IDENTIFYING OBJECT IN AN IMAGE AND GENERATING CAPTION FOR GIVEN IMAGE

Mrs.T.Kranthi

Assistant Professor, Department of CSE,

Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, India

B.Sai kumar

Department of CSE, Anil Neerukonda Institute of Technology and Sciences,

Visakhapatnam, India

A.Srikar

Department of CSE, Anil Neerukonda Institute of Technology and Sciences,

Visakhapatnam, India

A.Ephraim

Department of CSE, Anil Neerukonda Institute of Technology and Sciences,

Visakhapatnam, India

D.Ganesh teja

Department of CSE, Anil Neerukonda Institute of Technology and Sciences,

Visakhapatnam, India

ABSTRACT

Automatically generating the textual description for an image from an artificial system is known as image captioning. It uses each tongue process and computer vision to get the captions. It is a challenging artificial problem since it requires computer vision to understand the content of the image and a language model from the field of natural language processing to turn the understandings of the image into words. Many techniques proposed to solve this problem. Current approaches or techniques mainly focus on generating captions that are general about image contents.Describing images at a human level and applicable in real life environments is a challenging issue. In this paper, we would implement a neural network based method for image captioning and make it generate the captions for images at a human level by using image captioning, well grounded by the elements of images.

***Index Terms*---** Convolutional Neural Network, Recurrent Neural Network

1. INTRODUCTION

Given an image a quick glance is sufficient for a human to understand and describe what is happening in that image.Automatically generating this matter description from a synthetic system is that the task of image captioning. The task is simple the generated output is anticipated to explain what's shown within the image the objects gift, their properties, the actions being performed and therefore the interaction between the objects, etc.

As a problem that integrates vision and language understanding, its main challenges arise from the need of translating between two different, but usually paired, modalities. It was shown that just a fraction of a second is sufficient for a human to capture the meaning of the scene in order to be able to describe it accurately. This includes not only to discern most salient objects and their attributes but also reasoning about intricate relationships and interactions between them. Even more so, people describing an image usually rely on common sense knowledge for adding context, or are capable of using imagination for making descriptions vivid and interesting.

But to duplicate this behaviour in a man-made system may be a vast task, like any other image process drawback and thus the employment of advanced and advanced techniques such as machine learning to solve the task. It is a relatively new task to let a computer use a human-like sentence to automatically describe an image that is forwarded to it.

As a challenging and meaningful research field in artificial intelligence, image captioning is attracting more and more attention and is becoming increasingly important. Since much of human communication depends on natural languages, whether written or spoken, enabling computers to describe the visual world will lead to a great number of possible applications, such as producing natural human robot interactions, early childhood education, information retrieval, and visually impaired assistance, and so on.

Although much of research has gone through in this field, the current approaches mainly focus on generating captions that are general about image contents. However, they don't describe images at a human level so that they can be applicable in real-life environments. Hence in this paper we mainly focus on generating an image caption at a human level using one of the image captioning methods.

2. RELATED WORK

Recent analysis[2,3,4] has incontestable progressive image captioning results miss treat deep learning technique.These ways analyze visual data, acknowledge and classify objects and actions, and describe each still and video frames through captions. All these works use a supervised learning theme wherever pictures with corresponding captions area unit wont to train the network.Convolutional Neural Networks (CNNs) area unit deployed for visual feature extraction and algorithmic neural network based mostly architectures, either a simple recursive network or a Long-Short Term Memory (LSTM) based architectureare wont to learn the language mode land so generate descriptions. This Project work draws inspiration from their work, adapts some of the concepts used in the works and builds upon those techniques to help overcome their limitations in an attempt to improve results. This section briefly walks the readers through the approaches employed in the aforementioned researches and describes the concepts adapted.

The first work being delineate during this section is by Karpathy et al[3].The basic design for his model is shown fig 2.It uses a CNN that has been pretrained on ImageNet[5] and fine-tuned on data sets in ImageNet. In addition straight forward perennial network (SRN) is employed to operate as a caption generator .During training the SRN

is fed the image feature descriptor from the CNN in addition with the keyword START at the first time instance followed by each word in the ground through image caption from the coaching knowledge at every instance along with the hidden state from the previous time step.After coaching with enough examples. The SRN learns the language semantics and predict the next word with good accuracy based on either the previous word or the image features through the weight updates. During testing the image features descriptor extracted from the CNN is used as the first input to the SRN along with the keyword START. The first word of the image caption is excepted supported the image feature descriptor. The next prediction is created is created supported the previous prediction as input in conjuction with the previous hidden state.The process continues till the tip of sentence has been encountered .fig1 demonstrates however the project design makes a prediction on take look at image that includes a image of a person carrying a hat

