International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

VIDEO CAPTIONING AND GROUNDED ON PICTURE CAPTIONING TECHNIQUES USING DEEP LEARNING

[1]Kuna Naresh , , [2]Thatikonda Radhika, [3]Thakur Madhumathi, [4]Bhutam Shivani

[1] Assistant Professor, Department of CSE, Tkr college of engineering and Technology, Hyderabad
[1]Kuna48@gmail.com

[2]Assistant Professor, Department of IT, Tkr college of engineering and Technology, Hyderabad
[2]radhikathatikonda08@gmail.com

[3]Assistant Professor, Department of IT, Tkr college of engineering and Technology, Hyderabad
[3]madhu.thakur26@gmail.com

[4]Assistant Professor, Department of IT, Tkr college of engineering and Technology, Hyderabad
[4]bhutam.shivani@gmail.com

ABSTRACT:

Applications that seek to automatically generate captions or explanations for images and video frames have a lot to gain from using Deep Learning-based methodologies. Captioning images and videos is well recognized as a challenging academic subject within the field of image analysis. Automatic caption (or portrayal) age for pictures and recordings for individuals with fluctuating levels of visual hindrance; programmed production of metadata for pictures and recordings (ordering) for use via web indexes; broadly useful robot vision frameworks; and numerous others are instances of conceivable application spaces. Every one of these utilization cases can possibly essentially and emphatically influence a great many different applications. This isn't intended to be a thorough look of picture subtitling; rather, it is a concise outline of a few profound learning-based techniques for doing both picture and video inscribing. This study centers around the algorithmic similitudes among picture and video inscribing to take care of the two issues.

Keywords: — Deep learning, image captioning, video captioning, long short term memory, generative adversarial network

1. Introduction

A significant role for image processing in research and development is guaranteed to continue. Its utilization spread to numerous different fields, for example, picture acknowledgment [1] and scene understanding [2]. Before Deep Learning was grown, most examinations depended on imaging techniques that must be involved on unbending items in a controlled lab climate with particular equipment [3-12]. Ongoing years have seen a sensational improvement in the field of picture subtitling thanks to the presentation of profound learning-based convolutional brain organizations. We need to introduce a brief outline of current advances in profound learning-based picture and video subtitling here. Numerous gatherings of scholastics have been cooperating on profound learning model improvement [13], execution [14], and understanding [15] beginning around 2012. While the hypothesis and strategies for profound learning have been available for quite a long time, late advances have been rushed by the accessibility of a lot of computerized information and the utilization of strong GPUs. TensorFlow and PyTorch are just two instances of accommodating programming improvement libraries; the open-source local area; a few major

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

marked datasets (e.g., MSCOCO, Flicker, ACoS, LSMDC); [15], [16]; and stunning exhibits generally copy and copy the dramatic development of the profound learning area.

When asked to explain what they see in a still picture or a brief video clip, humans have a very difficult time doing so. Scientists in the field of computing have been looking at methods that combine the study of human language comprehension with the research of autonomously extracting and analysing visual information in order to build machines with this kind of capability. Picture and video subtitling are more work serious than picture acknowledgment as a result of the additional test of recognizing things and exercises in the picture and producing a concise, important expression in view of the data identified. The headway of this strategy has broad ramifications for a wide assortment of certifiable application spaces, for example, however not restricted to, assisting individuals with differing levels of visual impedance, independent vehicles, communication through signing interpretation, human-robot collaboration, programmed video captioning, video observation, and numerous others. This article presents a study of the cutting edge methods for subtitling photographs and recordings, with an accentuation on profound learning models.To evaluate the efficacy of the models and the produced captions, researchers utilise a variety of metrics [17–19].

