ACADEMICIA: An International Multidisciplinary Research Journal
ISSN: 2249-7137 Vol. 11, Issue 12, December 2021 SJIF 2021 = 7.492
A peer reviewed journal

A REVIEW ON DEEP LEARNING FOR VISUAL UNDERSTANDING

Ramesh Chandra Tripathi*

*Professor,
Department of Computer Science, Faculty of Engineering,
Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, INDIA
Email id: tripathi.computers@tmu.ac.in

DOI: 10.5958/2249-7137.2021.02657.4

ABSTRACT

Deep learning algorithms are a kind of machine learning that aims to find many layers of distributed representations. To address conventional artificial intelligence issues, a number of deep learning methods have recently been suggested. This article attempts to summarize the state-of-the-art in computer vision deep learning algorithms by emphasizing contributions and difficulties from over 210 recent research publications. It begins by providing an overview of the different classifiers and their recent developments, followed by a brief description of their applications in a variety of vision applications, including image classification, object detection, image retrieval, semantic segmentation, and human pose estimation. Finally, the article outlines future trends and difficulties in neural network based design and training.

KEYWORDS: *Deep Learning, Image, Human Pose Estimation, Artificial Intelligence, Training.*

1. INTRODUCTION

Deep learning is a branch of machine learning that uses hierarchical structures to learn high-level abstractions from data. It is a new technique that has seen widespread use in classic AI areas including semantic parsing, learning techniques, natural language, computer vision, and many more. The flourishing of transfer learning today may be attributed to three major factors: substantially improved chip processing capacities (e.g. GPU units), much reduced computer hardware costs, and significant improvements in machine learning techniques. In recent years, many deep learning methods have been widely studied and debated. Schmidhuber et al. used a historical timeline approach to highlight key ideas and technical achievements, while Bengio looked at the difficulties of deep learning research and suggested a few forward-looking research paths. Deep networks have been proven to be effective for computer vision applications because they can extract relevant information while discriminating simultaneously[1].

Deep learning techniques have been extensively used by various researchers in recent ImageNet Large Scale Visual Recognition Challenge (ILSVRC) contests and have obtained top efficiency ratings. This survey is aimed for standard neural computing, computer vision, and media researchers who are interested in the current status of machine learning in machine learning. It provides an overview of different deep learning algorithms and their applications, with a focus on those that may be used in the field of computer vision. The following is how the remainder of the article is organized: Convolutional Neural Networks, Restricted Boltzmann Machines,

Autoencoder. In these categories, several well-known models, as well as their advancements, are mentioned. In this part, we also discuss the contributions and limits of various models. we discuss the accomplishments of deep learning methods in picture classification, object recognition, image retrieval, classification techniques, and human posture estimation in different computer vision applications. In the pipeline of their frequently used datasets, the outcomes of these applications are shown and contrasted[2]. Despite the success of deep learning techniques, we still confront a number of difficulties when building and training deep networks. In this part, we'll go over some of the main difficulties that deep learning faces, as well as some of the underlying patterns that may emerge in the future[3].

2. DISCUSSION

1. Convolutional Neural Networks (CNNs):

Convolutional Neural Networks (CNN) is one of the most well-known deep learning methods in which several layers are robustly taught. It has been shown to be extremely successful and is the most widely utilized in a variety of computer vision applications. This depicts the general CNN architecture's pipeline. Convolutional layers, pooling layers, and fully connected layers are the three major neural layers that make up a CNN. Distinct layers have different functions. A forward stage and a backward stage are both used to train the network. A forward stage's primary objective is to represent the input picture in each layer using the current parameters (weights and bias). The loss cost is then calculated using the prediction output and the ground truth labels. Second, the backward stage uses chain rules to calculate the gradients of each parameter depending on the loss cost. The gradients are used to update all of the parameters, which are then readied for the next forward calculation. The network learning may be halted after a sufficient number of forward and backward rounds. Following that, we'll go through the roles of each layer, as well as recent advancements, before summarizing the most popular network training methods. Finally, we discuss a few well-known CNN models, as well as derivative models, and the current trend of utilizing these models in real-world applications.[4]

1.1 *Types of Layers*:

In general, a CNN is a hierarchical neural network in which Convolutional and pooling layers alternate, followed by fully connected layers. The function of the three layers will be presented in this section, as well as a brief review of recent advances in research on those layers.

• *Layers of convolution:*

A CNN uses various kernels to convolve the entire image as well as the intermediate feature maps in the convolutional layers, resulting in various feature maps. The convolution operation has three main advantages: In the same feature map, the weight sharing mechanism reduces the number of parameters. Local connection learns correlations between pixels in close proximity. Invariance with respect to the object's position. Some well-known research publications utilize the convolution operation as a substitute for fully linked layers to speed up the learning process because of the advantages it provides. The Network in Network (NIN) method is a novel way of dealing with convolutional layers, in which the main idea is to replace linear filters with nonlinear neural networks by replacing the conventional convolutional layer with a small multilayer perceptron consisting of multiple fully connected layers with nonlinear activation functions. In terms of picture categorization, this approach performs well[5].

