Ilkogretim Online - Elementary Education Online, Year; Vol 20 (Issue 5): pp. 1118-1127
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.05.121

Study of hierarchical learning and properties of convolution layer using sign language recognition model

Sagaya Mary J, Research Scholar, Christ University, Bangalore, Karnataka, India
Nachamai M, Data Scientist, Siemens Healthcare Private Limited, Bangalore, Karnataka, India
Dr.M.Vijayakumar, Professor, Department of MCA, K.S.R.College of Engineering, Tiruchengode, Tamilnadu, India
Dr. Chandra J, Associate Professor, Department of Computer Science, CHRIST (Deemed to be University)
Dr.Ravi Teja Bhima, Assistant Professor, Department of Computer Science and Engineering,GITAM Deemed-to-be-University, Hyderabad

Abstract: Convolution Neural Network (CNN) as a technique improves research minds to overcome the challenges of handcrafted feature extraction and classification. CNN be a part of the representation learning methods in deep learning architecture to discover the representation needed for detection and classification automatically. So far this technique has been thought as "black boxes", meaning that their inner working principles are mysterious and inscrutable. In order to understand the internal behavior of CNN, a model is developed on sign language recognition with 99.81%, 94.69%, 92.60% accuracy in train, test, and validation. While developing a model the inner principles of automatic feature extraction and the unique properties of convolution operations available in hierarchical CNN architecture are also learned. CNN is a multilayered network leading to feature learning and classification, it is necessary to understand how the features are learned from each layer and how it is transformed and fed into the next higher level layers without any human interventions.

Keywords: deep learning, convolution neural network, kernel, sparse connection, parameter sharing.

I. INTRODUCTION

In the past era, developing an object recognition system is quite complex since it requires careful engineering and significant knowledge to design preprocessing, feature learning and classification algorithm. The basic idea behind the algorithm is to convert the raw data into an applicable interior design or feature vector. Next, the vectors will be fed as input into the classifier which perhaps learns or categorize patterns in the output.

Current researches adopt the representation learning algorithm in which the machine is input with raw data and it has the capability to determine the representation required for learning or classification automatically. Deep learning is one of the representation learning methods [1] constructed with hierarchical layers, where each layer represents the features from elementary to abstract level. The composition of the transformation of each layer will have the ability to learn even more complex function**s**. Finally, the higher level features are converted into feature vector**s** in order to enhance the classification accuracy.

CNN is a part of the representation learning [2] techniques in deep learning architecture. In a hierarchical representation of CNN, the initial image is usually in terms of pixel array values. Learned features of the first layer of representation depict either the existence or non-existence of edges reflecting the distinct orientations and locations of the original image. The succeeding layer consistently finds out motifs by identifying the specific measures of edges with slight differences in the edge positions. The next layer combines motifs which correspond to the segments of well-known objects, and the following layers would recognize objects as consolidations of these segments [3].

Similar to deep learning, the key aspect of CNN utilizes the common learning procedure in each layer without human interventions. It is understood that image recognition in CNN learns the feature of each layer with many trainable phases organized in a consecutive manner. Hence, learning in convolution networks delivers feature hierarchy through simple architecture [4].

Classic CNN framework consists of alternatively arranged convolution layers, spatial pooling layers and fully connected layers (FC) [5]. These layers are valuable for feature learning (convolution, activation, and pooling layers) and classification (fully connected classification layer and softmax layer) [6].

In this paper, a CNN model is developed for sign language recognition with 99.81% in training, 94.69% in the testing and 92.60% accuracy in the validation. As we know CNN is a multilayered network leading to feature learning and classification, it is necessary to understand how the features are learned from each layer and how it is transformed and fed into the next higher level features without any human interventions. This paper is composed of the following sections, i) neuro-scientific background of CNN ii) feature learning in CNN iii) properties of convolution operation and iv) conclusions.

