ISSN: 0976-3104
SPECIAL ISSUE(SCMSA)
Jenisha et al.

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

A SURVEY OF NEURAL NETWORK ALGORITHMS USED FOR IMAGE ANNOTATION

Jenisha T. and Swarnalatha Purushotham
School of Computing Science and Engineering, VIT University, Vellore-632014, INDIA

ABSTRACT

In recent years, scalable image annotation has gained more popularity because of its usefulness in various applications like image recommendation system, context aware chat bots , image based query retrieval, object detection, object segmentation, pose estimation, matching patterns, generating synthetic images and many more. The usage of convolutional neural network for object detection and localization became popular in 2012 when then the neural network designed by Alex Krizhevsky et al. had record breaking performance with error rate of 16.4% in top 5 hit% in ImageNet 2012 challenge . In the year 2014 Goog LeNet neural network model from Google outperformed with error rate of 6.66% in ImageNet 2014 test dataset. This statics shows the accuracy is pretty high when deep convolutional network is used for object detection. In medical image the accuracy of detected object is crucial. Object detection using Convolution Neural Networks has given good accuracy of about 96.7% in ImageNet dataset. This research paper does literature review of different architecture styles researchers used to construct Convolutional Neural Network model for object detection, localization, and annotation. Visualizing neurons which was a complete black box till 2012 is described. Convolutional Neural Networks (CNN) is networks that share parameters across space. It also describes the entire process flow of automatic image captioning task. The performance improvements gained by different optimization techniques are described. It starts from data preparation to CNN and ends with automatic sentence generation using CNN.

Received on: 30th-Nov-2015
Revised on: 11th-March-2016
Accepted on: 26– March-2016
Published on: 10th –Aug-2016

KEY WORDS

CNN,annotation,deeplearning,classification,pooling, convolution

***Corresponding author: Email:** jenisha.t2011@vit.ac.in

INTRODUCTION

Current approaches to image annotation make essential use of deep neural networks because of available big data and GPU provided by Nvidia and other GPU infrastructure providers like amazon cloud GPU instance. The current computing performance of GPU K80 is amazingly good which takes only few hours to process 1 million images in contrast to CPU which takes several weeks to compute. To improve performance of annotation, large datasets are trained. Taking large datasets and training them prevents over fitting of data. ImageNet [1] one of the biggest object detection dataset available consists of 500K images and 251 basic categories like dog, tree, person etc. Microsoft Common Objects in Context [2] consists of 300K images and 80 object categories. To learn about millions of objects we need a learning model with good learning capability. Despite of having object detection dataset which consists of millions of images, to automatically annotate all objects present in real world is challenging. This means the training model should have prior knowledge to compensate for the objects not present in training dataset. Convolutional Neural network a supervised learning model comes with such learning capability.

Supervised learning is composed of two important subfields Classification and Regression. Classification is predicting discrete classes. Regression is finding the continuous outcomes (range) of target variables. Example for regression is predicting the house price in particular region. In image classification task for input image Ix predict which class is most suitable. In **Figure- 1** shows linearity layer of CNN. For Image Ix linear model as shown in Eqn. 1 is applied. W and b are computed using Stochastic Gradient Descent or other numerical optimization techniques which are detailed in coming sections.

THE IIOAB JOURNAL
ISSN: 0976-3104

$$Y=WI_x+b \qquad (1)$$

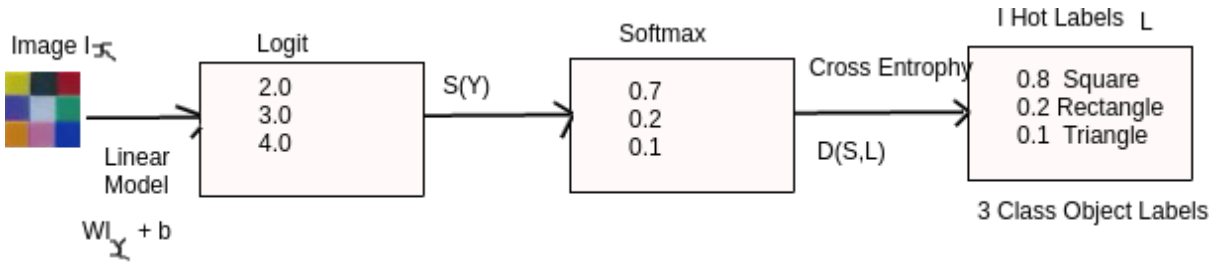


Fig: 1.Linearity Layer

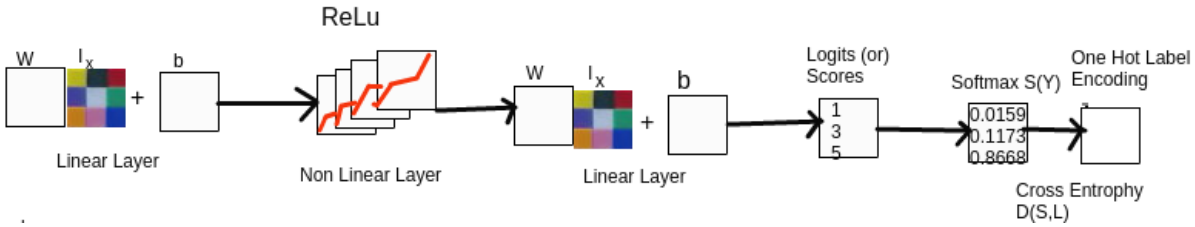


Fig: 2. Combining Linearity and Non Linear Layer to build CNN

The output is scores in terms of logistic regression. These are also called Logits. In CNN two linear models are connected by nonlinear layer. In **Figure-2** shows how two linear layers are connected using ReLu to form non-linear network. The first layer effectively consists of the set of weights and biases applied to X and passed through ReLUs. The output of this layer is fed to the next one, but is not observable outside the network; hence it is known as a hidden layer. The second layer consists of the weights and biases applied to these intermediate outputs, followed by the softmax function to generate probabilities.

Terminologies used in CNN are explained below.

Softmax

Softmax function converts scores into probability. Eqn. 2 gives proper probability for score *i*. Proper probability is probability when summed, all the probabilities of given *i* the value will be 1. When the size of output *Y* is high the confidence will be more for prediction. When the size of *Y* is small the confidence which class given object belongs to decreases.

$$S(y_i)=\frac{e_i^y}{\sum e_j^y} \qquad (2)$$

One Hot Encoding

In One Hot Encoding we give one for correct class and zero to other class labels. This method has drawback when we have billions of classes which makes one hot encoding inefficient. A better approach is to use embeddings. In embeddings we convert word to vector and the words that have same context has closer value. For more details about embeddings see section on Automatic Image Captioning Using CNN and RNN.

Cross Entropy

The way to measure distance between two probabilities is called cross entropy. Eqn. 3 is used to compute the distance between two probabilities.

