A Comprehensive Investigation on Image Caption Generation using Deep Neural Networks

HARAPRASAD NAIK[1]
PRAFULLA KUMAR BEHERA[2]
SASWATI SOUMYA TRIPATHY
LAXMIPRIYA BARIK

Utkal University
Department of Computer Science and Applications
Vani Vihar, Bhubaneswar, Odisha, India- 751004
[1]hnaik.cs@utkaluniversity.ac.in
[2]pkbehera.cs@utkaluniversity.ac.in

Abstract. Caption Generation from an Image is a task that involes Object Detection and Natural Language Processing. Research in the Object Detection has been progressing tremendously since a decade. Traditional approach of Object Detection includes three steps: (1) region selection (2) Feature Extraction and (3) Classification. But in recent research, Neural Networks are used to overcome the hand-crafted feature extraction along with the application of classification techniques through various algorithm such as SVM, AdaBoost and Deformable Part Model (DPM). To generate a Caption we can take the help of Neural Network specifically Recurrent Neural Network (RNN). In most of the Caption Generator a variation of RNN that is Long Short-Term Memory (LSTM) is used. Many researchers adopts either sampling or Beam Search to generate a valid sentence/caption. In this paper we have discussed the fundamental idea behind the presently available Image Caption Generator and compare their architectural design with their performance.

Keywords: DNNs, Convolutions, LSTM, Auto-Encoder, Fine-Tunning, Attentions Mechanism.

(Received April 1st, 2022 / Accepted May 1st, 2022)

1 Introduction

Image Description or Caption generation is the fundamental problem in Artificial Intelligence, which deals with the object recognition along with Natural Language Processing(NLP). This task gained the attention of computer vision research communities because of its direct relation and applications with society. For example, smart cites can have instant alarm system when any incident / unpleasant scene captured in the security cameras, or a blind person can get descriptive command when he/she face any obstacle in his/her way, Driver alertness detection[14]. As we have stated that Image Caption generation consist of two different domains of research, so in this paper we have investigated some research articles to find the common approaches they have used and finally summed off with an abstract idea of fundamental architectural design and working principle of presently available Image Caption Generator.However, before we jump into the in-depth details of any Image Caption Generator, we need to understand the image and its properties because in this research direction image is the principal ingredient. Now a days, the image that we may encounter in our daily life, are mostly captured through digital camaras. These images are digital Images which can be stored in the memory of a digital computer. That will lead to a new era of subject domain called as Computer Vision. The digital images are presented as a composition of pixels in the computer

memory. Here the pixel is the smallest unit of a digital image that presents the intensity of a signal value. The signal that we have received from the digital camera are analog in nature and they must be converted into a numerical value. In other words, there exists a signal-to-symbol converter, which is responsible for the above said conversion of signal intensity to the corresponding pixel value. These pixel values of a digital image can be stored in a two-dimensional array. However, the objects in the real world are in the form of 3-Dimention rather than 2-Dimention. It is customary to say that these 3- Dimensional objects are mapped into the 2-Dimentional Image. It is worth noting that, Computer vision is inverse optics or inverse graphics, which can be represented by using some techniques like shading and shadows, gradients and many more. In case of binary image or monochromatic image, each pixel may be represented as either 0 (Black) or 1 (White). However, if the Image is not the binary image, then the intensity of each pixel may be presented through a cell of the matrix and that range from 0 to 255(8bits). In other words, the colour images are represented using three colour planes each of which are 8-bit, i.e., total 24-bit per pixel. Nevertheless, there exists some medical images that are Grey-scale images, which makes the utilization of 2D matrix with higher pixel intensities (16 bits) to represent the image and these usually ranges from 0 to 65535. On the other hand, the colour images are presented through the Three-Dimensional Tensor to represent three intensity level of each pixel of the image. These colour images are presented through various colour model to present colour channels such as Lightness-A-B(LAB), Red-Green-Blue (RGB),HueSaturation-Value (HSV). It is worth noting that, some images can be presented through floating point or rational numbers. If we discuss the properties of a digital image, then resolution may came into the picture. Where resolution is the unit measure of the number of pixels in an unit area of the image. Similarly, spatial resolution represented by the form of density of element determined by number of independent pixels in an image. Whereas luminance resolution determined by number of bits per pixel resolved by digitizer. The image property can be compressible, manipulable as per requirement of the format like, .jpeg / .jpg , .mpeg , .gif , .tiff etc. To understand an image completely we need to concentrate not only over the image classification but also on object localization. The object localization is a subtask of object detection where we need to employ some object classification techniques. Since our image captioning task involves both the techniques from object detection and natural language processing, we need an encoder-decoder architecture by employing some Deep Neural Networks [24, 25, 15, 19, 18, 17] as caption generator. Further the remaining contents of this paper is organized as follows: The Section-2 contains the Literature Review, the Section-3 describes the general architecture of any Caption Generation Model. The sub-section 3.1 and 3.2 explains the fundamental idea about Neural Network and Attention Mechanism respectively. The Section-4 describes the Object Detection part and Section-5 is introducing the Caption Generation part. Finally, the Section-6 is presenting the list of Dataset generally used in caption generation model along with their benchmark metrics followed by the conclusion.

