DOIs:10.2017/IJRCS/202304019 --:-- Research Paper / Article / Review

Deep Learning for Computer Vision: Convolutional Neural Networks

[1] **Dr.Vineetha K. R.,** [2] **Kashyap Krishna K. R.**

[1] Associative Professor Nehru college of engineering and research centre, Thrissur, India
[2] Department of MCA, Nehru college of engineering and research centre, Thrissur, India
Email - vpvprakash@gmail.com , kashyapkrishnakr2000@gmail.com

Abstract: *Convolutional neural networks (CNNs) and its application to object detection in computer vision tasks are thoroughly discussed in this study. The study explains the various CNN components, their advantages and disadvantages, and their actual propensity for unexpected future developments. Also covered are transfer learning with CNNs and well-liked designs like YOLO and Faster R-CNN. In the paper, the benefits of CNNs are discussed, including their high accuracy, transfer learning, automated feature extraction, and effective image processing. It also highlights CNNs' susceptibility to hostile attacks and their substantial computing needs. The paper's conclusion lists several potential directions for further research and advancement, including weakly supervised learning, self-supervised learning, multitask learning, explainability, and continuous learning. In overall, this work is a valuable source for understanding CNNs and their potential to advance computer vision applications.*

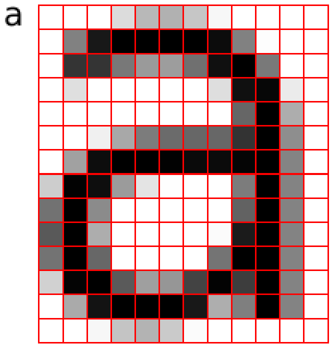
Key Words: *convolutional neural networks, CNNs, computer vision, object detection, image processing, transfer learning, automated feature extraction, weakly supervised learning, self-supervised learning, multi-task learning, memory-efficient architectures, advantages, disadvantages, future directions, explainability, noise robustness, high accuracy, adversarial attacks.*

1. INTRODUCTION:

This article provides an overview of convolutional neural networks (CNNs) and their use in computer vision applications, including object detection. Convolutional layers, convergence layers and fully connected layers are among the several explained parts of CNN. Other terms such as pitch, padding, activation function, dropout, learning rate, set size, optimization, and transfer learning are also explained. In addition, the essay discusses the advantages and disadvantages of CNNs, as well as future promises for continuous learning, explainability, weakly supervised learning, self-directed learning, multitasking learning, and memory-efficient design. Regarding Convolutional Neural Networks (CNN) and their use in computer vision applications, this article focuses on object detection. CNN networks have grown in popularity in recent years due to the efficiency and accuracy of their image processing. This work explains convolutional layers, pooling layers, and fully connected layers in detail, along with related ideas such as pitch, padding, activation function, stopping, learning rate, set size, optimization, and transfer learning. In addition to their future continuous learning, explainability, weakly supervised learning, self-directed learning, multitasking and memory-efficient design, the advantages and disadvantages of CNNs are also explored. To advance computer vision applications, the study aims to provide a comprehensive understanding of CNNs.

1.1 CONVOLUTIONAL NEURAL NETWORK

A type of neural network known as a convolutional neural network (also known as a CNN or ConvNet) is designed to process data with a grid-like topology, such as an image. The binary representation of visual data is a digital image. It consists of a series of pixels arranged in a grid pattern that indicate the brightness and color of each pixel.



1.0	1.0	1.0	0.9	0.6	0.6	0.6	1.0	1.0	1.0	1.0	1.0
1.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0	1.0	1.0
1.0	0.2	0.2	0.5	0.6	0.6	0.5	0.0	0.0	0.5	1.0	1.0
1.0	0.9	1.0	1.0	1.0	1.0	1.0	0.9	0.0	0.0	0.9	1.0
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0
1.0	1.0	1.0	0.5	0.5	0.5	0.5	0.5	0.4	0.0	0.5	1.0
1.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0
0.9	0.0	0.0	0.6	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0
0.5	0.0	0.6	1.0	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0
0.5	0.0	0.7	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.5	1.0
0.6	0.0	0.6	1.0	1.0	1.0	1.0	0.5	0.0	0.0	0.5	1.0
0.9	0.1	0.0	0.6	0.7	0.7	0.5	0.0	0.5	0.0	0.5	1.0
1.0	0.7	0.1	0.0	0.0	0.0	0.1	0.9	0.8	0.0	0.5	1.0
1.0	1.0	1.0	0.8	0.8	0.9	1.0	1.0	1.0	1.0	1.0	1.0

