2022

A Comprehensive Literature Review on Convolutional Neural Networks

Ehsan Ur Rahman Mohammed
University of Windsor

Narasimha Reddy Soora Dr.
Kakatiya Institute of Technology and Science

Sharfuddin Waseem Mohammed
Kakatiya Institute of Technology and Science

A Comprehensive Literature Review on Convolutional Neural Networks

Mohammed Ehsan Ur Rahman[1], Soora Narasimha Reddy[2], and Mohammed Sharfuddin Waseem[3]

[1]School of Computer Science, University of Windsor, Canada, rahman6a@uwindsor.ca
[2]Associate Professor, Department of Computer Science and Engineering, Kakatiya Institute of Technology and Science, India, snr.cse@kitsw.ac.in
[3]Assistant Professor, Department of Computer Science and Engineering, at Kakatiya Institute of Technology and Science, India, waseem.cse@kitsw.ac.in

ABSTRACT: The fields of computer vision and image processing from their initial days have been dealing with the problems of visual recognition. Convolutional Neural Networks (CNNs) in machine learning are deep architectures built as feed-forward neural networks or perceptrons, which are inspired by the research done in fields of visual analysis by visual cortex of mammals like cats. This work gives a detailed analysis of CNNs for the computer vision tasks, natural language processing, fundamental sciences and engineering problems along with other miscellaneous tasks. The general CNN structure along with its mathematical intuition and working, a brief critical commentary on the advantages and disadvantages, which leads researchers to search for alternatives to CNN's are also mentioned. The paper also serves as an appreciation of brain-child of past researchers for the existence of such a fecund architecture for handling multidimensional data and approaches to improve their performance further.

Keywords:
Artificial Intelligence, Computer Vision, Convolutional Neural Networks, Deep Neural Networks, Image Processing, Machine Learning, Machine Vision, Pattern Recognition

1. INTRODUCTION

The review in this paper covers a lot of topics and is structured as follows: Analysis of few related hardware research works, image processing (IP) as a parent of computer vision dealing with views of computer vision (CV) tasks as plain IP procedures, components and working of CNN, data related aspects and learning strategies used for CNN, mathematical intuition covers the mathematics dealing with modeling of IP tasks as models with complexity capable of understanding the visual data and analyze it, comparison of different CNNs and alternatives to CNNs, applications of CNNs [17] and an abstract outlined way of improvements; lags and leads shall shed some light on advantages and disadvantages of using CNNs for CV tasks and related studies; future scope presents current and predictions of upcoming trends in research and product creation in this domain.

It also serves as a critical introspective study of the various fronts of visual analysis tasks, applications of CNNs and ideas driving towards better CNNs. We shall also shed some light on biased opinions people have about their performance. The paper shall also make it compulsory for its readers analyzing about the convergence of the domains of CV and cognitive neuroscience as well as develop interests towards betterment of the representation of visual data, advancing the analysis process in terms of time and other computational resources. A tutorial survey of architectures, algorithms, and applications for deep learning (DL) is given in [10]. CNNs are one among the many other algorithms existing and upcoming that can tackle problems of CV and they shall also be observed in various other contexts which are by and large the scope of AI as well as cognitive neuroscience.

2. HARDWARE ADVANCES

Work [21] describes the importance of hardware [16] (Intelligence on Things (IoT)) in the domain of computer science, which is generally not given its due importance by the algorithm writers for AI and also the general public. As AI is penetrating into the lives of common public, it's very crucial to think about hardware specifications of devices used by them in order to give the customers their fair share. This dire need results in evolution of technologies which are applicable in mobile platforms like mobiles phones, tablets, etc. Most of the tasks in CNNs and other similar structured Artificial Neural Networks (ANN) are quantization of values, matrix operations such as multiplication, etc. and because convolution layers require lot of computation with a relatively small number of parameters. In the literature, researchers have used separate components for different layer of the ANN and a controller to synchronize them together. The most important breakthrough is that they drove inspiration from Recurrent Neural Networks (RNN) and time step delay and hence got rid of a global memory thereby reducing off-chip accesses. Qualitatively speaking, the results obtained by the authors from [86] conclude that these are energy efficient. The author in [22] explores the different architectures used in CV and sheds light on those aspects that matter the most to AI related tasks as well as affect their performance. AI algorithms [1] leads us to "online adaptation", inference that can be drawn is the basic idea of our online adaptation scheme is to use pixels with very confident

predictions as training examples. The implications that online adaptation has on big data analysis tasks and transfer learning [86] applications drives the point that it is important that the adaptation retains a memory of the positive class in order to create a counterweight to the many negative examples being added. Even embedded system applications [51] have to utilize this analytic technique so that it can be available at user's disposal readily [94]. The author in the paper [69] gives the summarization of research in CNNs for Visual Recognition, research trends and improvement trends which results in expanding the horizons of convolutions in forms of step-increasing convolution which not only learns kernels but also diminishes the dimensionality thereby becoming a good alternate to pooling layers, this behavior assists in reducing over-fitting; suggestions of removal of other forms of sub-sampling layers from the CNNs like FC layers and replacing them with global mean of the last activation map which embeds into itself the confidence levels of various categories (image classification); unsupervised pre-training can help in generalizing well and randomization of weights and fusion of feature maps from different levels in the network enhances their accuracies.

An exciting work presented by author in the paper [36] that paved way towards generation and understanding of an efficient training of DL-based models inspired by biological neurons [58]. The inventors have modelled a silicon circuit that is a miniature physical manifestation of a neural network (NN) [36]. It consists of some "neurons" which are connected in such a way that they can mimic the neocortex. The author in the paper [67] gives a detailed abstract level understanding of functioning of biological neurons and also serves as a guide for constructions of similar electrically simulated versions of neural silicon circuits also can be found as a useful resource describing different actions and types of connections among a cluster of neurons.

The DL era is in existence due to many factors [73], one of them being high computational power through dedicated hardware and processors and achieved faster CNNs training using algorithms on software level namely Dynamic Precision Scaling (DPS) which leaves advanced precision search and traces out sufficient precision using dynamic fixed point and flexible multiplier-accumulator (MAC)[1] i.e. a configurable MAC reduces the propagation delays for lower precisions, thereby gaining a 5.7 times speed-up.

3. COMPONETNS AND WORKING OF CNNS

Prior to CNNs and other DL architectures for CV, there have been traditional IP pipelines such as BRIEF (Binary Robust Independent Elementary Features) for feature extraction and dimensionality reduction [88], ORB (Oriented FAST and Rotated BRIEF) for feature matching and key-point descriptor [89], [116] scale invariant feature extractors and key point image descriptor [117], etc. Current technology uses the local feature extractors which are numerical representations of an image local structures and make-up and are not same as ML feature extractors, like Scale Invariant Feature Transformation (SIFT) and SURF [74]. In CBIRs (Content Based Image Retrieval) systems, images (query and those in the search space) are compared by matching their features and searching for a geometric operations/correlation that can associate the regions of both images.

CNN's [79] are used for large scale IP [11], and while speaking in the context of visual recognition [144] are made of layers including one input, one output and many hidden layers; those can be pooled, fully connected or Convolutional layers. Mathematically, convolution is a process performed by two functions to make them a new function which gives information about their effect on each other. Convolution is very much related to cross-correlation; hence convolution has its applications in multiple fields of science and technology like CV, NLP, image processing, signal processing, and other engineering disciplines.

Understanding receptive fields are crucial while considering the study of CNNs, because they provide understanding of the quantitative modelling done by the kernels on data. The size of the kernels which is bounding the spatial extent of it is called its receptive field, which is modifiable across the various layers of the NN. They are generally of small values thereby making weight-sharing and local patterns extractable. In the hidden layers, the effective receptive field is a function of receptive fields of the previous layers, hence enabling learning in CNN's with such a small number of parameters to be trained it's highly desirable to stack up many layers with small receptive fields. The layers apply a convolution operation to the input and passing its results to the next layers, basically the way layers of a NN work, this way they mimic the activity of the neurons in our brain that simulate and analyze visual data (stimuli).

Pooling layers - as the name suggests these layers maps multiple neurons or neuron clusters of previous/prior layer to a single neuron in the next layer. Many types of pooling layers exist like max pooling which uses the maximum valued neuron from the neuron cluster for mapping, average pooling uses the average value from the neuron clusters. Another important concept of CNNs is pooling, which is a form of non-linear down-sampling. There are several non-linear functions to implement pooling among which max pooling is the most common. The goals of pooling are to reduce the dimension of features and retain effective information to avoid over-fitting. Besides, pooling can resist rotation, translation, expansion, etc. The common operations adopted at the pool layer include maximum value sampling, average value sampling, summation area sampling and random area sampling. Some recent versions of pooling are mentioned in Table 1.

Fully-connected layers - are layers consisting of "neurons" that perform well on classification tasks as well as on tasks to learn features but require a very large number of neurons due to the large dimensionality of modern-day images, so as to perform well. Convolution operation reduces the number of parameters to be trained and makes it independent of the size of the input image, this also helps in dealing with vanishing gradients (gradients become zero and hence activation functions, tensors, filters become zero resulting in no learning, greater bias, lower variance) or exploding gradients (gradients exceeding the limits/ranges which are set for them to perform

well and try to increase variance and lower the bias, resulting in bad learning) problems which are faced in cases of traditionally used multi-layer NNs using back propagation for gradient calculation.

Convolutional layers - these layers do the actual work of Convolutional networks and hence derive their name from this layer. This layer performs the convolution operation (*) on different differentiable functions which are manifested as matrices in case of IP and CV, as images are inherently matrices (multi-dimensional) on the input values obtained from the prior layers. It contains these functions as filters which can be convolved independent of each other and give a feature map. These filters are initialized randomly and will be tuned as the training proceeds further. Deeper the network i.e. higher the number of hidden Convolutional layers, better is the performance.

A good description of the training of CNNs can be found in [42]. Convolution is basically an expensive mathematical operation when performed on data in matrix form, but when considered in the Fourier domain, the operation is point wise multiplication and nothing else, this makes training on massive datasets viable in meeting with real-time demands; learning at real-time has profound benefits such as turning model into a near-perfect data distribution sampler thereby enhancing the generalizability and accuracy of the model, [128] presents a variation of transfer learning called as transfer subspace learning wherein the new training data samples are fed into a PCA to capture maximum variance, output data is then fed into the existing network (training phase for subsequent samples).

1) Formal math definition of Convolution:

Convolution is defined as

$$(f * g)(t) \stackrel{\nabla}{=} \int_{-\infty}^{\infty} f(\tau) g(t-\tau) d\tau \quad (1)$$

Where f and g are two differentiable functions which can evaluate to constant values, f*g is the convolved function

2) Hyper-parameters:

The exponential number of hyper-parameters settings possible in any given CNN makes lead to the search of optimal set of hyper-parameters, building of complex and efficient hierarchical architectures like CNNs and/or exploring of the hyper-parameter space for some desirable feature/information becomes a hard task. This problem in one among the many and can be solved in polynomial time if NP-complete conjuncture is proven true. CNNs being stepwise processing DL algorithms, sequential model-based optimization (SMBO) problems are better to solve when compared to the optimization of objective functions like those in RNNs. Model searches like Grid and Random search are computationally very inefficient (expensive) and the alternatives have been presented in [45]. Knowing about choice of hyper-parameters set can be dealt with the concepts of cross validation.

