

A Comparative Review of Recent Architectures of Convolutional Neural Networks

Kalpna Devi¹, Aman Kumar Sharma²

¹Research Scholar, Computer Science Department, Himachal Pradesh University, Shimla, H.P., India

²Professor, Computer Science Department, Himachal Pradesh University, Shimla, H.P., India

Abstract: A Deep Convolutional Neural Network (CNN) is an important part of deep learning that has delivered admirable successes in various competitions related to Image Processing and Computer Vision. Certain attractive application fields of CNN vary from Image and Video Recognition, Image Segmentation and Classification, Medical Image Analysis, Natural Language Processing, and Object detection. One of the greatest powerful abilities of deep CNN is the various feature extraction in an automatic way. Recently, developments in the research of CNNs and attractive deep CNN architectures have been described due to the inherence of the huge quantity of data and refinement in hardware automation. A handful of encouraging concepts such as the use of distinct activation and loss functions, regularization, parameters optimization, and architecture modernization, derive progress in deep CNNs. However, the remarkable advancement in the representational ability of the deep CNN is accomplished by architectural modernization. Thus, this review paper presents a brief survey of the advances that can occur in the architecture of CNNs from the very first architecture to the recent one. This paper, therefore, targets the inherent anatomy in the newly disclosed deep CNNs architectures and accordingly describes the strengths and gaps of various deep CNNs architectures.

Keywords: Deep Learning, Convolutional Neural Networks, Anatomy, Representational Capacity, Residual Learning, and Channel Boosted CNN

1. Introduction

The advanced kind of Neural Network (NN) is the Convolutional Neural Network (CNN) [1] which has shown a wide future in Computer Vision (CV) associated tasks. CNNs are one of the dominant learning algorithms for finding image content and have established an ideal accomplishment in image segmentation, classification, detection, etc. [2]. Besides, in academia, promising results have been captured in industries or companies such as Google, Microsoft, AT&T, NEC, and Facebook have advanced active research groups for presenting progressive architectures of CNNs [3]. The appealing feature of CNN is its ability to make use of spatial or temporal connections in data. CNN is a feed-forward multi-layered hierarchical network, where every one-layer carry's out numerous variations [4]. Convolution operations assist in drawing out suitable features from locally linked data points. The output of the convolution filters is then assigned to the non-linearity function (activation function), which supports learning extraction as well as implanting non-linearity in the feature span. The output of the non-linear activation function comes after the sub sampling, which supports encapsulating the results and constructs the input constants to geometrical curvatures [5]. CNN learns from end to end of the back propagation algorithm by controlling the alteration in weights in accordance with the object. CNNs with automated feature extraction capability diminish the demand for a discrete feature extractor [6]. The attractiveness of deep CNNs is elementary due to its multilayered, hierarchical architecture, which provides the power to draw out low, mid, and high-level features.

Deep architectures usually have dominance over shallow architectures when handling convoluted learning problems. It lay out the potentiality of learning complex depictions at distinct elevations of extraction because of the heap-up of

the different linear and non-linear processing units in a layer-wise style. The use of CNNs is increased in image classification and segmentation tasks [7], because of its deep architecture which upgrades the representational ability of CNN.

After the ideal presentation of AlexNet in the ImageNet dataset in 2012 [7], CNN-based approaches became extensive. Later, the notable modernization in CNN has been progressive and is mostly related to the improvement of processing units and composition of the latest blocks. In 2013, the knowledge of the layered view of CNN was proposed which enhances the extraction of features. Then, moved in the direction of the extraction of features at low spatial resolutions in deep architecture named VGG [8]. Google deep learning proposed an inception block by constituting the concept of a break, alter and combine, which approves the divergence inside a layer [9]. The approach of skip connection was proposed by ResNet [10] for training deep CNNs. After that, this approach was used in Inception-ResNet then Wide-ResNet [11], and in ResNeXt [12]. Several architectural designs such as Pyramidal Net [13], Wide ResNet, ResNeXt, Xception [14], etc. are initiating the cardinality or incrementing the width. Hence, the basis of research moved from parameter optimization and connection remodeling, in the direction of the enhanced architectural layout of the network. The aforementioned movement appeared in a lot of advanced architectural designs such as spatial and feature-map-wise utilization, Channel Boosting, attention-based information processing, etc.

In recent years, a lot of fascinated surveys or review papers are assisted on deep CNNs, which consider different algorithms and applications of CNN [15][16] [17]. They mostly focused on the concepts such as the use of distinct activation and loss functions, parameters optimization, regularization, etc. But, in this review paper, we consider

Volume 11 Issue 7, July 2022

www.ijsr.net

Licensed Under Creative Commons Attribution CC BY

inherent anatomy available in the current and outstanding CNN architectures described from 2012-2022. We discuss the numerous CNN architectures with strengths and gaps and also mention the categories like width, depth, spatial exploitation, multi-path, attention, etc. This paper will inspire the researchers or the readers to promote the abstract observation of the design concepts of CNN and thus advance speed up the architectural modernization in CNNs.

