COMP 6961 Graduate Seminar Report:

Using Intrinsic Dimensionality to improve Dropout Regularization

Rutwikkumar Sunilkumar Patel (40160646)

August 05, 2021

Abstract

The speaker of the seminar, using Intrinsic dimensionality to improve dropout regularization, is Javier Fernandez Cruz who did research under the supervision of Professor Thomas Fevens. He starts by explaining the outline of his presentation, where the presentation is divided into four main parts namely, Introduction, Proposed Framework, Experimental Results and Conclusions, and future work.

To estimate intrinsic dimensionality, there are many theoretical and practical models used. The Maximum Likelihood Estimator (MLE), eminent among researchers and developers, is used in many applications to get very good results. He explains MLE in detail and why he used it in this experiment with a mathematical formula and a diagram. He then describes the applications of Intrinsic Dimensionality. He talks about Artificial Neural networks and shares the importance of Deep Learning, along with this he explains overfitting, underfitting, and good fit. Next, he describes how overfitting in the model can be eliminated.

He continues his presentation by explaining the whole process of Dropout in deep learning. He explains the proposal of the dropout technique by Hinton at el's two approaches to defining value. Few researchers have come up with different techniques like Variational, Biased and Crossmap dropout to assign different rates to weights. He explains how images, are the real challenge for researchers. He discusses his various experiments performed on different datasets and their respective results.

He concludes the presentation by stating some important points that should be considered while performing the experiments. He then talks about his future work to be done in this field of research.

Contents

List of Symbols and Abbreviations

AI - Artificial Intelligence
MLE - Maximum Likelihood Estimation
MDS - Multidimensional Scaling
MNIST - Modified National Institute of Standards and Technology Database
CIFAR - Canadian Institute For Advanced Research
SVHN - Street View House Numbers

List of Figures

1 Introduction

1.1 Background on Problem Domain

Most modern classic AI, data mining, machine learning, and many more, frequently are dealing with a large quantity of data that are usually characterized by a huge number of features. Although working with all these high dimensional datasets can produce very troublesome situations. To avoid such dangers of high dimensionality, the very first step in many practical scenarios is searching for a less complex representation of the data.

1.2 Problem Statement

In today's fast-moving world, where data is very crucial, storing data into different datasets is very important. When these stored data are used in various fields of computer science to study a trend, it becomes hard to process the large datasets. To reduce these datasets by using dimensionality reduction techniques and produce a new version of the dataset.

1.3 Methodology

To overcome the trouble caused by large datasets, the dataset to be reduced into an updated dataset in which the target dimension value should not be too less otherwise some important data might be lost, also it should not be too high otherwise some irrelevant and noisy data might persist in the datasets which make some algorithm of target application unstable. The optimal value must be such that the data should be reduced with minimum loss of information, this is known as Intrinsic Dimensionality[2].

1.4 Assumptions and Limitations

The latent feature that will be obtained from the reduced dataset must be such that no important data must be lost and should not hold more unnecessary data. The Early theoretical models and projection models, which are global methods, are unsuccessful in estimating with complex datasets.

2 Background

Today data is everywhere and is easily available from various applications and websites. Using these data in a proper way in various fields of computer science like Machine Learning, Artificial Intelligence, and many more, one can predict various trends. It is not mandatory that all the data that is store in a particular dataset is important, sometimes it may be junk for some specific usage. So, for this purpose, the datasets having huge quantities of data are reduced to a newer version of the dataset, which has fewer data than the original one but covers every important latent feature of the original dataset. The process of reducing the dataset with minimal loss of data is called Intrinsic Dimensionality.

Figure 2.1: A rich Image Dataset consisting tons of different images. Source:- A grid of images from the Microsoft Celeb(MS-Celeb-1M) dataset.

Figure 2.2: A Reduced Image Dataset, after the large dataset is being reduced by removing noisy and not useful data from it. Source:- Image Datasets

Figure 2.1 shows a rich image dataset and figure 2.2 shows a reduced image dataset. Here there will be images with less minimal loss and mostly all unique images would have been covered. These two figures collectively explain the concept of intrinsic dimensionality. Figure 2.1 shows a rich image dataset and figure 2.2 shows a reduced image dataset. Here there will be images with less minimal loss and mostly all unique images would have been covered. These two figures collectively explain the concept of intrinsic dimensionality.