Sequential Model-Based Optimization (SMBO), a probabilistic model which is promising in noisy environments (avoid over fitting). Bayesian training techniques are chosen so as to cope up with uncertainty, have flexibility in the model being generated and exploring of other analytical dimensions (of the model/objective function). Reinforcement learning [26], to deal with the problem of figuring out what model structure can give acceptable results for a given problem; using a non-synchronous approach for finding optimal network configuration by tuning the parameters autonomously and the paper [26] also proves the convergence of the procedure to an optimal solution for MNIST data set. In RL, the model's decisions (predictions) are reinforced by rewards which it receives from its environment [100] [102] [103]. RL becomes expensive in domains involving exploding among of dimensions in the input spaces, because this is directly proportional to the number of details that must be considered for taking a decision and receiving a reward; a max-pooling CNN as unsupervised mechanism but has been trained on class-labelled samples thereby maximizing the variance in the output representations of this set collected at random from the environment; transforms the HD visual input into a concise feature vector, [102] hence is paving way for DL architecture for neuro-evolution.

The paper [25], reminds us of the fact that machines and algorithms can perform learning only when data is available and which is sufficient to draw inference and make decisions on unseen data which is not a varying at enormous rates from the ones the model is trained on and [25] is among those efforts which are devoted to enlarging of existing datasets primarily due to unfair comparison issues associated with lack of data to model the universe, web images are labelled by a DCNN using the context surrounding the images in the WebPages, focusing on ability of web to serve our models with supervisory information with some leniencies. Though the goal is to minimize expected errors on previously unseen images, only empirical errors can be directly optimized on a set of labelled images with respect to a function space defined by a model. According to statistical learning theory, the gap between expected error and empirical error is determined by the sample size and model capacity. The gap becomes smaller with increasing sample size, and model design tries to minimize the expected error by defining a function space to minimize the empirical error and control the model capacity. Automatic data augmentation from massive web images, image search engines and graph building are the task directly linked with the availability of ever-increasing loads of visual data on the internet. DCNN [17] extracts image's visual information while the web provides an image's contextual information, which is complementary and can jointly provide additional information to an existing dataset. We can augment an existing dataset in a scalable, accurate, and informative way.

Edge detection [63] is the most prominent image processing operation performed to analyse and process images and also among the very first steps to perform visual recognition, this uses a filter of some dimension, a filter can also be called as a kernel because its properties are much similar to a kernel function used in calculation of transformations. This filter is convolving the input images through the use of convolution operation/function of your choice which must belong to the set of functions which are differentiable.

4. DATA-OIL RUNNING AI/ML/CV

Data in context of CNNs, its generation, storage format, computer internal representation, collection, pre-processing

and possible conclusive inferences that can be drawn from the visual data is discussed here. Firstly, GANs[142] and new training techniques for their stability[110], introduced by Ian et al., consists of two sub networks 1) generator that can take in random noise as raw data and produce some pattern-following data (images, video sequences called as synthetic images/data) which is then passed to the 2) discriminator which checks the quality of data produced and rejects those sample points (fake data), which do not match with the data samples already fed into it. (1), (2) work continuously and become better at their respective jobs. DCGANs [81] are a GAN model generates realistic images from the random noise samples. The loss functions used by generator can be minimax, heuristic non-saturation, maximum likelihood and discriminator obviously uses cross-entropy for binary classification of real and fake data.

There is so much data that can be fed into AI systems to ensure that small intricacies are also used to obtain the solutions (predictions/classification/actions) and for which data collection is important as well as mapping of relevance of a given sample of data to certain part of the system. While dealing with CV tasks it's easy to pre-process the data because of the variety of IP techniques available such as image enhancement/restoration, pixel values quantization etc. Pre-processing if taken from the perspective of a data scientist consists of data pre-processing methods like dealing with missing data, normalizing and scaling the data sample values.
Possible conclusive inferences that could be drawn from visual data are numerous and this can be concluded from the fact that most of the brain is used to process visual stimulus as well as that most of the data on internet contains images (texts, maps, video clips, satellite images, medical diagnostic images, etc.). Few of them to be mentioned which can lead to probable advances and widening of the scope of CV are:1) Multiplexed image representations, 2) dynamic task construction, etc. Expensive labelled data is utmost important for the training of CNNs for image classification task and it's not known to have any advantage when it comes to other CV tasks. The paper [5] deals with unsupervised learning [81] using CNNs by using a discriminative model trained with the cheaply acquired loads of image data. The feature representation learned by Exemplar-CNN is, by construction, discriminative and invariant to typical transformations. Surrogate data processing tasks are created using combinations of a set of IP algorithms as discussed already in the paper, which, when applied to raw data results in appearance if surrogate classes, which serve as training assistance to the discriminative convolution neural networks using loss functions followed by SoftMax and optimizing of the negative of the multinomial log likelihood of the output. Results were influenced by factors such as number of surrogate classes, samples per class, set of transformations and invariance properties of learned representations.

5. MATHEMATICS BEHIND CNNs

The intuition here is that these layers can actually compute results with or without the biological neuron's analogous "artificial neurons", the basic units of NNs. They consist of learnable filters with parameters. Every filter is spatially determined by its width and height but extends through the full depth and dimensions of the input volume i.e. the image. The paper [13] serves as an excellent guide to Convolutional arithmetic in context of Deep Learning.

The data fed to CNNs have some intrinsic structure made up of properties that are exploited for abstract representation learning, training, prediction and other NNs' tasks; properties being multiple dimensions, ordering or rather weighing the features axes, another axis which makes it handy for us to exploit different viewpoint of the data. Discrete convolutions preserve the ordering of the data; filters are sparse, and weight-sharing is high. Not only algorithms are important but also making the algorithms to deal with properly explainable, clean representations of data (input data), here dealing stands for training upon data or learning from data. Better generic representation learning [100] DL-based models [98] include probabilistic models, auto-encoders, manifold learning-based models, etc. Much of the workflow pipeline in visual recognition tasks such as pre-processing can be handled using the IP techniques as well they help in creating transformed versions of images which unravels better or requirement set of knowledge from the visual data. We shall come back to this argument/point for further elaboration in a different context later in the paper. The question to be addressed is "what makes one representation better than another".

Sparse coding, which can be taken as a close cousin of CNNs training strategy as a hierarchical feature learning [91], [66] for image data has increasingly become popular. Convolutional filter banks are proposed in [2] as a feature learning algorithm for CNNs having stages of trainable CFBs, mixed with non-linear functions similar to the activation functions, and spatial pooling. The problem with sparse encoding is that it performs inference on all patches having dimensions equal to that of the filters, independently thereby leaving convolving nature of filters and producing a high redundant representation of the image; which can be completely avoided by applying sparse coding simultaneously on all the possible patches of the image. Encoder module uses a fast feed forward non-linear objective function for faster computation of sparse code.

[2], [5] Inspired by sparse coding which is used to learn visual features hierarchically which tend to produce redundant codes. The paper [2] proposed unsupervised method for learning multi-stage hierarchies of sparse Convolutional features learnt over the entire image at once and producing a variety of codes from which highly diverse filters such as center-surround filters, corner detectors, orientation detectors, etc. can be constructed. The paper [5] outperforming many state-of-art unsupervised learning methods on the task of visual object recognition on benchmark datasets like DVS benchmark for object recognition, cites the criticality of tons of data generated from multiple data sources (statistically speaking, as a producer of different but related physical events, data representing same object but with variations of features its composed of) grabbing the meticulous details and feature variations which are crucial for endurance of unsupervised models.

Feature maps are the output of convolution operation which gives a certain level of understanding of the input in terms of its features and the deeper they are found in the network the higher is the level of retrieval of complex patterns. Kernels are referred as matrices in 2D and cuboids in 3D. The paper [13] uncovers

the detailed resource to understand the arithmetic involved in CNNs. Strides and padding are important settings that are taken care while performing convolution operations.

From [42], we know convolutions using discrete Fourier transformations are fast, but [6] learns the powerfulness of representation in the spectral domain; based on which quantizing the values called as spectral pooling which doesn't lose as much information as in other sub-sampling which deal with the matrix representation, enabling a stochastic regularization by random adjustment of the resolution. Few other generalization techniques used for CNNs are mentioned in Table 3.

6. LEADS

There are various reasons behind working with CNNs for visual recognition, and they include

1) CNNs are good at performing feature learning i.e. they can capture and learn only the relevant features from the input data and try to discard the useless information.

2) Weight sharing is the next important advantage, which is why we need not bother about complexity at each level/stage or layer of the neural network.

3) CNNs can perform feature extraction and works at par with data with pre-trained features included in them.

4) The impressive effectiveness of CNNs in capturing or forming of robust generic representation of visual data for a given task excites it to be transferred to other problems of visual recognition. An analysis of the factors [41] that account for transferability for a generic ConvNet representation. Transfer learning saves a great deal of learning and leverages the training done to pre-trained networks available in the market.

5) The parameters of the ConvNet like the weights, distribution of the training data, structure, dimensionality reduction, etc., which are adjusted to account for improvement in performance on various visual recognition tasks by minimizing the discrepancy between source and destination tasks and their correlation based on the observations from the proposed[41] factors. Enlisted are the practices to implement a transfer learning in an effective way.

6) [143] sensor networks and their applications.

7. UNDERSTANDING AND VISUALIZING CNNS

Understanding CNNs by visualizing [85]: The visualization of the intermediate hidden layers gives the understanding of functionality of feature layers and hence can be used for diagnostic purposes. While comparing performance of various DL models by allowing us to observe the evolution of features during training, intuitively getting the stimuli which can activate neurons of a given layer require an operation which is opposite to that of the convolution, hence [85] proposes a DeConvNet.

Paraview is a data visualization software, [75] utilizes the same framework along with Matplotlib to create Tensor View for visualizing the learning features present in CNNs at different layers. Visualization tasks help achieve the mission of interpretability and explain-ability which are becoming crucial in the days to come due to influx of technologies and applications such as autonomous driving which are built on top of CNNs or in general any other learning algorithms.

Basically, all ANNs act like a black-box containing mathematically strong models capable of mapping the inputs to outputs. Learning mechanism is hence possible through methods relating to visualizing its hidden layers. Visualizing those layers whose space contains the pixel projections is feasible; while coming to the hidden layers finding of the stimulus which activates the correct set of neurons descriptive of the learning that took place on the model is crucial and involvement of Gradient Descent algorithm with better initialization of weights is critical. [85] Deals with complexities of visualizing and effectively solving them. Parallel to this we need to understand that this lack of transparency can be a big setback to the growth and development of AI algorithms. Some gradient descent algorithms from recent literature can be looked at in Table 2.