The remaining of the paper is organized as: section 2 discusses the architectural revolutions of CNN and also mentions the strengths and gaps of each CNN model in table 2. Finally, section 3 concluded with future work.

2. Architectural Revolutions in CNN

Different redesigns in CNN architecture have been made from 1989 to recent times. It is observed that the recasting of processing units and the composing of new blocks are responsible for the improvement in the performance of CNN.

A. LeNet

LeNet [18] is the first and the most popular CNN architecture which came in the year 1998 as shown in Figure 1. LeNet was originally developed to categorize handwritten digits from 0-9 of the MNIST Dataset. It is made up of seven layers, each with its own set of trainable parameters. It accepts a 32×32 -pixel picture, which is rather huge in comparison to the images in the data sets used to train the network. RELU is the activation function that has been used. The traditional fully connected multi-layered neural network has the main drawback that it examines each pixel as distinct input and employs a transformation to it, which causes a significant computational load, specifically at the time of Gardner and Dorling 1998 [19]. LeNet utilized the fundamental basis of the image that the neighboring pixels are interrelated to each other and the theme of the feature is scattered across the whole image. LeNet was the first CNN architecture, which diminished the number of parameters and learned features from raw pixels automatically.

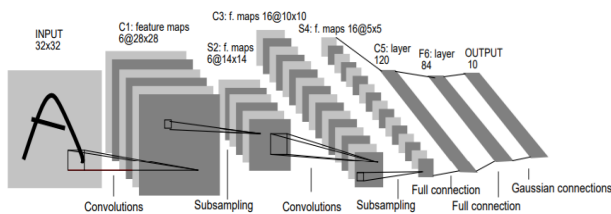


Figure 1: LeNet Architecture [18]

B. AlexNet

LeNet [18] though, initiate the history of deep CNNs, but it was restricted to hand digit identification tasks and didn't accomplish well to all categories of images. AlexNet [20] is designed as the first deep CNN architecture, which displayed revolutionary outcomes for image classification and identification tasks. AlexNet was presented by Krizhevsky et al., which improved the learning capability of CNN by constructing it deeper and by applying numerous parameter optimization approaches. The network is similar to the LeNet Architecture, but it includes a lot more filters than the original LeNet, allowing it to categorize a lot more

objects. It deals with over fitting by using "dropout" rather than regularisation. The basic architectural design of AlexNet is shown in Figure 2. In accumulation to this, ReLU was used as a non-immersing activation function to recover the converging rate by improving the problem of vanishing gradient [21] to some level.

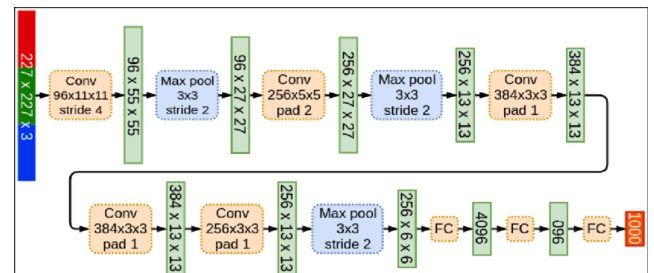


Figure 2: AlexNet architecture with 5 convolutions, 3 max-pooling, and three fully connected layers [19]

C. ZfNet

Earlier in 2013, the learning technique of CNN was built primarily on the hit-and-trial method. Due to this, the improvement in the performance of deep CNNs on complex images was limited. In 2013 Zeiler and Fergus presented an attractive multilayer Deconvolutional Neural Network (DeconvNet) that made out famous as ZfNet [22]. ZfNet was constructed to statistically determine network performance. The purpose of network activity determination was to follow the performance of CNN by examining neuron activation. Its architecture is shown in Figure 3.

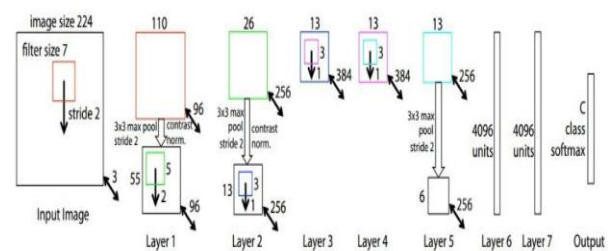


Figure 3: ZfNet Architecture with 5 convolutional layers with filter size 7×7 , max-pooling, dropout, and 3 fully-connected layers [22].

D. VGG

Simonyan et al. [23] presented an uncomplicated and efficient design concept for CNN architectures called VGG. It was based on a study on how to make such networks denser. There are different versions for VGG networks according to the layer number, such as VGG-13, VGG-16, and VGG-19. VGG was designed with 19 layers deep to recreate the association between depth and network illustration ability as compared to AlexNet and ZfNet. ZfNet, which was a frontline network of the 2013- ILSVRC competition, suggested that small-size filters can recover the performance of CNNs. Placed on these remarks; VGG changed the 11×11 and 5×5 filters with a heap of 3×3 filters. The usage of small-size filters allows an additional advantage of low computing complexity. These discoveries set a new drift in research to work with smaller-size filters in CNN. The main drawback related to VGG was the use of 138 million parameters, which make it computationally expensive and tough to organize on low-resource systems.