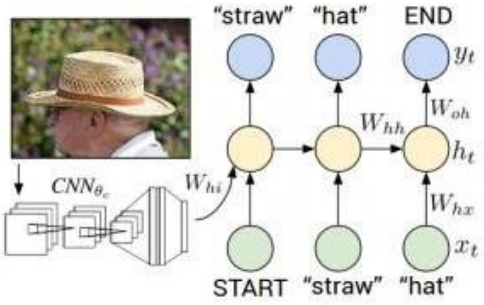

Fig 1. Architecture of image description model proposed by Karpathy et al. [3]

Before the coaching and testing, Karpathy preprocessed the words by mapping them into identical vector areas because the image feature vectors extracted from the CNN such the dot product of a word vector with its corresponding image vectors is maximized. This has been achieved through AN RCNN as planned by Girshick et al in [6], that identifies the highest nineteen regions/objects in a picture and generates twenty image feature vectors by passing these nineteen regions together with the complete image through a CNN.A SRN design, known as two-way algorithmic Neural Network (BRNN) [7] is employed to map every word into identical vector area because the image feature vector supported the contextual info close the word in each directions and also the feature vector of the word's corresponding image.

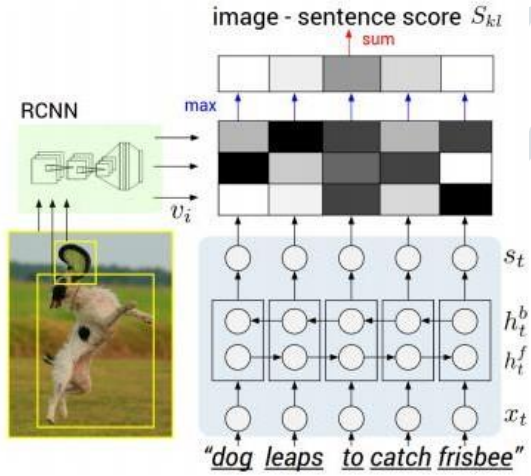

Fig 2. RCNN/BRNN based word and image feature vector embedding. [3]

This is illustrated in fig 3, whenever an image of a dog catching disk is passed to the RCNN. The example has 3 regions of interest: dog ,disk and also the entire image. The word dog, catch and leaps correlate well with the image feature vector of dog. maximizing the image-sentence scores, which is the dot product of image feature vector and word vector. Similarly the image feature vector disk features high correlation with the word disk. Higher scores are indicated with in the image with lighter shades whereas darker shades indicate lower image sentence scores .Because of this preprocessing step for word vectors and also the cooling of all the layers in CNN that is that the image feature extracting stage this model isn't end-to-end trainable .Also while multi-modal embedding is an important start as other researches show, learning it offline through a separate model is probably unnecessary

3. PROPOSED SYSTEM

The task of image captioning aims to develop visual systems that generate textual descriptions about objects in images.Given a picture, break it right down to extract the various objects, actions, and attributes, and at last generate a present sentence (caption/description) for the image. A description must contain not only the objects contained in an image, but it also must express how these objects relate to each other as well as their attributes and the activities they are involved in. The description must also be presented in a semantically correct format in a natural language like English. Hence we also need a language model in addition to the visual understanding.Thus the matter boils right down to 2 things - image analysis to urge options, so a language model to come up with important captions

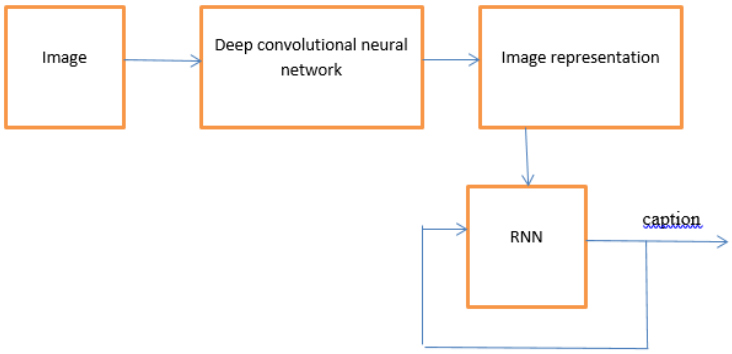

Fig:3 architecture

3.Implementation details

In this paper two summarization algorithms are implemented which mainly focuses on research papers of the given area. The two algorithms are, as follows

