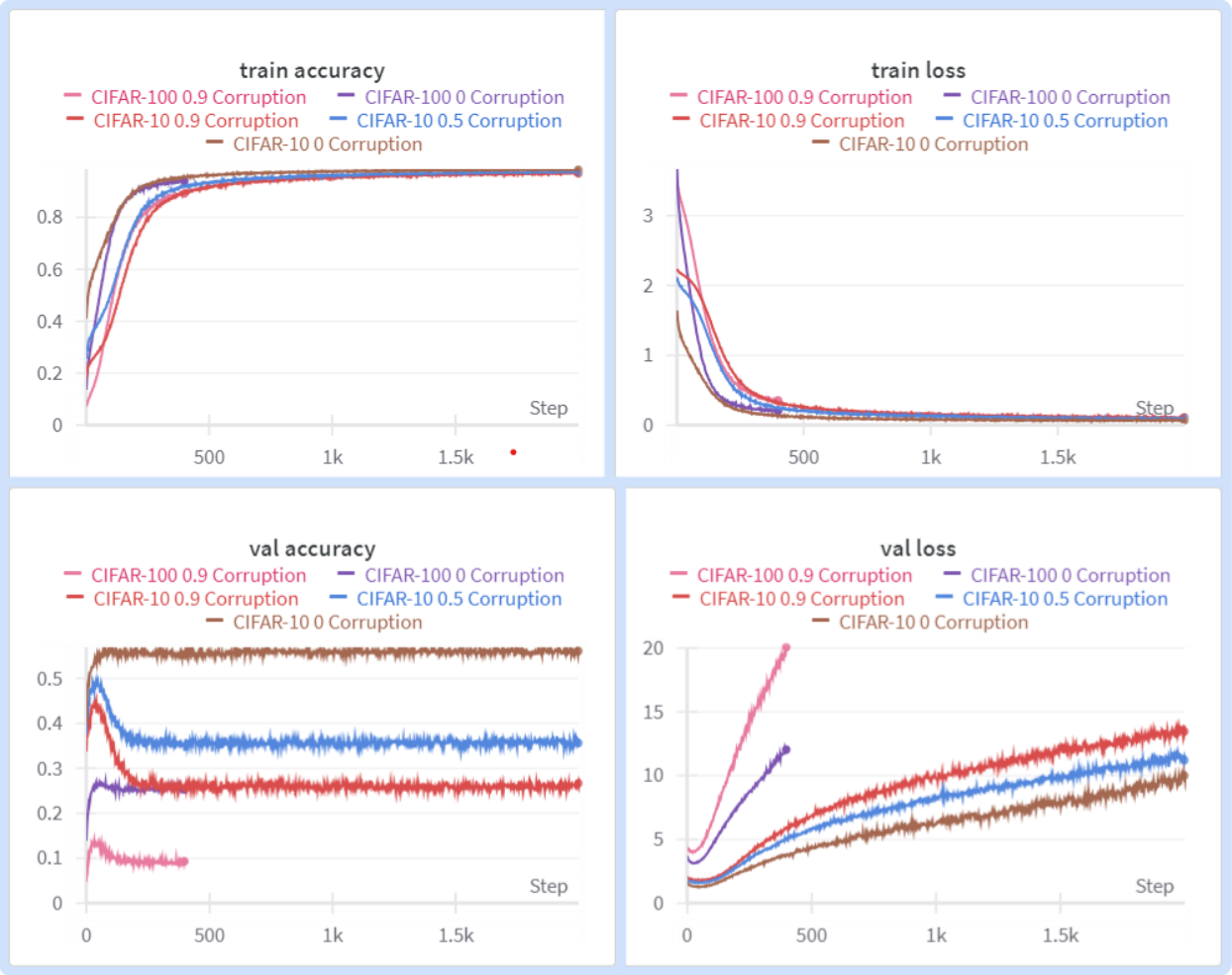
Regularization techniques used during training are largely attributed to this performance. However, through systematic and extensive experimentation we show that these models fail to perform in scenarios where there is noise in the data in the form of random labelling or even noise in pixels themselves. Despite having used traditional approaches to tackle generalization issues, these models fit such noisy data to the extent of 100% train accuracy.

## Expressivity of neural networks

> 💡 A very simple two layer neural network can express any labelling of the training data.

