DL

ZHIHAN WANG
University of Southampton
School of Electronic and Computer Science
Email: zw3u18@soton.ac.uk

ALEX
University of Southampton
School of Electronic and Computer Science
Email: zw3u18@soton.ac.uk

SHAUNAK
University of Southampton
School of Electronic and Computer Science
Email: zw3u18@soton.ac.uk

***Abstract*—Abstract goes here.The paper we are going to reproduce is [2]**

II. part III

I. Introduction

Optical character recognition has been well studied on constrained domains, such as document processing, but is still challenging in unconstrained domains, such as natural photographs. In the paper [2], an equally hard sub-problem, arbitrary multi-character text recognition in photos captured at street level, has been discussed.

The paper employed DistBelief, a software framework that can utilize computing clusters with thousands of machines to train large models[1] to implement large-scale deep neural networks on publicly available Street View House Numbers(SVHN) dataset and finally achieved over 96% accuracy in recognizing street numbers, and 97.84% accuracy on per-digit recognition tasks. After then, this trained model was implemented to solve CAPTCHA puzzles where text is deliberatly distorted and used to distinguish humans and robots, and achieved a 99.8% accuracy.

[2] contributes a lot and the results even reached human level performance at specific thresholds. The detailed information goes below.

A. Architecture

There are three steps in traditional approaches to recognize multi-digit numbers from photos, localization, segmentation, and recognition. The paper [2] proposed a unified model to integrate these three steps via the use of a deep convolutional neural network that operates directly on the image pixels and achieved an end-to-end prediction.

1) Basic methods: The task of street number recognition is that given an image, the numbers in the image should be identified. The basic method used here is to train a probabilistic model of a predicted sequence output given an image. Let $\mathbf{S}$ represent the output sequence and X represent the input image. The goal is to learn a model of $P(\mathbf{S}|X)$ by maximizing $\log P(\mathbf{S}|X)$ on the training dataset. The probability of a specific sequences **s=**$s_1, s_2, ..., s_n$ is given as below, in which n is the number of digits in the image.

$$P(\mathbf{S} = \mathbf{s}|X) = P(L = n|X) \prod_{i=1}^{n} P(S_i = s_i|X)$$

At prediction time,

$$\mathbf{s} = (l, s_1, s_2, ..., s_l) = \underset{L,S_1,S_2,...,S_L}{\arg\max} \log P(S|X)$$

2) CNN structure: The best model trained on the SVHM dataset in [2] is with 11 hidden layers, consisting of 8 convolutional layers, 1 locally connected hidden layers and 2 densly connected hidden layers. All connections are feedforward and there are not skipped layers. The first hidden layer contains maxout units[3] and each unit is with three filters while other layers contain ReLU. Each convolutional layer includes max pooling with window size 2×2 and subtractive normalizaiton with window size 3×3. The stride at each layer alternates between 1 and 2; therefore, zero padding is used to preserve representation size. The size of all the kernels is in 5×5. As significant overfitting can be seen, dropout applied to all hidden layers.

B. Performance

[2] shows that the performance of model increases with the depth of the convolutional network. Two experiments were used to prove this conclusion. The first one confirmed that depth is necessary for good performance and the second one with a accuracy graph demonstrated that smaller models even with more parameters cannot reach the same level of the performance as deeper models.

C. Comparison with previous work

Images recognition networks trained in the previously published papers generally have 2 to 4 convolutional layers followed by 1 or 2 densely connected layers and classification layers. But the model in [2] used more convolutional layers as referred above. This is because earlier layers are used to solve localizaiton and segmentation firstly, and then the results are processed to later layers to recognize. This model achieves an end-to-end prediction.

D.

II. Evaluation

III. Limited knowledge

IV. Conclusion

Acknowledgment

This paper is supported by University of Southampton. I would like to thank Dr. for his guidance and patience during the entire work.

References

[1] Dean, Jeffrey, et al. "Large scale distributed deep networks." Advances in neural information processing systems. 2012.

[2] Goodfellow, Ian J., et al. "Multi-digit number recognition from street view imagery using deep convolutional neural networks." arXiv preprint arXiv:1312.6082 (2013).

[3] Goodfellow, Ian J., et al. "Maxout networks." arXiv preprint arXiv:1302.4389 (2013).