Special Issue Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, www.ijtrd.com

Convolution Neural Network and its Application

[1]Logeswari.M, [2]Amrutha.H
[1]Assistant Professor, [2]Student,
[1,2]Department of Computer Science, St.Joseph's College of Arts and Science for Women,(Hosur, Tamilnadu, India)

Abstract - Deep Learning has been used extensively in a wide range of fields. In deep learning, Convolution Neural Networks are found to give the most accurate results in solving real world problems. In this paper, we give a comprehensive summary of the applications of CNN in computer vision and natural language processing. We delineate how CNN is used in computer vision, mainly in face recognition, scene labeling, image classification, action recognition, human pose estimation and document analysis. Further, we describe how CNN is used in the field of speech recognition and text classification for natural language processing. We compare CNN with other methods to solve the same problem and explain why CNN is better than other methods.

Keywords — Deep Learning, Convolution Neural Networks, Computer Vision, Natural Language

I. INTRODUCTION

Convolution Neural Network (CNN) is a deep learning architecture which is inspired by the structure of visual system.[1]This network is widely considered as a predecessor of CNN and it was based on the hierarchical organization between neurons transformations image.[2][3] Established the framework of CNNs by developing a multi-layer artificial neural network called as LeNet-5. LeNet-5 was used to classify handwritten digits and could be trained with the back propagation algorithm [5] which made it possible to recognize patterns directly from raw pixels thus eliminating a separate feature extraction mechanism. But even with all these advantages, due to the lack of large training data and computational power at that time, Recurrent Neural Networks (RNN) are generally applied to solve Natural Language Processing (NLP) problems. CNNs have also been applied to the problem of speech recognition which essentially is a major researched task in NLP. Speech which is the spectral representation of spoken words consists of several hundred variables and generally face problems of over fitting when trained using fully connected feed-forward networks. They also do not have built-in invariance with respect to translations. These architectures also entirely ignore the topology or hierarchy of the input. On the other hand, in CNN, shift variance is automatically obtained and it also forces the extraction of local features thus improving the performance with respect to traditional architectures [10]. To give a comprehensive review of the development of CNNs in the fields of Computer Vision and Natural Language Processing. Represent of overview of the CNN architecture and application.

II. CNN ARCHITECTURE OVERVIEW

CNN architecture differs from the traditional multilayer perceptron (MLP) to ensure some degree of shift and distortion invariant [1][2].They combine three architectural ideas to do the same:

Local Receptive Fields Shared weights

Spatial and Temporal Sub sampling

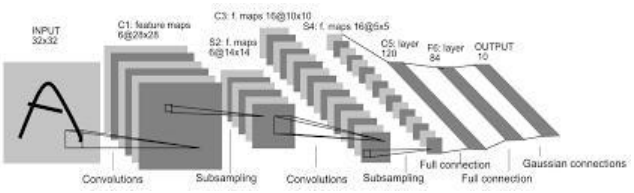

Figure.1 Convolution Neural Network.

Convolution networks are trainable multistage architectures with each stage consisting of multiple layers. The input and output of each stage are sets of arrays called as feature maps [3]. In the case of a colored image, each feature map would be a 2D array containing a color channel of the input image, a 3D array for a video and a 1D array for an audio input. The output stage represents features extracted from all locations on the input. Each stage generally consists of a convolution layer, non-linearity and a pooling layer. A single or multiple fully-connected layers are presents.

A. Convolution layer

This layer is the core building block of a CNN. The layer's parameters consist of learnable kernels or filter which exit end through the full depth of the input. Each unit of this layer receives inputs from a set of units located in small neighborhood in the previous layer.

B. Non-linearity Layer

These functions introduce nonlinearities which are desirable for multi-layer networks. The activation functions are typically sigmoid, tanh and ReLU. Compared to other functions Rectified Linear Units (ReLU)] are preferable because neural networks train several times faster.

C. Pooling Layer

The Convolution layer may be followed by the pooling layer which takes small rectangular blocks from the convolution layer and subsamples it to produce a single maximum output from the block. Pooling layer progressively reduces the spatial size of the representation, thus reducing the parameters to be computed. It also controls over fitting.

D. Fully Connected Layer

There maybe one or more fully-connected layers that perform high level reasoning by taking all neurons in the previous layer and connecting them to every single neuron in the current layer to generate global semantic information.

III. HISTORY AND APPLICATION

A. Computer Vision

Convolution neural networks are employed to identify the hierarchy or conceptual structure of an image. Instead of feeding each image into the neural network as one grid of numbers, the image is broken down into overlapping image tiles that are each fed into a small neural network.