8. PROBLEMS WITH CNNS AND TACKLING THEM IN CONTEXT OF COMPUTER VISION

1. Overfitting: Happens when training dataset size is insufficient for good learning (not memorization) and hence leads to memorization of patterns in data, this makes the model rigid and less intuitive to gain knowledge of delicate patterns in it. Large datasets available now-a-days eased this problem
2. Underfitting: When models got a lot of data to train upon and hence made inference from it which was so general that it doesn't reach to the expected level of performance even though the train and test error are almost the same. These two results from setting of poor trade-off between bias and variance.
3. Exploding gradients: When gradients grow so much that those connections and/or units gain the ability to manipulate entire model's performance
4. Vanishing gradients: When gradients shrink so much that those connections and/or those units become irresponsive towards the learning procedure
5. It's important to realize that mathematics despite centuries of research is not in position to give us exact/accurate and precise solutions to equations involving multiple variables (independent in case of pattern related tasks such as classification, recognition, predictions, etc.) but rather there is a small set of rather pseudo-intelligent techniques coming under the branch of numerical methods that give us the ability to at least get close to the point of finding approximate solutions to such problems. This is evident at a variety of places including image/signal processing where we use Fourier and other transforms/filters to recover data from the signals, get to accurate predictions about missing values of specific data having the knowledge of prior.
6. Lower comprehension (sub sampling): When CNNs become adaptive towards regular & simple tasks or were victims of high bias then it becomes tougher to analyse and

find out better ways to do feature extraction and use that knowledge for conceptually

7. General discussion is between considering a problem as deterministic or stochastic and whether data plays a crucial role or algorithm.
8. Sub sampling which is performed due to the confined receptive field caused by selection of filter which are not taking all of the input volume but only a part of it (and also having no overlapping with other filters) may result in loss of precise global and local spatial relationships which are very much required for identity recognition.
9. There is another kind of sub-sampling too which refers to the use of sub-sampling layers like pooling, FC, etc., which are indeed not learning any parameters and hence not contributing vigorously to the model being formed by the CNNs. These terms are varying in meaning according to their traditional and modern usage.
10. They have harder time in extrapolating their comprehension of the features of images to radically different and new viewpoints.
11. The CNN pipeline is not guaranteed to produce results which are robust to transformations like rotation or scale invariance and so on which hold a lot of worth in many of the CV tasks such as retrieval, etc.

All these reasons make CNNs doomed to perform well on the ever increasing and newer problems of computer vision.

But the authors of the paper subscribe to the viewpoint that even RNNs are a type of CNNs having connections among the nodes of different layers which make the network dynamic in training [101], dynamism is crucial because anatomically speaking the connections among neurons of neocortex are not only feed-forward, top-down, and so on but also consists of synapses which are recurrent.

1. A new approach is suggested by Hinton et al., recently which is called as "capsule networks". This model when challenges the viewpoint that simpler models are better, all models are incorrect, but some are useful. Other neural network models which can serve as alternate to CNNs for some tasks are mentioned in Table 4.

2. Specialized neural networks to recognize objects in new situations with less data [80] than what the regular neural nets use for recognizing.

3. [80] is a mathematical work entailing the detailed description of dataset size's influence on the performance of CNN.

4. Using DNN with different hyper-parameters sometimes tends to outperform the traditional CNNs in terms of visual recognition's performance evaluation [38]. Hyper-parameters may include: using Nesterov Momentum which is a slightly different version of the momentum update that has recently gained much popularity; for optimization using a group of methods in context of deep learning which are based on Newton's method.

5. A Convolutional kernel network [8] tends to approximate the features learned through the various kernels. They are simpler to train without eating out on its accuracy. Invariance [116] which is utmost important in tasks like image classification, is being dealt with a natural tool i.e. kernels.

9. APPLICATIONS OF CNNs IN COMPUTER VISION

CNNs are applied extensively in CV tasks [12], [29] because of its network topology which integrates within itself two major ideas, one being weight sharing which reduces the parameters to be learned thereby memory and computation-wise efficient and another is it being locally constrained which supports it to improve upon errors due to local distortions. The papers discussed under this section have been sort out to be presented here based on their performance, as well as potential to widen the scope of applications of CNNs. Hence it's suggested to look at the results/evaluation/performance sections of respective papers for convincing yourself as well as datasets for corresponding CV tasks. A list of CV applications is deliberated below in brief.

1) Super-resolution (SR) [14] [52] is a method to scale a lower-resolution image to a higher resolution one. CNNs are good at doing this and the reason to quote it first in the applications is because SR can be found in various image analysis tasks in which CNNs are also involved. This is achieved through aligning multiple column-blocks to get multi-scale features, foundation of SR, up sampling with transposed convolution. Cascading of these blocks is done passing the raw pixel data through at least one layer of convolution. Image reconstruction selects an optimized networks-based loss function which can convert the bicubic up sampled image to a high-quality HR image output.
2) The paper [20] gives an abstract related to understanding of sentiment analysis under the umbrella of NLP and working of a DCNN as a model for the same. Ensemble theory points out that, combining models can improve performance by complimenting each other's mechanisms, herein CNNs are doing and RLSTM are used for memory-based training which progresses with each time-step rather than layers which is how we can make a language model.
3) Next big task is image classification[15][71][72] and indeed the use of CNNs have brought back the old shine in vigour in the research areas of DL and the pivotal work done by Fei-Fei Li can be summarised as creation of DCNN [17] to ace the ILSVRC[7]. To make CNNs more robust to the transformations in image data such as translation, scaling and rotation, [78] randomly transforms the activation maps of CNNs during training; which implies that each kernel produces a better transform-invariant answer because of the multitude of transform levels of its input feature maps. [15] Presents a simple image classification framework.
4) Face recognition: Viola-Jones is considered to be state-of-art in doing this task it uses edge descriptions of an image to detect faces [61]. The paper [83] makes clear that object detection [129] is a multiple instance learning problem, because an image possibly comprises of multiple patches/image segments of different sizes which have to be extracted prior to application of object detection algorithms [62], which makes the training set have high sparsity in terms of detection of useful information. The paper [83] suggests that cascading of detectors, boosting algorithms, and Viola-Jones enhances performance. Cascading helps ML algorithms to select better features in the early rounds of training, this can be easily mimicked by CNNs.

5) Scene labelling: firstly, recognition of single and isolated objects though occluded is a big leap, a more complex task is to analyse these objects for a higher level of understating that may comprise of knowing about relations between the various objects, labelling them accurately, describing minute details of individual objects, scene summarization in videos, etc. Deep context based architectures deal with 3D objects by embedding their contexts into the topology of CNNs (geometric DL) thereby capable of performing scene understanding. The paper [4] presents a related task of producing answers to questions that can be generated from image data's statistics. This task falls into the region of intersection of NLP and CV and is witness to the ability of CNNs to perform well in both domains through a multi-world approach of semantic parsing of the question belongs to and segmentation of images to produce the answer belonging to CV. The paper [4] also discussed about the multimodal convolution layer, sentence CNN, and image CNN.
6) Scene understanding or Natural Image Understanding [77][4][62]: Scene understanding can be manifested creatively as modelling of questions based on the image and answers extracted to them from the same image, this is perfectly illustrated in [4] where according to the general understanding of the task three separate CNNs, the first for image decoding, second for question formation based on the decodes and the third for joint representational learning though multimodal convolution layer for assimilation of possible answer words for the extracted question. A better strategy could have been to use of multimodal convolution layer in all of the CNNs and training which can be similar to that of the RNNs but not exactly the same. Deep scenes can be understood as scenes composed of objects having object detectors which are automatically discovered due to available representations of the learned scene categories. The challenge in this task is the obscurity of the underlying representation of the objects generally being part-based representation which doesn't always extract meaningful parts of the objects but there are certain problems of complexity, simplification of the input images, identification of the semantics of the internal units, emergence of the objects as the internal representation. The automation of the tasks of total scene understanding like classification, annotation, object and semantic segmentation, visual question and answering, etc. is currently possible through hierarchical generative model.
7) Facial morphing [19]: Goal is to avoid break-ins to modern-day biometric verification systems happening through the use of morphed images, by detecting morphed images through extraction of image features learnt which can have differential effects when images are morphed.
8) Visual recognition: Visual recognition [5] is one of the great practical applications of CNNs. It is defined as finding meanings in visual content (images, videos, scenes, etc.) and analysing images for scenes, objects[43], food[44], faces [33][61][137], and other contents[39][40][60] to understand what is happening in a scene. [112] Explores the feasibility of production of sufficiently generic features from the training of deep CNNs on many object recognition tasks in a fully supervised manner through a series of quantitative and qualitative tests based on visualizing semantic clusters and contrasting with the current state-of-art models and evaluated on tasks like subcategory recognition and domain adaptation.
9) Pattern mining [60]: Unlike traditional pattern mining schemes, [60] discovers patch patterns from image-caption pairs generated from news events; which are human recognizable, representative and discriminatory (Representative refers to the tendency of this patch to appear in many of the images for a target class, and discriminative refers to the fact that it should be unique and not appear frequently in images outside of the target category) as well. The multiple modalities applied here are semantics from captions of the images and the images themselves. Obtained patterns are named according to the associated image captions. The paper [60] not only proposes an automated image patch pattern mining technique but also evaluates the same based on the effectives of the method to produce patterns aligned with the strong semantics specifying the high-level events.
10) Image forgery detection [3] can happen with the availability of high-level representations of images because CNNs work well and intuitively speaking they have conglomerated features of image which provides adaptation to multiple tampering manipulations. We will get back to sparse type of coding of data here, because in order to remove redundancy in data and retention of characteristics of data, ReLU activation is applied.
11) Action recognition: Model composed of alternative layers of volumetric Convolutional layers to extract features and recurrent layers to collect contexts (semantic information abstractions formed to contribute as clues for actions, from the input, wherein RF captures small part of image and throughout the deep expanse of the network collecting the optical flow energy) as they are doing temporal modelling with preservation of hierarchical contexts which have evolved from the beginning of the training of the video [32]. Action recognition [34] had helped the deep learning community to conclude that improvement of performance in CNNs as well as in other common networks is greatly enhanced when they are deep (consists of multiple layers and/or nodes per layer). This task deals with context which serves as abstraction for semantic clues within images helpful in extracting higher order features and/or activity (in context of action recognition). The problems posed while doing AR with CNNs comes from the confined RF as well as its local connectivity which leaves out on the excessively important global spatial relationships among the pixels of the image. Depth avoids losing of context topologies and trajectory-based hierarchically formed information. Human action recognition is detailed in [113]. Action recognition is closely associated with other CV task such as action prediction which is of paramount importance in scenarios wherein prediction/detection of a certain action can be very helpful like accidents on a road predicted by a self-driving car, potential violent attack by some individual which is visual surveillance, video retrieval are becoming harder because the associated text and other data with the video is not descriptive of it example of their applications are video-games, and human-robot interaction (the working of robots

gives us challenges in representing the image data which is perceivable by them and can model our world appropriately into their understanding – [111]).