II. NEURO-SCIENTIFIC BACKGROUND OF CNN

2.1. Hierarchical structure of visual transmission

Starting from early 1959 to 1968, neuropsychologists-Hubel and Wiesel collaborated for several years to discover the basic fact of vision system. Their great experiment on recording the individual neurons of cat's and monkey's visual and striate cortex laid the foundation for developing the modern deep learning architecture[7][8][9]. Later on many researches have been done to gain more knowledge on deep hierarchy in visual cortex. To understand the basic overview of deep hierarchy in primate visual cortex, the study of Kruger et al in 2012[10] is explained here. They extracted data from macaque monkeys since most of the regions (cortical and motor areas, primary sensory) are homologous to human brain. The visual process of the neuron as shown in figure 2.2 starts with both left and right retina and the entire connections of the retina are passed to the Lateral Geniculate Nucleus (LGN) a visual area, before the destination point of the visual cortex. The occipital cortex ensures the generic feature representation in terms of various aspects of processing visual information from V1-V4 and Middle Temporal (MT) or V5 visual area retinotopically (connecting visual cortex with the receptive field). In feature representation, the size of the receptive field increases as the complexity of the features goes to the next higher levels of the hierarchy. The ventral stream bounded with TEO and TE is responsible for object recognition and categorization and the size of the receptive field is higher than in occipital cortex. Finally, the ventral stream achieves the abstract features that fit the level of the object for the explicit object class. Overall, the neurons in the primary visual regions extract simple image features over small local sections of visual spaces and the neurons produce complex features when the simple images are shift into higher visual spaces. The higher level features are more invariant to feature size, rotation or position. This way the hierarchical process of visual system of the brain is efficiently handled. Hence, the hierarchical structure of the visual system is deep in nature.

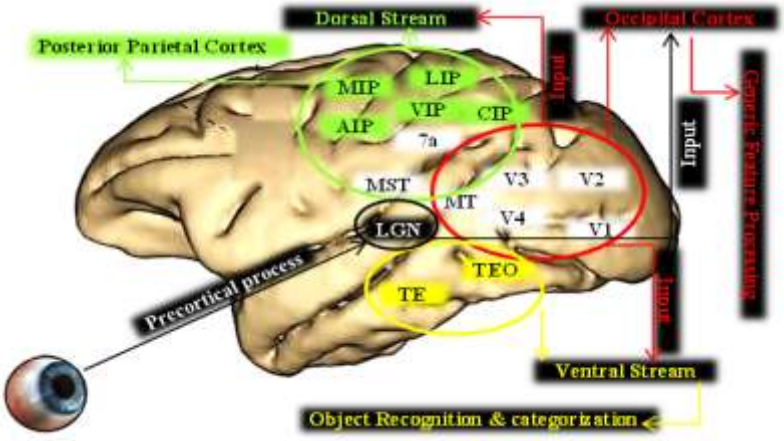


Figure 1 Primate visual cortex

2.1.1. Diffusion of light in a pre-cortical pathway

The visual transmission of pre-cortical pathway is explained in figure 2. Initially images in the form of light, stimulate the light sensitive tissue-retina which resides in the back of the eye. The photoreceptors present in the innermost layer of the retina converts the visual information (light) into neural signals. These signals are transmitted to retinal ganglion cells through sensory neurons called bipolar cells [11].

Further, the retinal ganglion cells carry visual signals to various brain areas together with LGN which conveys the sensory information to visual cortex at the back of the brain through optic nerves [12]. The neurons of the retina perform some preprocessing operations for the image without affecting the way it is represented.

