Christopher Rico
MSDS 422
Assignment 7
11/7/19

Introduction

For website providers who seek to automatically label images, automated content recognition for user-submitted content is an immeasurably beneficial application of machine learning techniques. The ability to automatically tag and categorize new user-submitted content can assist in a website's maintenance of community standards, as well as allow for increased searchability and analysis of end-user interests. Furthermore, automated content recognition has the potential to eliminate the need for manual review of user-submitted content, freeing up labor resources. Convolutional neural networks (CNN), considered to be the state of the art in image recognition, are a cutting-edge machine learning approach modeled after biological brain architecture that allow for highly abstract comprehension of image contents. CNN performance, however, is highly dependent on the architecture and training hyperparameters of the model. In furtherance of exploring the most accurate model configuration for image recognition agnostic of the training time, several CNN modeling topologies were attempted on the cats and dogs image set.

Research Design

The cats and dogs dataset is a well-known image dataset that consists of 37,500 color images of dogs or cats in multiple different angles, levels of zoom, and background settings. The machine learning competition Kaggle has broken up these 37,500 images into a 25,000 image labeled training set and a 12,500 image test set, which were the training and test sets used to assess the CNN approaches. For this research, models will be optimized agnostic of training time: only model accuracy is relevant. As Kaggle reports model performance based on log loss, this will be the metric used to compare models to one another.




Christopher Rico
MSDS 422
Assignment 7
11/7/19

Programming Work

Model construction and evaluation was completed entirely in the cloud, using Google Colaboratory as a development environment and TensorFlow as a CNN modeling environment. TensorFlow is an industry standard neural net library built by Google, with the capacity to model convolutional neural nets.

To begin, four different CNN models were instantiaed. The configuration combinations tested were combinations of 3 or 4 convolutional layers, and a dropout rate of 0.5 or 0. Each CNN was trained on the 25,000 image training set. Then, training set accuracy and loss were reported, and probability predictions on the 12,500 image test set were output to CSV files for submission and evaluation on Kaggle.com.

Results and Recommendation

Both training and test set loss results are fairly consistent across CNN topologies. Training time increases modestly as the complexity of the model increases, which is to be expected. The CNN with 3 convolutional layers layers and a dropout rate of 0.5 yields a test set log loss of 1.2187, which is the best among the configurations tested.

It is recommended to use a CNN with 3 convolutional layers and a layer of dropout at a rate of 0.5 for image recognition applications. That said, it would be prudent to explore further model configurations and methods of training set augmentations, as there may be a more optimal configuration not examined.

Convolutional Layers	Dropout	Processing Time	Training Set Loss	Test Set Loss
3	Yes	1042.92	0.51773	1.21871
3	No	1037.19	0.49702	1.35034
4	Yes	1060.15	0.51461	1.43362
4	No	1071.95	0.49070	1.46427

A shared version of the interactive Jupyter notebook used to run this analysis can be found at:
https://colab.research.google.com/drive/1VTAz28jI21Z6bVAbAB35dg6ASLiSl9-b

Kaggle profile with test scores can be found at:
https://www.kaggle.com/christophrico