ISSN:0975-9646

Applications of Convolutional Neural Networks

Ashwin Bhandare[#1], Maithili Bhide[*2], Pranav Gokhale[*2], Rohan Chandavarkar[*2]

[#]*Department of Information Technology, Pune Institute of Computer Technology*
Savitribai Phule Pune University, Pune

[*]*Department of Computer Engineering, Pune Institute of Computer Technology*
Savitribai Phule Pune University, Pune

***Abstract*— In recent years, deep learning has been used extensively in a wide range of fields. In deep learning, Convolutional 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 labelling, 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, Convolutional Neural Networks, Computer Vision, Natural Language**

I. Introduction

Convolutional Neural Network (CNN) is a deep learning architecture which is inspired by the structure of visual system. In 1962, Hubel and Wiesel [1] in their classic work on cat’s primary visual cortex found that cells in the visual cortex are sensitive to small sub-regions of the visual field called as receptive field. These cells are responsible for detecting light in the receptive fields. Neocognitron proposed by Fukushima [2] was the first model which was simulated on a computer and was inspired from the works of Hubel and Wiesel. This network is widely considered as a predecessor of CNN and it was based on the hierarchical organization between neurons for the transformation of image. LeCun et al. [3, 4] 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 backpropagation 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, LeNet-5 failed to perform well on complex problems such as video classification.

Since the advent of GPGPUs and their use in machine learning [6], the field of CNN has gone through a renaissance phase. Several publications have established more efficient ways to train convolutional neural networks using GPU computing [7-9]. Krizhevsky et al. proposed a deep CNN architecture called as AlexNet [10] which demonstrated significant improvement in image classification task. AlexNet is very similar to the classic LeNet-5 albeit a deeper structure. Following the success of AlexNet several publications such as GoogleNet [11], VGGNet [12], ZFNet [13] and ResNet [14] have shown to improve its performance.

Recurrent Neural Networks (RNN) are generally applied to solve Natural Language Processing (NLP) problems [15] since RNNs resemble how human beings process language. But recently, CNNs which seem less intuitive in solving such problems than RNNs have been applied to solve NLP problems such as sentiment analysis, spam detection or topic categorization. CNNs have achieved state-of-art or competitive results. 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 overfitting when trained using fully connected feed-forward networks [16]. 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.

In this paper we try to give a comprehensive review of the development of CNNs in the fields of Computer Vision and Natural Language Processing. We give an overview of the CNN architecture in Section II. Next, we discuss the applications of CNN and illustrate the increased efficiency that it provides in Section III. Finally, we conclude the paper in Section IV.

II. CNN Architecture overview

CNN architecture differs from the traditional multilayer perceptrons (MLP) to ensure some degree of shift and distortion invariance [16]. They combine three architectural ideas to do the same:

- Local Receptive Fields
- Shared weights
- Spatial and Temporal Subsampling

We have mentioned many CNN architectures in the literature but their basic components are very similar. Let us consider the typical convolutional network architecture for recognizing characters in Fig. 1 [3]

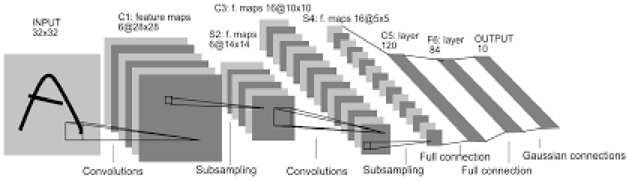

Fig. 1 Convolutional Neural Network

Convolutional 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. In the case of a coloured image, each feature map would be a 2D array containing a colour 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 present after several convolution and pooling layers.

A. Convolution layer

This layer is the core building block of a CNN. The layer's parameters consist of learnable kernels or filters which extend through the full depth of the input. Each unit of this layer receives inputs from a set of units located in small neighbourhood in the previous layer. Such a neighbourhood is called as the neuron's receptive field in the previous layer. During the forward pass each filter is convolved with input which produces a map. When multiple such feature maps that are generated from a multiple filters are stacked they form the output of the convolution layer. The weight vector that generates the feature map is shared which reduces the model complexity.

B. Non-linearity Layer

This is a layer of neurons which apply various activation functions. 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) [17] 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 [19-21]. Pooling layer progressively reduces the spatial size of the representation, thus reducing the parameters to be computed. It also controls overfitting. Pooling units apart from the maximum function can also perform other functions like average [18] or L2-norm pooling.

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. Applications of CNN

A. Computer Vision

Convolutional 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.

Convolutional neural networks are trainable multi-stage architectures [3], [4] with the inputs and outputs of each stage consisting of sets of arrays called feature maps. If the input is a colour image, each feature map is a 2D array containing a colour channel of the input image, for a video or a volumetric image it would be a 3D array. Each feature extracted at all locations on the input is represented by a feature map at the output. Each stage is composed of a filter bank layer, a non-linearity layer and a feature pooling layer. A typical CNN is composed of one, two or three such 3-layer stages, followed by a classification module.

1) *Face Recognition*: Face recognition constitutes a series of related problems-

- Identifying all the faces in the picture
- Focussing on each face despite bad lighting or different pose
- Identifying unique features
- Comparing identified features to existing database and determining the person's name

Faces represent a complex, multi-dimensional, visual stimulus which was earlier presented using a hybrid neural network combining local image sampling, a self-organizing map neural network and a convolutional neural network. The results were presented using Karhunen-Loe`ve transform in place of the self-organizing map which performed almost as well (5.3% error versus 3.8%) and a multi-layer perceptron which performed poorly (40% error versus 3.8%). [22]

2) *Scene Labelling:* Each pixel is labelled with the category of the object it belongs to in scene labelling. Clement Farabet et al proposed a method using a multiscale convolutional network that yielded record accuracies on the Sift Flow Dataset (33 classes) and the Barcelona Dataset (170 classes) and near-record accuracy on Stanford Background Dataset (8 classes). Their method produced 320 X 240 image labelling in under a second including feature extraction. [24].

