Classification Via Inference: Hello, World of Deep Neural Networks

Lincoln Stein

Abstract—This report discusses the creation of a model to perform multi-class object classification utilizing machine learning. Specifically, two deep neural networks are trained to perform object classification utilizing the Nvidia Digits software framework. These models are trained on a dataset provided as part of the Udacity Robotics Software Nanodegree and a custom dataset. A discussion of each implementation covers problem background information, data acquisition, model selection, training and inference, model results, and possibilities for future work.

1 Introduction

CLASSIFICATION is a foundational task for many applications of robotics and artificial intelligence, from object recognition in images or video to semantic analysis used in advanced spam filters. Traditional techniques used to create a classifiers through human generated models is a difficult and long process. The relatively recent advances in technology that enable implementation of neural networks has created a very attractive alternative. Classifiers created through the training of neural networks achieve astounding accuracy in their tasks, require less human labor, and can be adapted to many different problem sets. This work discusses the process of creating a neural network that can reliably perform classification of objects in images. There have been numerous works previously in this area that follow a similar approach. This work is a documentation of what is effectively the 'Hello, World!' equivalent of neural networks and inferencing. The goal is to create two networks. The first classifies objects passing a camera on a conveyor and could be used as the basis for automated sorting. The second is intended to classify basic dining utensils which could be utilized by a home assistant robot to put kitchenware in its proper storage area or be able to set a table before dinner.

2 Background / Formulation

The first step in creating a neural network to perform any task is to specify the output of the neural network. This work is concerned with classification of images so when given an input image the network should return a single predicted object class. Next is to examine the data we have for training our neural network in order to determine the required network parameters and any preprocessing that needs to be performed. As this work is focused on the overall process of creating a inference based classifier from a trained neural network, the already proven GoogLeNet architecture is used as the basis for both networks. This network architecture was chosen because it was originally designed to perform the task of image classification and has shown great results in this domain. This 22 layer deep convolutional neural network was designed for the ImageNet 2014 challenge. [1] It uses a series of convolutional filters to learn features of the images. Pooling and 1x1 convolutions are used to increase depth and reduce the the dimensionality. The final process uses a softmax to obtain our predicted class. The network was trained with a 0.001 base learning rate using SGD This network architecture was chosen for both datasets because they are simply different formulations of the same problem.

3 Data Acquisition

The conveyor dataset consists was created from 7570 RGB images that were sized to 256x256 pixels captured from a camera positioned over the conveyor. Examples of this dataset can be seen in Fig. 1. The dining utensil dataset was captured with a mobile phone at 1920x1080 resolution. Lighting was varied, combining both natural and fluorescent to represent the varied lighting conditions in which the model is expected to perform. The photos were captured including various combinations of the tabletop and place mats, from a variety of angles to ensure the model is not artificially constrained in learning a robust model of the classes. These images were then cropped to the item of interest for the three classes and the resulting image sized to 256x256 using the Digits dataset creation pipeline resulting in a dataset of approximately 600 images. Examples of this dataset are shown in Fig. 2.

4 Results

The conveyor classifier was trained for 10 epochs, a graph of this is shown in Fig. 3.

The provided evaluation reported a 75.4% accuracy as shown in Fig. 4.

The dining utensil classifier was trained for 200 epochs, its training graph is shown in Fig. 5.

The results from testing this model for each class through the Digits interface are shown in Fig. 6 through Fig. 8. The model exhibits high confusion between knife and fork but correctly identifies the spoon.

