Understanding the Literature: Gradient-Based Learning Applied to Document Recognition

Prepared by Yang Liu (13174420)

1. Introduction:

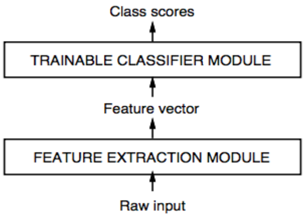
The purpose of this report is to understand and critique the paper ‘Gradient-Based Learning Applied to Document Recognition’, written by Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. This paper was published by the proceedings of the IEEE in Nov 1998. It introduced convolutional neural networks (CNN) for handwritten digit recognition and compared with some other traditional techniques. In terms of the whole document recognition system, the graph transformer network (GTN) was then introduced, where CNN was integrated as a module. The report will be divided into 8 sections: introduction, content, innovation, technical quality, application and X-factor, presentation and references.

2. Content:

2.1 Background

In the past few years, machine learning techniques especially neural networks have played a crucial role in the design of recognition systems. Building effective and efficient learning techniques can be served as an important factor in the success of pattern recognition applications. The pattern recognition process could be divided as two modules: feature extraction module and trainable module (Figure 1), where the performance of the second part is heavily determined by the accuracy of feature extraction. Traditional techniques applied hand-crafted algorithms to extract features via heuristics, while due to the variability and richness of raw input, it’s almost impossible and costly to achieve accurate information retrieving. Furthermore, the handcrafted methods are usually task-specific, which limits the generalization ability for new problems.

Figure 1:



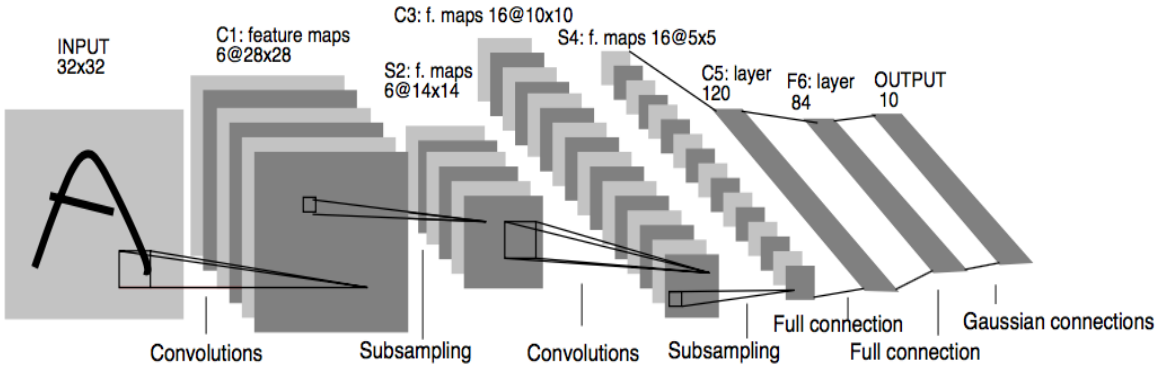
2.2 Main Point

The main point of this paper is that developing automatic recognition systems would be more promising than relying on hand-crafted heuristics. More specifically, the authors pointed out that convolutional neural networks (CNN) could be used to extract features automatically, that the requirement for hand-crafting could be eliminated, and compared with traditional methods to validate their argument. Besides that, instead of manually integrating independently designed modules, they introduced Graph Transformer Networks, a unified design paradigm, to enable joint training of all modules for a global learning objective.

2.3 Convolutional Neural Networks and LeNet-5

The first contribution of this paper is the application of convolutional neural networks for character recognition. Before introducing CNNs, the authors first illustrated the weakness of using fully connected neural networks (FCs). Although the features can also be extracted by FCs, the high dimensional image inputs makes it computational expensive to train. And the topology of images will be ignored to fit the input shape of FC networks, while in the images the pixels are spatially correlated, that capturing these local correlations may be helpful for recognition. CNN is then introduced to handle these problems via three crucial architectural ideas: 1) local receptive fields 2) shared weights and 3) spatial or temporal sub-sampling. The idea of local receptive fields is based on the principle of image local correlation such as shift, scale and distortion invariance. By using it elementary visual features, such as oriented edges, end-points and corners can be extracted as feature map for subsequent layers. The weights for a convolutional kernel are shared to perform the same operation on different parts of an image, thus the spatial invariance can also be kept. Generally a convolutional layer consists of several kernels, that at each location multiple features can be extracted. After extracting the feature maps, before feeding into subsequent layers, sub-sampling would be performed to reduce not only the sensitivity to shifts and distortions, but also the resolution. This operation, to some extent, is a kind of dimension reduction method which can make the extraction process more efficient.

