International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 06 | June 2021 www.irjet.net p-ISSN: 2395-0072

Survey on Deep Learning based Computer Vision Applications

Ms. Yugandhara A. Thakare [1], Mr. Anup W. Burange[2], Ms. Ankita V. Pande[3] Ms. Ashwini P. Ghatol[4]

[1]*Assistant professor, CSE Dept, Sipna C.O.E.T, Amravati*
[2]*Assistant professor, IT Dept, PRMIT&R, Amravati*
[3]*Assistant professor, CSE Dept, Sipna C.O.E.T, Amravati*
[4]*Assistant professor, CSE Dept, Sipna C.O.E.T, Amravati*

Abstract - A computer vision is the field where 3D scene can be reconstructed or interpreted by using basic 2D images. The field of Computer vision has been changing rapidly with the consistent growth in powerful technology like deep learning along with neural networks which can extract many required information from images directly. With the advent of deep learning in computer vision applications like face recognition system, self-driving cars, image captioning etc. making rapid progress within a very short span. Advancements in machine learning and deep learning has made computer vision technology more accurate and reliable too. With the advanced deep learning algorithms, computer vision has been highly effective in real world scenarios. The use of convolutional neural network in computer vision has made it suitable for many industrial applications and made it as a reliable technology to trust as investment for companies which are looking to automate their works and tasks

Key Words: **Image captioning, Segmentation, Object detection.**

1.INTRODUCTION

Humans have the inbuilt vision capability which uses experimental knowledge which it gains from day to day activities. This knowledge help them to contextualize or visualize the data within the view field. Human's eyeballs captures the visual parameters for e.g.. image of dog and the prior knowledge of that visual parameters about that image or similar to that image relates it to the dog. This ability is due to our very powerful visual perception system which is closely related to our mind and memory. Mind provides higher reasoning ability to convert this visual data into meaningful context through the experience of day to day activities. These powerful human abilities are not available to machines but can be imitate through machine learning and deep learning algorithms. To imitate such expertise to machines is difficult task and lot of work is going on in this field of research to make it convenient.

Earlier computer vision techniques were dependent on substantial manual work to create rule based classification techniques which were able to predict and classify certain groups of pixel arrangement. For e.g. to detect the image of dog, programmer codified every component of dog into computer as fixed rules so that it could be able to detect these features in an image. In past decades this field of computer vision depends on this troublesome, manually-created feature detectors for sorting and classifying an image. These procedures were not flexible and was very difficult for making change into it also it used to take lot of time for each new object of detection. This model tends to fail when the number of classes needed to classify increases or when the quality of image degrades. Simple changes in size of object, rotating direction would cause to system to get halt or stop. Advances in machine learning and deep learning gives the new directions and way to this computer vision field. Today's deep neural network can train the process which uses very large dataset by countless training cycles which can teach the machine entirely how actually dog looks. In training phase algorithm can automatically extracts the appropriate features of 'dog' to predict. This process build a model which can be imposed to previously unseen images to generate more accurate classification and prediction. The below image illustrates the conflict between traditional machine learning process for image detection & recognition compared to deep learning approach.

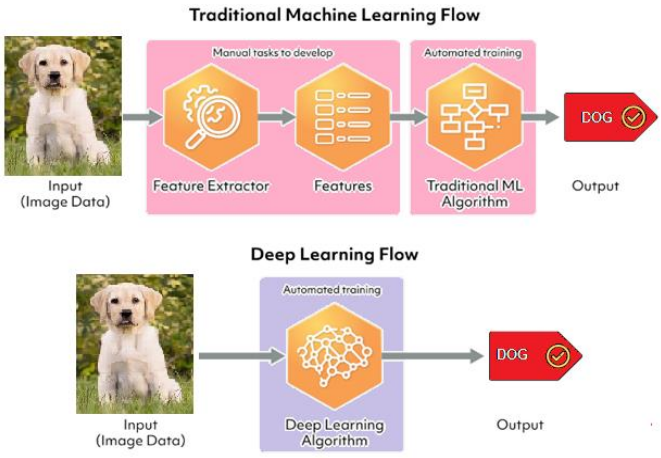

Fig -1: Deep Learning for Computer Vision

Deep learning is basically a subset of machine learning technique which consists of quite large neural network architectures. A perceptron which generally known as artificial neuron is basically a computational node which makes many inputs and creates a weighted summation to generate an output. A perceptron can be viewd as linear mapping between input and the output.