As soon as we see an image, our brain processes a lot of information. Each neuron has its own receptive field and is connected to other neurons so that it covers the entire visual field. In the same way that each neuron responds only within a limited region of the visual field, called the open field in the organic visual framework, each neuron in a CNN processes information only within its responsive field. The layers are first arranged to detect lines, curves and other simpler patterns. and more complex patterns (such as faces and objects) early. With CNN you can see computers.

2. LITERATURE REVIEW:

[1] The paper by K. Simonyan and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." arXiv preprint arXiv:1409.1556 (2014). The VGGNet architecture consists of 16-19 layers with small 3x3 convolutional filters and max-pooling layers. The authors show that this architecture achieves state-of-the-art performance on the ImageNet dataset, which contains over 1 million images and 1000 object categories. The paper also compares the performance of VGGNet with other deep learning architectures, such as AlexNet and ZFNet, and shows that VGGNet outperforms them. The paper highlights the importance of depth in neural network architectures for achieving high accuracy in image recognition tasks. The VGGNet architecture has since become a popular choice for various computer vision tasks, including object detection and segmentation.

[2] The paper by Saad Albawi; Tareq Abed Mohammed; Saad Al-Zawi "Understanding of a convolutional neural network" IEEE provides an overview of convolutional neural networks (CNNs) and their application in computer vision tasks. The paper explains the various components of CNNs, such as convolutional layers, pooling layers, and fully connected layers, as well as related concepts like stride, padding, activation function, dropout, learning rate, set size, optimizer, and transfer learning. The authors also discuss the advantages and disadvantages of CNNs, as well as their potential for advancing computer vision applications. The paper provides a valuable resource for understanding CNNs and their potential for advancing computer vision applications

[3] The paper "A Comparative Study of Convolutional Neural Network Architectures for Object Recognition in Images" by Sridharan Balasubramanian and Srinivasan Narasimhan. Published in Journal of Computer Science and Technology in 2020. The authors evaluate the performance of four different CNN architectures, namely AlexNet, VGGNet, GoogLeNet, and ResNet, on the CIFAR-10 dataset, which contains 60,000 images of 10 different object classes. The authors compare the accuracy, training time, and number of parameters of each architecture. The results show that ResNet achieves the highest accuracy, followed by GoogLeNet, VGGNet, and AlexNet. However, ResNet also has the highest number of parameters and longest training time. The authors conclude that the choice of CNN architecture depends on the specific requirements of the application, such as accuracy, computational resources, and training time. The paper provides a valuable resource for understanding the performance of different CNN architectures for object recognition in images.

3. METHOD:

How do convolutional neural networks work?

Convolutional neural networks differ from other neural networks in their superior performance with image, speech, or audio signal inputs. They have three main types of layers, which are:

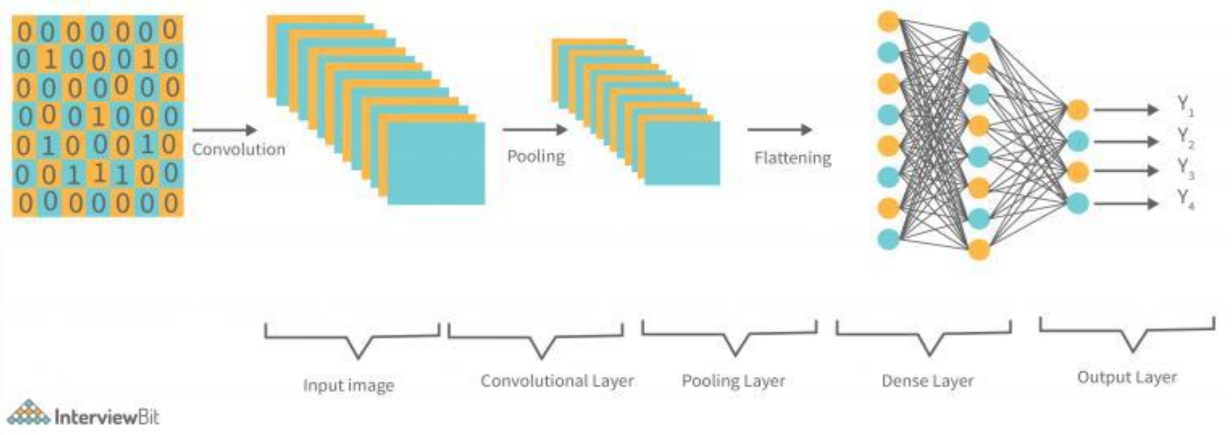
Convolutional layer

Pooling layer

Fully connected (FC) layer

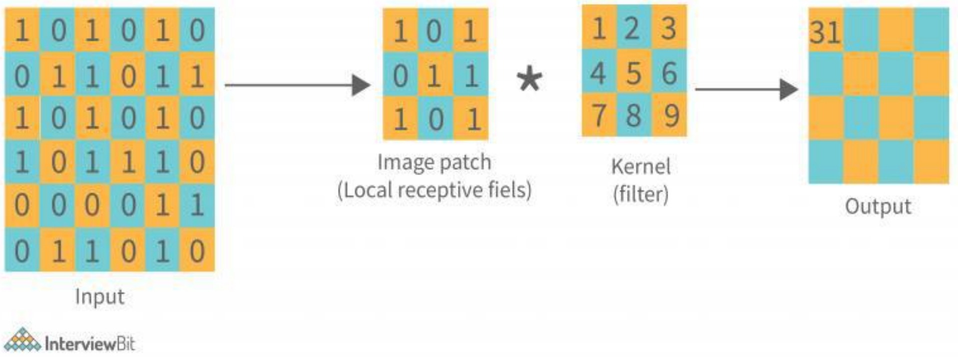
A convolutional layer is the first layer of a convolutional network. Although convolutional layers can be followed by more convolutional layers or convergence layers, the final layer is a fully connected layer. With each layer, the CNN increases in complexity and detects larger parts of the image. Earlier levels focused on simple features like colors and borders. As the image data moves through the layers of the CNN, it begins to recognize the larger elements or shapes of the object until it finally detects the target object.

❖ Typical CNN Architecture



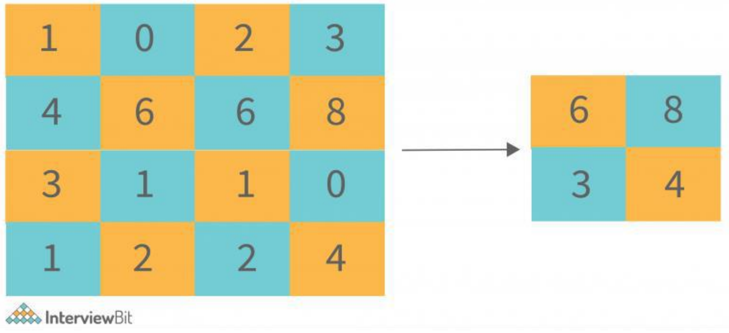
ConvNet's job is to compress the images into an easier to process format while preserving the essential elements for proper prediction. This is critical to designing an architecture that can learn features while being scalable to large data sets.

• Convolution layer (CONV): These are the basis of CNN and are responsible for performing **convolutional** operations. The kernel/filter is the component of this layer that performs the convolution operation (matrix). Until the entire image is scanned, the kernel makes horizontal and vertical adjustments based on the step rate. The core is smaller in size than pictured but has more depth. This means that if an image has three (RGB) channels, the height and width of the kernel will be spatially modest, but the depth will cover all three.