THE IIOAB JOURNAL
ISSN: 0976-3104

$$D(S,L)=-\sum_i L_i log(S_i) \quad (3)$$

Cross Entrophy is not symmetric. $D(S,L)\neq D(L,S)$
Composition of computations abstracted as layers in CNN is organised Directed Acyclic Graph(DAG). The different layers in CNN are Convolution layer, Non-linearity layer and Pooling layer. A featurer map is obtained after passing through these layers. These are commonly used layers found in literature. One or more layer like inception layer as discussed in subheading inception and a backpropagation layer in subheading backpropagation,and residual block discussed in resnet are introduced by researchers for performance tuning. **Figure-3** shows different layers of CNN.

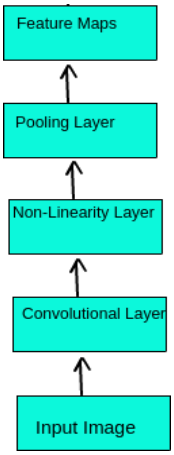


Fig:3. Different Layers of CNN

.....

Computing Weight and Bias

When computing weight *W* and bias *b*, the value should be more for correct class and less for incorrect class. The distance should be low for correct class and high for incorrect class. We need to minimize distance between similar labels and maximize distance between dissimilar labels. Eqn. 4 gives training loss over the entire training set. Training loss is computed by measuring the distance averaged over the entire training sets for all inputs and all labels. If we have 10,000 training images and 100 labels then distance between 10000 images and each label is calculated and average distance is taken as training loss *L*.

$$\mathcal{L}=\frac{1}{N}\sum_i D(S(WX_i+b),L_i) \quad (4)$$

Loss is function of weight and biases. We want to minimize loss. A simple numerical optimization approach to minimize loss is to use Gradient Descent method. To obtain numerical stability we subtract and divide each pixel by 128 as shown below.

RedPixel R= (*R*-128)/128,

GreenPixel G= (*G*-128)/128,

BluePixel B= (*B*-128)/128)

While computing Gradient Descent weights are initialized by randomly drawing weights from Gaussian distribution with mean μ zero and equal variance. The smaller values for standard deviation leads to uncertainty and larger values for standard deviation leads to overconfidence. In training its better to choose very small value as variance σ. Eqn. 5 and Eqn. 6 shows computing weight W and bias b used in 8

$$W=w-\alpha\Delta_w L \quad (5)$$

$$b=b-\alpha\Delta L' \quad (6)$$

THE IIOAB JOURNAL
ISSN: 0976-3104

Alternate numerical optimization methods to find optimum weight and bias are using Stochastic Gradient Descent (SGD), Ada Grad, Adam, Nesterov's Accelerated Gradient, RMSprop.Ada Grad is modification of SGD which implicitly does momentum and learning rate decay. Using AdaGrad makes less sensitive to hyper parameters.

This literature survey is organized as follows. Section 1 gives overview of CNN, section 2 describes evolution of CNN, section 3 describes data preparation for CNN, followed by each layers in CNN, then section on optimization next error computation methods are described. Next popular CNN architecture are described in section 12. Different CNN network models found in literature, then performance evaluation used to determine accuracy of object detection finally conclusion based on comparative study of different approaches.

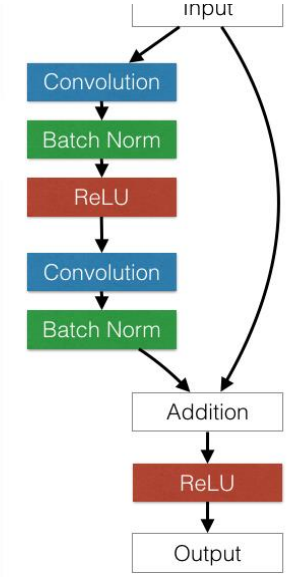
EVOLUTION OF CONVOLUTIONAL NEURAL NETWORKS

The research in neural network started early 1980's but lack of computing power the model didn't gain popularity. In the year 1998 [4] LeNet -5 model was designed to classify MNIST dataset . This model is used to read the digits in cheque and successfully used in banking industry. LeNet-5 model has 60,000 training and 10,000 test samples hand written digits are taken.Test error rate is 0.8 %. In 2012 AlexNet [5] was modeled with 7 hidden weight layers, 650K neurons and 60M parameters, and 630M connections to classify 1.2M images in 1000 categories. Imagenet 2012 dataset is used to train AlexNet. 4 layer CNN with activation unit ReLU is used which converged 6 times faster than equivalent network using tanh as activation unit. Alexnet decreased state of art error rate 47.1 % to 37.5 % for top 1 result and the error rate was further reduced from 28.2 % to 17 % for top 5 results. In the year 2014 GoogLeNet [7] introduced deep CNN architecture named as 'inception'. In 2015 Microsoft deep residual network model is designed based on the principle of residual learning. This model had 152 layers which is 8x deeper than state-of-the-art CNN layers. **Figure- 4** shows evolution of Convolutional Neural Networks. **Table- .1** shows summary of error rate decresed in Imagenet dataset over a period of time without using any external dataset. The error rate decresed as the cnn layer goes deeper and deeper.

Table: .1: Top 5 Error Rate and CNN Models

CNN Model	Year	Error Rate Top 5 in %
AlexNet	2012	16.4
Clarrifai	2013	11.7
GoogLeNet	2014	6.66
ResNet	2015	3.57

2015 ResidualNet



152 layers

THE IIOAB JOURNAL
ISSN: 0976-3104

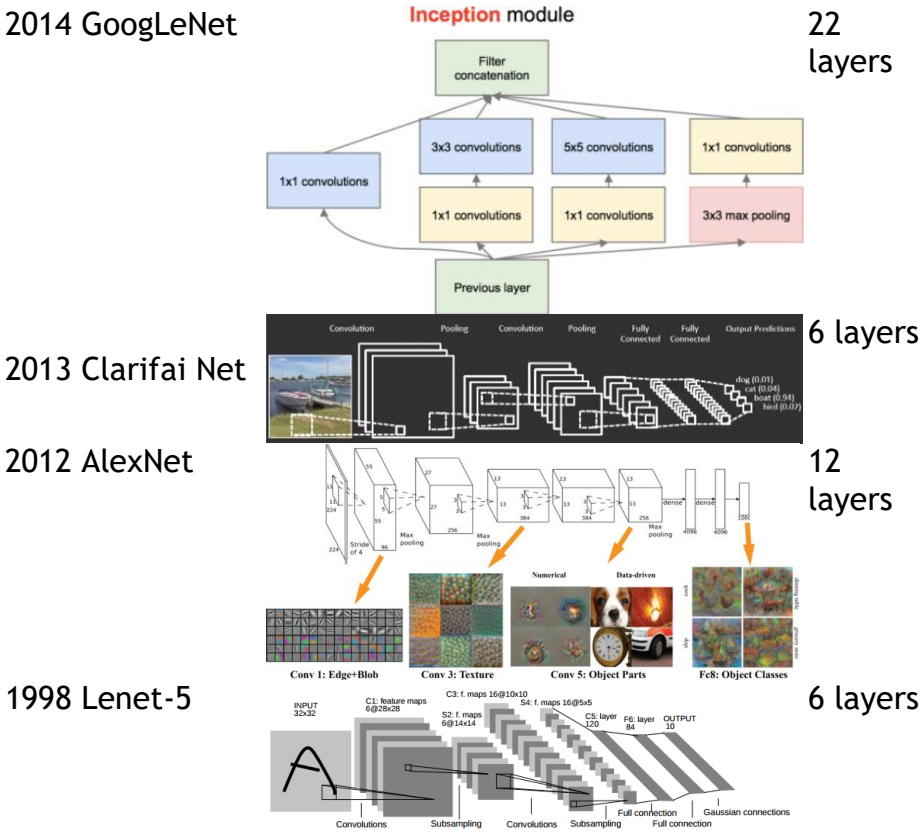


Fig: 4. Evolution of CNN

.....

