IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

IMAGE CAPTION GENERATION USING LSTMS AND CONVOLUTIONAL NEURAL NETWORKS

[1] Kanishya Gayathri D, [2]Prinslin L
[1]PG Student, [2]Assistant Profession
[1]Department of Computer Science and Engineering,
,[1]Agni College of Technology, Thalambur, Chennai, India

Abstract - Image caption generation is a captivating intersection of computer vision and natural language processing, with applications spanning assistive technology, content retrieval, and human-computer interaction. In this project, we delve into the fusion of Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) to address this intriguing challenge.

Our project centres on the holistic development of an image captioning system. We kick-started the journey with meticulous data collection and pre-processing, leveraging established datasets like MSCOCO. The images underwent resizing, normalization, and feature extraction using a pre-trained CNN, serving as our image encoder.

The nucleus of our innovation resides in the model architecture. We meticulously designed a two-tiered structure, comprising a CNN as the image encoder and an LSTM as the text decoder. The CNN's role is to extract salient image features, while the LSTM excels at generating coherent and contextually relevant captions, guided by a cross-entropy loss function during training. Teacher forcing further stabilizes our model's convergence.

Our system underwent rigorous evaluation employing a battery of metrics, including BLEU, METEOR, CIDEr, and ROUGE, benchmarking the generated captions against human-annotated references. This quantitative analysis provides an in-depth perspective on the system's performance and underscores areas for enhancement.

As we conclude this report, we reflect on the challenges encountered throughout our project's lifecycle and propose avenues for future research in the captivating realm of image caption generation. Our aspiration is to refine and enhance our system, enabling it to generate not just accurate but also contextually rich captions—an endeavor that advances the frontiers of multimodal AI applications.

In summary, this project showcases the symbiotic relationship between computer vision and natural language processing, underscoring the potential of CNN-LSTM architectures in the captivating domain of image captioning. Our work contributes to the burgeoning field of multimodal AI, fostering innovative applications across diverse domains.

Index Terms - Image Captioning, Convolutional Neural Networks, LSTMs, Deep Learning, Computer Vision, Natural Language Processing, MSCOCO Dataset, Data Preprocessing, Model Architecture, Training, Evaluation Metrics, Encoder-Decoder, Teacher Forcing, Cross-Entropy Loss, Metric-Based Evaluation, BLEU, METEOR, CIDEr, ROUGE, Multimodal AI, Visual Understanding, Text Generation, Image Description, Machine Learning, Deep Neural Networks.

I. INTRODUCTION

In the realm of artificial intelligence, the fusion of Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) has revolutionized image understanding and language generation. Our project, "Image Caption Generation using LSTMs and Convolutional Neural Networks," explores this captivating synergy at the crossroads of computer vision and natural language processing.

Image captioning, the process of generating textual descriptions for images, has diverse applications, from assisting the visually impaired to enriching content retrieval. Our project seeks to empower machines with the ability to comprehend images and express this understanding through coherent, contextually relevant captions.

This report offers an in-depth journey through our project's phases. We start with data collection and pre-processing, obtaining a dataset comprising images paired with human-generated captions. Next, we delve into our model architecture—a two-tiered system featuring a CNN as the image encoder and an LSTM as the text decoder. The CNN extracts image features, while the LSTM generates captions, trained using cross-entropy loss and teacher forcing.

Our system's performance is meticulously evaluated using metrics such as BLEU, METEOR, CIDEr, and ROUGE, benchmarked against human-annotated references. These metrics offer quantitative insights into caption quality, guiding model enhancements.

Throughout this report, we also discuss project challenges and future research avenues. Our work underscores the potential of CNN-LSTM architectures in bridging visual and textual domains, with implications across diverse applications.

Deep Learning

Deep learning is a subset of machine learning, which is essentially a neural network with three or more layers. These neural networks attempt to simulate the behavior of the human brain—albeit far from matching its ability—allowing it to "learn" from large amounts of data. While a neural network with a single layer can still make approximate predictions, additional hidden layers can help to optimize and refine for accuracy.

Deep learning drives many artificial intelligence (AI) applications and services that improve automation, performing analytical and physical tasks without human intervention. Deep learning technology lies behind everyday products and services (such as digital assistants, voice-enabled TV remotes, and credit card fraud detection) as well as emerging technologies (such as self-driving cars).

