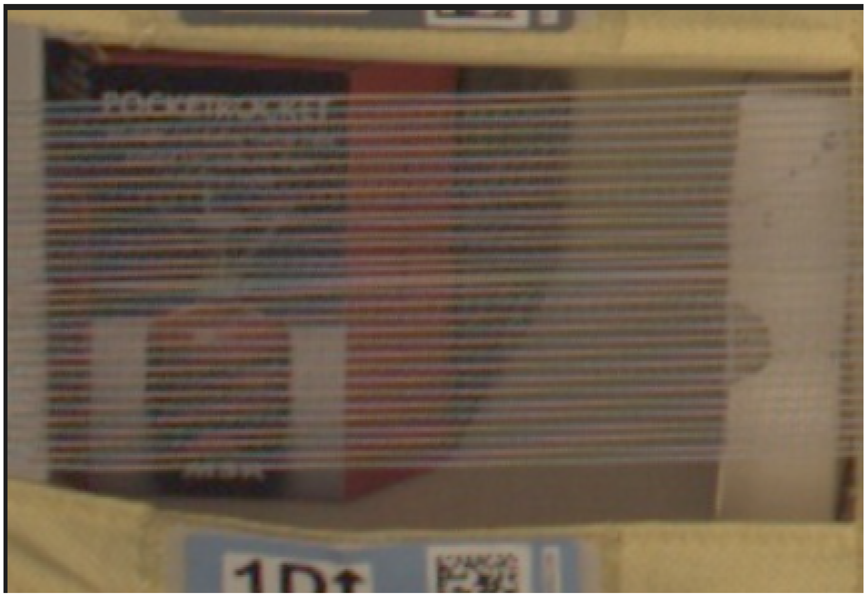
Warehouse Classifier

Using ResNet and ViT to classify images from the Amazon Warehouse dataset



Tony Ho

01.11.2023
Udacity Machine Learning Capstone

INTRODUCTION

For my Udacity Machine Learning Capstone project, I selected the Inventory Monitoring at Distributions Centers project. In distribution centers, there are robots that hold bins containing a varying number of items. However, there may be errors resulting in the expected number of items in the bin not matching the actual number. Detecting these errors early can result in substantial benefits when applied to a large-scale operation, such as in an Amazon Distribution Center. The actual number of items can be identified by evaluating an image of the bin, taken from a top-down view. This project aims to train a deep learning model that can classify an image of the bin's contents as a bin with the correct number of items. For example, if the image showed only two distinct items then it should be classified as a two-item bin.

This problem has previously been tackled using SVMs and CNNs, particularly in the paper Amazon Inventory Reconciliation with AI (Bertorello et. al, 2018). The researchers achieved ~56% validation accuracy using ResNet34 on the AWS Warehouse dataset. This project attempts the replicate the results using the same architecture using AWS resources.

Recently, the Transformer architecture has been shown to perform very well in the NLP domain. An implementation of a vision transformer has been proposed by Dosovitskiy et. al (2020). Transformers are known to perform well when given vast amounts of data. With a training size of ~200,000 images, I investigate the performance of using the ViT model architecture for image classification on the AWS Warehouse dataset.

PROBLEM STATEMENT

Since transformer architectures are a recent innovation in the deep learning domain, it would be worth investigating whether using such architectures is inherently better than architectures developed years ago. For this task, I train the same ResNet model as described in the paper Amazon Inventory Reconciliation with AI and determine a benchmark accuracy on the test dataset. The specific architecture is the ResNet34 model, which features CNN blocks with skip connections to combat the vanishing gradient problem found in neural networks that have many layers. This model features ~21 million parameters.

The transformer architecture used is the Vision Transformer (ViT) model developed by

Google. Specifically, I use the ‘google/vit-base-patch16-224’ model which features 16x16 patches trained on the ImageNet dataset where the images are 224x224. This model has ~86 million parameters. The Huggingface implementation of the model enables fine-tuning of this model on the AWS Warehouse dataset, taking advantage of any intrinsic benefits gained from training on the ImageNet dataset, which has 1.3 million images.