Recurrent architecture [25] for convolutional neural network suggests a sequential series of networks sharing the same set of parameters. The network automatically learns to smooth its own predicted labels. 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) by more than 30% relative to the previous best result on VOC 2012—achieving a mAP of 53.3% was suggested by researchers at UCB and ICSI. It was called as R-CNN: Regions with CNN features as it combined region proposals with CNN features. [26]

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 labelled with the class of its encoding object or region. Fully convolutional networks adapted from contemporary classification networks such as AlexNet[10], GoogleNet[12] and VGG net[11] achieve state-of-the-art segmentation of PASCAL VOC (20% relative improvement to 62.2% mean IU on 2012), NYUDv2, and SIFT Flow, while inference takes less than one fifth of a second for a typical image.[27]

Since the past two years Deep Convolutional Neural Networks (DCNNs) have greatly improved the performance of computer systems on problems in image classification (Krizhevsky et al., 2013; Sermanet et al., 2013; Simonyan & Zisserman, 2014; Szegedy et al., 2014; Papandreou et al., 2014). The work of George Papandreou, Liang-Chieh Chen et al. shows that responses at the final layer of DCNNs are not sufficiently localized for accurate object segmentation. A fully connected Conditional Radom Field (CRF) is used to overcome this problem and gives a 71.6% IOU accuracy in the test set for a new state-of-art at the PASCAL VOC-2012 semantic image segmentation task. [28]

3) *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.[29] Krizhevsky et al. [10] develop the AlexNet and achieve the best performance in ILSVRC 2012. Following the success of the AlexNet, several works made significant improvements in classification accuracy by reducing filter size[32] or expanding the network depth[12], [11].

A fast, fully parameterizable GPU implementation of CNN published benchmark results for object classification (NORB, CIFAR10) with error rates of 2.53%, 19.51%.[39] GPU code for image classification is upto two magnitudes faster than its CPU counterpart.[32], [33] Multi-column deep neural networks(MCDNN) can outperform all previous methods of image classification and demonstrate that pre-training is not necessary(though sometimes beneficial for small datasets) while decreasing the error rate by 30-40%[34]. Non-saturating neurons and efficient GPU implementation of the convolution operation resulted in a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry in the ILSVRC-2012 competition for classification of 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes [35].

Hierarchical Deep Convolutional neural Networks (HD-CNN) are based on the intuition that some classes in image classification are more confusing than other classes. It builds on the conventional CNNs which are N-way classifiers and follows the coarse-to-fine classification strategy and design module. HD-CNN with CIFAR100-NIN building block is seen to show a testing accuracy of 65.33% which is higher than the accuracy for other standard deep models and HD-CNN models on CIFAR100 dataset. [36]

Fine grained image classification systems are based on the principle of identifying foreground objects to extract discriminative features. Applying visual attention to fine grained classification task using deep neural network using the attention derived from the CNN trained with the classification task can be conducted under the weakest supervision setting where only the class label is provided in contrast to other methods that require object bounding box or part landmark to train or test. This method gives the best accuracy on CUB200-2011 dataset under the weakest supervision setting.[37]

4) 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. Earlier approaches involved construction of motion history images, use of Hidden Markov Models or more recently action sketch generation. The three-dimensional receptive field structure of the modified CNN model [38] provides translation invariant feature extraction capability, and the use of shared weight also reduces the number of parameters in the action recognition system. Researchers at Stanford University suggested an improvement to the common approaches in visual recognition which relied on SIFT[39] and HOG[40] using Independent Subspace Analysis (ISA) algorithm which is an extension of Independent Component Analysis (ICA) which is well-known for its use in natural image statistics[41]. ISA algorithm to learn invariant spatio-temporal features from unlabelled video data applied on the Hollywood2 and YouTube action datasets gives classification accuracy of 53.3% and 75.8% [42] respectively, which is approximately 5% better than previously published results.

Temporal pyramid pooling based convolutional neural network for action recognition [43] avoids the risk of missing important frames and requires much fewer training data while providing superior results on Hollywood2 and HMDB51 datasets. Two stream CNN architecture which incorporates both spatial and temporal networks gives competitive results on the standard UCF101 and HMDB51 video actions benchmarks.[44] Pose-based Convolutional Neural Network descriptor (P-CNN)[45] for action recognition targets human action recognition in videos. Performance of appearance-based (App) and flow-based (OF) P-CNN features show a cumulative accuracy of 73.4%maP for JHMDB-GT and 60.8%maP for the MPPII-Cooking Pose estimation datasets. R*CNN [46] trains action specific models and feature maps jointly achieving 90.2% mean AP on the Pascal VOC Action dataset outperforming all other approaches in the field by a significant margin.