2.1 Intrinsic Dimensionality Estimators

Intrinsic Dimensionality is an active subject for all researchers, which leads to the development of many theoretical and practical models. Early theoretical models such as Lebesgue, Hausdroff dimensions, box-counting and packing dimensions, and multidimensional scaling. Projection model like Principal component analysis and local variations of MDS tries to explicitly find the mapping that projects the data into appropriate space. Model approaches like locally linear embedding, nearest neighbor estimator explore the geometry of the dataset by estimating some statistics related to neighboring points. The last, Fractal based approaches analyze the space-filling capacity of a local neighborhood to infer an estimation. Some examples of this include Mind estimator, expansion and correlation dimensions, and some interesting nearest neighbor approaches like the incising balls and the maximum likelihood estimator. Among all these models, theoretical and projection models tackle the dimension using the entire dataset and therefore considered global methods. When some complex dataset is being presented to such models, they are unsuccessful. Local approaches like geometric and fractal models look to quantify the estimation in the vicinity of the provided dataset which provides them computational efficiency since they just deal with distances between neighboring points.

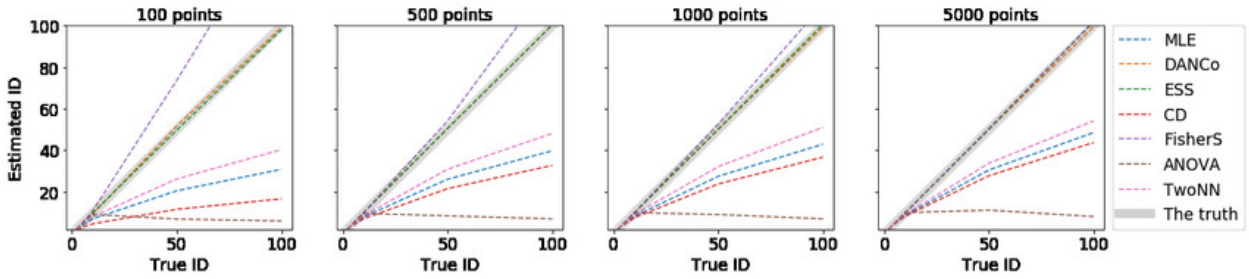

Figure 2.3: Intrinsic Dimentionality of different Intrinsic Dimensionality Estimators
Source:-https://www.researchgate.net/figure/Ability-of-various-estimators-to-quantify-high-intrinsic-dimensionality-from-different$_fig5_338989808$

The image shown above shows intrinsic dimensionality of datasets computed through different different intrinsic dimensionality estimators

2.2 Intrinsic Dimensionality – Maximum Likelihood Estimation

In statistics, Maximum Likelihood Estimation is a method of estimating and observing the parameters of a probability distribution by maximizing a likelihood function, by which the observed data is most probable under the assumed statistical model. The maximum likelihood estimate is known as the point in the parameter space that maximizes the likelihood function[3, 4]. It has become very popular among the pioneers of the computer science field and developers because it is intuitive and flexible. Moreover, it is being used in many applications to get very good results. It notes the distance between close neighbors and derives the final estimator using a Poisson process approximation by randomly sampling within a given radius, around each interest point. The likelihood equations deal an estimate of the intrinsic dimensionality in the neighborhood defined by equation 2.1,

$$m_k(x) = [\frac{1}{k-1}\sum_{j=1}^{k-1}\log\frac{T_k(x)}{T_j(x)}]^{-1} \tag{2.1}$$

where $T_k(x)$ and $T_j(x)$ are representing Euclidean nearest neighbors

Here in the place of using a simple average of all the local estimations in equation 2.1, a more refined and accurate direction proposes an average of inverses shown in equation 2.2,

$$m_k = [\frac{1}{n}\sum_{i=1}^{n} m_k(x_i)^{-1}]^{-1} \tag{2.2}$$

where n represents the total amount of neighborhoods sampled.

A note while using this technique is choosing a correct value for the K parameter. Because different values of it will result in different dimensionality estimates for a particular dataset.

The main principle of the likelihood estimation is to make deductions about the population that is most likely to have generated the sample, from a random sample of an unknown population, specifically the joint distribution of the random variables.
As Maximum Likelihood Estimation's quality and robustness of its approximations and its performance in terms of computational cost and run-time, all the experiments carried out were done using Maximum Likelihood Estimator.

