A Brief Survey on Deep Learning for Computer Vision: Challenges and Future Trends

Ms. Yugandhara A. Thakare[#1], Ms. Ankita V. Pande[*2]

[#] *Department of Computer Science & Engineering, Sipna COET,Amravati*

[*] *Department of Computer Science & Engineering, Sipna COET,Amravati*

[1]yugathakare@gmail.com

[2]ankita.pande2@gmail.com

Abstract

Deep learning is an evolving area of research which brought revolutionary advances in machine learning and computer vision. Deep Learning has broken the limits of what was possible in the domain of Digital Image Processing. Recently, most of the artificial intelligence problems have been solved by using deep learning algorithms. In the recent years deep learning has attracted much attention because of its state-of-the-art performance in different regions like computer vision, natural language processing, object perception, speech recognition etc. This paper presents a brief survey of deep learning methods and recent developments in computer vision, applications in different vision task. Finally, the paper summarizes the future trends and challenges in deep neural networks.

***Keywords*:** Deep learning, Computer vision, machine learning

1. INTRODUCTION

Deep Learning is a subset of machine learning. Deep Learning is based largely on Artificial Neural Networks (ANNs), a computing paradigm inspired by the functioning of the human brain. Like the human brain, it is composed of many computing cells or 'neurons' that each performs a simple operation and interacts with each other to make a decision [1]. Deep learning makes use of two strategies like supervised or unsupervised. These strategies will automatically learn hierarchical representations in deep architectures for classification task. The objective of this is to determine other abstract features in the higher levels of the representation, by using neural networks. Deep learning has been widely applied in various traditional artificial intelligence domains, like natural language processing [4], transfer learning [2], [3], computer vision [5], [6] and etc. Deep learning becomes a boom topic today because of vividly increased in processing abilities, low cost computing hardware and advancement in machine learning algorithm. [7]

2. DEEP LEARNING ARCHITECTURE AND RECENT DEVELOPMENTS

There are numerous deep architectures available in the literature and growing by the day. A fair comparison of these architectures is difficult jobs given different architectures have different advantages based on the application and the characteristics of the data involved. For example in computer vision, Convolutional Neural Networks and in sequence and time series modeling Recurrent Neural Networks are preferred.

2.1 Deep feed-forward networks:

Deep Feed-forward Neural network, the most basic deep architecture with only the connections between the nodes moves forward. Basically, when a multilayer neural network contains multiple numbers of hidden layers, we call it deep neural network [8]. An example of Deep Feed-Forward Network with n hidden layers is provided in Fig.1. Multiple hidden layers help in modeling complex nonlinear relation more efficiently compared to the shallow architecture. A complex function can be modeled with less number of computational units compared to a similarly performing shallow network due to the hierarchical learning possible with the multiple levels of nonlinearity [9]. Due to the simplicity of architecture and the training in this model, it is always a popular architecture among researchers and practitioners in almost all the domains of engineering. Back propagation using gradient descent [10] is the most common learning algorithm used to train this model.