3D CNN model for action recognition extracts features from both spatial and temporal dimensions by performing 3D convolutions thereby capturing motion information encoded in multiple adjacent frames. Intelligent video surveillance, customer attributes, and shopping behaviour analysis are examples of real-world environments in which 3D CNN outperforms the cube frame-based 2D CNN model, SPM cube gray, and SPM MEHI. [47] The 3D CNN model is most effective when the number of positive samples is less and achieves an overall accuracy of 90.2% as compared to 91.7% achieved by the HMAX model (Jhuang et al., 2007). The reconfigurable convolutional neural network[48] suggested by Wang et al. advanced the existing approaches and achieved an average accuracy of 81.2% on the CAD120 dataset, 60.1% on the OA1 dataset and 45.0% on the OA2 dataset which is significantly better than the previous accuracies shown by models suggested by Xia et al.[49] and Ji et al.[50]

The need of the hour in 3D deep learning models with extended connectivity using CNNs is scaling up the computations to support large datasets and accelerating the training on the models for high performance. This can be achieved using multi-core CPUs and GPUs by parallelizing the training of models and achieving data and model parallelism. The code scales up well on multi-cores and GPUs, with a speedup of 10x on CPUs and achieves almost 12x on GPUs compared to the serial version. M. Sai Rajeshwar and other researchers concluded that 3D-CNN code scales up best on CPUs when the convolution step is implemented with a highly parallel FFT based approach, thereby achieving the performance comparable to GPUs using OpenMP. [51]

5) *Human Pose Estimation:* Human-pose recognition is a long-standing problem in computer vision. This is primarily because of high dimensionality of the input data and the high variability of possible body poses. Traditional approaches [52] [53] tend to depend on appearance cues to predict the human pose rather than motion-based features. Because both the local evidence and the global structure are hand crafted, there is a large scope for errors. Hence this system tries to learn both the local features and the global structure using a convolutional neural network. This system [54] uses motion-based features to outperform existing state-of-the-art techniques to predict the human pose.

Deep Pose [58], developed in 2014, is the first application of Deep Neural Networks for human pose classification. In this model pose estimation is formulated as a joint regression problem. The location of each body joint is regressed to using as an input the full image and a 7-layered generic convolutional DNN. There are two advantages of this formulation. First, the DNN is capable of capturing the full context of each body joint – each joint regressor uses the full image as a signal. Second, the approach is substantially simpler to formulate than methods based on graphical models – no need to explicitly design feature representations and detectors for parts or explicitly design a model topology and interactions between joints. The results of this model for the four most challenging limbs namely, lower and upper arms and legs showed considerable improvement than the earlier approaches[59][60]. The performance for the estimation of legs improved a great deal from 0.74 to 0.78. In traditional systems like SIFT[61] or HOG[62] or Deep Pose[58] for human pose recognition much work is devoted to engineering the system to produce the vector representation that is sensitive to class (e.g. head, hands, torso) while remaining invariant to the various nuisance factors (lighting, viewpoint, scale, etc.) However, the non-rigid structure of the body, the necessity for precision (deep recognition systems often throw away precise location information through pooling), and the complex, multi-modal nature of pose contribute to the problems of the traditional networks. This paper [63] tries to use convolutional networks to learn human poses. In particular, they present a two-stage filtering approach whereby the response maps of CNN part detectors are denoised by a second process informed by the part hierarchy to finally make out the respective human pose. Recognition of human poses from 2D videos has been a long-standing task for machines. Earlier (HOG-KDE)[55], which used spatio-temporal features and 2DCNN-EM[56] were used for human pose modelling. However, these models used a very high number of frames in a sequence which in turn increased the cost of computation. 3D CNN model [57] applies CNN on RGB videos to obtain an output in 3 dimensional convolutions. The researchers use the time dimension in videos as the third dimension in convolutional space to achieve the same. State-of-art performance on a selected Human 3.6M dataset is obtained using this CNN architecture. An improvement of 11% was observed in human pose estimation from the earlier (HOG-KDE) [55].A heterogeneous multi-task learning architecture [66] was proposed with deep neural convolutional network for human pose estimation. This framework consists of two parts:

- Pose regression
- Body part detection using a sliding window classifier.

Human pose estimation has been done in traditional architectures [67] using depth maps but in today's world majority of the visual media is in the form of 2D images and hence it is extremely important to do human pose estimation using them. This approach [66] is better than the earlier ones because of the reduced number of parameters as compared to the fully connected models and their intuitive structure. The pose estimation models used earlier are sensitive to noise and graphical models and lack expressiveness to model complex human poses. In this Dual Source Deep CNN [71], the model takes a set of image patches as input and then learns the appearance of each part by considering their holistic view in the full body. Then joint detection and joint localization is done, the outputs of which are combined to form an estimate of the human pose. The proposed method performs better when compared with Tompson et al. [70] when the normalized distances are large and performs worst when the same distances are small. New hybrid architecture [64] called as the Spatial Model that combines a deep Convolutional neural network and a Markov random field is successfully applied to the long-standing task of human pose estimation. Spatial-Model does not have much impact on the accuracy of low radii threshold, however for large radii, the performance increases by around 8-12%. Thus, the unification of a novel CNN Part-Detector coupled with MRF inspired Spatial Model into a single learning framework outperforms all the existing architectures. State-of-the-art performance for human pose estimation has been achieved using Deep Convolutional Neural Networks. These networks use two approaches:

- Regressing heat maps of each body part
- Learning deep-structured output

However, in these methods the geometric constraints amongst body parts which are essential to ascertain joint consistency are missing.