Poisson distribution process can be used to model events where one has to know the number of meteorites greater than 1-meter diameter that strike the Earth in a year, a number of photons hitting detector in a particular time interval[5]. There are some assumptions and validations that must be true while using Poisson distribution.

The main advantages of Maximum likelihood are that it provides a consistent methods to solve the problem of parameter estimation. It can be used in reliability analysis to censored the data under different censoring models. Moreover, their minimum variance become unbiased when the sample size is increased. The demerits of likelihood function are that they are heavily biased when a small sample is considered. Also, they are sensitive to the choice of initial values[6].

2.3 Intrinsic Dimensionality - Applications

Intrinsic Dimensionality has many applications in different fields of computer science. The applications of intrinsic dimensionality is not only restricted to the dimensionality reduction problem but also cover a wide range of barriers where it can be used. In the identification of some data from a particular dataset which tends to be very different from the rest of the data in that dataset, it can be achieved by using intrinsic dimensionality. Intrinsic dimensionality can be used Some of them are Analysis of Search Indices, Adversarial Attack, and Feature selection/reduction.

Apart from these, we might see intrinsic dimensionality being used exclusively in Big Data and in many Machine Learning applications. Having an estimate of dataset's intrinsic dimensionality sometimes become very important in choosing Machine Learning Methodology and its applications to any of its given problem. While designing and configuring machine learning models, intrinsic dimensionality provides intuition of the nature of data. Moreover, it plays a vital role in model selection and also understanding model behavior.

2.4 Deep Learning - An Introduction

Artificial Intelligence in computer science is a field that is progressing steeply day by day. To cater various functionalities to the naive users and to understand the data from various ongoing trends, different datasets, and many more, it is inevitable for engineers, data scientists to project these data and fetch out some important functionality or a money-making trend for a new business. It learns from by itself from the training data available and then predicts exciting trend which is used in decision making for big and small businesses[7].

There are mainly three layers namely the input layer, hidden layer, and output layer. The hidden layer can be more than one. There are specific weights awarded to each input layer and this processing is done in hidden layers which in the end result in output[8]. It has vast importance in fields of Natural Language Processing, Computer vision, Industrial Processing, and many more.

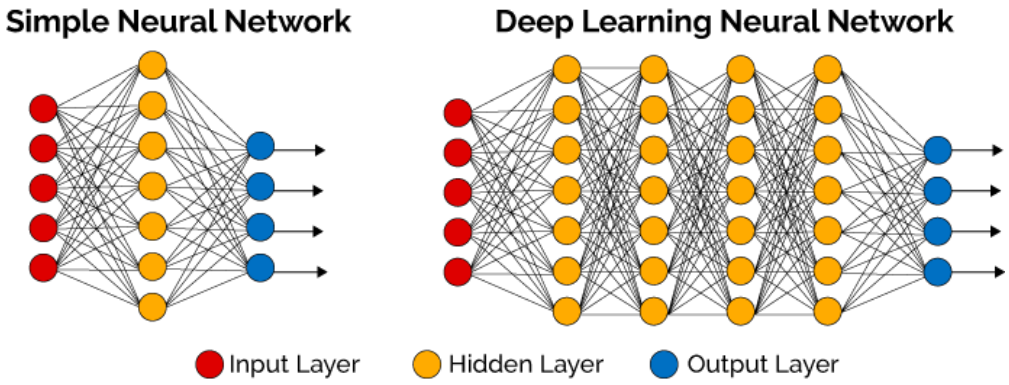

Figure 2.4: Deep Neural Network. Source:- Journal of Educational Computing Research 57(4):073563311875701 DOI:10.1177/0735633118757015

3 Technical Contributions

3.1 Deep Learning - A deep dive

When preparing to design a deep learning model, knowing what type of data, the quantity of it, and its structure. These things are very important in determining aspects of the designed model like a type of network, architecture, and size, which are vital elements to design a model.

When very limited data is available to feed the training data causing underfeeding issues or a very simple model is being designed, underfitting can be observed. this is being observed when no relationship between the input and output can not be obtained[9]. The result of such training data will be very poor and biased.

A model with too much noisy data and an exaggerated number of parameters leads to overfitting the data[10]. The result of overfit models will be very good on training data but worse on new data. Network Capacity can be increased to overcome underfeeding of data like it is always a good practice to build a large network that has more data and in the end, some techniques can be used to overcome the overfitting of data in a model.