2 Literature Review

There exist an URL, http://www.arxiv-sanity.com/ which presents the list of recently published article in arXiv Repository. Most of these articles are from the field of Machine Learning, Computer Vision, Artificial Neural Network and Natural Language processing. So, from this webpage we got the Idea of potentiality of research in the domain of Computer Vision. The Object Detection and Image Classification are the two principal field of research in Computer Vision. Then we decided to review the available literature in this field. During our review we got two papers, which are as follows:

- A Comprehensive Survey of Deep Learning for Image Captioning by Hossain et al. published on ACM Computer Survey, 2019.
- A Systematic Literature Review on Image Captioning by Staniute et al. published on Applied Science, MDPI Journal, 2019.[22]

From these two journal Papers, we came to know that the very first work “A Multimodal Neural Language Models” on Image Caption generation was done by Kiros et al. in 2014 at International Conference on Machine Learning held at Beijing, China. This was the only one conference paper on this field in 2014. Later on, in the year 2015, only four papers are published, and most of them are highly cited article. Afterwards the research on the field of Image Caption Generation was captured with flying colors, and we may see in the academic year 2017–18, 57 Research paper has been published and the graph of popularity of these papers exponentially growing.(For further reference see Figure-1)

3 Basic Architecture of Image Captioning

The Image Caption Generation system is originally inspired from the Machine Translation.[26][17] When

Sl.No.	Authors	Contents
1	Karpathy et. al.	proposed an alignment model which is combination of CNN and Bidirectinal RNN alongwith a structured objectives.
2	He et. al.	presented a residual learning framework to ease the training of network.
3	Soh et. al.	proposed an generative CNN-LSTM model that beats human performance.
4	Mikolov et. al.	introduced the skip-gram model.
5	vinyals et. al.	proposed the generative model based on a deep recurrent architecture that combines computer vision and machine learning to generate sentences.
6	Zhang et. al.	presented a layer flexible RNN that applies the attention mechanism.
7	Krizhevsky et. al.	trained a deep CNN by implementing dropout, which will reduce overfitting.
8	Lindh et. al.	address the limitations of generic caption and trained the image retrival model by unsupervised learning, which is capable of generating more divergent and specific caption.
9	Baker et. al.	proposed metaQNN, which is a meta-modeling algorithm based on reinforcement learning.
10	Prelec et. al.	proposed an algorithm “suprisingly popular algorithm”, that claim knowledge extraction from crowd.
11	Karras et. al.	proved the GAN can be implemented using Actor-Critic Model, which will improve qality, stability and variation of image manipulation.

Table 1: The List of research papers Reviewed

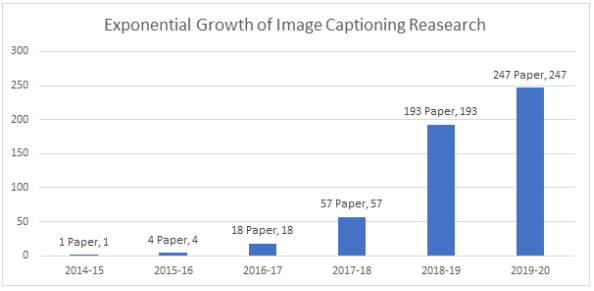

Figure 1: Graph to show exponencial growth of research in the domain of Image Caption Generation

Machine Translation gained its polpularity, where two RNNs are used in the form of Encoder-Decoder architecture. Here the Encoder RNN usually receives a set of words(Sentence) as input and the decoder RNN evaluates the desired sentence. This idea of Machine Translation inspired the advances of Image Caption Generation by replacing the encoder RNN with a Deep Convolutional Neural Network(CNN). It has been observed in last few years that many researchers used CNN as encoder, because it can extract the features from the input image and capable of embedding these features to a fixed length vector. However, the Fundamental idea behind any image captioning system comprises of Object Recognition along with Caption Generation.