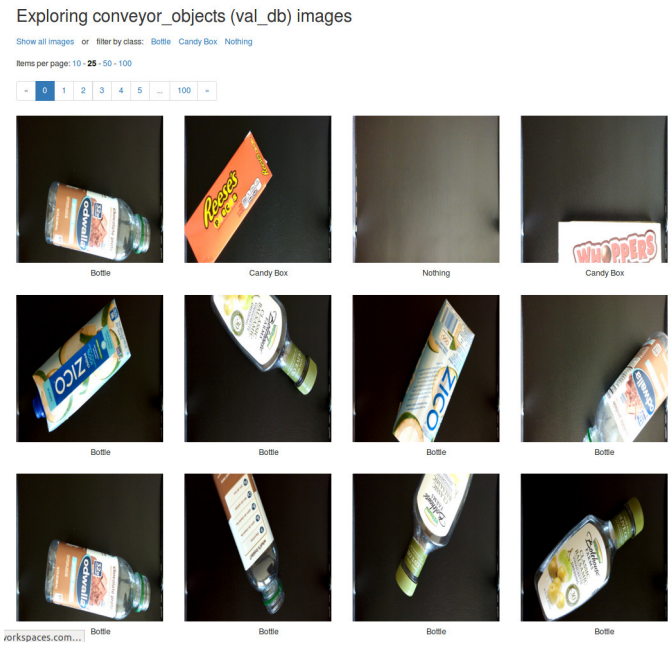

Fig. 1. Conveyor Dataset

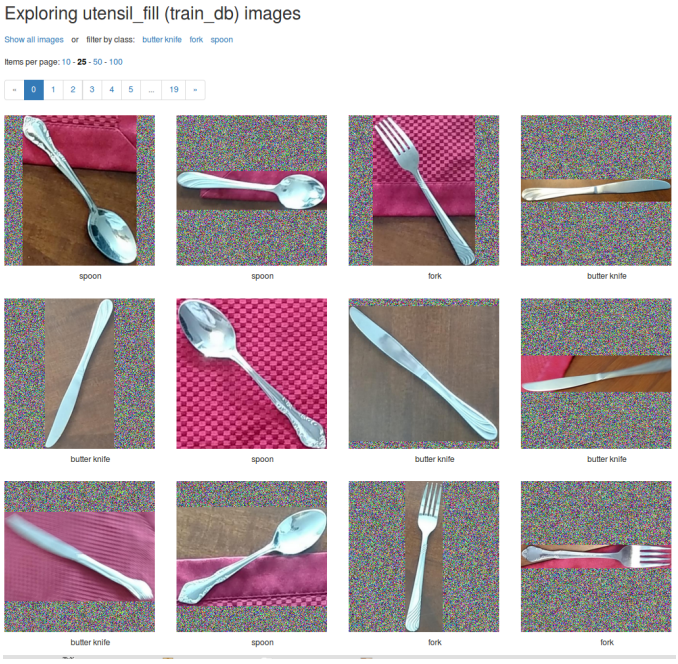

Fig. 2. Utensil Dataset

5 Discussion

The Conveyor classifier exhibits acceptable accuracy of over 75% likely due to the size of the dataset despite its uneven distribution of examples. It completes the inference in about 5 ms as well. The known success of the GoogLeNet architecture provides a proven platform and succeeds in adapting to this particular problem set. The variety of images and relatively stable background enable the network to learn the required features to differentiate the objects effectively.

The utensil classifier exhibits relatively poor results, only successfully classifying one class. While the architecture has been shown to create effective models, there are several factors that may lead to such performance. The biggest factor is likely dataset and subsequent class examples are comparatively small which gives the model less information to learn. The resizing of the images may have distorted the details, making it difficult for the model to differentiate as well. In this application, the accuracy of the inference would be more important than the speed because there is not a need for high speed response; the robot could take an extra second to determine the class and the impact on the application would be rather minimal.

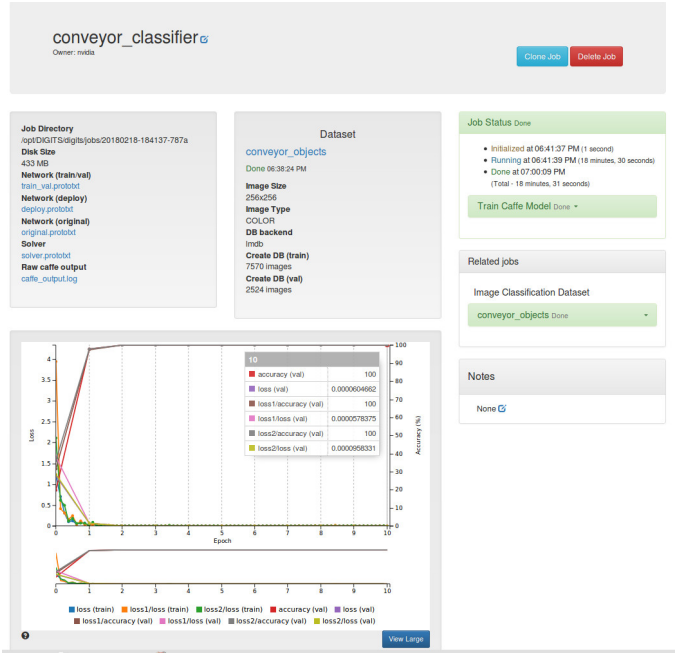

Fig. 3. Conveyor Classifier Training Graph

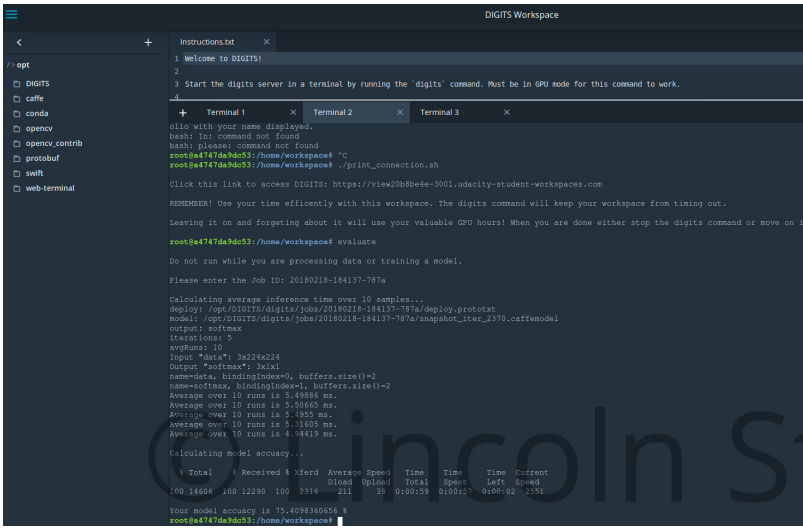

Fig. 4. Conveyor Classifier Evaluation Results

6 Conclusion / Future Work

As a first step into developing robotic inference, I think this project was overall successful. The model for the supplied

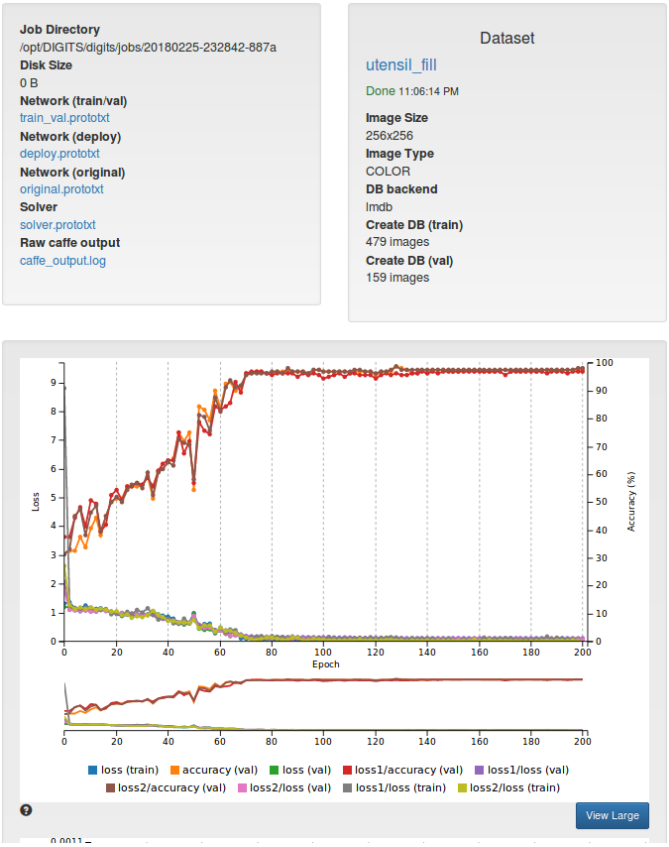

Fig. 5. Utensil Classifier Training Graph

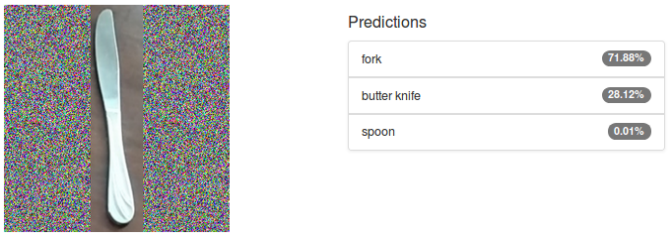

Fig. 6. Butter Knife Class Test Inference

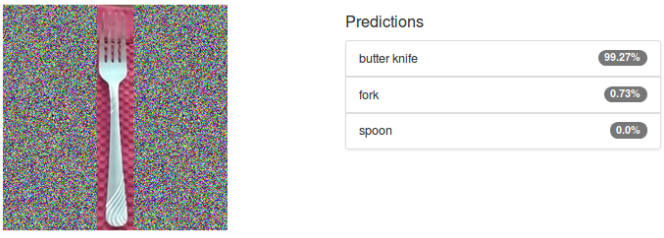

Fig. 7. Fork Class Test Inference

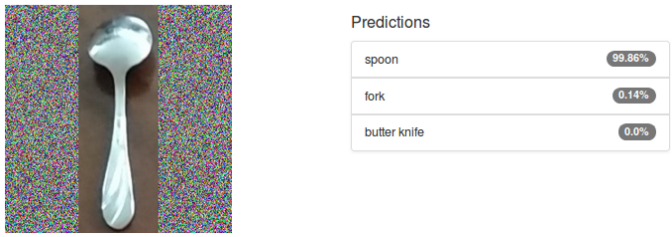

Fig. 8. Spoon Class Test Inference

dataset met specifications for accuracy at 75.4% and speed at 4.9ms while validating the execution of training and model selection. The attempt to create a dining utensil classifier provided a great experience to develop hands on intuition about the extended work flow to create a model. While it is currently not up to the necessary accuracy for implementation,it is a good starting point. Future work is planned to expand the dataset and modify the model parameters in order to achieve over 90% accuracy. This dataset was actually created as a subset of a object detection dataset and will serve as the foundation for creating multi-class DetectNet based model. This is intended to allow much more advanced functionality, such as answering the question:'Is the table set for dinner?'

REFERENCES

[1] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, D. A. Scott Reed, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *IEEE Conference on Computer Vision and Pattern Recognition, 2015*, October 15 2015.