12) Human pose estimation: This activity can be thought of as a follow up of the one the application mentioned before where agent is a human being. Human pose estimation is mentioned separately because it's crucial to analyse the actions of humans, for example if we look at tasks such as violence detectors, security threats detection in real-time through the video inputs available because of surveillance systems. This application has challenges arising out of high dimensionality in the input as well as the multiple possible human body poses which doesn't even imply the same emotional state of his and/or intention behind having the particular posture. It's more of a learning challenge than feature extraction one. Human pose estimation is briefly explained with an innovative idea in [108] where the thing that is done is prediction of 3D positions of body joints from image without considering temporal information, wherein an intermediate body parts representation that transforms a difficult problems into a simpler per-pixel classification task, and this problem puts light on the importance of availability of dataset encompassing a wide variety of data to make the model invariant to occlusions, body shape and size, clothing, etc.
13) Video classification [53]: Harder task than image classification as CNNs have to expand their connectivity in the temporal domain as well for extracting information across sequence of frames forming a video clip, CNNs have proved to be efficient local spatio-temporal information extractors and the authors have proposed a two parallel streams of processing one learning the features of low-resolution frames and another on the high-resolution frames; fusion of their outputs can take place earlier, later or a mix of both. On similar lines we have infusion with object detection is called video object segmentation [97]. Recent works utilize recurrent neural networks (RNNs) with long-short term memory (LSTM) units at the end CNN-based features of each frame to exploit semantic combinations [5], [25], [35]. Aggregating the information at every hierarchical stage of a NN obtained due to the flow of information which is raw (pixels) and generated (feature maps) due to understanding and extraction of continually higher order features till the end of the CNN makes use of evolutional learning (inferential-based analysis). AR is to recognize the actions of one or more agents by observing a series of frames (video) giving information on the agents' actions changes in its environmental conditions. Dann in the paper [32] encourage the readers to think on the similar lines but to broaden the applications of this CV task and apply to it other domains such as medical diagnosis wherein different cells/tissues/organs shall serve as agents. Scientific research can get a great boost if observation of results (generally done by human lab workers) can become a task of such AI-driven systems using this particular task of data classification and categorization and this way more information and precise functioning of biological agents (cells), chemical agents (substrates in reactions) and physical processes can be assessed in greater details and opens up more domains of research as well as allows scientists to make conclusive remarks regarding the feasibility of physical phenomena going on elsewhere in the universe, chemical reactions and effects of drugs on the animals and giving priority to certain stages of the aforementioned processes. Video object segmentation using online adaptation is discussed in [97].
14) Contour detection: The paper [1] shows that contour detection can be used to draw statistical inference. We can also have multiple views of contour detection as discussed in [1] that include non-parametric, adaptive learning and supervised mid-level information, but we here discuss its possibility through CNNs. A simple combination of CNNs and nearest neighbour search, as they are good at feature extraction with different layers of the model can learn different levels of feature abstractions. Numerous image patches extracted on which contour detection is to be performed is classified based on the basis of selection for further processing by using a random forest classifier. The performance on domain-specific tasks can be improved through pre-training of CNNs on contour datasets.
15) Document analysis: Generally, it falls into many categories like OCR (Optical Character Recognition), OMR (Optical Mark Recognition). Document Analysis is a discipline that combines image analysis and pattern recognition techniques to process and extract information from documents from different sources. Sources include either raster formats, which are obtained after scanning paper-based documents, or electronic formats such as ps, html, pdf, etc. Text spotting and text-based image retrieval is done in [118] using a combination of proposal generation methods which tend to reduce the humongous task to repeated application of CNN on patches of varying sizes and possibly having texts present in them at different orientations. On similar lines we have many state-of-art algorithms to perform object detection such as Single Shot Detector (SSD), You Only Look Once (YOLO), Faster-RCNN, which accomplish this task of text detection which has many applications like translator for signboards and restaurant menus in a foreign language can assist a traveller & Long videos, news clips, etc. can be summarized based on the text present in them. A combination of powerful represent-ability of MLP (Multi-Layer Perceptron) and unsupervised feature learning constructing a robust common framework to train highly-accurate text detection and character recognizer modules. End2end text recognition with CNNs is discussed in [115]. The level of character recognition is enhanced in [55]with the use of rotation-based patterns generated by rotating the character images of the dataset, two possibilities are explored in this context , the first one being multiple CNNs given different rotation-based samples of an original dataset and using ensemble of them using techniques of winner takes all with respect to node-wise responses if all the different CNNs give different predictions or choosing of the class response which is highest and the second one being passing all the diverse distorted samples to a single network. Empirical result can be stated as CNNs can recognize a large variety of patterns by incorporating patterns having some deviations from the norm, but the level of distortion that can't hinder the recognition ability to a great level is still not defined by the study [55], it has also deployed a

unique way of back-propagation which is a mixture of upsampling, inverse convolution, etc.

16) Neural style transfer [46]: Styling the semantic content of an image can be done provided we have better semantic representations of the image. This can easily be made using CNNs used in object recognition [64] tasks (higher level abstraction of image information). This task reassures that "learning representations" are crucial for "learning procedures" i.e. ML algorithms or NN models because in such visual recognition tasks image representations [111] which give the ability to independently model the semantic variations of an image (scene/visual) is a prerequisite. The work in [46] creates a new image by combining the semantic content of an image with the overall texture pattern of few well-known artworks. Texture shall be extracted from the artwork image and applied to the target image without changing its edge/object orientation information, etc. Comparison between the target and image produced is done to back-propagate error using gradient calculation techniques to make better representation possible. Closely associated with this task are style classification and object recognition.
17) Bone age assessment[92][27]: Medical task of analysing the maturity of the skeleton can be possible by CNNs. Detection of targeted regions in the skeleton are feasible by faster-region proposal based CNNs. This task helps in guiding the growth of a new-born and equally important for diagnosis of various bone related disorders, etc. According to the medical community rather than using radiological parameters, bone age assessment using hand's radiograph is the most accurate. This task reflects the importance of speed and accuracy in learning procedures, as many works before [27] were accurate but not applicable in busy clinics. Along with predicting good estimates of bone age, it also gives a prediction about the acceleration stages of puberty. Domain knowledge is of paramount importance when dealing with similar task which helps in expanding the range of applications and formulating ways to make DL-models to represent and process various activities of humans thereby taking a step towards AI.
18) Video processing [53]: Multiple approaches for extending the connectivity of a CNN in time domain to take advantage of local spatio-temporal information and suggest a multi-resolution, foveated architecture as a promising way of speeding up the training. By modifying the architecture to contain two separate streams of processing: a context stream that learns features on low-resolution frames and a high-resolution fovea stream that only operates on the middle portion of the frame, we treat every video as a bag of short, fixed-sized clips. Videos are inherently difficult to train upon, because they are difficult to collect, label, store and process. Temporal connectivity pattern are the three broad connectivity pattern categories (Early Fusion, Late Fusion and Slow Fusion) which work best for video-based action recognition due to the ability of providing additional motion information.
19) Semantic image segmentation: The papers [1] [54] [134] [138] [141] explains about the dense pixel-wise class probability map is obtained by successive operations of un-pooling, de-convolution, and rectification. Here, the algorithm poses semantic segmentation as instance wise segmentation problem i.e. the network takes a sub-image potentially containing object called as an instance as an input and produces pixel-wise class prediction as an output. Semantic segmentation on a whole image is obtained by applying the network to each candidate proposals extracted from the image and aggregating outputs of all proposals to the original image space. The proposed network is trained to perform semantic segmentation for individual instances. Given an input image, we first generate a sufficient number of candidate proposals, and apply the trained network to obtain semantic segmentation maps of individual proposals. Then we aggregate the outputs of all proposals to produce semantic segmentation on a whole image. Optionally, we take ensemble of our method with Fully Convolutional Networks [17] to further improve the performance. Single-Shot Detector, You Only Look Once, Faster-RCNNs are the current state-of-art architectures for this task. Deep de-convolution network such as auto-encoders apart from serving as image data generators, image modifiers, super resolution algorithms, speech generators (Generative Adversarial Networks (GANs), Deep Convolutional GANS, Super Resolution GANs) do semantic segmentation of images i.e. identification of pixel-wise class labels and segmentation of image using masks.
20) Image compression [135]: Image compression performed on JPEG images can be detected as well as forgery localization can be done using quantization of histograms of discrete cosine transform (DCTs) coefficients which are reflecting of differences between single-compressed and double-compressed areas.
21) Image restoration [56]: Annotating documents is costly affair and [99] envisions and reaps from the inclusion of a context aware language models in the task of extraction of named entities from historical manuscripts wherein CNNs used for Visual Recognition and Long Short Term Memory models for modelling of interrelated words; as an instance of image retrieval.

There are various other applications of CNNs in domains such as NLP (Table 5), scientific applications(Table 7) and those other than CV, NLP and scientific (Table 6).

10. FUTURE SCOPE AND IMPROVEMENTS

Some of the few possible future avenues to carry research upon are 1) the feasibility and analysis of using centered formula for derivative calculation in gradient descent; 2) use of relative error for the comparison between analytic and numerical gradients; 3) analysis of kinks in the objective (kinks stand for the non-differentiable parts of the function) function; 4) feasibility and analysis of possible improvements by the use only few data points; etc. Self-supervised learning in a way can be implemented as [70] wherein the network structure can be modulated without any learning or change in the values of the weights thereby improving its performance. The authors have used the strategy of an intuitive combination of the error gradient of each pixel by using a modified ground truth (GT) referred to as pseudo GT, binarizing the prediction output for each class, and scaling factor for the computed error gradient which are then passed backwards during back-propagation.

Critical ways to approximate ideal gradient directions are inferred, useful interactions which are outcome of overlapping receptive fields also add to the learning procedure.

The exploitation of both supervised and unsupervised DL networks jointly is yet to be done, and there is much work in the literature in that direction which shows that layer(s) from one of the two separately trained supervised and unsupervised feature learning through DL networks[48] should be replaced by other one if the output of individual layers and hence features learnt by individual layers at the corresponding level (i.e. both DL networks must have same number of layers) and it has shown improved accuracy and reduced training expenditure on benchmark datasets like ImageNet for image classification.

A reliable approach for improvement of the performance of CNNs by a few percent is to train multiple independent models and add their test predictions. With increase in the number of models in the ensemble (using of combinations of learning algorithms for better performance), the performance also monotonically improves with very negligible enhancement to the current method [26]. Speeding up of algorithm is really important as CNNs are slow at processing the data. Using of optimization techniques such as dynamic precision scaling [24] [73] and dropout [22] help in resolving over-fitting along with speeding up evaluation of outputs.

CV tasks are very much dependent on much of the IP works accomplished till date, in domains of pre-processing, feature learning, data transformations, etc. But this availability of various methods in IP hinder the goal and progress towards full realization of AI which is to make machines “understand the world around us” and there has to be some constraint/threshold which may help in deciding what kind of assistance shall be provided to AI/Ml algorithms/systems alike for better performance.