The Convolutional Network/Convolutional Neural Network was first proposed by LeCun *et al.(1989)* and it was initially designed to process the data, which are in a grid-like structure(e.g Matrix). The CNN performs a convolution operation over kernel and Input data(Matrix). (Refer Figure-2) Mathematically, the discrete convolution operation may be stated as:

$$S(t) = (x * w)(t) \tag{1}$$

where w is a *probability density function.*

These Convolutions are of *3* different types:

1. Naive Convolution
2. Dilated Convolution
3. Transposed Convolution

In most of the encoder-decoder architecture, the transposed convolutions are implemented. Similarlly, The Recurrent Neural Network has the cycles in the network. In 1986 *Hopfield and Tank* proposed a Neural Network, which was called as Hopfield Network, Later on, David RumelHart modified the Hopfield Net and proposed a new network called as Reccurent Neural Network in 1986. Now a days, most of the encoder-decoder architecture are based on *LSTM (Long Short Term Memory)*, which is a variation of RNN proposed by *Hochreiter and Schmidhuber*[8] in 1997. Here the object recognition are most often carried out by the the pattern recognition, in this paper we have highlighted

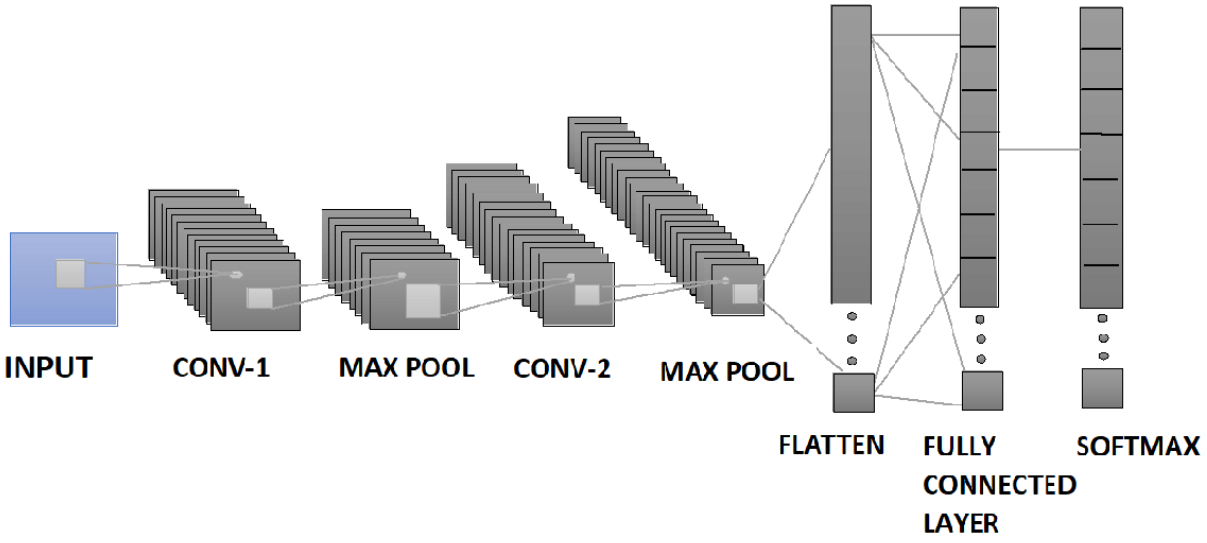

Figure 2: Architecture of CNN with Max-Pooling

some of the basic methods to classify an object, that is much more crucial to study in terms of computer vision. The pattern recognition may also be carried out using many techniques such as graph matching, neural networks, genetic algorithm, simulated annealing, and fuzzy logic. However it is important to note that object recognition will not be possible without using knowledge. in artificial Intelligence the knowledge are presented to a computer system using logic. these logic are two types Predicate Logic and Propositional Logic. the production rules may be framed using either of these two logic. Now a days, most of the image captioning system are based on Neural Networks. For example, AlexNet, VGGNet, ResNet, GoogleNet, DenseNet are often used as feature extractor that are purely based on some Neural Network. It is evident that, in most of the image recognition system, an encoder and a decoder are used to represent an image in computer memory. According to the Andrej's thesis [10] there exists two strategies of image encoding, and these are:

- Global Image Encoding.
- Fragment Level Image Encoding.

