Deep Learning Architecture : A Literature Survey

[1]Srivalli Devi S, [2]Dr.A.Geetha
[1]Ph.D Research Scholar, [2]Assistant Professor & Head
[1]PG & Research Department of Computer Science
[1]Chikkanna Government Arts College, Tiruppur, India

Abstract : Deep learning is a sub field of machine learning. Learning can be of supervised, semi-supervised and unsupervised. There are different types of architectures for deep learning . In this paper we are giving an overview of different architectures that are widely used and their application area. Deep learning is applied in many areas such as image processing, speech recognition, data mining, natural language processing, social network filtering, machine translation, bioinformatics and drug design.

IndexTerms **- Deep learning ;deep learning architecture; machine learning**

I. INTRODUCTION

There are numerous architectures and algorithms for deep learning. Deep Learning algorithms consists of varied models in contrast to machine learning algorithm. This is because of the flexibility that neural network provides when building a full fledged end-to-end model[1]. Deep learning architectures as stated by [1] are Recurrent neural networks, Long short-term memory /Gated Recurrent Unit networks, Convolutional neural networks, Deep belief networks, Deep stacking networks.

In this paper we had done a survey on the following deep learning architectures and their applications. Reference[2] states the following architectures are top ten architectures that a data scientist should know: AlexNet, VGG Net, GoogleNet, ResNet, ResNeXT, RCNN(Region based CNN), YOLO(You Only Look Once), SqueezeNet, SegNet, GAN(Generative Adversarial Network).

II. DEEP LEARNING VS. MACHINE LEARNING

Reference[3] states the difference between deep and machine learning as machine learning analyzes data and crunches numbers, learns from it, and uses that to make a prediction/truth/determination depending on the scenario. The machine is being trained, or training itself, on how to perform a task correctly after learning from all the data it has analyzed. It's building its' own logic and solutions. Machine learning can be done with a bunch of different algorithms like: Random Forest & Decision Tree: A collection or ensemble of simple tree predictors, each capable of producing a response, like Netflix suggesting movies based off of your star ratings.

Linear Regression: Predicts the value of a categorical outcome with limitless outcomes, like figuring out how much you can sell a car for based on the market.

Logistic Regression: Predicts the value of a categorical outcome with a limited number of possible values, like figuring out if you can sell a car for a certain cost.

Classification: Puts data into different groups, like filing documents or emails.

Naive Bayes: A family of algorithms that all share a common principle, that every feature being classified is independent of the value of any other feature, like predicting happiness in photos of children.

There are also two types of machine learning algorithms, supervised learning and unsupervised learning.

Supervised learning requires a human to input the data and the solution, but allows the machine to figure out the relationship between the two. This is extremely helpful in mathematical situations.

Unsupervised is putting in random numbers/data for a certain situation and asking the computer to find a relationship and solution. It's kind of like shooting a target in the dark, we won't know what we hit until we put the lights on.

So, machine learning eliminates the need for someone to continuously code or analyze the data themselves to solve a solution or present a logic.

The prime difference between machine and deep learning is deep learning crunches more data than machine learning. So, if we have a little bit of data, machine learning is the way to go but if you're drowning in data deep learning is our answer [3].Deep learning algorithms are powerful and they need a lot of data to give you the best solution/outcome. Deep learning algorithms need powerful machines, machine learning algorithms don't.

Reference [3] states that deep learning algorithms do complicated things, like matrix multiplication, which require a graphic processing unit (GPUs). They also try to learn high-level features, so in the case of facial recognition the algorithm will get the image pretty close to the raw version in replication whereas machine learning's images would be blurry. Another powerful feature, it forms an end-to-end solution instead of breaking a problem and solution down into parts. It is composed of the machine learning algorithms, neural networks, and AI.

III. ALEXNET

AlexNet is the first deep architecture which was introduced by one of the pioneers in deep learning – Geoffrey Hinton and his colleagues. It is a simple yet powerful network architecture. When broken down, AlexNet seems like a simple architecture with convolutional and pooling layers one on top of the other, followed by fully connected layers at the top.The things which set apart this model is the scale at which it performs the task and the use of GPU for training. In 1980s, CPU was used for training a neural network. Whereas AlexNet speeds up the training by ten times just by the use of GPU[4]. Here is a representation of the architecture as proposed by the authors.