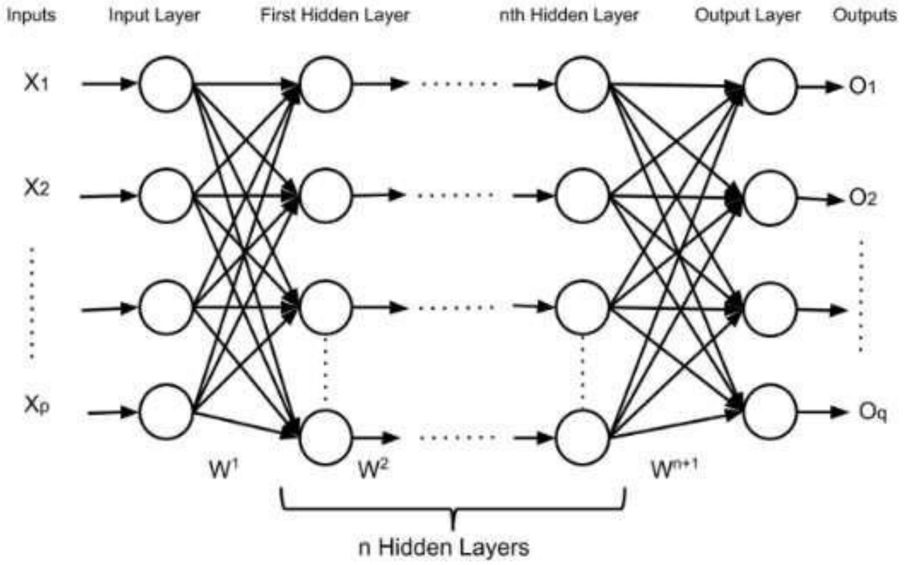

Fig.1 Deep feed forward neural network with n hidden layers, p input units and q output units with weights w.[36]

2.2 Restricted Boltzmann machines:

Restricted Boltzmann Machines (RBM) [11] can be interpreted as stochastic neural networks. An RBM is a popular deep learning framework due to its ability to learn the input probability distribution in supervised as well as unsupervised manner. Restricted Boltzmann Machine is a variation of Boltzmann machines with the restriction in the intra-layer connection be-tween the units, hence the term restricted. It is an undirected graphical model containing two layers, visible and hidden, forming a bipartite graph. Different variations of RBMs have been introduced in the literature in terms of improving the learning algorithms, provided the task. Temporal RBMs [12] and conditional RBMs [13] are applied to model multivariate time series data and to generate motion captures. Each node in RBM is a computational unit that processes the input it receives to make stochastic decisions whether to transmit that input or not. An RBM with m visible and n hidden units is provided in Fig. 2.

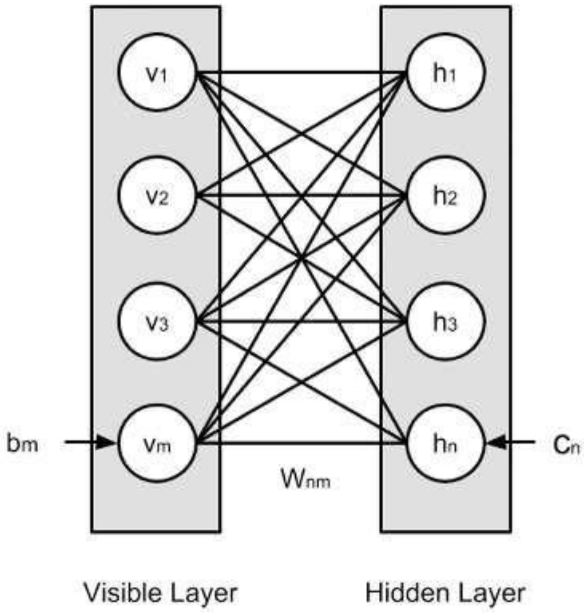

Fig.2 RBM with m visible units and n hidden units [36]

2.3 Deep belief networks:

Deep belief network (DBN) is a generative graphical model composed of multiple layers of latent variables. The latent variables are typically binary, can represent the hidden features present in the input observations. The connection between the top two layers of DBN is undirected like an RBM model, hence a DBN with 1 hidden layer is just an RBM. The other connections in DBN except last are directed graphs towards the input layer. DBN is a generative model; hence to generate a sample from DBN follows a top-down approach.

2.4 Auto encoder:

An auto encoder is a three-layer neural network, as shown in Fig. 3, which tries to reconstruct its input at its output layer. Hence, the output layer of an auto encoder contains the same number of units as the input layer. The hidden layer typically contains less number of neurons compared to the visible layer and tries to encode or represent the input in a more compact form. It shares the same idea as an RBM, but it typically uses deterministic distributions instead of stochastic units with particular distributions as is the case with RBMs. Like feed forward neural networks, an auto encoder is typically trained using the back propagation algorithm. The training consists of two phases: encoding and decoding. In the encoding phase, the model tries to encode the input into some hidden representation using the weight metrics of the lower half layer, and in the decoding phase, it tries to reconstruct the same input from the encoding representation using the metrics of the upper half layer. Several variations of auto encoders are introduced with differing properties and implementations to learn more efficient representation of the data under consideration. One of the popular variations of an auto encoder that is robust to input variations is the de noising auto encoder [14, 15, and 16]. The model can be used for compact representations of input with the number of hidden layers less than the input layer.