In Global Image Encoding, the Convolutional Neural Network (CNN) is used to encode a raw image into the vector. Here the CNN is used as a Feature extractor function, that is $CNN_{\theta_c}(I)$, where I is the image pixel and θ_c is the number of parameters passed. Mathematically the encoding can be represented as:

$$V = W[CNN_{\theta_c}(I)] + b \tag{2}$$

where the b is the bias and W is the weight and V is the output vector.

On the other hand, In Fragment Level Image Encoding, the Region-based CNN (R-CNN) is used to encode the image in to a set of vectors V_r and mathematically we can represent the feature extractor function as:

$$V_r = W[CNN_{\theta_c}(I)] + b \tag{3}$$

where the b is the bias and W is the weight and V_r is the set of output vectors

Analogous to the image encoding, the sentence/ caption may be encoded by encoder, where each word may be encoded with the one-hot encoding technique [10]. The Recurrent Neural Network (RNN) is used as a decoder in the encoder-decoder model *(figure-1)* and since it has the general problem of gradient vanishing, LSTM (Long Short-Term Memory) [8] can substitute RNN. Since LSTM posses long memory structure, it overcomes the problem of gradient vanishing problem [27].

3.1 The Neural Network

According to the pioneer work of *McCulloch and Pitts(1943)*, the Neural Network is the composition

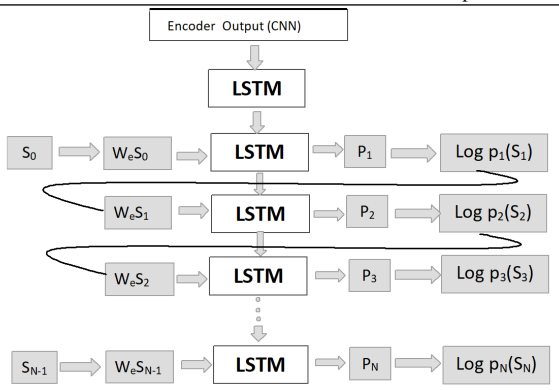

Figure 3: The Caption Generation of LSTM

of neurons / (McCulloch-Pitts) neurons. Therefore, The Deep Neural Networks(DNNs) are the architectural models where many (Hidden)layers of these interconnected neurons are composed to form a more complex networks.The Neural Networks gained its popularity since 1980, when most of the Neural Networks are used for Pattern Recognition. all neural networks may be categorised in two different classes such as Feed Forward Neural Network and Back Propagated Neural Network. The Neural Network having multiple hidden layers may be termed as Deep Neural Network[13] and the famous DNNs includes Convolutional Neural Network, Recurrent Neural Network etc. however these nueral networks requires a large amount of training Data set for the improvement of its performance. The Dataset such as Flicker, MSCOCO[12], ImageNet may be used to train the Deep Neural Network such as DeepCNN. The concept of the neural network is popular and efficient in image captioning until big data and deep learning is introduced. Image description generation method based on encoder-decoder model. An encoder is the CNN. The last layer that is the convolutional layer extracts the feature of an image. A decoder is the RNN, that is used for image description generation. RNN has more importance in the era of deep learning. LSTM is the special RNN architecture that able to solve the problem of gradient disappearance and has long-term memory. Inspite of this encoder-decoder architecture of Image Captioning system, Now a days, a new mechanism grabbing the attention of the researchers and that is Attention Mechanism.

3.2 The Attention Mechanism

The attention Mechanism has a complex cognitive ability, stemming from the human brain. Attention is the ability of self-selection. In the neural network model, the attention mechanism allow the neural network to focus on its sub inputs to find the specific feature of an image. we may categorise these attention mechanism in following categories:

Soft Attention Soft Attention is generally applied on machine translation, which focuses on calculating the weighted sum of entire-region of an image. A deterministic model can be formed by calculating weighted attention vector.

Hard Attention Rather than focusing on the entire region, Hard Attention focuses on a partilular location. this location may be selected randomly based on either by random sampling or by maximum sampling.

Multi Head Attention Generally, Input information are stored in the form of Key-value pair, But in Multi Head Attention multiple keys and values are used, where each focuses on a particular location of input Image.