AlexNet is still used as a starting point for applying deep neural networks for all the tasks, whether it be computer vision or speech recognition. The below Fig.1 is the architecture for AlexNet

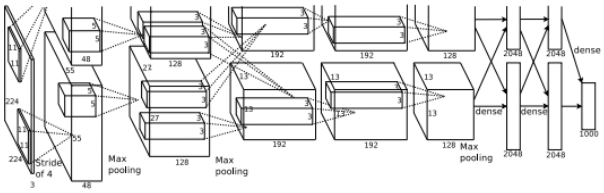

Figure 1.CNN architecture, Source: Adapted from [4]

IV. VGGNET

The VGG Network was introduced by the researchers at Visual Graphics Group at Oxford (hence the name VGG). This network is specially characterized by its pyramidal shape, where the bottom layers which are closer to the image are wide, whereas the top layers are deep. As the figure 2 depicts, VGG contains subsequent convolutional layers followed by pooling layers. The pooling layers are responsible for making the layers narrower. In their paper, they proposed multiple such types of networks, with change in deepness of the architecture.The advantages of VGG are: It is a very good architecture for benchmarking on a particular task. Also, pre-trained networks for VGG are available freely on the internet, so it is commonly used for various applications. On the other hand, its main disadvantage is that it is very slow to train if trained from scratch. Even on a decent GPU, it would take more than a week to get it to work[1][5].

In 2014, 16 and 19 layer networks were considered very deep (although we now have the ResNet architecture which can be successfully trained at depths of 50-200 for ImageNet and over 1,000 for CIFAR-10).Simonyan and Zisserman found training VGG16 and VGG19 challenging (specifically regarding convergence on the deeper networks), so in order to make training easier, they first trained smaller versions of VGG with less weight layers (columns A and C) first[5]. Figure 2 shows the Table 1 of Very Deep Convolutional Networks for Large Scale Image Recognition, Simonyan and Zisserman (2014)

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Figure 2.shows Table1 of Very Deep Convolutional Networks for Large Scale Image Recognition, Simonyan and Zisserman (2014).

V. RESNET(RESIDUAL NETWORK

Unlike traditional sequential network architectures such as AlexNet, OverFeat, and VGG, ResNet is instead a form of "exotic architecture" that relies on micro-architecture modules (also called "network-in-network architectures").The term micro-architecture refers to the set of "building blocks" used to construct the network. A collection of micro-architecture building blocks (along with your standard CONV, POOL, etc. layers) leads to the macro-architecture (i.e,. the end network itself).First introduced by He et al. in their 2015 paper, Deep Residual Learning for Image Recognition, the ResNet architecture has become a seminal work, demonstrating that extremely deep networks can be trained using standard SGD (and a reasonable initialization function) through the use of residual modules.

Further accuracy can be obtained by updating the residual module to use identity mappings, as demonstrated in [6].Architecture is shown in the below Fig.3

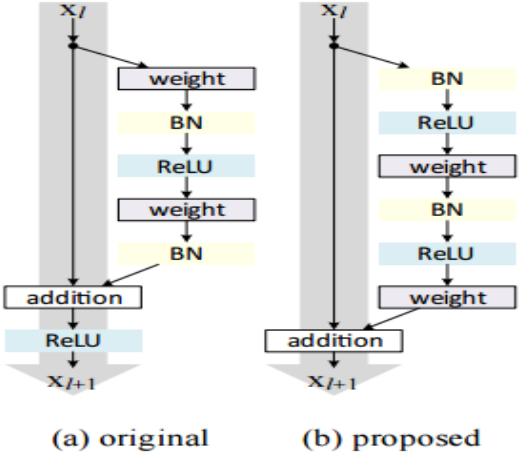

Figure 3. (Left) The original residual module. (Right) The updated residual module using pre-activation . Adapted from [6]

VI. GOOGLENET

GoogleNet (or Inception Network) is a class of architecture designed by researchers at Google. GoogleNet was the winner of ImageNet 2014, where it proved to be a powerful model.In this architecture, along with going deeper (it contains 22 layers in comparison to VGG which had 19 layers), the researchers also made a novel approach called the Inception module. In a single layer, multiple types of "feature extractors" are present. This indirectly helps the network perform better, as the network at training itself has many options to choose from when solving the task. It can either choose to convolve the input, or to pool it directly.