Model	Number of Parameters (millions)	Top-5 Error Rate (%)
VGG-11	133	10.4
VGG-13	133	10.5
VGG-16	133	9.9
VGG-19	134	9.4
VGG-16 (Conv1)	138	8.8
VGG-16	144	9.0

Figure 4: VGG Architecture with layers 11, 13, 16, and 19[23]

E. GoogleNet

GoogleNet [24] was the champion of the 2014-ILSVRC challenge and is also called Inception-V1. The model has comprised of a basic unit referred to as an "Inception cell" in which we perform a series of convolutions at different scales and subsequently aggregate the results. It combines multi-scale convolutional changes using split, transform and merge ideas. The architecture of the inception block is shown in Figure 5. The network used a CNN inspired by LeNet but implemented a novel element which is dubbed an inception module. To conquer the problem of unnecessary information by using sparse connections and reduce the cost by neglecting feature maps that were not appropriate. Furthermore, the density of the connections was decreased by using global average pooling at the last layer, in place of using a fully connected layer.

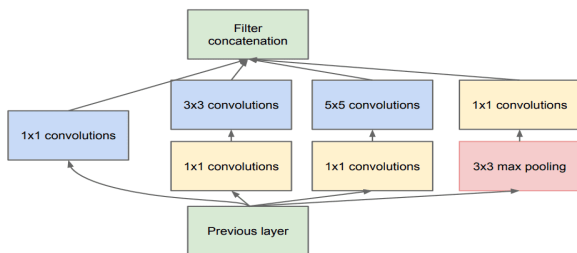


Figure 5: Elementary architecture of the inception block viewing the split, transform, and merge idea [24]

F. Highway Network

Is based on the knowledge that the learning capability can be enhanced by rising the network depth. But, the rise in depth of a network recovers performance generally for complex problems. But it is also concerned with problems of slow training of the network and convergence speed. In deep networks, because of the large number of layers, the error of back propagation may consequence in small gradient values at lower layers [25]. To resolve this problem, in 2015 Srivastava et al. [26] presented a deep CNN, named Highway Networks. These Networks make use of depth for learning augmented feature representation and presenting a new cross-layer interconnected mechanism for the fruitful training of the deep networks. Thus, Highway Networks are classified as depth as well as multi-path-based CNN architectures. In Highway Networks, the gating mechanism allows for computation paths along which information can flow across many layers without attenuation. They denote those paths as information highways as shown in Figure 6.

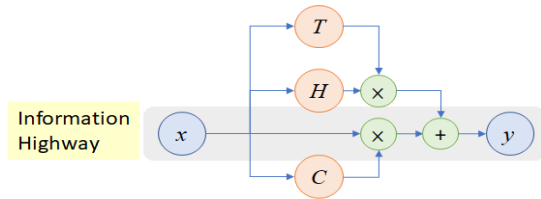


Figure 6: Highway Circuit [26]

G. ResNet

For the extension of deep networks, He et al [27] introduced a novel architecture called Residual Neural Network (ResNet). ResNet introduces the idea of residual learning in CNNs and arises with an effective framework for the training of deep networks. ResNet presented 152-layers deep CNN, which was the winner of the 2015-ILSVRC competition. Figure 7 shows the architecture of the residual block of ResNet, which displays that there is a direct link that skips several model levels. By using the idea of "skip connection" and abundant of batch-normalization for training 100s of layers successfully without the problems caused by vanishing/exploding gradient. ResNet showed less computational complexity and was 20 and 8 times deeper than AlexNet [20] and VGG [23] respectively. In the famous image recognition benchmark dataset named COCO [28], ResNet gained 28% improvement.

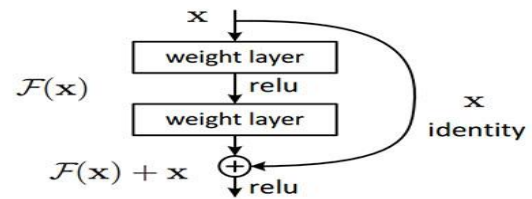


Figure 7: The basic structural unit of ResNet [27].

H. Wide Residual Networks (WRNs)

It is concerned that the feature reuse problem is the main drawback of residual networks which makes these networks very slow to train. To tackle this limitation, Zagoruyko and Komodakis proposed a novel architecture named Wide Residual Network (WRNs) [29], with decreased depth and increased width of residual networks. Wide Residual Networks introduce an additional factor k (which controls the width of the network), which increased the width. It also showed that as compared to the depth of the residual network, the widening of the layers might provide a more effective way of performance improvement. Zagoruyko and Komodak is demonstrated that even a simple 16-layer-deep wide residual network outperforms in accuracy and efficiency all previous deep residual networks, including thousand-layer deep networks, achieving new state-of-the-art results on CIFAR, SVHN, COCO, and significant improvements on ImageNet.