In the previous method [64] the parameter space for learning the spatial constraints with convolutional kernels is too large, thus making the learning difficult. Thus DCNN can be trained more effectively by incorporating DCNN with deformable mixture of parts model into an end to end framework. This framework [65] improves the performance in several widely used benchmarks like number of message passing layers, components investigation, qualitative evaluation, etc. Similar for face tracking, a face tracker [69] is proposed, adjusted to each person's face chrominance value. Based on the face bounding box CNNs are employed to find out the face orientation and alignment. Employing local tracking techniques [68] would usually be futile in order to get accurate results because of variations in the surroundings with respect to lighting, equipment, resolution, etc. thus not being able to accurately estimate the head pose. An image is used as the first layer of CNN, in the second the output of the first layer is convolved with a series of filters in the second. In the third the output of the layer is subsampled. This approach was tested on the Boston Face Dataset and error dropped down from 5.7% to 4.6% as compared to the earlier architectures

6) *Document Analysis:* Many traditional handwriting recognizers [72] 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. This model [73] uses a combination of CNNS, Markov models, EM algorithm and encoding of words into low resolution images to recognize handwriting. Using this system error rates dropped to 4.6% and 2.0% from 5% for word and character errors respectively. This model [74] was implemented using two different practices: a database was expanded by adding a new generic collection of elastic distortions. Convolutional neural networks were used. The highest performance was achieved on the MNSIT [3] data set using elastic distortion and convolutional neural networks. As compared to the 2 layer Multi-layer perceptron models error rate of 1.6% this model achieves an error rate of 0.4%, thus significantly improving the performance. This is because MCP suffered from the problem of inverse problem or transformation invariance. Though OCR and other systems[75][76] have already reached high recognition rates for printed documents recognition of characters in natural scene images is a challenging task for the system because of complex background, low resolution, non-uniform light etc. Thus this model proposes complex colour scene text image recognition, based on specific convolution neural network architecture. Using this system [77] the average recognition rate is found to be an improved 84.53% ranging from 93.47% for clearer images and 67.86% for seriously distorted images using the ICDAR 2003 dataset. In this system [78] a convolutional neural based text detection system was presented which learnt to automatically extract its own feature set instead of using a hand crafted one. The network also learnt to multiline or badly localized text. This proposed approach outperforms the other systems [79] with 54% and 61% precision and recall rates. In the earlier systems, in a cluttered background and a free environment, detecting text is a challenging task as the image background and the text itself are unpredictable. In this architecture [78] feature extraction and classification are performed jointly in the first step. It is a 2 step process which are as follows:

- Select an appropriate set of features
- Learning takes place

Character recognition in complex natural scenes can be quite challenging mainly because of non-contrasting backgrounds, low resolution, de-focused and motion blurred images. Previous approaches in classifying characters used multiple hand-crafted features [80] and template-matching [81]. In contrast, CNNs learn features all the way from pixels to the classifier. This approach [86] is a supervised learning approach whereas the earlier algorithms all were trained unsupervised.

LRCN [84] is a class of models that is spatially and temporally deep and can be applied to a variety of computer vision tasks. In this paper 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 averaging. Recurrent convolutional models are doubly deep. They can be compositional in spatial as well as temporal layers. Such models have advantages when target concepts are complex. This approach compared with the TACoS [83] multilevel approach achieves a performance of 28.8% whereas the TACoS achieve a performance of 26.9%. Recognizing arbitrary multiple digits from Street view imagery is a very challenging task. This difficulty arises due to the wide variability in the visual appearance of text in the wild on account of a large range of fonts, colours, styles, orientations, and character arrangements. Traditional approaches [85] to solve this problem typically separate out the localization, segmentation, and recognition steps. However in this paper, a unified approach is followed using CNN which is directly applied on the image pixels. The DistBelief [86] implementation of neural networks is used in order to train CNN on high quality images. This approach increases the accuracy of recognizing complete street 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 {localising and recognising text in natural scene images {and text based image retrieval is proposed in this work [86]. 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 [90] 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 Extremal Regions and sliding window based methods to achieve over 78% F-measure which is significantly higher than previous methods [87][88][89]. 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

Convolutional 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 [3, 4]. Speech essentially is a raw signal and its recognition is one of the most important tasks in NLP. We 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.

1) Speech Recognition: Convolutional Neural Networks have been used recently in Speech Recognition and has given better results over Deep Neural Networks (DNN). In 2015, 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[92].

The researchers used CNN and obtained relative 4% Word Error Rate Reduction (WERR) when trained on 1000 hours of Kinect distant, over DNN trained on same size. They used CNN structure with maxout 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. Dimitri Palaz et al. [93] show that in CNN, direct modelling can be done for the relationship between raw speech signal and the phones and Automatic Speech Recognition (ASR) system comparable to standard method can be developed. In this paper, it is shown that features of such ASR systems are less affected by noise than the MFCC features.

Distant Speech can be captured by either a Single Distant Microphone (SDM) or Multiple Distant Microphone (MDM). The main issues which need to be tackled in Distant Speech Recognition are multiple acoustic sources and constant reverberation. In 2014, Pawel Swietojanski et al. [94] found that WER is improved using CNN by 6.5% relative to DNN for distant speech recognition and 15.7% over a Gaussian Mixture Model. In case of cross-channel convolution, the WER enhances by 3.5% compared to DNN and 9.7% over GMM. The researchers inferred that in SDM the WER produced is comparable to DNN trained on beamforming across 8 microphones. In MDM the max-channel pooling gave better results than multi-channel convolution. Generally Recurrent Neural Network (RNN) is used in Distant Speech Recognition by many people for accurate results. Suyoun Kim et al. at Carnegie Mellon University embedded an attention mechanism in RNN to get a desired output in one step. By comparing the result obtained on ChiME-3[95] Challenge Task, they conclude that it is similar to beamforming using purely data driven method. Researchers at Stanford University combined the approaches of RNN and CNN for better output. CNN is used for frame level classification and RNN is used with Connectionist Temporal Classification for decoding frames in a sequence of phenomes. They obtained 22.1% on TIMIT dataset [96] with CNN and phone sequence has error of 29.4%.