Data Preparation for Convolutional Neural Network

Color images have height, width and depth. In RGB color channel an image is represented as Red, Blue and Green intensity. The intensity value of R, G, B are represented as three matrices of same height and width. **Figure- 5** is RGB image of size 38x38x3.When this image is represented in matrix form there are 3 matrix one for each R,G,B and matrix size is 38 x 38.

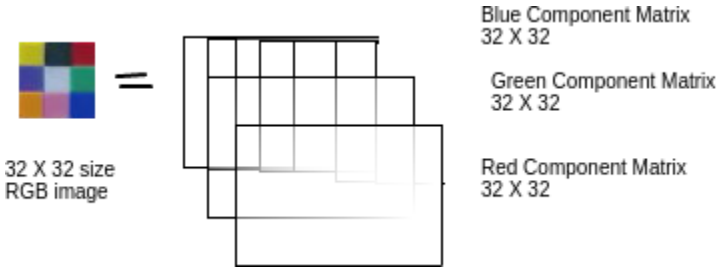


Fig: 5. RGB Components of an Image

.....

If color doesn't matter, it might help to reduce the complexity of the problem by combining color channels into a single monochromatic channel. Taking the average (R+G+B)/3 is one way of doing it; however there are other transformations that might be more effective/closer to, how human's perceive color (e.g. converting to YUV and using the Y channel). In luminosity method of converting RGB to monochrome green is given more weightage as green is more sensitive for human eye as shown in Eqn. 7.

THE IIOAB JOURNAL ISSN: 0976-3104

$$I_{(gray)} = 0.21R + 0.72G + 0.07B \quad (7)$$

The gray image is represented as single column vector. For CNN both color image and monochromatic image can be used. The images taken should be translation invariance meaning even if the image is rotated the features extracted from the image should be able to detect object correctly. When the extracted features are translation invariant, the position of the object in the image doesn't matter. The object located anywhere top, left, bottom, right can be detected with translation invariant features.

Dataset Used to Benchmark CNN Models

There are few standard dataset that can be used to benchmark our algorithms with state-of-art algorithms. Commonly used dataset for natural image annotation are Microsoft COCO dataset, ImageClef dataset, CIFAR dataset.

1. Microsoft COCO Dataset with Image Captions is used by researchers for image recognition, Image Segmentation and Image Captioning.
2. Imagenet dataset can be used for object detection,object localization,Object detection from video and Scene Classification for Video.

Data Layer for CNN

While training the CNN model common practice is 70 % of image is used during training, 20 % during validation and 10 % during testing. HDF5 file format is popularly used while handling large image data.

Data_layer= AsyncHDF5DataLayer(name=``train_data``,source=``data/train.txt``,batch_size=64,shuffle=true")

Above code is to read HDF5 format image data in TensorFlow Framework. It reads 64 images in single batch from data/train.txt. Shuffle the image is set to true to get faster convergence.

Convolution Layer

Natural image has stationary properties. That means the statistical properties of images remains the same in different parts of the images. This suggests that the features we have learned in one part of the image can be used to learn in other parts of the image. This is done using convolution. Suppose we have learnt features from 8 X 8 patch sampled from 1024 X 1024 pixel image. Train a sparse encoder on 8 X 8 or a X b patch small patch sampled from bigger image X. From the sampled patch x_{small} the number of features learned is k

$$f=\sigma(W^{(1)}x_{small}+b^{(1)}) \quad (8)$$

σ - sigmoid function given by weight $W^{(1)}$ and $b^{(1)}$ from visible unit to hidden units.
For every x_{small} in image I_x compute f_s as shown in Eqn. 8. The size of convoluted image $f_{convoluted}$ array is computed according to Eqn. 9

$$f_{convoluted}=kX(r-a+1)X(c-b+1) \quad (9)$$

where, r - number of pixel array rows in image I_x

c - number of pixel array columns in image I_x

a - number of pixel array rows in sampled patch of the image x_s

b - number of pixel array columns in sampled patch of the image x_s

k – number of image channel used.
For patch size 7 X 7 pixel from 38 X 38 image, k=3 ,then array size of convoluted image will be 3 X 32 X 32 when using valid convolution. In valid convolution those part of convolution computed without using zero padded edges is returned. Its observed that CNN first learns lines, then edges, shapes and finally objects.

THE IIOAB JOURNAL ISSN: 0976-3104

Non-Linearity Layer

Two linear netwoks are connected using nonlinear function such as sigma, tanh or ReLu as shown in Eqn. 10 & 11.An activation function layer or non-linearity layer applies an entry wise non-linearity map. More useful insights are obtained by increasing depth of the layers than increasing width of layers.

$$Sigmoid\sigma(x)=\frac{1}{1+e^{x}} \quad (10)$$

$$ReLu\sigma(x)=max\{x,0\} \quad (11)$$

Pooling Layer for CNN

Pooling is done for dimensionality reduction and make image translation invariant. A convoluted image is divided into *m*m* sub regions. Usually the value of *m* is less than 5 for large image. From each sub regions take max value or maxpooling or average value for meanpooling. Pooling also avoids overfitting. The outcome of pooling is pooled convoluted features. **Figure- 6** shows sample of pooling image.

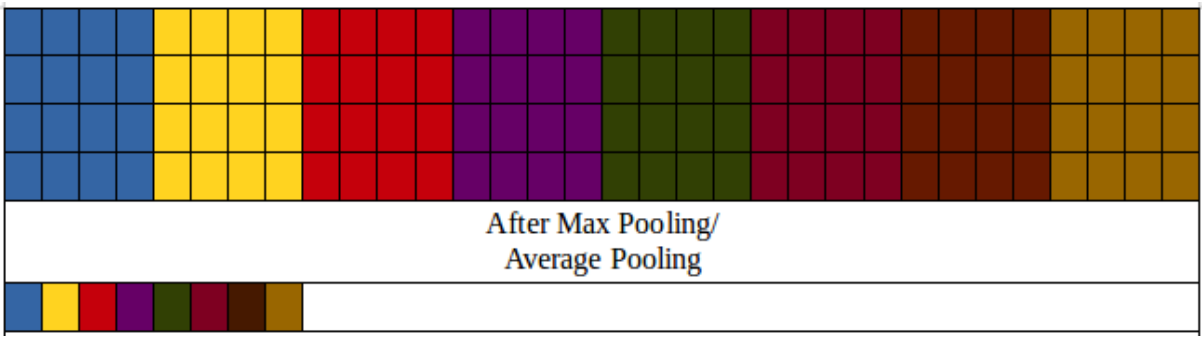


Fig: 6. Pooling Convoluted Image of size 32 X 32

.....