The final architecture contains multiple of these inception modules stacked one over the other. Even the training is slightly different in GoogleNet, as most of the topmost layers have their own output layer. This nuance helps the model converge faster, as there is a joint training as well as parallel training for the layers itself.

The advantages of GoogleNet are :

GoogleNet trains faster than VGG.

Size of a pre-trained GoogleNet is comparatively smaller than VGG. A VGG model can have >500 MBs, whereas GoogleNet has a size of only 96 MB

GoogleNet does not have an immediate disadvantage per se, but further changes in the architecture are proposed, which make the model perform better. One such change is termed as an Xception Network, in which the limit of divergence of inception module.

VII. RESNEXT

ResNeXT extends the VGG-style strategy of repeating layers of the same shape, which is helpful for isolating a few factors and extending to any large number of transformations. We set the individual transformation Ti to be the bottleneck shaped architecture , as illustrated in Fig. 4 (right). In this case, the first 1×1 layer in each Ti produces the low dimensional embedding[8]. The following Fig.3 shows the architecture.

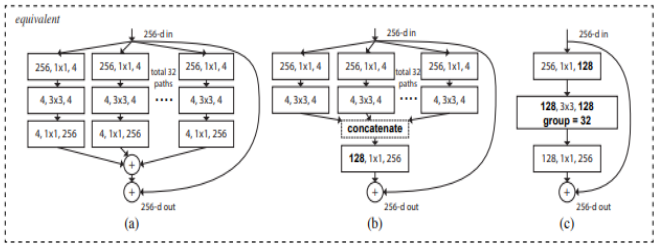

Figure 4. Figure 3 of [8]

VIII. Region Based CNN (RCNN)

Region Based CNN architecture is said to be the most influential of all the deep learning architectures that have been applied to object detection problem. To solve detection problem, what RCNN does is to attempt to draw a bounding box over all the objects present in the image, and then recognize what object is in the image. The following Fig.5 shows the architecture for RCNN

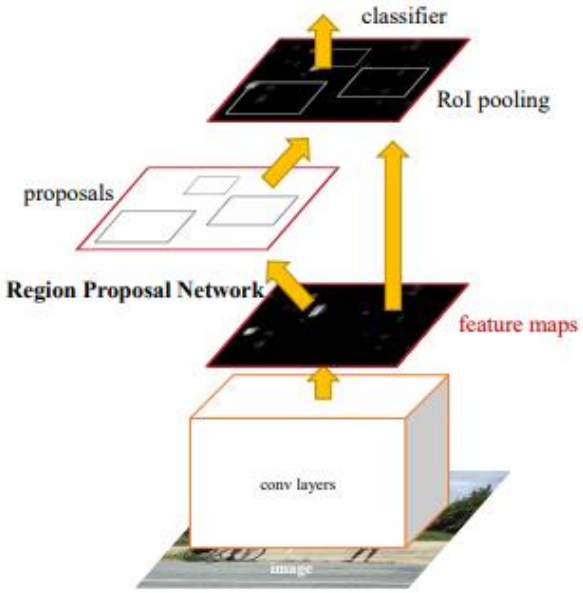

Figure 5. Figure 2 from [9]

IX. YOLO(YOU LOOK ONLY ONCE)

YOLO is the current state-of-the-art real time system built on deep learning for solving image detection problems. As seen in the below given image, it first divides the image into defined bounding boxes, and then runs a recognition algorithm in parallel for all of these boxes to identify

which object class do they belong to. After identifying this classes, it goes on to merging these boxes intelligently to form an optimal bounding box around the objects. The following Fig.6 shows the architecture .

