Transfer Learning for Scene Recognition

Datuluna Ali Dilangalen and Hans Gustaf Capiral
Department of Computer Science
College of Engineering
University of the Philippines, Diliman

***Abstract*—In this project, we will attempt to apply the power of pre-trained convolutional neural networks to detect scenes within an image. For our dataset we will be using the MiniPlaces Dataset. Using this dataset we will apply Transfer Learning and retrain Google's Inception V3 Model.**

I. Introduction

A scene recognition engine would be a useful tool for digital marketing and social media platforms like Facebook, Instagram and Pinterest to narrow down the interests of its users. Along with a product recommendation engine, it can recommend items appropriate for recent events or recommend nearby shops that a user might also be interested in.

II. Short of Review of Related Studies

The main references that were used for this project are: Ondiekis, Convolutional Neural Networks for Scene Recognition[1], Donahue, et. al. , DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition [2] and Zhou, et.al, Places: A 10 million Image Database for Scene Recognition [3].

Ondiekis[1] research uses a convolutional neural network built from scratch and separately two models on the MIT Indoor67 and SUN 397 dataset, comparing the differences in properties between indoor-centric and outdoor-centric datasets.

Donahues, et. al. [2] research on the other hand evaluates whether features extracted from the activation of a deep CNN model trained on a large fixed set of object recognition tasks can be repurposed for a new generic but closely related task.

Zhous, et.al. [3] research involves the use of a multi-million dataset to train state-of-the-art CNN models and evaluate the performance of the models when trained on a very large dataset.

III. Methodology and Results

A. Dataset

For this project we will be training our model with the MiniPlaces Challenge Dataset, which is a subsample of the Places2 dataset. It contains 100,000 training images, 10,000 test images, and 10,000 validation images, each image resized to 128*128.

B. Transfer Learning

Building a CNN model from scratch and training it would need a lot of computational power, i.e. a certain number of powerful GPUs and time. In order to shorten the time we would need to complete a model for this project, we used Transfer Learning.

Transfer learning is a machine learning technique where a model that is trained to do a certain task is repurposed to do a different yet related task. In our project we used a pretrained model of Googles Inception V3 model, which is an image classifier trained on the ImageNet database that contains around 1.2 million images with 1000 categories.

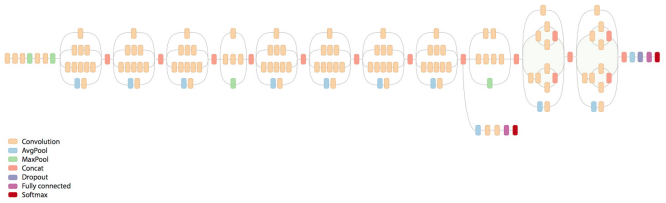

Fig. 1. Inception V3 Model

We retrained the Inception V3 model using the MiniPlaces Dataset for 4000 training steps (2.5 hours approx.) with a learning rate of 0.00001, training batch size of 2000 images per training step, and validation batch size of 1000 images per training step. The dataset was divided into a ratio of 8:1:1, for the training, test, and validation sets. After the last training step we got a training set accuracy of 55.05%, test set accuracy of 50.60% and an average validation set accuracy of 47.81%.

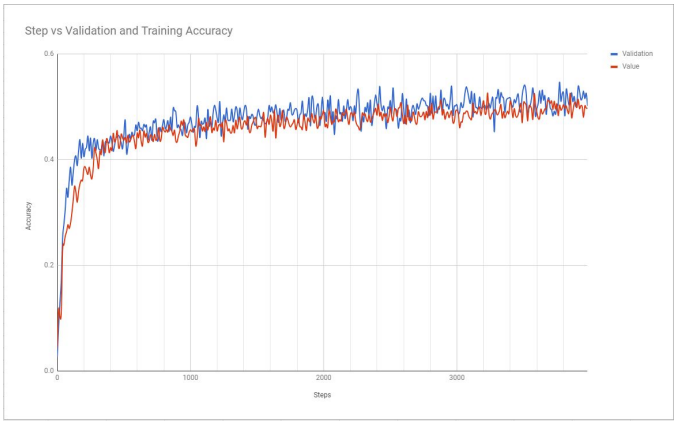

Fig. 2. TensorBoard Accuracy Graph

Accuracy wise our model is better by 8-10%, compared to the Indoor 67 and SUN 397 Trained CNNs when it was tested

CNN	MIT Indoor 67	SUN Dataset
Indoor 67 Trained	43.89	58.33
SUN 397	42.09	67.02

Fig. 3. Results from [1]

on the Indoor 67 dataset [1], but could not top the scores when tested with the SUN 397 Dataset. Our model probably got a higher accuracy than his model because we were using a pretrained model of Inception V3 and that our dataset was neither indoor-centric nor outdoor-centric.

C. Pitfalls

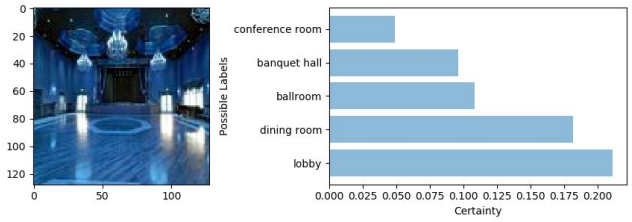

Fig. 4. Misclassification of ballroom

For this image, the correct classification would be ballroom, but the model classified it as a lobby. It probably assumed that the image is a lobby because there are fancy hotel lobbies that have large spaces and hanging chandeliers.

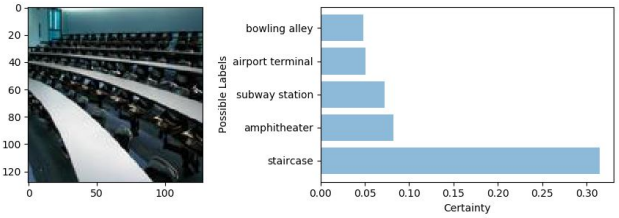

Fig. 5. Misclassification of classroom

The correct classification for this image, is actually a classroom but the model classified it as a staircase. Looking at it from afar, it does seem as though the desks were stairs, and that it had an ascending pattern.

IV. Conclusion

We were able to create a subpar scene recognition engine using transfer learning with the Inception V3 model. There weren't any significant improvements from previous studies, since they used state-of-the-art CNN models and extensively large datasets. But even so, our model shows that it is possible to create a scene recognition engine in less time by using transfer learning and retraining a pre-existing model on new categories, which is better than CNN models built and trained from scratch.

References

[1] B. Ondieki, *Convolutional Neural Networks for Scene Recognition*, Stanford University.

[2] J. Donahue, et.al. , *Decaf: A deep convolutional activation feature for generic visual recognition*, International conference on machine learning, pp. 647-655, 2014.

[3] B. Zhou, et.al., *Learning deep features for scene recognition using places database*, Advances in neural information processing systems, pp. 487-495, 2014.

[4] MiniPlaces Dataset http://6.869.csail.mit.edu/fa17/miniplaces.html