Deep learning neural networks, or artificial neural networks, attempts to mimic the human brain through a combination of data inputs, weights, and bias. These elements work together to accurately recognize, classify, and describe objects within the data.

Deep neural networks consist of multiple layers of interconnected nodes, each building upon the previous layer to refine and optimize the prediction or categorization. This progression of computations through the network is called forward propagation. The input and output layers of a deep neural network are called visible layers. The input layer is where the deep learning model ingests the data for processing, and the output layer is where the final prediction or classification is made.

Another process called backpropagation uses algorithms, like gradient descent, to calculate errors in predictions and then adjusts the weights and biases of the function by moving backwards through the layers in an effort to train the model. Together, forward propagation and backpropagation allow a neural network to make predictions and correct for any errors accordingly. Over time, the algorithm becomes gradually more accurate.

The above describes the simplest type of deep neural network in the simplest terms. However, deep learning algorithms are incredibly complex, and there are different types of neural networks to address specific problems or datasets. For example,

Convolutional neural networks (CNNs), used primarily in computer vision and image classification applications, can detect features and patterns within an image, enabling tasks, like object detection or recognition. In 2015, a CNN bested a human in an object recognition challenge for the first time.

Recurrent neural network (RNNs) are typically used in natural language and speech recognition applications as it leverages sequential or times series data.

Image to Speech Using Deep Learning

Here, we will caption the input images then convert those text to speech. So we use to methods for doing the whole work.Image caption includes the multi-level use of image information. From the target in the image, the relationship between the targets, to the description of the image, and the construction of the scene graph, all belong to the category of image description research. Each task in image caption has great research value and great practical application value. Image caption methods can be divided into template-based methods, retrieval-based methods and deep learning-based methods. The template-based method first obtains some visual concepts for the image, and then generates a sentence through sentence templates, syntactic rules, or combined methods. Retrieval-based methods usually need to save a large database, and then obtain a sentence or a group of sentences through image retrieval, and then obtain a complete image description. The image description based on deep

learning mainly uses the structure of codec to complete the image caption task. Further, the effect of image description can be improved through the attention machine or other methods of enhancing the deep learning model.

In order to better complete the image caption task, on the one hand, the model needs to use the information in the image more efficiently and rationally, and on the other hand, the problem of inconsistency between model training and testing needs to be solved. In order to solve these two problems, curriculum learning, reinforcement learning and other methods

are introduced into the image caption task. In the course of training, the actual labeled words and the words generated by the model will be sampled according to a certain proportion as the input of the decoder, which can alleviate the problem of inconsistent input during training and testing to a certain extent.Reinforcement learning is to introduce feedback of test indicators during training in the form of agents to alleviate the problem of inconsistency between the optimized loss function and the test indicators during training. This paper introduces context coding as well as reinforcement learning. and makes full use of image features to describe the image, so that the generated description sentence is more conform to content of the image.

The bulk of recent research into human speech has focused on neural network techniques to improve speech understanding and recognition or to provide simplified, high-quality textto-speech (TTS) models that can directly convert written text into natural speech. The most highly developed domain for such research has a focus on spoken English and is

based on native-speaker adult voice data samples. Automated speech recognition (ASR) is a core element of modern consumer technology user interfaces employed in smart-speaker and voice command interfaces.

II. LITERATURE SURVEY

1.Kulkarni G, Premraj V, Dhar S, et al. Baby talk: Understanding and generating simple image descriptions[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society

It posits that visually descriptive language offers computer vision researchers both information about the world, and information about how people describe the world. The potential benefit from this source is made more significant due to the enormous amount of language data easily available today. We present a system to automatically generate natural language descriptions from images that exploits both statistics gleaned from parsing large quantities of text data and recognition algorithms from computer vision. The system is very effective at producing relevant sentences for images. It also generates descriptions that are notably more true to the specific image content than previous work.

A significant amount of this language describes the world around us, especially the visual world in an environment or depicted in images or video. Such visually descriptive language is potentially a rich source of information about the world, especially the visual world, and training data for how people construct natural language to describe imagery. This paper exploits both of these lines of attack to build an effective system for automatically generating natural language – sentences – from images. It is subtle, but several factors distinguish the task of taking images as input and generating sentences from tasks in many current computer vision efforts on object and scene recognition.