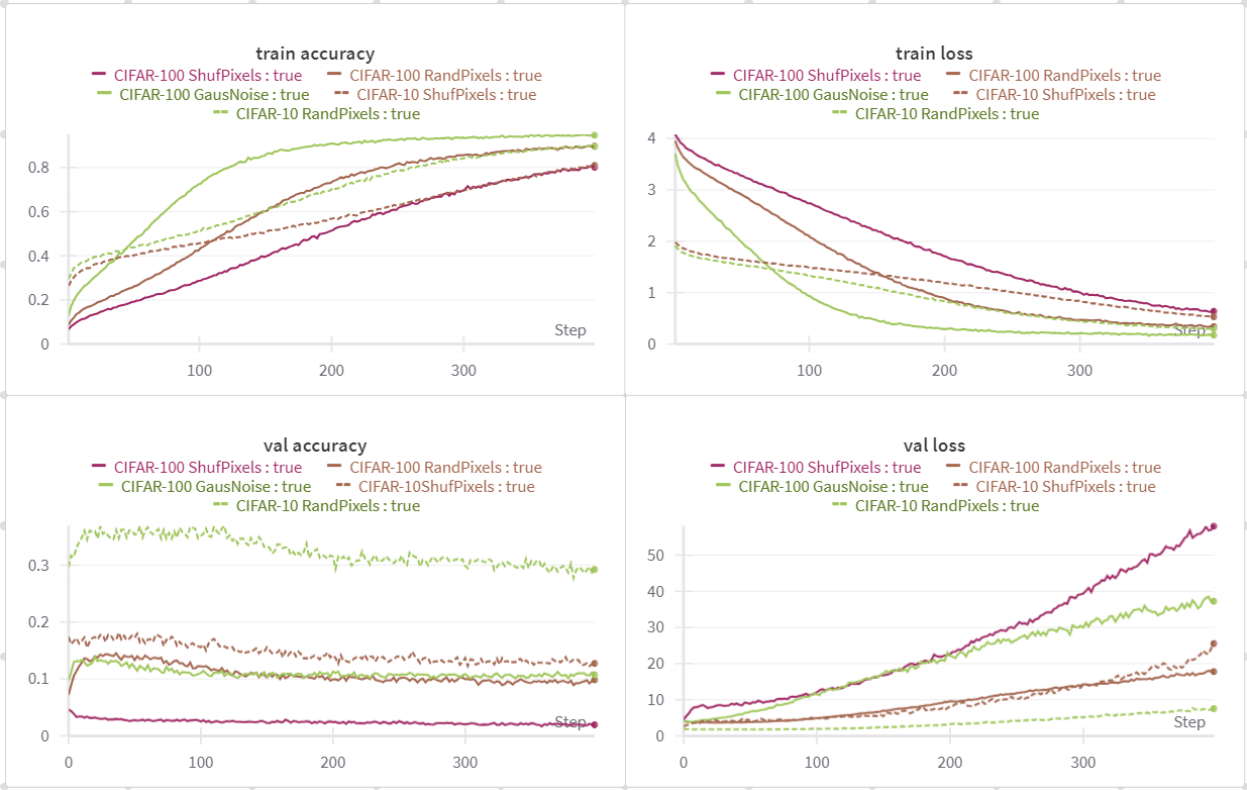
The methodology is a variant of  the well-known randomization test from non-parametric statistics (Edgington & Onghena, 2007). In the first set of experiments , we train a 2-layer simple ReLU MLP with a copy of the CIFAR-10 and CIFAR-100 datasets where the true labels were replaced by random labels . We vary the corruption level between 90% (close to truly random) and 50%. Further , we replace true images by random and shuffled pixels . We also experiment by adding gaussian noise to images.

### Random Labelling




As one can see , a simple MLP can express any labelling of data as the train loss approaches 0 for all the cases whether noisy or not with training accuracy approaching 100%

## Noisy Images




As one can see , even after replacing true images by Random Pixels , Shuffled Pixels (Image coherent noise) and Gaussian Noise (Image independent noise) , a 2-layer MLP fits the data and approaches 0 training loss or 100% training accuracy

Randomizing labels and Degrading Images is solely a data transformation, leaving all other properties of the learning problem unchanged. However , the above experiments show that by doing such augmentations alone we can force the generalization error of a model to jump up considerably without changing the model, its size, hyperparameters, or the optimizer.

**Universal approximation theorem** states that

$$For\ every\ \hat{f} \in C(K)\ with\ K \subset R^d\ and\ every\ \varepsilon > 0\ , there\ exists\ a\ neural\ network\ \hat{f}_{H,\varepsilon}\ such\ that\ |\hat{f}_H - \hat{f}|_\infty \leq \varepsilon$$

Simply , it states that even a simple neural network with depth-2 and a suitable activation function can approximate any modeling function. The above experiments verify the theorem. This also establishes the fact that effective capacity of neural networks is sufficient for **memorizing** the entire data set !!

# Role of Explicit regularization

We saw above that model architecture alone doesn't improve generalization performance on noisy datasets , then the question remains : **What does ? .**

Traditional approaches suggest using explicit regularization methods

> *Explicit regularization* techniques are those specifically and solely designed to constrain the effective capacity of a given model in order to reduce overfitting. Furthermore, explicit regularizers are not a structural or essential part of the network architecture, the data or the learning algorithm and can typically be added or removed easily.