REFERENCES

1. Liu, Na, Ye Yuan, Lihong Wan, Hong Huo, and Tao Fang. "A Comparative Study for Contour Detection Using Deep Convolutional Neural Networks," in *Proc. of the 2018 10th International Conf. on Machine Learning and Computing*, 2018, pp. 203-208.
2. Kavukcuoglu, Koray, Pierre Sermanet, Y-Lan Boureau, Karol Gregor, Michaël Mathieu, and Yann L. Cun. "Learning convolutional feature hierarchies for visual recognition," in *Proc. Advances in Neural Information Processing Systems*,Vancouver, British Columbia, Canada, 2020, pp. 1090-1098.
3. Huang, Na, Jingsha He, and Nafei Zhu. "A novel method for detecting image forgery based on convolutional neural network," in *Proc. 17th IEEE International Conf. On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conf. On Big Data Science And Engineering (TrustCom/BigDataSE)*, 2018, pp. 1702-1705.
4. Ma, Lin, Zhengdong Lu, and Hang Li. "Learning to answer questions from image using convolutional neural network," in *Proc. 13th AAAI Conf. on Artificial Intelligence*, 2016.
5. Dosovitskiy, Alexey, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. "Discriminative unsupervised feature learning with convolutional neural networks," in *Proc. Advances in neural information processing systems*, 2014, pp. 766-774.
6. Rippel, Oren, Jasper Snoek, and Ryan P. Adams. "Spectral representations for convolutional neural networks," in *Proc. Advances in neural information processing systems*, 2015, pp. 2449-2457.
7. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Advances in neural information processing systems*, 2012, pp. 1097-1105.
8. Mairal, Julien, Piotr Koniusz, Zaid Harchaoui, and Cordelia Schmid, "Convolutional kernel networks," in *Proc. Advances in neural information processing systems*, 2014, pp. 2627-2635.
9. Bengio Yoshua, "Learning deep architectures for AI," *Foundations and trends in Machine Learning*, Vol. 2, pp. 1-127, 2009.
10. Deng, Li., "A tutorial survey of architectures, algorithms, and applications for deep learning," *APSIPA Trans. on Signal and Information Processing*, vol. 3, pp. 1-29, 2014.
11. Wu, Sai, Mengdan Zhang, Gang Chen, and Ke Chen, "A new approach to compute cnns for extremely large images," in *Proc. ACM on Conf. on Information and Knowledge Management*, Singapore, 2017, pp. 39-48.
12. Aarthi, R., and S. Harini, "A survey of deep convolutional neural network applications in image processing," *Int. J. Pure and Appl. Math*., vol. 118,pp. 185-190, 2018.
13. Vincent Dumoulin and Francesco Visin, “A guide to convolution arithmetic for deep learning,” *CoRR*, vol. abs/1603.07285, 2016.
14. Shuai, Yuan, Yongfang Wang, Ye Peng, and Yumeng Xia, "Accurate Image Super-Resolution Using Cascaded Multi-Column Convolutional Neural Networks," in *Proc. IEEE International Conf. on Multimedia and Expo (ICME)*, 2018, pp. 1-6.
15. Guo, Tianmei, Jiwen Dong, Henjian Li, and Yunxing Gao, "Simple convolutional neural network on image classification," in *Proc. IEEE 2nd International Conf. on Big Data Analysis (ICBDA)*, 2017, pp. 721-724.
16. Kim, Sunwoong, and Rob Rutenbar, "Accelerator Design with Effective Resource Utilization for Binary Convolutional Neural Networks on an FPGA," in *Proc. IEEE 26th Annu. International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, 2018, pp. 218-218.
17. Aloysius, Neena, and M. Geetha, "A review on deep convolutional neural networks," in *Proc. International Conf. on Communication and Signal Processing (ICCSP)*, 2017, pp. 0588-0592.
18. Zhu, Boqing, Changjian Wang, Feng Liu, Jin Lei, Zhen Huang, Yuxing Peng, and Fei Li, "Learning environmental sounds with multi-scale convolutional neural network," in *Proc. 2018 International Joint Conf. on Neural Networks (IJCNN)*, 2018, pp. 1-8.
19. Mohammed Ehsan Ur Rahman and Mohammed Sharfuddin Waseem, “A Novel Framework for Detection of Morphed Images using Deep Learning Techniques,” unpublished.
20. S. Chen, C. Peng, L. Cai, and L. Guo, "A deep neural network model for target-based sentiment analysis," in *Proc. International Joint Conf. on Neural Networks (IJCNN)*, Rio de Janeiro, 2018, pp. 1-7.
21. Shin, Dongjoo, Jinmook Lee, Jinsu Lee, Juhyoung Lee, and Hoi-Jun Yoo, "An energy-efficient deep learning processor with heterogeneous multi-core architecture for convolutional neural networks and recurrent neural networks," in *Proc. IEEE Symposium in Low-Power and High-Speed Chips (COOL CHIPS)*, 2017, pp. 1-2.
22. Shin, Dongjoo, and Hoi-Jun Yoo, "DNPU: An energy-efficient deep neural network processor with on-chip stereo matching," in

Proc. Hot Chips: A Symposium on High Performance Chips Hot Chips, 2017.
23. Ruder, and Sebastian, "An overview of gradient descent optimization algorithms," *CoRR*, vol. abs/1609.04747, 2016.
24. Bhandare Ashwin, Maithili Bhide, Pranav Gokhale, and Rohan Chandavarkar, "Applications of convolutional neural networks," *International J. of Computer Science and Information Technologies*, vol. 7, pp. 2206-2215, 2016.
25. Bai Yalong, Kuiyuan Yang, Tao Mei, Wei-Ying Ma, and Tiejun Zhao, "Automatic data augmentation from massive Web images for deep visual recognition," *ACM Trans. on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 3, pp. 1-20, 2018.
26. Neary Patrick, "Automatic hyperparameter tuning in deep convolutional neural networks using asynchronous reinforcement learning," in *Proc. IEEE International Conf. on Cognitive Computing*, 2018, pp. 73-77.
27. Wang Shuqiang, Yanyan Shen, Dewei Zeng, and Yong Hu, "Bone age assessment using convolutional neural networks," in *Proc. International Conf. on Artificial Intelligence and Big Data*, 2018, pp. 175-178.
28. Kim Chang Hoon, Espoir K. Kabanga, and Sin-Jae Kang, "Classifying malware using convolutional gated neural network," in *Proc. 20th International Conf. on Advanced Communication Technology*, 2018, pp. 40-44.
29. LeCun Yann, Koray Kavukcuoglu, and Clément Farabet, "Convolutional networks and applications in vision," in *Proc. IEEE international symposium on circuits and systems*, 2010, pp. 253-256.
30. Duvenaud David K., Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *Proc. Advances in neural information processing systems*, 2015, pp. 2224-2232.
31. Rai Siddhant, Akshayanand Raut, Akash Savaliya, and Radha Shankarmani, "Darwin: convolutional neural network based intelligent health assistant," in *Proc. Second International Conf. on Electronics, Communication and Aerospace Technology*, 2018, pp. 1367-1371.
32. Wang Jinzhuo, Wenmin Wang, Ronggang Wang, and Wen Gao, "Deep alternative neural network: Exploring contexts as early as possible for action recognition," in *Proc. Advances in Neural Information Processing Systems*, 2016, pp. 811-819.
33. Wu Yue, Jun Li, Yu Kong, and Yun Fu, "Deep convolutional neural network with independent softmax for large scale face recognition," in *Proc. 24th ACM international conf. on Multimedia*, 2016, pp. 1063-1067.
34. Xi Rui, Mengshu Hou, Mingsheng Fu, Hong Qu, and Daibo Liu, "Deep dilated convolution on multimodality time series for human activity recognition," in *Proc. International Joint Conf. on Neural Networks*, 2018, pp. 1-8.
35. Zhang Longfei, and Yanming Guo, "Delving into Fully Convolutional Networks Activations for Visual Recognition," in *Proc. 3rd International Conf. on Multimedia and Image Processing*, 2018, pp. 99-104.
36. Hahnloser Richard HR, Rahul Sarpeshkar, Misha A. Mahowald, Rodney J. Douglas, and H. Sebastian Seung, "Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit," *Nature*, vol. 405, no. 6789, pp. 947-951, 2000.
37. Srivastava Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The j. of machine learning research*, vol. 15, no. 1, pp. 1929-1958, 2014.
38. Nebauer Claus, "Evaluation of convolutional neural networks for visual recognition," *IEEE trans. on neural networks*, vol. 9, no. 4, pp. 685-696, 1998.
39. Kinnikar Ashwini, Moula Husain, and S. M. Meena, "Face recognition using Gabor filter and convolutional neural network," in *Proc. International Conf. on Informatics and Analytics*, 2016, pp. 1-4.
40. Do Nhu-Tai, Soo-Hyung Kim, Hyung-Jeong Yang, Guee-Sang Lee, and In-Seop Na, "Face tracking with convolutional neural network heat-map," in *Proc. 2nd International Conf. on Machine Learning and Soft Computing*, 2018, pp. 29-33.
41. Azizpour Hossein, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson, "Factors of transferability for a generic convnet representation," *IEEE trans. on pattern analysis and machine intelligence*, vol. 38, no. 9, pp. 1790-1802, 2015.
42. Michael Mathieu, Mikael Henaff, and Yann LeCun, "Fast training of convolutional networks through FFTs," in *Proc. 2nd International Conf. on Learning Representations*, Banff, AB, Canada, April, 2014.
43. Ren Shaoqing, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proc. Advances in neural information processing systems*, 2015, pp. 91-99.
44. Kagaya Hokuto, Kiyoharu Aizawa, and Makoto Ogawa, "Food detection and recognition using convolutional neural network," in *Proc. 22nd ACM international conf. on Multimedia*, 2014, pp. 1085-1088.
45. Murugan Pushparaja, "Hyperparameters optimization in deep convolutional neural network/bayesian approach with gaussian process prior," *CoRR*, vol. abs/1712.07233, 2017.
46. Gatys Leon A., Alexander S. Ecker, and Matthias Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE conf. on computer vision and pattern recognition*, 2016, pp. 2414-2423.
47. Sainath Tara N., Brian Kingsbury, Abdel-rahman Mohamed, George E. Dahl, George Saon, Hagen Soltau, Tomas Beran, Aleksandr Y. Aravkin, and Bhuvana Ramabhadran, "Improvements to deep convolutional neural networks for LVCSR," in *Proc. IEEE workshop on automatic speech recognition and understanding*, 2013, pp. 315-320.
48. Nguyen Kien, Clinton Fookes, and Sridha Sridharan, "Improving deep convolutional neural networks with unsupervised feature learning," in *Proc. IEEE International Conf. on Image Processing*, 2015, pp. 2270-2274.
49. Maurya Ajay K., Divyesh M. Varade, and Onkar Dikshit, "Effect of Pansharpening in Fusion Based Change Detection of Snow Cover Using Convolutional Neural Networks," *IETE Technical Review*, pp. 1-11, 2019.
50. Özbulak, Gökhan, and Hazim Kemal Ekenel, "Initialization of convolutional neural networks by Gabor filters," in *Proc. 26th Signal Processing and Communications Applications*, 2018, pp. 1-4.
51. Wang, Xing, Him Wai Ng, and Jie Liang, "Lapped convolutional neural networks for embedded systems," in *Proc. IEEE Global Conf. on Signal and Information Processing*, 2017, pp.1135-1139.
52. Wang Qiang, Huijie Fan, Yang Cong, and Yandong Tang, "Large receptive field convolutional neural network for image super-resolution," in *Proc. IEEE International Conf. on Image Processing*, 2017, pp. 958-962.
53. Karpathy Andrej, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE conf. on Computer Vision and Pattern Recognition*, 2014, pp. 1725-1732.
54. Noh Hyeonwoo, Seunghoon Hong, and Bohyung Han, "Learning deconvolution network for semantic segmentation,"