- Convolution neural networks
- Recurrent neural networks

3.1 .Convolution neural networks

Convolutional Neural Networks (CNNs) are a specific form of FNNs that explicitly assume the inputs to the network be structured samples, such as audio signals or image pixels which can be filtered. These architectures usually specialize in solutions for pc vision applications, like classification, localization and segmentation of pictures and videos .So far it has been assumed that layers in FNN are fully-connected, thus making each input contribute to the output of all hidden layers.If a fully-connected FNN were to be used for associate application that uses associate input from a VGA camera, whose customary resolution would be 640x480x3, then each hidden neuronshall have 921,600 weights for the connections between the input and initial hidden layer alone.An image of this dimension would need the primary hidden layer to own thousands of neurons.The model would have a billion weight parameters just for the connections between input and hidden layer. This is unacceptable both in terms of the computational power and memory requirements.

3.1.1.Convolution layer

To prevent the networks from having too many parameters, the fully-connected layers are replaced by convolutional layers in a FNN, leading to CNN models. In convolutional layers (CONV), the hidden neurons are replaced with convolutional filters .Instead of resolution for somatic cell weights, we solve for a family of filters, each filter having its own weights. The convolutional layers arrange the neurons in a 3D fashion using the height, width and depth for the signal being processed. Fig 4shows a comparison of a fully-connected conventional FNN and a CNN. Each layer in the depth dimension, aka depth slice, of the CONV layer is analogous to a filtered signal used for digital image processing, where each filtered signal came from a learned filter, whose weights shall be learned throughout the coaching method

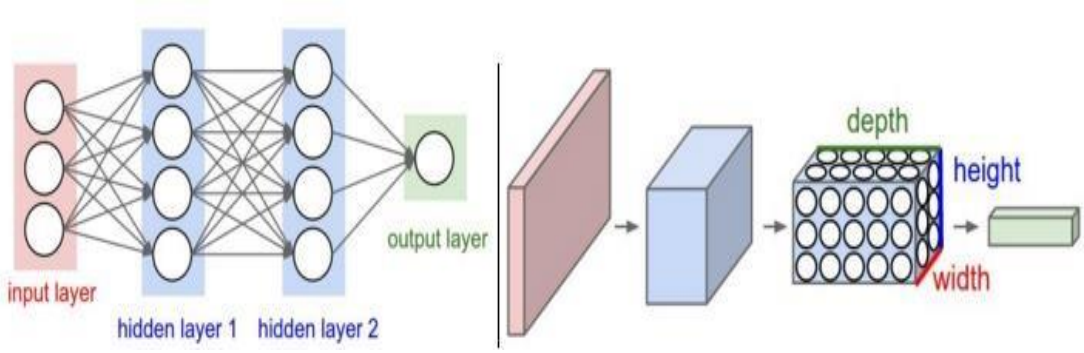

Fig 4. Comparison of FNN and CNN.

3.1.2.Depth

The depth of the convolutional layer, determines the number of different neurons that process the same receptive fields which is called the depth column, with a different set of weights. For example, in ancient gray scale image process, filter could also be of size 5x5.If the image were and color RGB image, the filter would be extended to 5x5x3. The underlying idea is similar to connecting the same input node being processed by multiple hidden nodes in traditional FNN architectures. The objective of having multiple neurons processing the same receptive field is to identify and capture different features for the same input region. Each filter applied to the input image (regardless of the depth), outputs a single output plane. The number of filters, and thus the depth of the convolutional layers are increased as the network moves from input to output as the network switches from capturing simple features to more complex features within images. The depth of the convolutional layer should not be confused with the depth of the CNN which is the number of hidden layers in a CNN.