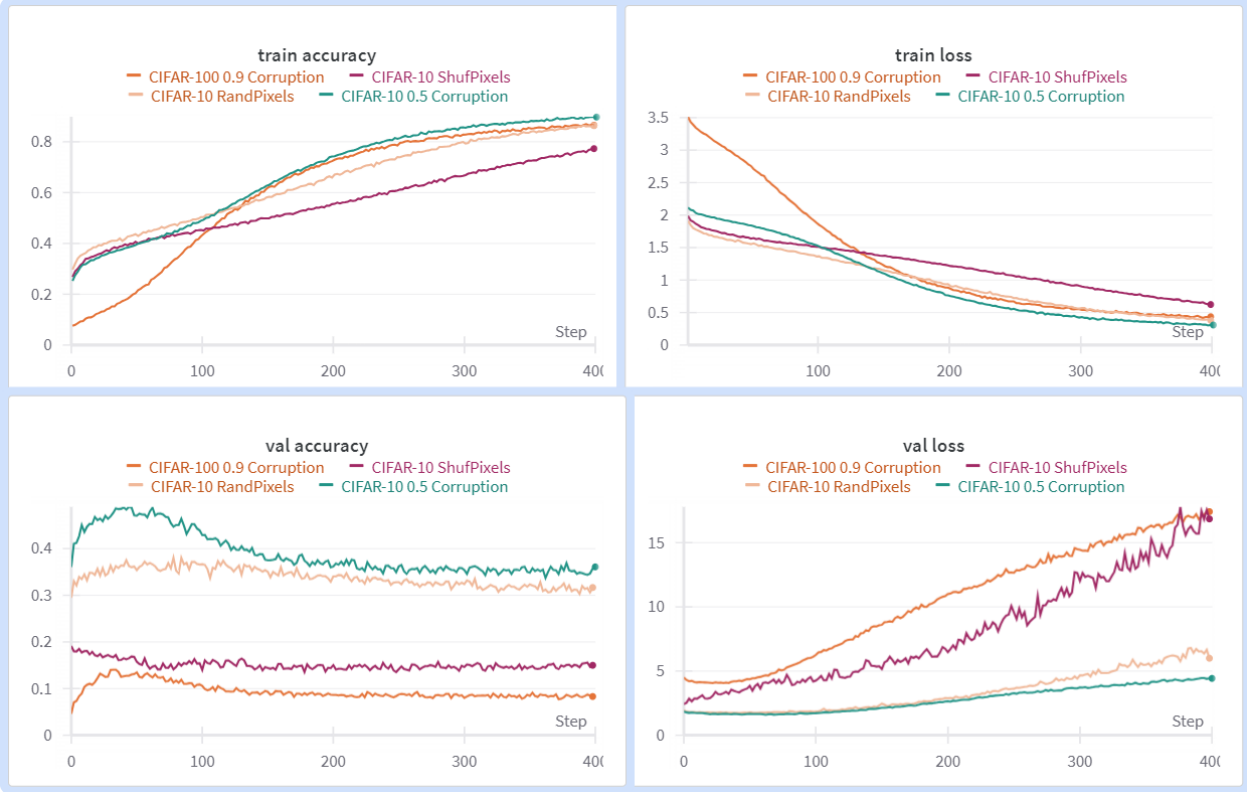
We experiment with different explicit regularization methods on noisy copies of CIFAR-10 and CIFAR-100 datasets. Noise is similar to the ones described in the previous section.

## Method - 1 : Dropout

### Motivation

Dropout was first described in “Dropout: A Simple Way to Prevent Neural Networks from Overfitting” with the key idea being to randomly “drop” units from the neural network during training. It hypothesized a way of approximately combining exponentially many different neural networks providing a smaller hypothesis class to search on. Dropout simulates a sparse activation from a given layer, which interestingly, in turn, encourages the network to actually learn a sparse representation as a side-effect.

### Experimentation




When experimented with different types of noise on CIFAR-100 and CIFAR-10 , we observe that the model still approaches 0 training loss / 100% training accuracy .

### Conclusion

Dropout may improve generalization error in some cases , but doesn’t stop the model to memorize the data along with its noise. Its absence doesn’t necessarily mean a poor generalization performance.

> 💡 Dropout’s explicit regularization effect doesn’t improve the generalization error in noisy scenarios. Its working in some cases is due to the dropout noise it injects , similar to the mini-batch noise, which works as an implicit regularize and helps it generalize.

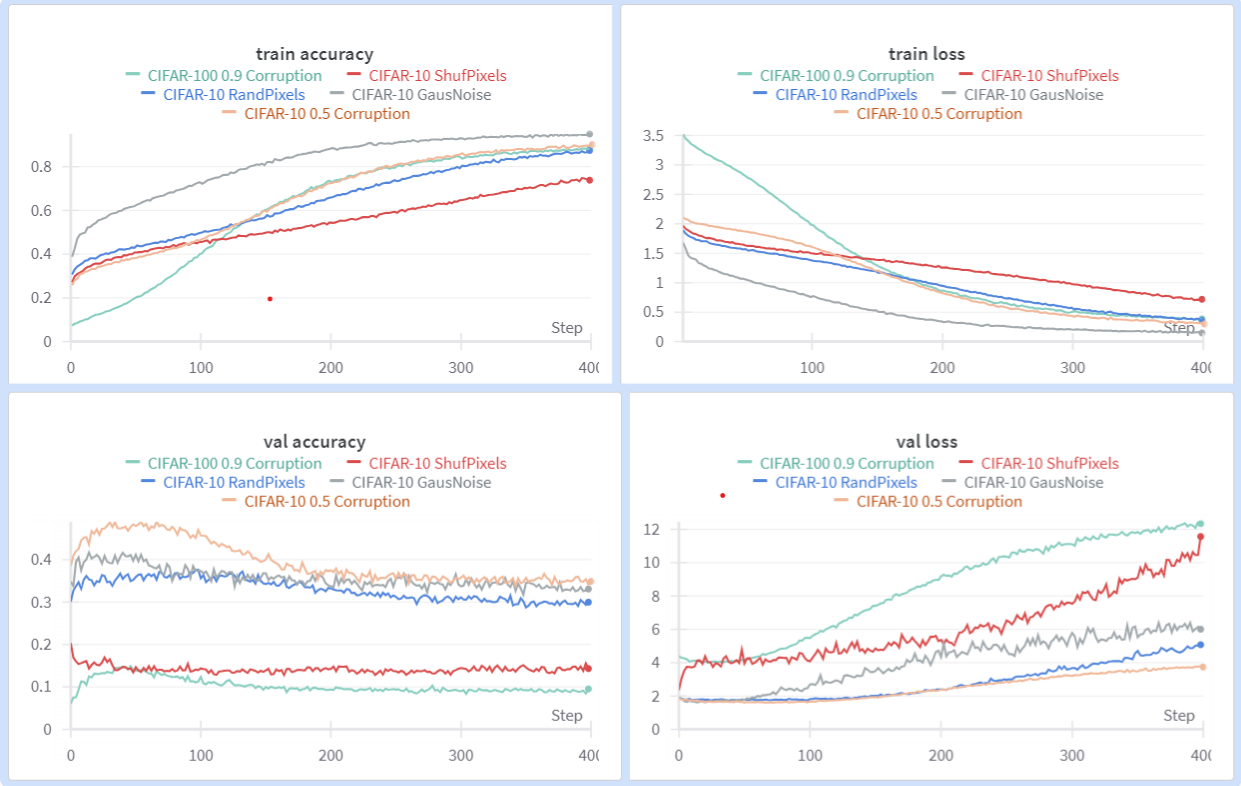
Mini-batch noise in SGD is due to the randomness it introduces during the training process which allows it escape the local minima and explore other parts of the loss landscape . Dropout, on a higher level, also “randomly” drops out some nodes which can be seen as a form of noise.