in *Proc. IEEE international conf. on computer vision*, 2015, pp. 1520-1528.
55. Akhand M. A. H., Mahtab Ahmed, MM Hafizur Rahman, and Md Monirul Islam, "Convolutional neural network training incorporating rotation-based generated patterns and handwritten numeral recognition of major Indian scripts," *IETE J. Research*, vol. 64, no. 2, pp. 176-194, 2018.
56. Zhao Hang, Orazio Gallo, Iuri Frosio, and Jan Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. on computational imaging*, vol. 3, no. 1, pp. 47-57, 2016.
57. Kabanga Espoir K., and Chang Hoon Kim, "Malware images classification using convolutional neural network," *J. of Computer and Communications*, vol. 6, no. 1, pp. 153-158, 2017.
58. McCulloch, W.S., and Pitts, W., "A logical calculus of the ideas immanent in nervous activity," *Bulletin of Mathematical Biophysics*, vol. 5, pp. 115–133, 1943.
59. Niepert Mathias, Mohamed Ahmed, and Konstantin Kutzkov, "Learning convolutional neural networks for graphs," in *Proc. International conf. on machine learning*, 2016, pp. 2014-2023.
60. Li Hongzhi, "Multimodal visual pattern mining with convolutional neural networks," in *Proc. ACM on International Conf. on Multimedia Retrieval*, 2016, pp. 427-430.
61. Farfade Sachin Sudhakar, Mohammad J. Saberian, and Li-Jia Li, "Multi-view face detection using deep convolutional neural networks," in *Proc. 5th ACM on International Conf. on Multimedia Retrieval*, 2015, pp. 643-650.
62. Zhou Bolei, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Object detectors emerge in deep scene cnns," in *Proc. 3rd International Conf. on Learning Representations*, 2015, pp. 1-12.
63. Etemad Elham, and Qigang Gao, "Object localization by optimizing convolutional neural network detection score using generic edge features," in *Proc. IEEE International Conf. on Image Processing*, 2017, pp. 675-679.
64. Alom Md Zahangir, Mahbubul Alam, Tarek M. Taha, and Khan M. Iftekharuddin, "Object recognition using cellular simultaneous recurrent networks and convolutional neural network," in *Proc. International Joint Conf. on Neural Networks (IJCNN)*, 2017, pp. 2873-2880.
65. Kim Edward J., and Robert J. Brunner, "Star-galaxy classification using deep convolutional neural networks," *Monthly Notices of the Royal Astronomical Society*, Vol. 464, No. 4, pp.4463-4475, Feb. 2017.
66. Lee Honglak, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proc. 26th annual international conf. on machine learning*, 2009, pp. 609-616.
67. Hubel David H., and Torsten N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *The J. of physiology*, vol. 148, no. 3, pp. 574-591, 1959.
68. Visin Francesco, Kyle Kastner, Kyunghyun Cho, Matteo Matteucci, Aaron Courville, and Yoshua Bengio, "Renet: A recurrent neural network based alternative to convolutional networks," *CoRR*, vol. abs/1505.00393, 2015.
69. Wei Qingdong, Fengjing Shao, and Ji Liu, "Research Summary of Convolution Neural Network in Image Recognition," in *Proc. of the International Conf. on Data Processing and Applications*, 2018, pp. 39-44.
70. Sankaranarayanan Swami, Arpit Jain, and Ser Nam Lim, "Guided perturbations: Self-corrective behavior in convolutional neural networks," in *Proc. IEEE International Conf. on Computer Vision*, 2017, pp. 3562-3570.
71. Guo Tianmei, Jiwen Dong, Henjian Li, and Yunxing Gao, "Simple convolutional neural network on image classification," in *Proc. 2nd International Conf. on Big Data Analysis*, 2017, pp. 721-724.
72. Howard Andrew G, "Some improvements on deep convolutional neural network based image classification," in *Proc. 3rd International Conf. on Learning Representations*, 2014.
73. Na Taesik, and Saibal Mukhopadhyay, "Speeding up convolutional neural network training with dynamic precision scaling and flexible multiplier-accumulator," in *Proc. International Symposium on Low Power Electronics and Design*, 2016, pp. 58-63.
74. Bay Herbert, Tinne Tuytelaars, and Luc Van Gool, "Surf: Speeded up robust features," in *Proc. European Conf. on Computer Vision*, Berlin, 2006, pp. 404-417.
75. Chen Xinyu, Qiang Guan, Xin Liang, Li-Ta Lo, Simon Su, Trilce Estrada, and James Ahrens, "Tensorview: visualizing the training of convolutional neural network using paraview," in *Proc. 1st Workshop on Distributed Infrastructures for Deep Learning*, 2017, pp. 11-16.
76. Hattikatti Pratiksha, "Texture based interstitial lung disease detection using convolutional neural network," in *Proc. International Conf. on Big Data, IoT and Data Science*, 2017, pp. 18-22.
77. Li, Li-Jia, Richard Socher, and Li Fei-Fei, "Towards total scene understanding: Classification, annotation and segmentation in an automatic framework," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2009, pp. 2036-2043.
78. Shen Xu, Xinmei Tian, Anfeng He, Shaoyan Sun, and Dacheng Tao, "Transform-invariant convolutional neural networks for image classification and search," in *Proc. 24th ACM international conf. on Multimedia*, 2016, pp. 1345-1354.
79. Albawi Saad, Tareq Abed Mohammed, and Saad Al-Zawi, "Understanding of a convolutional neural network," in *Proc. International Conf. on Engineering and Technology*, 2017, pp. 1-6.
80. Du Simon S., Yining Wang, Xiyu Zhai, Sivaraman Balakrishnan, Russ R. Salakhutdinov, and Aarti Singh, "How many samples are needed to estimate a convolutional neural network," in *Proc. Advances in Neural Information Processing Systems*, 2018, pp. 373-383.
81. Radford Alec, Luke Metz, and Soumith Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. 4th International Conf. on Learning Representations*, 2016.
82. Lin Wen-Hui, Hsiao-Chung Lin, Ping Wang, Bao-Hua Wu, and Jeng-Ying Tsai, "Using convolutional neural networks to network intrusion detection for cyber threats," in *Proc. IEEE International Conf. on Applied System Invention (ICASI)*, 2018, pp. 1107-1110.
83. Zhang Cha, John C. Platt, and Paul A. Viola, "Multiple instance boosting for object detection," in *Proc. Advances in neural information processing systems*, 2006, pp. 1417-1424.
84. Amato Giuseppe, Fabrizio Falchi, and Lucia Vadicamo, "Visual recognition of ancient inscriptions using convolutional neural network and fisher vector," *J. Computing and Cultural Heritage (JOCCH)*, vol. 9, no. 4, pp. 1-24, 2016.
85. Brinkmann Eva-Maria, Martin Burger, Julian Rasch, and Camille Sutour, "Bias reduction in variational regularization," *J. Mathematical Imaging and Vision*, vol. 59, no. 3, pp. 534-566, 2017.
86. Yang Tien-Ju, Yu-Hsin Chen, and Vivienne Sze, "Designing energy-efficient convolutional neural networks using energy-aware pruning," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2017, pp. 5687-5695.

87. Flamary Rémi, "Astronomical image reconstruction with convolutional neural networks," in *Proc. 25th European Signal Processing*, 2017, pp. 2468-2472.
88. Calonder Michael, Vincent Lepetit, Christoph Strecha, and Pascal Fua, "Brief: Binary robust independent elementary features," in *Proc. European conf. on computer vision*, Berlin, 2010, pp. 778-792.
89. Rublee Ethan, Vincent Rabaud, Kurt Konolige, and Gary Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. International conf. on computer vision*, 2011, pp. 2564-2571.
90. Lin Sangdi, and George C. Runger, "GCRNN: Group-constrained convolutional recurrent neural network," *IEEE trans. on neural networks and learning systems*, vol. 29, no. 10, pp. 4709-4718, 2017.
91. Kavukcuoglu Koray, Pierre Sermanet, Y-Lan Boureau, Karol Gregor, Michaël Mathieu, and Yann L. Cun. "Learning convolutional feature hierarchies for visual recognition," in *Proc. Advances in neural information processing systems*, 2010, pp. 1090-1098.
92. Lee Hyunkwang, Shahein Tajmir, Jenny Lee, Maurice Zissen, Bethel Ayele Yeshiwas, Tarik K. Alkasab, Garry Choy, and Synho Do, "Fully automated deep learning system for bone age assessment," *J. of digital imaging*, vol. 30, no. 4, pp. 427-441, 2017.
93. Hinton Geoffrey E., Alex Krizhevsky, and Sida D. Wang, "Transforming auto-encoders," in *Proc. International conf. on artificial neural networks*, Berlin, Heidelberg, 2011, pp. 44-51.
94. Silva Cecilia F., and Clauirton A. Siebra. "An investigation on the use of convolutional neural network for image classification in embedded systems," in *Proc. IEEE Latin American Conf. on Computational Intelligence (LA-CCI)*, 2017, pp. 1-6.
95. Stivaktakis Radamanthys, Grigorios Tsagkatakis, Bruno Moraes, Filipe Abdalla, Jean-Luc Starck, and Panagiotis Tsakalides, "Convolutional neural networks for spectroscopic redshift estimation on euclid data," *IEEE Trans. Big Data*, Aug. 2019.
96. Mirmohammadsadeghi Moein, Samer S. Hanna, and Danijela Cabric, "Modulation classification using convolutional neural networks and spatial transformer networks," in *Proc. 51st Asilomar Conf. on Signals, Systems, and Computers*, 2017, pp. 936-939.
97. Voigtlaender Paul, and Bastian Leibe, "Online adaptation of convolutional neural networks for the 2017 davis challenge on video object segmentation," in *Proc. The 2017 DAVIS Challenge on Video Object Segmentation-CVPR Workshops*, vol. 5, no. 6. 2017.
98. Bengio Yoshua, Aaron Courville, and Pascal Vincent, "Representation learning: A review and new perspectives," *IEEE trans. on pattern analysis and machine intelligence*, vol. 35, no. 8 , 2013, pp. 1798-1828.
99. Toledo J. Ignacio, Manuel Carbonell, Alicia Fornés, and Josep Lladós, "Information extraction from historical handwritten document images with a context-aware neural model," *Pattern Recognition*, vol. 86, pp. 27-36, 2019.
100. Shankar Tanmay, Santosha K. Dwivedy, and Prithwijit Guha, "Reinforcement learning via recurrent convolutional neural networks," in *Proc. 23rd International Conf. on Pattern Recognition (ICPR)*, 2016, pp. 2592-2597.
101. Krizhevsky Alex, "One weird trick for parallelizing convolutional neural networks," *CoRR*, vol. abc/1404.5997, 2014.
102. Koutník Jan, Jürgen Schmidhuber, and Faustino Gomez, "Evolving deep unsupervised convolutional networks for vision-based reinforcement learning," in *Proc. Annu. Conf. on Genetic and Evolutionary Computation*, 2014, pp. 541-548.
103. Mnih Volodymyr, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *Proc. International conf. on machine learning*, 2016, pp. 1928-1937.
104. Gibert Daniel, "Convolutional neural networks for malware classification," *Masters of Artificial Intelligence thesis, Department of Computer Science*, University Rovira i Virgili, Tarragona, Spain, 2016.
105. Young Tom, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria, "Recent trends in deep learning based natural language processing," *IEEE Computational intelligenCe magazine*, vol. 13, no. 3, pp. 55-75, 2018.
106. Ding Chen, Ying Li, Yong Xia, Wei Wei, Lei Zhang, and Yanning Zhang, "Convolutional neural networks based hyperspectral image classification method with adaptive kernels," *Remote Sensing*, vol. 9, No.6, pp. 1-15, June 2017.
107. DeVries Phoebe MR, T. Ben Thompson, and Brendan J. Meade, "Enabling large-scale viscoelastic calculations via neural network acceleration," *Geophysical Research Letters*, vol. 44, no. 6, pp. 2662-2669, 2017.
108. Shotton Jamie, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake, "Real-time human pose recognition in parts from single depth images," in *Proc. Computer Vision and Pattern Recognition*, 2011, pp. 1297-1304.
109. Garbin Daniele, Elisa Vianello, Olivier Bichler, Quentin Rafhay, Christian Gamrat, Gérard Ghibaudo, Barbara DeSalvo, and Luca Perniola, "HfO 2-based OxRAM devices as synapses for convolutional neural networks," *IEEE Trans. Electron Devices*, vol. 62, pp. 2494-2501, 2015
110. Zhang Tianyuan, and Zhanxing Zhu, "Interpreting adversarially trained convolutional neural networks," in *Proc. 36th International Conf. on Machine Learning*, 2019.
111. Khasanova Renata, and Pascal Frossard, "Geometry Aware Convolutional Filters for Omnidirectional Images Representation," in *Proc. International Conf. on Machine Learning*, 2019, pp. 3351-3359.
112. Donahue Jeff, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *Proc. International conf. on machine learning*, 2014, pp. 647-655.
113. Kong Yu, and Yun Fu, "Human action recognition and prediction: A survey," *CoRR*, vol. abs/1806.11230, 2018.
114. Bjorck Nils, Carla P. Gomes, Bart Selman, and Kilian Q. Weinberger, "Understanding batch normalization," in *Proc. Advances in Neural Information Processing Systems*, 2018, pp. 7694-7705.
115. Wang Tao, David J. Wu, Adam Coates, and Andrew Y. Ng, "End-to-end text recognition with convolutional neural networks," in *Proc. 21st International Conf. on Pattern Recognition*, 2012, pp. 3304-3308.
116. Lowe David G, "Distinctive image features from scale-invariant keypoints," *International j. of computer vision*, vol. 60, no. 2, pp. 91-110, 2004.
117. Vedaldi Andrea, and Brian Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," in *Proc. 18th ACM international conf. on Multimedia*, 2010, pp. 1469-1472.
118. Jaderberg Max, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Reading text in the wild with convolutional neural networks," *International J. of Computer Vision*, vol. 116, no. 1, pp. 1-20, 2016.
119. He Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Spatial pyramid pooling in deep convolutional networks for