3.1.3.Stride

While the depth is determined by the number of input planes to a filter, the stride determines the step value across and down the image as the convolution is performed. The filter width, height, depth, and stride are used to construct the 3D convolutional layer. A unit stride implies the need for introducing new depth columns for spatial regions of the image that are a unit distance apart. The stride should be chosen carefully as low stride values lead to a higher number of resolution per each filtered image, with a high overlap in the receptive fields leading to an increased redundancy in weights. Contrarily, higher stride values yield lower resolution filtered images, at the cost of an increased risk in rapid loss of vital information due to many input parameters contributing to a relatively smaller set of parameters

3.1.4.Output volume of CONV

The output volume of each CONV layer is the dimensions of the output of convolutional layer, is calculated using (11), (12) . Let Hin, Win, Din and $Hout$, $Wout$, $Dout$ be the height, width and depth of input and output of a given convolutional layer. In addition, let it be assumed that the hyperparameters receptive field, depth, stride and zero padding size are given by Hrf x Wrf, K, S and P respectively.Then the output volume parameters may be obtained by the subsequent equations.

$$Hout = (Hin - Hrf + 2 * P + 1)/S$$

$$Wout = (Win - Wrf + 2 * P + 1)/S$$

$$Dou = K$$

The stride value S needs to be picked such that H_{out}, W_{out} are integral values.

3.1.5.Parameter Sharing

In observe, there square measure a awfully few applications that value pel values at totally different|completely different} locations in a picture with different filter values.Thus, a parameter sharing scheme would lead to a great improvement in terms of the computational power, training time and memory requirements. Now that there is only one set of weights per filter for all the pixel values, the output of the CONV layer can be computed as a 3D convolution between the input and the filter weights.This is really the explanation for naming this specific FNN architectures as Convolutional Neural Networks.

3.1.6.Benefits

Based on what has been discussed so far the number of neurons in the convolutional layer shall be $Hout * Wout * Dout$ and each of these neurons has $Hrf * Wrf * Din + 1$ weight parameters. Considering the previous VGA input image with dimensions 640x480x3 with a stride of 5, a receptive field of 5x5, a filter size of 100, and a zero padding size of 0, the output volume becomes 127x95x100 and each of the neuron in the CONV has 5*5*3+1,i.e. 76 weights.Thus the convolutional layer shall have ninety one,694,000 weight parameters that is incredibly high

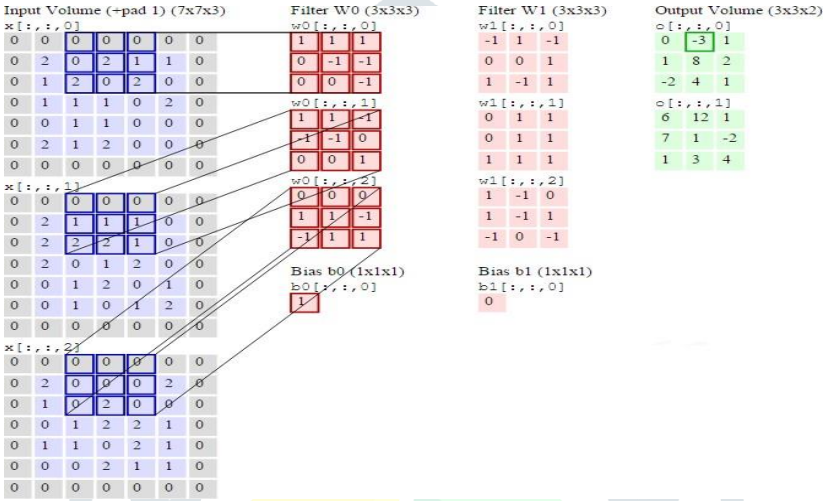

Fig 5. Output computation for CONV illustrated using a 5x5x3 input.

computation for $Dout$ filters every having $Hrf * Wrf * Din +$ one weight parameters. This reduces, the parameters of the illustrative model to 7600 from ninety one,694,000 that may be a vast improvement.Fig eighteen shows output computation for a convolutional layer with inputs of size 5x5x3, receptive field of 3x3, zero size of one, depth two and stride two.

3.2. CNN Architecture

CNNs are made up of four kinds of layers. The main constituent is the convolutional layer, CONV. The focus in this section will shift to the other three layers that constitute the CNNs. They are RELU layers (RELU), Pooling layers (POOL) and absolutely Connected layer (FC).