2. LiteratureSurvey

[1] J. Redmon and A. Farhadi, ‘‘YOLOv3: An incremental improvement,’’ 2018, arXiv:1804.02767. [Online]. Available: http://arxiv.org/ abs/1804.02767

As promised, here are the new and improved YOLO! We improved the overall quality by making several minor adjustments to the design. In addition, we trained a whole new network that performs admirably. The size is little larger than before, but the accuracy has improved. Don't worry however, it's still lightning quick. YOLOv3 is three times as quick as SSD at a resolution of 320 320 and achieves the same level of accuracy. For mAP detection, YOLOv3 works well when compared to the standard.5 IOU. Scores 57.9 AP50 in 51 ms on a Titan X, which is similar to RetinaNet's 57.5 AP50 in 198 ms but is 3.8 times faster. As always, you can get the code at https://pjreddie.com/yolo/.

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, ‘‘The cityscapes dataset for semantic urban scene understanding,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 3213–3223.

The ability to comprehend intricate city street sceneries visually is useful in many contexts. Object distinguishing proof has benefited extraordinarily from huge scope datasets, particularly with regards to profound learning. In any case, no current dataset satisfactorily catches the intricacy of genuine metropolitan conditions for semantic translation of metropolitan scenes.

Cityscapes is a benchmark suite and enormous scope dataset proposed by the creators to prepare and assess strategies for semantic characterization at both the pixel and

occasion levels.The stereo video segments that make up Cityscapes were shot on the streets of fifty different cities. 5000 of these photos have fine-grained comments at the pixel level, while the other 20,000 pictures have coarse explanations that will be utilized by calculations that influence a lot of ineffectively marked information. Our concentrate altogether outperforms earlier endeavors as far as dataset size, explanation detail, scene variety, and occupation intricacy.In addition to providing a thorough review of the benchmark, our supporting empirical research also includes a detailed examination of the dataset's features.

[3] H. R. Arabnia and M. A. Oliver, ‘‘Fast operations on raster images with SIMD machine architectures,’’ in Computer Graphics Forum, vol. 5, Hoboken, NJ, USA: Wiley, 1986, pp. 179–188, doi: 10.1111/j.1467- 8659.1986.tb00296.x.

In this paper, we describe a raster image data format and accompanying algorithms for efficiently processing such data. All of the procedures and data structures have been fine-tuned to work best with SIMD parallel CPUs. Generating simple shapes like lines, discs, and circles is straightforward. It is possible to merge, resize, and translate images. Similarities between this data format and runlength encoding may be seen in both their designs. An ICL DAP has been modified to run the algorithms and do the necessary tasks in real time (on non-trivial images).

[4] S. M. Ehandarkar and H. R. Arabnia, ‘‘Parallel computer vision on a reconfigurable multiprocessor network,’’ IEEE Trans. Parallel Distrib. Syst., vol. 8, no. 3, pp. 292–309, Mar. 1997.

Here, we provide a novel reconfigurable architecture that makes advantage of a multiring multiprocessor network. Having every hub in the organization have a low level of availability and a little organization width is an immediate impact of the plan's adaptability.In this research, we describe the hardware of the reconfiguration switch and the mathematical aspects of the network topology. The most principal equal procedure on an organization's geography are characterized and investigated. It is shown that the plan works for an extensive variety of 2D cross section geographies, incorporating those with simply a solitary component of the Boolean hypercube. It is shown that a huge gathering of calculations for both the 2D cross section and the Boolean n-3D shape decipher well onto the proposed engineering without causing critical execution punishments. It is shown that the idea might be utilized to a wide assortment of issues, including the quick Fourier change (FFT), edge discovery, format coordinating, and the Hough change. Consequences of timing explores different avenues regarding a computer based model are accommodated normal rudimentary and optional level vision calculations.