2.Ordonez V, Kulkarni G, Berg T L. Im2Text: describing images using 1 million captioned photographs[C]// International Conference on Neural Information Processing Systems.

Producing a relevant and accurate caption for an arbitrary image is an extremely challenging problem, perhaps nearly as difficult as the underlying general image understanding task. However, there are already many images with relevant associated descriptive text available in the noisy vastness of the web. The key is to find the right images and make use of them in the right way! In this paper, we present a method to effectively skim the top of the image understanding problem to caption photographs by collecting and utilizing the large body of images on the internet with associated visually descriptive text. We follow in the footsteps of past work on internet vision that has demonstrated that big data can often make big problems – e.g. image localization, retrieving photos with specific content, or image parsing – much more bite size and amenable to very simple nonparametric matching methods. In our case, with a large captioned photo collection we can create an image description surprisingly well even with basic global image representations for retrieval and caption transfer. In addition, we show that it is possible to make use of large numbers of state of the art, but fairly noisy estimates of image content to produce more pleasing and relevant results. People communicate through language, whether written or spoken.

3.Vinyals O, Toshev A, Bengio S, et al. Show and Tell: A Neural Image Caption Generator.

Automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. In this paper, we present a generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image. Being able to automatically describe the content of an image using properly formed English sentences is a very challenging task, but it could have great impact, for instance by helping visually impaired people better understand the content of images on the web. This task is significantly harder, for example, than the well-studied image classification or object recognition tasks, which have been a main focus in the computer vision community. Indeed, a description must capture 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. Moreover, the above semantic knowledge has to be expressed in a natural language like English, which means that a language model is needed in addition to visual understanding. The model is trained to maximize the likelihood of the target description sentence given the training image. Experiments on several datasets show the accuracy of the model and the fluency of the language it learns solely from image descriptions. Our model is often quite accurate, which we verify both qualitatively and quantitatively.

4.Aneja J, Deshpande A, Schwing A. Convolutional Image Captioning.

Image captioning is an important task, applicable to virtual assistants, editing tools, image indexing, and support of the disabled. In recent years significant progress has been made in image captioning, using Recurrent Neural Networks powered by long-short-term-memory (LSTM) units. Despite mitigating the vanishing gradient problem, and despite their compelling ability to memorize dependencies, LSTM units are complex and inherently sequential across time. To address this issue, recent work has shown benefits of convolutional networks for machine translation and conditional image generation [9, 34, 35]. Inspired by their success, in this paper, we develop a convolutional image captioning technique. We demonstrate its efficacy on the challenging MSCOCO dataset and demonstrate performance on par with the LSTM baseline , while having a faster training time per number of parameters. We also perform a detailed analysis, providing compelling reasons in favor of convolutional language generation approaches.Image captioning, i.e., describing the content observed in an image, has received a significant amount of attention in recent years. It is applicable in various scenarios, e.g., recommendation in editing applications, usage in virtual assistants, for image indexing, and support of the disabled. With the availability of large datasets, deep neural network (DNN) based methods have been shown to achieve impressive results on image captioning tasks [16, 37]. These techniques are largely based on recurrent neural nets (RNNs), often powered by a Long-Short-Term-Memory (LSTM) component.

5.Jie Hu, LiShen,Samuel Albanie,GangSun,Enhua Wu:Squeeze-and?Excitation Networks.journal version of the CVPR 2018 paper,accepted by TPAMI.cs

Convolutional neural networks are built upon the convolution operation, which extracts informative features by fusing spatial and channel-wise information together within local receptive fields. In order to boost the representational power of a network, several recent approaches have shown the benefit of enhancing spatial encoding. In this work, we focus on the channel relationship and propose a novel architectural unit, which we term the "Squeezeand-Excitation" (SE) block, that adaptively recalibrates channel-wise feature responses by explicitly modelling interdependencies between channels. We demonstrate that by stacking these blocks together, we can construct SENet architectures that generalise extremely well across challenging datasets. Crucially, we find that SE blocks produce significant performance improvements for existing state-ofthe-art deep architectures at minimal additional computational cost.