Scaled Dot-Product Attention This attention function can be presented by a key, value, and query value. It is faster and more space-efficient than multi-head attention, if it becomes a scaled-down attention function.

Global Attention Global Attention evaluates the weight distribution of the model by comparing current hidden layer of decoder with each hidden layer of encoder.

Local Attention It is the combination of soft and hard attention. As compared to other attention techniques, it reduces the cost of attention mechanism.

Adaptive attention Adaptive attention is applicable to extract spatial data and provides Visual Sentinel.

Semantic attention It selects the semantic value and include them into the hidden state along with the output of the LSTM.

Spatial and channel wise attention It selects the semantic attributes to derive the necessity of the sentence context.

Areas of attention It represents the state of dependencies between predicted words, image region, and the state of the language model(RNN).

Deliberate Attention Deliberate Attention observes the habits of peoples and generates image caption accordingly.

4 The Object Detection

In earlier days, a virtual premitive recognizer was used to generate caption for videos[2]. these systems are based on rule-based system where an And-Or graph is used along with handcrafted features. Similarly, Object Detection can also be applied over the images where the traditional process of object detection involves three basic steps, which includes: (1)region selection, (2)fea-

ture extraction and (3)object classification. A multi scale sliding window is used to select the region where the objects are located. This process of selecting the region is also called as object localization. Now at this stage, to recognize an object in a particular image, one need to extract the visual features. These visual features of an image may be categorized in two different class according to their usage. The features such as SURF, SIFT, ORB are categorized as Point-like features, where as the features such as LBP, HOG, Haar may be categorized as Face-like features. More clearly the former are the features that are used for the algorithm where the point like information are required. For Example, a fingerprint matching algorithm need this kind of feature. On the other hand, the features like LBP,HOG and Haar are used for Face Recognition. However, the features descriptor may generates a large variations in feature extraction due to objects view point, occlusions, poses, and light condition. The final step is the object classification, where the classifiers are used to classify an object from another object present in an image. The popular classifier includes the name of SVM, AdaBoost etc. However since this traditional approch posses two drawbacks such as : (1) redundancy in bounding box generation, (2) the involvement of handcrafted feature. there was a scope of advancement in this area of research, therefore when Deep Learning becomes polpular in 2006, many Deep Neural Networks are used in this field to recognize an object. Furthermore, the Object Detection may be categorised in two different categories: (1) Generic Object Detction and (2) Salient Object Detection.

4.1 Generic Object Detection:

It intends to locate and classify the object present in one of the images, with rectangular rounded bounding box labeling to show the existence in a confidence manner. The arrangement of Generic object detection is defined as two types:

1. Region Proposal Based Framework
2. Regression/Classification Based Framework

4.1.1 Region Proposal Based Framework

It consists of two-steps which is equivalent to the attention mechanism within a certain range. At first it scans the whole image by focusing on the interested region. It applies CNN in sliding window method to predict the bounding boxes directly from the location. The region proposal based methods includes the models such as: SPP-net, R-CNN, Fast R-CNN, Faster R-CNN, Mask R-CNN, R-FCN, and FPN.

R-CNN: In 2014 Ross Girshick has proposed R-CNN[3]. It applies the selective search that generates 2000 region proposals for each image.R-CNN improves the feature extraction and the quality of the anchor box (Bounding Box).

SPP-Net:SPP-Net was proposed by He *et.al.* and it is based on the theory of "Spatial Pyramid Maching" . SPP-Net reuses the feature mas at 5th convolution Layer and presets as fixed length feature map.

Fast R-CNN: Fast R-CNN is identical to SPP-Net in terms of architecture. Here the extracted feature map are processed through a sequence of fully connected layer and finally it produces two separate output layer.

Faster R-CNN: Faster R-CNN was proposed by Ren *et. al.* and as per his proposal a Region Proposal Network will be used. This Region Proposal Network will be responsible for producing object boundaries. More clearly, an Image is fed into an Region Propasal Network , which will produce a set of rectangular proposal.

R-FCN:R-FCN was proposed by Li *et. al.* which is the acronym of Region-based Fully Convolutional Network. It produces m^2 number of score maps with a fixed grid size of m x m. Note that, the score map used here is a position sensitive score map. Here, for each Region-Of-Interest(ROI) a m^2 score map can be produced.