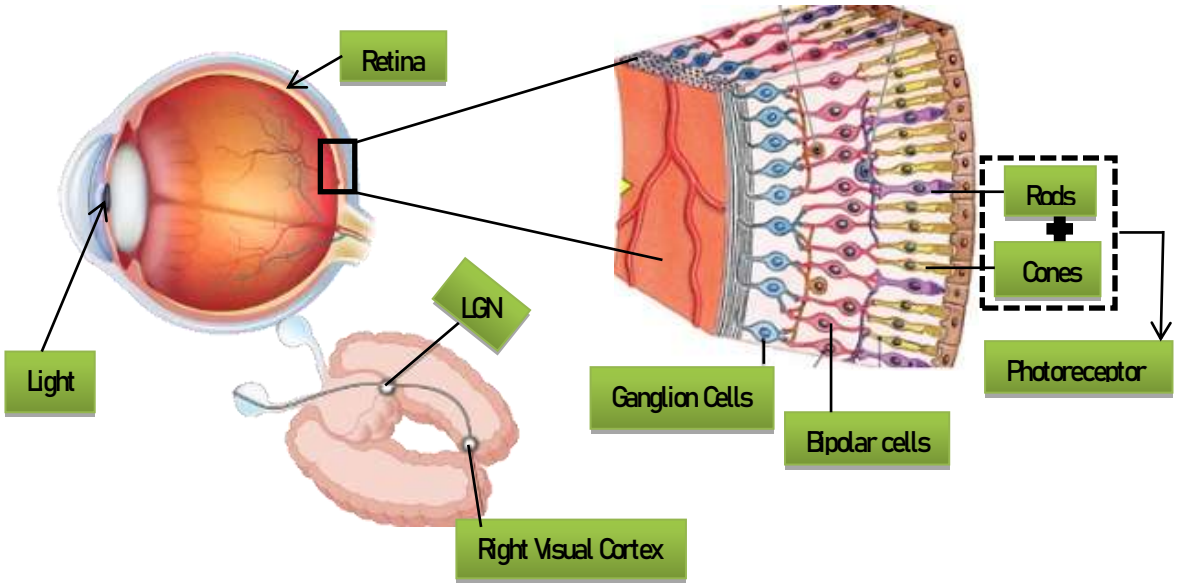


Figure 2 Visual information transmissions in a pre-cortical pathway

2.1.2. Neural transmission from V1-V4and MT

As shown in figure 3, the primate visual cortex is not organized in a rigorous successive hierarchy but there are some shortcuts observed in the levels. In order to reach the ventral pathway, there is a stream flowing from V1-V2-(V3)-V4 and to obtain dorsal pathway, the stream flows only from V1-V2-MT.

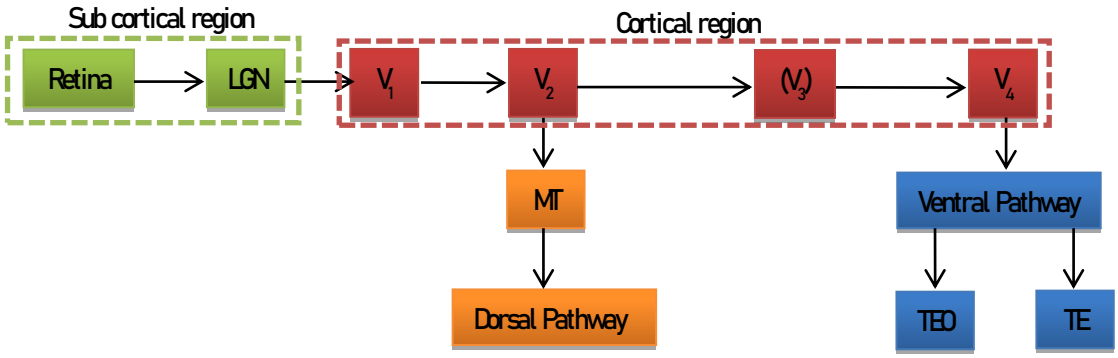


Figure 3 Visual transmissions from occipital cortex to dorsal and ventral pathway

2.1.3. Object detection in ventral stream

The cells of retina, LGN, visual cortex and Inferior Temporal (IT) cortex make the pathway to carry the visual information. On the basis of learning level, these cells are divided into upper and lower level pattern recognizer. Normally visual cortex is a lower level pattern recognizer when compare with IT since the capacity of learning of visual data is lower. The higher level pattern recognizer-IT does the prediction of the input data [13]. The input from V4 cells to the ventral stream-an IT cortex is partitioned into TEO and TE. The orientation and shape selective TEO neurons generally respond to very simple shape features. The receptive field of TEO is small (3-5 degree) compared to TE (10-20 degree) since TE is more

responsible for feature configuration. A trademark of the cells in ventral stream is robust in its responses even in scaling and position changes [14]. The overall characteristics of IT are shown in figure 4.