3.2.1.Pooling Layers

computing the output volume for the CONV layer requires a careful choice of architectural specifications such that the parameters of the output volume always yield integral outputs. Also, it is important to consider the fact that the aforementioned equations are used recursively over multiple CONV layers where the output of the first CONV layer becomes the input to the second and so on until the end. Instead of going through the painstaking process of solving these equations, it is much simpler to fix the stride to 1 and the receptive field to some constant for all the convolutional layers and adjust the padding size such that the input and output always have the same spatial dimensions .using this system to change the planning method as against [3] that will it the sophisticated approach.However, now that more researchers are preferring the simpler approach; it is essential to have a mechanism through which the spatial features can be downsized when moving away from the input layer towards the output layer thus effectively moving

away from more number of simpler feature to less number of complex features. This can be achieved by using pooling methods. The pooling layer reduces the spatial dimensions of the output volume and keeps the number of weight parameters in check. The pooling operation, works on each depth slice of the input and down samples it.The pooling operation uses 2 parameters receptive field and stride.

Let H_{ip}, W_{ip}, D_{ip} and H_{op}, W_{op}, D_{op} be the height, width and depth of input and output of a given pooling layer. In addition, let it be assumed that the receptive field and stride are x W_{rf} and S respectively. Then the output parameters of the POOL layer can be obtained by the following equations.

$$Hop = (Hip - Hrf + 1) / S$$

$$= (Wip - Wrf + 1) / S$$

$$D_{op} = D_{ip}$$

Large receptive fields are usually not used as that may throw away loads of information .The reduction in the number of parameters shouldn't be at the cost reduced accuracies of the CNNs. Some of the foremost common pooling techniques are mentioned below.

3.2.2.Max Pooling

The max pooling technique replaces all the elements of the receptive field in the input with the maximum element in the receptive field for the output. Then it moves with the specified stride to the next receptive field in the input. The most common values are 3x3 receptive fields with a stride of 2 and 2x2 receptive fields with a stride of 2. The former is referred to as overlapping max pooling, while the latter goes by non-overlapping max pooling. The latter is the most commonly employed pooling technique. Fig 7(a).provides visual image for down sampling through pooling in conjunction with Fig 7(b).which illustrates non-overlapping max pooling with an example.

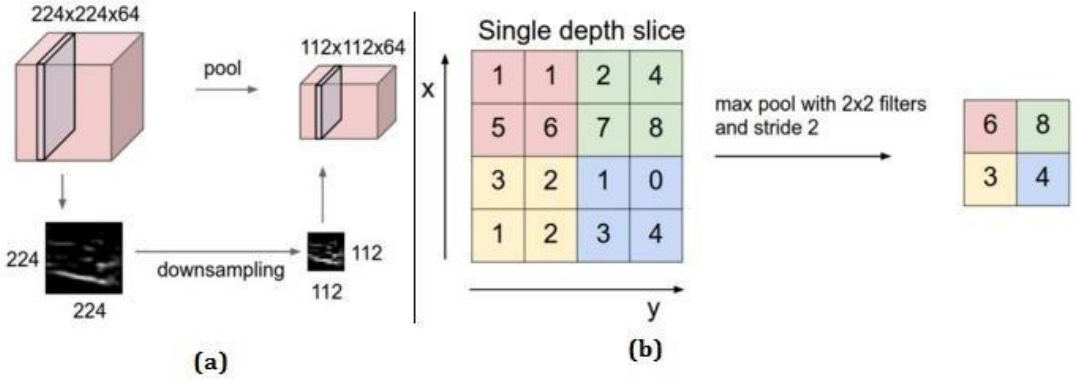

Fig 6. Downsampling the output size through pooling.

(a) Visualization of down sampling of an image using non-overlapping max pooling. (b) Illustrative example of non-overlapping max pooling.

3.2.3.Average Pooling

The average pooling method replaces the receptive field with a single element whose value is equal to the mean of all the elements in the receptive field.This technique has been used traditionally however isn't any longer favored because it has been through empirical observation incontestible that goop pooling outperforms average pooling.This is most likely due to the fact that max pooling retains the most prominent information while averaging blurs out details during downsamplingL2 PoolingThe L2 pooling method computes the L2 norm of all the elements in the receptive field and replaces the receptive field with this value. The L2 norm is just the square root of the sum of squares of all elements in the receptive field.

3.2.4.Average Pooling

The average pooling method replaces the receptive field with a single element whose value is equal to the mean of all the elements in the receptive field.This technique has been used traditionally however isn't any longer favored because it has been through empirical observation incontestible that goop pooling outperforms average pooling.This is most likely due to the fact that max pooling retains the most prominent information while averaging blurs out details during downsamplingL2 PoolingThe L2 pooling method computes the L2 norm of all the elements in the receptive field and replaces the receptive field with this value. The L2 norm is just the square root of the sum of squares of all elements in the receptive field.