Figure 2:



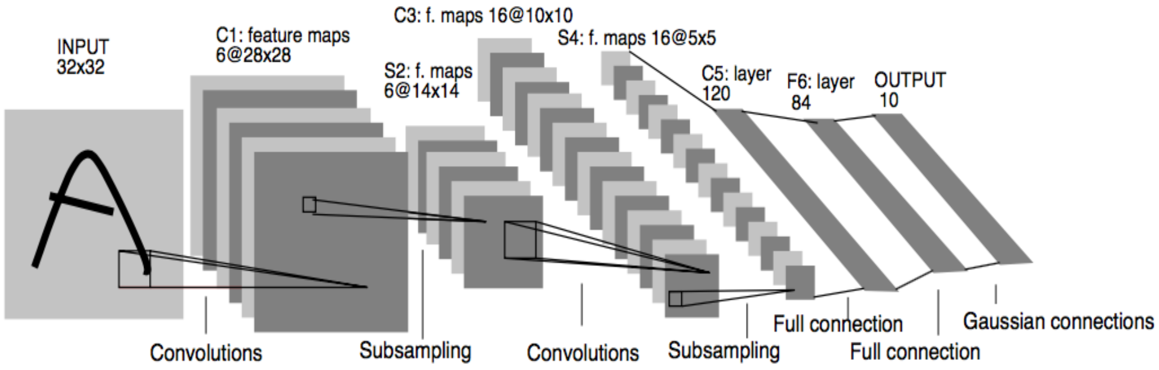


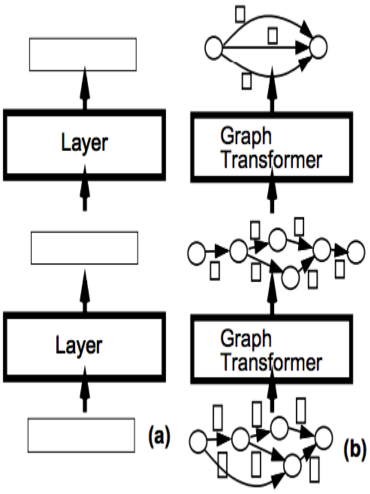
Figure 2 shows the proposed architecture, LeNet-5, which is a typical convolutional network for feature extraction. After preprocessing, the 32*32 input is processed by 6 5*5 kernels (C1) to get 6 28*28 feature map, then 2*2 average pooling operation (S2) is applied as sub-sampling, and the result becomes 6 14*14 feature maps. Similar sequential operation (C3 -> S4 -> C5) is performed again and the output is then flattened to feed into classical FC layer. Before feeding into final output layer, tanh is used as squashing function. Finally, the output layer is composed of Euclidean Radial Basis Function units (RBF), to predict which class should be assigned to this input.

In order to validate their method, this paper introduced MNIST dataset, where the task was to predict the true class of hand-writing digits. Several versions of LeNet-5 were compared first, and then LeNet-5 was compared with traditional classifiers such as linear classifier and nearest neighbour classifier. The results shown that LeNet-5 achieved promising low error rate and comparable memory requirement, and its noisy resistance ability was rather robust as well.

2.4 Graph Transformer Networks

Besides LeNet-5, the authors also showed how it could be used for building an integrated document recognition system. For large and complex systems it’s essential to build it out of multiple smaller modules. Compared with traditional framework, which applies fixed-size vectors to represent the data, the proposed design framework, graph transformer networks (GTN), represents the states as graphs or mixtures of probability distributions over structured collections of vectors, as shown in Figure 3. In this way the flexibility of application can be improved. As gradient based learning procedures can be generalized to that graph form, different modules can be trained jointly for one global learning objective. In this paper, the authors proposed two application cases to validate the effectiveness of GTN: an online-handwriting recognition system and a bank check reading system. In the first one, the discriminative Viterbi training GTN framework is used on both heuristic over-segmentation method and space displacement neural network (SDNN LeNet-5).The result shows that both methods have lower error rate by applying global training, while SDNN achieves much better performance than HOS. Similar experiment result can be seen on the second system.

Figure 3:



3. Innovation:

They pointed out the main problem faced and compared many approaches. LenCun et al also did many experiments to evident the results. Although the results did not achieve a breakthrough due to the technique limitation at that time. The ideas of this paper also give further researches guidance.