## Method - 2 : Weight Decay

### Motivation

Weight decay (also known as *l2-regularisation*) is a technique applied to the weights of a neural network . We minimize a loss function compromising both the primary loss function and a penalty on the L2 Norm of the weights. Its usage comes from the traditional knowledge that we want to bias our model towards the smoother , simpler ones such that it is able to generalize.

### Experimentation




When experimented with different types of noise on CIFAR-100 and CIFAR-10 , we observe that the model still approaches 0 training loss / 100% training accuracy

### Conclusion

> 💡 Weight decay is also rarely useful as an explicit regularizer but modifies the optimization dynamics enhancing the ever-present implicit regularization of SGD via the loss stabilization mechanism.

The role of weight decay is largely misunderstood in classical deep learning where it is seen as an explicit regularizer. However , its regularization effect is due to the following reason :

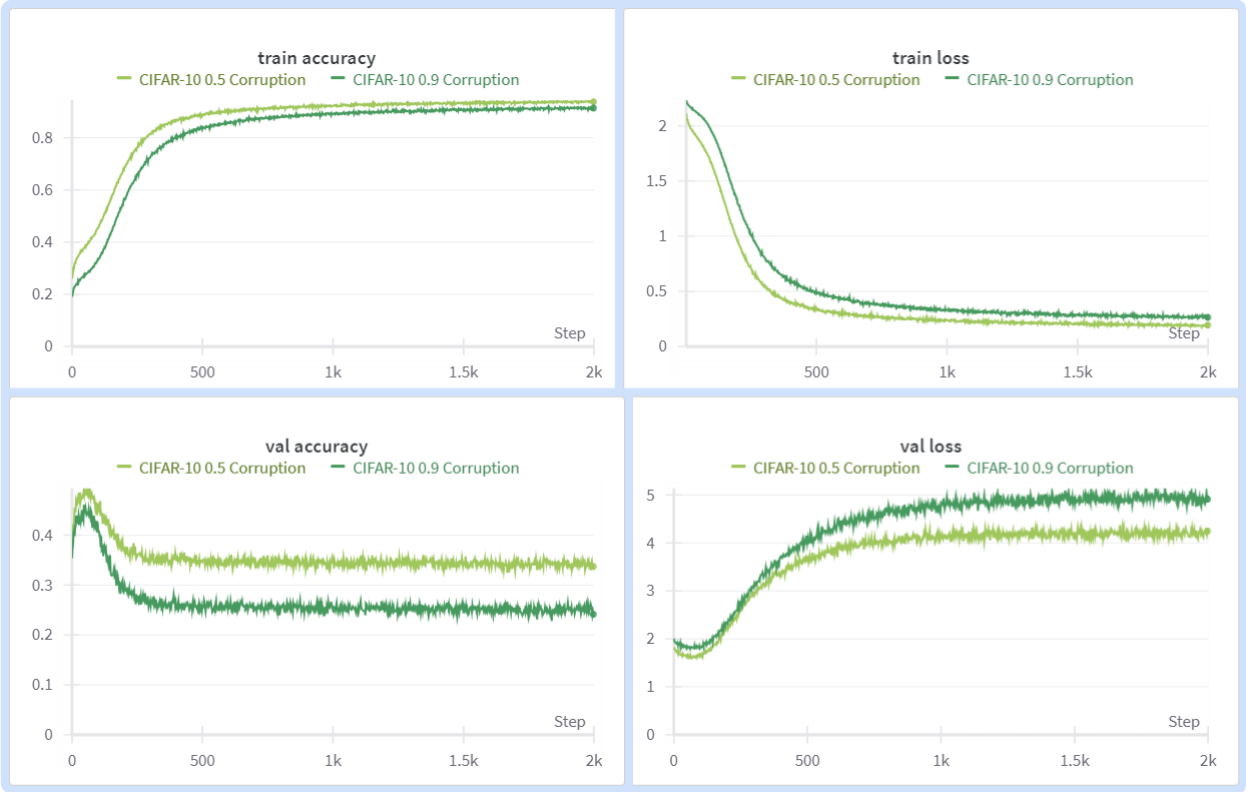
*Weight decay maintains parameters norm in a small bounded interval. The resulting projected noise-driven process induces an implicit regularization effect.*

However , there is enough evidence that SGD already converges to a solution with minimum norm resulting in weight decay not able to stop training of noisy datasets

## Method - 3 : Weight decay + Dropout

In practice , deep neural network models are trained by employing both weight decay and dropout. Here are some experiments with the same.