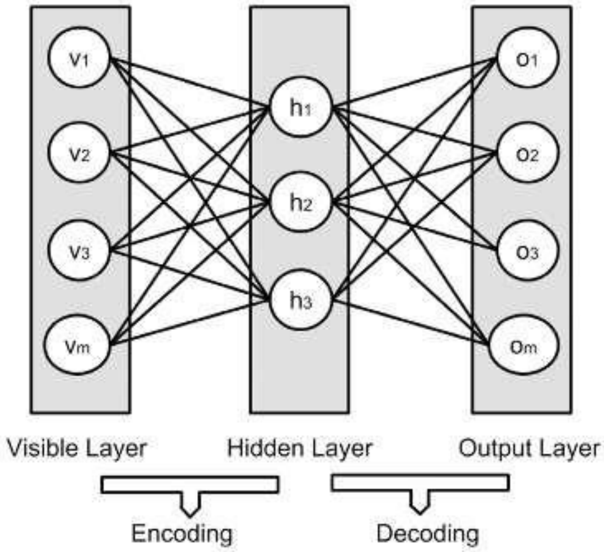

Fig. 3. Autoencoder with 3 neurons in hidden layer[36]

2.5 Convolutional neural networks:

Convolutional Neural Networks (CNN) is a class of neural networks inspired from the human visual system. The Convolutional Neural Networks (CNN) is one of the most notable deep learning approaches. Generally, a CNN consists of three main neural layers, which are Convolutional layers, pooling layers and fully connected layers.
Different kinds of layers play different roles.

2.5.1. Types of layers

Usually, a CNN is a hierarchical neural network whose convolutional layers alternate with pooling layers, followed by some fully connected layers. In this, we will present the functions and brief review of recent advances that have appeared in research on those layers.

2.5.1.1. Convolutional layers. In the convolutional layers, a CNN exploits several kernels to convolve the entire image as well as the intermediate feature maps, generating various feature maps etc.

2.5.1.2. Pooling layers. Normally, a pooling layer follows a convolutional layer, and can be used to decrease the dimensions of feature maps and network parameters. Similar to convolutional layers, pooling layers are also translation invariant, because their computations take neighbouring pixels into account. The most commonly used strategies are Average pooling and max pooling.

2.5.1.3 Fully-connected layers perform like a traditional neural network and contain about 90% of the parameters in a CNN. It enables us to feed forward the neural network into a vector with a predefined length.

The basic architecture of CNN is shown in Fig. 7, which contains multiple convolutions and pooling layers with a fully connected layer at the end. Convolution layers extract important features from the input image in consideration of the spatial relationship between the input pixels whereas pooling layers reduce the dimensionality of the feature map while retaining the feature information [17]. The fully connected layer connects the network with the discriminative layer (output layer), which ultimately provides the desired output. CNNs are particularly useful in extracting image descriptors using latent spatial information. An image has several characteristics like edges, contours, strokes, textures, gradients, orientation, and color. A CNN breaks down an image in terms of these types of simple properties and learns them as representations in different layers [18]. CNNs are popular in computer vision tasks such as image detection [19, 20], image segmentation [21, 22], image classification [23] and image super-resolution reconstruction [24, 25]. Several CNN architectures have been developed considering real-time application requirements while simultaneously meeting high accuracy thresholds. R-CNN (Region-based CNN) [21] and YOLO (You Only Look Once) [26] are examples of such recent architectures. The naive approaches of CNN [27] are computationally very expensive as it considers a massive number of region proposals to locate an object within an image. R-CNN however, is a region-based CNN, that overcomes the limitation of naive CNN by selecting the regions of interest (ROI) with a selective search and limits the proposal regions to 2000 [21].For the application of R-CNN to real-time processing, the authors later proposed Fast R-CNN [28]. Apart from these CNN architectures, there are several variations of existing classical ones such as AlexNet [31], VGGNet [29], ResNet[30] etc.