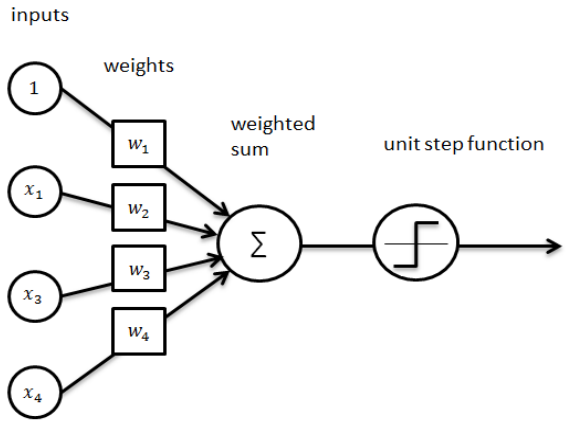

Fig -2: Neural network activation functions

1.1 Activation functions

These are generally known as mathematical functions that can predict the range of output values of a perceptron.

Different types of Activation Functions:

Sigmoid: Sigmoid is a curved step function and thus can be differentiable. It is useful in the field of binary classification and conditions where there is a need to convert any value to probability. It confines the value of a perceptron to [0,1], which isn't symmetric.

tanh: It is hyperbolic tangent function, which is also called as the tanh function, confines the output between [-1,1] and thus symmetry is preserved. A significant point to consider here is that symmetry is a advantageous property during the propagation of weights.

The Rectified Linear Unit (ReLU): It is defined as a function y=x, which is able to produce the output of a perceptron, irrespective to what supplied through it. When the output value given is negative, then it maps the output to 0. Hence it is defined as a function max(0, x), where x denotes the output of the perceptron.

1.2 Various Types of Neural Networks

A. Artificial Neural Network (ANN)

ANNs are the combination of perceptrons and activation functions linked together. ANN forms the non-linear basis to map between input and output by forming hidden layers of internally connected neurons. The dimensionality of mapping depends on the number of hidden layers in neural networks. The dimension in which the output is being mapped will be higher if number of layers are high. Difficulty of using ANN in computer vision.

Difficulty of using ANN in computer vision.

Neurons within the same layer don't share connections and ANN deals with entirely connected layers which if used with images may cause over fitting and it may also results in larger size because of large number of neurons. One solution can be to increase the model size so as to increase the requirement of neurons, we need to select an image with many proportions of height, width and depth.

B. Convolutional Neural Network: It is a deep learning algorithm which can take input as an image and can allot weights and biases to different objects in an image to distinguish it from other. In contrast to other classification algorithms the requirement of pre-processing in Convolutional Neural Network is much minimum. Filters which manually designed to be used in primitives methods, while CNN has the ability to learn these filters. The structure of CNN is inspired by the association of visual cortex and analogous to the connectivity model of neurons in human brain [2].

Advantages

- The use of CNNs are inspired by the fact that they can capture or can easily able to learn the relavent features from an image/video at different stages which is equivalent to human brain. This is called as feature learning and conventional CNNs are not able to do this.

- Weight sharing is another important feature of CNNs. CNNs are more efficient in terms of memory and complexity. In comparisons with other neural networks CNN would be less complex and can save memory compared to neural network.

C. *Recurrent neural network (RNN):* RNN is one of the type of artificial neural network which makes use of time series data or sequential data. Such kind of deep learning algorithms are mostly used in kind of ordinal or temporal problems like speech recognition,image captioning, language translation, natural language processing etc [1].

Advantages of using RNN.

- CNNs are not at its best for modeling sequential information, so the this problem can be solved with a network that models the sequential patterns.
- RNNs has the ability to solve that problem, RNN executes the similar tasks by adding a feedback element, which takes the output of the previous data in a series to its next input.

- This feature helps the network to learn the important correlation between the current data point and the previous data point.

D. *Long short-term memory (LSTM):* Long short-term memory (LSTM) has the ability to store a huge number of data points for larger span of time, and that is why, it works well enough with capturing long-term efficiencies [3].

2. APPLICATIONS OF COMPUTER VISION

A. Classification: To which class an particular image belongs can be identified by image classification. It will help in getting better result and confidence of task performed if we classified the image which belongs to certain class.

B. Object Detection/Localization: If we want to know which part of the image, and it belongs to which particular class, only the classification is not good enough. Object detection- is a process that can detect a very important object in a picture and give it to the same image-wrapper. This feature is useful in applications related to knowledge of the relative distances and directions to objects. For example, in autonomous vehicles, where pedestrians or other objects have been located, it will help you to make better decisions. The Intersection over Union-level indicator is used for the calculation of the efficiency of object detection algorithms.

C. Segmentation: Approach to the problem is different by segmentation, it consider pixel-wise classification. This provides information about the finer details like the boundaries of objects. These details can be helpful for building other applications.