3. ExistingSystem

A significant role for image processing in research and development is certain to persist. Visual recognition [1] and scene comprehension [2] are only two examples of the numerous fields that may benefit from it. The majority of researchers prior to the development of Deep Learning relied on imaging techniques that were only applicable to rigid objects in a laboratory setting using specialised hardware [3–12]. Recent years

have seen a dramatic improvement in picture captioning thanks to the widespread use of deep learning-based convolutional neural networks. In this article, we want to provide a summary of current developments in deep learning-based image and video captioning. Since 2012, a large number of researchers have laboured to enhance the design [13], implementation [14], and interpretation [15] of deep learning models. The hypothesis and procedures of profound learning have been around for a really long time, however the coming of computerized information and the consideration of strong GPUs have sped up its encouraging as of late.

4. ProposedSystem

Humans have a significant challenge when asked to describe what they see in a still picture or a short video clip. Researchers in the field of artificial intelligence have been trying to figure out how to combine the study of human language understanding with automated visual information extraction and analysis. Picture and video subtitling need more improvement than picture acknowledgment due to the additional test of recognizing things and activities in the image and producing a compact, important sentence in view of the data distinguished.

Humans have a significant challenge when asked to describe what they see in a still picture or a short video clip. To construct machines that can do this, specialists in the field of software engineering have been taking a gander at ways of consolidating examination into how PCs process language with examination into how PCs can naturally remove and dissect visual data. Picture and video subtitling need more improvement than picture acknowledgment in view of the additional test of recognizing things and activities in the image and producing a concise, significant sentence in light of the data distinguished. The headway of this technique has sweeping ramifications for a wide assortment of certifiable application spaces, for example, however not restricted to, assisting individuals with changing levels of visual impedance, independent vehicles, communication via gestures interpretation, human-robot connection, programmed video captioning, video observation, and numerous others. This article presents an outline of current procedures for inscribing photographs and films, with an emphasis on profound learning models.

5.Implementation

Professor Hinton was the first to propose DL, an advanced sub-field of ML that uses many layers of representation to simplify the modelling of different types of complicated ideas and connections. Successful applications of DL include language detection, picture processing, and the pharmaceutical industry. This has inspired studies on how DL theory may be used to the intrusion detection classification issue. DL's superior performance when more data is added is the defining feature that sets it apart from more conventional ML approaches. Since DL algorithms need a lot of training data to become effective, they aren't a good fit for issues with limited amounts of data. Because of its high throughput, DL is well-suited to solving the difficulties of raising accuracy and decreasing false-positive alarm rate, which have arisen as a result of the growing amount of datasets utilised in IDS research. Moreover, when the amount of the datasets rises, the input space and the attack categorization dimension also increase in

size.Therefore, misclassification is common, leading to a rise in the false positive alert rate and having a detrimental effect on the overall performance of the system. So it's important to put in place systems that can pick and choose which characteristics to use for the categorization process. Many areas of machine learning study now centre on feature engineering (FE). FE's feature selection techniques may be broken down into three distinct types: filter models, wrapper models, and hybrid models. The filter model does not rely on any particular classifier and instead relies on the underlying structure of the data. The hybrid approach is a mix of the wrapper and filter algorithms, whereas the wrapper approach just examines the performance of the classification algorithm. Due to the computational complexity of the latter two methods, this research proposes an alternative, filter-based solution. Here are some of the main things that this project has accomplished:

- An FEU (Feature Extraction Unit) is shown. The FEU uses filter-based algorithms to produce feature subsets that are both effective and efficient.
- With the NSL-KDD dataset, we compare the performance of several classification strategies for IDS without the FEU, including k-nearest neighbour (KNN), support vector machine (SVM), decision tree (DT), random forest (RF), and naïve bayes (NB). We also look at the efficacy of combining these algorithms with the FEU.
- We present a feed-forward deep neural network (FFDNN). Using the FEU and the NSL-KDD dataset, we analyse its effectiveness. After looking at how the FEU-FFDNN stacks up against KNN, SVM, DT, RF, and ND, it becomes clear that it is the best option for intrusion detection systems. Moreover, experimental data show that the accuracy of an FFDDN classifier is directly related to the number of neurons utilised and the depth of the network.

