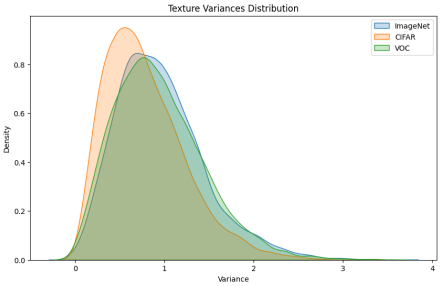
Dataset generalization via continual learning-inspired training schemes.

Abstract:
The ability of machine learning (ML) models to effectively generalize data is often a challenge for ML practitioners when creating models. Datasets can have various ways in which they differ - some can be measured quantitatively (such as texture, average RGB value, etc), and some are more qualitative (such as object angle, occlusions, etc). ML models' ability to generalize and perform well on unseen data is important in many applications. One of these applications is with models in consumer products, as it can be hard to collect this data or to effectively anticipate all of the different use cases. We explored this idea by using a continual learning-inspired training scheme, where we used an EfficientNet model (pre-trained on ImageNet) and trained it on 7 classes common between ImageNet, CIFAR-10, and PASCAL VOC 12. The first round of training was on ImageNet and then measured performance on all three datasets. The second round of training was on ImageNet and CIFAR-10, and performance was measured again. We found that by adding new datasets over time and analyzing performance per class, we gained insights into how the model generalizes to different datasets as new data is added.

Additional Details:
Although datasets may have similar classes and collection methods, they can still have differences in style that can impact how well the model generalizes to other sources of data (such as other datasets or data from new agents). In **Figure 1** and **Figure 2**, we show some quantitative measures to capture these differences. In **Figure 1**, we see that the datasets all have different texture variances and edge densities, with CIFAR being the most distinct of the datasets. In **Figure 2**, we can view the RGB channel distributions, further showing that the datasets are different from each other, even though their classes are the same.



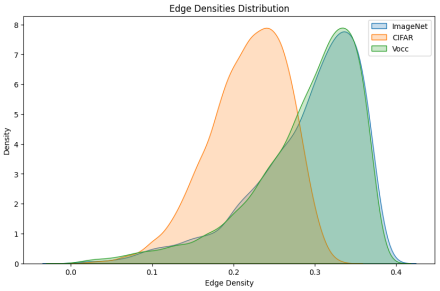


Figure 1

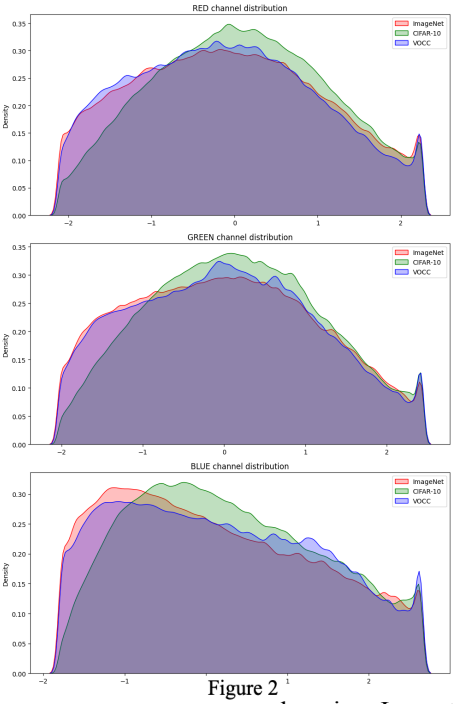
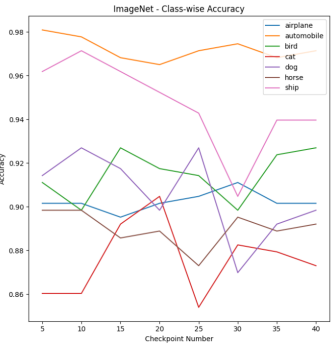


Figure 2

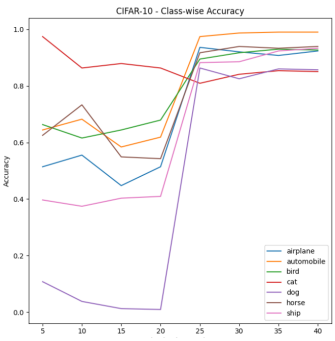
This project was run on Google Colab, which although provided with resources better than our local machines, we were still constrained in terms of data processing and computational power. Thus, for this project, we focused on the 7 common classes between ImageNet, CIFAR-10, and PASCAL VOC 12 (airplane, automobile, bird, cat, dog, horse, ship). As resources were limited, PASCAL VOC 12 was downloaded offline first to remove extraneous classes. ImageNet was downloaded by class from the ImageNet website, except for the dogs class, which was extracted by using the Stanford Dogs Dataset, which is pulled from ImageNet, and combined all the classes to form a general dog class within ImageNet. These two datasets were uploaded into the Colab runtime to be used in model training and evaluation.

