Machine Learning Engineer Nanodegree

Capstone Project Proposal

Derrick Chang
March 1st, 2017

Domain background

Human brain is pretty good at gathering high-level information from images. However, these tasks are often challenging for machines. Computer vision is an interdisciplinary field seeks to automate tasks that the human visual system can do. One of these interesting tasks is to recognize digits in natural scenes.

The Street View House Numbers (SVHN) Dataset [1] is a collection of images contain house numbers in natural scenes created by Google. For Google, being able to read house numbers from images means improving map quality since every house number picture is geotagged. In this project, I will use deep learning approach based on Goodfellow [2] to recognize house numbers in the SVHN Dataset.

The reason I chose this project topic is because I want to learn more about deep learning, and image recognition is one of the fields which deep learning performs very well. My goal is to build an end-to-end system for house numbers recognition, so that I can have a concrete hand-on experience for deep learning.

Problem statement

This problem is a supervised classification task. Given an image which contains a house number, the model should output a sequence of digits as a prediction. This problem is similar but more difficult than classifying digits for the MNIST dataset [3], because of the following two reasons:

1. The SVHN dataset contains images from natural scene, where perspective, lighting conditions, and other objects can cause distraction.

2. For a house number prediction to be considered as correct, all the digits in the house number should match the target label.

Generally speaking, house numbers range from 1 to 99999, and most of the house numbers in the SVHN dataset have 2-4 digits. Consider the size of our dataset and the distribution of house numbers' lengths, it would be impractical (way too many) if we define 99999 classes for this problem.

Datasets and inputs

The SVHN Data can be downloaded from: http://ufldl.stanford.edu/housenumbers/
The dataset comes in two formats:

Format 1: Original Images (*.png) with character level bounding boxes (Figure 1)

- Train Data: 33401
- Test Data: 13068
- Extra Data: 202353

For each set of images, there is a 'digitStruct.m' file stores metadata for every single digit in an image.

- Bounding boxes information: [Top, Left, Width, High]
- Digit class: [1-10], one for each digit. (Digit '1' has label 1, '9' has label 9, and '0' has label 10).

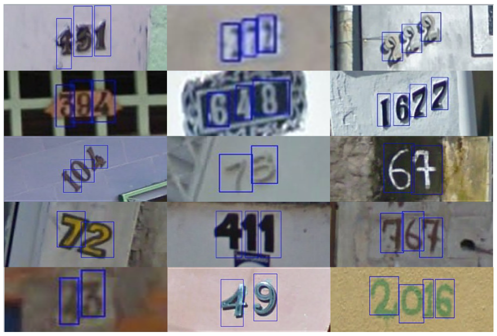

Figure 1: Examples of images from the SVHN dataset (format 1)

Format 2: MNIST-like 32x32 images centered around a single character (Figure 2)

- Train Data: 32x32x3x73257 Train Label: 73257x1
- Test Data: 32x32x3x26032 Test Label: 26032x1
- Extra Data: 32x32x3x531131 Extra Label: 531131x1

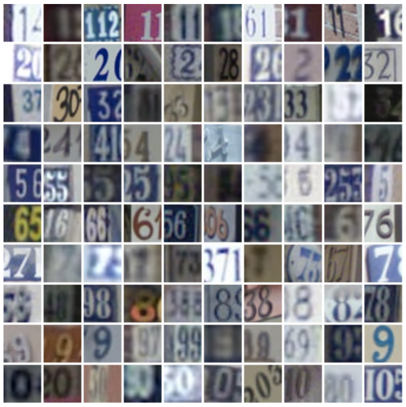

Figure 2: Examples of images from the SVHN dataset (format 2)

In this project, only the Training Data and Testing Data from the original images (Format 1) will be used.

Solution statement

In this project, I will be assuming that house numbers are 1-5 digits long and defining 11 classes for each digit. Class [0-9] represents digit value [0-9], and class 10 represents N/A. The final model will take an image which contains a house number with 1-5 digits long, and output a sequence of 1-5 digits as a prediction. I will be building this model using deep convolutional neural networks based on Goodfellow's approach [2].

For this capstone project, the deliverables that I will be producing are:

1. Source Code
2. Project Report (this document)

Benchmark model

The benchmark that I will be using is the performance presented by Goodfellow et al. They reported an accuracy of 96% for multi-digit classification on the SVHN dataset. Also, human have 98% accuracy on this dataset.

Evaluation metrics

For a house number prediction to be considered as useful, it needs to be completely correct. No partial credits should be given if any of the digits in a house number is wrong. For example, a house number '3589' may be a few streets away from '9589' even though there is only a single digit mismatch.

As a result, a good metric for this project is defined as:

$$accuracy = \frac{\text{total instances of correctly predicted house number}}{\text{total house number}} * 100\%$$

Notice that the house number here means the entire house number sequence instead of a single digit.

For example, if we have 3 image samples with their labels and predictions shown as follows:

	Image#1	**Image#2**	**Image#3**
Labels	1234	1998	151
Predicts	1234	1999	51
Correctness	v	x	x

The accuracy is: 1/3 x 100%= ~33.33%

Project design

The outline of tasks is as follow:

1. Collect SVHN data and parse metadata (bounding boxes and labels)
2. Preliminary analysis of data
3. Preprocess data and split the data into training, validation and test set
4. Build the CNN model
5. Fine tune the CNN model
6. Test and evaluation