I. DenseNet

DenseNet [30] was proposed to resolve the problem of vanishing gradient same as solved by Highway Networks and ResNet. Recent work had shown that convolutional networks can be substantially deeper, more accurate, and more efficient to train if they contain shorter connections between layers close to the input and those close to the output. Huang et al [30] embraced that observation and introduced the Dense Convolutional Network (DenseNet),

which connects each layer to every other layer in a feed-forward manner. The traditional convolutional networks with L layers had L connections—one between each layer and its subsequent layer. DenseNet had $L(L+1)/2$ direct connections. DenseNet alleviated the vanishing-gradient problem, strengthened feature propagation, encouraged feature reuse, and substantially reduce the number of parameters. The proposed architecture was evaluated on four highly competitive object recognition benchmark tasks (CIFAR-10, CIFAR-100, SVHN, and ImageNet).

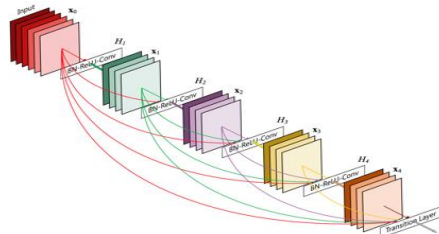


Figure 8: A 5-layer dense block with a growth rate of $k = 4$. Each layer takes all preceding feature maps as input [30]

J. Pyramidal Net

In previous Deep CNNs such as AlexNet, ResNet, and VGG, the heap of many convolutional layers can increase the depth of feature maps. The convolutional layer or block comes after a sub-sampling layer that decreases the spatial dimension. Therefore, Han et al [31] claimed that the learning ability of deep CNNs is restricted due to an extreme increase in the feature-map depth (no. of channels) and at the same instant, the loss of spatial information takes place. Han et al proposed the pyramidal Net to increase the learning ability of ResNet [27]. In this research, instead of sharply increasing the feature map dimension at units that perform down sampling, Han et al gradually increase the feature map dimension at all units to involve as many locations as possible. This proposed design had proven to be an effective means of improving generalization ability. Furthermore, they proposed a novel residual unit capable of further improving the classification accuracy with new network architecture. It was called a pyramidal Net because of its regular increase in the depth of features mapped in a top-down manner. The major difference between PyramidalNets and other network architectures is that the dimension of channels gradually increases, instead of maintaining the dimension until a residual unit with down sampling appears. A schematic illustration is shown in Figure 9.

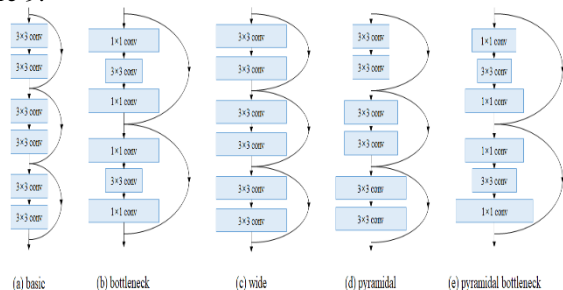


Figure 9: Schematic illustration of (a) basic residual units [27], (b) bottleneck residual units [27], (c) wide residual units [29], (d) pyramidal residual units, [31] and (e) pyramidal bottleneck residual units [31].

K. Xception

Xception [32] in the deep convolutional architecture can be considered as an extreme Inception, which applies the depth wise convolution design. To manage the computational complexity, Xception builds the original inception block larger and substitutes a single dimension (3×3) which comes after a 1×1 convolution in place of the multiple spatial dimensions (1×1 , 5×5 , 3×3). Xception constructs the network more efficient in computation, by separating spatial and feature-map (channel) relationships. The Architecture of the Xception block is shown in Figure 10.

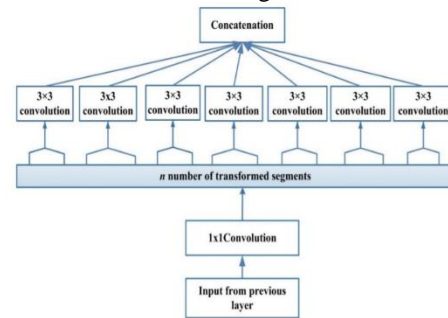


Figure 10: Xception block's architecture [32]

L. ResNeXt

To solve the limitation of ResNet regarding reducing the error rate, ResNeXt [33], which is also called the Aggregated Residual Transform Network, was introduced. Xie et al. used the idea of the split, transform, and merge in a powerful but easy way by initiating a new term; cardinality [24]. Cardinality is an added dimension, which states the size of the set of transformations [34] [35]. The main goal of ResNeXt is to handle large inputs and improve the accuracy of the network by increasing cardinality without the need of constructing the deeper layers of the network. ResNeXt exploited the deep similar topology of VGG and basic GoogleNet architecture by setting the spatial resolution to 3×3 filters within the split, transform, and merge block. Whereas, it used residual learning to improve the convergence of deep and wide networks [23] [24]. The building block for ResNeXt is shown in Figure 11.