The paper Amazon Inventory Reconciliation with AI achieved ~56% classification accuracy on their test set. The goal is to train the ResNet34 model and the ViT to similar or better accuracy in a sufficient number of epochs with AWS resources that are reasonably accessible to a student.

DATA

The AWS Warehouse dataset features 535,234 image and json file pairs. Each image has a top-down view of a bin that will contain 0-5 items. For this project, only images with 1-5 items are used. The images are resized to 224x244 and the pixel data is normalized. The train dataset size is 198,000 and the test dataset size is 22,000. Each image was also horizontally flipped to augment the dataset.

Class	1	2	3	4	5	Total
Train	23,397	43,403	50,813	45,193	35,194	198,000
Test	2,590	4,700	5,596	5,191	3,923	22,000

An example of an image from the dataset

METHOD

During training, the images are randomly horizontally flipped with a probability of 0.5. For the ResNet34 model, the images are normalized with a mean of (0.53, 0.4495, 0.3624) and a standard deviation of (0.1691, 0.1476, 0.1114). For the ViT model, since the model has already been trained on the ImageNet dataset, the normalization values are the default of (0.5,0.5,0.5) for both the mean and standard deviation.

The ResNet model was trained for 10 epochs with a batch size of 128 and an initial learning rate of 0.001. The training was spread across 4 instances of ml.m5.2xlarge for a total of 154,096 seconds, which amounts to ~42.8 hours. Using spot instances, the billable time is 74,024, which amounts to ~20.6 hours.

The challenges for this model involved understanding how to implement data distributed training across multiple CPU instances. For this architecture, the decrease in training time was directly correlated to the number of instances. Doubling the number of instances reduced the real-time training by half, scaling very efficiently. The real training time for this task across 4 instances is ~10 hours.

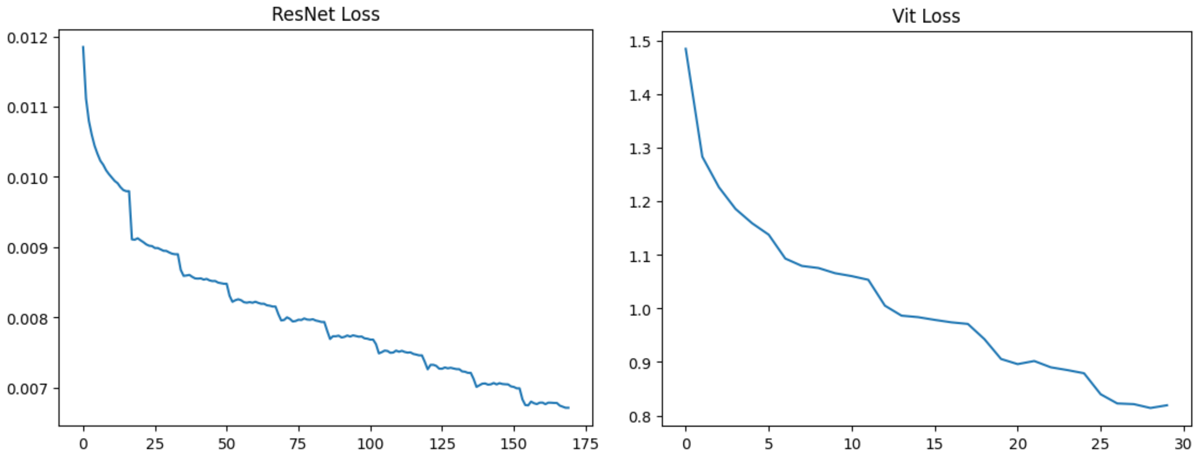
The ViT model was trained for 5 epochs with a batch size of 64 and an initial learning rate of 0.00002. The training uses 1 instance of ml.p3.2xlarge for a total of 11,483 seconds, which amounts to ~3.2 hours. Spot instances were not practical since there were frequent errors pertaining to insufficient capacity for p3 instances, resulting in training jobs timing out after 24 hours.