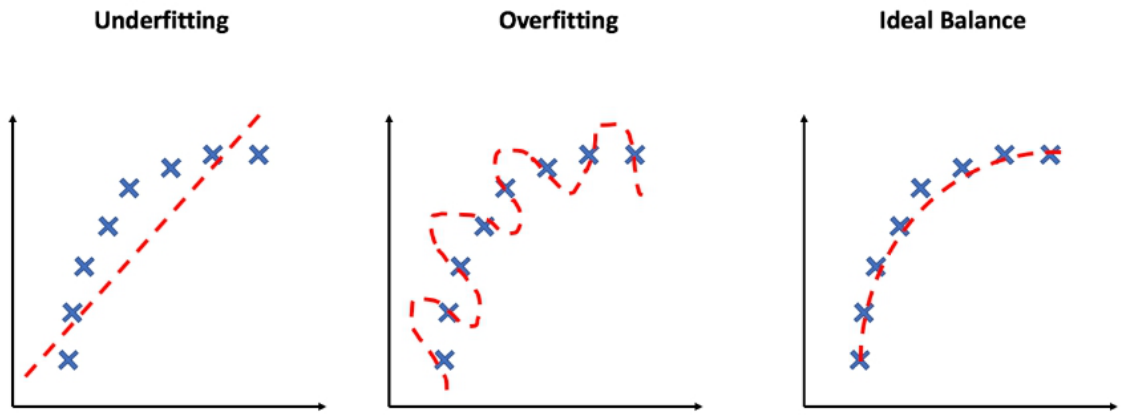

Figure 3.1: Visual demonstration of Underfitting, Overfitting and Ideal balance.
Source:- https://subscription.packtpub.com/book/data/9781838556334/7/ch07lvl1sec82/underfitting-and-overfitting

3.2 Deep Learning - Overfitting

To avoid the overfitting model, a simple solution is to expand the available data through different techniques. This can be done with the help of Sampling Techniques, Data Augmentation, and many more different methods/techniques. Secondly, the training of data should stop using the stopping approach when overfeeding of data is discovered. Thirdly, ensemble learning can be used where the weighted outputs were averaged to give the final result. Finally, Regularization is also a technique that could be used, in this, the complexity of the model is reduced during training. Here dropout simplifies the network and exploits some concepts of model combination and network. Let's deep dive a little more in dropout.

3.3 Deep Learning - Dropout

The technique of dropout is very simple. It just eliminates the units and their connections in the network temporarily, during the training phase. It offers a very computationally cheap and very effective regularization method that is used to prevent the overfitting of data in deep neural networks.

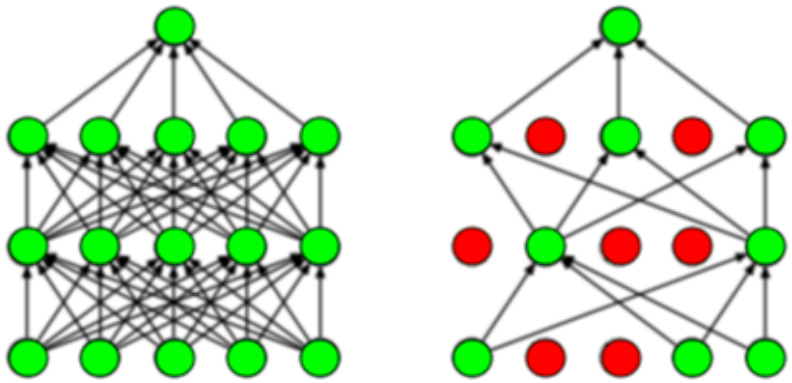

Figure 3.2: Deep Learning: Dropout.
Source: https://medium.com/analytics-vidhya/a-simple-introduction-to-dropout-regularization-with-code-5279489dda1e

As seen in figure 3.1, some layer's output nodes are intentionally dropped out. Here some probability is assigned to some output node and is then calculated. This makes the network layers of the model co-adapt to correct the error from prior layers, which results in a more robust model. A common value is a probability of 0.4 for keeping the output of hidden layer node and a value up to 0.8 for retaining visible layers inputs[11].

The various approaches of dropout consist of Adaptive dropout, Maxout Networks, and Concrete dropout. The proposal of the dropout technique by Hilton et al, suggested two approaches

namely, the Hyper-parameter search algorithm to find the optimal value for hyperparameters, and the second is per-defined results.

