Image Caption Generator Using CNN and LSTM

[1] Neha Kumar, [2] Rishabh Anand, [3] Samay M Shetty, [4] Shikhar Jaiswal

[1,2,3,4] Student
Information Science and Engineering
RVITM, Bangalore, India

Abstract - In this study, we employ CNN and LSTM models to detect and analyse image captions. Picture subtitle age is a framework that fathoms regular language handling and PC vision principles to perceive the association of the picture in English. In this examination paper, we mindfully seek after various significant ideas of photo subtitling and its natural cycles. We discuss Keras library, numpy and jupyter scratch pad for the creation of this paper. We likewise discuss flickr_dataset and CNN utilized for photograph order.

Index Terms - CNN, LSTM, image captioning, deep learning.

I. INTRODUCTION

Everyday, we come across numerous photographs in our environment, social media, and the media. Only human beings are capable of recognizing images. Humans can identify photos without their prescribed captions, but machines require training images before they can produce captions for photos automatically. Image captioning has various applications, including providing real-time feedback through a camera feed to assist visually impaired individuals who use text-to-speech technology, improving social media engagement by generating captions for pictures in social feeds, and converting messages to speech. Additionally, it can aid children in identifying chemicals as they learn the language. Every picture on the internet can have captions, which makes browsing and indexing more quickly and thoroughly. There are several applications for image captioning, including those in biomedicine, business, the military, and online search. Social media platforms like Instagram and Facebook can automatically create subs/CC from photographs. This study paper's main objective is to gain some understanding of deep learning techniques. For image classification, We utilize two particular approaches for our methods: CNN and LSTM.

II. LITERATURE SURVEY

TECHNIQUES FOR IMAGE CAPTIONING

CNN- CNN is crucial for working with images, and convolutional brain frameworks are explicit significant brain frameworks that can give data that has a data shape, for example, a 2D grid. It examines images from the left corner to the right corner and all the way through to eliminate prominent elements, then solidify the component to characterize images. It can handle edited, scaled, turned, and decoded images. A substantial learning computation, the convolutional brain framework gets the data picture, assigns priority to several elements/fights in image, and remembers the respective from one another.

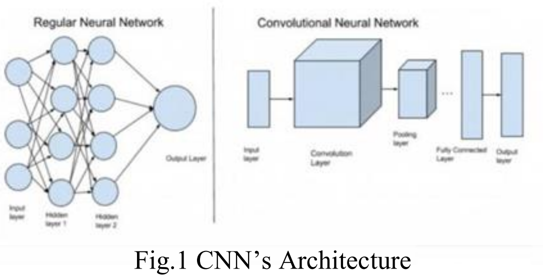


Fig.1 CNN's Architecture

When compared to other request estimations, ConvNet's pre-treatment need is irrelevant. ConvNets is effective for learning these channels/features even while diverts are hand-planned in crude approaches with sufficient preparation. The method of the association of the visual cortex drives the development of the bent framework, which mirrors the neuronal organisation plan within the human psyche. Individual neurons predominantly react to changes in a limited area of the visible field referred to as the receptive field. These receptive fields collectively cover the entirety of the visual areas.

CNN's Architecture

When viewing large images and videos, an unaltered natural brain architecture, where in one layer all neurons converge with following layer which also consists of neurons, is wasteful. The scope of limitation utilising a recognised brain framework will be in the tonnes for a normal size picture with numerous picture components called pixels and 3-tone tones (RGB, or red, green, and blue tone), and that can lead to overfitting.

CNN uses a 3D strategy in which every change in neurons separates a small section or "feature" of the image in order to compel realistic levels of restrictions and recognition of the brain structure on large bits of picture. Each social event of the neurons invests significant work in differentiating one specific aspect of the picture, such as a nose, left ear, mouth, or leg, as opposed to all neurons skirting their selections to the next brain layer. The final result serves a purpose and demonstrates the wisdom with which every capability was selected as a class characteristic.