Speech Emotion Recognition (SER) has been an important application in recent times in human-centred signal processing [97]. CNN is used to learn salient features of SER. In 2014, Qirong Mao et al. 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. In 2015, W. Q. Zheng et al. [98] used 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 are unaffected by the noise signals. Yanmin Qian et al. [99] used the best pooling, padding and input feature map selection strategies and evaluated on two tasks: Aurora4 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 Aurora4. In 2014, Samuel Thomas et al. [100] 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. Their experiments show that CNN can adapt well in novel channel scenarios with limited supervised adaption data. In 2015, Rafael M. Santos et al. [101] proposed a system with CNN- HMM model which did not use additional method for treatment of noise but was still robust. It reduced the Equal Error Rate (EER) in word recognition tasks in presence of additive noise.

2) 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. Word embedding is a set of language modelling and feature learning technique in NLP where words from the vocabulary are mapped to vectors of real-numbers in a low-dimensional space [102]. Word embedding techniques like word2vec [102] and GloVe [103] are generally used. Word2vec vectors were trained on 100 billion words of Google News which are available publicly. 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. Location invariance and local compositionality which are the main ideas behind CNNs make sense in case of computer vision applications but they do not make much sense in the case of NLP. The location where a word lies in the 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, CNNs are generally applied only to classification tasks such as topic categorization or sentiment analysis. Convolution and pooling operations lose information about the order of words. So applications like PoS tagging or entity extraction where sequence is important are harder to fit into the traditional CNN architecture

The work of Yoon Kim [104] evaluates 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 hyperparameter 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. Their baseline CNN-rand model which contains randomly initialized words did not perform well but their simple CNN-static word containing pre-trained static words produced comparable results and state –of-art results in one task. The CNN-non-static model gave competitive results and improved the performance by almost 2% with respect to other models which were mainly based on RNN, Autoencoders and Support Vector Machines (SVM). The CNN-multichannel which contained two sets of word vectors provided two state-of-art results. Thus the models in this paper gave 4 out of 7 state-of-art results. A similar but a more complex model was proposed by Kalchbrenner et al. [105] 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. Wang et al. [106] added a semantic clustering to this network and their results validated the model's effectiveness.

Johnson and Zhang [107] trained 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. In [108] the authors extended their model with an additional unsupervised learning which learns embeddings of text regions. Their model outperformed the previous best results on IMDB by nearly 0.9%. Their approach works well for long texts but their performance on short texts is not clear.

While building CNN architecture various hyperparameters have to be considered like input representations (word2vec, one-hot or GloVe), size and number of filters, pooling functions and activation functions. In [109], the authors perform a sensitivity analysis by varying various hyperparameters in CNN and explore their impact on model performance. The authors drew some important conclusions from their study. Non-static word embeddings should be avoided for sufficiently large datasets, filter region size impact the performance drastically and 1-max pooling outperforms other pooling strategies are some of the important conclusions from this study.

[110] performs relation extraction and classification tasks. The author uses word vectors along with the relative position of the words with respect to the entities of interest as an input to the convolution layers. The works of Sun et al. [111] takes into consideration the semantic representations of context, mention and entity and encodes them into continuous vector space and ultimately prove to be of maximum advantage for entity disambiguation. Their models showcase an improvement of nearly 2-3% on two datasets. [112] explores a similar model.

Gao et al. [113] 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. The work of Shen et al. [114] is on similar lines. They both illustrate the ways to learn semantically meaningful representations of sentences. The latter out-performs the previous state-of-art semantic models.

Semantic embeddings from hashtags [115] is a CNN architecture which predicts hashtags for Facebook posts and at the same time generate meaningful embeddings for words and sentences. These embeddings are then successfully applied to document recommendation task.

Recently there has been on-going research in applying CNNs directly to characters. Cicero et al. [116] learns character-level embeddings, and joins them with pre-trained word embeddings. Then this model uses a CNN for Part of Speech tagging. It has produced results for two languages: English having accuracy of 97.32% on Penn Tree-bank WSJ corpus and Portuguese with 97.47% accuracy on the Mac-Morphus Corpus where the error has been reduced by almost 12% as compared to the best previous result. Zhang et al. [117] [119] in his two publications uses CNN to learn directly from characters. With the help of a relatively deep network they apply the model to text categorization and sentiment analysis. The results vary according to the size of the dataset, choice of alphabet and whether the texts are curated or not. They also throw light on the potential of character level CNN models. The publication Character-aware neural language model [119] employs CNN whose output is given to a long short-term memory (LSTM) RNN. Inspite of having 60% less parameter it performs on par with existing state-of-art results on the English Penn Treebank. It outperforms the previous word-level models on languages with rich morphology. The results prove that character-level input is enough for language modelling.

IV. 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 fields

References

[1] Hubel, David H., and Torsten N. Wiesel. "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex." *The Journal of physiology* 160.1 (1962): 106-154.

[2] Fukushima, Kunihiko. "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position."*Biological cybernetics* 36.4 (1980): 193-202.

[3] Le Cun, B. Boser, et al. "Handwritten digit recognition with a back-propagation network." *Advances in neural information processing systems*. 1990.