3.4 Deep Learning - Images

One most significant data being governed applications of deep learning or dropout is widely and extensively used are images, which still present a big challenge for researchers and are the main focus of study. The image dataset takes every image's pixel value as an input in the model. However, it is not that much important to learned every pixel representation as not every pixel draws an important feature. From studies, it is been studied that the image datasets have very low intrinsic dimensionality value as there are a small amount of identified features of image representation in the datasets. Should there be some interpretation of the dataset's intrinsic dimensionality to get a better quantity of units by reducing the network size? For this dropout, a regularization technique can be used to reduce the units during the training of the model to prevent overfitting.

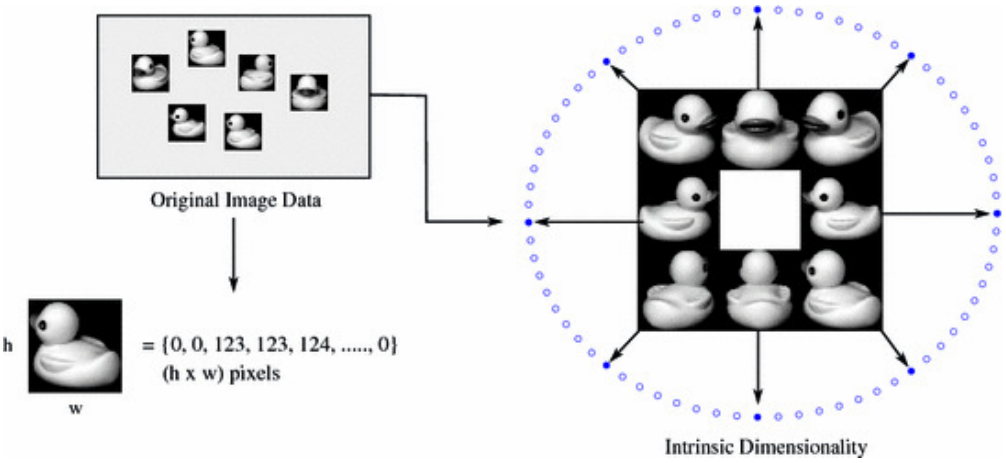

Figure 3.3: Image Intrinsic Dimensionality.
Source: https://link.springer.com/chapter/10.1007/978-3-319-03943-5_4

3.5 Proposed Framework

Considering a dropout for the approach will need defining and computing a quantity p' dependent on the intrinsic dimensionality of the dataset and the dimensions of an image, which can be utilized as the rate of the dropout layers. Here, they considered several formulations to compute this quantity and used them to validate the effectiveness. Results obtained from this test depict that computing the percentage of intrinsic dimensionality against the dimensions of an Image provides a good estimation for this value. Thus they define P' as shown in equation 3.1.

$$p' = \frac{m_k}{(c\ x\ wx\ h)}\ x\ 100 \tag{3.1}$$

where m_k is the intrinsic dimenstionality of the dataset D.
;c is the number of channels[12]
w is the Image's width
and h is the Image's height

3.6 Experiments

The experiments conducted, used five eminent datasets used for classification problems of Image. These datasets are namely, MNIST, CIFAR-10, CIFAR-100, SVHN, and ImageNet. The first task was to define the intrinsic dimensionality of these datasets, for the estimation, they have used MLE with some predefined value of k. Along with this, p' derived in equation 3.1 is also used for every MLE configuration. The value of p' is not the same for all, but is distinct which helps in getting some exciting trends from the obtained result.

Dataset	MNIST		CIFAR-10		CIFAR-100		SVHN		ImageNet	
	ID	p'	ID	p'	ID	p'	ID	p'	ID	p'
MLE (k=3)	7	0.30	13	0.42	11	0.36	9	0.29	26	0.02
MLE (k=5)	11	0.47	21	0.68	18	0.59	14	0.46	38	0.03
MLE (k=10)	12	0.51	25	0.81	22	0.72	18	0.59	43	0.03
MLE (k=20)	13	0.55	26	0.85	23	0.75	19	0.62	43	0.03

Figure 3.4: Experiment Data. Source: Original Seminar

Next, applying the p' values as the dropout rates is the followed work to be done. He has selected deep learning models that use dropout regularization for each and every dataset and replicate every single item of the original data unless the measurement accuracy is outperformed for each of the datasets.

Two scenarios were designed for training data, wherein the first scenario was to take a 10% random sample from the dataset at least twice the epochs as proposed by the original work. The second scenario reestablished the reported reference in terms of data use and the number of epochs. He trains each model in different scenarios under the raw configuration and using the new dropout rates.

The result that will be deduced from by comparing all the results, the importance of the new rate for the model's performance. He explains the results of each dataset briefly. Here is the explanation of the result of each dataset,

3.6.1 Experimental Results - MNIST

The result reported for the MNIST dataset, accuracy has been reported for training and testing the model in both scenarios under different configurations. In scenario 1, overfitting was observed, with some exceptional results on the training set and poor results on the testing set. In both the scenarios the best performance was observed when the dropout rate was 0.47.

Model	Epochs	Accuracy	SOTA
EnsNet	1300	99.85	99.87

	Scenario 1		Scenario 2	
	training	test	training	test
base (0.35)	92.07	77.11	99.96	99.85
k=3 (0.30)	94.06	66.29	99.95	99.42
k=5 (0.47)	93.91	81.17	99.97	99.90
k=10 (0.51)	91.54	72.29	99.83	99.01
k=20 (0.55)	90.55	69.37	99.76	98.84

Figure 3.5: MNIST Dataset Result. Source: Original Seminar

4 Conclusions and Future Work

4.1 Conclusion

The natural image dataset's intrinsic dimensionality value is low relative to the high-dimensionality pixel representation of images. From the experiment performed on five different datasets, it has been observed that the estimations of intrinsic Dimensionality done by MLE on these datasets are directly dependent on the value of its respective k.

It has been proved that the intrinsic dimensionality estimations are independent of any selection data points i.e data sampling[13]. Also, exploiting the relation between intrinsic dimensionality can result in better model performance. While experimenting the highly regularized models consume more time to converge.

It can be concluded that the intrinsic dimensionality of a dataset plays a vital role not only in designing the model and understanding its behavior but also in model configuration, modern artificial intelligence applications. The dropout rates were calculated for the particular value of k for MLE, in the experiment performed on different datasets. This is because exceptional results were obtained when the value of k is 5. The results of the experiment prove that the tune configurations have a major impact on the performance of the model mainly in overfitting conditions. Many future works can be done in this field and the future seems brighter.

4.2 Future Work

The experiments performed were exclusively for Image classification but it can also be used for other types of data and the results of this experiment are to be performed by the result obtained from the experiment performed of Image Classification.

Improvement can be done in the computation method of p' by considering different aspects or providing a more meaningful interpretation. In the future obtaining more stable and reliable Intrinsic dimensionality estimation could breach the gap between Intrinsic dimensionality and different aspect of learning.

Bibliography

[1] Some content of report is referred from the original seminar and also some images and experimental result table is cited from it.

[2] Intrinsic Dimension
https://en.wikipedia.org/wiki/Intrinsic_dimension

[3] Maximum Likelihood Estimation
https://en.wikipedia.org/wiki/Maximum_likelihood_estimation

[4] Rossi, Richard J. (2018). Mathematical Statistics : An Introduction to Likelihood Based Inference. New York: John Wiley & Sons. p. 227. ISBN 978-1-118-77104-4.

[5] Poisson Distribution
https://en.wikipedia.org/wiki/Poisson_distribution

[6] Merit and demrits of Maximum Likelihood Estimations
https://www.itl.nist.gov/div898/handbook/eda/section3/eda3652.htm

[7] Deep Learning - Overview
https://www.investopedia.com/terms/d/deep-learning.asp

[8] Artificial Neural Network - Layers
Mohammed Imran, Sarah A. Alsuhaibani, in Intelligent Data Analysis for Biomedical Applications, 2019

[9] Underfitting
https://www.ibm.com/cloud/learn/underfitting

[10] Overfitting
https://www.ibm.com/cloud/learn/overfitting

[11] Deep Learning - Dropout
Dropout: A Simple Way to Prevent Neural Networks from Overfitting, 2014.

[12] Number of Channels of an Image explanation
https://datascience.stackexchange.com/questions/64278/what-is-a-channel-in-a-cnn

[13] Data Sampling
https://searchbusinessanalytics.techtarget.com/definition/data-sampling#: :text=Data%20sampling%20is%20a%20statistical,larger%20data%20set%20being%20examine