D. Image caption: An image caption in the process of the creation of the captions on the image. It makes use of object classification, models, and LSTMs to produce the signature. The LSTMs caption, the model will help you to find out of clear information, an image and its corresponding information. This application makes use of both computer vision and natural language processing to create signatures .The image captioning is used to gain an understanding of the relative position of an object in an image with other objects in this picture [3].

E. Generative models: Generative models works with training networks to generate images based on their perceptive or learning. During the training phase, for example, if the dataset is about cars, the generative model learns about the features that make up the car. So, after learning the features of the car, it can create new cars afterwards. An exciting application of the same is the generation of cars with different features. The application for the same is in the field of training deep learning models. Since these models are data greedy, collecting huge amounts of data is a tiresome task. As an alternative, generating data according to our requirements is a superior option to train the model accordingly.

F. Video analysis: Video analysis-refers to the provision of real-time results and for the creation of models, the conclusions and results of the performance of the application. For example, you will learn about sports events such as cricket, soccer, football, etc., etc. can be automated. Real-time stats, player positions, and the best way to beat the strongest team, these are just a few examples.

3. LITERATURE SURVEY

A. Feature Extraction:

Bin Jiang, et al. [4] proposed a deep-learning based real-time, cross-media, search method. In this method, the processes the previous evaluation of the pattern of samples for the determination of the need for recognition of the algorithm. For large databases, analysis, it may be useful to determine the failure rate of an image, which can be useful for machine learning. The obtained results are promising and show that the analysis is accomplished with a high degree of accuracy, the cross-field image of the text, using a shorter amount of time to search.

Voshihiro Hayakawa et al. [5] carried out an experiment in which a constant number of characters is derived from the scripture, sign, image, with the help of a multi-layered NN. The author has also explored the possibility of an effective use of the physical symptoms, such as a timer with the POWER. After the analysis of a series of repeated movements of the body, the experimental results showed that there were no signs of improvement if a Gaussian filter is applied to the overlap of the training data, as a kind of pre-processing.

B. Human activity recognition

Human activity recognition has become a lively research topic in the field of computer vision. Thanks to the application, and its use in a wide range of applications in areas such as robot learning, human-computer interaction, intelligent control, and a medical diagnosis. The detection of the human activities in videos is one of the most challenging tasks in computer vision, mainly due to the inflexible real-world movement of the scenarios, and the large amount of data, which will need to be incorporated in.

Bagautdinov et al. [6] in their work they took care of the order, with the help of a recurrent neural network that corresponds to the human level. The proposed architecture is to learn from the beginning to the end to make it rich, a suggestion that the card is fixed with a new inference rules to identify more people who are distracted by their shared activities together.

Zhu et al. [7] proposed a method for the identification of the actions by the addition of a mixed-norm of the control function of a deep LSTM center network. One of the most popular deep learning methods are used for the processing of your pictures/images is convolutional

neural network (CNN). There are a variety of works that use of a 2D-CNN, which is to make use of the spatial correlation among video frames, and then connect the outputs with the help of different strategies.

C. Image Captioning

Mao et al. [8] suggested that it is a word that is listed in this picture and the last to be released of the words use the RNN language model. In their structure, they have used deep convolutional network to extract image features, and the model as a result of the words, with the same features, images, and the soft words, they are used in RNN, along with the multimodal part. The RNN language model is mainly composed of a tray, the input layer to layer, in addition to a repeat of the layer.

Socher et al. [9]. A method is proposed, in which one of the sentences or phrases in the form of a compound of vectors to represent himself as a recursive neural network dependency tree, which is used to search for a name. The acquired multi-attribute, in the shared space, using a max-margin. With the use of deep neural networks, the performance of the subtitle in view of the methods is then enhanced by improving and customizing new products.

The author, Tong Li et al. [10] have a proposed a new method, called the sentimental picture, captioning, which is able to generate a signature with a built-in feeling that is reflected in image. In comparison with the digital duties of subtitles, the images, which requires a pre-defined styles to your picture to be their new method for the automatic analysis of the risks that may be associated with an image from the original. Proposed to be an Integral Ideal, the Caption on the image "(InSenti-Known), extracts the contents, and the information about the object out of the picture. This method adds information about the content and the mood to the formation of the ideal of the sentences by using the attention mechanism. A two-step strategy was proposed, which consists of sensitive questions, and emotional feedback, to allow for the ability to easily model is created and the appropriate suggestions, with built-in sentence style. In their model, it is a tough one to understand the content and the mood of the image, and at the same time, in a caption to make sense of the image.

At present, the encoder-decoder needs to be used, and the structure of the self-monitoring mechanism for labels and pictures. It is used to enhance the appearance of the features in the image to the encoder, as well as to capture the most relevant information in the language box. However, the existing methods often indicate the weights for all of the candidate vectors, which, in turn, means that all of the vectors have been targeted on the basis of its content. The modern workings of consciousness, not to pay attention to the internal object, the distribution of the attention, but it will only recognize the inter-object interactions.