Traditional Machine Learning Classifiersused

A. SUPPORT VECTORMACHINE:

One of the most investigated and generally utilized strategies for AI (ML) applied to Big Data is support vector machines (SVM).An application of supervised machine learning, support vector machines (SVMs) classify data. SVM is appropriate for managing both direct and non-straight issues. To sort information, support vector machines (SVMs) make a hyperplane, or numerous hyperplanes, in a high-layered space and afterward pick the hyperplanes that split information most effectively.

B. K-NEARESTNEIGHBOR:

Another machine learning technique for data classification is called K-Nearest Neighbor (KNN). KNN is characterized as follows, with its base in the typical Euclidean distance between focuses in a space: Given two focuses x and y in some space P, we might communicate d(x, y) as the distance between them.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

$$d(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

Assuming there are n total occurrences, we say that. Using the Euclidean distance between the instance x0 and the k closest samples in the training set, the KNN technique assigns x0 the label of the k most similar neighbours inside the space.

C. NAIVE BAYES:

Simple classification algorithms inspired by Bayes' Theorem are known as Naive Bayes (NB) classifiers. An NB classifier takes as input a dataset and "naively" assumes feature independence. Assume that a given instance X has n characteristics for classification, and that this instance is represented by the vector X = (x1, ..., xn). For NB to determine X's class Ck, it conducts the following:

$$p(C_k|X) = \frac{p(X|C_k)p(C_k)}{P(X)}$$

And the equation below is used to determine X's class:

$$y = \underset{k\in\{1,\ldots,K\}}{\operatorname{argmax}}\ p(C_k)\prod_{i=1}^{n} p(X_i|C_k)$$

If y is the expected label, then the expression is as follows.

D. DECISION TREE AND RANDOMFOREST:

In the fields of data mining and machine learning, the Decision Tree (DT) method is often used. To forecast the class of unknown records, the DT method takes a dataset containing labelled cases (training) and creates a predictive model in the form of a tree [14]. The three primary parts of a directed tree (DT) are the root node, the internal nodes, and the category nodes. Top-down categorization techniques lead to the best possible selection when the proper leaf node category is identified. On the other hand, a Random Forest classifier uses several DTs to classify a single dataset.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Implementing Feed Forward Deep NeuralNetwork

Complex challenges in ML and DL are often addressed by using deep neural networks (DNNs). An artificial neuron (AN) is the building block of a DNN; it is conceptually similar to a biological neuron in the human brain. An AN adds up the data it receives and sends that total on to the next stage. Each AN uses an activation function to improve learnability and approximation because to the non-linear nature of real-world issues. This may be a Sigmoid-like activation function $\sigma = 1 /1+e-t$, a Rectified Linear Unit (ReLU): $f(y) = \max(0, y)$, or an hyperbolic tangent: $\tanh(y) = 1 - e -2y/1 +e-2y$.

All of the above-mentioned activation functions are not without their limitations, and their optimum performance depends on the nature of the underlying issue. DNNs might have somewhere in the range of three secret layers to tens or many secret layers, while customary ANNs have an info layer, one to three secret layers, and a result layer, as shown in the picture beneath. How much an ANN is "profound" not set in stone by any one measurement. We will characterize a DNN as a brain network with at least two secret layers for this review. Data is sent from the info layers to the concealed levels and afterward from the secret layers to the result layers in a Feed Forward DNN.Similarly positioned neurons in the same layer are unable to interact with one another. As seen in the image below, each AN in the current layer has a direct connection to every neuron in the layer above it.