Figure 4 Receptive field and features learned in Inferior Temporal cortex

2.2 Role of receptive fields in visual transmission process

The term "receptive field" was originally coined by Sherrigton in 1906. He defined the region of the body part where a stimulus could produce a reflex. Further, Hartline in 1938 extended the term "receptive field" for the spot of the retina where any modification in light brightness affected the rate of input density of a retinal ganglion cells. The receptive fields of retina and thalamus are small and have simple structure as of two concentric circles. Such concentric arrangements were named as "center-surround" by Kuffler in 1953 in which on–center ganglion cells responded to light specks with dark background whereas off-center ganglion cells responded to dark spots with light background. Hubel and Wiesel in their series of studies (1959, 1961, 1962 and 1965) explained the presence of receptive field in the embedded cortical regions and they elucidated how the level of learning of receptive field increased up to the complex level when the cortical regions went deeper. The receptive fields in cortical regions were sound to be much more complicated than subcortical regions in terms of size and structure. Few cortical receptive fields appeared like thalamic receptive fields while others had extended sub-regions that responded to either dark or light spot and some zones did not respond to spots at all. In 1962, Hubel and Wiesel classified the cortical cells based on the receptive field structure in which if the cells had isolated sub-regions for the response of either light or dark spots then it were called simple cells and the cells which did not have separate sub-regions were known as complex cells. Visual receptive fields were in 2-D whose size can range from a few tiny arc to tens of degrees and sometimes it could also include the depth thence described as 3-D in planar space [15][16]. Receptive field and features learned in each stage of visual processing is explained in figure 5.

SUB CORTICAL REGIONS

Photoreceptors

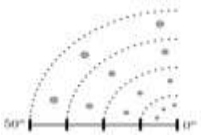
Receptive field

Features Learned

Rods and Cones

LGN

Receptive Field



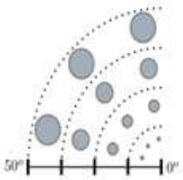
Features Learned

Luminance sensitive and single opponent cells

V1-1st Cortical Area

Features are sensitive and more complex than LGN

Receptive field



Features learned in V1

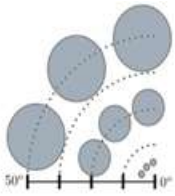
Linear oriented patterns Edges, gratings, line endings

End stopped or hyper complex End of a bar or edge or the border of a grating

Absolute disparity

V2-2nd Cortical Area

Receptive field



Features learned in V2

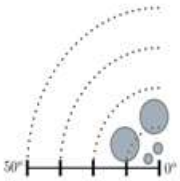
Texture & luminous defined contours with orientation tuning

Border ownership

V4-4th Cortical Area

Orientation, color, disparity and simple shapes

Receptive field



Features learned in V4

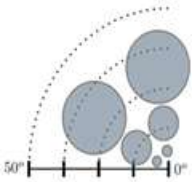
Curvature Selectivity

Contour with certain curvature or vertices in a particular angle

Color hue and luminance invariance

MT Area (Visual motion and binocular depth processing)

Receptive field



Features learned in MT

2D motion(compute motion by combining V1 features)

Motion Gradients &motion defined shapes

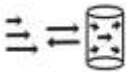


Figure 5 Features learned and the size of Receptive field in each stage

III. METHOD

3.1 Feature learning in CNN architecture

Convolution networks are multilayered trainable architecture constructed of multiple stages. The learning process happens in three stages, convolution or filter bank layer, non-linear mapping layer, and a pooling layer. The first phase accepts various input modalities (1D, 2D, and 3D) and goes through a set of filters generating several feature maps, which are processed into a point wise non-linear mapping and further down-sampled to reduce the spatial dimensionality.

3.1.1 Convolution operation

The linear convolution operation in CNN is different from the traditional matrix multiplication in a fully connected neural network [17]. From mathematical perspective, convolution is an operation of two functions (input and kernel) to produce the third function (feature map) that illustrates how the shape of one is reformed by the other [18]. While convolving, the kernel slides over the neighborhood pixels of the original image and extracting or learning the feature that is represented in a feature map as the re-estimated pixel. This process is explained in figure 6. The 2D convolution operation is performed by sliding 3D kernel over the 3D input image with height, width and depth (number of channels) volumes. Though the kernel and input are in a 3D volume, the output will be a 2D feature map since the kernel passes through the height, width and not the depth (d). This can be understood from the equation (1) where the kernel ($W_{l,m,d}$) pixels are multiplied with the set of units of small neighborhood raw input pixels ($I_{j,k,d}$) bounded with the size of the kernel. The summed product is further added with the constant bias (b) value and is stored as a 2D feature map in ($F_{j,k}$) neuron. The kernel (W) in machine learning is handcrafted based on the domain knowledge whereas in deep learning, it is an additional learnable parameter of the particular dimensions of the matrix [19] from the input image.

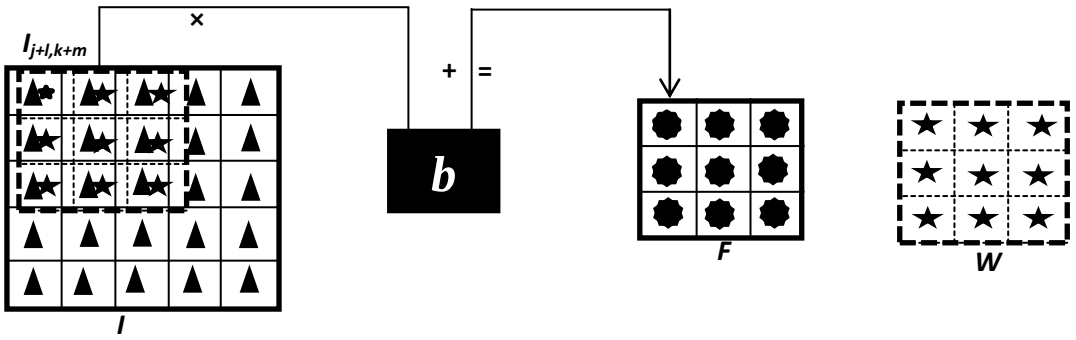


Figure 6 The process of forming a neuron in the feature map through convolution operation

$$F_{j,k} = b + \sum_{0}^{l}\sum_{0}^{m} W_{l,m,d} I_{j+l,k+m,d} \quad (1)$$

3.1.2 Complex learning in hierarchical convolution layers

CNN is a computational model that comprises deep subsequent layers for learning the representation present in the raw data in a hierarchical manner. This hierarchical structure learns the internal representation of the data by receiving the representation from the previous layer. Since the computation of convolution is crucial in learning, it is treated as the core for building the CNN structure. The convolution process in the elementary level of the model connects local regions of the input *(100×100×3)* pixels into their weights *(3×3×3)* through scalar product to determine the output neurons *(100×100×6)* in the form of edges, corners and endpoints at particular orientation and location. Further, continuous edgesforming a contour (motifs), and the representations inside the contour in the form of parts of the object, and the objects from the combination of parts of the objects are learned by increasing the feature map into *16, 32, 64* and *128.* The robustness with shift, distortions etc., of convolution layers enriches as the hierarchy moves from simple to complex level. CNN is best suited for its learning ability to automatically learn accurate features [20]. After receiving the input from one layer, it learns the features, and learned features are passed to next layer without any human interventions [21]. Convolution layers

constructed for learning a sign language is plotted in figure 7. To efficiently use the convolution layer one should have the knowledge of properties of convolution layer in CNN network.

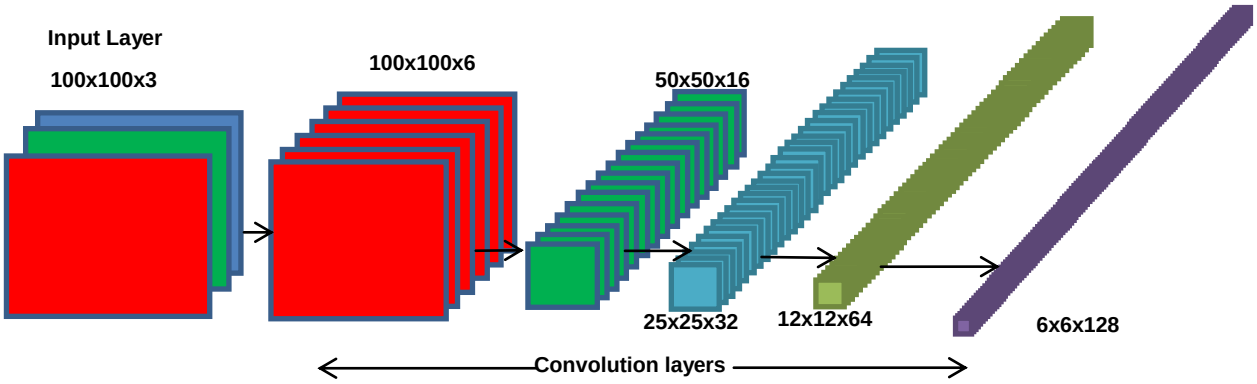
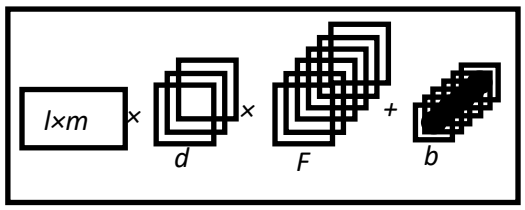


Figure 7. The overall Convolution layer architecture for learning the features

3.2 Properties of convolution layer

3.2.1 Kernel

Kernel or filter is a user-defined array portion that contains the subset representation extracted from the input. During convolution, the learned parameters in a kernel scan the entire input in a linear manner by placing its same weight over the neighborhood pixels of the image. The convolved features are plotted as neurons in feature maps.The successive neurons in a feature map are computed in the manner the kernels move on the basis of stride – a hyperparameter. The accuracy of the neuron is confirmed by controlling the weight with the additional signed vector value (bias) which is constant for the whole feature map. Since the portion of the kernel is extracted from the input, some of the neurons learn similar features. The hierarchy of the feature maps in a layer learns different sets of features through different kernels. Gradual increment in sets of features in successive hierarchical layers improves the levels of learning from elementary to abstract level. To understand the mathematical calculation of parameter learning, first convolution layer of the model is plotted in figure 8a in which ($l \times m \times d$) denotes the size of the kernel, F represents the size of the feature map introduced in first layer and b represents the bias value. The depth of the feature map of each layer is determined by the number of kernels introduced in each layer. The process of hierarchical learning in CNN is explained in figure 8b, it is observed that the depth level of the kernel increases from elementary level to abstract level. In the beginning, since only the elementary level features are learned and the unlearned features are preserved in feature maps as the size of input image. As we know, the depth of the feature map is directly proportional to the number of kernel**s** and inversely proportional to the size of the feature map**s**. The shrinkage of the size occurs by preserving new features with previously learned features. When the learning reaches to abstract level, it ultimately diminishes the size of the feature map and analogously increases the depth where the feature maps are so formed from the deep stacking of the small parts of the object.



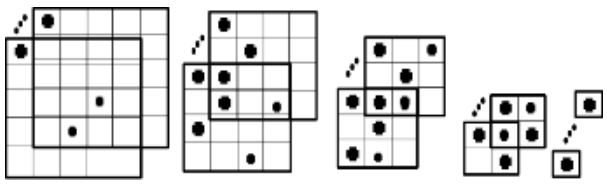


Figure 8a Parameter learning in 1st convolution layer *8b Hierarchical learning in CNN architecture*

3.2.2 Sparse connection

In a traditional feed-forward neural network, an extremely dense connection was required to find out a neuron in the feature map. Such network used matrix multiplication where a neuron describes the interaction between each input unit(X) and each output unit(Y) (connection from X_1 to X_n exists in the first hidden layer itself). Therefore, a single neuron was required $(X \times Y)$ parameters and $O(X \times Y)$ times. Convolution operation, however, have sparse connection which connects only the neighborhood pixels as

shown in figure 9a to find a neuron in the feature map (no connection from X_1 to X_n in the first hidden layer). To find a neuron, it requires $(K\times Y)$ parameters and only $O(K\times Y)$ time where K denotes the limited number of connections (kernel). The elegant deep CNN shows the indirect connection of larger portion of the input when it goes deeper. Thus the multi-level learning (connection exists from X_1 to $X_n$in the second layer in 5×5 image) of local neighborhood pixels (sparsely) reduces the number of parameters as well as improves the learning accuracy. Convolution operation is explained in equation (1) where $2D$ tensors of $F_{j,k}$ is computed from the range of the original image ($I_{j+k,l+m}$) which is termed as $X_i$in the input layer of figure9b and is mapped as F_i (where $0<i \leq n$) in the hidden layer. In figure 9b, it can be noticed that $X_1$and X_n are not directly connected with one another to find out any of the neurons from F_1 to F_n. The figure 13b illustrates the clear picture of sparse connection, we have considered a 5×5 input image whose pixels are entirely connected to a single output neuron (O_1) through single hidden layer. The equation (2) calculates $O_{j,k}$ from the features($F_{j,k}$) learned in the hidden layer.

$$O_{j,k} = b + \sum_{0}^{l}\sum_{0}^{m} W_{l,m,d} F_{j+l,k+m,d} \tag{2}$$

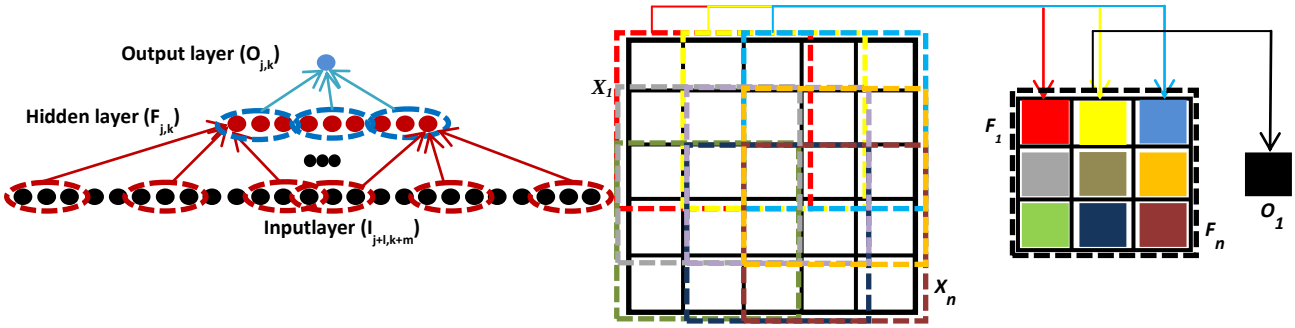


Figure 9a: Sparse connection in 5×5 input pixel (Ij+l,k+m) to compute a neuron in feature map(Fj,k)

9b: The interaction of O1 neuron with the entire pixels through a hidden layer

3.2.3 Weight sharing

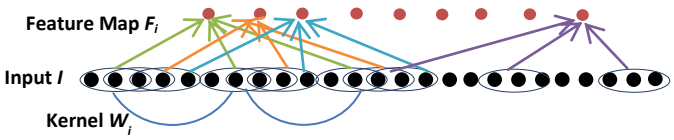


Figure 10: Process of weight sharing to find a feature map

$$F_i = I \times W_i \tag{3}$$

In a traditional neural network, the kernel parameters learned once is not used again to compute the output of the layer. Parameter sharing(figure 10) a unique property of convolution layers handle the kernel in a different way in which a set of unique kernel parameters, learned only once instead of learning separate sets of parameters for every location. The property of weight sharing can be renamed as tied weight since the value of the weight applied in one input is tied to the value of a weight applied elsewhere. It is explained through the equation 3, in which the weight parameter (W_i) is same for the entire image (I) to find out every neuron in the feature map (F_i). Hence the run time of parameter sharing in convolution layer is still same as $(O(K\times N))$ since one set of parameters are used for learning all the locations. Further, it reduces the memory locations by storing only k parameters which is comparatively very small of several orders of magnitudes of the actual size of the input. The weight parameters changed in the second feature map to learn the different representations. The replication of the same set of weight over the entire image extracts only the similar local features presented anywhere on the input plane. The precise feature extracted on one layer is combined into the next level of local features to form higher

order features [22]. On the whole, the properties of convolution layers are greatly effective in computation and judicious storage allocation than dense matrix multiplications.