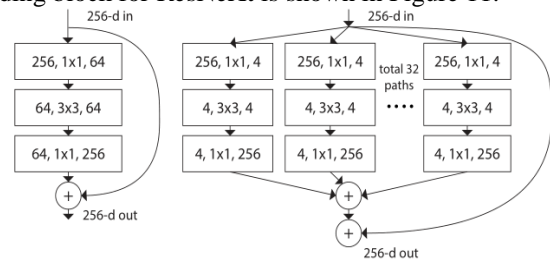


Figure 11: Left: A block of ResNet [19]. Right: A block of ResNeXt with cardinality = 32, with roughly the same complexity. [33]

M. Squeeze-and-Excitation Networks

Several recent approaches have shown the benefit of enhancing spatial encoding to boost the representational power of a network. Hu et al. [36] proposed a novel architectural unit, termed the “Squeeze-and-Excitation” (SE) block, that dynamically readjusts channel-wise feature replies by definitely modeling interrelationships between channels. This new block was known as SE-block (shown in Figure 12), which abolishes less significant feature-maps, but gives more weightage to the class specifying feature

maps. By heaping these blocks together, build SENet architectures that generalize enormously well across challenging datasets.

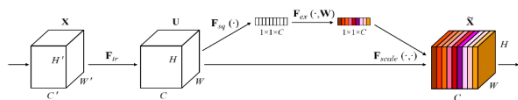


Figure 12: A Squeeze-and-Excitation block [36]

N. Competitive-Squeeze-and-Excitation Network

One of the major drawbacks of ResNet [27], is to use a residual unit to supplement the identity mapping and enable very deep convolutional architecture to operate well. However, residual architecture had to be proved diverse and redundant. To overcome the drawback of residual architecture Hu et al. [37] proposed a competitive squeeze and excitation based on the SE-block [36], known as the Competitive SE (CMPE-SE) Network. Re-scaled the value for each channel in the CMPE-SE structure will be determined by the mapping of the residual and identity, which enabled expansion and of the meaning of channel relationship modelling in residual blocks. Furthermore, Hu et al. designed a novel inner-imaging Competitive SE block to shrink the consumption and re-image the global features of intermediate network structure. Compared to the typical SE building block, the composition of the CMPE-SE block is illustrated in Figure 13.

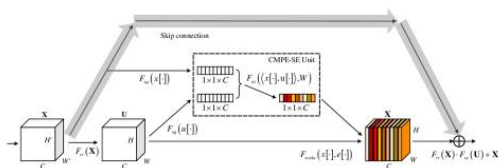


Figure 13: Competitive Squeeze-Excitation Architecture for Residual block [37]

O. Channel Boosted CNN using Transfer Learning (TL)

In 2018, Khan et al. [38] presented a new CNN architecture called Channel Boosted CNN (CB-CNN) built on the concept of boosting the number of channels for refining the representational ability of the network. Figure 14 shows the Block diagram of CB-CNN. Channel boosting is accomplished by synthetically creating additional channels (known as auxiliary channels) to get across auxiliary deep generative models and then utilizing them across the deep discriminative models. For refining the representation of the data, Khan et al. make use of the power of Transfer Learning (TL) and deep generative learners [39], and [40]. These

learners enhance the representational capacity of deep CNN-based discriminators.

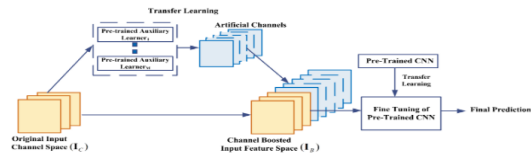


Figure 14: The basic building block of Channel Boosted deep CNN (CB-CNN) [38]

P. Residual Attention Neural Network (RAN)

Motivated by attention mechanisms and recent advances in the deep neural network, Wang et al. [41] presented a Residual Attention Network (RAN) to improve the network’s feature representation. The Residual Attention Network is a convolutional network that adopts a mixed attention mechanism in a “very deep” structure by stacking multiple Attention Modules which generate attention-aware features. The attention-aware features from different modules change adaptively as layers go deeper. The attention module is divided into stalk and mask divisions that implement a bottom-up, top-down learning approach. The gathering of two different learning approaches into the attention module allows fast feed-forward operation and top-down attention feedback in one feed-forward process. The bottom-up feed-forward design produces low-resolution feature maps along with strong semantic facts. Whereas, top-down design produces compressed features to make a conclusion of each pixel.

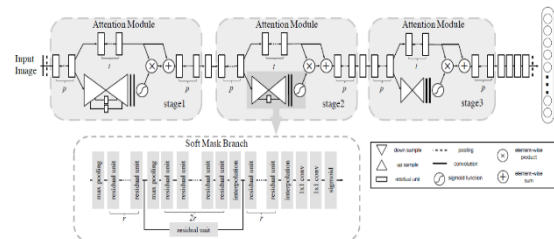


Figure 15: Architecture of Residual Attention Network [41]

Q. High-Resolution Network (HRNet)

Sun et al. [42] proposed a novel architecture, namely High-Resolution Network (HRNet) which can maintain high-resolution representations throughout the whole process. The first stage starts from a high-resolution subnetwork and constantly adds high-to-low resolution subnetworks one by one to form more stages and connect the multi-resolution subnetwork in parallel.