Poolinglayer oolingLayer(name=``pool``,kernel=(2,2),stride(2,2),bottoms=[:conv],tops=[:pool])

Network Layer for CNN

Fully connected layer is also known as dense layer or inner product layer or linear layer is computed as following

$$y_i=\sum w_{ij}x_j \quad (12)$$

Backpropagation Layer

Backpropagation layer takes advantage of chain rule.

$$[g(f(x))]'=g'(f(x))\times f'(x) \quad (13)$$

It decompose convolution, nonlinear and pooling operations into $f^{conv}, f^{nonlinear}, f^{pool}$ elements whose derivatives with respect to inputs are known by symbolic computations. It backpropagate error signals corresponding to a differentiable cost function by numeric computation. The propagated error signal is used to compute local minima.

Error Computation for CNN

The common criteria for error computation are recall with respect to precision and carried out for different image transformations. In **Figure-7** shows precision and recall computation is shown.

THE IIOAB JOURNAL
ISSN: 0976-3104

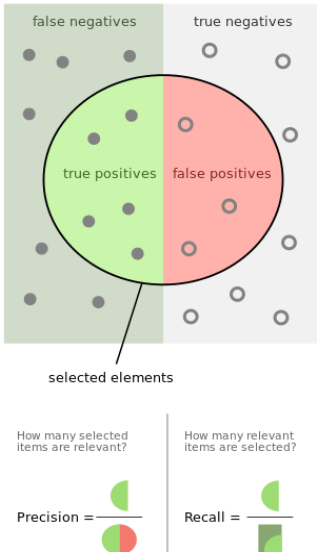


Fig:7.Precision and Recall.Image Credit : Wikipedia, Precision and recall,2016

Eqn.14 shows recall calculation. Recall is used to measure sensitivity or true positive rate. Its equivalent to computing hits or how many images classified correctly.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \quad (14)$$

Eqn. 15 shows precision calculation. Precision is used to measure Positive Predicted Value (PPV).Its equivalent to computing correct rejection.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \quad (15)$$

The other approaches are computing true negative rate and accuracy. Eqn. 16 is used to compute true negative rate.

$$TrueNagativeRate(TNR) = \frac{TrueNegative}{TrueNegative + FalsePositive} \quad (16)$$

Eqn. 17 is used to compute Accuracy.

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + TrueNegative + FalsePositive + FalseNegative} \quad (17)$$

A combination of prediction and recall is called F-score. Eqn. 18 shows computing F-score.

$$F = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (18)$$

Optimization Techniques in CNN

CNN model success is heavily on designing correct training loss function which can be optimized in less time. Challenge is the optimization models are very difficult to scale. Few optimization methods used in CNN are Gradient Descent, Stochastic Gradient Descent(SGD), Ada Grad and L-BFGS, Nesterov's Accelerated Gradient and Adam.

Stochastic Gradient Descent Optimizing Weights W and Bias B

If computing Training Loss $\mathcal{L}$ takes 1× memory computing Gradient Descent takes 3× memory. One way is to choose random sample. Then compute Training Loss $\mathcal{L}$ for that sample. Next take derivative of that sample. Repeat this step for different random samples multiple times till numerical stability is obtained. This is called Stochastic

THE IIOAB JOURNAL
ISSN: 0976-3104

Gradient Descent which performs better than Gradient Descent and scales well. But in practice it has it's own limitation. **Figure- 8** shows SGD with Momentum which can easily navigate to local minimum even if the area of surface curves more steeply in one direction.

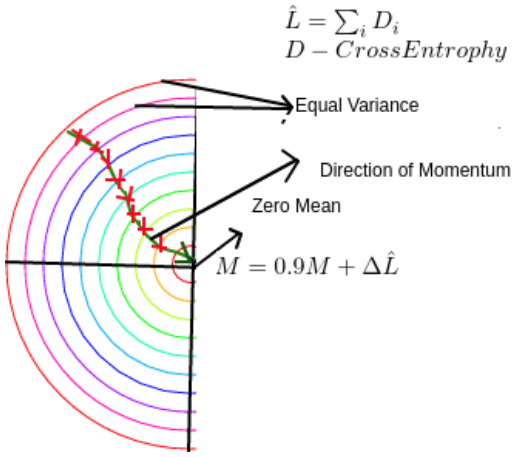


Fig: 8. SGD with Momentum

REGULARISATION

Deep learning technique is successful only if we have right amount of data to train the model. Network with right size of data is very hard to optimize. In practice we train network that is very big for model then find to try the best. In **Figure- 9** shown when to stop training process a time when the accuracy rate doesn't change considerable even after repeated training.

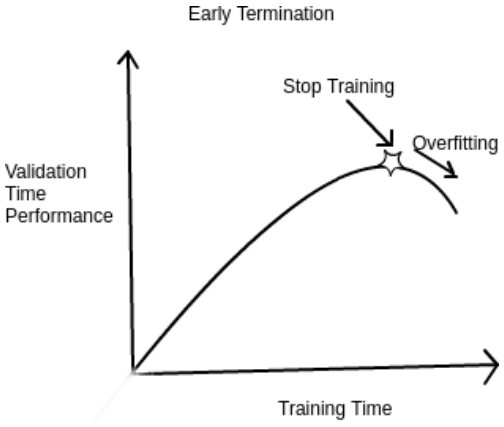


Fig: 9. Early Termination of Training Data in CNN

Regularization is adding artificial constraints to network which implicitly reduces network parameter. In L2-Regularization L2 Norm is added to training loss $\hat{L}$ as shown in Eqn. 19

$$\hat{L} = L + \beta \frac{1}{2}\|w\|_2^2 \qquad (19)$$

THE IIOAB JOURNAL
ISSN: 0976-3104

L2 Norm is sum of the squares of individual weights in the vector.

$$L2Norm = \frac{1}{2}(w_1^2 + w_2^2 + w_3^2 + w_4^2 + ... + w_n^2) \quad (20)$$

Regularizing through early stopping , results in fast training and good generalization performance. A small-enough step-size w.r.t learning rate is often sufficient for state-of-the-art performance. One-vs-rest strategy is a flexible option for large-scale image classification.

Dropout

The parameter that passes from one network to another network is called activators. Take half activations flowing through network randomly and destroy them randomly again. If one activation get smashed there is always one or more activations which do the same job that are not destroyed. Dropouts prevent overfitting and make it more robust . If dropout doesn't work we should increase network size. Randomly destroy these activators in each layer. In **Figure- 10** the value 0.2 and 0.9 are eliminated during training. Scale all activation parameters by 2 and subtract them at output to determine accuracy improved using dropout.