[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. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.

[11] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[12] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556*(2014).

[13] Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." *European Conference on Computer Vision*. Springer International Publishing, 2014

[14] He, Kaiming, et al. "Deep residual learning for image recognition." *arXiv preprint arXiv:1512.03385* (2015).

[15] Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks." *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013

[16] LeCun, Yann, and Yoshua Bengio. "Convolutional networks for images, speech, and time series." *The handbook of brain theory and neural networks*3361.10 (1995): 1995.

[17] Nair, Vinod, and Geoffrey E. Hinton. "Rectified linear units improve restricted boltzmann machines." *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. 2010.

[18] T. Wang, D. Wu, A. Coates, and A. Ng, “End-to-end text recognition with convolutional neural networks,” in ICPR, 2012.

[19] J. Yang, K. Yu, Y. Gong, and T. Huang, “Linear spatial pyramid matching using sparse coding for image classification,” in CVPR, 2009.

[20] Y. Boureau, J. Ponce, and Y. LeCun, “A theoretical analysis of feature pooling in visual recognition,” in ICML, 2010.

[21] M. Ranzato, F. J. Huang, Y. Boureau, and Y. LeCun, “Unsupervised learning of invariant feature hierarchies with applications to object recognition,” in CVPR, 2007.

[22] Lawrence, Steve, et al. "Face recognition: A convolutional neural-network approach." *IEEE transactions on neural networks* 8.1 (1997): 98-113

[23] LeCun, Yann, Koray Kavukcuoglu, and Clément Farabet. "Convolutional networks and applications in vision." *ISCAS*. 2010.

[24] Farabet, Clement, et al. "Learning hierarchical features for scene labeling."*IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013): 1915-1929.

[25] Pinheiro, Pedro HO, and Ronan Collobert. "Recurrent Convolutional Neural Networks for Scene Labeling." *ICML*. 2014.

[26] Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014.

[27] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[28] Chen, Liang-Chieh, et al. "Semantic image segmentation with deep convolutional nets and fully connected crfs."*arXiv preprint arXiv:1412.7062*(2014)

[29] Gu, Jiuxiang, et al. "Recent Advances in Convolutional Neural Networks."*arXiv preprint arXiv:1512.07108* (2015)..

[30] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in ECCV, 2014.

[31] Ciresan, Dan C., et al. "Flexible, high performance convolutional neural networks for image classification." *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*. Vol. 22. No. 1. 2011.

[32] D. Strigl, K. Kofler, and S. Podlipnig. Performance and scalability of gpu-based convolutional neural networks. Parallel, Distributed, and Network-Based Processing, Euromicro Conference on, 0:317–324, 2010.

[33] R. Uetz and S. Behnke. Large-scale object recognition with CUDA-accelerated hierarchical neural networks. In IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS), 2009.

[34] Ciregan, Dan, Ueli Meier, and Jürgen Schmidhuber. "Multi-column deep neural networks for image classification." *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012.

[35] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.

[36] Yan, Zhicheng, et al. "HD-CNN: hierarchical deep convolutional neural network for image classification." *International Conference on Computer Vision (ICCV)*. Vol. 2. 2015.

[37] Xiao, Tianjun, et al. "The application of two-level attention models in deep convolutional neural network for fine-grained image classification."*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

[38] Kim, Ho-Joon, Joseph S. Lee, and Hyun-Seung Yang. "Human action recognition using a modified convolutional neural network." *International Symposium on Neural Networks*. Springer Berlin Heidelberg, 2007.

[39] D. Lowe. Object recognition from local scale-invariant features. In ICCV, 1999.

[40] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In CVPR, 2005.

[41] A. Hyvarinen, J. Hurri, and P. Hoyer. Natural Image Statistics. Springer, 2009.

[42] Le, Quoc V., et al. "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis." *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011.

[43] Wang, Peng, et al. "Temporal pyramid pooling based convolutional neural networks for action recognition." *arXiv preprint arXiv:1503.01224* (2015).

[44] Simonyan, Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." *Advances in Neural Information Processing Systems*. 2014.

[45] Chéron, Guilhem, Ivan Laptev, and Cordelia Schmid. "P-CNN: Pose-based CNN features for action recognition." *Proceedings of the IEEE International Conference on Computer Vision*. 2015.

[46] Gkioxari, Georgia, Ross Girshick, and Jitendra Malik. "Contextual action recognition with r* cnn." *Proceedings of the IEEE International Conference on Computer Vision*. 2015.

[47] Ji, Shuiwang, et al. "3D convolutional neural networks for human action recognition." *IEEE transactions on pattern analysis and machine intelligence*35.1 (2013): 221-231.

[48] Wang, Keze, et al. "3D human activity recognition with reconfigurable convolutional neural networks." *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014.

[49] Xia and J. Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In CVPR, pages 2834–2841, 2013

[50] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell., 35(1):221–231, 2013.

[51] Rajeswar, M. Sai, et al. "Scaling Up the Training of Deep CNNs for Human Action Recognition." *Parallel and Distributed Processing Symposium Workshop (IPDPSW), 2015 IEEE International*. IEEE, 2015.

[52] Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: CVPR. (2008)

[53] Weiss, D., Sapp, B., Taskar, B.: Sidestepping intractable inference with structured ensemble cascades. In: NIPS. (2010)

[54] MoDeep: A Deep Learning Framework Using Motion Features for Human Pose Estimation Arjun Jain, Jonathan Tompson, Yann LeCun and Christoph Bregler New York University arXiv:1409.7963v1 [cs.CV] 28 Sep 2014

[55] Bugra Tekin, Xiaolu Sun, Xinchao Wang, Vincent Lepetit, and Pascal Fua. Predicting people's 3D poses from short sequences. arXiv preprint arXiv:1504.08200, 2015.