Convolution neural networks[3][4] are trainable multi-stage architectures, with the inputs and outputs of each stage consisting of sets of arrays called feature maps. A typical CNN is composed of one, two or three such 3-layer stages, followed by a classification module**.**

- *Face Recognition*:

Face recognition constitutes a

Series of related problems-

Identifying different faces

Using on each face despite bad

lighting or different pose

Identifying unique features

Comparing identified features to existing database and determining the person's name

- *Scene Labeling:*

Each pixel is labeled with the category of the object it belongs to in scene labeling. Their method produced 320 X 240 image labeling in under a second including feature extraction.

As the context size increases with the built-in recurrence, the system identifies and corrects its own errors. A simple and scalable detection algorithm that improves mean average precision (MAP). [5] Fully convolutional networks trained end-to-end, pixels-to-pixels address the shortcomings of prior approaches of CNNs which were used for semantic segmentation in which each pixel was labeled with the class of its encoding object or region.

- *Image Classification:*

Compared with other methods CNNs achieve better classification accuracy on large scale datasets due to their capability of joint feature and classifier learning., several works made significant improvements in classification accuracy by reducing filter size or expanding the network depth.

- *Action Recognition***:**

The difficulties in developing an action recognition system are to solve the translations and distortions of features in different patterns which belong to the same action class. The three-dimensional receptive field structure of the modified CNN model provides translation invariant feature extraction capability.

- *Document Analysis*:

Many traditional handwriting recognizers use the sequential nature of the pen trajectory by representing the input in the time domain. However, these representations tend to be sensitive to stroke order, writing speed and other irrelevant parameters. This system AMAP preserves the pictorial nature of the handwriting images.

[8]LRCN is a class of models that is spatially and temporally deep and can be applied to a variety of computer vision tasks long term.

recurrent CNNs are proposed which is a novel architecture for visual recognition and description. Current models like TACoS assume a fixed spatio -temporal receptive field or simple temporal average ingraditional approaches to solve this problem typically separate out the localization, segmentation, and recognition steps. However a unified approach is followed using CNN which is directly applied on the image pixels. The implementation of neural networks is used in order to train CNN on high quality images. This approach increases the accuracy of recognizing complete str***n***eet numbers to 96% and accuracy of recognizing per digit to 97.84%.

The traditional OCR techniques can't be used for text detection in natural scene images because of variability in appearances, layout, fonts and styles as also inconsistent lighting, occlusions, orientations, and noise and background objects. An end-to-end system for text spotting (localizing and recognizing text in natural scene images) and text-based image retrieval is proposed in this work. This system is based on region proposal mechanism for detection and DCNN for recognition. This system is fast and scalable as compared to the earlier OCR systems. A novel framework is proposed to tackle this problem of distinguishing texts from background components, by leveraging the high capability of DCNN. This approach takes advantage of Maximally Stable External Regions and sliding window-based methods to achieve over 78% F-measure which is significantly higher than previous methods. While the recognition of text within scanned documents is well studied and there are many document OCR systems that perform very well, these methods do not translate to the highly variable domain of scene text recognition. When applied to natural scene images, traditional OCR techniques fail holistically, departing from the character-based recognition systems of the past. The deep neural network models at the Centre of this framework are trained solely on data produced by a synthetic text generation engine.

B. Natural Language Processing

Convolution Neural Networks have been traditionally applied in the field of Computer Vision. CNNs have provided major breakthroughs in image classification. From its inception, CNNs have been used to extract information from raw signals. Speech essentially is a raw signal and its recognition is one of the most important tasks in NLP. To explore how CNNs have been used in speech recognition in recent years. Recently, CNNs have also been applied to the tasks of sentence classification, topic categorization, sentiment analysis and many more.

- *Speech recognition:*

Convolution Neural Network share been used recently in Speech Recognition and has given better results over Deep Neural Networks (DNN). In researchers in Microsoft Corporation indicated four domains in which CNN gives better results than DNN. They are:

Noise robustness,

Distant speech

Recognition

Low- Footprint models

Channel-mismatched training-test conditions. The researchers used CNN and obtained relative 4% Word Error Rate Reduction (WERR) when trained on 1000 hours over DNN trained on same size. They used CNN structure with max out units for deploying small-footprint models to devices to get 9.3% WERR from DNN. There are certain factors of CNN due to which they give better results in Speech Recognition. Robustness of CNN is enhanced when pooling is done at a local frequency region and over-fitting is avoided by using fewer parameters to extract low-level features. Recognition (SER) has been an important application in recent times in human -cantered signal processing. CNN is used to learn salient features of SER. trained CNN for SER in two stages. In the first stage, local invariant features (LIF) are learned using sparse auto- encoder (SAE) and in the second stage, LIF is used as an input to salient descriptive feature analysis. The system developed was robust enough and was stable in complex scenes Deep CNN for SER using labelled training audio data. They used principal component analysis (PCA) technique to tackle the interferences and decrease the dimensionality. The model consisted of 2 convolution and 2 pooling layers which attained 40% classification accuracy. It performed better than SVM based classification using hand-crafted acoustic features.