6.J. Xu, X. Tan, Y. Ren, T. Qin, J. Li, S. Zhao, and T. Y. Liu, ''LR Speech: Extremely low-resource speech synthesis and recognition,'' in Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Assoc. Comput. Mach. New York, NY, USA

Speech synthesis (text to speech, TTS) and recognition (automatic speech recognition, ASR) are important speech tasks, and require a large amount of text and speech pairs for model training. However, there are more than 6,000 languages in the world and most languages are lack of speech training data, which poses significant challenges when building TTS and ASR systems for extremely low-resource languages. In this paper, we develop LR Speech, a TTS and ASR system under the extremely low-resource setting, which can support rare languages with low data cost. LR Speech consists of three key techniques: 1) pre-training on rich-resource languages and fine-tuning on low-resource languages; 2) dual transformation between TTS and ASR to iteratively boost the accuracy of each other; 3) knowledge distillation to customize the TTS model on a high-quality target-speaker voice and improve the ASR model on multiple voices. We conduct experiments on an experimental language (English) and a truly low-resource language (Lithuanian) to verify the effectiveness of LR Speech. Experimental results show that LR Speech

7.P. K. Muthukumar and A. W. Black, ‘‘A deep learning approach to data-driven parameterizations for statistical parametric speech synthesis,’’ Carnegie Mellon University Pittsburgh, Pittsburgh, NA, USA, 2014.
Nearly all Statistical Parametric Speech Synthesizers today use Mel Cepstral coefficients as the vocal tract parameterization of the speech signal. Mel Cepstral coefficients were never intended to work in a parametric speech synthesis framework, but as yet, there has been little success in creating a better parameterization that is more suited to synthesis. In this paper, we use deep learning algorithms to investigate a data-driven parameterization technique that is designed for the specific requirements of synthesis. We create an invertible, low-dimensional, noise-robust encoding of the Mel Log Spectrum by training a tapered Stacked Denoising Autoencoder (SDA). This SDA is then unwrapped and used as the initialization for a Multi-Layer Perceptron (MLP). The MLP is fine-tuned by training it to reconstruct the input at the output layer.

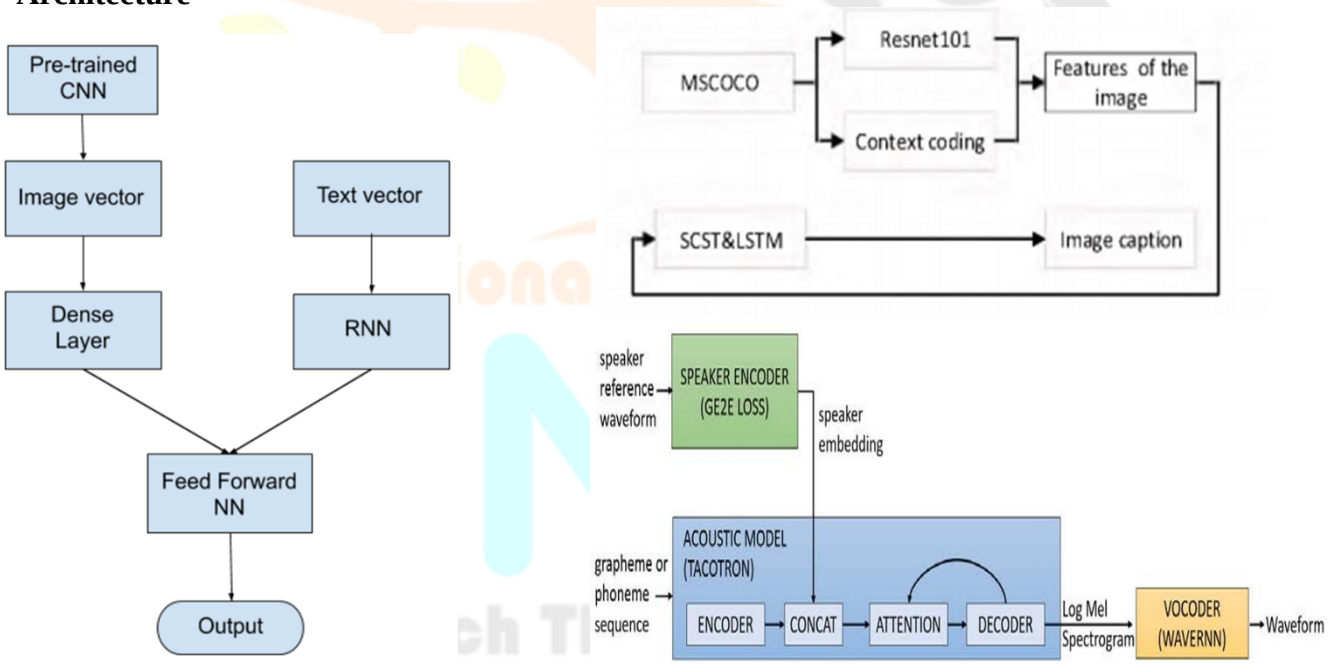
III. Existing System and Proposed System