The author's Weitao Jiang et al. [11] offer a Multi-Gate of Attention (MGA) unit, which extends the traditional with the same attention from the rest with an optional Attention to the Weight of a Gate (AWG) module is a Self-Gated (SG) module. The components of the first reduction of the weight of the notes that are to be assigned to the active object. More recently it is taken into account within the object, the distribution of the attention-and the solution-relevant information contained in the vector of characteristics of the object. It has also been suggested that, in the areas of pre-and in-the layers, in order to make it a simple transformer to the architecture, and the right to increase the image parameters. It is integrated with a MN-a unit with a pre-low-standard transformer architecture is in the image encoder and the AWG-language-decoder module is to present an advanced Multi-gateway-Attention Network (MGAN). They have conducted experiments on the MS COCO dataset, which shows that the MGAN is far superior to the majority of contemporary practice.

A visual dialogue task that consists of several rounds of dialogue with a wide range of visual content, which could be related to issues, relationships, or the high-level semantics. The major challenges in the visible Window, The aim is to explore a more complex and semantically rich image that is adaptive to participate in the visual content in different sorts of questions.

Jing-Yu, et al. [12] proposed a new method of image output, in both the visual and semantic perspectives, is presented. A visual representation allows you to capture, view the information at the photo by including objects and their visual relationships, as a semantic concept that makes it possible to use the agent-to-understand, high-level visual semantics of the entire image to a local regions. Dual encoder module, as well as Visual Dialogue (DualVD), has also been proposed, which is able to custom make the choice of the correct information, visual, and semantic representations in a hierarchy of modes. In order to demonstrate the effectiveness of the DualVD, they are offering two new visual discourse of the models Late Fusion framework and The Memory Network. In the proposed models, this will allow you to get the most reliable results on three standard data sets.

4. CONCLUSION

In this paper we addressed the different applications of computer vision and we also mentioned how deep learning is making computer vision applications better in terms of efficiency and usage. Deep learning algorithms that we mentioned like CNN, RNN and LSTM has made many improvements in the working of these applications. Image captioning is one the promising application of computer vision which has used deep learning to get meaningful insights but still there are many fields in which

deep learning can be beneficial for getting intelligent insights.

REFERENCES

[1] Jianqiong Xiao; Zhiyong Zhou, "Research progress of RNN language model" IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA) 2020.

[2] Hideaki Yanagisawa; Takuro Yamashita; Hiroshi Watanabe, "A study on object detection method from manga images using CNN." International Workshop on Advanced Image Technology (IWAIT) 2018.

[3] Minsi Wang; Li Song; Xiaokang Yang; Chuanfei Luo, "A Parallel-fusion RNN- LSTM architecture for image caption generation." IEEE International Conference on Image Processing (ICIP) 2016.

[4] B. Jiang, J. Yang, Z. Lv, K. Tian, Q. Meng, and Y. Yan, "Internet cross-media retrieval based on deep learning," Journal of Visual Communication and Image Representation, vol. 48, pp. 356–366, 2017.

[5]Y. Hayakawa, T. Oonuma, H. Kobayashi, A. Takahashi, S. Chiba, and N. M. Fujiki, "Feature extraction of video using deep neural network," in Cognitive Informatics & Cognitive Computing (ICCI* CC), 2016 IEEE 15th International Conference on. IEEE, 2016, pp. 465–470.

[6] Timur Bagautdinov, Alexandre Alahi, Franc¸ois Fleuret, Pascal Fua, Silvio Savarese, "End-to-End Multi-Person Action Localization and Collective Activity Recognition",arXiv:1611.09078v1 [cs.CV](2016).

[7] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks," in Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[8] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). In International Conference on Learning Representations (ICLR)).

[9] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, A. Y. Ng, Grounded compositional semantics for finding and describing images with sentences, TACL 2 (2014) 207–218.

[10] T. Li, Y. Hu and X. Wu, "Image Captioning with Inherent Sentiment," 2021 IEEE International Conference on Multimedia and Expo (ICME), 2021, pp. 1-6.

[11] W. Jiang, X. Li, H. Hu, Q. Lu and B. Liu, "Multi-Gate Attention Network for Image Captioning," in *IEEE Access*, vol. 9, pp. 69700-69709, 2021.

[12] J. Yu, X. Jiang, Z. Qin, W. Zhang, Y. Hu and Q. Wu, "Learning Dual Encoding Model for Adaptive Visual Understanding in Visual Dialogue," in IEEE Transactions on Image Processing, vol. 30, pp. 220-233, 2021.