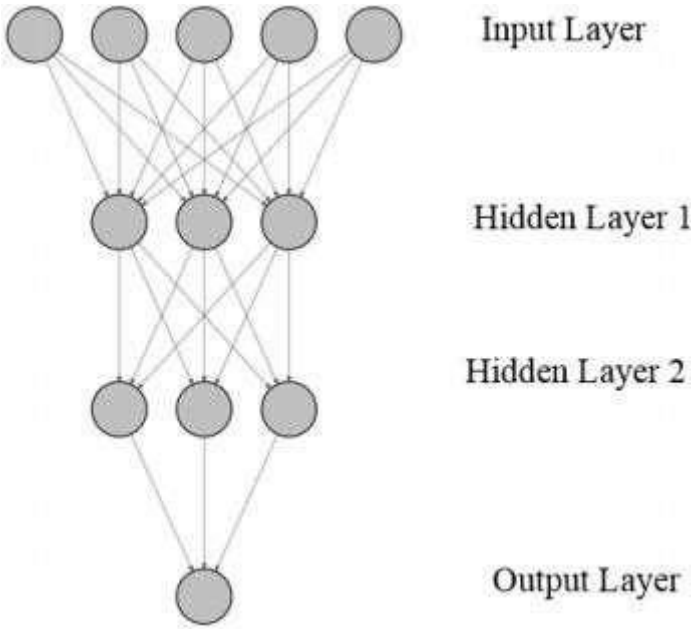

Fig 5.6: Feedforward Deep Neural Network.

Dataset used forImplementation

The NSL-Knowledge Discovery and Data mining (NSL-KDD), a revised version of the KDDCup 99, is used for the proposed system's training, analysis, and testing. The NSL-KDD is commonly considered a benchmark dataset for use in intrusion detection

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

systems across both wired and wireless networks (IDS). The five basic categories that make up the class name of the NSL-single KDD are "Normal," "Probe," "Denial of Service," "User to Root," and "Remote to User" (R2L). In addition, as shown in Table, the NSL-KDD consists of 41 characteristics, of which three are categorical and 38 are numeric.

- The datasets is download from kaggle
- https://www.kaggle.com/sensor-vehicle/download
- There are 101 columns in the data set representing various car sensors.

No.	Feature Name	Category	No.	Feature Name	Category
f1	duration	numeric	f22	is_guest_login	numeric
f2	protocol_type	nonnumeric	f23	count	numeric
f3	service	nonnumeric	f24	srv_count	numeric
f4	flag	nonnumeric	f25	serror_rate	numeric
f5	src_bytes	numeric	f26	srv_serror_rate	numeric
f6	dst_bytes	numeric	f27	rerror_rate	numeric
f7	land	numeric	f28	srv_rerror_rate	numeric
f8	wrong_fragment	numeric	f29	same_srv_rate	numeric
f9	urgent	numeric	f30	diff_srv_rate	numeric
f10	hot	numeric	f31	srv_diff_host_rate	numeric
f11	num_failed_logins	numeric	f32	dst_host_count	numeric
f12	logged_in	numeric	f33	dst_host_srv_count	numeric
f13	num_compromised	numeric	f34	dst_host_same_srv_rate	numeric
f14	root_shell	numeric	f35	dst_host_diff_srv_rate	numeric
f15	su_attempted	numeric	f36	dst_host_same_src_port_rate	numeric
f16	num_root	numeric	f37	dst_host_srv_diff_host_rate	numeric
f17	num_file_creations	numeric	f38	dst_host_serror_rate	numeric
f18	num_shells	numeric	f39	dst_host_srv_serror_rate	numeric
f19	num_access_files	numeric	f40	dst_host_rerror_rate	numeric
f20	num_outbound_cmds	numeric	f41	dst_host_srv_rerror_rate	numeric
f21	is_host_login	numeric			

5.1Modules

Supervised Classification (Training Dataset)

There is a 70:30 ratio between training and assessment information. After applying learning algorithms to the training data, predictions are made on the test data set.

Supervised Classification (Test Dataset)

Thirty percent of the entire data is used in the test dataset. Application of supervised learning algorithms to test data, followed by a comparison of the resulting output to the real output, is performed.