3.2.4.Fully Connected layers

As previously talked about, fully connected (FC) layers are hidden layers where all the input nodes connect and contribute to all the output nodes. A fully connected layer can thus be represented as a special case of a convolutional layer where the receptive field of the filters is equal to the spatial dimensions of the input, with a padding size of zero and no stride, thus producing an output volume of 1x1xK, where K is the number total number of neurons in the FC layer. This relation between the two helps in implementing both FC and CONV layers the same way for CNNs.Now that all the layers involved in a CNN architecture have been discussed, it is time to evaluate the architecture of a typical CNN. A typical CNN architecture is shown inFig 7.

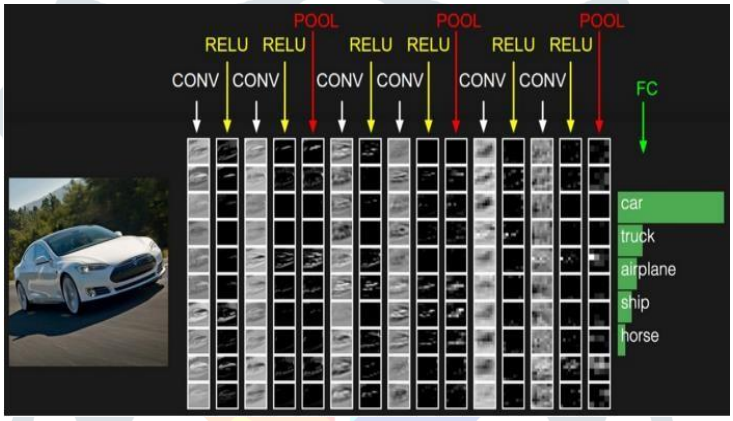

Fig 7 : **CNN architecture for a typical image classification problem.**

Typically, the POOL layer is not used after each CONV and RELU layers.This is as a result of exploitation multiple convolutions with smaller receptive field area unit typically most popularover one CONV layer with a bigger receptive field. CONV layers with smaller receptive field has a similar result as exploitation one convolutional filter with giant receptive field, with the addedbenefit of having a lower range of parameters overall.To demonstrate this they have replaced a 7x7 convolutional filter with a 3x3 convolutional filter and used the 3x3 filter thrice.Performing a 3x3 convolution thrice would cowl a similar space as a 7x7 filter would.However a 7x7 filter would have 49 parameters and all the three 3x3 filters combined would have 27 parametersThus, smaller filters perform a similar job with abundant fewer parameters.Furthermore using more number of CONV layers with smaller filters to do the same job, will increase the depth of the CNN architecture, and will increase the non-linearity introduced in the data leading to better classification results. Despite all these advantages, a CONV layer with large receptive field can be used in the first layer, if the spatial co-ordinates of input to the CNN is very high and needs to be reduced in the output volume

3.3.Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are ANNs wherein the neurons are allowed to form cyclical connections with themselves and are allowed to connect with other neurons within the same layer. A baseline RNN is depicted in Fig 8. Two specific RNN architectures include the Simple Recursive Networks (SRNs) and Long Short Term Memories (LSTMs), each of which is described and analyzed in the sections that follow.

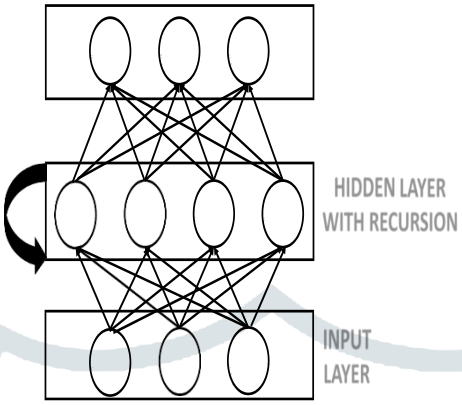

Fig 8. RNN displaying the characteristic cyclical connections.