2.6 Recurrent neural networks:

RNNs are form of feed-forward networks spanning adjacent time steps such that at any time instant a node of the network takes the current data input as well as the hidden node values capturing information of previous time steps.Fig.9 shows a Recurrent Neural Network architecture. During the back propagation of errors across multiple time steps the problem of vanishing and exploding gradients take place which can be avoided by Long Short Term Memory (LSTM) Networks. The amount of information to be retained from previous time steps is controlled by a sigmoid layer known as ‘forget’ gate whereas the sigmoid activated ‘input gate’ decides upon the new information to be stored in the cell followed by a hyperbolic tangent activated layer to produce new candidate values which is updated taking forget gate coefficient weighted old state’s candidate value. Finally the output is produced controlled by output gate and hyperbolic tangent activated candidate value of the state.

2.7 Generative adversarial networks:

This is a novel framework for Generative Adversarial Nets with simultaneous training of a generative and a discriminative model. Fig.4. Shows an architecture of a Generative Adversarial Network.

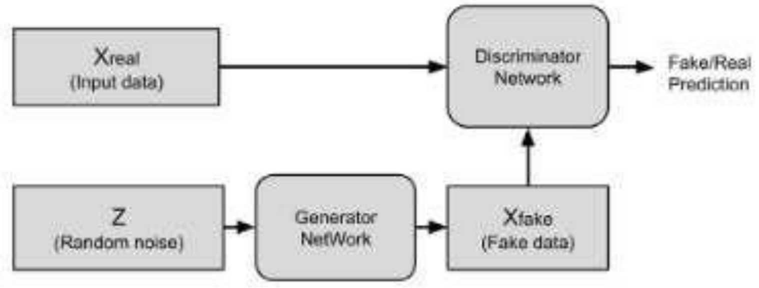

Fig. 4. A generative adversarial network architecture.[36]

Wu et al. [32] proposed a 3D Generative Adversarial Net-work (3DGAN) for three dimensional object generation using volumetric convolutional networks with a mapping from probabilistic space of lower dimension to three dimensional object space so that the 3D object can be sampled or explored without any reference image. As a result high quality 3D objects can be generated employing efficient shape descriptor learnt in an unsupervised manner by the adversarial discriminator. Vondricket al. [33] came up with video recognition/classification and video generation/prediction model using Generative Adversarial Network (GAN) with separation of foreground from background employing spatiotemporal convolutional architecture.

3 APPLICATIONS OF DEEP LEARNING IN COMPUTER VISION

Deep learning reveals strong advantages in the feature extraction. Deep learning has been broadly used in the field of computer vision and among others.

3.1 Image search

One of the areas wherever computer vision has created vast progress is image classification and object detection. A neural network trained on enough labelled knowledge will be able to detect and highlight a wide range of objects with impressive accuracy. When we upload photos in Google Photos, it uses its computer vision algorithms to annotate them with content information regarding scenes, objects, and persons. We are able to then search images based on this information.

3.2 Medical image processing

Before deep learning, creating computer vision algorithms that could process medical images needed intensive efforts from software engineers and subject matter experts. They had to collaborate to develop code that extracted relevant features from radiology images and then examine them for diagnosis. Deep learning algorithms provide end-to-end solutions that make the process very easier. The engineers produce the correct neural network structure and then train it on x-rays, MRI images or CT scans annotated with the outcomes. The neural network then finds the relevant features associated with each outcome and can then diagnose future images with impressive accuracy. Computer vision has found its way into many areas of medicine, including cancer detection and prediction, radiology, diabetic retinopathy.

