

Understanding the impact of dataset size on training accuracy

Project Hypothesis

- Question: We are curious to study the impact of increasing the number of images during training (i.e. dataset size) on the model accuracy on a test dataset.
- Expected answer: We expect increasing dataset size will increase test accuracy.

Dataset

- Description: We will be using the CIFAR-10 dataset. The dataset contains 60000 32x32 colour images. Labels: Each image comes with a category label for the object in the image.
- Publicly available? Yes, the dataset is publicly available. Here is the link: <https://www.cs.toronto.edu/~kriz/cifar.html>
- Did you find any code-bases using this dataset? Yes, here's a link: https://pytorch.org/tutorials/beginner/blitz/cifar10_tutorial.html
- Do you have the compute power to train/test on this dataset? Yes, this is a small dataset (unlike ImageNet), so Google Colab is sufficient. (Note that ImageNet is also OK if you have access to other GPUs outside of class. If not, avoiding large datasets like ImageNet is a good idea).

Model Input

Model will be fed in CIFAR-10 images.

Model Output

Output will be the category of image.

Model Architecture

We will use a standard ResNet architecture. Here is a code-base using that architecture:

https://pytorch.org/hub/pytorch_vision_resnet/

Evaluation Metric

Accuracy will be measured as % of images correctly classified.