Table 1: Performance comparison of the recent architectures of different categories

Architecture Name & Year	Main Beneficiation	Category	parameters	Error Rate	Depth
LeNet [18], 1998	First popular CNN Architecture	SpatialExploitation	0.060 M	MNIST: 0.95	5
AlexNet [20] 2012	- Deeper and wider than LeNet - Uses RELU, Dropout & overlap pooling	Spatial Exploitation	60 M	ImageNet: 16.4	8
ZfNet [22], 2014	-visualization of intermediate layers	Spatial Exploitation	60 M	ImageNet: 11.7	8
VGG [23], 2014	-Homogeneous topology -Uses small size kernels	Spatial Exploitation	138 M	ImageNet: 7.3	19
GoogleNet [24], 2015	-Introduced block concept -Split transform and merge concept	Spatial Exploitation	4 M	ImageNet: 6.7	22
Highway Networks [26], 2015	-Introduced the concept of multi-path	Depth & Multi-path	23 M	CIFAR-10: 7.76	19
ResNet [27], 2016	-Residual learning -Identity mapping-based skip connections	Depth & Multi-path	25.6 M 1.7 M	ImageNet: 3.6 CIFAR-10: 6.43	152 110

Wide-Residual Networks [29] 2016	-Width is increased and depth is decreased	Width	36.5 M	CIFAR-10: 3.89 CIFAR-100: 18.85	28
Xception [32], 2017	-Depth-wise convolution followed by point-wise convolution	Width	22.8 M	ImageNet: 0.055	126
Residual Attention Network [41], 2017	-Introduced an attention mechanism	Attention	8.6 M	CIFAR-10: 3.90 CIFAR-100: 20.4 ImageNet: 4.8	452
ResNeXT [33], 2017	-Cardinality -Homogenous topology -Grouped convolution	Width	68.1 M	CIFAR-10: 3.58 CIFAR-100: 17.3 ImageNet: 4.4	29 101
Squeeze & Excitation Network [36], 2017	-Model interdependencies between feature-maps	Feature-map Exploitation	27.5 M	ImageNet: 2.3	152
DenseNet [30], 2017	-Cross-layer information flow	Multi-path	25.6 M 25.6M 15.3 M 15.3 M	CIFAR-10+: 3.46 CIFAR-100+: 17.18 CIFAR-10: 5.19 CIFAR-100: 19.64	190 190 250 250
PyramidalNet [31], 2017	-Increases gradually per unit	Width	116.4 M 27.0 M 27.0 M	ImageNet: 4.7 CIFAR-10: 3.48 CIFAR-100: 17.01	200 164 164
Channel Boosted CNN [38], 2018	-Boosting of original channels with additional information-rich generated artificial channels	Channel boosting	-	-	-
Competitive Squeeze and Excitation Network (CMPE-SE-WRN-28) [37], 2018	- Residual and identity mappings both are used for rescaling the feature-map	Feature-map Exploitation	36.92 M 36.90 M	CIFAR-10: 3.58 CIFAR-100: 18.47	152 152
High-Resolution Network (HRNet) [42], 2019	-High-resolution representation	Width	-	-	-

Table 2: Major strengths and drawbacks associated with the implementation of CNN architectures

Architecture Name	Strengths	Drawbacks
LeNet [18], 1998	-Exploited spatial correlation to reduce the computation and number of parameters. -Automatic learning of feature hierarchies	- Poor scaling to diverse classes of images -Large size filters -Low-level feature extraction
AlexNet [20] 2012	-Low, mid, and high-level feature extraction using large and small size filters -Give an idea of deep and wide CNN architecture -Introduced regularization in CNN -To deal with complex architectures, started parallel use of GPUs.	-Inactive neurons in the 1 st and 2 nd layer. -Aliasing artifacts in the learned feature maps due to large filter size.
ZfNet [22], 2014	-Introduced the idea of parameter tuning by visualizing the output of the intermediate layers -Reduced both the filter size and stride in the first two layers of AlexNet	- Extra information processing is required for visualization
VGG [23], 2014	-Proposed an idea of an effective receptive field -Gave the idea of simple and homogenous topology	- Use of computationally expensive fully connected layers
GoogleNet [24], 2015	-Introduced the idea of using Multiscale Filters within the layers -Gave a new idea for a split-transform-merge strategy -Reduce the number of parameters by using the bottleneck layer, global average pooling at the last layer, and Sparse Connections -Use of auxiliary classifiers to improve the convergence rate	-Tedious parameter customization due to heterogeneous topology -may lose the useful information due to bottleneck representational
Highway Networks [26], 2015	- Introduced training mechanism for deep network -Used auxiliary connections in addition to direct connections -Mitigates the limitations of deep networks by introducing cross-layer connectivity	- Parametric gating mechanism, difficult to implement - Gates are data dependent and thus become parameter expensive
ResNet [27], 2016	-Use of identity-based skip connections to enable cross-layer connectivity -Information flow gates are data-independent and parameter-free -Can easily pass the signal in both directions, forward and backward	-Many layers may contribute very little or no information -Relearning redundant feature maps may happen
Wide-Residual Networks [29] 2016	-Shows the effectiveness of parallel use of transformations by increasing the width of ResNet and decreasing its depth - Enables feature reuse -Have shown that dropouts between the convolutional layer are more effective	-Over fitting may occur - More parameters than thin deep networks
Xception [32], 2017	-Introduce the concept that learning across 2D followed by 1D is easier than learning filters in 3D space - Depth-wise separable convolution is introduced - Use of cardinality to learn good abstraction	-High computational cost
Residual Attention Network [41], 2017	-Generates attention aware feature-maps -Easy to scale up due to residual learning -Provides different representations of the focused patches -Add soft weights on features using bottom-up top-down feed-forward attention	-Complex model