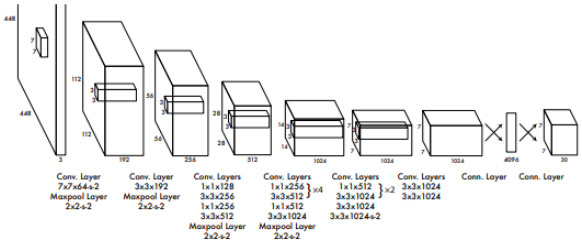

Figure 6. Figure 3 as in [10]

X. SQUEEZE NET

The squeezeNet architecture is one more powerful architecture which is extremely useful in low bandwidth scenarios like mobile platforms. This architecture has occupies only 4.9MB of space, on the other hand, inception occupies ~100MB! This drastic change is brought up by a specialized structure called the fire module. Below image is a representation of fire module. The following Fig. 7 and Fig. 8 shows whats squeeze net and its architecture.

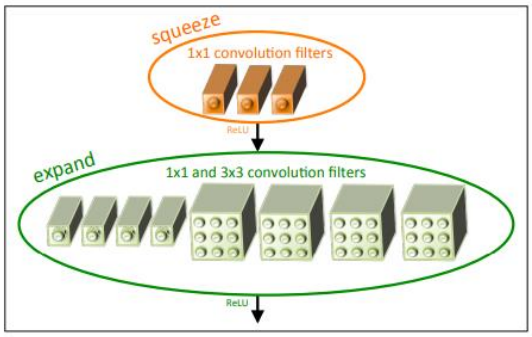

Figure7 . Figure 1 of [11]

The architecture of SqueezeNet is given below

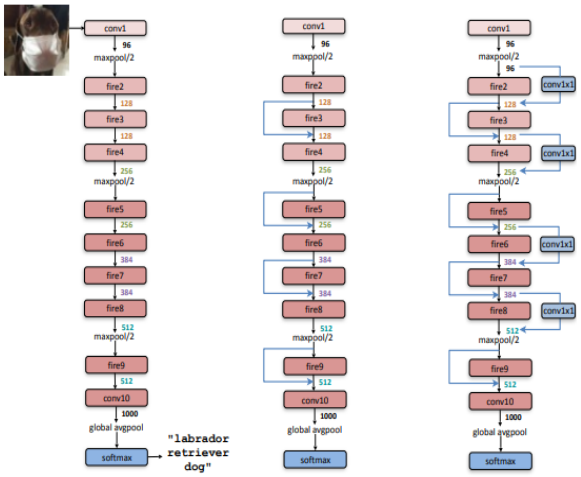

Figure 8. Figure 2 of [11]

XI. SEGNET

SegNet is a deep learning architecture applied to solve image segmentation problem. It consists of sequence of processing layers (encoders) followed by a corresponding set of decoders for a pixelwise classification . Below image summarizes the working of SegNet. One key feature of SegNet is that it retains high frequency details in segmented image as the pooling indices of encoder network is connected to pooling indices of decoder networks. In short, the information transfer is direct instead of convolving them. SegNet is one the the best model to use when dealing with image segmentation problems. The following Fig.9 shows its architecture.

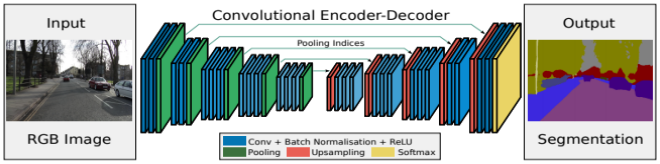

Figure 9. Figure 2 of [12]

Abbreviations and Acronyms

RCNN-Region based CNN

YOLO-You Only Look Once

CNN-Convolutional Neural Networks

ILSVRC- ImageNet Large Scale Visual Recognition Challenge

XII. MERITS AND DEMERITS OF THE EXISTING DEEP LEARNING ARCHITECTURE

The merits and demerits of the existing deep learning architecture are formulated as a table and given below