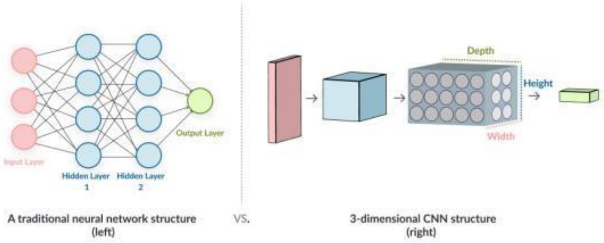

Input Layer
Hidden Layer 1
Hidden Layer 2
Output Layer
Depth
Height
Width
A traditional neural network structure (left)
VS.
3-dimensional CNN structure (right)

Fig.2 CNN's Working

Working of CNN

As discussed previously, a brain network that is fully connected and where the contribution of each layer is associated with the contribution of the subsequent layers is beneficial for the main task. According to CNN, the neurons in a cell may be linked to a specific region in the preceding cell rather than all neurons being connected in an identical manner.

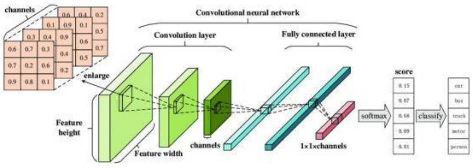
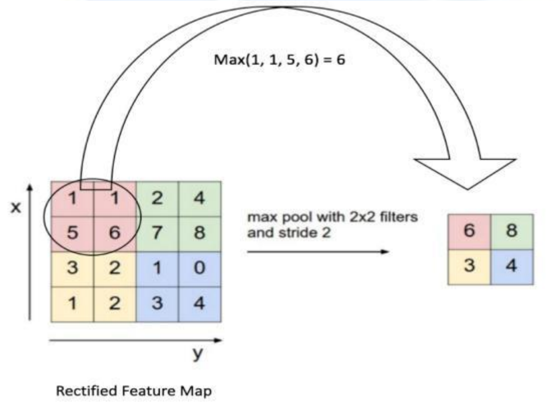


Fig.3 Layers generated from the features of the image

This assists in diminishing the intricacy of the brain organization and obtaining less figuring with driving. Using numbers to represent each pixel in a standard image is a common practice in modern personal computers. At the point when one and large analyse two pictures the pixel is checked upsides of every other pixel. This strategy just assists with looking at two indistinguishable pictures just yet when we keep various pictures to think about the correlation fizzles. In CNN picture correlation happens piece by piece.



1	1	2	4
5	6	7	8
3	2	1	0
1	2	3	4

6	8
3	4

Fig.4 Include guide of CNN

The primary reason for utilizing the CNN algorithm is that it is the only algorithm capable of accepting images as input and generating a feature map based on similarities and differences between pixels. CNN defines the pixels and creates a network, which is known as a feature map. A feature map is a collection of similar pixels placed in a distinct category. These networks play a critical role in identifying the content of the object in the input image.

A bit more on CNN

The CNN's model consists of three distinct types of layers.

1.Convolutional

2.Pooling

3.Fully connected

In the principal layer, the info picture is perused the CNN, and on that establishment a component map is made. From that component map, it fills in as a contribution to the accompanying layers, i.e for the Pooling layer. In the pooling layer, the element map is separated into extra more straightforward parts to look at the setting of the image cautiously. This layer makes the element map more thick in order to find the most basic data about the image.

The first and second layers i.e Convolutional and Pooling they're polished so often, contingent upon the image as to get the densed data about the image. The extra thick component map is made due to these two layers. Furthermore, this densed highlight map is used by the last layer i.e Completely Associated.

The layer is responsible for conducting characterization, which involves sorting pixels based on their proximity and contrast. The characterization process is carried out with utmost precision in order to capture the essence of the image, aiding in the identification of objects, individuals, and other elements present within it.

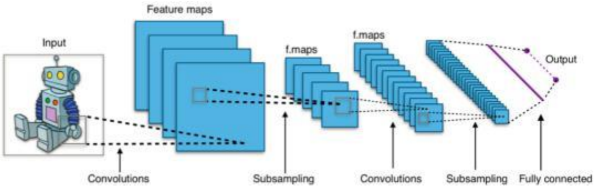


Fig.5 Scanned picture layers

The implementation of these layers in CNN allows for easy identification and tracking of the image's features, without plagiarism. The extraction of crucial elements from fixed-length inputs is transformed into fixed-sized outputs. CNN methods have multiple applications.

Computer vision— In field of medical science, CNNs are exclusively utilized for image analysis to effortlessly examine the inner structure of the body. Similarly, in the realm of mobile phones, CNNs have been applied for various purposes, such as determining the age of an individual and unlocking the phone via picture analysis. Furthermore, industries utilize this technology to establish patents or copyrights for distinctively captured images.

Pharmaceuticals discovery— The process of identifying drugs and pharmaceuticals has become widely prevalent. It involves analyzing chemical characteristics to determine the most effective drug for a specific condition.

LSTM's Origin

In the year of 1997, Sepp Hochreiter and Jurgen Schmidhuber, two researchers from Germany, first discovered LSTM. It plays a vital role in the Deep Learning field of recurrent neural networks. What makes LSTM unique is its ability to not only store input data, but also predict future datasets using its own data. The LSTM network holds onto stored data for a specific time period and uses that information to forecast or provide future values for the data. This is why LSTM is preferred over traditional RNN in this field.

The Problem persisting with RNN's

RNN's (Recurrent Neural network) are an essential component of "deep learning" algorithms that are utilized to tackle a variety of intricate computer-related tasks such as speech recognition and item classification. RNNs are designed to handle a series of sequential activities where each scenario is based solely on data from previous occurrences, enabling them to address a diverse range of complex problems.

We aim to prioritize the use of RNNs that possess extensive data collections and superior capabilities. These RNNs have the potential to effectively address real-life challenges such as inventory forecasting and speech recognition enhancement. However, the Vanishing Gradient issue has hindered the practical application of RNNs for problem-solving purposes.

How can we address the issue of the Vanishing Gradient problem in RNNs?

To address the issue at hand, Long short-term memory (LSTM) will be utilized, which falls under the category of Recurrent Neural Networks (RNNs). LSTMs were specifically designed to overcome the problem of Vanishing Gradients. What sets LSTMs apart is their ability to retain data values over an extended period of time, effectively tackling the vanishing gradient problem.

LSTMs are designed to always have errors, allowing them to continually learn data values across multiple time steps. This iterative process of learning makes backpropagation simpler over time and across layers.

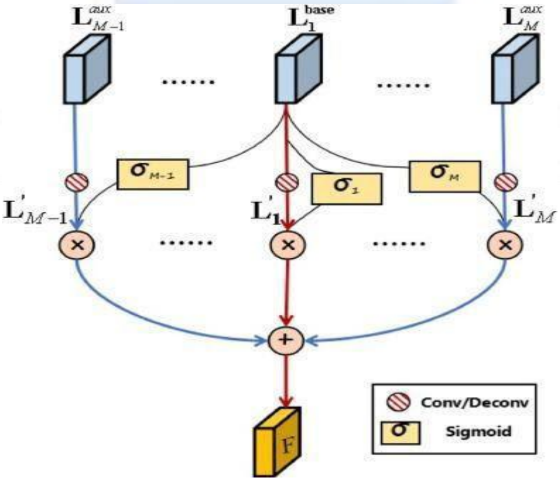


Fig.6 LSTM's gates

According to the above diagram presented, LSTM architecture employs multiple gates for data storage, processing, and transfer to the final gate. In contrast, Recurrent Neural Networks (RNNs) are utilized to transmit data directly to the final gate without undergoing any manipulation. However, the different gates present in the LSTM network provide the capability for multiple forms of data

manipulation, such as storage and analysis. Additionally, the LSTM gates are independently capable of analysing and interpreting data, and have the ability to open or close as needed.
The comprehension of LSTM gates retaining information for an extended period confers an advantage to the LSTM in comparison to RNN's.

LSTM's Architecture
LSTM's architecture is simple and consists of three main gates that allow for the retention of information over long periods, overcoming the limitations of Recurrent Neural Networks (RNN's). The three primary gates of the LSTM enclosures include:

Forget gate — The primary function of the forget gate is to filter out irrelevant data that will not be necessary to complete a specific task in the future. The optimization of data in the LSTM is a critical aspect, and the gate plays a vital role in achieving this goal by filtering out redundant information.

Input gate — The commencement of LSTM commences with the input gate, which is responsible for receiving information from the user and relaying it to other gates.

Output gate — The function of this gate is to present the intended outcome effectively**.**

Uses of LSTM Networks
LSTMs are extensively employed in a diverse range of deep learning tasks, primarily involving the forecasting of data based on prior information. Two notable examples include predicting text and forecasting stock market trends.

Text Prediction - The LSTM algorithm is frequently utilized for text prediction due to its ability to effectively analyze long-term memory. The sophisticated understanding of the LSTM algorithm enables it to predict the succeeding words in sentences, resulting in highly accurate text prediction. This predictive capability is the outcome of the LSTM network's ability to store information such as word usage, word styling, and contextual relevance, which is subsequently used to anticipate forthcoming words. The input data is preserved for future use, further increasing the efficiency of the algorithm. A prime example of text prediction is chatbots, which are commonly used by e-commerce webpages and mobile apps.

Trends on Stock market - The utilization of LSTM in the stock market entails the storage of data concerning market trends and behavior at specific instances, facilitating predictions of future market fluctuations. Accurately forecasting stock market variations is challenging, as they are unpredictable. The LSTM model necessitates extensive training to deliver precise values to users, requiring the accumulation of large amounts of data over extended periods, which may last for days.

A bit more on LSTM
LSTMs are a component of RNNs that can retain a greater quantity of data than RNNs. In today's world, LSTMs are extensively used in all domains. The fundamental LSTM diagram illustrated below has three primary gates, namely the forget, input and the output gates respectively, all of which possess the ability to retain data and supply the necessary output When referring to LSTM networks, the three gates are always mentioned.

III. CONCLUSIONS

In this overview, we have compiled all aspects of the image caption generation task, discussed the model framework proposed in recent years to solve the description task, focused on the algorithmic essence of different attention mechanisms, and summarized how the attention mechanism is applied. We summarize the large datasets and evaluation criteria commonly used in practice.

Although image caption can be applied to image retrieval, video caption, and video movement and the variety of image caption systems are available today, experimental results show that this task still has better performance systems and improvement. It mainly faces the following three challenges: first, how to generate complete natural language sentences like a human being; second, how to make the generated sentence grammatically correct; and third, how to make the caption semantics as clear as possible and consistent with the given image content.

IV. FUTURE ENHANCEMENT

For future work, we propose the following four possible improvements:
(1) An image is often rich in content. The model should be able to generate description sentences corresponding to multiple main objects for images with multiple target objects, instead of just describing a single target object.
(2) For corpus description languages of different languages, a general image description system capable of handling multiple languages should be developed.
(3) Evaluating the result of natural language generation systems is a difficult problem. The best way to evaluate the quality of automatically generated texts is subjective assessment by linguists, which is hard to achieve. In order to improve system performance, the evaluation indicators should be optimized to make them more in line with human experts' assessments.
(4) A very real problem is the speed of training, testing, and generating sentences for the model should be optimized to improve performance.

V. REFERENCES

[1] Abhaya Agarwal and Alon Lavie. 2008. Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with human rankings of machine translation output. In Proceedings of the ThirdWorkshop on Statistical Machine Translation. Association for Computational Linguistics, 115–118.

[2] Ahmet Aker and Robert Gaizauskas. 2010. Generating image descriptions using dependency relational patterns. In Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics, 1250–1258.

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In European Conference on Computer Vision. Springer, 382–398.

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and vqa. arXiv preprint arXiv:1707.07998 (2017).

[5] Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. 2018. Convolutional image captioning.

[6] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, Trevor Darrell, Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, et al. 2016. Deep compositional captioning: Describing novel object categories without paired training data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In International Conference on Learning Representations (ICLR).

[8] Shuang Bai and Shan An. 2018. A Survey on Automatic Image Caption Generation. Neurocomputing. ACM Computing Surveys, Vol. 0, No. 0, Article 0. Acceptance Date: October 2018. 0:30 Hossain et al.

[9] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, Vol. 29. 65–72.