Journal of Research Administration
Society of Research Administrators International

EXPLORING GENERATIVE AI APPROACHES FOR CRAFTING IMAGE CAPTIONS: A RESEARCH INQUIRY INTO ETHICAL, ENGAGING CONTENT GENERATION

Kabil C.A, Vairachilai.S
School of Computing Science and Engineering, VIT Bhopal University, Kothri Kalan, Sehore, 466114, Madhya Pradesh, India

Abstract
In the age of social media, the demand for high-quality content with engaging captions is paramount for audience en- gagement and community building. This research explores the transformative impact of Generative AI, exceeding initial expectations. The study focuses on CNN and RNN, specifically LSTM, for image detection and caption generation. The model's central purpose is to detect objects in images, predict their relationships, and produce meaningful captions. The novelty lies in enhancing these generated captions through the utilization of an open-source API key, connecting to an ad- vanced Large Language Model (LLM) like Chat GPT-4. This method not only enhances engagement by generating captions that are relevant, informative, engaging and ensures the generation of ethical content by incorporating safeguards against bias, misinformation, and harm.
Keywords: Image captioning, Generative AI, CNN, LSTM, Neural Network, API Key

1. Introduction

In recent times, Generative AI has had a surprisingly posi- tive impact on society, going beyond our initial expectations. Numerous studies have explored its capabilities, ethical consid- erations, and how people interact with it.

In the world of social media, where sharing quality content with appealing captions is important for grabbing attention and good engagement, this research implements the use of Genera- tive AI to create high-quality, engaging, ethical, and interesting captions and hashtags. This initiative is aimed at assisting con- tent creators, especially those passionate about social media, in producing captivating content

Similar to the principles of image recognition, this research employs Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), specifically Long Short-Term Mem- ory (LSTM), to analyze images and generate relevant captions or titles based on the image content. The primary goal is to identify different objects within images, predict their relation- ships, and produce meaningful captions

A key aspect of this research is to further refine the generated captions. This is achieved by sending them to a transformer- based language model like Chat GPT-4 using the open-source API provided by OpenAI. This additional step ensures that the captions and hashtags are not only more interesting but also stick firmly to ethical standards

The model relies on image processing techniques, includ- ing CNN for recognizing objects and patterns within images, and LSTM for crafting coherent and contextually relevant cap- tions. The final step involves utilizing a Large Language Model (LLM) to generate concise and captivating

captions and hash-tags designed for various social media platforms like Instagram, Facebook, and WhatsApp.

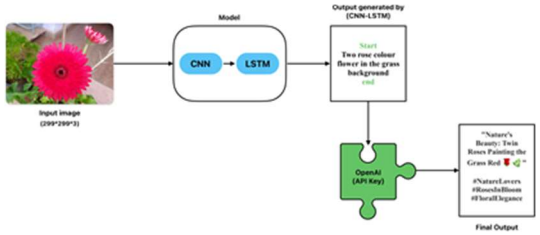


Figure 1: Model Implementation Flow Chart

This model illustrates the deep learning Neural Network ar- chitecture, comprising two main components. The ImageNet dataset is used to train a specific CNN model called Xception. Xception's primary role is to extract features from images for processing visual information. It analyzes the content of im- ages, recognizing objects and patterns within them. The output from CNN serves as the foundation for generating language- based content.

This model uses a language generating RNN in addition to CNN. This RNN takes the visual information from CNN and transforms it into complete sentences, which can be used as captions or descriptive sentences for various purposes. Impor- tantly, these sentences can be further converted into both cap- tions and hashtags, suitable for social media posts.

2. Contribution

The main finding of this research paper is mentioned below:

2.1. Data Cleaning and Preprocessing

1. The process initiates by loading both the text file and the image file into distinct variables.
2. The main task of data cleaning involves removing punc- tuation marks, converting the whole text to lowercase, re- moving stop words and removing words that contain num- bers.
3. Another aspect of Preprocessing the data involves tokeniz- ing our vocabulary with a unique index value.
4. These preprocessing tasks can be achieved by using the Keras.preprocessing module.

2.2. Tokenizing the text

1. Tokenization is the process of dividing a text into individ- ual tokens. In the case of image captions, tokens are often words or subwords.
2. For example, "A cat is sitting on a mat" might be tokenized into the following tokens: ["A", "cat", "is", "sitting", "on", "a", "mat"].

2.3. Ethical Caption Refinement

1. This is achieved by employing a transformer-based lan- guage model such as Chat GPT-4 through the open-source API provided by OpenAI.

2. By combining image processing techniques (CNN and LSTM) with a Large Language Model (LLM), this re- search offers a holistic and comprehensive approach to im- age captioning

2.4. Model Evaluation

1. Using language evaluation metrics like BLEU Score, ME- TEOR, ROUGE, and CIDEr to quantitatively measure the quality and relevance of generated captions compared to human authored captions.