The challenge for this model involved using the Huggingface transformers module instead of the pure PyTorch/Torchvision modules used for the ResNet model. The transformers module implements the ViT model architecture and enables fine-tuning on a custom dataset more easily than the Torchvision implementation. However, the Huggingface datasets module is required, and to load image folders, a newer version of this module was needed. To address this, a Docker image can be used to extend an existing image provided by AWS. Also, this model requires a GPU instance, so it was necessary to request service limit increases for ml.p3 instances. These instances are expensive, so it was very valuable to establish a workflow for a local GPU instance for debugging and iterating.

Spot instances for ml.p3 instances proved to be challenging to use since it would take very long to get an instance. Interruptions were also frequent and made it impossible to train a model within 24 hours. However, training this model on a GPU instance was relatively very fast, completing an epoch in ~30 minutes when using the ml.p3.2xlarge instance. Using the ml.p3.16xlarge would complete training on 5 epochs in ~40 minutes, but the price would be 8x the cost per hour compared to the 2xlarge instance. A 5-epoch training run on the ml.p3.2xlarge instance would be practical for this task since it would complete in ~3 hours of real-time and have a relatively modest cost.

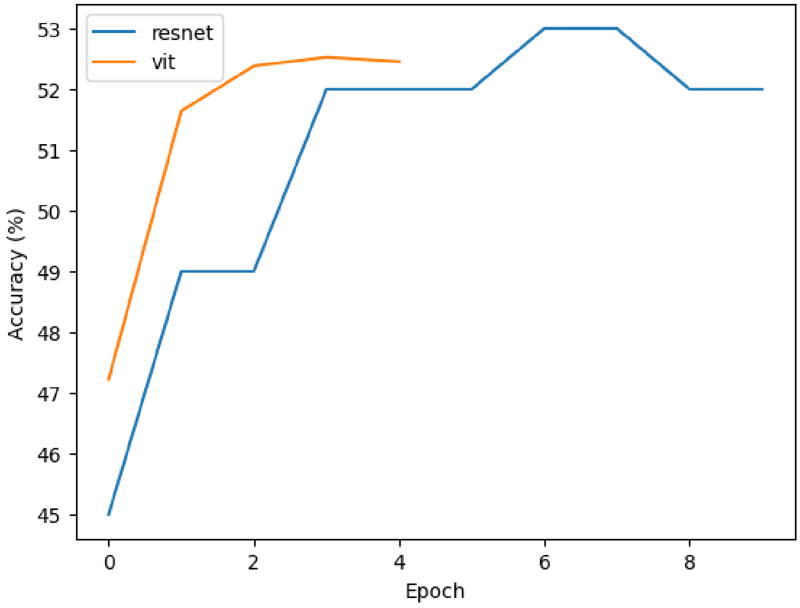
RESULTS

Training loss:



These graphs show the training loss evenly sampled throughout their respective training runs.

Evaluation Accuracy:



After each epoch of training, the model is evaluated on the test set. The accuracy plateaus around 52-53%.

CONCLUSION

Since evaluation accuracy for both training runs plateau, it seems it is sufficient to train the ResNet model for 10 epochs and the ViT model for 5 epochs. The training loss for both models don't plateau, suggesting that further training is likely to overfit to the dataset.

These results are similar to the result of the ResNet34 model trained in the Amazon Inventory Reconciliation with AI. Since the evaluation accuracy for the ResNet and ViT models are very similar, it seems to show that the transformer architecture is not inherently better for this image classification task.

However, for equal performance, training the ViT model on the optimal p3 instance kept resource cost to a minimal while delivering results much sooner.

REFERENCES

1. Rodriguez Bertorello, P., Sripada, S., & Dendumrongsup, N. (2018). Amazon Inventory Reconciliation Using AI. Available at SSRN 3311007.

2. Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28

3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.