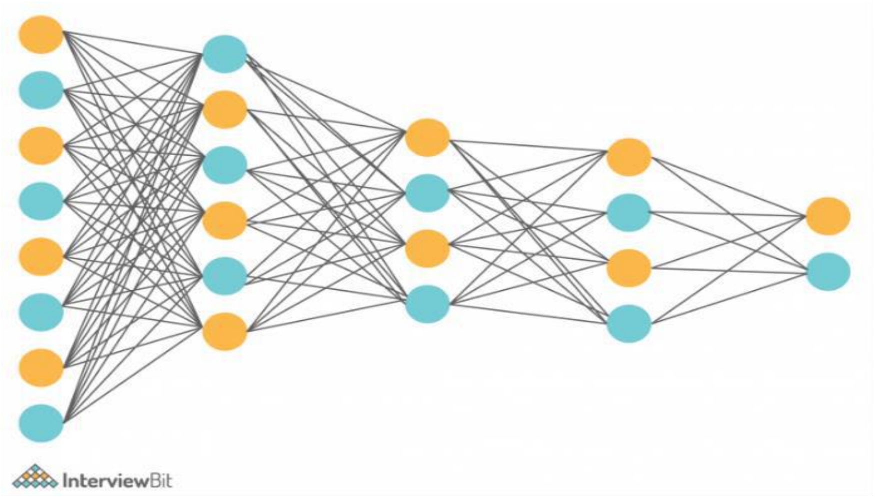


In addition to convolution, convolutional layers have another important component called a non-linear activation function. The outputs of linear operations such as convolution are routed through a nonlinear activation function. Although previously smooth nonlinear functions such as sigmoid or hyperbolic tangent function (tanh) were used because they are mathematical representations of the behavior of biological neurons. The rectified linear unit (ReLU) is the most commonly used nonlinear activation function today. f(x)=max(0,x).

• **Connection** Layer **(HALF):** This layer is responsible for dimensional reduction. This helps to reduce the computer power required to process the data. Aggregation can be divided into two types: maximum aggregation and average aggregation. The max-split function returns the maximum value of the area covered by the image kernel. The average value of all the values in the covered image part is returned by averaging.



• Fully connected (FC) layer: The fully connected (FC) layer works with flattened input, which means that every input is connected to every neuron. The flattened vector is then sent through some additional FC layers where **math** operations are usually performed. From this moment the classification process begins. FC layers are often found at the end of CNN architectures, if present.



Along with the above layers, there are some additional terms that are part of a CNN architecture.

- ✓ ⎞ **Pitch: Pitch** refers to the number of pixels by which the filter is moved during the convolution **function. Greater spacing** results in smaller printable feature maps and faster computation, but at the cost of lower spatial resolution and **potential** loss of important features.
- ✓ ⎞ Padding: Padding refers to the process of adding zeros around the input image or feature map before convolution to ensure that the output has the same spatial dimensions as the input. Padding helps prevent important features from **disappearing from** the edges of the image.
- ✓ ⎞ Activation function: An activation function is applied to the output of each CNN neuron to introduce nonlinearity into the network. Common activation functions include ReLU, sigmoid, and tanh.
- ✓ ⎞ **Deletion: Deletion** is **a regularization** technique that randomly drops a certain percentage of neurons in a network during training. This helps avoid overfitting and improve **generality.**

- ❖ ⎞ Learning: **Learning** determines the step size in each iteration of the optimization process during training. A higher **learning** rate can speed up **convergence,** but also increase instability and optimization **minima.**
- ❖ ⎞ Set size: **The set** size refers to the number of samples used in each training iteration. A larger set can improve the stability of the optimization process and **allow** faster computation, but it can also require more memory and cause **redundancy.**
- ❖ ⎞ Optimizer: Optimizer is used to update CNN weights during training. Common optimizers include stochastic gradient descent (SGD), Adam, and RMSprop.
- ❖ ⎞ Transfer Learning: Transfer learning **means** using a pre-trained CNN to initialize the weights of a new CNN. This can save time and improve performance, especially when limited training data is available for a new task.

