Capstone Proposal

Jianxun Gao
November 1st, 2017

Domain Background

Recognizing the sequence of numbers from a image is a challenging job, even harder with nature image due to visual artifacts, such as distortion, occlusion, directional blur, cluttered background or different viewpoints. Since recent years the deep learning had made significant progress[LeNet][AlexNet][VGG][GoogLeNet]. So the path to the solution is a bit more clearer.

Moreover, deep learning is a fast-growing field of AI focused on using neural networks for complex practical problems. Deep neural networks are used nowadays for object recognition and image analysis, for various modules of self-driving cars, for chatbots and natural language understanding problems.

In this project, I am going to train a classifier to recognize a sequence of the digits by building a deep convolutional neural networks (CNNs).

Problem Statement

The goal of this project is to build a model that provide a solution for recognizing the sequences of digits in a image. The proposed the model is going to train on the SVHN dataset (http://ufldl.stanford.edu/housenumbers) collected from house numbers in Google Street View.

The tasks will be:

1. download the dataset
2. preprocessing/analysis the dataset
3. train the different CNNs architecture
   LeNet-5
   AlexNet

VGG

Inception

4. compare the results

Datasets and Inputs

The dataset being considered for the project should be the Street View House Numbers (SVHN) Dataset. SVHN is a real-world image dataset public available for developing machine learning and object recognition algorithms.

The dataset overview:

- 10 classes, 1 for each digit. Digit `'1'` has label 1, `'9'` has label 9 and `'0'` has label 10.
- 73257 digits for training, 26032 digits for testing, and 531131 additional, somewhat less difficult samples, to use as extra training data

In total, the dataset comprises over 600,000 labeled characters, and has been made available in two formats:

- Full Numbers - the original, variable-resolution, color house-number images as they appeared in the image file. Each image includes a transcription of the detected digits as well as character level bounding boxes.
- Cropped Digits - character level ground truth - in this MNIST-like format all digits have been resized to a fixed resolution of 32-by-32 pixels. The original character bounding boxes are extended in the appropriate dimension to become square windows, so that resizing them to 32-by-32 pixels does not introduce aspect ratio distortions.

For the project, I am going to using Cropped Digits as input to train a convolutional neural networks classifier to recognize a sequence of digits up to 5 of them.

Solution Statement

The past few years of computer vision research has been on study to form effective convolutional neural networks[Inception][ResNet]. Why convolutional neural network structure are so useful and it works so well in computer vision, there are two main advantages of convolutional layers over just using fully connected layers, which are parameter sharing and sparsity of connections. Because of these two mechanisms, a neural network has a lot fewer parameters, which allows it to be trained with smaller training sets, and it's less prone to be over-fitting.

And a convolutional structure helps the neural network encode the fact that an image shifted a few pixels should result in pretty similar features, and should probably be assigned the same output label. I am trying to using the basic building blocks such as convolutional layers, pooling layers, and fully connected layers of components to form effective convolutional neural networks.

Benchmark Model

According the paper[2], the accuracy of human performance on this dataset as 98.0%. Yuval Netzer et al. on the paper[2] achieved 90.6%. Goodfellow et al. on the paper[3] achieved 97.84% accuracy. The proposed model in the project should not be lower than 90% and set up LeNet-5 or AlexNet as baseline mode and to see what can be improved from there.

Evaluation Metrics

Accuracy is commonly used as a statistical measure of how well a binary classification test correctly identifies or excludes a condition. That is, the accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined.

Accuracy = (TP + TN) / (TP + TN + FP + FN)

The proposed model is going to using this accuracy metic as evaluation purpose to test against SVHN-test dataset.

Project Design

In this project, the theoretical workflow for approaching a solution given the problem should be

1. analyze and preprocess the dataset: splitting to train, validation, and test.

2. define the loss function: categorical cross-entropy loss function.

3. define the optimization algorithm: mini-batch gradient descent with momentum or mini-batch with Adam mode.

4. building neural networks: using the basic building blocks such as convolutional layers, pooling layers, and fully connected layers of components.

5. train the network.

6. collect the results and analyze the results.

References
[1] Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). ImageNet classification with deep convolutional neural networks.
[2] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, Andrew Y. Ng, Reading Digits in Natural Images with Unsupervised Feature Learning
[3] Ian J. Goodfellow, Yaroslav Bulatov, Julian Ibarz, Sacha Arnoud, Vinay Shet. Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks