Convolutional Neural Networks and Architectures: A Brief Overview

A brief study of different architectures of CNN for Image Retrieval

[1]Vaibhav Kumar Dubey, [2]Punit Kumar Johari

[1]PG Scholar, [2]Assistant Professor

[1]Department of CSE & IT,

[1]Madhav Institute of Technology and Science, Gwalior, India

Abstract: Deep Convolutional Neural Networks (CNNs) are specialized Neural Networks capable of proving state-of-the-art results based on various benchmarks. CNN has somewhat powerful learning ability which is achieved by using various feature extraction levels that can learn the structural or hierarchical representations from the data. With the exponential increase of large amount of data and certain improvements in the computation power of hardware's there has been an increment in the research of CNN which has resulted in development of various architectures of CNN. Recent trend in developing deep CNN architectures has improved the performance on various computer vision related tasks based on innovations in ideas behind architectures and parameter optimization. Due to these various architectural designs has been explored based on different activation and loss functions, regularization, parameter optimization and restructuring the processing units. The latter has proved to be the major improvement in the representational capacity of CNN along with using block as a structural unit instead of a layer. This paper thus focuses on different versions of CNN architectures reported recently, covers the understanding of elementary concepts of CNN Components and throw some lights on application and challenged in CNNs.

***Index Terms* – Deep Learning, Convolutional Neural Networks, VGG, AlexNet, ResNet and DenseNet.**

I. Introduction

Artificial Intelligence (AI) has introduced specialized research area for Machine Learning which powers the computers to learn by making the relationships among the data and make decisions without being programmed explicitly. Various Machine Learning algorithms has been developed in order to replicate the human sensory response such as vision and speech but have failed to achieve their satisfactory responses [1-5]. The nature of Computer Vision and its challenges gave rise to development of Neural Networks (NN) [6] which replicates the Visual Cortes processing and is called Convolutional Neural Network (CNN).

The design and architecture of Convolutional Neural Network was inspired by Hubel and Wiesel's work (1962) [7] which was based on visual cortex structure. CNNs has been proven to be best for understanding the content of the images and have also shown the state-of-the-art results on various image processing tasks such as image segmentation, image recognition, image detection and various image retrieval related tasks [8]. In various industries and top MNCs such as Microsoft, Google and Facebook have made various groups for exploring new architectures of CNN [9].

CNN has been divided into several learning stages comprising of combination of convolutional layer, subsampling layers and nonlinear processing units [10]. Each layer performs several transformations using a group of several filters [6]. The operation of convolution involves extraction of locally co-related features by diving the image into smaller number of slices. Thus, it becomes capable of learning interesting features. The output produced by these convolutional layers is then assigned as an input to the nonlinear processing units which helps in learning abstraction and embeds non linearity in the feature space. This generates various patterns of activations for various responses and thus make self-capable of differentiating semantic studies in images. The out of these nonlinear functions is then followed up by subsampling, which makes the input invariant to distortions at geometrical level.

CNN has the capability of extracting low level, mid-level and high-level features. Higher level features which are generally considered as abstract features are combinations of low level and mid-level features. CNN have automatic feature extraction which makes it reduce the use of sperate feature extractor. Hence it can learn a well internal representation from raw pixels distinctive processing.

During the training phase, CNN learns by following backpropagation algorithm following, regularly changing of weights with respect to the input. The relevant costs function which used optimization also use backpropagation algorithm which is similar to the response based on human brain learning.

For dealing with complex learning problems, Deep architecture has more advantage over shallow architecture. At different level of abstraction, the grouping of multiple linear and nonlinear processing units in a layer wise structure gives deep networks the power to learn complex representation. By increasing the depth, the representational capacity of CNN can be enhanced and thus become more useful in image classification and segmentation [11]. Along with supervised learning, deep convolutional neural network has ability of learning the useful representation from huge amount of un-labeled data. Transfer learning has introduced a next level concept where feature including both low-level and high-level CNN features can be transferred to generic recognition tasks [12,13]. The main advantage of CNN is automatic feature extraction, weight sharing, multi-tasking and structural learning [14,15]. Along with the mentioned the major contribution towards CNN architectures is that the learning process can be visualized in a layer wise structure.

There has been proposal of innovations of various architectures of CNN since 2012. The innovations have been categorized as regularization, parameter optimization, structural reformulation, etc. Although main improvement in CNN is mainly due to processing units and new block designing. Application based on CNN became more popular after the remarkable performance of AlexNet on ImageNet dataset [11]. Similarly, introduction of new concept on layer wise visualization of features was bought up by Zeiler and Fergus [16] which made the trend on extracting features on low spatial resolution using VGG [17]. Recently Google has introduced a popular idea of split, transform and merge which then known as Inception block. The inception block gave an idea of branching within a layer, which makes extraction of features at different spatial scales [18]. ResNet [19] introduces skip connection in 2015 became popular for training of Deep CNNs. Skip connection later on was used by various other Nets such as Inception-ResNet [20], ResNext [21] etc. Later, the research was then routed towards improvement of architectural designs or layer structure of the network which resulted in various new architectural ideas such as channel boosting, information processing based on attention etc. in Convolutional Neural Network [22-24].

Various studies have been conducted on Deep CNNs in the last few years which has explored the basic components of CNN and their alternatives. The survey in [25] has reviewed various architectures and their components from 2012 to 2015. Also, surveys has been made on different applications of CNN [22,26]. This paper has been organized in the order as: Section 1 introduces the underlying basics concepts of CNN and their resemblance along with their contribution in Computer Vision. Section 2 provides an overview on CNN components. Section 3 discuss the recent and trendy innovations in CNN architectures and divides the CNN into various classes. Section 4 discuses the application of CNN. Section 5 discusses the challenges of CNNs. The last section denotes the overall conclusion..

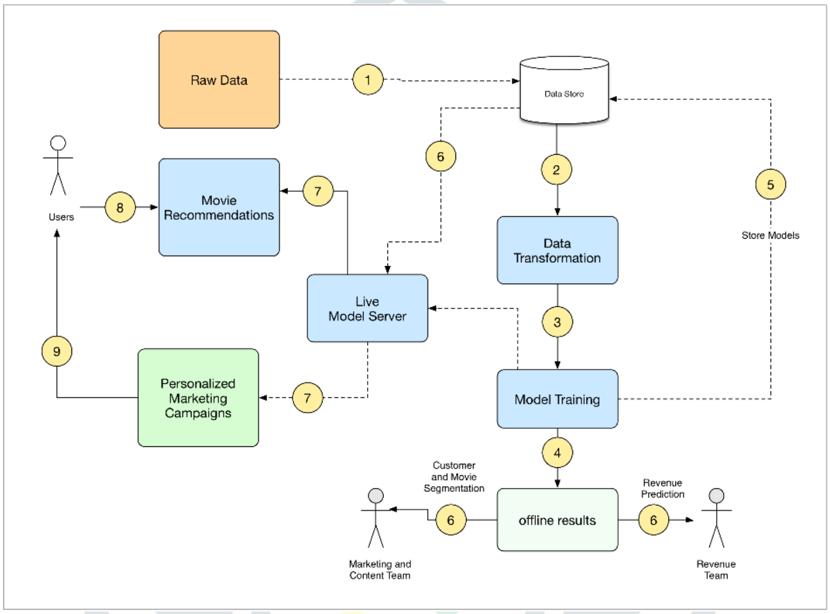

Figure 1: Basic ML System [27].

II. Basics Of CNN Components

In the present era CNN is mostly used techniques of Machine Learning especially in computer vision related tasks and has proven to be state-of-the-art results to ML related tasks. A Block diagram in Figure 1 has been shown. CNN is widely used in the extraction of features and selection stages of model due to is possession of good features as well as strong discrimination ability.

The architecture of CNN generally covers alternate lavers of convolution followed by pooling along with one or more fully connected layers in the end. Sometimes Fully connected layers are replaced by global average pooling layers. Along with different learning stages, for optimizing the performance of CNN various units such as dropouts and batch normalization are also embedded [28]. The structural arrangements of CNN components play a vital role for designing architecture for getting enhanced performance.

2.1 Convolutional Layer

It consists of sets of various convolutional Kernels where the act of kernel is performed by each neuron. Image's respective field are associated with these kernels. It is performed by slicing the image into small blocks which are also called as receptive fields and convolving them to the specifies sets of weights where multiplication of filter elements with receptive fields elements are done. This is represented in equation (1) [31].

$$F_l^k = (I_x^y * K_l^k) \qquad (1)$$

I_x^y is input image where *x*,*y* shows spatial locality, K_l^k shows *lth* convolutional kernel of the *kth* layer. Slicing of image into small blocks helps to find the correlated values of pixels locally. Various features are extracted by sliding the convolutional filter on the image with the same weights. The operation of convolution is then further categorized based on padding type, filter size and type and convolutional direction [29].

2.2 Pooling Layer

After extraction of features, the approximate position is preserved but it is less important. Down sampling or pooling like convolutional is interesting local operation. It adds up the same type of information about receptive field and outputs the dominant response within this local region [30]. This is expressed in equation (2) [31].

$$Z_l = f_p(F_{x,y}^l) \qquad (2)$$

Zi represent *ith* output feature map. $F_{x,y}^l$ shows *ith* input feature map where f_p () defines type of operation of pooling. Operation of pooling helps to extract the combination of features that are invariant to small distortions and translational shifts [32,33]. Pooling also helps in increment of generalization by reducing the overfitting along with feature map size reduction which regulates the network complexity. Max, average, overlapping and L2 are some of the different types of pooling operation which are used for extracting translational invariant features [34,35].

2.3 Activation Function

Such function helps in a decision making and also helps in learning some complex patterns. To accelerate the learning process it is mandatory to select appropriate activation functions. This is defined in equation (3) [31].

$$T_l^k = f_A(F_l^k) \qquad (3)$$

F_l^k is an output of a convolutional operation. f_A () adds nonlinearity and shows the transformed output T_l^k for *kth* layer. Different activation functions such as tanh, sigmoid, ReLU, maxout and other variants [34,37] are used to inculcate nonlinear combinations of features. To overcome the vanishing gradient problem, ReLU and its variant are mostly preferred over other activations [37].

2.4 Batch Normalization

To overcome the problems related to internal covariance shift within feature maps, batch normalization is performed. It is a change in the distribution of hidden units' values, which slows the convergence enforcing small learning rate values and need careful initialization of parameters. The equation (4) [31] is shows as Batch normalization for transformed feature map T_l^k

$$N_l^k = \frac{T_l^k}{\sigma^2 + \sum_i T_l^k} \qquad (4)$$

N_l^k represents normalized feature map and 0sigma represents variation in feature map. It merges the distribution of feature map by bring them to unit variance and zero mean [38]. Along with it, it acts as a regulating factor and smoothen the gradient flow which ultimately improves the generalization of network by not relying on dropout.

2.5 Fully Connected Layer

It is used for classification tasks at the end of the network. It takes input from the previous layer and globally analyses output from all previous layers [39]. It forms a nonlinear combination of selected features, which are then used for data classification. It is a global operation unlike pooling and convolution [40].

III. Innovation in CNN Architectures.

Several modifications have been done in CNN architectures from 1989 to till date. These modifications are done based on regularization, parameter optimization structural reformulation etc. Based on the kind of architectural improvement CNN are broadly classified into seven different which are shown in Figure 2. In this paper we have covered only few architectures which are most widely used for computer vision related tasks.

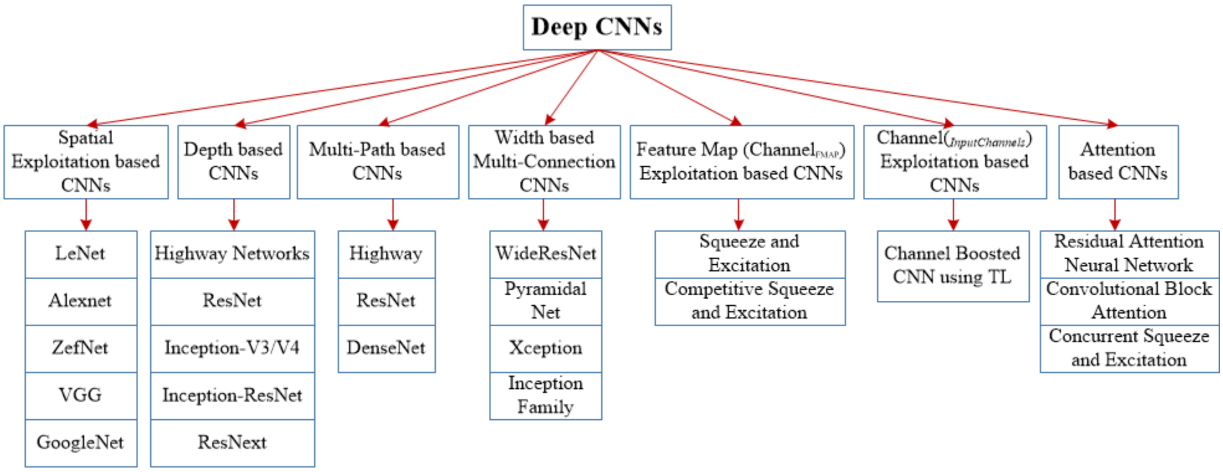
Deep CNNs
Spatial Exploitation based CNNs
Depth based CNNs
Multi-Path based CNNs
Width based Multi-Connection CNNs
Feature Map (Channel$_{FMAP}$) Exploitation based CNNs
Channel($_{InputChannels}$) Exploitation based CNNs
Attention based CNNs
LeNet
Alexnet
ZefNet
VGG
GoogleNet
Highway Networks
ResNet
Inception-V3/V4
Inception-ResNet
ResNext
Highway
ResNet
DenseNet
WideResNet
Pyramidal Net
Xception
Inception Family
Squeeze and Excitation
Competitive Squeeze and Excitation
Channel Boosted CNN using TL
Residual Attention Neural Network
Convolutional Block Attention
Concurrent Squeeze and Excitation

Figure 2: CNN Architecture based on Various Schemes [31]

CNN consists of huge number of parameters like neurons or number of processing units, filter size, padding, stride, learning rate etc. [41 42]. CNN performs better on coarse- and fine-grained level details sue to different size of filter levels of graduality which uses filters of small size to extract fine grained and large size coarse grained information.

3.1 LeNet

LeNet was proposed by LeCun in 1998 [43]. It is the first CNN which showed the state-of-the-art performance on hand character recognition tasks. It is capable of identifying characters without being affected by rotation and variations of scale and position and small distortions. LeNet consists of five alternative layers of convolutional and pooling along with two fully connected layers and thus is a feed-forward neural network. In early system due to low computation power and speed LeNet exploited the basis of image that the neighboring pixels are corelated to each other and distributed on entire image. Hence convolution with learning parameters are effective way to get similar features at various location with few parameters. LeNet was the first of its CNN to reduce the number of parameters along with computation and ingeniously learned features.

3.2 AlexNet

AlexNet [11] has been considered to be the first deep CNN architecture, which showed handy performance for image classification and recognition tasks. AlexNet was proposed by Krizhevesky et al [11] who powered up the learning capacity and making it deeper by applying some number of parameter optimization strategies. The Architectural design of AlexNet is shown in figure 3. AlexNet was trained with two NVIDIA GTX 580 GPUs to overcome the hardware limitations and get the benefits from it. In AlexNet, the extraction of feature stages was increased to 7 from 5 (LeNet) to make it more capable of diverging the categorization of images. Although depth improves the generalization, but the major drawback was overfitting. Later Krizhevesky et al explored the concepts of Hinton [44,45] in which the algorithms skip randomly some of the transformational units during the phase of training to enforce the algorithms to learn robust feature characteristics. To overcome the problem of vanishing gradient ReLU [46] was embedded as a non-saturating activation function and to improve the convergence rate. There were implementation of local response normalization and overlapping subsampling for improving the generalization by reducing the overfitting. Use of large filters such 11X11 and 5X5 was additional adjustment. AlexNet has its own importance since it has efficient learning approach and made various researchers to start a new phase on CNNs.

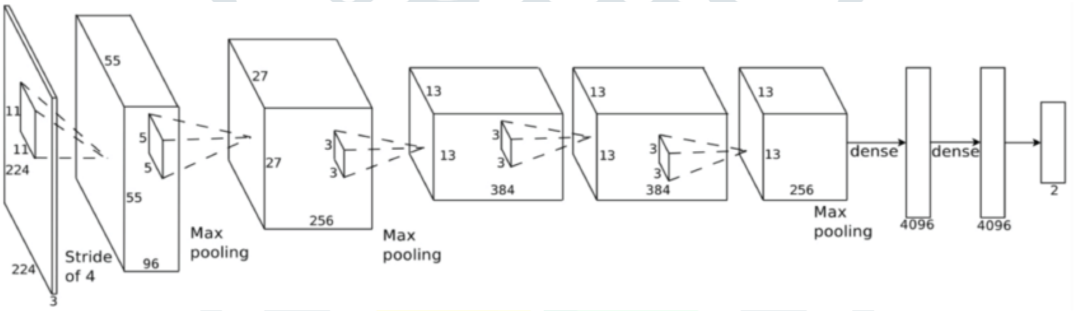

Figure 3: AlexNet Architecture [47].

3.3 VGG

Simonyan and Zisserman proposed a new simple and effective architecture design for CNN which was called as VGG after the successful use of CNN for the purpose of image recognition. It is modular in layer patterns [17]. VGG is 19 layers deeper than AlexNet [11]. VGG has replaced the 11x11 filter and 5x5 filters with a 3x3 filters and proved to be the concurrent placement to induce the effects of huge size filter. It suggested that parallelly placing the small size filters will be equally effective as placing a large size filter such as 5x5 and 7x7. It gives an additional advantage of low computation complexity by reducing the parameters. It made the researcher to motivate and find new domain on the small size filters. VGG helps in regulating the complexity by using 1x1 filter between the convolution layers which also learns a linear combination of feature map produced as a result. For better tuning of the network, max pooling [48] is used after the convolution operation and padding is used to conserve the image size. VGG has proved to be effective in image classification and problems related to localization. Although the limitation with this architecture is high computational costs.

3.4 GoogleNet

GoogleNet is also known as Inception V1. The main goal of this architecture is to achieve high accuracy with less computational power. Inception module or blocks was introduced in this architecture. It integrates multi scale convolutional transformation by using split, transform and merge idea of extracting the features. Figure 4 shows a block diagram of inception block. It incorporates block of different sizes including 1X1, 3X3, 5X5 which captures information based in spatial resolutions and channel information. In GoogleNet, the traditional convolutional layer was replaced by small blocks with micro Neural Network which was proposed by Network in Network Architecture [44]. This idea helped in solving the problems based on learning the diversifying variations included in the same category of different images. The focus of this architecture was to make the parameters of CNN more efficient. It regulates the computation by summing up the bottleneck layer using 1X1 filter of convolution. It uses the connection of sparse where every output channel is not connected to input channels. This was done to overcome the problem of information which is redundant in nature and cut the cost by neglecting channels that were not useful or relevant. The density of the connection was reduced by using average pooling at the last layer instead of using fully connected layers. GoogleNet introduced the idea of auxiliary learners to boost up the convergence rate. But the major drawback was its heterogeneous structure that needs to be tuned from every module. Some other limitation was that GoogleNet reduces the feature space which lead to the loss of useful data.

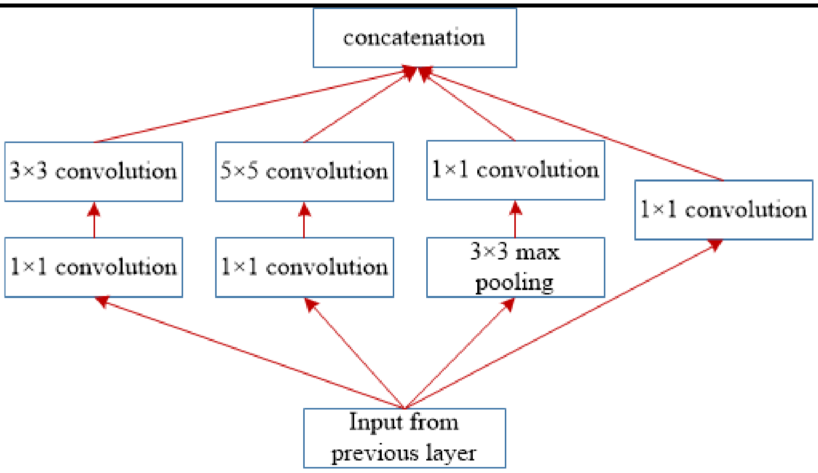

Figure 4: Basic Architecture of Inception Block [31].

3.5 ResNet

He et al [19] proposed one of the optimal methods which is called as Residual neural Network or ResNet to resolve the problems during the phase of training the Deep Networks. It is 152 layers deeper. Figure 5 shows a basic Residual block. It has less computation complexity then previous Network like AlexNet [11] and VGG [17] which is 20 and 8 times deeper respectively. The [19] showed the Residual Neural Network with 50,101 and 152 layers which showed more accuracy than a typical 34 layers plain network. It proved more efficient on popular image dataset named as COCO [49] which was 28% more efficient. Later on, with such a tremendous performance it became an important in various computer vision related tasks such as image recognition and localization.

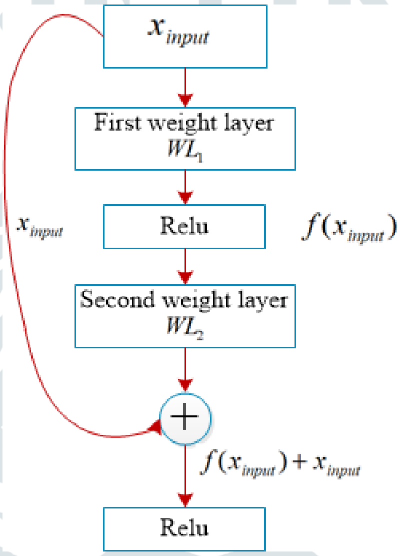
x_{input}
First weight layer WL_1
x_{input}
Relu
$f(x_{input})$
Second weight layer WL_2
$+$
$f(x_{input}) + x_{input}$
Relu

Figure 5: A basic Residual block [31].

3.6 ResNext

Also coined by the term Aggregated Residual transform network is an improvement over the type of network which is so called inception network [21]. A new term called cardinality was coined based on the idea of split, transform and merge. It is another dimension which denotes the size of transformations [50,51]. Inception network not only improved capability of learning of CNN but they do make network resources effective. ResNext has been derived from VGG and ResNet [17,18,19]. It utilizes the deep homogenous characteristics of VGG and GoogleNet by fixing spatial resolution to 3x3 filters within split, transform and merge blocks. Figure 6 shows a ResNext model. It used multiple transformation and defined them as cardinality. Increasing the cardinality improves the performance of the model. The optimization of training was done through skip connections while complexity was regulated using low embedding 1X1 filter.

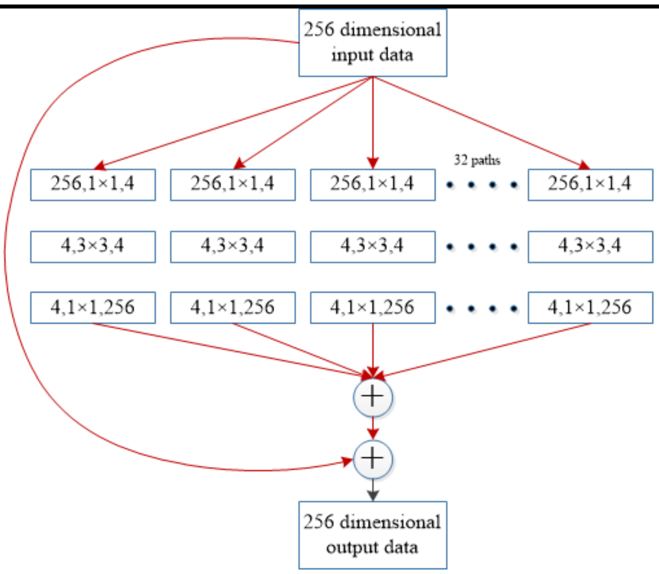

Figure 6: Building block of ResNext [31].

3.7 DenseNet

This architecture was proposed to solve the problem of vanishing gradient [19,52]. DenseNet uses connectivity through cross layer in a modified way. It connects each layer to each other layer in a feed forward mechanism. Hence feature maps of all previous layer was used as inputs to all subsequent layer. It makes the effect of cross layer depth wise convolutions. The network becomes capable to differentiate between information which is added to network and information which is preserved since it concatenates the previous layers features despite adding them. It has narrow layer structure and becomes costlier while increasing the number of feature maps. Reduction in overfitting on tasks on smaller training sets implements a regularizing effect.

IV. Applications of CNN

CNN is used in almost every different ML application which are linked to computer vision tasks such as segmentation, classification, regression, object detection, object recognition etc. It needs a large amount of data for learning. CNN has almost abundant the labeled data like detection of text, faces, pedestrians, medical image segmentation etc. Some of the trendy applications are discussed as:

4.1 Motion Recognition

Motion recognition deals with identification of various actions, activities and motions in a human body. Some of the areas covered are pose estimation, Face detection, action recognition etc. One of the major challenging tasks is Face detection. Farfade et al [53] has proposed a method for Deep CNN for face recognition from different facial expression. Zhang et al. [54] developed multitask cascaded CNN to perform face recognition. Human pose detection is another major challenge in Computer Vision due to high variability in the pose of body.

4.2 Natural Language Processing

The method converts the human language into something that can be understand by the computer. Speech recognition, language modeling and analysis etc. are some of the NLP based applications. It has introduced a new concept of sentence modelling which to get the semantic of the sentence. Usually traditional method just analyses the data based on features or words and ignore the core depth of the sentence. Dynamic CNN and dynamic K-Max Pooling is used by author in [55]. Collobert et al [56] has proposed another method based on CNN architecture to perform multitasking at the same time related to MLP like slicing, recognizing entity like name, role modelling, modelling of languages etc. Xue et al [57] proposed another method which performs matching between two sentences and hence can be implemented to different kind of languages.

4.3 Image Classification

One of the widely use of CNN is image classification [58,59] specially in medical imaging where operations like diagnosing of cancer using histopathological images are used [60]. Spanhol et al [61] has used the CNN for breast cancer image diagnosis and then the outputs are compared with the network model that is trained on the dataset that has manual feature descriptor crafted by in-hand [62]. Wahab et al [63] has proposed another method for the same in which two phases are done. In first hard non-mitosis samples are identifies. In Second data augmentation is done in order to cope with the class skewness problem.

4.4 Object Detection

This method involves identification of different objects in the images. R-CNN (Region-Based CNN) is used for object detection. In a work by Ren et al. [65] full convolution is used for feature extraction as a space to identify the boundaries and object located at various positions. Dai et al [66] proposed a new method based on region-based object detection using fully convolution network. To learn the semantic feature another work was proposed by Gidaris et al [67] based on multi region based Deep network.

4.5 Speech Recognition

With the improvement in hardware resources compared with the historical times the utilization and training of Deep neural network with huge data becomes possible and thus also performed better with speech recognition related tasks. The work proposed by Hamid et al [64] proposed a CNN based independent speech recognition model. Results have shown a reduction of ten percent of error rate compared with the previous one. Various CNN architecture is being innovated which are both based on full and limited number of weights sharing within the convolutional layer. Performance is evaluated at pre training phase by initializing the network.

V. CNN CHALLENGES

Although Deep Convolutional Neural Network has made a good takeover and got good performance on data there has also been some challenges that are associated with the use of Deep Neural network and their architecture related to Computer Vision tasks.

Addition of even small amount of noise in the image make the network incapable to classify the original and its perturbed data differently.

Some Challenges while training the deep neural network includes lack of explain-ability and interpretability. Sometimes it becomes difficult to verify them with respect to computer vision tasks. It is important to know the characteristics of the feature extracted before the classification. Although this can be solved sometimes with the help of feature visualization.

The learning mechanisms is supervised learning and hence there is a need of large and annotated data. The performance of CNN is sometimes affected by hyperparameter selection. A slight change in the values can affect its performance, thus it is mandatory to have careful selection of parameters which can solve some issues related to optimization strategy.

One of the major challenges in Deep CNN is the requirements of powerful hardware resources such as GPUs. Hence to get the proper training and efficient result of the Deep system it is required to also have good resources of hardware.

VI. CONCLUSION

Convolution Neural network has made an outstanding progress in Image processing domain. Various researches have been going on to improve the performance of the Deep CNN which are related to computer vision tasks. The paper has summarized some of the important and widely used architectures of Deep CNN, there applications and challenges.

A lot improvement has been done based on the exploring the depth and other structural information. The new phase in innovation of Deep CNN architectures has introduced the effective block architectures. These blocks act as an auxiliary learner. They play a vital role in increasing the performance of CNN by problem aware learning. Also, these block-based structure motivates learning in a structural manner making the architecture more elementary and understandable.

REFERENCES

[1] O. Chapelle, P. Haffner and V. N. Vapnik, "Support vector machines for histogram-based image classification," in IEEE Transactions on Neural Networks, vol. 10, no. 5, pp. 1055-1064, Sept. 1999.
doi: 10.1109/72.788646

[2] Lowe, David. (1999). Object Recognition from Local Scale-Invariant Features. Int Conf Comput Vis. 2. 1150 - 1157 vol.2. 10.1109/ICCV.1999.790410.

[3] Bay H., Tuytelaars T., Van Gool L. (2006) SURF: Speeded Up Robust Features. In: Leonardis A., Bischof H., Pinz A. (eds) Computer Vision – ECCV 2006. ECCV 2006. Lecture Notes in Computer Science, vol 3951. Springer, Berlin, Heidelberg

[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 2005, pp. 886-893 vol. 1.
doi: 10.1109/CVPR.2005.177

[5] Timo Ojala, Matti Pietikäinen, David Harwood, A comparative study of texture measures with classification based on featured distributions, Pattern Recognition, Volume 29, Issue 1, 1996, Pages 51-59,
ISSN 0031-3203, https://doi.org/10.1016/0031-3203(95)00067-4.

[6] Y. LeCun, K. Kavukcuoglu and C. Farabet, "Convolutional networks and applications in vision," Proceedings of 2010 IEEE International Symposium on Circuits and Systems, Paris, 2010, pp. 253-256. doi: 10.1109/ISCAS.2010.5537907

[7] H Hubel, D & N Wiesel, T. (1971). Aberrant visual projection in the Siamese cat. The Journal of physiology. 218. 215-243. 10.1113/jphysiol.1971.sp009603.

[8] Ciresan, D.C., Giusti, A., Gambardella, L.M., & Schmidhuber, J. (2012). Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images. NIPS.

[9] Li Deng; Dong Yu, "Deep Learning: Methods and Applications," in Deep Learning: Methods and Applications , , now, 2014, pp**.**

[10] Jarrett, K., Kavukcuoglu, K., Ranzato, M., & LeCun, Y. (2009). What is the best multi-stage architecture for object recognition? 2009 IEEE 12th International Conference on Computer Vision, 2146-2153.

[11] Krizhevsky, Alex & Sutskever, Ilya & E. Hinton, Geoffrey. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems. 25. 10.1145/3065386.

[12] Saeed, Aqsa & Khan, Asifullah. (2018). Adaptive Transfer Learning in Deep Neural Networks: Wind Power Prediction using Knowledge Transfer from Region to Region and Between Different Task Domains**.**

[13] Pan, Sinno & Yang, Qiang. (2010). A Survey on Transfer Learning. Knowledge and Data Engineering, IEEE Transactions on. 22. 1345 - 1359. 10.1109/TKDE.2009.191.

[14] Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M.S. (2016). Deep learning for visual understanding: A review. Neurocomputing, 187, 27-48**.**

[15] Liu, Weibo & Wang, Zidong & Liu, Xiaohui & Zeng, Nianyin & Liu, Yurong & E. Alsaadi, Fuad. (2016). A survey of deep neural network architectures and their applications. Neurocomputing. 234. 10.1016/j.neucom.2016.12.038.

[16] Zeiler M.D., Fergus R. (2014) Visualizing and Understanding Convolutional Networks. In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8689. Springer, Cham

[17] Simonyan, Karen & Zisserman, Andrew. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 1409.1556**.**

[18] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1-9.

[19] He, Kaiming & Zhang, Xiangyu & Ren, Shaoqing & Sun, Jian. (2015). Deep Residual Learning for Image Recognition. 7.

[20] Szegedy, Christian & Ioffe, Sergey & Vanhoucke, Vincent. (2016). Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. AAAI Conference on Artificial Intelligence**.**

[21] Xie, S., Girshick, R.B., Dollár, P., Tu, Z., & He, K. (2017). Aggregated Residual Transformations for Deep Neural Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5987-5995.

[22] Khan, Asifullah & Sohail, Anabia & Ali, Amna. (2018). A New Channel Boosted Convolutional Neural Network using Transfer Learning.

[23] Woo, S., Park, J., Lee, J.-Y. & Kweon, I. S. CBAM: Convolutional Block Attention Module. (2018)

[24] Wang, F. et al. Residual attention network for image classification. Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017 2017–Janua, 6450–6458 (2017)

[25] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, Tsuhan Chen, Recent advances in convolutional neural networks, Pattern Recognition, Volume 77, 2018, Pages 354-377, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2017.10.013.

[26] Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M. et al. Journal of Big Data (2015) 2: 1. https://doi.org/10.1186/s40537-014-0007-7

[27] Pentreath, Nick, et al. “Machine Learning with Spark - Second Edition.” O'Reilly | Safari, Packt Publishing, Apr. 2017, www.oreilly.com/library/view/machine-learning-with/9781785889936/b219ea46-001c-470d-b742-f9bf1e89697a.xhtml**.**

[28] Bouvrie, J. 1 Introduction Notes on Convolutional Neural Networks. (2006)**.**

[29] Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444 (2015).

[30] Lee, C.-Y., Gallagher, P. W. & Tu, Z. Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree. in Artificial Intelligence and Statistics 464–472 (2016).

[31] Khan, Asifullah & Sohail, Anabia & Zahoora, Umme & Saeed, Aqsa. (2019). A Survey of the Recent Architectures of Deep Convolutional Neural Networks**.**

[32] Scherer, Dominik & Müller, Andreas & Behnke, Sven. (2010). Evaluation of pooling operations in convolutional architectures for object recognition. 92-101. 10.1007/978-3-642-15825-4_10.

[33] Ranzato, M., Huang, F.J., Boureau, Y., & LeCun, Y. (2007). Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition. 2007 IEEE Conference on Computer Vision and Pattern Recognition, 1-8.

[34] Wang, T., Wu, D.J., Coates, A., & Ng, A.Y. (2012). End-to-end text recognition with convolutional neural networks. Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), 3304-3308.

[35] Boureau, Y. Icml2010B.Pdf. (2009). doi:citeulike-article-id:8496352

[36] Xu, B., Wang, N., Chen, T., & Li, M. (2015). Empirical Evaluation of Rectified Activations in Convolutional Network. CoRR, abs/1505.00853.

[37] Hochreiter, Sepp. (1998). The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 6. 107-116. 10.1142/S0218488598000094**.**

[38] Ioffe, Sergey & Szegedy, Christian. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.

[39] T. Yoshioka et al., "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Scottsdale, AZ, 2015, pp. 436-443.
doi: 10.1109/ASRU.2015.7404828

[40] Rawat, W., & Wang, Z. (2017). Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. Neural Computation, 29, 2352-2449.

[41] Shin, Hoo-chang & Roth, Holger & Gao, Mingchen & Lu, Le & Xu, Ziyue & Nogues, Isabella & Yao, Jianhua & Mollura, Daniel & Summers, Ronald. (2016). Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. IEEE Transactions on Medical Imaging. 35. 10.1109/TMI.2016.2528162.

[42] Kafi, Mojtaba & Maleki, M & Davoodian, N. (2015). Functional histology of the ovarian follicles as determined by follicular fluid concentrations of steroids and IGF-1 in Camelus dromedarius. Research in Veterinary Science. 99. 10.1016/j.rvsc.2015.01.001.

[43] LeCun, Y. et al. Learning algorithms for classification: A comparison on handwritten digit recognition. Neural networks Stat. Mech. Perspect. 261, 276 (1995).

[44] Srivastava, Nitish & Hinton, Geoffrey & Krizhevsky, Alex & Sutskever, Ilya & Salakhutdinov, Ruslan. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research. 15. 1929-1958.

[45] Dahl, G. E., Sainath, T. N. & Hinton, G. E. Improving deep neural networks for LVCSR using rectified linear units and dropout. in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on 8609–8613 (IEEE, 2013)

[46] Xu, B., Wang, N., Chen, T., & Li, M. (2015). Empirical Evaluation of Rectified Activations in Convolutional Network. CoRR, abs/1505.00853.

[47] Glomerulus Classification with Convolutional Neural Networks - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/AlexNet-CNN-architecture-layers_fig1_318168077 [accessed 11 May, 2019]

[48] Huang, F. J., Boureau, Y.-L., LeCun, Y. & others. Unsupervised learning of invariant feature hierarchies with applications to object recognition. in Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on 1–8 (2007)

[49] Lin TY. et al. (2014) Microsoft COCO: Common Objects in Context. In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8693. Springer, Cham

[50] Sharma, A. & Muttoo, S. K. Spatial Image Steganalysis Based on ResNeXt. 2018 IEEE 18th Int. Conf. Commun. Technol. 1213–1216 (2018). doi:10.1109/ICCT.2018.8600132

[51] Han, W., Feng, R., Wang, L., & Gao, L. (2018). Adaptive Spatial-Scale-Aware Deep Convolutional Neural Network for High-Resolution Remote Sensing Imagery Scene Classification. IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, 4736-4739.

[52] G. Huang, Z. Liu, L. v. d. Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 2261-2269.
doi: 10.1109/CVPR.2017.243

[53] Farfade, S.S., Saberian, M.J., & Li, L. (2015). Multi-view Face Detection Using Deep Convolutional Neural Networks. ICMR.

[54] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," in IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499-1503, Oct. 2016.
doi: 10.1109/LSP.2016.2603342

[55] Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A Convolutional Neural Network for Modelling Sentences. ACL.

[56] Collobert, Ronan & Weston, Jason. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. Proceedings of the 25th International Conference on Machine Learning. 160-167. 10.1145/1390156.1390177.

[57] Xue, Xiaobing & Yin, Xiaoxin. (2011). Topic modeling for named entity queries. 2009-2012. 10.1145/2063576.2063877.

[58] P. Sermanet, S. Chintala and Y. LeCun, "Convolutional neural networks applied to house numbers digit classification," Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), Tsukuba, 2012, pp. 3288-3291.

[59] Z. M. Long et al., "Modeling and Simulation for the Articulated Robotic Arm Test System of the Combination Drive", Applied Mechanics and Materials, Vol. 151, pp. 480-483, 2012

[60] Cireşan D.C., Giusti A., Gambardella L.M., Schmidhuber J. (2013) Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks. In: Mori K., Sakuma I., Sato Y., Barillot C., Navab N. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013. MICCAI 2013. Lecture Notes in Computer Science, vol 8150. Springer, Berlin, Heidelberg

[61] F. A. Spanhol, L. S. Oliveira, C. Petitjean and L. Heutte, "Breast cancer histopathological image classification using Convolutional Neural Networks," 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, 2016, pp. 2560-2567.

[62] F. A. Spanhol, L. S. Oliveira, C. Petitjean and L. Heutte, "A Dataset for Breast Cancer Histopathological Image Classification," in IEEE Transactions on Biomedical Engineering, vol. 63, no. 7, pp. 1455-1462, July 2016.
doi: 10.1109/TBME.2015.2496264

[63] Wahab, Noorul & Khan, Asifullah & Lee, Yeon Soo. (2017). Two-phase deep convolutional neural network for reducing class skewness in histopathological images based breast cancer detection. Computers in Biology and Medicine. 85. 10.1016/j.compbiomed.2017.04.012.

[64] O. Abdel-Hamid, A. Mohamed, H. Jiang and G. Penn, "Applying Convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition," 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, 2012, pp. 4277-4280.

[65] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017.

[66] Dai, Jifeng & Li, Yi & He, Kaiming & Sun, Jian. (2016). R-fcn: Object detection via region-based fully convolutional networks.

[67] Gidaris, Spyros & Komodakis, Nikos. (2015). Object Detection via a Multi-region and Semantic Segmentation-Aware CNN Model. 10.1109/ICCV.2015.135.