[56] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Kosta Derpanis, and Kostas Daniilidis. Sparseness meets deepness: 3D human pose estimation from monocular video. arXiv preprint arXiv:1511.09439, 2015.

[57] Agne Grinciunaite, Amogh Gudi, Emrah Tasl,i Marten den Uyl Human Pose Estimation in Space and Time using 3D CNN.arXiv:1609.00036v1,2016

[58] DeepPose: Human Pose Estimation via Deep Neural Networks Alexander Toshev Christian Szegedytoshev@google.com szegedy@google.com

[59] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. International Journal of Computer Vision, 61(1):55–79, 2005

[60] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. Computers, IEEE Transactions on, 100(1):67–92, 1973

[61] D. G. Lowe. Object recognition from local scale-invariant features. In Computer vision, 1999. The proceedings of the seventh IEEE international conference on, volume 2, pages 1150–1157. Ieee, 1999.

[62] N. Dalal and B. Triggs. "Histograms of oriented gradients for human detection." "In Computer Vision and Pattern Recognition, 2005." CVPR 2005. IEEE Computer Society Conference on, volume 1, pages 886–893. IEEE, 2005. 2

[63] Arjun Jain, Jonathan Tompson, Mykhaylo Andriluka, Graham W. Taylor, Christoph Bregler. "Learning Human Pose Estimation Features with Convolutional Networks." arXiv:1312.7302v6 [cs.CV] 23 Apr 2014

[64] Jonathan Tompson, Arjun Jain, Yann LeCun, Christoph Bregler. "Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation." New York University

[65] Wei Yang Wanli Ouyang∗ Hongsheng Li Xiaogang Wang. "End-to-End Learning of Deformable Mixture of Parts and Deep Convolutional Neural Networks for Human Pose Estimation."

[66] Sijin Li, Zhi-Qiang Liu, Antoni B. Chan. "Heterogeneous Multi-task Learning for Human Pose Estimation with Deep Convolutional Neural Network." 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops

[67] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. "Real-time human pose recognition in parts from single depth images." CVPR, 2011.

[68] JURE KOVAC, P. P., AND SOLINA, F. 2003. "Human skin colour clustering for face detection." In IEEE International Conference on Computer as a Tool, vol. 2.

[69] Stylianos Asteriadis, KostasKarpouzis, Stefanos Kollias. "Face Tracking and Head Pose Estimation using Convolutional Neural Networks." Image, Video, Multimedia Lab, National Technical University of Athens, Greece

[70] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. "Joint training of a convolutional network and a graphical model for human pose estimation." in NIPS, 2014.

[71] Xiaochuan Fan, Kang Zheng, Yuewei Lin, Song Wang. "Combining Local Appearance and Holistic View: Dual-Source Deep Neural Networks for Human Pose Estimation." arXiv:1504.07159v1 [cs.CV] 27 Apr 2015

[72] Guyon et al. "H.1991 design of a neural network character recognizer for a touch terminal Pattern Recognition." 24(2):105-119

[73] Yoshua Bengio, Yann LE Cun, Donni Henderson. "Globally Trained Handwritten Word recognizer using Spatial representation, Convolutional Neural Networks and Hidden Markov Models."

[74] Y. Simard, Dave Steinkraus, John C. Platt. "Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis Patrice." Microsoft Research, One Microsoft Way, Redmond WA 98052

[75] C. Garcia and M. Delakis. "Convolutional face finder: A neural architecture for fast and robust face detection." *IEEE Transactions*on pattern analysis and machine intelligence, 26(11),November 2004.

[76] M. Yokobayashi and T.Wakahara. "Segmentation and recognition of characters in scene images using selective binarization in color space and gat correlation." *Eighth International Conference*on Document Analysis and Recognition ICDAR'05, 1:167–171, 29 Aug.-1 Sept. 2005

[77] Zohra Saidane and Christophe Garcia. "Automatic Scene Text Recognition using a Convolutional Neural Network." Orange Labs 4, rue du Clos Courtel BP 9122635512 Cesson S´evign´e Cedex – France

[78] "Manolis Delakis and Christophe Garcia." Text detection with convolutional neural networks.

[79] Lucas, S., Panaretos, A., Sosa, L., Tang, A., Wong, S.,Young, R., Ashida, K., Nagai, H., Okamoto, M., Yamamoto, H., Miyao, H., Z., J., Ou, W.-W., Wolf, C.,Jolion, J.-M., Todoran, L., Worring, M., and Lin, X.(2005). ICDAR 2003 robust reading competitions:entries, results, and future directions. *International*Journal on Document Analysis and Recognition, 7(2-3):105–122.

[80] T. E. de Campos, B. R. Babu, and M. Varma. "Character recognition in natural images." In *Proceedings of the International* Conference on Computer Vision Theory an Applications, Lisbon, Portugal, February 2009.

[81] T. Yamaguchi, Y. Nakano, M. Maruyama, H. Miyao, and T. Hananoi. "Digit classification on signboards for telephone number recognition." In *ICDAR*, pages 359–363, 2003.

[82] Pierre Sermanet, Soumith Chintala and Yann LeChun. "Convolutional Neural Networks Applied to House Numbers Digit Classification."