Table 1: Merits and demerits of the deep learning architectures

S.No	*Architecture*	*Merit*	*Demerit*
1	AlexNet	Basic algorithm	Need to be implemented with other architecture
2	VGGNet	used for Benchmarking any application	very slow to train if trained from scratch.
3	ResNet	1)To accelerate the speed of training of the deep networks 2) Instead of widen the network, increasing depth of the network results in less extra parameters Reducing the effect of Vanishing Gradient Problem 4) Obtaining higher accuracy in network performance especially in Image Classification	1) Vanishing Gradients 2)Optimization Difficulty
4	GoogleNet	Simplicity: the network consists of 9 identical and relatively simple blocks Parallelism: the network layers within each block are structure in 4 parallel pathway Computation and memory efficiency: because of the parallel network implementation and the dimension reduction layers in each block, the model size is contained within 27Mb npy file, and its execution time beats VGG or ResNet on commodity hardware.	Lower accuracy: the high efficiency comes at a small cost of the model accuracy
5	ResNeXT	High accuracy: ResNet achieves one of the best performance accuracy, beating VGG and GoogleNet in ILSVRC 2012 testset	Relative complex model: although simple in concept, ResNet implementation is highly complicated due to the extensive use of shortcut path that skips layers and pooling, normalizations operations. This increases debugging and innovation cost.

6	RCNN	Store Information Learn Sequential Data	-High computational cost. - If you don't have a good GPU they are quite slow to train (for complex tasks). -They use to need a lot of training data
7	YOLO	Speed (45 frames per second—better than realtime) Network understands generalized object representation (This allowed them to train the network on real world images and predictions on artwork was still fairly accurate). faster version (with smaller architecture)—155 frames per sec but is less accurate.	spatial constraint limits the number of nearby objects that our model can predict. Our model struggles with small objects that appear in groups
8	SqueezeNet	same accuracy of AlexNet, SqueezeNet can be 3 times faster and 500 times smaller.	Computationally expensive. Multiple step pipeline. Requires feature engineering. Each step in the pipeline has parameters that need to be tuned individually, but can only be tested together. Resulting in a complex trial and error process that is not unified. Not realtime
9	SegNet	Best when used for image segmentation	Only 2 bits are needed for each window of 2x2, slight loss of precision
10	GAN	GANs are a good method for training classifiers in a semi-supervised way. GANs generate samples faster than fully visible belief nets because there is no need to generate the different entries in the sample sequentially.	It’s hard to learn to generate discrete data, like text. • Compared to Boltzmann machines, it’s hard to do things like guess the value of one pixel given another pixel.

CONCLUSION

The existing deep learning architectures are studied and the merits and demerits of those are analyzed.

REFERENCES

[1] https://www.ibm.com/developerworks/library/cc-machine-learning-deep-learning-architectures/index.html

[2] https://www.analyticsvidhya.com/blog/2017/08/10-advanced-deep-learning-architectures-data-scientists/

[3] https://www.kairos.com/blog/the-best-explanation-machine-learning-vs-deep-learning

[4] https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[5] Karen Simonyan, Andrew Zisserman "Very deep convolutional networks for large-scale image recognition",2014, arXiv:1409.1556v6

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Identity mappings in deep residual networks" (Submitted on 16 Mar 2016 (v1), last revised 25 Jul 2016 (this version, v3)).arXiv:1603.05027v3

[7] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, Zbigniew Wojna, "Rethinking the inception architecture for computer vision", (Submitted on 2 Dec 2015 (v1), last revised 11 Dec 2015 (this version, v3)) , arXiv:1512.00567v3

[8] Saining Xie, Ross Girshick, Piotr Dollar , Zhuowen Tu, Kaiming He,UC San Diego,Facebook AI Research, "Aggregated Residual Transformations for Deep Neural Networks" arXiv:1611.05431v2 [cs.CV] 11 Apr 2017

[9] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks" , (Submitted on 4 Jun 2015 (v1), last revised 6 Jan 2016 (this version, v3))

[10] Joseph Redmon ,Santosh Divvala, Ross Girshick, Ali Farhadi "You Only Look Once: Unified, Real-Time Object Detection"

[11] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, Kurt Keutzer , "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size" , (Submitted on 24 Feb 2016 (v1), last revised 4 Nov 2016 (this version, v4)) arXiv:1602.07360v4 [cs.CV] 4 Nov 2016

[12]Vijay Badrinarayanan, Alex Kendall, Roberto Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation", arXiv:1511.00561v3 [cs.CV] 10 Oct 2016

[13]Ian J. Goodfellow, Jean Pouget-Abadie∗ , Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair† , Aaron Courville, Yoshua Bengio, "Generative Adversarial Nets" , arXiv:1406.2661v1 [stat.ML] 10 Jun 2014.