IV. CONCLUSION

A deep learning model is a special sort of representation learning technique that determines the explanatory factors or features. One of the crucial elements for its state of the art results is the usage of CNN architecture which could alternate convolutional layers and pooling layers. Computer vision based CNN achieves accurate result in complex processes like automatic feature extraction and classification. Although CNN have been successful in giving solutions for all computer vision tasks, their internal operations are mysterious and inscrutable. This paper is a solution to understand how the features are extracted from the conventional hierarchical layers and the unique properties of convolution layer. This study helps us to get the clear idea of constructing the architecture in order to avoid trial and error process.

REFERENCES

1. LeCun, Y. et al., (2015) 'Deep learning', nature, Vol. 521, No. 7553, p.436.
2. Hasanpour, SH. et al., (2018) 'Towards Principled Design of Deep Convolutional Networks: Introducing SimpNet', arXiv preprint, arXiv:1802.06205.
3. LeCun, Y. et al., (2010) 'Convolutional networks and applications in vision', In ISCAS, Vol. 2010, pp. 253-256.
4. Farabet, C. et al., (2013) 'Learning hierarchical features for scene labeling' , IEEE transactions on pattern analysis and machine intelligence, Vol. 35, No. 8, pp. 1915-1929.
5. Lin, M. et al., (2013) 'Network in network', arXiv preprint, arXiv:1312.4400.
6. Doukali, F. (2018) 'Convolutional Neural Networks (CNN, or ConvNets). Medium'. [online] https://medium.com/@phidaouss/convolutional-neural-networks-cnn-or-convnets-d7c688b0a207.
7. D. Hubel and T. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. Phyiology, 160:106–154, 1962.
8. D. Hubel and T. Wiesel. Receptive fields and functional architecture of monkey striate cortex. Journal of Physiology, 195(1):215–243, 1968.
9. Goodfellow, Ian, et al. Deep learning. Vol. 1. Cambridge: MIT press, 2016.
10. Kruger, N., Janssen, P., Kalkan, S., Lappe, M., Leonardis, A., Piater, J., Rodriguez-Sanchez, A.J. and Wiskott, L., 2012. Deep hierarchies in the primate visual cortex: What can we learn for computer vision?. IEEE transactions on pattern analysis and machine intelligence, 35(8), pp.1847-1871,(2012).
11. https://www.coursehero.com/sg/introduction-to-psychology/sight-and-visual-perception/
12. https://cdn.the-scientist.com/assets/articleNo/36746/iImg/15573/5431d6ac-bc46-403b-a4db-15adfcf95154-pg33-thebasics.jpg?_ga=2.39979745.2129381710.1585290326-1452409736.1585290326
13. M. Ahmad, J. Joe and D. Han, "CortexNet: Convolutional Neural Network with Visual Cortex in human brain," 2018 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia), Jeju, 2018, pp. 206-212, doi: 10.1109/ICCE-ASIA.2018.8552151.
14. Riesenhuber, M. and Poggio, T. Hierarchical models of object recognition in cortex. Nature neuroscience, 2(11), p.1019(1999).
15. http://www.scholarpedia.org/article/Receptive_field
16. Martinez, L. M., & Alonso, J. M. (2003). Complex receptive fields in primary visual cortex. The neuroscientist, 9(5), 317-331.
17. Ronao, C.A. and Cho, S.B., 2015, November. Deep convolutional neural networks for human activity recognition with smartphone sensors. In International Conference on Neural Information Processing, Springer, Cham, 46-53.
18. Convolution. (2019). Retrieved from https://en.wikipedia.org/wiki/Convolution
19. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W. and Jackel, L.D., 1989. Backpropagation applied to handwritten zip code recognition. Neural computation, 1, 541-551.
20. Mei, S. et al., (2017)' Invariant feature extraction for image classification via multi-channel convolutional neural network', In Intelligent Signal Processing and Communication Systems ,International Symposium IEEE , pp.491-495

21. Patel, P and Bhatt, D. (2018) 'A Quick Look at Image Processing with Deep Learning - Open Source For You'. [online] https://opensourceforu.com/2017/11/a-quick-look-at-image-processing-with-deep-learning/.
22. Deep Learning (CS7015): Lec 11.3 Convolutional Neural Networks. (2019). Retrieved from https://www.youtube.com/watch?v=PmZp5VtMwLE