4. Object Detection with CNNs:

Convolutional Neural Networks (CNN) are used for object detection tasks to identify and locate objects in an image. Object detection includes both the classification of objects and the detection of their location in an image. Here is an overview of how CNNs are used for object detection:

- ○ ⎞ Input: CNN takes an input image and performs **the** necessary **preprocessing** steps such as resizing or normalization.
- ○ ⎞ Convolutional Levels: The input image goes through a series of convolutional levels to extract features important for object detection.
- ○ ⎞ **Regional** Recommender Network (RPN): RPN is a subnet **whose task is to recommend** regions of an image that are likely to contain objects. RPN generates a set of candidate regions that can contain objects by predicting the coordinates of the **interfaces** of the regions of interest.

-) Non-maximum damping: The candidate bounding boxes generated by RPN may overlap, so a non-maximum damping algorithm is used to select the most accurate and non-overlapping bounding boxes.
-) Classification and **localization:** The **final** result of an object recognition network includes both classification and localization. The classification layer predicts the object class of each selected bounding box, while the localization layer predicts the exact coordinates of each bounding **box object.**
-) Training: In the training phase, the object recognition network is presented with a set of labeled training images and the network weights are iteratively updated using an optimization algorithm to minimize the difference between the predicted output and the actual label. The loss function used for object detection contains both classification and localization terms.

5. Popular architectures such as YOLO and Faster R-CNN

- ❖ YOLO is a one-stage target detector known for its speed and accuracy. The YOLO architecture consists of **several** convolutional layers followed by a fully connected layer that creates a grid of bounding boxes and class probabilities for the input image. YOLO processes the entire image **simultaneously** and **simultaneously** predicts **the boundaries** and class probabilities **of** all objects in the **image.** This makes YOLO very fast and efficient, but can sometimes result in lower accuracy compared to **two-phase** detectors like Faster R-CNN. •
- ❖ The faster R-CNN, on the other hand, is a two-stage object detector known for its accuracy and flexibility. The Faster R-CNN **architecture** consists of two main components: a region proposal network (RPN) and a detection network. **The** RPN **creates** a set of candidate regions in the input image that are likely to contain objects, while **the** expression network performs classification and localization for each candidate region separately. Both the RPN and the detection network are implemented using CNNs, and the architecture can be trained **anywhere** to simultaneously optimize both the region proposal and target detection tasks.

6. Transfer Learning with CNNs :

Pretrained CNNs can be used for transfer learning, which uses a pretrained model on a large dataset as a starting point for a new task or dataset. This can be particularly useful when limited training data is available for a new task or dataset, as a pre-trained model can provide a good initialization of the weights and help improve **generalizations.**

There are two main **ways** to **use pre-trained** CNNs in transfer learning: fine-tuning and feature extraction.

) **Finally,** tuning means taking a pre-trained model and retraining it with a new **dataset.** This usually means replacing the last layers of the pre-trained model with new fully connected layers and then training the entire model with the new data set. The weights of the pretrained model are usually frozen at the start of training and only the weights of the new layers are updated. Frozen layers can also be fine-tuned as **the** training progresses, allowing them to refresh their **difficulty.** Fine-tuning can be useful if the new data set is similar to the original data set on which the pretrained model was trained.

) Feature extraction involves using a pre-trained model as a fixed feature extractor, where the features learned **from** the pre-trained model are extracted and then used as input to a new classifier or regressor. This usually means removing the final fully connected layers from the pretrained model and using the output of the other layers as features for the new task. A new classifier or regressor is then trained using these features as input. **The function** can be useful when the new data set is significantly different from the original data set on which the pre-trained model was trained. Both fine-tuning and feature extraction can be used with several pre-trained CNN architectures such as VGG, ResNet and Inception. Transfer learning using pretrained CNNs has been shown to be effective **in** many computer vision tasks, including image classification, object **recognition,** and semantic segmentation.

