Nirajan Koirala
CSC 9010_001

The 1990 paper presents a new type of architecture for image recognition tasks. This paper shows how neural networks trained using back-propagation can be applied to image-recognition tasks without a complex feature extraction stage. The 1998 paper further describes about the similar architecture of CNNs to recognize isolated characters from images. Since the architecture of a network strongly influences generalization performance, an architecture with sufficient amount of a priori knowledge about the task can be built to achieve high number of correct generalizations. Such an architecture can be directly fed with images instead of feature vectors and can deal with large amount of low-level information. The three main ideas incorporated in the architecture of CNN are local receptive fields, shared weights and spatial subsampling. These ideas are described using an LeNet-5 network which is able to recognize the alphanumeric characters.
Authors in the 1990 paper have used a database consisting of 9298 segmented numeral digitized from handwritten zipcodes. The database is augmented with 3349 printed digits coming from 35 different fonts to help models in training stage. In the preprocessing stage, segmentation of digits is performed by humans, and their normalization to a 16 by 16 pixel is carried out using a linear transformation. After this recognition task is carried out using a multi-layer neural network in which are the networks are adaptive but heavily constrained and are trained using backpropagation. The input of the network is a 16 by 16 image and output consists of 10 units for each individual digit. Weights of the neurons in the network are iteratively modified to minimize the objective function and gradient descent is used for this procedure. Since gradient descent requires one to compute the gradient of the objective function with respect to weights, backpropagation is used to measure this gradient.

An approach with fully connected net would result in a conflict where network is either highly overfitted or underfitted so locally connected nets are used. Feature mapping is carried out by scanning the input image and the states of the neuron after scanning is stored at corresponding locations in the feature map. The units in a feature map are constrained to perform same operation on different parts of the image and increase the model’s robustness. This type of restriction in the receptive fields of hidden units to local parts of image help CNN extract local features. Reduction in the number of free parameters is also achieved using this process since large number of units in an image have similar weights. These kind of local feature maps can be applied to subsequent hidden layers as well to extract features with increasing complexity and abstraction. The resolutions of feature maps is reduced as well by adding additional layers which perform local averaging and subsampling. The second hidden layer of LeNet-5 is a subsampling layer and it helps to reduce the sensitivity of output to shifts and distortions. This extra layer helps reduce the precision on the position of features which is helpful for the model to generalize well later on the test set. CNNs can be seen synthesizing their own features extractors as all weights are learned with backpropagation. Although, multi-layer nets resemble purely statistical techniques, specifying the knowledge by constraining the model introduces some high-level structure in the learning process. Hence, the architecture designed for recognizing handwritten digits can work without any kind of predetermined feature extractors and it is adaptive to other tasks as well like recognizing handwritten alphanumeric characters. Today, these type of convolution networks are used for online handwriting recognition, face recognition and a variant of these networks known as TDNNs are used for phoneme and spoken word recognition tasks as well as signature verification.

The intended audience for these papers are researchers and people already having a good understanding of deep-neural networks. The main points main in the papers are how CNNs are better than regular neural-networks for image recognition and how the architecture of CNN helps them perform good at image-classification tasks.

Strengths:
Both the papers have done a good job of explaining CNN and its architecture using appropriate examples.
The 1998 paper also talks about the similarity of different characters like 0 and O and 1 and I and discusses methods for dealing with this type of confusion like using a linguistic processor.

Weaknesses:
It would have been easier to follow description of the architecture of the CNN in 1990 paper if the figure was printed within the same page where the description takes place. Also in the same paper it would have been helpful if some equations for the objective and the non-linear squashing function were provided.

Points of Confusion:
The working of subsampling layer is a little complicated. The third idea of the architecture of CNN, spatial or temporal subsampling in the 1998 paper is difficult as well. Also, paper 1990 paper talks about introducing high level structure in the architecture which is confusing.

Discussion questions:

a) How spatial or temporal sub-sampling takes place in CNNs?
b) How the backpropagation algorithm is modified to take account of weight sharing?
c) How does the different layers in CNN (convolutional, pooling and fully connected) communicate with each other?