FPN:FPN is the acronym of Feature Pyramid Network, which is based on bottom-up approch. In FPN, feature pyramid can be built by down sampling the feature map with stride of two. Since the feature pyramid can retrive simantics from all levels of convolution layer, it is independent from the architecture of Convolutional Neural Network.

Mask R-CNN: Mask R-CNN can detects an object and perform instance segmentation for each instance of object. actually these are the two separate task, which requires parallel processing. Therefore Mask R-CNN adds an extra level to predict segmented mask in pixel to pixel mode. Note that Mask R-CNN is flexible enough for instance recognition along with other task such as "human pose estimation" with little modification.

4.1.2 Regression / Classification based Framework

These Frameworks are on-step framework and these are based on pixel to anchor box mapping strategy along with class probabilities. In this section we will present two significant framework known as (1) You Only Look Once (YOLO), (2) Single Shot multibox Detector (SSD).

YOLO: YOLO was proposed by Redmon *et. al.* that can predict anchor box along with their confidence fac-

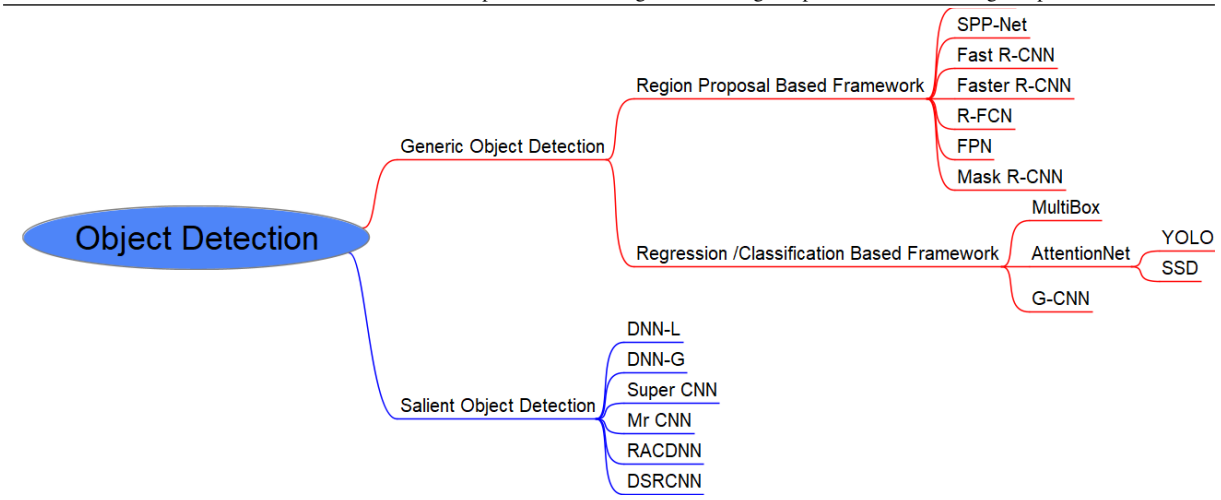

Figure 4: Different Categories of Object Detction

tor. YOLO can also maps a probability map for each class (Class Probability Map). YOLO comprises of twenty four number of convolutional layer followed by two fully connected layer.

SSD: SSD was proposed by Liu *et.al.* and SSD was inspired by multiple anchor box adoption. The SSD appends predition from feature maps to bounding box in order to handle objects with different size. in the architecture og SSD, we can observe that multiple feature layers are added to th VGG-Net16 taht are responsible to predict the offset of the anchor box with their confidence factor.SSD always trained with a weighted sum of localization Loss and Confidence Loss.

5 Composition of Captions

Several researches have been carried out since last few years in the field of object annotation, and Natural Language Processing, for example, according to the S. Li *et al.* n-gram language model along with Maximum Likelihood Estimation (MLE) can be used to predict the next possible words from the corpus of possible words, and merged together to form a caption / sentence. Similarly, a research paper of G. Kulkarni and his colleague proposed "Babytalk: understanding and generating simple image description", which states that Hidden Markov Model(HMM) along with the Conditional Random Field (CRF) can be used for Image Description generation. Part of Speech Tagging and Bag of words are also the two mostly explored field in this direction of research. Patrov et. al[16] twelve POS classes are defined, that are also called as Universal POS tags, and are as follows:

1. Noun (nouns)
2. Verb (verbs)
3. ADJ (adjectives)
4. ADV (adverbs)
5. PRON (Pronouns)
6. Det (Determinates)
7. ADP (Prepositions and Postpositions)
8. NUM (numerals)
9. CONJ (conjunctions)
10. PRT (particles)
11. ?.? (punctuation mark)
12. X (Others)

