Machine Learning Engineer Nanodegree

Capstone Proposal

Gap Kim
June 6, 2018

Proposal

Domain Background

Over the last years, deep learning methods have been shown to outperform previous state-of-the-art machine learning techniques in many fields such as visual, audio, medical, social, and sensor. In particular, object recognition has gained tremendous interest by engineers and scientists in artificial intelligence and computer vision. Deep learning allows computational models of multiple processing layers to learn and represent data with multiple levels of abstraction mimicking how the brain perceives and understands multimodal information, thus implicitly capturing intricate structures of large-scale data [1]. Among various methods developed in object recognition, Convolutional Neural Network (CNN) is of interest for this project. CNNs were inspired by the structure of a visual system and were shown to significantly outperform traditional machine learning approaches in computer vision and pattern recognition [2]. CNNs have been used in variety of fields, which includes but are not limited to object detection [3], face recognition [4], and action/activity recognition [5].

In this study, a deep learning model based on CNNs is proposed for Google Landmark Recognition Challenge from Kaggle [6]. A technology that can accurately predict landmark labels directly from image pixels can broadly benefit applications in various areas such as photo management, maps, aviation, and satellite images.

Problem Statement

One of great obstacles to landmark recognition research is the lack of large annotated datasets. Hence Google Landmark Recognition Competition presents the largest worldwide dataset to date and challenges to build models that recognize the correct landmarks from the test images. The Landmark recognition presents a dataset with a very large numbers of classes (15,000 classes), but the number of training examples per class may not be very large. This makes the problem different from image classification challenges like the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) where the aim is to recognize 1000 general object categories.

Datasets and Inputs

The original Landmark Recognition Challenge dataset provides two files, train.csv and test.csv. The training set images each depict exactly one landmark. Test images may depict no landmark, one landmark, or more than one landmark. Each image has a unique id (a hash) and each landmark has a unique id (an integer). Due to restrictions on distributing the actual image files, the dataset contains a url for each image.

The original number of images for training and test dataset are 1,225,029 and 117,703, respectively with a total of 14,951 unique landmark ids. Use of full dataset may require large storage capacity (in the order of hundred GB) and computational power not easily available to an individual. Therefore a subset of dataset will be used for this study. First, top 100 landmark ids most frequently appearing are identified among 14,951 landmark ids. Then, 2% of images from each of the 100 landmark ids are downloaded. The procedure is similar to stratified sampling applied to the top 100 frequent classes to preserve the ratio of images among the classes.

Approximately 8100 images will be eventually used for the reduced dataset. This dataset is split into training, validation, and test dataset for building the CNN models.

Solution Statement

The CNN model will be built using the framework of TensorFlow with Keras library. The CNN will employ three main types of neural layers: (i) convolutional layers, (ii) pooling layers, and (iii) fully connected layers. The convolutional layer utilizes various filters to generate feature maps. The pooling layer reduces the spatial dimensions of the input volume and helps to alleviate overfitting problems. The fully connected layer eventually converts 2D feature maps into 1D feature vector and is forwarded into total number of landmark ids for classification. Regularization techniques such as dropout, batch normalization [7], and data augmentation are considered for developing an optimal CNN model for the problem. Accuracy will be used as the model performance evaluation metric.

Benchmark Model

The size of the dataset is approximately 8100 images with 100 landmark ids. However, the frequency of the 100 landmarks appearing in the dataset varies significantly. In other words, the expected accuracy may vary depending on the selected test dataset. Thus the accuracy of the proposed benchmark model, given the test dataset, can be calculated as the following:

(i) Given the test dataset and 100 unique landmark ids, the probability of correctly classifying a landmark id, $P(id)$ can be calculated as,

$$P(id) = n_{id}/N$$

where n_{id} is the number of the landmark id in the test dataset, and and N is the total number of images in the test dataset. For example, if landmark id = 99 appears 50 times in test dataset size of 800, $P(99) = 50/800$.

(ii) The expected value of correct number of landmark id classified is:

$$E(x) = \sum_{x=0}^{x=N} (x \cdot P(id(x)))$$

(iii) Finally, the expected accuracy for correctly classifying the landmark id is:

$$E(accuracy) = \frac{E(x)}{N}$$

Given the test dataset, expected accuracy can be computed, which will be used as the benchmark model. Since there are 100 unique landmark ids, expected accuracy by random guessing will be very low.

Evaluation Metrics

The overall evaluation metric that will be used for the benchmark model and the solution model is accuracy.

$$\text{accuracy} = \frac{\text{Number of correctly predicted class}}{\text{Total number of predictions}}$$

Project Design

The project will take on a subset of dataset from Google Landmark Recognition Challenge. The competition challenges to build models that can recognize the correct landmark where the dataset contains a very large number of classes (~15,000 classes) but the number of training examples per class may not be very large.

Due to computational limitation, a reduced dataset is proposed with 100 unique landmark ids in approximately 8,100 images. Stratified sampling is employed so that the ratios of images among the classes are preserved. A deep learning model utilizing the CNN will be built using the framework of TensorFlow with Keras library. The CNN model will employ three basic neural layers, convolutional, pooling and fully connected layers. To improve the accuracy and efficiency, regularization techniques such as dropout, batch normalization, and data augmentation will be considered when building the optimal CNN for the challenge. The batch normalization technique, which transforms mean activation close to 0 and the activation standard deviation close to 1, has shown to improve accuracy with fewer training steps in image recognition problems. Many times, landmark images may be rotated, take up only a portion of the whole image, and may be out of center. The data augmentation technique may help improve accuracy for such images. In addition, data augmentation strategy can help with those landmark ids with only small number of training images. The performance of CNN models will be compared with the accuracy computed for the benchmark model.