### Experimentation




Even after employing both weight decay and dropout, MLPs are still able to fit random labels.

# Role of Implicit regularization

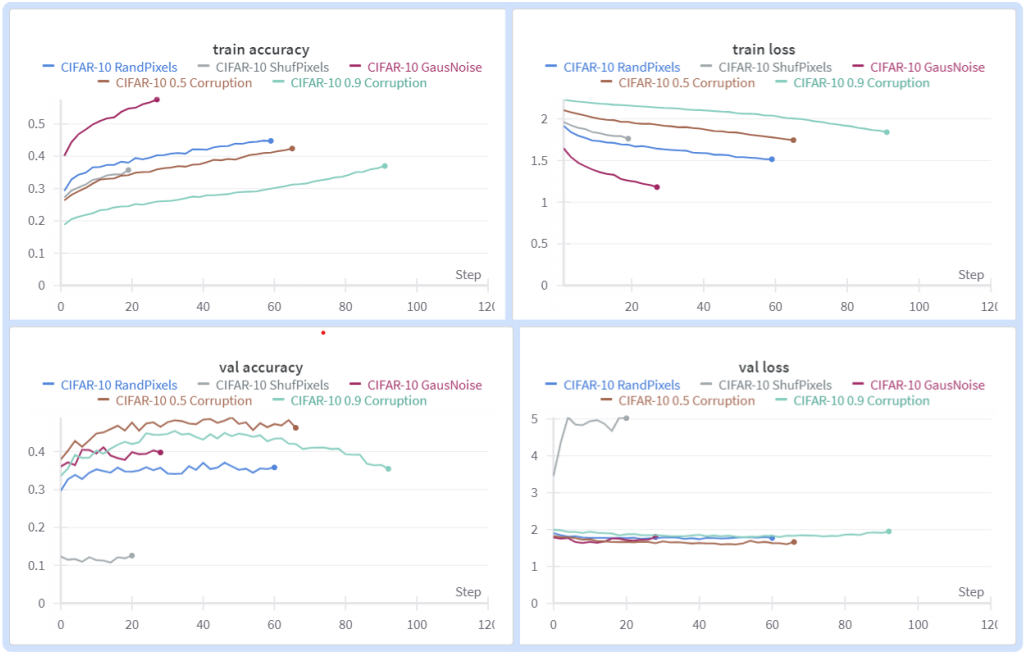
Implicit regularization is when machine learning algorithms, without explicitly using regularization terms, naturally produce models with simpler structures, helping prevent overfitting and enhancing generalization.

## Early Stopping

### Motivation

Early Stopping is a traditional regularization strategy which essentially involves saving and adjusting the current optimal parameters throughout the training process. When updates to these parameters no longer lead to enhancements (after a specified number of iterations) in validation, training is halted, and the most recent optimal parameters are employed. This technique acts as a form of regularization by constraining the optimization process to a narrower range of parameter values.

### Experimentation




### Conclusion

The above results show that early stopping prevents training noisy labels and noisy images by stopping the optimization process when the validation loss stabilizes

> 💡 Early stopping is crucial for gradient descent to converge to a sparse model, a phenomenon that we call *implicit sparse regularization*.

Implicit sparse regularization refers to the idea that stopping the training early can lead to simpler models with fewer parameters, which can be beneficial for preventing overfitting and improving generalization. . It implicitly penalizes overly complex models and encourages simpler models, achieving a form of sparse regularization.