ResNeXT [33], 2017	-Introduced cardinality to avail diverse transformations at each layer -Easy parameter customization due to homogenous topology -Uses grouped convolution	-High computational cost
Squeeze & Excitation Network [36], 2017	-It is a block-based concept -Introduced a generic block that can be added easily to any CNN model due to its simplicity -Squeezes less important features and vice versa	-In ResNet, only considers the residual information for determining the weight of each channel
DenseNet [30], 2017	-Introduced depth or cross-layer dimension -Ensures maximum data flow between the layers in the network - Avoid relearning redundant feature-maps -Low and high-level features are accessible to decision layers	-Large increase in parameters due to an increase in the number of feature maps at each layer
PyramidalNet [31], 2017	-Introduces the idea of increasing the width gradually per unit -Avoid rapid information loss -Covers all possible locations instead of maintaining the same dimension till the last unit	-High spatial and time complexity -May become quite complex if layers are substantially
Channel Boosted CNN [38], 2018	-It boosts the number of input channels for improving the representational capacity of the network -Inductive Transfer Learning is used in a novel way to build a boosted input representation for CNN	-Increase in computational load may happen due to the generation of auxiliary channels
Competitive Squeeze and Excitation Network (CMPE-SE-WRN-28) [37], 2018	-Uses feature-map-wise statistics from both residual and identity mapping-based features -Makes a competition between residual identity feature maps	-Doesn't support the concept of attention
High-Resolution Network (HRNet) [42], 2019	-able to maintain high resolution -repeated multi-scale fusion to boost high-resolution representation	-

3. Conclusion and Future Scope

This paper reviews the evolution of the architectures of the CNN, especially working out the outline of the processing units, and therefore put forward the structure for the recent architecture of the CNN. We cover the history of CNN from its starting to till date and also mention the standards which are used for improving the ability latest CNN. Table 1 views the performance criteria of each CNN model, and showed that how the error rate is decreased by using the optimization techniques. The hyper parameters such as activation function, size of filters, loss functions, optimizers, number of neurons per layer, etc. of deep CNNs are one of the future tasks for researchers in developing genetic algorithms. The approach of pipeline parallelism which overwhelms the hardware restrictions can be utilized to increase the training of CNN, without tuning hyper parameters, which is the future research area. The attention mechanism is not only apprehended the facts from images but also keeps its background connection with other parts of the image. In the future, in the belated phases of learning, the analysis may be achieved out in an angle that assures the spatial significance of objects along with their perceptive features.

References

- [1] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, Nov. 1998.
- [2] Ciresan, D., U. Meier, J. Masci, and J. Schmidhuber. "Multi-column deep neural network for traffic sign classification. Neural Networks," In *The International Joint Conference on Neural Network, IDSIA-USI-SUPSI| Galleria*, vol. 2. 2012.
- [3] Deng, Li, and Dong Yu, "Deep learning: methods and applications," *Foundations and trends® in signal processing* 7, no. 3–4, pp. 197-387, 2014.
- [4] LeCun, Y., Kavukcuoglu, K. and Farabet, C., "Convolutional networks and applications in vision," In *Proceedings of 2010 IEEE international symposium on circuits and systems, IEEE*, pp. 253-256), May 2010.
- [5] Scherer, D., Müller, A. and Behnke, S., "Evaluation of pooling operations in convolutional architectures for object recognition," In *International conference on artificial neural networks*, Springer, Berlin, Heidelberg, pp. 92-101, Sep. 2010.
- [6] Najafabadi MM, Villanustre F, Khoshgoftaar TM, et al, "Deep learning applications and challenges in big data analytics," *J Big Data* vol 2, pp.1–21, 2015.
- [7] Krizhevsky, A., Sutskever, I. and Hinton, G.E., "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097-1105, 2012.
- [8] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*. Sep. 4 2014.
- [9] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z., "Rethinking the inception architecture for computer vision," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818-2826, 2016.
- [10] He K., Zhang X., Ren S. and Sun, J., "Deep residual learning for image recognition," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778, 2016.
- [11] Zagoruyko, S. and Komodakis, N., "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.
- [12] Xie, S., Girhick, R., Dollár, P., Tu, Z. and He, K., "Aggregated residual transformations for deep neural networks," In *Proceedings of the IEEE conference on*