visual recognition," *IEEE trans. on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904-1916, 2015.
120. Evgeny A.Smirnov, Denis M.Timoshenko, and Serge N. Andrianov, "Comparison of Regularization Methods for ImageNet Classification with Deep ConvolutionalNeural Networks," *AASRI Procedia*, vol. 6, pp. 89-94, May 2014.
121. Murugan Pushparaja, and Shanmugasundaram Durairaj, "Regularization and optimization strategies in deep convolutional neural network," *arXiv*:1712.04711, 2017.
122. Johnson Rie, and Tong Zhang, "Semi-supervised convolutional neural networks for text categorization via region embedding," in *Proc. Advances in neural information processing systems*, 2015, pp. 919-927.
123. Kalchbrenner Nal, Edward Grefenstette, and Phil Blunsom, "A convolutional neural network for modelling sentences," in *Proc. 52nd Annual Meeting of the Association for Computational Linguistics*, 2014.
124. Dos Santos, Cicero, and Maira Gatti, "Deep convolutional neural networks for sentiment analysis of short texts," in *Proc. 25th International Conf. on Computational Linguistics: Technical Papers*, 2014, pp. 69-78.
125. Shen Yelong, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil, "A latent semantic model with convolutional-pooling structure for information retrieval," in *Proc. 23rd ACM international conf. on information and knowledge management*, 2014, pp. 101-110.
126. Abdel-Hamid Ossama, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533-1545, 2014.
127. Severyn Aliaksei, and Alessandro Moschitti, "Modeling relational information in question-answer pairs with convolutional neural networks," *CoRR*, vol. abs/1604.01178, 2016.
128. Anuse Alwin, and Vibha Vyas, "A novel training algorithm for convolutional neural network," *Complex & Intelligent Systems*, vol. 2, no. 3, pp. 221-234, 2016.
129. Chen Linkai, Feiyue Ye, Yaduan Ruan, Honghui Fan, and Qimei Chen, "An algorithm for highway vehicle detection based on convolutional neural network," *EURASIP J. on Image and Video Processing*, vol. 2018, no. 1, pp. 109, 2018.
130. Islam Jyoti, and Yanqing Zhang, "Brain MRI analysis for Alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks," *Brain informatics*, vol. 5, no. 2, pp. 2, 2018.
131. Aykanat Murat, Özkan Kılıç, Bahar Kurt, and Sevgi Saryal, "Classification of lung sounds using convolutional neural networks," *EURASIP J. on Image and Video Processing*, vol. 2017, no. 1, pp. 65, 2017.
132. Costa Pedro, and Aurélio Campilho, "Convolutional bag of words for diabetic retinopathy detection from eye fundus images," *IPSJ Trans. on Computer Vision and Applications*, vol. 9, no. 1, pp. 1-6, 2017.
133. Yamashita Rikiya, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into imaging*, vol. 9, no. 4, pp. 611-629, 2018.
134. Chen Liang-Chieh, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE trans. on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834-848, 2017.
135. Wang Qing, and Rong Zhang, "Double JPEG compression forensics based on a convolutional neural network," *EURASIP J. on Information Security*, vol. 2016, no. 1, pp. 1-23, 2016.
136. Bian Peng, Wanwan Li, Yi Jin, and Ruicong Zhi, "Ensemble feature learning for material recognition with convolutional neural networks," *EURASIP J. on Image and Video Processing*, vol. 2018, no. 1, pp. 1-64, 2018.
137. Zafar Umara, Mubeen Ghafoor, Tehseen Zia, Ghufran Ahmed, Ahsan Latif, Kaleem Razzaq Malik, and Abdullahi Mohamud Sharif, "Face recognition with Bayesian convolutional networks for robust surveillance systems," *EURASIP J. on Image and Video Processing*, vol. 2019, no. 1, pp. 10, 2019.
138. Long Jonathan, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE conf. on computer vision and pattern recognition*, 2015, pp. 3431-3440.
139. Turan Bilal, Taisuke Masuda, Anas Mohd Noor, Koji Horio, Toshiki I. Saito, Yasuyuki Miyata, and Fumihito Arai, "High accuracy detection for T-cells and B-cells using deep convolutional neural networks," *ROBOMECH J.*, vol. 5, no. 1 , pp. 1-29, 2018.
140. Tóth László, "Phone recognition with hierarchical convolutional deep maxout networks," *EURASIP J. on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1-13, 2015.
141. Chen Liang-Chieh, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs, " in *Proc. 3rd International Conf. on Learning Representations*, ICL, san Diego, CA, USA, May 7-9, 2015.
142. Wang Jason, and Luis Perez, "The effectiveness of data augmentation in image classification using deep learning," *CoRR*, Vol. abs/1712.04621, 2017.
143. Zhang Zhong, and Donghong Li, "Transfer deep convolutional activation-based features for domain adaptation in sensor networks," *EURASIP J. on Wireless Communications and Networking*, vol. 2018, no. 1, pp. 49, 2018.
144. Hijazi Samer, Rishi Kumar, and Chris Rowen, "Using convolutional neural networks for image recognition," *Cadence*, 2015, Available: https://ip.cadence.com/uploads/901/cnnwp.pdf.
145. Bengio Yoshua, "Practical recommendations for gradient-based training of deep architectures," *Lecture Notes in Computer Science*, vol. 7700, pp. 437-478, 2012.
146. Wan Li, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus, "Regularization of neural networks using dropconnect," in *Proc. International conf. on machine learning*, 2013, pp. 1058-1066.
147. Li Zuhe, Yangyu Fan, and Weihua Liu, "The effect of whitening transformation on pooling operations in convolutional autoencoders," *EURASIP J. on Advances in Signal Processing*, vol. 2015, no. 1, pp. 1-11, 2015.
148. Salcic Zoran, Stevan Berber, and Paul Secker, "FPGA prototyping of RNN decoder for convolutional codes," *EURASIP J. on Advances in Signal Processing*, vol. 2006, No. 1, Dec. 2006.
149. Chen Sheng, and Hongxiang He, "Stock prediction using convolutional neural network," *IOP Conf. Series: Materials Science and Engineering*, vol. 435, no. 1, pp. 1-9, Nov. 2018.

Appendix:

Table 1: Some Types of pooling layers

Name of the pooling layer	Tasks suitable for	Intuition
Spatial pyramid pooling[119], Volumetric pyramid pooling	Semantic segmentation	Extraction of important image patch using weights(learning)
Second-order pooling[50]	Visual Recognition	It doesn't assume unimodal distributions to fully capture statistics of conv. simulations of samples with complex contents, a global Gated Mixture of Second-order Pooling (GM-SOP) is a simple direct is ensemble of multiple models for task summarization.
[147] Whitening transformation assists in dimensionality reduction by discarding the redundancy among neighboring pixels, pooling does the same but by lowering the resolution of the image, [147] in context of Convolutional auto-encoders which are unsupervised learning algorithms wherein the network learns filters that can minimize the reconstruction error for reconstruction which is done in encoding part of the network; an adaptive pooling mechanism is employed to gain information about the effect of whitening on pooling operations in different conditions; the paper also houses technical uses of pooling.		

Table 2: Some Gradient descent (GD) variations

Name of the gradient descent variant	Batch GD	Stochastic GD	Mini batch GD
Speed [145]	Slow as for one update gradients are calculated for complete training dataset	Very slow as there's an update involved with calculation of gradient descent which is done for every training data sample	It provides an update for every mini batch of n training example each
Traceability [23]	Intractable	Intractable	Tractable
Convergence [23] [145]	Guaranteed to converge to global minimum for convex objectives and local minima for non-convex	Convergence is mostly dependent on the value of learning rate	Convergence is optimal and speeded up

Table 3: Some techniques for regularization of Neural Network Models

Techniques of generalization/regularization [120]]121]	**Why?**
Dropout[37]	At each epoch during the training a random sampling technique is used to activate neuron, this doesn't enforce the network to memorize the training data.
Dropconnect [146]	It is same as dropout except that the rather than neurons' activations getting sampled, it's the weights or connections between the neurons (the strength of the activations of neurons or the contribution of neuron to the learning).
Batch normalization [114]	Making learning less dependent on hyper-parameter choices, making end-to-end training feasible through the use of normalization layer and works even in situations with bad weight initializations.
Ensemble model averaging [130]	The final performance metrics is a combination (weighted combination of all predictions, average of all the predictions, etc.) from various models that have performed learning of the same task. Tasks are more accurate, and over-fitting is minimized.
L2 norm	Addition of regularization term penalizes big coefficients and tries to minimize them to zero, although not making them exactly to zero, it seeks to reduce the MSE by adding some bias and, at the same time, reducing the variance. Remember high variance correlates to an over-fitting model and is called as ridge regression.
Elastic net constraints	Combines L1 and L2 norm.
Max-Norm	Network parameters are less memory expensive and belong to a reasonable range of values, because an upper bound constraint is applied to the norm of all incoming weights for all neurons.
Early stopping	The learning is regulated (stopped or slowed down) when there is no further improvement of performance on the validation set (). It works well with iterative gradient-descent-based algorithms.
Data augmentation	Avoids over-fitting by increasing the amount of training data, which possible by performing simple image processing operations on the images, creating synthetic data, using some machine learning tools for producing data having some features exaggerated when compared to others thereby becoming new training examples.
Variance regularization [84]	Variational regularization is solving of the standard variation problem along with a consecutive de-biasing step which minimizes the data fidelity on a group of images which have a common thread running through them can be based on various aspects of visual data. The de-biasing is achieved by deploying infimal convolutions which try to reduce Bregman distance by moving towards optimality of this variational model.
DropBlock[28]	DropBlock is a spatially extending form of a structured dropout unit, where units in a continually connected region of an activation map are dropped at once; when applied in skip connections with the convolution layers increases its and more robust to hyper-parameter choices and beats dropout when exhaustively tested on various datasets for different tasks.