6. CONCLUSION

Numerous approaches for automatically creating captions for still photos and movies have been suggested and demonstrated in recent years. Despite contributing to technological progress, these models have limitations in terms of accuracy owing to basic limits. Many of the previously presented models approach picture captioning and video captioning

independently, each using their own unique algorithms and methods. This article has focused on approaches to video captioning that are grounded on picture captioning techniques. Subsequently, one might say that the video subtitling process is a blend of the outline of still picture inscriptions. Due to the previously mentioned, this study will exclusively talk about the algorithmic equals among picture and video subtitling. This article won't profess to be a thorough assessment of both picture and video inscribing, since there is algorithmic cross-over between the two. Also, just profound learning-based calculations were considered here. Those deep learning algorithms. In general, it's challenging to compare various deep learning models used in the captioning of images and videos. This is because researchers employ a wide variety of picture datasets, parameters, classification algorithms, preprocessing techniques, structure combinations, and so on. Despite their obvious dissimilarities, the authors of this research instead examined their commonalities.

Having a mechanism for accurately and reliably captioning videos and images in real time has several practical uses. Science tries to give robots eyes. Machines acquire visual perception first. Then, they improve our eyesight. Because of the machines' intelligence, we won't just utilise them; we'll work with them in ways we haven't even thought of yet. Assistive technologies such as image and video captioning systems may greatly benefit those with visual or auditory disabilities. The captions may be utilised by search engines as meta-data, expanding the search engine's capabilities in novel ways. Many different kinds of recommendation systems may benefit from the inclusion of captions.

References

[1] J. Redmon and A. Farhadi, ‘‘YOLOv3: An incremental improvement,’’ 2018, arXiv:1804.02767. [Online]. Available: http://arxiv.org/ abs/1804.02767

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, ‘‘The cityscapes dataset for semantic urban scene understanding,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 3213–3223.

[3] H. R. Arabnia and M. A. Oliver, ‘‘Fast operations on raster images with SIMD machine architectures,’’ in Computer Graphics Forum, vol. 5, Hoboken, NJ, USA: Wiley, 1986, pp. 179–188, doi: 10.1111/j.1467- 8659.1986.tb00296.x.

[4] S. M. Ehandarkar and H. R. Arabnia, ‘‘Parallel computer vision on a reconfigurable multiprocessor network,’’ IEEE Trans. Parallel Distrib. Syst., vol. 8, no. 3, pp. 292–309, Mar. 1997.

[5] H. Valafar, H. R. Arabnia, and G. Williams, ‘‘Distributed global optimization and its development on the multiring network,’’ Neural, Parallel Sci. Comput., vol. 12, no. 4, pp. 465–490, 2004.

[6] D. Luper, D. Cameron, J. Miller, and H. R. Arabnia, ‘‘Spatial and temporal target association through semantic analysis and GPS data mining.,’’ in Proc. IKE, vol. 7, 2007, pp. 25–28.

[7] R. Jafri and H. R. Arabnia, ‘‘Fusion of face and gait for automatic human recognition,’’ in Proc. 5th Int. Conf. Inf. Technol., New Generat., vol. 1, Apr. 2008, pp.

167–173.

[8] H. R. Arabnia, W.-C. Fang, C. Lee, and Y. Zhang, ‘‘Context-aware middleware and intelligent agents for smart environments,’’ IEEE Intell. Syst., vol. 25, no. 2, pp. 10–11, Mar. 2010.

[9] R. Jafri, S. A. Ali, and H. R. Arabnia, ‘‘Computer vision-based object recognition for the visually impaired using visual tags,’’ in Proc. Int. Conf. Image Process., Comput. Vis., and Pattern Recognit. (IPCV). Steering Committee World Congr. Comput. Sci., Comput. Eng. Appl. Comput. (WorldComp), 2013, p. 1.

[10] L. Deligiannidis and H. R. Arabnia, ‘‘Parallel video processing techniques for surveillance applications,’’ in Proc. Int. Conf. Comput. Sci. Comput. Intell., Mar. 2014, pp. 183–189.