Existing System
Deep learning-based image captioning methods can generate captions from both visual space and multimodal space. Understandably image captioning datasets have the corresponding captions as text. In the visual space-based methods, the image features and the corresponding captions are independently passed to the language decoder. In contrast, in a multimodal space case, a shared multimodal space is learned from the images and the corresponding caption-text. This multimodal representation is then passed to the language decoder.

Proposed System
The image description based on deep learning mainly uses the structure of codec to complete the image caption task. Further, the effect of image description can be improved through the attention machine or other methods of enhancing the deep learning model.In order to better complete the image caption task, on the one hand, the model needs to use the information in the image more efficiently and rationally, and on the other hand, the problem of inconsistency between model training and testing needs to be solved. In order to solve these two problems, curriculum learning, reinforcement learning and other methods are introduced into the image caption task.

Advantages
Here we use context coding as well as reinforcement learning. and makes full use of image features to describe the image, so that the generated description sentence is more conform to content of the image.

IV. Architecture



V. Problem Description

Image Caption Generation involves predicting a sequence of words(sentence) from an input image. It is essentially an advanced image classification problem with a lot of potential applications like:

- Aiding visually impaired people who rely on sounds and texts to describe a scene
- Targeted marketing on social media applications like Facebook and Instagram
- A slightly (not-so) long term use case would be explaining what happens in a video, frame by frame

We were motivated to select this project after reading through the work of Andrej Karpathy and Marc Tanti. The idea of generating a description given an image was fascinating to us.

VI. Dataset

We are using Flickr8K dataset for image caption generation available in Kaggle. Flickr8K.zip comprises of two folders:

- Flickr8K_Images: Contains a total of 8000 jpeg images of different shapes and sizes. We are using 6000 images for training, 1000 images for testing and 1000 images for validation.
- Flickr8K_TextData: Contains text files describing the images in the train, validation and the test sets. Each image has a total of 5 captions i.e. a total of 40000 captions.

Captions:
a beagle and a golden retriever wrestling in the grass
Two dogs are wrestling in the grass
Two puppies are playing in the green grass.
two puppies playing around in the grass
Two puppies play in the grass

VII. Approach and Methodology

In this project we deal with two types of data namely text and images. Various pre-processing steps are applied to the captions and images as discussed below:

Text Pre-Processing

- Tokenized each caption and converted all the words to lower-case so that the model would not treat words like “Hello” and “hello” differently.
- Removed alpha-numeric characters as they won’t hold much significance in the data
- Removed punctuation marks from the description, as predicting even the punctuation correctly would be a deviation from the main motive of this project and would bring in linguistic complexities which can be treated as a different project.

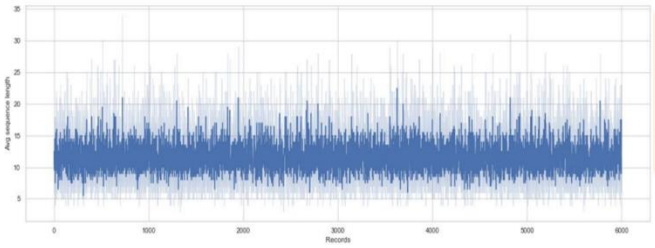
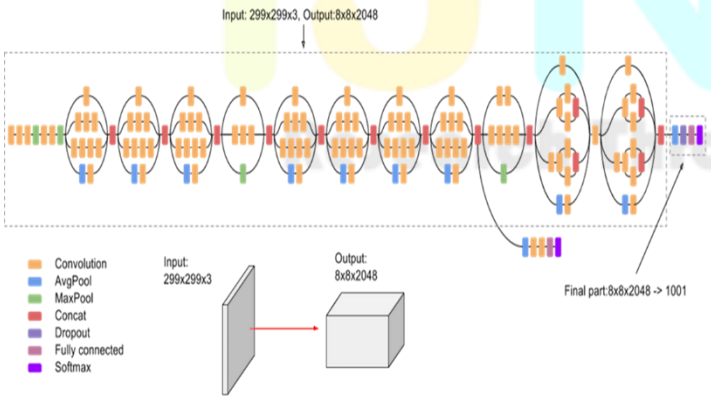


