Project 2:

Simple Image Corruptions Can Destroy Classification

In this project I investigate how resnet50 can be corrupted by simple image transformations, namely gaussian noise and rotation. I chose gaussian noise and rotation because they are simple, and one would expect models to be robust to them. Many models trained on ImageNet (like ResNet), apply some amount of random noise and rotation during training. I would expect this to make the models more robust to these simple transformations, but that is not what I saw in my experiments. I used ResNet 50 because ResNet 18 incorrectly (albeit understandably) classified the image with no corruption.

I added noise drawn from a gaussian distribution with mean 0 and standard deviation 0.2 \* i where i is the index of the image (i.e. 0, 1, ..., 9) to the original image. All of the images are preprocessed according to the standard ImageNet process prior to being corrupted.

A picture containing text, grass

Description automatically generated

Figure : Each image shows the predicted class followed by the model's confidence in that class.

As you can see the model initially classified the image correctly, but as the amount of noise increases the network seems to guess somewhat randomly from other dog classes. Interestingly, the model shows highest confidence in the correct prediction when some level of noise is applied (standard deviation of 0.2 and 0.4 respectively), this could be a consequence of training with noise. It is also worth noting that all the predictions, except for the final image, are different dog breeds. This hopefully indicates that although the gaussian noise is corrupting the image, that the model can still correctly classify more general classes (e.g. dog).

From my professional work, we often deploy neural networks to more challenging scenarios than ImageNet. They often have variable angles to the target, distance from target, and are non-stationary, which makes ImageNet pretrained networks perform quite poorly. In practice we retrain and finetune these networks, but for this project I wanted to investigate how simply rotating an image would affect its classification. It’s worth noting that while most classes should be invariant to rotation, not all are, so I picked a class from the former.

Graphical user interface, application

Description automatically generated

Figure : Each image is rotated by 36 degrees and shows the predicted class followed by the model's confidence in that class.

As you can see, the model correctly classifies the image when it is near the canonical rotation and quickly loses accuracy with rotations about 40 degrees. Unlike the previous corruption, the predicted classes are mostly far from the correct class (golden retriever) or more general class (dog). This simple example illustrated two main issues with the current computer vision research paradigm: neural networks are overly dependent on spurious features created from pixel colors; benchmark datasets are too refined and do not capture the amount of variation found in practice. Research into self-supervised learning, where models more often attend to salient image features, is being done to address the first concern, but as far as I know little is being done to produce more robust open-source datasets. There are some difficult datasets being used for unsupervised domain adaption tasks, which is progress in the right direction, but still far from an ideal classification benchmark.

Chart, line chart

Description automatically generated

Figure : How the model's confidence in predicting golden retriever changes as we progressively corrupt the image.

For both corruptions, the model’s confidence in predicting the correct class decreases as we further corrupt the image. For future work, I want to see how self-supervised networks fair against these corruptions. They should be more robust to the corruptions, but I assume they will still suffer significant classification performance degradation.