7. ADVANTAGES OF CNN:

- ✓) Efficient image processing - One of the main advantages of CNNs is their ability to process images efficiently. This is because they use a technique called convolution, where a filter is applied to the image to extract features relevant to the task. This allows CNN to reduce the amount of data to process, making **it** faster and more efficient than other types of algorithms.
- ✓) High **Accuracy -** Another advantage of CNNs is their ability to achieve high accuracy. This is because by analyzing large data sets, they can learn to recognize complex patterns in images. This means they can be trained

to recognize specific objects or features with high accuracy, making them ideal for tasks such as face or object recognition.

- ✓) Noise robust – CNNs are also robust to noise, meaning they can still recognize patterns in images even if they are distorted or corrupted. This is because they use multiple layers of filters to extract features from images, making them more robust to noise than other types of algorithms.
- ✓) Transfer Learning **-** CNNs also support transfer learning, **which means** they can be trained for one task and then perform another task with little or no additional training. This is because the extracted features of CNNs are often general enough to be used for a variety of tasks, making them a versatile tool for many different applications.
- ✓) **Automatic** Feature Extraction **-** Finally, CNNs automate the **process of** feature **extraction, which means** they can learn to recognize patterns in images without manually designing features. This makes them ideal for tasks where the features **relevant to the task** are not known **a priori,** as the CNN can learn to identify the relevant features through training.

8. DISADVANTAGES OF CNN :

- ✓) High computational requirements: **The** high computational requirements **of CNNs is** one of their main **disadvantages.** This is because **CNNs require** a lot of memory and processing power to train and **run, as they usually** have **many** layers and parameters. **Therefore,** they **may** not be **usable** in some situations where resources are limited.
- ✓) Difficulty working with small **data sets:** CNNs also need a lot of data to **work accurately.** This is **because** they practice recognizing patterns in images by looking at **many** examples of those patterns. **If** the dataset is too small, **the CNN may overfit, meaning** it becomes too specialized for the training dataset and **performs** poorly with new data.
- ✓) **In addition,** large datasets are required for high accuracy **of** CNNs. This is **because** they figure out how to **identify patterns** in **images** by **studying multiple]cases** of **these** examples. **If** the dataset is too small, **the CNN may overfit, meaning** it becomes too specialized for the training dataset and **performs** poorly with new data. **The** lack of interpretability **of CNN** is another **disadvantage. That said, it's** hard to **understand** how CNN **decides** on its **picks.** This can be a problem in applications where **it is important to know** why a certain decision was **made.**
- ✓) Vulnerability to adversarial attacks: CNNs are also **vulnerable** to adversarial attacks, which involve manipulating input data **to trick** the network into making **bad** decisions. In applications **such as** autonomous vehicles, where safety is **paramount,** this **can be** a **major problem.**

9. FUTURE SCOPES OF CNN :

- ✓) Continuous **learning:** Continuous learning **means** learning new information without forgetting **the** previously learned information. CNNs can struggle with continuous learning because they are prone to catastrophic forgetting when learning new data. An important future direction is the development of CNN architectures that can learn gradually without forgetting previously learned information.
- ✓) Explainability: CNNs are often considered black-box models, making **interpretation of decision making difficult.** Developing methods to explain the decision-making process of CNNs can help increase confidence in these models, increase their adoption, and provide insight into their inner workings. Weakly Supervised Learning: In many cases, **extracting** pixel-level annotations from training data can be difficult, making weakly supervised learning an important future direction. CNN architectures that can learn from weak annotations, such as image-level or **bounding box-level** annotations, can help reduce annotation costs and increase the applicability of CNNs.
- ✓) Self-supervised learning: Self-supervised learning involves training a CNN **on sample tasks** using unsupervised data to learn useful representations. Self-directed learning has shown promising results and may be useful in areas where labeled information is scarce. **An important future direction is the** development of CNN architectures **using** self-directed **learning.**

- ✓ Multi-task learning: Multi-task learning involves training a CNN to perform multiple related tasks simultaneously using shared representations. Multi-task learning can reduce the number of parameters and training time and lead to more reliable models. The development of **multi-task learning** CNN architectures is an important future direction.
- ✓ Memory efficient architectures: CNNs can be computationally expensive and require a lot of memory. Developing more memory-efficient CNN architectures can help reduce computational requirements and improve their scalability.