- computer vision and pattern recognition, pp. 1492-1500), 2017.
- [13] Han, D., Kim, J. and Kim, J., "Deep pyramidal residual networks," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5927-5935, 2017.
- [14] Chollet, F., "Xception: Deep learning with depthwise separable convolutions," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251-1258, 2017.
- [15] O'donovan, P., Leahy, K., Bruton, K. and O'Sullivan, D.T., "Big data in manufacturing: a systematic mapping study," *Journal of Big Data*, 2(1), pp.1-22, 2015.
- [16] Alom, M.Z., Taha, T.M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M.S., Hasan, M., Van Essen, B.C., Awwal, A.A. and Asari, V.K., "A state-of-the-art survey on deep learning theory and architectures," *Electronics*, 8(3), p.292, 2019.
- [17] Zhang, Q., Zhang, M., Chen, T., Sun, Z., Ma, Y. and Yu, B., "Recent advances in convolutional neural network acceleration" *Neurocomputing*, vol. 323, pp.37-51, Jan 2019.
- [18] LeCun, Y. and Bottou, L., YB and Haffner, P., "Gradient-based learning applied to document recognition," *Proc. IEEE*, 1998.
- [19] Gardner, M.W. and Dorling, S.R., "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences," *Atmospheric environment*, 32(14-15), pp.2627-2636, 1998.
- [20] Krizhevsky, A., Sutskever, I. and Hinton, G.E., "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.
- [21] Hochreiter, S., "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02), pp.107-116, 1998.
- [22] Zeiler, M.D. and Fergus, R., "Visualizing and understanding convolutional networks," CoRR, abs/1311.2901. *arXiv preprint arXiv:1311.2901*, 2013.
- [23] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [24] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., "Going deeper with convolutions," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1-9, 2015.
- [25] Huang, G., Sun, Y., Liu, Z., Sedra, D. and Weinberger, K.Q., "Deep networks with stochastic depth," In *European conference on computer vision* Springer, Cham., pp. 646-661. October 2016
- [26] Srivastava, R.K., Greff, K. and Schmidhuber, J., "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.
- [27] He, K., Zhang, X., Ren, S. and Sun, J., "Deep residual learning for image recognition," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778, 2016.
- [28] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L., "September. Microsoft coco: Common objects in context," In *European conference on computer vision* Springer, Cham, pp. 740-755, 2014.
- [29] Zagoruyko, S. and Komodakis, N., "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.
- [30] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ , "Densely connected convolutional networks," *Proc - 30th IEEE Conf Comput Vis Pattern Recognition, CVPR 2017*, pp. 2261-2269, 2017.
- [31] Han, D., Kim, J. and Kim, J., "Deep pyramidal residual networks," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5927-5935, 2017.
- [32] Chollet, F., "Xception: Deep learning with depthwise separable convolutions," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251-1258, 2017.
- [33] Xie, S., Girshick, R., Dollár, P., Tu, Z. and He, K., "Aggregated residual transformations for deep neural networks," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492-1500, 2017.
- [34] Liu, X., Chi, M., Zhang, Y. and Qin, Y., "Classifying high-resolution remote sensing images by fine-tuned VGG deep networks," In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium* IEEE, pp. 7137-7140, 2018.
- [35] Sharma, A. and Muttou, S.K., "Spatial image steganalysis based on resnext," In *2018 IEEE 18th International Conference on Communication Technology (ICCT)* IEEE, pp. 1213-1216, October 2018.
- [36] Hu, J., Shen, L. and Sun, G., "Squeeze-and-excitation networks," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132-7141, 2018.
- [37] Hu, Y., Wen, G., Luo, M., Dai, D., Ma, J. and Yu, Z., "Competitive inner-imaging squeeze and excitation for residual network," *arXiv preprint arXiv:1807.08920*, 2018.
- [38] Khan, A., Sohail, A. and Ali, A., "A new channel boosted convolutional neural network using transfer learning," *arXiv preprint arXiv:1804.08528*, 2018.
- [39] Hamel, P. and Eck, D., "Learning features from music audio with deep belief networks," In *ISMIR*, Vol. 10, pp. 339-344, August 2010.
- [40] Vincent, P., Laroche, H., Bengio, Y. and Manzagol, P.A., "Extracting and composing robust features with denoising autoencoders," In *Proceedings of the 25th international conference on Machine learning*, pp. 1096-1103, July 2008.
- [41] Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X. and Tang, X., "Residual attention network for image classification," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156-3164, 2017.
- [42] Sun, K., Xiao, B., Liu, D. and Wang, J., "Deep high-resolution representation learning for human pose estimation," In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5693-5703, 2019.