This paper designed a convolutional neural network architecture, LeNet-5, for handwritten character recognition. Compared it with traditional methods that manually extracted the features, the CNN-based method achieved better accuracy while the memory requirement was comparable. Although LeNet-5 is not the first work to propose CNN or apply NN for image recognition (earlier work includes fully-connected NNs, LeNet-1, LeNet-4 et al.), in this work the author systematically provided the main advantages of convolutional neural networks for extracting features from high dimensional inputs. From current point of view, the design of LeNet-5 is important for that it sheds light on the development of deep learning these years. Inspired by this work, a lot of CNN variants have been proposed in different areas such as AlexNet, VGG and ResNet for recognition, RCNNs and YOLOs for object detection. Furthermore, CNN-based feature extraction has become an essential backbone for deep learning related methods and applications including but not limited to computer vision.

In the perspective of application, this paper proposed GTN design paradigm, which integrated CNNs with other modules to be a complete document recognition system. By using gradient-based learning, this paradigm enables joint training with global learning objective, which achieves better performance than system composed of independently trained modules.

Another contribution could be the MNIST dataset used in the experiment. Till now it’s still a classical dataset to validate the effectiveness of neural network related methods.

In terms of reference, this paper has very profound research significance, with over 20000 citations.

4. Technical quality:

This paper is of high technical quality. The architecture of LeNet-5 and GTN cases are very clear that the implementation details of each layer (for LeNet) and each module are illustrated, which would be helpful for the readers to understand the mechanism, reproduce the result and modify for new applications. The authors performed thorough experiments to validate the effectiveness of the LeNet and GTN design paradigm. The experiments includes not only the error rate, which is the core measurement for recognition, but also the resource requirement and noise resistance. In additional to many traditional methods, different variants of proposed architecture are compared and ablation study is also performed.

This paper was published in 1998, a potential technical weakness would be that all the experiments were performed under rather small data size due to limited amount of data and computational capability. Recently, with the development of deep learning techniques, more variants are perprosed and larger scale experiments are performed, going beyond this paper.

5. Application and X-factor:

5.1 The application of online handwritten recognition

The proposed technique is appropriate for document recognition. More specifically, the use of CNNs makes it possible to automatically extract features from high dimensional input data, which is more flexible than traditional methods that heavily rely on hand-crafting features. The convolution operation keeps the local correlation and spatial invariance, makes it reasonable for image data. Additionally, the sub-sampling operation decrease the number of weights need to be learnt, makes the network more efficient to train and execute.

In terms of performance, compared with traditional methods, the proposed architecture achieves lower error rate, while the memory requirement and noise resistance are applicable. Two cases of GTN show how the CNN-based recognition module can be integrated with other modules as a whole system. Compared with independently trained modules, The globally training paradigm makes the deployment of system more flexible.

5.2 Other application domains

In my view the most promising part of this paper is the apply of CNN for automatic feature extraction, which alleviates the need for hand-crafting. Inspired by this, CNN has been used in almost every everywhere in computer vision, such as recognition, segmentation, object detection and tracking [3]. In some domains, the recent proposed CNN-based methods have exceed human's performance. Going beyond computer vision, CNN has been applied in other domains like natural language processing [23,8], audio [7], reinforcement learning, [12,13] and recommendation systems [21].

5.3. Further improvement

With the development of deep learning techniques and GPU computational cabability, it is possible to design much deeper architecture with better accuracy than LeNet-5, such as VGG19 [16] and ResNet152 [5].

However, for mobile devices with limited compute capability, designing compact but effective architecture could be promising. There has been some existing work such as Mobilenet family [6,17] and Shufflenet [24,14]. Another improvement could be compress the size of CNNs and remove abundant part, such as filter pruning [9] and soft weight sharing [22]. In terms of recognition and identification, it would be promising to improve the generalizability of system to make it effective for unseen images. Some existing work includes transfer learning [1,2,11], few shot learning [4,18,20] or even zero-shot learning [15,19].

5.4 Discussion in class

This work is worthwhile to be presented and discussed in class for that it clearly shows why CNN is useful for feature extraction and how it could be applied for document recognition. By understanding this paper it could be helpful for the students to know the mechanism of recognition systems and develop their own modifications and applications. However this paper contains too much knowledge such as prior terminations, the architecture of CNNs and LeNet-5, the design of GTNs and case studies. This makes it hard to illustrate all knowledge in class discussion, thus it would be more suitable to focus on one or just some of the parts, such as discuss the CNN and its variants only.

5.5 X-factors

It’s interesting that the proposed LeNet achieved comparable memory requirements and promising low error rate with limited data and devices in1990s. It’s worthwhile to understand their design philosophy for that even nowadays it is still hard to deploy a large network on a small device with limited computing and memory capability. Another interesting factor would be that maybe we can go beyond recognition and try to dig the possibility of causal reasoning, such as question answering and image understanding. This could make the system more intelligent, not just classification.