Applications of CNN in Real World :

- **Object detection:** Object detection **tasks that aim** to identify and **find** objects **in** an **image often use** CNNs. Robotics, security **systems** and self-driving cars are all possible **applications.**
- **Image Segmentation:** Image segmentation **uses** CNNs and aims to divide an image into **several** regions or segments based on **pixel properties.** Video processing, satellite **image analysis** and medical imaging are all examples of applications.
- **Image Classification:** Image classification, **where image** content is used to **determine** a label or **class,** is a common application for CNNs. Image **tags,** content-based image **search** and **face** recognition are all examples of applications.
- **Video analytics: Activity detection, where** the **goal** is to classify the **activities** performed in a video **series,** and video object tracking, **where** the **goal** is to track objects as they move through **the video. Both** use CNNs for video analysis tasks.
- Natural **language processing: natural** language processing tasks **such as** sentiment analysis, which **aims** to classify **text** as positive, **negative** or neutral, and text classification, which **aims** to classify text into **different** categories or **topics. ,** both use CNNs.
- **Speech-to-speech: CNNs can perform tasks such as** recognizing spoken words or **sentences** and converting them **to text.** Applications **include trivial utilities, storage management,** and **instant messaging** programming

10. RESULT:

This article provides a comprehensive overview of convolutional neural networks (CNNs) and their application to computer vision tasks, particularly object detection. It explains the various components of CNNs, such as convolutional layers, pooling layers, and fully connected layers, along with related concepts such as pitch, padding, activation function, deletion, learning rate, set size, optimization, and transfer learning. We also discuss the advantages and disadvantages of CNNs and their future potential for continuous learning, explainability, weakly supervised learning, self-directed learning, multi-task learning, and memory-efficient architectures. The advantages of CNN networks are efficient image processing, high accuracy and robustness against noise. They also support transfer learning and automate the feature extraction process. However, CNNs require a lot of data to be accurate and can overestimate if the dataset is too small. They are also vulnerable to adversarial attacks and have high computing requirements. The article also discusses two popular object recognition architectures, YOLO and Faster R-CNN, and explains how CNNs are used for object recognition tasks. This highlights the importance of CNNs in advancing computer vision applications and identifies future research and development areas. Overall, the paper provides a valuable resource for understanding CNNs and their potential for developing computer vision applications.

11. CONCLUSION:

In summary, convolutional neural networks (CNN) have become increasingly popular in recent years due to their ability to efficiently process images and achieve high accuracy. The paper provides a comprehensive overview of CNNs and their application to computer vision tasks, especially object detection. The article explains the various components of a CNN, such as convolutional layers, pooling layers, and fully connected layers, along with related concepts such as pitch, padding, activation function, deletion, learning rate, set size, optimization, and transfer learning. We also discuss the advantages and disadvantages of CNNs and their future potential for continuous learning, explainability, weakly supervised learning, self-directed learning, multi-task learning, and memory-efficient architectures. The paper highlights the importance of CNNs in advancing computer vision applications and identifies areas for future research and development. Overall, the paper provides a valuable resource for understanding CNNs and their potential for developing computer vision applications.

REFERENCES:

1. K. Simonyan and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." arXiv preprint arXiv:1409.1556 (2014).
2. Saad Albawi; Tareq Abed Mohammed; Saad Al-Zawi “Understanding of a convolutional neural network” IEEE
3. Sridharan Balasubramanian and Srinivasan Narasimhan ; "A Comparative Study of Convolutional Neural Network Architectures for Object Recognition in Images" by. Published in Journal of Computer Science and Technology in 2020.
4. https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939
5. https://aspiringyouths.com/advantages-disadvantages/convolutional-neural-network-cnn/
6. https://www.interviewbit.com/blog/cnn-architecture/
7. https://www.ibm.com/in-en/topics/convolutional-neural-networks