Similarly, according to the Andrej karpathy [10] a word embedding matrix may be used for encoding a caption from an one-hot vector I, and mathematically the sentence may be presented as $S_t = W_w I_t$, where W_w is the matrix of parameters to learn during backpropagation[11]. in most of the caption generator model Beam search is used to find a probable sentense/ caption. in [26] it was presented as:

$$\theta = arg \ max_\theta \sum_{(I,S)} log \ p(S|I;\theta) \qquad (4)$$

where θ is the parameter, I is the input Image and S is the partially completed Sentence. Then a joint probability over many partially completed sentence will be applied as:

$$\log \ p(S|I) = \sum_{(t=0)}^{N} \log \ p(S_t|I, S_0, S_1, ..., S_(t-1)) \tag{5}$$

6 Datasets and Results

Since Image captioning is a multimodal task of image classification and natural language processing, it always grabs the attention of the Computer Vision Researchers and in last decade plenty of research studies have been devoted in this field. For example, according to the research paper of Moses Soh [21], there exists two different approach on image captioning and these are top-down approach and bottom-up approach. The architecture we have discussed in previous section fall on the category of former one. In our previous paper(Tripathy et.al), we have used the PBOD[4] and GDX-ray dataset for scanned x-ray images.the accuracy of our proposed model can be observed in the figure-5 given below of this section. Google Neural Image Caption (NIC) Generator has illustrated the benchmark of BLUE-4 by 27.7 which is better than human baseline performance(**Refer Table-2**). In this context it is worth noting that, there exist many other evaluation models / Benchmarks such as: METEOR[9], ROUGE[20], CIDEr[7], SPICE[27]. In his paper he has implemented a generative CNN-LSTM model and made experiments over the MSCOCO dataset [12]. To alleviate overfitting attention mechanism may be used with a pre-trained mask R-CNN model in conjunction with beam search[1]. According to the Marc Tanti et. al.[23] RNN could be used as an encoder in caption generation and this can be achieved by bottom-up approach [12]. He explained that, image features can be injected into the RNN and merged together to form the caption. Similarly, Thomas Mikolov *et al.* proposed a skip-gram model, which is a learning method of learning the vector of words from an unstructured corpus of words. Later on, the word2vec[6] (word to vector) software was developed by the same author on the basis of skip-gram model. There exist many other datasets which includes ImageNet[11], PASCAL VOC[3] and SUN over which benchmarks are established earlier. Most of these datasets are labelled through crowd workers.

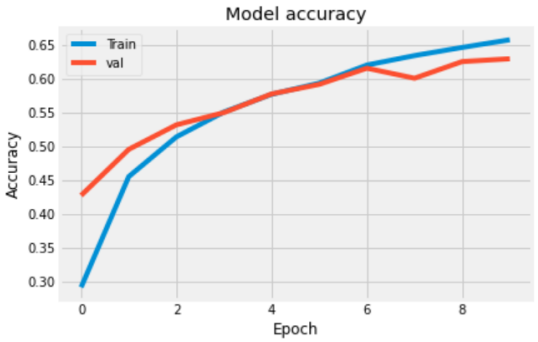

Figure 5: Accuracy of our model for Threat Detection

Table 2: List of Datasets and Number of Images.

Name	*Train*	*Valid*	*Test*
Cifar	60,000	50,000	10,000
MSCOCO	82,783	40,504	40,775
PASCALVOC	2,501	2,510	4,952
Flicker8k	6,000	1,000	1,000
Flicker30k	28,000	1,000	1,000
SUN	76,128	21,750	10,875
IMAGENET	1,281,167	50,000	1,00,000

7 Conclusion

In this paper we have compared various architecture of object detection mechanism and also reviewed the caption generation methods such as, PoS tagging, beam search etc. The sole part of any of the Image caption generator is comprises of these two parts. However, some DNNs such as GooLeNet beats the human level performance in object detection, In our review we found that if the Tabu Search[5] applied over the Beam Search(size=1) for caption generation then the model can perform better.

References

[1] Biswas, R., Barz, M., and Sonntag, D. Towards Explanatory Interactive Image Captioning Using Top-Down and Bottom-Up Features, Beam Search and Re-ranking. *KI - Künstliche Intelligenz*, (0123456789):1–14, 2020.