3.3 Self-driving cars

Cars that can route roads without human drivers have been one of the longest standing dreams and biggest challenges of the AI community. Today, we're still very far from having self-driving cars that can navigate any road on several lighting and weather circumstances. But we have made a lot of advancement thanks to developments in deep neural networks. The biggest challenge of making self-driving cars is to enabling them to make sense of their surroundings.

3.4 Autonomous weapons

Computer vision can also give senses to weapons. Military drones can use AI algorithms to recognize objects and preference out targets.

4. CHALLENGES IN DEEP LEARNING

1. With increasing readiness of data as well as powerful and distributed processing units, Deep Learning architectures can be effectively applied to major industrial problems. However, deep learning is traditionally big data driven and lacks efficiency to learn abstractions through clear verbal definitions [34] if not trained with huge training sample sizes i.e. challenge occur with scarcity of data. One shot learning [35] is also bringing in new avenues to learn from very few training examples which have already started showing progress in language processing and image classification tasks. More generalized techniques are being developed in this domain to make DL models learn from sparse or fewer data representations.
2. As of rendered a data explosion in recent times and while more data equates to more training examples, a vital issue for machine learning practitioners is to separate the good from the bad. While data may be sourced in structured or semi-structured ways, filtering out bad instances and/or instances which are uncorrelated to the learning objective still remains a key challenge requiring further research. So, data preprocessing is an overhead task.
3. Regardless of the accomplishment of Deep learning techniques across several domains, most of the neural architectures used so far are specifically trained to learn and perform a particular job and not a variety of disparate jobs at once. One of the major research challenges is to create general multi-purpose architectures that can encapsulate the learning from different domains

5. CONCLUSION

In this paper we have presented a broad review of deep learning. The state-of-the-art architecture are discussed and analyzed in detail with different applications in the computer vision domain. Despite the promising results reported so far, there is significant room for further advances which has been discussed in this paper. This paper describes challenges and summarizes the new trends in deep neural networks.

REFERENCES

[1] O'Mahony, N., Murphy, T., Panduru, K., et al.: Adaptive process control and sensor fusionfor process analytical technology. In: 2016 27th Irish Signals and Systems Conference (ISSC), pp. 1–6. IEEE (2016)

[2] D.C. Ciresan, U. Meier, J. Schmidhuber, Transfer learning for Latin and Chinese characters with deep neural networks, in: Proceedings of the IJCNN, 2012.

[3] J.S.J. Ren, L. Xu, On vectorization of deep convolutional neural networks for vision tasks, in: Proceedings of the AAAI, 2015.

44 Y. Guo et al. / Neuro computing 187 (2016) 27–48[4] T. Mikolov, I. Sutskever, K. Chen, et al., Distributed representations of words and phrases and their compositionality, in: Proceedings of the NIPS, 2013.

[5] D. Ciresan, U. Meier, J. Schmidhuber, Multi-column deep neural networks for image classifification, in: Proceedings of the CVPR, 2012.

[6] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classifification with deep convolutional neural networks, in: Proceedings of the NIPS, 2012.

[7] L. Deng, A tutorial survey of architectures, algorithms, and applications for deep learning, APSIPA Trans. Signal Inf. Process. 3 (2014) e2.

[8] L. Deng, D. Yu, et al., Deep learning: methods and applications, Found. Trends Signal Process. 7 (3–4) (2014) 197–387.

[9] Y. Bengio, et al., Learning deep architectures for ai, Found. Trends Mach. Learn. 2 (1) (2009) 1–127.

[10] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning internal representations by error propagation, Tech. rep., California Univ San Diego La JollaInst for Cognitive Science, 1985.

[11] A. Fischer, C. Igel, An introduction to restricted Boltzmann machines,in: Iberoamerican Congress on Pattern Recognition, Springer, 2012, pp.14–36

[12] I. Sutskever, G. Hinton, Learning multilevel distributed representations for high-dimensional sequences, in: Artificial Intelligence and Statistics,2007, pp. 548–555.