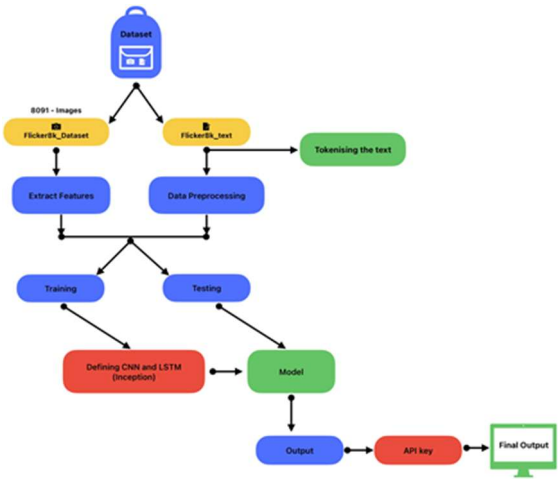


Figure 2: Work Flow Chart

3. Proposed Model

In the context of literature review, it is evident that the traditional CNN-LSTM model exhibits limitations, leading to the generation of captions that may contain inaccuracies and language-related issues, making them unsuitable for popular so- cial media platforms such as Instagram, Facebook, and What- sApp. This can be overcome by creating a promising solution: integrating Generative AI, specifically OpenAI's Chat GPT, by using its API key. This integration serves as a bridge between the LSTM-generated captions and the creation of captivating and engaging captions for social media and various publica- tions. By using the capabilities of Chat GPT, to enhance the quality of the generated captions, ensuring they not only pro- duce accurate information but also possess the intriguing and appealing qualities needed for social media engagement. This innovative approach underscores our commitment to improving the accessibility and utility of AI-generated content, ultimately broadening its applications across diverse platforms.

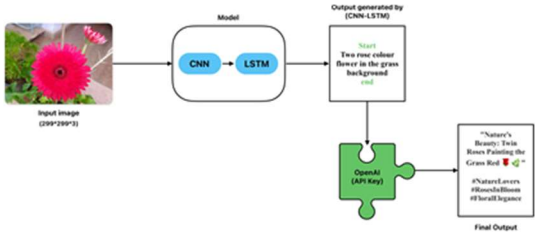


Figure 3: Proposed Model

4. Literature Survey

In the field of artificial intelligence (AI), the task of generat- ing image captions has been a subject of extensive research in recent years. But this research has primarily focused on produc- ing concise and limited-length descriptions for images. What's been notably missing is the ability to connect these descriptions into cohesive short stories, which require the skill of connecting multiple sentences to make a narrative that flows seamlessly. In response to this gap in the field, our research, as outlined in the paper published on August 12, 2021, with the current version dated August 19, 2021, authored by Min, Kyungbok
; Dang, Minh; Moon, Hyeonjoon. [1], introduces an innova- tive framework. This framework employs an encoder-decoder structure to facilitate the generation of short story captions (SS- Cap) by leveraging both a conventional image caption dataset and a meticulously curated corpus of narrative text.

The core of this paper methodology lies in the encoder- decoder architecture, which allows us to seamlessly process im- ages and transform them into comprehensive, contextually co- herent story captions. By Capturing the synergy between visual inputs and the words of the manually collected story corpus, this approach empowers the creation of image-driven short sto- ries that are not only meaningful but also narratively engaging.

This research extends the boundaries of image captioning by using encoder-decoder method, more immersive narratives in- spired by visual content. This innovation holds great promise for diverse applications, from enhancing storytelling in multi- media to advancing content generation, and it underscores the potential of AI to bridge the worlds of visual and textual data in novel and creative ways.

In the research outlined in the paper by Megha J Pan- icker, Vikas Upadhayay, Gunjan Sethi, and Vrinda Mathur (published in January 2020). [2], the focus is on two dis- tinct deep learning models for image captioning: the Con- volutional Neural Network-Recurrent Neural Network (CNN-RNN) Based Image Captioning and the Convolutional Neural Network-Convolutional Neural Network (CNN -CNN) Based Image Captioning. In the CNN-RNN Based framework, the methodology is structured as follows. Convolutional Neural Networks (CNNs) are employed for image encoding, and Re- current Neural Networks (RNNs) are responsible for the de- coding process. To break it down, CNNs play a crucial role in converting images into vectors, effectively referred to as image features. These image features are subsequently fed as inputs into the Recurrent Neural Networks. This approach uses the strength of CNN in their ability to extract meaningful image features, capturing essential visual information from the im- ages. These features, in the form of vectors, can be used as the foundation for the subsequent storytelling process. The RNNs, known for their sequential data processing capabilities, inter- pret these vectors and generate coherent and contextually rele- vant captions. The synergy between CNNs and RNNs, in this context, marks a significant advancement in the field of image caption generation. This innovative approach ensures that im- ages are not described but narratively

written, opening doors to applications in content generation, accessibility, and multime- dia storytelling.
By making use of this powerful combination of neural net- works, the research authors find an efficient means of convert- ing images into expressive narratives, enhancing the ways of interacting with and understanding visual content.