• *Layers that pool:*

A pooling layer is often employed after a convolutional layer to decrease the size of feature maps and network parameters. Pooling layers, like convolutional layers, are translation invariant because their calculations take into consideration adjacent pixels. The most frequently utilized methods are average pooling and maximum pooling. The output maps for 8 8 feature maps are reduced to 4 4 dimensions using a max pooling operator with size 2 2 and stride. Boureau et al. presented a thorough theoretical study of the performance of max pooling and average pooling. Scherer et al. compared the two pooling procedures and discovered that max-pooling may result in quicker convergence, better invariant feature selection, and improved generalization. Various rapid GPU implementations of CNN variations have been published in recent years, the majority of which use the max-pooling approach. Among the three layers, the pooling layer has received the most attention. There are three well-known methods to pooling layers, each with its own set of goals[6].

2. A plan for training:

Deep learning has the benefit of being able to construct deep architectures to learn more abstract knowledge when compared to shallow learning. The high number of factors added, however, may lead to another issue: over fitting. Several regularization techniques, like the stochastic pooling method described above, have recently developed in defense against over fitting. In this part, we'll go over a few more regularization methods that may affect training results[7].

3. CNN's structure:

Some well-known CNN models have developed as a result of recent advances in CNN methods in the computer vision area. In this part, we'll go through the most frequently used CNN models before summarizing their features and applications. Alex Net, a major CNN architecture with five convolutional layers and three fully linked layers, is a significant CNN architecture. The network would repeatedly convolve and pool the activations after inputting one fixed-size (224 224) picture, then transmit the results onto the fully-connected layers. The network was built using Image Net and used a variety of regularization methods including data augmentation, dropout, and so on. Alex Net took first place in the ILSVRC2012 competition, igniting a wave of interest in deep convolutional neural network designs. Nonetheless, this model has two significant flaws: 1) it requires a fixed picture resolution; there is no obvious explanation for why it works so effectively. In 2013, Zeiler et al. proposed a new visualization method for gaining insight into the intermediate feature's inner workings. They were able to identify architectures that outperformed Alex Net on the Image Net classification benchmark thanks to these visualizations, and the resultant model, Clarifai, won first place in the ILSVRC2013 competition.

In response to the need of a fixed resolution, developed a novel pooling method, spatial pyramid pooling, to overcome the picture size constraint. Despite their differences in construction, the SPP-net may improve the accuracy of a range of published CNN architectures. There are other methods to exploring deeper networks, in addition to the widely used CNN structure (five convolutional layers + three fully linked layers). VGG enhanced the network's depth by adding additional convolutional layers and using extremely tiny convolutional filters in all levels, in contrast to Alex Net. Similarly, Szegedy et al. presented GoogLe Net, a model with a somewhat deep structure (22 layers) that placed first in the ILSVRC2014 competition. Despite the fact that

many models have achieved top-tier classification results, CNN-related models and applications are not restricted to image classification. New frameworks have been developed based on these models to handle additional difficult problems such as object recognition, semantic segmentation, and so on. There are two well-known derivative frameworks: RCNN (Regions with CNN features) and FCN (completely convolutional network), which are primarily intended for object recognition and semantic segmentation[8].

4. Inference of activation:

We must infer the feature activations from a collection of weights. The Iterative Shrinkage-Thresholding Method (ISTA) is a common sparse coding inference algorithm that uses a gradient step to optimise the reconstruction term, followed by a sparsity term with a closed form shrinkage operation. Despite its simplicity and effectiveness, the algorithm has a serious flaw: it converges slowly. The issue is partially addressed by the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA), which maintains the computational simplicity of ISTA but converges faster thanks to the addition of a "momentum" component in the dynamics (the convergence complexity increased from O1=t to O1=t2). Both the ISTA and FISTA inferences require iterative optimization (i.e. LASSO), which is computationally intensive. Kavukcuoglu et al., on the other hand, used a feed-forward network to mimic the sparse codes, significantly speeding up the inference process. Furthermore, marginal regression was used to replace the LASSO optimization step in, successfully scaling up the sparse coding system to huge dictionaries[4].

5. Developments:

We will discuss several well-known algorithms related to sparse coding, in particular those that are employed in computer vision problems, in this subsection, as we have quickly indicated how to create the sparse representation given the objective function. Sparse coding SPM (ScSPM), which is an extension of the Spatial Pyramid Matching (SPM) technique, is one example of a sparse coding methodology. Unlike the SPM, which employs vector quantization (VQ) to describe images, the ScSPM uses sparse coding (SC) and multi-scale spatial max pooling. SC's codebook has an overabundance of options, and each feature can only activate a tiny number of them. SC has a significantly smaller reconstruction error than VQ owing to the less stringent restriction[9].

