Machine Learning Engineer Nanodegree

Capstone Proposal

Devinder Saini
December 2, 2016

Domain Background

Multi character and multi digit recognition is a computer vision problem which is more common than recognizing single digits or characters. It inherently involves multi step detection and data representation challenges. The general field of Optical Character Recognition has a lot of motivations like cheque processing, package routing in post offices etc. It will be even more important with upcoming trends like self driving cars, delivery drones, augmented reality applications etc.
Street View House Numbers Dataset is one such corpus of real world images of house numbers extracted from Google Street View images. The SVHN Dataset solution has been used for automatic geotagging of house numbers in Google Street View. A lot of research has been done on the subject, and a lot of algorithms have been proposed [1] [2][3], though most of them deal with single cropped digit recognition.

Problem Statement

The problem of identifying multi digit sequences consists of two parts - identifying each digit correctly, and identifying all digits. If X represents the input image, Y represents the output sequence and $y_1\ y_2\ y_3\ \cdots\ y_n$ are the individual digits, we need to maximize the probability that the predicted sequence Y' matches Y

$$P(Y'=Y|X)$$

The problem can be divided as first identifying the number of digits in the scene and then identifying each digit correctly.

Datasets and Inputs

The dataset to be used in the project will be Street View House Numbers Dataset. It is a real world dataset take from Google Street View images. The dataset has variable resolution and all kinds of colors, lighting, orientation and fonts. The dataset has two formats, full house numbers with all digits and 32x32 pixel cropped digits. Each format has separate training and test sets, and another extra set of somewhat less difficult samples. The dataset is not balanced, there are a lot more samples with two and three digit numbers than the rest.

In our setup, we will not use the cropped digit format. We also won't be using the extra set because of huge computational requirements. We'll use most of training set for training and a fraction of it for validation, and the test set for evaluation. Each image in the full numbers set consists of a targeted group of numbers varying in length from 1 to 6 digits. To use this dataset in our setup, we'll flatten the data so that each sample point represents an individual digit. The training set consists of 33,402 samples. By flattening it we obtain 73,292 samples.

We'll also augment the data by adding an additional feature *i,* index of the digit in the sequence. The input images will be resized to a standard size of (32,77) using bilinear filtering. It is small enough to be processed by convolutional networks with reasonable speed and is a good representative of the general aspect ratio of images in the dataset. Our inputs to the model during training will be image X and index i, while the outputs will be count n and label y.

Solution Statement

If any individual digit, or the length of predicted sequence is wrong, then a completely different number will be obtained and will be wrong. So we can express the probability of detecting a sequence correctly as:

$$P(Y'=Y|X) = P(n'=n|X). \prod_{i=1}^{n} P(y'_i=y_i|X)$$

During training phase we need to maximize the log likelihood that given *X* and true sequence *Y, Y'* is equal to the true sequence *Y.*
We will divide the problem into two parts, predicting the number of digits in the image and identifying each digit. When we've identified the number of digits in the image, we will loop through that range and provide the index as an additional input to the label detector. To train the counter, we'll use images as input and counts as output. To train the label detector, we'll use images and indices as inputs and labels as output. The counter and label detector will use a shared convolutional neural network and use separate fully connected layers after that.

Benchmark Model

We will evaluate the performance of our model with model proposed by Goodfellow, Bulatov, Ibarz and Shet [4]. They use the same concept of counter and sequential label detection that we use here, but they use a series of cropping and expansion of input images to augment the dataset. The benchmark model achieves an accuracy of 96.03% on SVHN dataset. Another multi digit sequence detection is model proposed by Guo, Tu, Lei and Li [5] uses a combination of Convolutional Neural Networks and Hidden Markov Model with embedded Viterbi training. It achieves an accuracy of 81%.

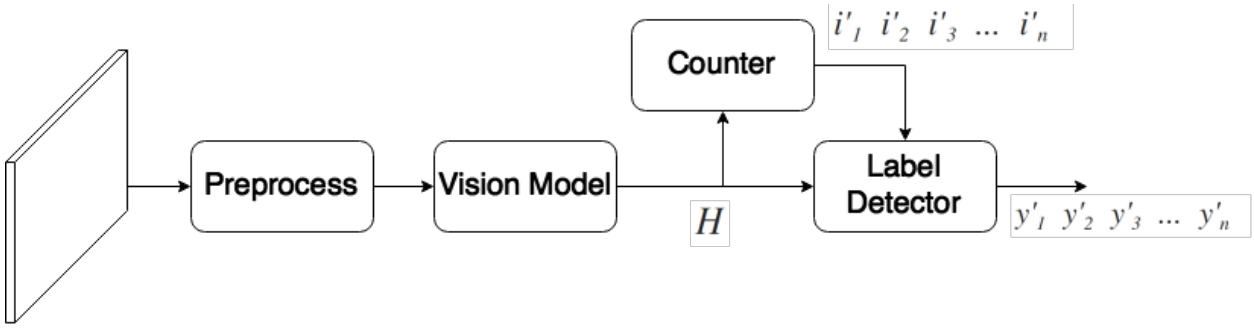
Evaluation Metrics

To evaluate our model and to compare with the benchmark, we'll use accuracy as evaluation metric. We will only consider a prediction accurate if the predicted sequence on numbers completely matches the true sequence.

Project Design

Our design will first preprocess each image to a standard 32 by 77 pixels. We'll create a shared vision model that will feed convoluted images to a counter model and a label detector. This vision model will have a series of convolutional, dropout and pooling layers, and a fully connected layer in the end. This shared model approach will process each image with convolution filters only once and produce a dense tensor which can be in-

gested by both counter and label detector. The label detector will take an additional input *i,* the index of the digit to be detected.



The idea of having a shared vision model is to process the input image only once for both digit counting and label detection. The vision model will produce a convolution tensor that can be further processed by both counter and label detection models. Counter will generate a single output *n ,* the number of digits in the image. We can then pass indices 1 to *n*, with the convoluted tensor *H* to label detector, which will generate the indexed sequence.

References

[1] K. Wang, B. Babenko, and S. Belongie, “End-to-end scene text recognition,” in Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011, pp. 1457–1464.

[2] Alsharif, O. and Pineau, J. (2013). End-to-end text recognition with hybrid hmm maxout models. Technical report, arXiv:1310.1811.

[3] L. Neumann and J. Matas, “A method for text localization and recognition in real- world images,” in Computer Vision–ACCV 2010. Springer, 2010, pp. 770–783.

[4] J. Goodfellow, Y. Bulatov, J. Ibarz, S. Arnoud, V. Shet, “Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks”, arXiv:1312.6082v3.

[5] Q. Guo, Dan Tu, Jun Lei, Guohui Li, “Hybrid CNN-HMM Model for Street View House Number Recognition”, in Computer Vision-ACCV 2014, DOI 10.1007/978-3-319-16628-5_22