A key aspect of this project is to develop a training scheme that is inspired by continual learning. In continual learning, the model does not show all of the data at once. Instead, as the model trains, the data distribution changes over time[1]. We decided on a simple, two-step scheme to emulate the data distribution changing over time. Phase 1 involves training on the reduced ImageNet dataset, and evaluating on all three datasets. Phase 2 involves training on ImageNet and CIFAR-10, and evaluating all three datasets. Each step was 20 training epochs. CIFAR was picked for the second as it was the most differentiable dataset of the three (per **Figures 1** and **2**), and ImageNet for phase 1 as the model was pre-trained on this data, so the model performance on ImageNet would already be decent, so it would not be a fair comparison to evaluate the model performance on ImageNet as if it was

[1] van de Ven, G.M., Tuytelaars, T. & Tolias, A.S. Three types of incremental learning. *Nat Mach Intell* 4, 1185–1197 (2022). https://doi.org/10.1038/s42256-022-00568-3



an unseen dataset. In **Figure 3**, we can see that as the model trains in phase 1, it can generalize fairly well to the PASCAL VOC dataset, but the accuracy does dip slightly as time goes on. When we enter phase 2 and introduce CIFAR, the PASCAL VOC accuracy starts to pick up again. In **Figure 4**, we can view the performance across time for each dataset, by class. Here, we can see that some of the classes in VOC and CIFAR always did well, even without direct training during phase 1, and some classes struggled more than others even after extensive training within ImageNet. This seems to imply that generalization problems may also be class-specific, and even dependent on the nature of the class and how data is collected for it. For example, the dog class always seems to struggle, which may be an artifact of the fact that there are many more distinct dog breeds than there are cat breeds.



Something else we wanted to look at was the activation of the model for the different classes and datasets. In **Figure 5**, the top row shows the t-SNE of the last model layer during the first 20 epochs during phase 1, and the bottom row is during phase 2. In phase 1 we see something interesting - that although the classes are grouped in the same general area, there are clear separations between datasets, even within the same class. This is seen in the orange sections and the green and blue sections in the phase 1 graphs. However, during phase 2, the class separations become much cleaner, and there is no more separation by dataset, implying there is greater cross-dataset generalization.

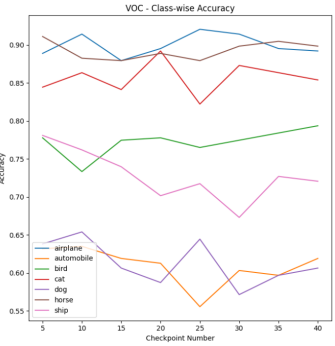
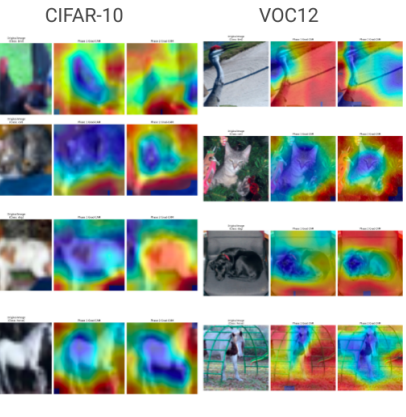


Figure 4

In **Figure 6**, we can see the model activations on a few sample images from each class. The far left column is the raw image, the middle is after phase 1 training, and the right is after phase 2 training. As we can see below, as the model trains more, the features of the images become much more focused on key features, implying that there is also some greater generalization occurring within the model.



As we can see here, but including more varied data into the model, the model starts to learn more specific class features, improves generalization, and overall becomes more robust. Some improvements here include changing the ratio of the datasets as time goes on, and also potentially changing the ratio of the classes depending on which classes need improvements. We also experimented with a vision transformer model, but that did not perform as well as we had hoped as a result of our smaller datasets, so increasing the scale of the datasets used is also an area for improvement.

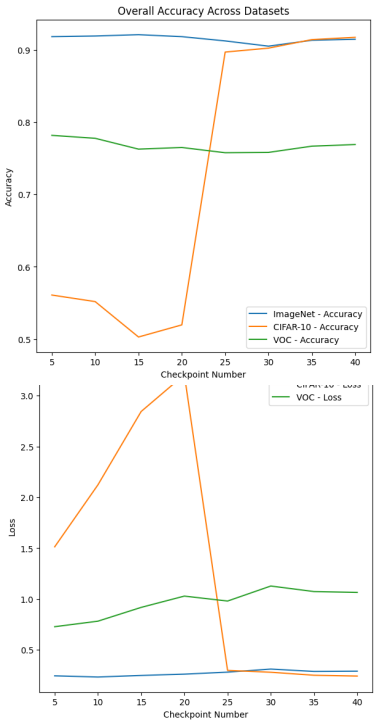


Figure 3

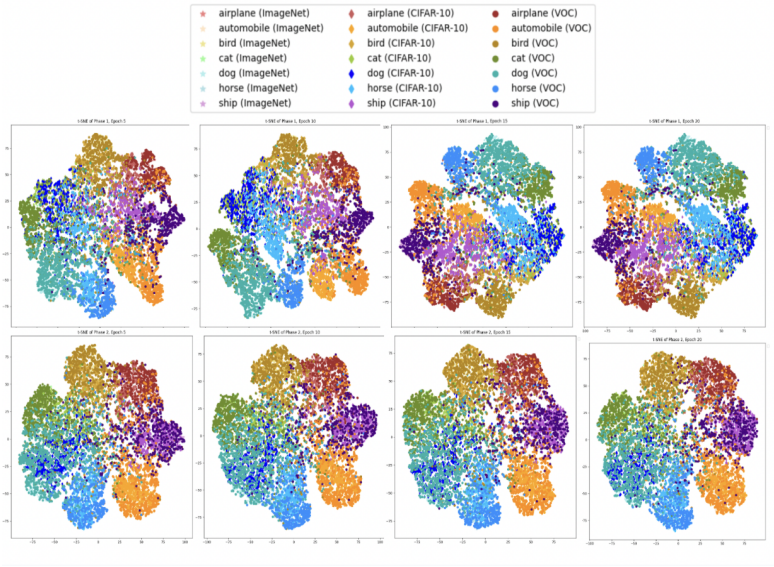


Figure 5