There are a total of nine properties. In further detail, 'generalization' refers to whether the method has been shown to work in a variety of media (e.g., text, pictures, audio) and applications, such as voice recognition and visual recognition. The capacity to train a deep model without supervision annotation is referred to as "unsupervised learning." The capacity to automatically learn features based on a data collection is known as "feature learning." The terms ‘real-time training' and ‘real-time prediction' relate to the speed with which learning and insinuating processes are completed. The terms ‘Biological understanding' and ‘Theoretical justification' indicate if the method has substantial biological or theoretical basis. If the method has been proven to be resistant to changes such as rotation, scaling, and translations, it is said to be invariant. The capacity to learn a deep model with a limited number of instances is referred to as a "small training set." It's essential to keep in mind that the table only depicts current results, not possible futures or specialized niche situations.

6. *Theoretical understandings:*

Although deep learning techniques have shown promise in solving computer vision problems, the underlying theory is not well known, and there is no clear understanding of certain architectures will perform better than others.

It's difficult to know which architecture, how so many layers, or how many nodes in each layer are appropriate for a given job, and it's also tough to select reasonable parameters like the learning rate, the secularizer's intensity, and so on. Historically, architectural design has been decided on an ad hoc basis.

A theoretical approach for finding the optimum number of feature maps was presented by Chu et al. This theoretical approach, however, only worked for very tiny receptive fields. Created a modeling technique that revealed the role of intermediate feature layers to better comprehend the behavior of well-known CNN designs. It opened up new opportunities for improved architectural designs by exposing characteristics in interpretable patterns. RCNN tried to find the CNN learning process in addition to visualizing the features. During the training phase, the performance was evaluated layer by layer, and it was discovered that the convolutional can learn more generic features and transmit the majority of the CNN representational capacity, while the top fully-connected levels are domain-specific. Agrawal et al. examined the impact of several frequently used techniques on CNN performance, such as fine-tuning and pre-training, in addition to evaluating CNN features, and offered evidence-based intuitions for applying CNN models to computer vision issues. Despite advances in deep cognitive approach, there is still a lot of space for improvement in terms of developing and optimizing CNN structures to improve desired characteristics like invariance and class discrimination[10].

3. CONCLUSION

This article provides a thorough overview of deep learning and proposes a classification system for analyzing the current material. Convolutional Neural Networks, Restricted Boltzmann Machines, Autoencoder, and Sparse Coding are the four categories it classifies deep supervised learning under based on the fundamental model they are built from. The four classes' current state-of-the-art methods are thoroughly explored and evaluated. The article focuses on advances in CNN-based methods for computer vision tasks since they are the most widely used and appropriate for pictures. The accuracy of certain CNN-based algorithms has already surpassed that of human raters, according to several recent publications. Despite the encouraging outcomes to far, there is still a lot of opportunity for improvement. The underlying theoretical basis, for example, does not yet explain under what circumstances they would perform well or outperform other methods, or how to identify the optimum structure for a given job. This article discusses these issues and highlights current trends in deep neural network design and training, as well as potential future avenues that might be pursued.

REFERENCES

1. Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," Neurocomputing, 2016, doi: 10.1016/j.neucom.2015.09.116.

2. H. Wang and D.-Y. Yeung, "Towards Bayesian Deep Learning: A Survey," arXiv Prepr., 2016.

3. W. Hao and D. Y. Yeung, "Towards Bayesian Deep Learning: A Framework and Some Existing Methods," IEEE Trans. Knowl. Data Eng., 2016, doi: 10.1109/TKDE.2016.2606428.

4. R. Memisevic, "Learning to relate images," IEEE Trans. Pattern Anal. Mach. Intell., 2013, doi: 10.1109/TPAMI.2013.53.

5. C. W. Tyler and L. T. Likova, "The role of the visual arts in the enhancing the learning process," Frontiers in Human Neuroscience. 2012, doi: 10.3389/fnhum.2012.00008.

6. T. Molnar, "Spectre of the Past, Vision of the Future – Ritual, Reflexivity and the Hope for Renewal in Yann Arthus-Bertrand's Climate Change Communication Film 'Home,'" M/C J., 2012, doi: 10.5204/mcj.496.

7. Y. Zhang, "A foundation for the design and analysis of robotic systems and behaviors," 1994.

8. J. A. Laub, "Assessing the servant organization; Development of the Organizational Leadership Assessment (OLA) model. Dissertation Abstracts International," Procedia - Soc. Behav. Sci., 1999.

9. M. Zahedi, "De-/reconstruction of geometrical forms," Int. J. Des. Objects, 2017, doi: 10.18848/2325-1379/cgp/v11i04/1-10.

10. R. W. Zhao, Z. Wu, J. Li, and Y. G. Jiang, "Learning semantic feature map for visual content recognition," in MM 2017 - Proceedings of the 2017 ACM Multimedia Conference, 2017, doi: 10.1145/3123266.3123379.