[2] Erhan, D., Szegedy, C., Toshev, A., and Anguelov, D. Scalable object detection using deep neural networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2155–2162, 2014.

[3] Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[4] Gittinger, J. M., Suknot, A. N., Jimenez, E. S., Spaulding, T. W., and Wenrich, S. A. Passenger baggage object database (PBOD). *AIP Conference Proceedings*, 1949(April), 2018.

[5] Glover, F. Tabu Search: A Tutorial, 1990.

[6] Goldberg, Y. and Levy, O. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *ArXiv*, (2):1–5, 2014.

[7] Gonthier, N., Gousseau, Y., and Ladjal, S. An analysis of the transfer learning of convolutional neural networks for artistic images. *ArXiv*, 2020.

[8] Hochreiter, S. and Schmidhuber, J. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997.

[9] Johnson, J., Karpathy, A., and Fei-Fei, L. DenseCap: Fully convolutional localization networks for dense captioning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem:4565–4574, 2016.

[10] Karpathy, A. Connecting Images and Natural Language. *Ph.D. Thesis*, (August), 2016.

[11] Karpathy, A. and Fei-Fei, L. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676, 2017.

[12] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8693 LNCS(PART 5):740–755, 2014.

[13] Miikkulainen, R., Liang, J., Meyerson, E., Rawal, A., Fink, D., Francon, O., Raju, B., Shahrzad, H., Navruzyan, A., Duffy, N., and Hodjat, B. Evolving deep neural networks. *Artificial Intelligence in the Age of Neural Networks and Brain Computing*, pages 293–312, 2018.

[14] Nissimagoudar, P. C., Nandi, A. V., and M, G. H. Driver alertness detection using CNN-BiLSTM and implementation on ARM-based SBC. *Infocomp Journal of Computer Science*, (2):1–9, 2020.

[15] Nunes, R. D., Rosa, R. L., and Rodríguez, D. Z. Performance improvement of a non-intrusive voice quality metric in lossy networks. *IET Communications*, 13(20):3401–3408, 2019.

[16] Petrov, S., Das, D., and McDonald, R. A universal part-of-speech tagset. *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, pages 2089–2096, 2012.

[17] Rodríguez, D. Z., da Silva, M. J., Silva, F. J. M., and Junior, L. C. B. Assessment of transmitted speech signal degradations in rician and rayleigh channel models. *INFOCOMP Journal of Computer Science*, 17(2):23–31, 2018.

[18] Rodríguez, D. Z. and Junior, L. C. B. Determining a non-intrusive voice quality model using machine learning and signal analysis in time. *INFOCOMP Journal of Computer Science*, 18(2), 2019.

[19] Rodríguez, D. Z., Rosa, R. L., Almeida, F. L., Mittag, G., and Möller, S. Speech quality assessment in wireless communications with mimo systems using a parametric model. *IEEE Access*, 7:35719–35730, 2019.

[20] Russakovsky, O. Scaling Up Object Detection. *Ph.D. Thesis*, (August), 2015.

[21] Soh, M. Learning CNN-LSTM Architectures for Image Caption Generation. *Nips*, (c):1–9, 2016.

[22] Staniute, R. and Šešok, D. A systematic literature review on image captioning. *Applied Sciences (Switzerland)*, 9(10), 2019.

[23] Tanti, M., Gatt, A., and Camilleri, K. P. What is the role of recurrent neural networks (RNNs) in an image caption generator? *INLG 2017 - 10th International Natural Language Generation Conference, Proceedings of the Conference*, pages 51–60, 2017.

[24] Terra Vieira, S., Lopes Rosa, R., Zegarra Rodríguez, D., Arjona Ramírez, M., Saadi, M., and Wuttisittikulkij, L. Q-meter: Quality monitoring system for telecommunication services based on sentiment analysis using deep learning. *Sensors*, 21(5):1880, 2021.

[25] Vieira, S. T., Rosa, R. L., and Rodríguez, D. Z. A speech quality classifier based on tree-cnn algorithm that considers network degradations. *Journal of Communications Software and Systems*, 16(2):180–187, 2020.

[26] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. Show and tell: A neural image caption generator. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June:3156–3164, 2015.

[27] Wang, H., Zhang, Y., and Yu, X. An overview of image caption generation methods. *Computational Intelligence and Neuroscience*, 2020, 2020.