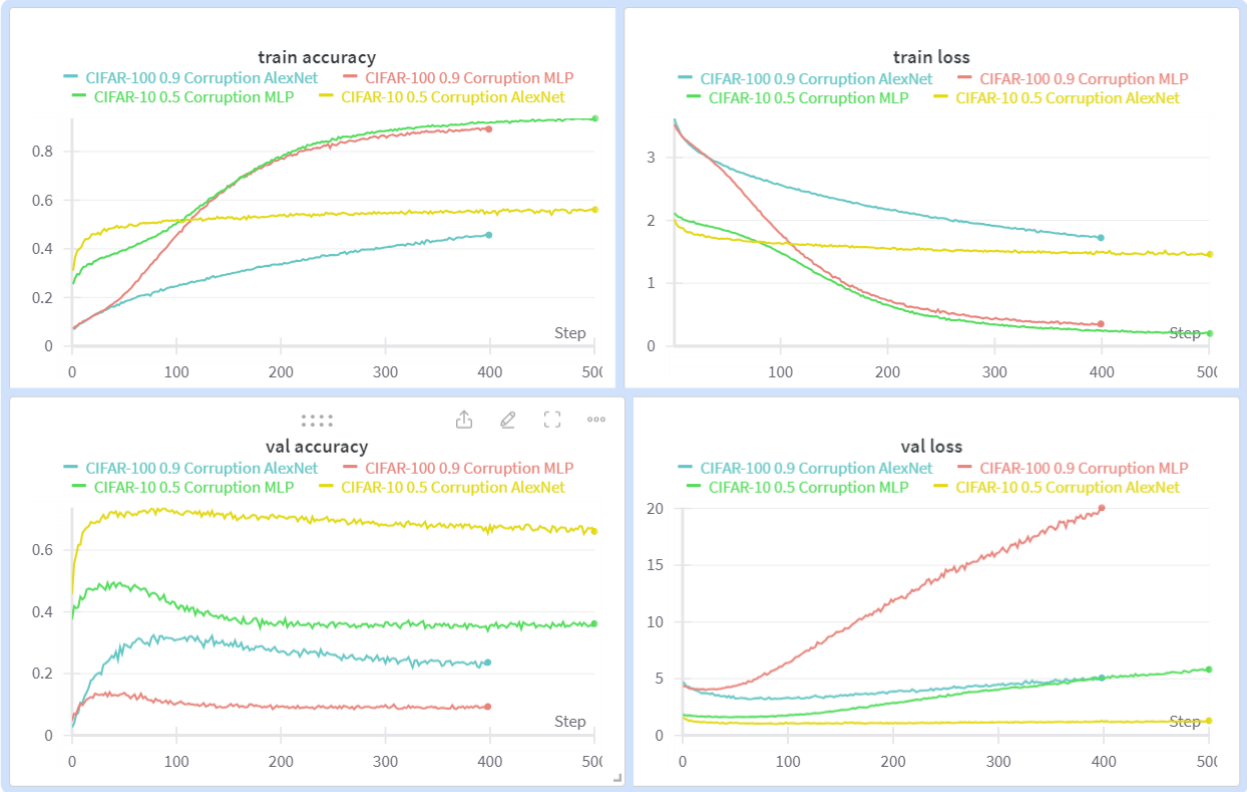
## CNN-Model

### Motivation

CNN(Convolutional neural networks) models are a class of models which try to preserve the spatial structure of the images intact and employ weight sharing similar to convolution of signals by moving the same kernel over the entire image. They are translational invariant also. Historically , CNN's have achieved much better than MLPs in image related tasks.

### Experimentation

### Random Labelling




CNN Model does a better job at avoiding fitting the random label noise in the data than MLP.

## Image Noise

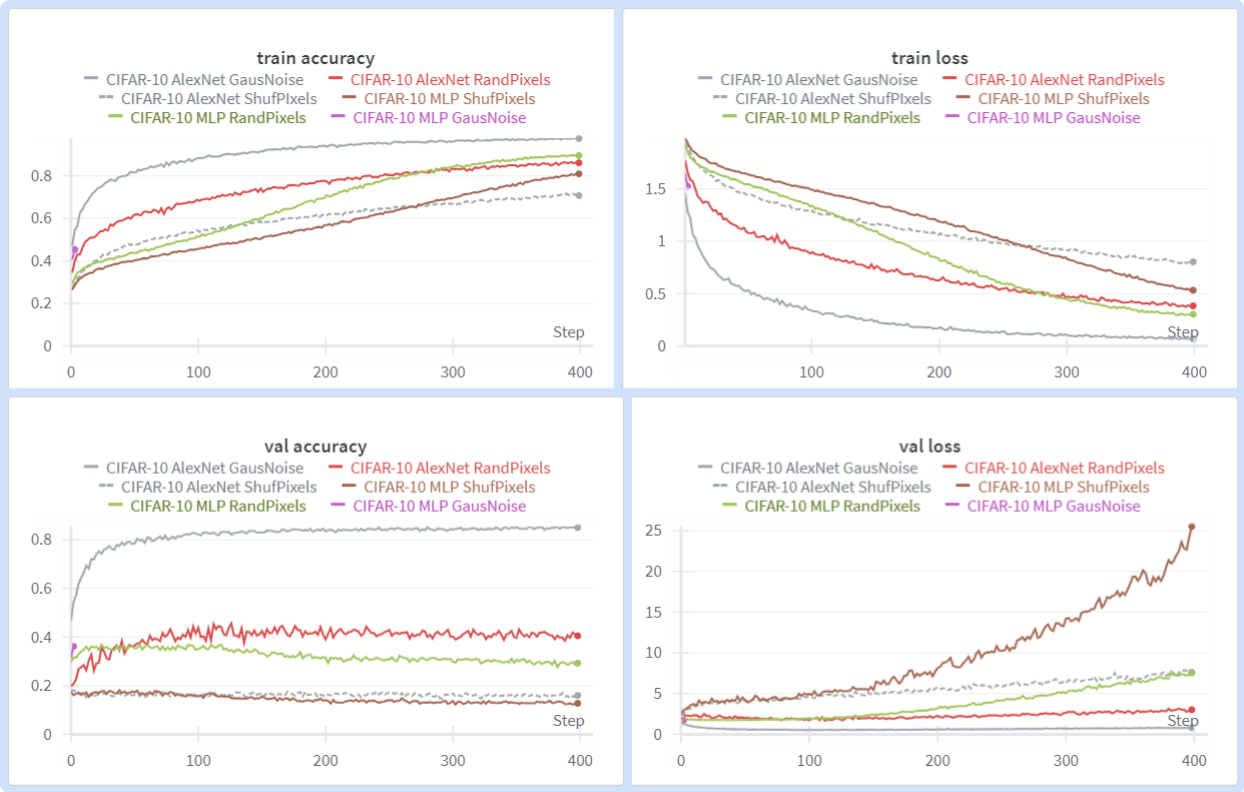



Image noise starts converging faster than random labels even for CNN model, thinning the difference between the generalization capability of MLP.

## Conclusion

The intuition behind using CNN was to preserve the spatial structure of the images and maintain a spatial hierarchy, however, their regularization properties are seldom appreciated.

> 💡 CNN's have this amazing intrinsic property of weight sharing which allows them to restrict the function hypotheses to a finite amount of functions compared to MLPs which assume that every neuron in 1 layer is related in some way to every other neuron of the next layer.