Figure 2: Average description length for each image

Image Pre-Processing and Feature Vector Generation

- Since the input to the neural network should be an image vector, we first resized all the images to a fixed dimension of 299x299x3 using OpenCV, and then converted each image into a fixed-length vector by employing transfer learning. We used a pre-trained InceptionV3 CNN model and VGG16 CNN model for comparison purposes.
- InceptionV3 model was trained on ImageNet dataset for the purpose of classification. Since our aim is to generate a fixed-length vector of size 2048, we removed the last softmax layer from the pertained model and extracted a 2048 length vector for each image.



4.5 Sequential Caption Injection

Image	Partial Caption	Target Word
Image	startseq	a
Image	startseq a	young
Image	startseq a young	boy
.....	.....	.....
Image	startseq a young boy wearing a helmet and riding a bike in a park	endseq

For each image we will train the model by temporally injecting incremental sequences of the description. We split each image-description pair into multiple such incremental sequences. The first sequence uses input description as 'startseq' and first word in the description as the target word. The target word acts as a label and helps training our model. The second sequence uses the description "startseq <first-word>" as the input description and second word in the description as the target word.

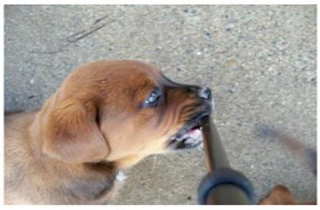
VIII. Result

We used two different pre-trained CNN models to generate the image feature vectors, namely InceptionV3 and VGG16. We did not observe any significant difference in the model results, the reason being these networks were extensively trained on huge image datasets like ImageNet and have very less difference in performance.

We used three different recurrent neural networks to generate captions, namely Simple Recurrent Neural Network, Gated Recurrent Unit and Long Term-Short Term Memory. We observed that LSTM was performing the best out of all the networks. We also used a Feed Forward Neural Network (FFNN) out of curiosity and observed that the whole model behaved as a non-efficient image classifier rather than a generative model.

Actual Caption:
a man wearing a red life jacket is holding a purple rope while skiing

Predicted Caption:
man in white and white and white shorts leash on swing



Actual Caption:
a dog is chewing on metal pole

Predicted Caption:
dog is standing in its mouth

Actual Caption:
a young hockey player playing in the ice rink

Predicted Caption:
chasing player in motorcycle is playing chasing

IX. Future Work

The idea of a neural network generating a sentence with proper grammar by looking at an image is utterly fascinating. We were successfully able to generate some good captions, some funny captions and some not so good captions as seen from the results. The main reason behind the funny captions was due to the model being confused by a predominant presence of a color or object in the image that sidelines the main subject and actions in the
image. This can be fixed using visual attention techniques like soft visual attention or gaussian visual attention that help elevate the importance of the subject and actions in the image.

X. CONCLUSIONS

In the ever-evolving landscape of artificial intelligence, our project has shed light on the immense potential of merging Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) for image caption generation. As we draw this project to a close, several key takeaways and conclusions have emerged.

Firstly, the synergy between CNNs and LSTMs is a powerful force in the realm of computer vision and natural language processing. The use of CNNs for image feature extraction and LSTMs for sequential caption generation has yielded promising results. This architecture has paved the way for machines to not only "see" images but also describe them in a manner that mirrors human understanding.

In conclusion, "Image Caption Generation using LSTMs and Convolutional Neural Networks" represents a significant stride in the journey to bridge the gap between visual and textual understanding. It underscores the transformative potential of AI in transforming how we interpret and interact with images. As we move forward, our project serves as a stepping stone towards more advanced and contextually aware image captioning systems, promising a future where machines truly understand and describe the visual world.

XI. REFERENCES

[1] Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3128-3137).

[2] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3156-3164).

[3] Chen, X., Fang, H., Lin, T. Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft COCO captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325.

[4] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In International Conference on Machine Learning (ICML) (pp. 2048-2057).

[5] Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). DenseCap: Fully convolutional localization networks for dense captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4565-4574).