1. **How do you make a dataset bigger for Deep learning?**

Ans: Data Augmentation is the technique, and it depends on the Neural Network model and Quality & Quantity of Data. Data augmentation is a technique to artificially create new training data from existing training data. It helps us to increase the size of the dataset and introduce variability in the dataset, without collecting new data. The neural network treats these images as distinct images anyway. So, to get more data, we need to make minor alterations to our existing training data. Here we are specifically talking about Image Data Augmentation. These alterations include flipping the image horizontally, vertically, padding, cropping, rotating, scaling and few other translations.

https://research.aimultiple.com/data-augmentation/#:~:text=Data%20augmentation%20is%20useful%20to,performs%20better%20and%20more%20accurate.

* 1. **Do we need large datasets for Deep Learning?**

Deep Learning does not really require large sample sizes, Neural Networks are not smart and treat images with minor alterations as distinct.

* 1. **How much data is needed for Deep Learning?**

Depends on the Learning Curve, which is dependent on the factors of classification method, complexity of the classifier and how well the classes are separated.

Rule of Thumb, the sample size is usually discussed in relation to model complexity (number of cases: number of variates), whereas absolute bounds on the test sample size can be given for a required precision of the performance measurement.

1. **How do you use the dataset with synthetic data to train your deep learning algorithm?**

Ans: Dataset with a mix of Raw and Synthetic data is a good dataset to train on. Using too much Synthetic data depends on application but limits prediction accuracy.

1. **Why do you use an 80% (Training) – 20% (Test) data split?**

Ans: Power law is driving the train and test split; Pareto Principle seems to be driving this split. that 80% of outcomes are due to 20% of causes.

**Math behind the split**

Ex: Taking the first rule of thumb (i.e., Validation set should be inversely proportional to the square root of the number of free adjustable parameters), you can conclude that if you have 32 adjustable parameters, the square root of 32 is ~5.65, the fraction should be 1/5.65 or 0.177 (v/t). Roughly 17.7% should be reserved for validation and 82.3% for training.

Paper that explains this well, **A scaling law for the validation-set training-set size ratio**

Scaling Law: In 1997, a new method was discussed in a paper called A scaling law for the validation-set training-set size ratio (Guyon).

Here, they reference “the best training/validation split for a specific problem: preventing overtraining of neural networks. They find that the fraction of patterns reserved for the validation set should be inversely proportional to the square root of the number of free adjustable parameters. Our result generalizes and confirms their result.” (Guyon, 1997).

In essence, the split is determined by how many unique features are in the dataset (not including the target) and not the number of observations, which is contrary to most opinions on the subject.

https://towardsdatascience.com/finally-why-we-use-an-80-20-split-for-training-and-test-data-plus-an-alternative-method-oh-yes-edc77e96295d

1. **How much Synthetic data is acceptable?**

Ans:

*A scaling law for the validation-set training-set size ratio*

We derived a formula for splitting the training database into training set and validation set valid for large training databases and small error rates.

If we call N is the number families of recognizers, hmax the largest complexity of those families, f the validation set size and g the training set size, the ratio f/g scales like ln N/hmax. For instance, for N 10, if hmax 100, 25% of the training data should be reserved for validation whereas if hmax 1000, only 10% is sufficient.

*What Makes Good Synthetic Training Data for Learning Disparity and Optical Flow Estimation?*

Augmentation has inherent limits, and we achieved the best results when using both augmentation and as much raw data as possible.

*RarePlanes: Synthetic Data Takes Flight*

When using just synthetic data, the model was only about 55% percent accurate, whereas when it only used real-world data that number jumped to 73%. But by making real-world data 10% of the training sample and using synthetic data for the rest, the model’s accuracy came in at 69%.

Literature review

Next steps,

Copy paste into the R template