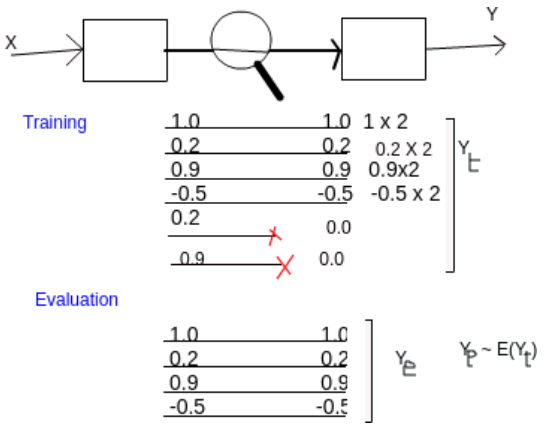


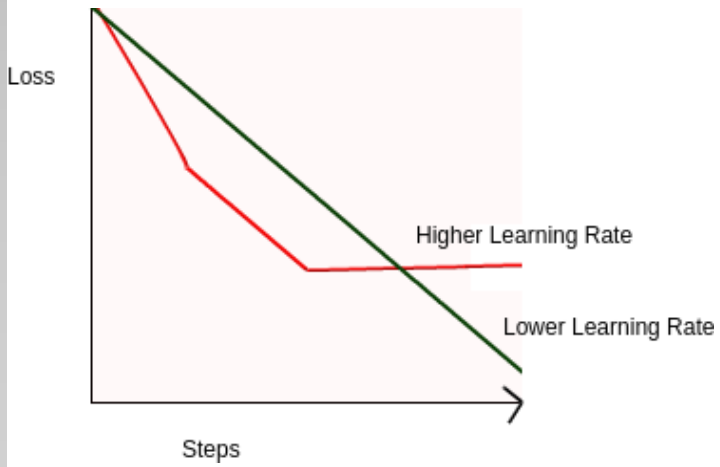
Fig: 10. Droupout: Redundant Activation Smashed to Aoid Overfitting

.....

Dropout's doesn't work well with RNN and LSTM. **Momentum** Momentum techniques leads to better convergence rate. Take running average of gradient to find momentum.

Learning Rate Tuning

In practice the training is done with lower learning rate 0.2. With higher learning rate the image classification model tends to learn quickly but doesn't converge. Training loss is average cross entropy. We do to minimize distance between similar labels and maximize distance between dissimilar labels. Training loss is computed as shown in Eqn. 4. As shown in **Figure- :11** lower learning rates stabilize over a period of time.



THE IIOAB JOURNAL
ISSN: 0976-3104

Fig:11. Learning Rate Turning

CNN MODELS IN LITERATURE

Convolutional Neural Networks are designed to take raw pixel as input and visualize the pattern present in the image with minimal preprocessing. The CNN learns at each layer starts from learning lines then edges then detect shapes and finally objects. Ross Girshick, Donahue et al. [3] proposed model to combine region information to localize and segment images. This approach is called R-CNN. A detailed description of famous CNN models described below.

LeNet

Lecun,Bottou et al. in 1998 [4] proposed LeNet5. The Lenet-5 model is capable of recognising hand written digits with extreme variability. Graphical depiction of LeNet-5 model is shown in **Figure- :12**. Convolution Layer and max pooling layer forms the core of LeNet5 model. 32 X 32 handwritten image is convoluted and 6 features map is formed. Then maxpooling is done to reduce size of feature map to 14 X 14. A non linearity layer is inserted between layer 1 and layer2. Tanh function is used as nonlinearity function. After nonlinearity function feature map is convoluted and 10 X 10 size feature map is formed. The last layer is fully connected layer. It creates a fully connected bipartite graph between input layers and output layers which are class labels in case of supervised classification of images.

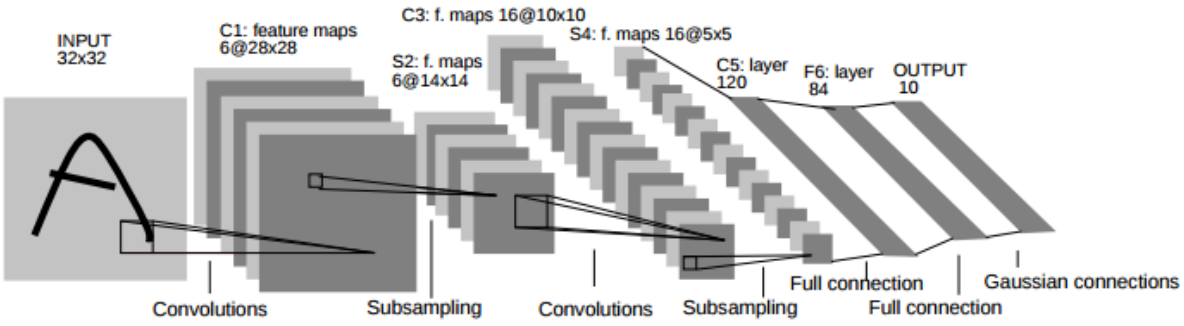


Fig:12. LetNet5 Architecture Image Credit: [4]

AlexNet

Krizhevsky, Sutskever et al. in 2012 [5] proposed Alexnet which won ImageNet competition.It uses the same model as LeNet-5 but a bigger model, mored data, and GPU implementation. It was able to classify 1000 objects. The accuracy of this model is high when compared to previous years hand engineered features which existed for decades. It had 7 hidden layers and 60,000,000 parameters trained on GPU for 2 weeks. **Figure- 13** shows graphical depiction of Alexnet architecture with visualization of features learned below. Neurons learned to detect simple edge and blob in layer1, texture pattern in layer 3 and object classes in layer 5.