This minimization of the hypothesis class delays the model from learning “any” arbitrary labelling of the data model . However , when trained for a long enough time , CNNs can still converge to 100% accuracy.

> 💡 When we combine SGD with CNN models , there is proof that the architecture returns solutions that are biased to have sparsity in the frequency domain and this bias changes with the depth of the architecture. This is another inductive bias which may act as an implicit regularizer in the case of CNNs

CNNs are also sparsity inducing regularizers, which induce sparsity more aggressively as the depth increases. This randomness in sparsity helps them generalise.

> 💡 However, the results suggest that CNNs are able to fit random pixels , shuffled pixels and gaussian noise in the image almost similar to MLPs.

This might be because with such augmentations, the inputs are more separated from each other than natural images that originally belong to the same category, therefore, easier to build a network for arbitrary label assignments.

## SGD

> 💡 SGD acts as an implicit regularizer.

While training models with SGD , the models become bias towards solutions with small norm . This implicit bias acts as a regularizer by penalising complex models that are prone to overfitting.  The randomness generated when selecting a mini-batch in SGD induces some SGD some noise which is also considered an implicit regularizer.

However , there is empirical evidence that not all models with smaller norm generalize well ( An example is provided in the paper “UNDERSTANDING DEEP LEARNING REQUIRES RETHINKING GENERALIZATION”.

However , the experiments in these section which are trained on SGD suggest that even models trained using SGD can fit random labelling of data perfectly without any explicit regularization when trained for a long enough time. The above experiments are consistent with the observations that :

1. We do not need to change the learning rate schedule
2. Once the fitting starts, it converges quickly

# Towards a model which handles noise: IOC-NNS

## Motivation

IOC-NNs ( Input Output Convex Neural Networks) impose a convexity constraint on the standard architecture i.e modelling the problem such that the resultant function is a convex function of the input . On a higher level, this is expected to minimize the hypothesis class of the model which is considered to be the traditional approach for preventing overfitting.

The convexity constraints include restricting the weights (for all but the first layer) to be non-negative and using a non-decreasing convex activation function. The final output y is convex with respect to the input x by ensuring two conditions:

(a) $w_{ij}^{2k+1} \geq 0$ and

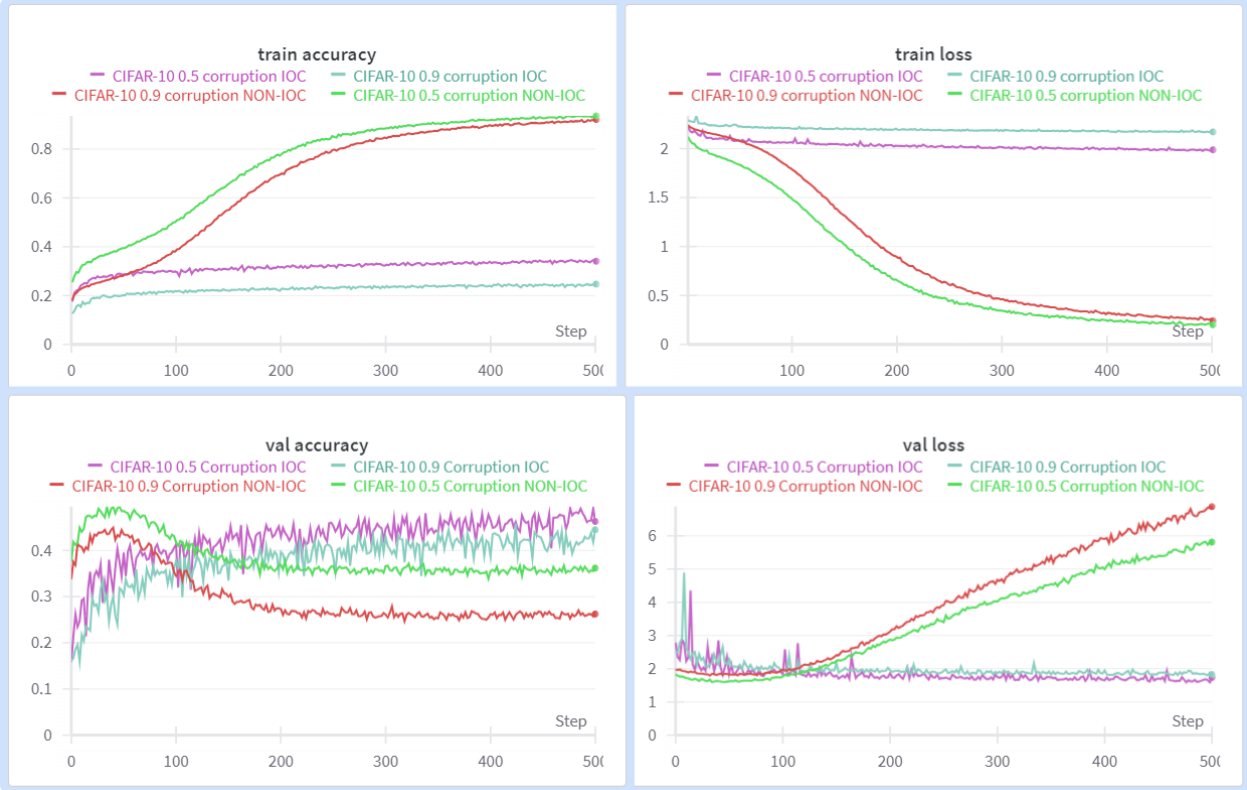
(b) φ is convex and a non-decreasing function , where φ represents activation function