Handling the unwanted noise is a major challenge in speech recognition. Researchers have built systems over the years which a

unaffected by the noise signals and used the best pooling, padding and input feature map selection strategies and evaluated on two tasks: Aurora Task and AMI meeting transcription task to test robustness. The architecture obtained 10% relative reduction over traditional CNN on AMI and 17% relative improvement over LSTM-RNN on Auroraal pointed out reasons for performance degradation in traditional DNN for speech activity detection (SAD) and used supervised data from novel channels to adapt the filters in the CNN layers to improve the performance.

- *Text classification*

NLP tasks deal with sentences and documents which are represented as a matrix at the input. Each row of a matrix corresponds to a token, which essentially is a word or in some approaches a character. Thus, each row is a vector that represents a token. These vectors are generally low-dimensional representations called as word embeddings.

Further the convolution is computed with constant or varying filter sizes and a feature map is generated. Pooling is performed over each feature map. A final feature vector is generated which is followed by a final layer which performs the necessary task at hand like classification or categorization. The location where a word lies whole sentence is of utmost important. Words which are close to one another in a sentence may not be connected in terms of meaning which is quite contrary to pixels in a specific region of an image which may be a part of a certain object Thus, [9] CNNs are generally applied only to classification tasks such as topic categorization or sentiment analysis.

CNN architecture for sentence level classification tasks on 7 datasets. In a series of experiments, the CNN is trained on top of pre-trained word vectors and with a bit of hyper parameter tuning, the research has acquired state of art results. The architecture is simple with the input layer contains sentences which are made up of word2vec word embeddings. This is followed by a convolutional layer, a max- pooling layer and a SoftMax classifier. The SoftMax layer gives the output as the probability distribution over different labels. The work adds more evidence to the fact that unsupervised pre-training of word vectors is an important factor in deep learning for NLP.

Which had a global pooling operation called as k-Max Pooling. Out of the four tasks on which the model was experimented it provided excellent performance on first three tasks and reduced the error by 25% with respect to the strongest baseline added a semantic clustering to this network and their results validated the model's effectiveness. CNN without applying any pre-trained word vectors i.e. they used high dimensional data directly. In the second method the authors employ a bag-of-words conversion in the convolution layer. Both the methods outperformed the previous methods by reducing the error rate by almost 2% and 1.5% respectively proposes a special type of deep neural network with convolutional structure for text analysis for recommending target documents to the user based on the document the user is reading. The network which is trained on a large set of web transitions, maps source-target document pairs to feature vectors, minimizing the distance between source and target documents. They both illustrate the ways to learn semantically meaningful representations of sentences. The latter out-performs the previous state-of-art semantic model

CONCLUSION

Results observed in the comparative study with other traditional methods suggest that CNN gives better accuracy and boosts the performance of the system due to unique features like shared weights and local connectivity. CNN is better than other deep learning methods in applications pertaining to computer vision and natural language processing because it mitigates most of the traditional problems. We hope that this paper gives a better understanding of why CNN is used in various applications and help others in future to use CNN in other field

References

[1] Y. LeCun, B.Boser, J. S. Denker, D.Henderson, R. E. Howard, Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in NIPS'89.

[2] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, 1998.

[3] S. Lyu and E. P. Simoncelli, "Nonlinear image representation using divisive normalization," in CVPR, 2008.

[4] LeCun, Yann, et al. "Gradient-based learning applied to document recognition." Proceedings of the IEEE 86.11 (1998): 2278-2324.

[5] Hecht-Nielsen, Robert. "Theory of the backpropagation neural network. "Neural Networks, 1989. IJCNN., International Joint Conference on. IEEE, 1989.

[6] Steinkrau, Dave, Patrice Y. Simard, and Ian Buck. "Using GPUs for machine learning algorithms." Proceedings of

the Eighth International Conference on Document Analysis and Recognition. IEEE Computer Society, 2005

[7] Chellapilla, Kumar, Sidd Puri, and Patrice Simard. "High performance convolutional neural networks for document processing." Tenth International Workshop on Frontiers in Handwriting Recognition. Suvisoft, 2006.

[8] Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets." Neural computation 18.7 (2006): 1527-1554.

[9] Bengio, Yoshua, et al. "Greedy layer-wise training of deep networks. "Advances in neural information processing systems 19 (2007): 153.

[10] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Image Net classification with deep convolutional neural networks." Advances in neural information processing systems. 201