THE IIOAB JOURNAL
ISSN: 0976-3104

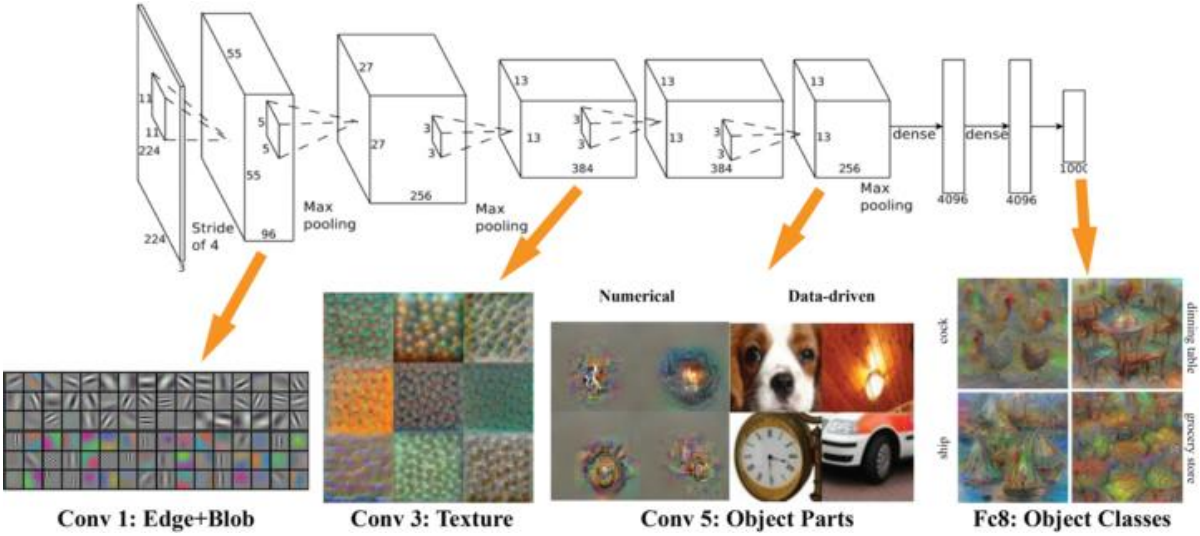


Fig:13. AlexNet Architecture Src: [6]

……..

GoogLeNets

Inception Module Szegedy,Liu et al. [7] which are parallel convolution and pooling layers merged with previous convolution stage. **Figure- 1** is graphical depiction of GoogLeNet Architecture is shown. Idea is at each layer of convnet you can make a choice have a pooling operation, have a convolution, then decide is it 1 X 1 convolution or 3 X 3 convolution or 5 X 5 convolution. All of these are actually beneficial to the modelling power of the network.Average pooling and 1 X 1 convolutions are better than convnets that simply use a pyramid of convolutions.

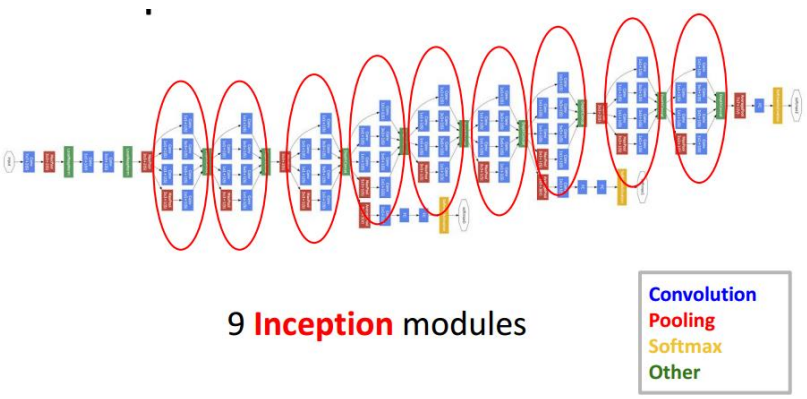


Fig: 14: GoogLeNet Architecture Src: [7]

……

ResNets

In 2015 Kaiming He, Xiangyu Zhang et al. [8] proposed Residual Neural Networks wich solves the probem of optimizing Feed Forward Neural Network beyond certain depth.They add residual block to overcome this challenge. A residual block is constructed by by passing few convolutional layers at a time. **Figure- :15** is graphical depiction of ResNet architecture at the left and a modified approach at middle and without Relu at right.

THE IIOAB JOURNAL
ISSN: 0976-3104

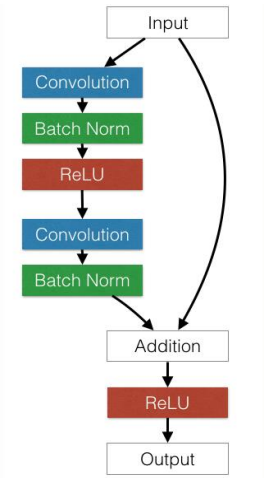


Fig:15.ResNets Architecture Image Credit: Tourch, Training and investigating Residual Nets [8]

.....

Decision Forest Deep Neural Networks DNDF and other CNN models

Kontschieder,Fiterau et al. [9] model with 12 layers beats GoogLeNet's 6.67% top-5 error on ImageNet to 6.38% accuracy.In [10] Oquab, Bottou et al. mentions weakly supervised model is used to recognize objects from cluttered scenes. Saliency inspired neural network model by Erhan, Szegedy et al [11] is used to localize images. Wang,Huang et al [12] used weekly supervised model using latent category learning. Ontology based hierarchial image annotation was proposed by Zarka,Ammar et al. [13]. Part based CNN method was used by zhang2014,Donahue et al [14] for category detection.

AUTOMATIC IMAGE CAPTIONING USING CNN AND RNN

In 2015, Klein,Lev et al [15] used Fisher Vectors derived from HGLMMs (Hybrid Gaussian-Laplacian Mixture Model) to represent sentence. Donahue,Anne et al [16] used Learning Long term dependency called LSTM to avoid the problem of vanishing gradients.A problem which occurs if we need to generate long sentence. This problem was solved by using memory unit which stores past state. Text embedding is done to convert words into vectors that can be used for auto image captioning.

Text Embedding

There are two challenges in text embedding. First is all words used in real world may not be present in training sets. For example the word 'tiffin'meaning light meal appears chiefly in Indian documents. While training a restaurant recommendation system the training data contains the word 'lunch' and 'breakfast' but not the word 'tiffin'. To do this mapping perfectly we need to know word context. The second challenge is sharing the weights between semantically related words. For example the words 'daughter' and 'princess' are related to each other. In **Figure-16** the semantic similarity similar words for the word 'daughter' is depicted. The word 'princess', 'sister' shares weight 0.7697988749 and sister shares weight 0.7702152133 which is very close to the weight of the original word 'daughter' which has weight of 0.7847946882 as shown in **Table- 2**

Table: 2.Word Similarity Based on Semantic Meaning for Word 'Daughter'

Word	Vector
princess	0.7697988749

THE IIOAB JOURNAL
ISSN: 0976-3104

sister	0.7702152133
cousin	0.7743015289
grandmother	0.775909245
daughters	0.7847946882
niece	0.7886790037
mother	0.7911210656
widow	0.8072112799
son	0.8378984928
wife	0.8501636982

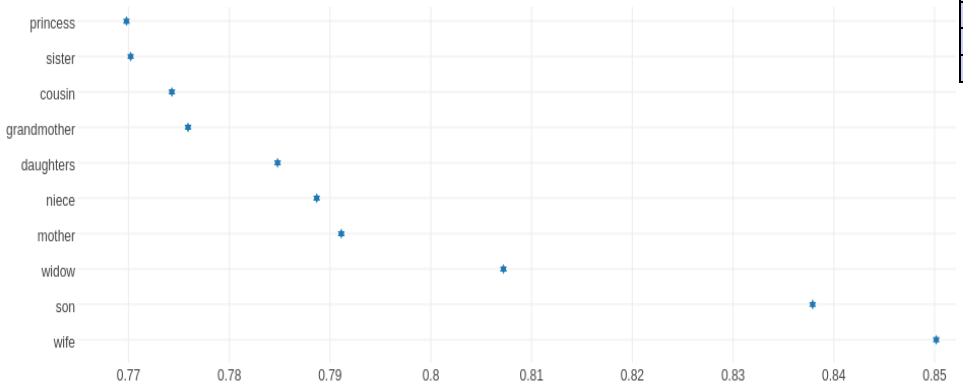


Fig: 16. Word2Vec similarity for word 'daughter'

In. Eqn.21 computing distance between two words are shown.

$$CosineSimilarity = \frac{V_{w1} V_{w2}}{\|V_{w1}\| \|V_{w2}\|} \qquad (21)$$

In sampled softmax take random samples of the target and predict its nearest neighbors. This increases the performance of the model. To generate sentence Recurrent Neural Networks are used. CNN share weights across space. RNN shares weight across time. Gated Recurrent Network (GRU) model is recently becoming popular to generate sentence. In **Figure- 17** the process flow of word2vec model is depicted.