6. Presentation:

The presentation of this paper is rather decent. The author provided thorough information about not only the theory, implementation but also detailed application cases for proposed architecture and framework. Furthermore, the basic terminologies such as data-driven learning, gradient-based learning and back-propagation are also illustrated at the very beginning, makes it easy to understand even for a newbee in this area. For the researchers and technicians, the depth is reasonable that it could be helpful for them to propose new variants or build applications based on this work. The writing style is clear and the experiment is valid, makes this work convincing and easy to follow.

7. References:

[1] Geng, M., Wang, Y., Xiang, T., & Tian, Y. 2016, ' Deep transfer learning for person re-identification ', *arXiv preprint arXiv*:1611.05244.

[2] Ghazi, M. M., Yanikoglu, B., & Aptoula, E. 2017, ' Plant identification using deep neural networks via optimization of transfer learning parameters ', *Neurocomputing*, 235, pp. 228-235.

[3] Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., ... & Chen, T. 2018, ' Recent advances in convolutional neural network ', *Pattern Recognition*, 77, pp.354-377.

[4] Gidaris, S., & Komodakis, N. 2018, ' Dynamic few-shot visual learning without forgetting', *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 4367-4375.

[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016, ' Deep residual learning for image recognition ', *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778.

[6] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. 2017, ' Mobilenets: Efficient convolutional neural networks for mobile vision applications ', *arXiv preprint arXiv*:1704.04861.

[7] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., ... & Slaney, M. 2017, ' CNN architectures for large-scale audio classification ', *In 2017 ieee international conference on acoustics, speech and signal processing (icassp)* pp. 131-135.

[8] Kim, Y. 2014, ' Convolutional neural networks for sentence classification ', *arXiv preprint arXiv:*1408.5882.

[9] Luo, J. H., Wu, J., & Lin, W. 2017, ' Thinet: A filter level pruning method for deep neural network compression ', *In Proceedings of the IEEE international conference on computer vision,* pp. 5058-5066.

[10] LeCun, Y., Bottou, L., Bengio, Y.& Haffner, P. 1998, 'Gradient Based Learning Applied to Document Recognition',*Proceedings of the IEEE*, pp.2278-2324.

[11]Long, M., Zhu, H., Wang, J., & Jordan, M. I. 2017, August, ' Deep transfer learning with joint adaptation networks ', *In Proceedings of the 34th International Conference on Machine Learning-Volume 70,* pp. 2208-2217.

[12] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. 2013 , ' Playing atari with deep reinforcement learning ', *arXiv preprint arXiv:*1312.5602.

[13] Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., ... & Kavukcuoglu, K. 2016, ' Asynchronous methods for deep reinforcement learning ', I*n International conference on machine learning*, pp. 1928-1937.

[14] Ma, N., Zhang, X., Zheng, H. T., & Sun, J. 2018, ' Shufflenet v2: Practical guidelines for efficient cnn architecture design ', *In Proceedings of the European Conference on Computer Vision (ECCV*), pp. 116-131.

[15] Romera-Paredes, B., & Torr, P. 2015, ' An embarrassingly simple approach to zero-shot learning ', *In International Conference on Machine Learning,* pp. 2152-2161.

[16] Simonyan, K., & Zisserman, A. 2014, ' Very deep convolutional networks for large-scale image recognition ', *arXiv preprint arXiv*:1409.1556.

[17] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. 2018, ' Mobilenetv2: Inverted residuals and linear bottlenecks ', *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 4510-4520.

[18] Snell, J., Swersky, K., & Zemel, R. 2017, ' Prototypical networks for few-shot learning ', *In Advances in Neural Information Processing Systems,* pp. 4077-4087.

[19] Socher, R., Ganjoo, M., Manning, C. D., & Ng, A. 2013, ' Zero-shot learning through cross-modal transfer ', *In Advances in neural information processing systems,* pp. 935-943.

[20] Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., & Hospedales, T. M. 2018, ' Learning to compare: Relation network for few-shot learning ', *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* pp. 1199-1208.

[21] Tang, J., & Wang, K. 2018, ' Personalized top-n sequential recommendation via convolutional sequence embedding ', *In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 565-573.

[22] Ullrich, K., Meeds, E., & Welling, M. 2017, ' Soft weight-sharing for neural network compression ', *arXiv preprint arXiv:*1702.04008.

[23] Wang, P., et al 2015, ' Semantic Clustering and Convolutional Neural Network for Short Text Categorization ', *Proceedings ACL 2015*, pp.352–357.

[24] Zhang, X., Zhou, X., Lin, M., & Sun, J. 2018, ' Shufflenet: An extremely efficient convolutional neural network for mobile devices ', *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6848-6856.