Table 4: Some commonly used alternative models to CNNs

Name of the alternate model[9]	**Definition**	**Tasks good at**	**Similarities with CNNs**	**Differences with CNNs**	**Performance metrics**
RBMs [66]	A generative model, which has stochasticity in its nature giving it learning ability making use of Bayesian probability distributions of training data.	dimensionality reduction, classification, collaborative filtering, feature learning and topic modeling.	Less number of parameters to be learnt.	Estimating of joint probabilities of the data accurately.	Accuracy, Precision and recall.
Autoencoders [93]	A type of unsupervised artificial neural networks capable	Dimensionality reduction.	Capable of performing transfer learning.	Falls into the unsupervised domain of ML and hence back-	Normalized "score" measuring the wellness of data

Name of the alternate model[9]	Definition	Tasks good at	Similarities with CNNs	Differences with CNNs	Performance metrics
	of generating efficient data coding.			propagation of errors is not guided.	representation, Precision and recall.
GANs [142]	Unsupervised neural network models used to synthesize data.	Generation of data.	Using of significantly smaller number of parameters to do the task of generation of data.	Falls into the unsupervised domain of ML and hence back-propagation of errors is not guided.	Inception score to measure the quality of generated images and their diversity, Precision and recall.
GANs [142]	Unsupervised neural network models used to synthesize data.	Generation of data.	Using of significantly smaller number of parameters to do the task of generation of data.	Falls into the unsupervised domain of ML and hence back-propagation of errors is not guided.	Inception score to measure the quality of generated images and their diversity, Precision and recall.
Deep belief Nets [66]	A generative graphical model constituting multiple RBMs (Stacked RBMs) .	dimensionality reduction, classification, collaborative filtering, feature learning and topic modeling.	Layers are good feature detectors and are hierarchical making them fit for classification tasks with supervision.	Estimating of joint probabilities of the data accurately and difficulty involved in doing the same.	Accuracy, Precision and recall.
CKNs [8]	CNNs with the ability of learning to approximate the generated kernel feature maps on the training instances.	Task at which CNNs are good.	Completely similar.	Invariance is encoded by a reproducing kernel.	Same as those of CNNs.
RNNs [148]	A class of ANNs with network structured in such a way that assists learning complemented by memory.	Handwriting recognition Speech recognition.	Neurons are connected to decompose the data feature's hierarchical behavior into subprograms [55][64][68].	Temporal learning algorithm.	Outputs of fitness/reward functions, Precision and recall.

Table 5: Some task in Natural Language Processing which use CNNs

Task in Natural Language Processing	Brief description
Sentence modeling	[123] Dynamic CNNs uses a dynamic k-max pooling layer, which makes it suitable for input sentences of varying length and generating feature graphs over sentences that are capable of capturing relations over short and long-range; the network can extend to any language
Environmental sounds [131]	[18] A multi-scale convolution operation can produce better audio representation fine tuning the frequency resolution and learning filters cross all frequencies, and to leverage the waveform-based features and spectrogram-based features together a two-phase method is used to fuse the features.
Speech recognition [140]	[126] The goal of automated speech recognition systems is the engendering of transcriptions of human speech into words; which is a daunting task to due high variability due to many factors like different accents, speaker's voice's attributes, environmental disturbances, and so on.

Task in Natural Language Processing	Brief description
Text classification/ Categorization	[122] A semi-supervised learning model with CNNs for text categorization; wherein word embedding is learnt for small text regions from unlabeled data which is then fed into a supervised CNN and suggestive of its usefulness for this task even though the training takes in unlabeled data samples
Sentiment analysis	Tasks involving sentiment analysis also require effective extraction of aspects along with their sentiment polarities [55][105][20][124]
Question answering	[127] CNNs formulates an optimum representation of Q&A word sequences by the use of relational information embedded into the matches between the set of words from the question and answer which is nothing but embeddings that are resulting in an increment in accuracy.
Document summarization	DCNNs learnt hierarchical features at document level and sentence level to capture low-level lexical features from the sentences into high-level semantic clues at the document level. [105]
Information retrieval	The CNN was used for making projections of queries and documents to be searched for, to a fixed-dimensional context space, and ranking of documents to be retrieved is done using cosine similarity; the model extracts rich contextual structures taking into consideration a temporal context window in the word sequences which is at n-gram level and max-pooled to form a unique sentence vector. [105] [125]
Machine translation	Sequential information should not lose its long-term dependency for machine translation to happen and hence they are not well suited for CNNs, speaking with regards to structures they possess, which lack this feature but the task is partially accomplished by considering both the similarity in meanings of the translation pair and their contexts in general to get good results amongst the benchmark datasets. [105]

Table 6: Task belonging to different domains other than CV and NLP accomplished using CNNs

Misc. tasks	Brief description
Cyber and computer security [104][82][28][57]	The generalization power of DL-based techniques is better compared to traditional ML-based approaches, DL based system can even detect zero-day malware; CNNs have the capacity of detecting even polymorphic malware. CNNs have also been able to utilize malware converted into grey-scale images and avoiding erroneous results production and adversarial attacks unlike other malware detection frameworks.
Financial stock price prediction [149]	Providing the DCNN with opening, high, low, closing price and stock's volume for prediction of financial stock prices resulting in accuracy and precision saturating up to human level.
Health assistant [31][76][130][131] [132][133][139][47]	[132] presented neural architecture is inspired by a bag-of-visual-word (BoVW) named Convolutional BoVW, CNNs can be seen as a cascaded pipeline capable of learning multiple tasks (feature extraction, coding and classification) as an end-to-end network, which is used to train a CNN, for detection of diabetic retinopathy from eye's fundus images, details reveal the need of human supervision in terms of variations in the patch size, selection of proper patch size and other parameters which are dataset/data and domain specific. [130] apart from a crucial med-app is also suggestive of a classical ML strategy to be leveraged even in DL era, which is creation of ensembling systems, helping a binary classification model of Alzheimer's diagnosis into detection of multiple stages of the disease too. A known fact to ponder again here is that many crucial medical indicators are implicitly extracted from MRI scans, all due to the hierarchical feature extraction taking place in CNNs. [47] This problem not only helps in diagnosis but helps in forming a better understanding of the human mind which can add to the generation of new functions capable of mimicking brain activity (CNNs are an example), rearing from the ability of convolutions to apprehend more complex information such as the EM neural signals(electroencephalography (EEG) or magneto encephalography (MEG)) produced by variations among multitude of factors of generation , hence the constructed algorithm is a multivariate (rendering insights into not only temporal but also spatial patterns) assumption-based one. [31] packs in an all-round healthcare system wherein based on the messages from the user, a message intent classifier is a trained analogous to model for text classification using CNNs which labels the message as belonging to a specific category which serves as an indicator of patient's health status. Limitations in availability of important indications in data about the understanding of useful extractable information can be overcome by CNNs and [139] serves as a good example of the same, wherein the limited amount of whole blood samples and sampling techniques and poor 3D spatial resolution still makes the high-accuracy detection and classification of T and B cells obtained from sub-microliter of blood using micro fluidic chip.

Misc. tasks	Brief description
Time series classification [90]	Time series analysis is difficult due to the necessity of order preservation, but CNNs have the capacity to achieve the same and used for financial stock price prediction.
Modulation classification [96]	Modulation levels and types is fundamental to signal processing applications, [96] proposes a framework to classify the types of modulations in a signal using a Spatial Transformer Network used in spatial invariant generation of image samples and here they do the same with the signals having channel influences which are then passed to CNNs for actual classification task.
An Image Processing application[49]	The problems of devising methods such as pan-sharpening to deal with spectral and spatial information preservation in images present as multispectral bands and panchromatic images , can be effectively handled using CNNs (ResNet) and thereby a methodology to predict the change in snow cover from satellite images is improvised and has suggested the potential disadvantages of performing the same task with other conventional techniques such as wavelet transforms, cluster analysis, etc.
Material Recognition[136]	[136] Ensemble learning of knowledge-based classifiers on their probability scores fed onto the extracted feature outputs of a CNN were efficient in dealing with the wide variety of textures, appearances varying with the illumination, colors (for example, different colored cement blocks) of numerous materials for their recognition which can be expanded into a segmentation and pixel level prediction problem as well. The important fact that manifolds of learning and feature analysis are possible at once in CNNs indicated by the fact that CNN generalizes and outperforms many methods like 3D texture recognition using bidirectional feature histograms, visual discriminative object-specific info, etc. to name a few.

Table 7: Scientific applications

Name of the Science/Engg. Related tasks	Description
Astronomy/Astrophysics	[87][95][65] deals with reconstruction of astronomical images which are but a convoluted output of observed object and a point spread function (PSF) as objects in the space are always moving away from every other object; using CNNs(generally considered to be a convex optimization problem with proximal splitting GD with super linear convergence) proven to be tractable. A priori in form of PSF is an added advantage of this domain-specific problem, PSNR is used to evaluate the network. Influence of more complex undiscovered PSF from the application of modern radio-interferometric celestial observation devices, and extensions to hyper spectral imaging with 3D image reconstruction which will require efficient image reconstruction strategies. Data generation technique can use more realistic simulators for radio-interferometry and other optical observations. [95] spectroscopic redshift (integral to measuring of celestial radial distances and positions of bodies in the space) estimation using CNNs (designed regression problem) reinforces the fact that low-level discrepancies prominent in datasets such as redshift in astronomical images (spectroscopic and photometric data) can be captured precisely which is hard with the noise rendered in the low signal-to-noise regions of such data by them
Chemistry	CNNs capable of training directly on graph [30] data are introduced to solve the problem of standard molecular feature extraction based on their fingerprints, intuition being interpretability and hence reliability/explain-ability of data-driven features [59]
Hyperspectral image analysis	[106] explores the hyperspectral features of a data and ways to include it in CNNs which are known for spatial correlation extraction, texture and object understanding based on pixel values. Intrinsic sequential data structure of a hyperspectral pixel is considered for proposed model; which prescribes the idea of utilizing intrinsic structure and modelling data as sequences wherever opportunity prevails.
Electrical and Electronics	Marriage of hardware and software level technologies is admirable as presented in [109]
Geology	[107] brings to light the processing of computationally intensive large-scale viscoelastic earthquake cycle models, and hence imprecise nature of the results can be attributed to the trade-off set between computation and spatial/temporal resolution. NNs minimize the risk and adds to the understanding of earthquake physics by delivering quick, reliable and precise calculations and inferences from earthquake data