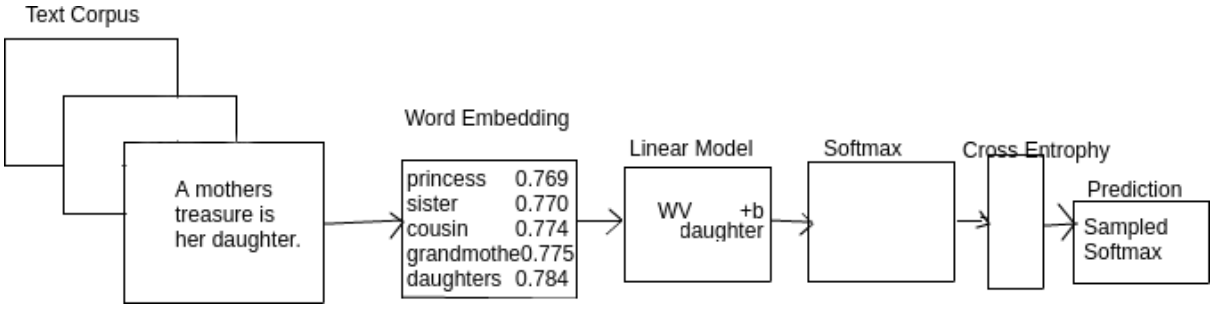


Fig: 17. Word2Vec Process Flow for word 'daughter'

Sutskever,Vinyals et al. [17] sequence to sequence mapping is done using LSTM. The sentence generated can also be used for translation. Luong,Sutskever et al [18] out-of-vocabulary (OOV) word emitted during sentence generation is fed to french dictionary for translation. The advantage is the model can be used to translate in different languages at no additional cost. Vinyals, Toshev et al. [19] has solved the challenge of connecting computer vision to natural language processing. They have proposed model named NIC that can automatically view an image and generate description in English by connecting CNN with RNN as shown in **Figure- 18**

THE IIOAB JOURNAL
ISSN: 0976-3104

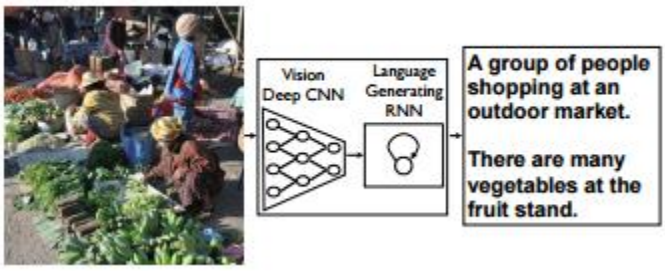


Fig: 18: CNN followed by Language Generating RNN. Image Credit : Vinyals, Toshev et al. 2014 [19]

Karpathy,Li et al. [20] proposed alignment model a combination of CNN image regions and bidirectional RNN over sentences for automatic image description. This is used for region level annotation.Lebret,Pinheiro et al. [21] has combined CNN with phrases to generate sentence. Using sematic information for video annotation is shown in [22].

VISUALIZING CNN

In [23] has shown a method to deconvolve neural network and plot the output. As the images passes through these layers depending on the features seen in input image the corresponding neuron gets activated as shown in **Figure- 6** At each layer the transform function from one neuron to other neuron is highlighted. **Figure- 19** shows learned features at Layer 4 and Layer 5. It has learned lines, edges, shapes and finally objects. Visualizing CNN has improved performance by fine tuning parameters at each layers. Clarifai model which first introduced Visualizing CNN is constructed using six layers. It used backpropagation technique to deconvolve and plot intermediate layers.

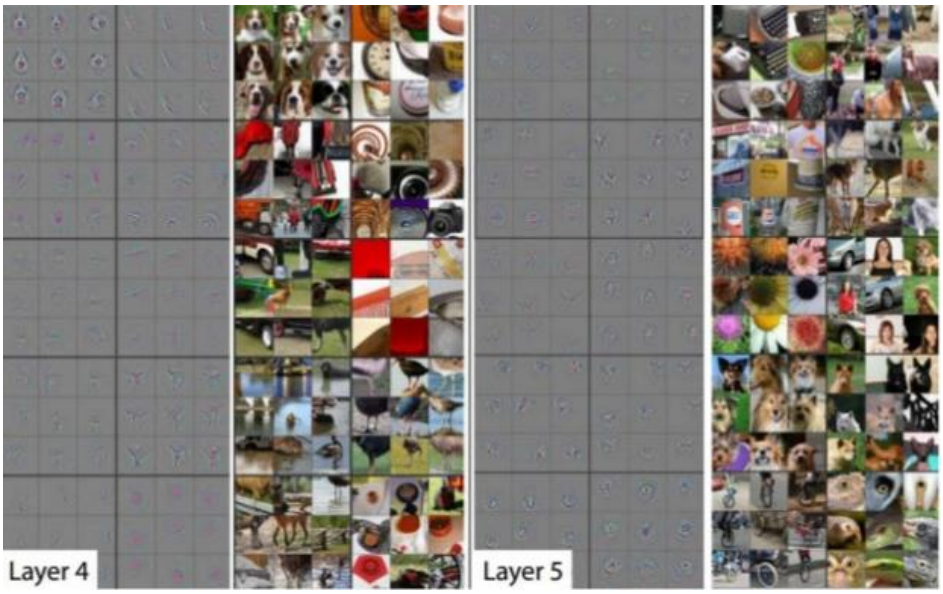


Fig: 19. Visualizing and Understanding CNN. Image Credit : zeiler & Fergus, ECCV2014 [23]

THE IIOAB JOURNAL
ISSN: 0976-3104

PERFORMANCE EVALUATION FOR CNN

Entire dataset is divided into training, testing and validation. Testing set is subset of training set which is not used during training. Change the parameters and measure the performance with validation set. Most researches use 70% of data for training set, 10% or more than 30,000 images whichever is higher as validation set and 20% as testing set. A change of 30 image in validation set to improve the accuracy by 1% or more is statistically significant size of validation set. To change the accuracy from 92% to 95% minimum 90 image parameters in validation set has to be changed. The math is as follows $\frac{3\times3000}{100}$=90 when we achieve 3% improvement in accuracy with 90 image parameters fine turned that indicates that accuracy is increasing.