3.3.1,Simple Recurrent Networks

A Simple Recurrent Network is a basic RNN with both cyclical and in layer connections. The architecture of a SRN can be depicted as shown in the fig 9a and fig 9b. Both these figures represent the same architecture. While the former depicts the conventional representation with the recursive connection, the latter gives an insight into the working of an RNN by depicting what happens during each time step and how the previous output of the hidden layer impacts the output of the current hidden output, along with the current input. As the output is depend on the previous hidden state(s), the output of the previous time step is impacting the current output

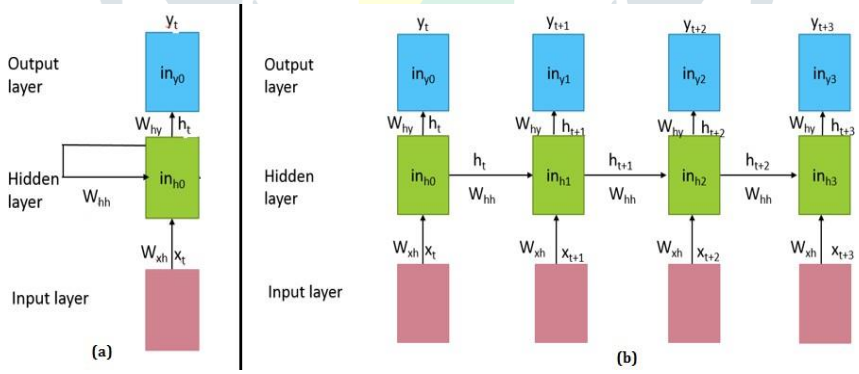

Fig 9. SRN architecture with one hidden layer.

(a) SRN architecture with all weight parameters, inputs and outputs labelled.

(b) Visualization of the impact of previous hidden states on current output using an unrolled SRN.

4. RESULTS AND PERFORMANCE EVALUATION

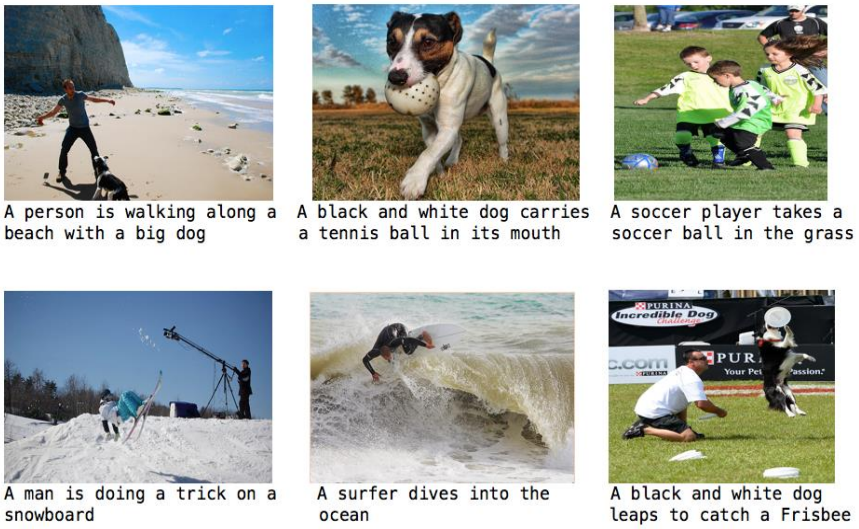

Fig 10: results

5. CONCLUSION AND FUTURE WORK

Conclusion

Image captioning has become a most recent field of research since it includes the task of computer vision and also natural language processing. It has been shown in this work we have successfully used for generating the captions of images by using neural network methods. All the existing techniques reflect some issues. The proposed method overcomes all those issues and can be used as best of all techniques.

Future Scope

Automatic image captioning is a relatively new task, thanks to the effort s made by researchers in this field, great progress has been made. In our opinion there is still much room to im- prove the performance of image captioning. due to the lack of paired image-sentence training set, research on utilizing unsupervised data, either from images alone or text alone, to improve image captioning will be promising. Fourth, current approaches mainly focus on generating captions that are general about image contents. However Research on solving image captioning problems in various special cases will also be interesting.

6.ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their careful reading of this paper and for their helpful comments

7.REFERENCES

[1] O. Russakovsky, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, pp. 211-252, 2015

[2] J. Donahue, *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," *arXiv preprint arXiv:1411.4389*, 2014.

[3] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *arXiv preprint arXiv:1412.2306*, 2014

[4] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," *arXiv preprint arXiv:1411.4555*, 2014.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 248-255

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014, pp. 580-587

[7] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *Signal Processing, IEEE Transactions on*, vol. 45, pp. 2673-2681, 1997.

[8] F.-F. Li and A. Karpathy. (2015, 29 Oct). *Convolutional Neural Networks for Visual Recognition*.

[9] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *The Journal of physiology*, vol. 195, pp. 215-243, 1968.

[10] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of physiology*, vol. 160, p. 106, 1962.