In the below experiments , we make 2 changes to the standard training process

1. $w_{ij} = e^{w_{ij}-5}\ \forall w_{ij} < 0$ except for the first layer
2. Use ELU activation function instead of ReLU

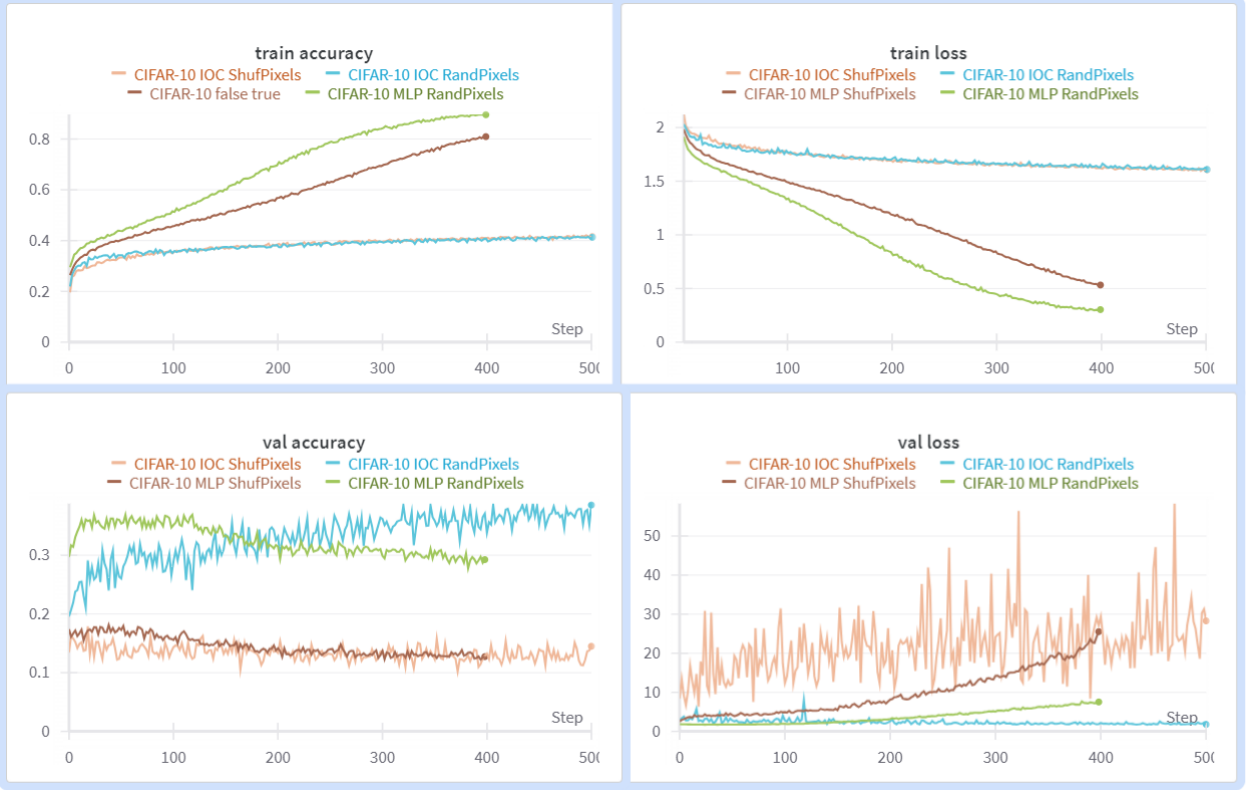
## Experimentation

### Random Labelling




IOC-NNs prevent the fitting of random labelling and the training curve flattens as compared to MLPs which approach 0 training loss

## Image Noise




IOC-NNs do not fit noisy images as compared to MLPs which approach 0 training loss

## Conclusion

> 💡 IOC-NNs act as self-regulariser and restrict the training process in presence of noise

We use an ELU activation to allow negative values; this is a minor but a key change from previous efforts that rely on ReLU activation.

The representation capacity of the resulting network reduces, and hence regularization comes in effect. This effectively helps in improving the generalization and controlling the over-fitting, as clearly observed in our empirical studies.

However , there is still little to no explanation why these networks actually work and what kind of regularisation they impose which makes them robust to data noise.

# Conclusion

- The paper aims at posing a very important question of formalising the methods which generalise large, over-parameterised neural network models.
- We first prove through empirical studies that these networks in effect are rich enough to memorise the entire training set, i.e give any arbitrary labelling of training data
- We then explore the regularisation techniques which are suggestive of making the model robust to data noise and observe that commonly acknowledged methods fail to generalise noisy data.
- We then describe a model (IOC-NN) and empirically show that it is able to make the model robust to training noise.
- However, all these experiments are still empirical and we have yet to discover a precise formal measure under which these enormous models are simple.

# References


- https://arxiv.org/pdf/2007.04759.pdf
- https://openreview.net/pdf?id=ByJWeR1AW
- https://www.cs.toronto.edu/~rsalakhu/papers/srivastava14a.pdf
- https://proceedings.mlr.press/v119/wei20d/wei20d.pdf
- https://arxiv.org/pdf/2310.04415.pdf#:~:text=We conclude that weight decay,dynamics in a desirable way.&text=We discuss the most related,comparisons later in the paper.
- https://proceedings.neurips.cc/paper/2020/file/f21e255f89e0f258accbe4e984eef486-Paper.pdf
- https://arxiv.org/abs/2006.05103