APPLICATIONS OF CNN

CNN is used for gaming, image completion, object detection, object segmentation, to do pose estimation, for pattern matching and artificial image/music synthesis. Mastering the game of Go using neural networks and Monte-Carlo tree search [24] which achieved 99.8% winning rate. It used Monte-Carlo simulation to play sub games. Research on CNN for Image Completion is useful for morphing images and reconstructs images. CNN is used in identifying drugs.

CONCLUSION

The abundance of research on Convolutional Neural Network suggests that without much preprocessing robust neural network can be modeled for object detection and localization. After 2012 more papers has been published using Convolutional Neural Network model. Instead of detecting general category say dog, the species can also be detected like 'German Shepard'. These annotated images are used for automatic scene description, image retrieval based on query and many more. Localize the image and then annotate improves performance of image annotation. The error rate of Image annotation has significantly reduced to 6.7% in 2014. More number of research papers contributed for object detection and localization using convolutional neural networks in last two years when compared to other approaches like graph based object detection or object detection through statistical models. Also the number of participants in ImageNet challenge is also showing upward trend. Results get better with more data, bigger models, more computation, better algorithms, new insights and improved techniques. Automatic image annotation achieves higher accuracy rate by combining CNN output with RNN and Natural Language Processing techniques gives state-of-the art results. The challenge is correctly understanding all the regions in image and understanding the context to generate sentence are still active research areas. The open problems are unsupervised learning in CNN, Reinforcement learning in CNN, Highly multi-task and transfer learning, Automatic learning of model structures.

CONFLICT OF INTERESTS

Authors declare no conflict of interest.

ACKNOWLEDGEMENT

None.

FINANCIAL DISCLOSURE

None.

REFERENCES

[1] UNC Vision Lab. Large scale visual recognition challenge 2015 (ilsvrc2015). http://www. image-net.org/challenges/LSVRC/2015/results, 2015. [Online; accessed 18-Dec-2015].

[2] Tsung-Yi Lin, Michael Maire, Serge J Belongie, et al[2014]. Microsoft COCO: common objects in context, CoRR, abs/1405.0312

[3] Ross B Girshick, Jeff Donahue et al. [2013]. Rich feature hierarchies for accurate object detection and semantic segmentation, 10.1109/CVPR.2014.81

[4] Y Lecun, L Bottou, Y Bengio et al .[1998] Gradient-based learning applied to document recognition. Proceedings of the *IEEE*, 86(11):2278–2324.

[5] Alex Krizhevsky, Ilya Sutskever et al.[2012] Imagenet classification with deep convo- lutional neural networks. In Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States., p. 1106–1114

THE IIOAB JOURNAL
ISSN: 0976-3104

[6] Donglai Wei et a, mNeuron: A Matlab Plugin to Visualize Neurons from Deep Models, http://vision03.csail.mit.edu/cnn_art/index.html#v_single, . [Online; accessed 19- Jun- 2016].
[7] Christian Szegedy, Wei Liu et al. [2014], Going deeper with convolutions. CoRR, abs/1409.4842,p.1-9.
[8] Kaiming He, Xiangyu Zhang, et al. [2015], Deep Residual Learning for Image Recognition, CoRR, abs/1512.03385
[9] P. Kontschieder, M. Fiterau, A. Criminisi and S. R. Bulò, [2015], "Deep Neural Decision Forests," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, p. 1467-1475.doi: 10.1109/ICCV.2015.172
[10] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, [2015], Is object localization for free? - weakly-supervised learning with convolutional neural networks. In Computer Vision and Pattern Recognition (CVPR), IEEE Conference p. 685–694
[11] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. [2014], Scalable object detection using deep neural networks. In Computer Vision and Pattern Recognition (CVPR), IEEE Conference , p 2155–2162
[12] Chong Wang, Kaiqi Huang, Weiqiang Ren, Junge Zhang, and S. Maybank[2015], Large-scale weakly supervised object localization via latent category learning. Image Processing, IEEE Transactions on, 24(4):1371–1385.
[13] Mohamed Zarka, Anis Ben Ammar, and Adel M. Alimi. Regimvid [2015],Imageclef scalable concept image annotation task: Ontology based hierarchical image annotation. In Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, p.8-11.
[14] Ning c, Jeff Donahue, Ross B Girshick. Trevor Darrell.[2014] Part-based R-CNNs for Fine-grained Category Detection, CoRR, abs/1407.3867
[15] Klein, B. and Lev, G. and Sadeh, G. and Wolf, L.[2015], Associating neural word embeddings with deep image representations using Fisher Vectors,Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on, p.4437-4446
[16] Jeff Donahue, Lisa Anne Hendricks et al. [2015] ,Long-term recurrent convolutional networks for visual recognition and description. CoRR, abs/1411.4389.
[17] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le.[2014] Sequence to sequence learning with neural networks. CoRR, abs/1409.3215
[18] Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals, and Wojciech Zaremba. [2014] Addressing the rare word problem in neural machine translation. CoRR, abs/1410.8206
[19] Oriol Vinyals , Alexander Toshev, Samy Bengio, Dumitru Erhan [2014] Show and Tell: {A} Neural Image Caption Generator, CoRR, abs/1411.4555
[20] Andrej Karpathy and Fei-Fei Li.[2014] Deep visual-semantic alignments for generating image descriptions. CoRR, abs/1412.2306
[21] R´emi Lebret, Pedro O. Pinheiro, and Ronan Collobert [2015] Phrase-based image captioning. CoRR, abs/1502.03671
[22] Virginia Fernandez Arguedas, Qianni Zhang, Krishna Chandramouli, Ebroul Izquierdo [2013] Vision Based Semantic Analysis of Surveillance Videos. Semantic Hyper/Multimedia Adaptation, p.83-125
[23] Matthew D. Zeiler and Rob Fergus [2013] Visualizing and understanding convolutional networks. CoRR, abs/1311.2901, 2013.
[24] David Silver, Aja Huang, Chris J Maddison, et al.[2016] , Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):p.484–489,
[25] Vincent Vanhoucke, Udacity, Deep Learning, ud730, https://classroom.udacity.com/courses/ud730/lessons/6370362152/concepts/63798118150923, 2015, [Online; accessed 29-Jun-2016].

ABOUT AUTHORS

Jenisha T *is a research scholar at VIT University, Vellore, India. Her area of interest includes Machine Learning,Computer Vision, Deep Learning, Cognitive Computing and Big Data Analytics. She did her B.Tech from St. Xaviers Catholic College of Engineering, Nagercoil,India. She completed her M.Tech in Information Technology in the year 2010.*

Dr. Swarnalatha Purushotham *works as Associate Professor in the School of Computer Science and Engineering,VIT University,Vellore - 632 014,Tamil Nadu, India. Her area of interest includes Digital Image Processing and Artificial Intelligence. She has published more than 40 papers in National and International Journals.*