[83] A. Rohrbach, M. Rohrbach,W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele. "Coherent multi-sentence video description with variable level of detail." In GCPR, 2014

[84] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach. “Long-term Recurrent Convolutional Networks for Visual Recognition and Description.”

[85] Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Le, Q., Mao, M., Ranzato, M., Senior, A., Tucker, P., Yang, K., and Ng, A. Y. (2012). “Large scale distributed deep networks.” In NIPS’2012.

[86] Ian J. Goodfellow, Yaroslav Bulatov, Julian Ibarz, Sacha Arnoud, Vinay Shet. “Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks.”

[87] Yin, X.C., Yin, X., Huang, K., Hao, H.W.: “Robust text detection in natural scene images. IEEE Trans. Pattern Analysis and Machine Intelligence (to appear).”

[88] Neumann, L., Matas, J.: “On combining multiple segmentations in scene text recognition.” In: ICDAR (2013)

[89] Huang, W., Lin, Z., Yang, J., Wang, J.: “Text localization in natural images using stroke feature transform and text covariance descriptors.” In: ICCV (2013)

[90] Weilin Huang1,2, Yu Qiao1, and Xiaoou Tang. “Robust Scene Text Detection with Convolution Neural Network Induced MSER Trees.” 2,11 Shenzhen Key Lab of Comp. Vis and Pat. Rec.,

[91] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, Andrew Zisserman. “Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition.”

[92] Huang, Jui-Ting, Jinyu Li, and Yifan Gong. "An analysis of convolutional neural networks for speech recognition." *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015.

[93] Palaz, Dimitri, and Ronan Collobert. "Analysis of cnn-based speech recognition system using raw speech as input." *Proceedings of Interspeech*. No. EPFL- CONF-210029. 2015.

[94] Swietojanski, Pawel, Arnab Ghoshal, and Steve Renals. "Convolutional neural networks for distant speech recognition." *IEEE Signal Processing Letters* 21.9 (2014): 1120-1124.

[95] Kim, Suyoun, and Ian Lane. "Recurrent Models for Auditory Attention in Multi-Microphone Distance Speech Recognition." *arXiv preprint arXiv:1511.06407* (2015).

[96] Song, William, and Jim Cai. "End-to-End Deep Neural Network for Automatic Speech Recognition."

[97] Mao, Qirong, et al. "Learning salient features for speech emotion recognition using convolutional neural networks." *IEEE Transactions on Multimedia* 16.8 (2014): 2203-2213.

[98] Zheng, W. Q., J. S. Yu, and Y. X. Zou. "An experimental study of speech emotion recognition based on deep convolutional neural networks." *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015.

[99] Yanmin Qian, Mengxiao Bi, Tian Tan and Kai Yu. “Very Deep Convolutional Neural Networks for Noise Robust Speech Recognition”. IEEE 2016.

[100] Thomas, Samuel, et al. "Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions." *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014.

[101] Santos, Rafael M., et al. “Speech Recognition in Noisy Environments with Convolutional Neural Networks.” 2015 Brazilian Conference on Intelligent Systems (BRACIS). IEEE, 2015.

[102] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).

[103] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global Vectors for Word Representation." *EMNLP*. Vol. 14. 2014.

[104] Kim, Yoon. "Convolutional neural networks for sentence classification." *arXiv preprint arXiv:1408.5882* (2014).

[105] Kalchbrenner, Nal, Edward Grefenstette, and Phil Blunsom. "A convolutional neural network for modelling sentences." *arXiv preprint arXiv:1404.2188*(2014).

[106] Wang, Peng, et al. "Semantic clustering and convolutional neural network for short text categorization." Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Vol. 2. 2015.

[107] Johnson, Rie, and Tong Zhang. "Effective use of word order for text categorization with convolutional neural networks." *arXiv preprint arXiv:1412.1058* (2014).

[108] Johnson, Rie, and Tong Zhang. "Semi-supervised convolutional neural networks for text categorization via region embedding." *Advances in neural information processing systems*. 2015.

[109] Zhang, Ye, and Byron Wallace. "A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification." *arXiv preprint arXiv:1510.03820* (2015).

[110] Nguyen, Thien Huu, and Ralph Grishman. "Relation extraction: Perspective from convolutional neural networks." *Proceedings of NAACL-HLT*. 2015.

[111] Sun, Yaming, et al. "Modeling mention, context and entity with neural networks for entity disambiguation." *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. 2015.

[112] Zeng, Daojian, et al. "Relation Classification via Convolutional Deep Neural Network." *COLING*. 2014.

[113] Gao, Jianfeng, et al. "Modeling interestingness with deep neural networks." U.S. Patent Application No. 14/304,863.

[114] Shen, Yelong, et al. "A latent semantic model with convolutional-pooling structure for information retrieval." *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 2014.

[115] Weston, Jason, Sumit Chopra, and Keith Adams. "# TagSpace: Semantic embeddings from hashtags." (2014).

[116] dos Santos, Cícero Nogueira, and Bianca Zadrozny. "Learning Character-level Representations for Part-of-Speech Tagging." *ICML*. 2014.

[117] Zhang, Xiang, Junbo Zhao, and Yann LeCun. "Character-level convolutional networks for text classification." *Advances in Neural Information Processing Systems*. 2015.

[118] Zhang, Xiang, and Yann LeCun. "Text understanding from scratch." *arXiv preprint arXiv:1502.01710* (2015).

[119] Kim, Yoon, et al. "Character-aware neural language models." *arXiv preprint arXiv:1508.06615* (2015).