[13] G.W. Taylor, G.E. Hinton, S.T. Roweis, Modeling human motion using binary latent variables, in: Advances in Neural Information Processing Systems, 2007, pp. 1345–1352

[14] Y. Bengio, N. Boulanger-Lewandowski, R. Pascanu, Advances in optimizing recurrent networks, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013, pp. 8624–8628.

[15] Y. Bengio, E. Laufer, G. Alain, J. Yosinski, Deep generative stochastic networks trainable by back prop, in: International Conference on Machine Learning, 2014, pp. 226–234.

[16] Y. Bengio, Deep learning of representations: Looking forward, in: International Conference on Statistical Language and Speech Processing, Springer,2013, pp. 1–37.

[17]I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016,http://www.deeplearningbook.org.

[18]Y. LeCun, Y. Bengio, G. E. Hinton, Deep learning, Nature 521 (7553)(2015)436–444.doi:10.1038/nature14539.URLhttps://doi.org/10.1038/nature1453

[19] D. Tomè, F. Monti, L. Baroffio, L. Bondi, M. Tagliasacchi, S. Tubaro, Deep convolutional neural networks for pedestrian detection, Signal Process.:Image Commun. 47 (2016) 482–489.

[20] Z.-Q. Zhao, P. Zheng, S.-t. Xu, X. Wu, Object detection with deep learning: A review, IEEE Trans. Neural Netw. Learn. Syst. (2019).

[21] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings ofthe IEEE Conference on Computer Vision and Pattern Recognition, 2014,pp. 580–587.
[22]K. He, G.Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2961–2969.
[23] Y. Wang, Z. Wang, A survey of recent work on fine-grained image classification techniques, J. Vis. Commun. Image Represent. 59 (2019)210–214.
[24] M.D. Zeiler, D. Krishnan, G.W. Taylor, R. Fergus, De convolutional net-works, in: 2010 IEEE Computer Society Conference on Computer Visionand Pattern Recognition, IEEE, 2010, pp. 2528–2535. [108] H. Noh, S. Hong, B. Han, Learning de convolution network for semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1520–1528.
[25] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.
[26] S. Fidler, R. Mottaghi, A. Yuille, R. Urtasun, Bottom-up segmentation for top-down detection, in: Proceedings of the IEEE Conference on ComputerVision and Pattern Recognition, 2013, pp. 3294–3301.
[27] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448. [117] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, in: Proceedings of the IEEE, 1998, pp.2278–2324.
[28] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, CoRR abs/1409.1556.[119] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR),2015,pp.1–9,http://dx.doi.org/10.1109/CVPR.2015.7298594.
[29] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR),2016,pp.770–778, http://dx.doi.org/10.1109/CVPR.2016.90.
[30] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional net-works, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham, 2014, pp.818–833.
[31] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, in:NIPS'12, Curran Associates Inc., USA, 2012, pp. 1097–1105, URL http://dl.acm.org/citation.cfm?id=2999134.2999257.
[32] J. Wu, C. Zhang, T. Xue, W.T. Freeman, J.B. Tenenbaum, Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling, in: NIPS, 2016.
[33] C. Vondrick, H. Pirsiavash, A. Torralba, Generating videos with scene dynamics, in: NIPS, 2016.
[34] G. Marcus, Deep learning: A critical appraisal, CoRR abs/1801.00631.
[35]O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, D. Wierstra, Match-ing networks for one shot learning, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, CurranAssociates Inc., USA, 2016, pp. 3637–3645. URLhttp://dl.acm.org/citation.cfm?id=3157382.3157504
[36] S. Sengupta, S. Basak, P. Saikia et al., A review of deep learning with special emphasis on architectures, applications and recent trends, Knowledge-Based Systems(2020), doi: https://doi.org/10.1016/j.knosys.2020.105596.