In the research article by Shikha Gupta (published on De- cember 18, 2021, and last modified on September 5th, 2023). [3], the authors delve into the application of two widely rec- ognized deep learning methods for image processing and cap- tion generation. These methods, Convolutional Neural Net- works (CNN) and Recurrent Neural Networks with Long Short- Term Memory (RNN-LSTM), are well-established for their ef- fectiveness in this domain. The main part of how they do it involves two steps by training a CNN model, specifically Xcep- tion, using the ImageNet dataset. Xception plays a pivotal role in extracting essential features from images, essentially cap- turing the visual details that form the foundation of generat- ing meaningful captions. These extracted features serve as a rich source of information. Afterwards, the authors employ an LSTM model in the second step. This model takes the image features extracted by Xception and works in generating coherent image captions. The LSTM's inherent sequential data pro- cessing capabilities make it good for this task, as its crafts sen- tences that not only describe but also narrate the details of the images. What sets this particular paper apart is its utilization of the Flickr dataset, comprising a substantial pool of approxi- mately 80,100 images for both training and testing phases. The authors have used supervised learning, a robust technique, to ensure the most favorable outcomes. The research findings in- dicate an impressive accuracy rate of over 70%.
Furthermore, the authors propose that to achieve even higher accuracy, it is advisable to expand the dataset by incorporating more than 100,000 images for training and testing. This rec-ommendation underscores the importance of sample data for training deep learning models effectively. In conclusion, this re- search underscores the effectiveness of the CNN-LSTM frame-work in the world of image caption generation. It offers not only a detailed guide for implementation but also valuable insights into dataset size and its impact on model performance.

In the research paper by Aishwarya Maroju, Sneha Sri Doma, and Lahari Chandarlapati, published on September 20, 2021,[4] the authors present a novel approach to image caption gen- eration by utilizing the power of ResNet50 and LSTM net- works. At the core of their methodology lies a two-fold pro- cess designed to improve the efficiency and accuracy of im- age understanding and caption creation. To begin, they employ ResNet50, a pre-trained deep learning model renowned for its prowess in feature detection. Unlike traditional CNNs, which are typically trained from scratch, ResNet50 is a pre-trained model. Comprising 50 layers of convolutional neural network architecture, the choice of ResNet50 is considered by its high performance and accuracy in image classification, as well as its ability to extract high-quality image features. When compared to conventional CNNs and even VGG architectures, ResNet stands out as a superior performer. Following the feature ex- traction phase, the authors employ LSTM networks to extract image

captions efficiently. LSTM networks excel in processing sequential data, making them a great fit for generating coherent and contextually relevant captions based on the image features extracted by ResNet50. This approach represents a significant step forward in image captioning, as it combines the strengths of ResNet50's feature extraction capabilities with the sequence generation of LSTM networks. The combination of these two components produces higher accuracy and efficiency in the im- age captioning process.

The findings of this study underscore the potential of ResNet50 as an asset in image understanding and caption- ing, highlighting its superiority over traditional CNN and VGG models. This research contributes to the evolving of deep learn- ing and image analysis, promising more accurate and contex- tual image captions.

5. Limitations of Current Approaches

As seen in the literature survey, there are some limitations in the existing model. Each existing model has its own disad-

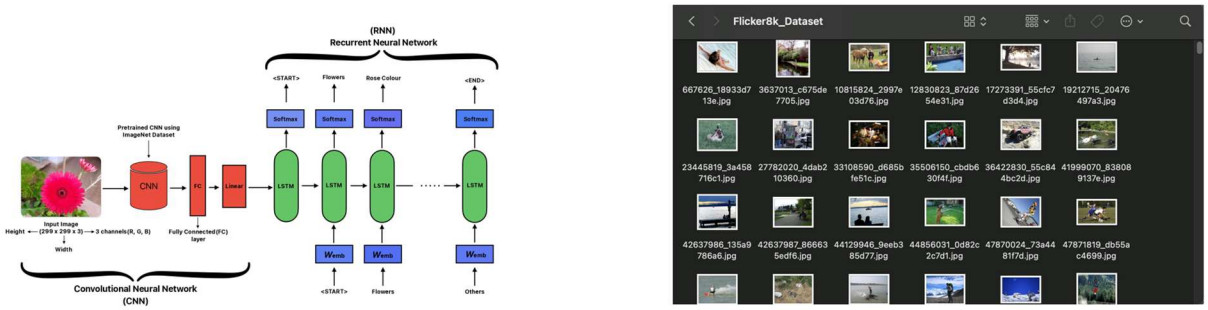


Figure 4: Existing Model

vantage, making the model less efficient and less accurate when the results are generated. The observed drawbacks in all the existing models are as follows:

The captions generated by each of the models in this research paper exhibit a recurring issue—broken or inaccurate English phrasing that makes them unsuitable for social media sharing or publication. These captions often lack the clarity and language accuracy necessary for effective communication. The challenge of generating fluent and contextually precise captions is a crit- ical one in the field of AI-driven image captioning. While the models showcase the potential for automation in this creative domain, their output falls short in achieving the level of lan- guage fluency and accuracy required for practical use.

Addressing this limitation represents a crucial area for fur- ther research and improvement. Improving the language qual- ity of generated captions is not only integral to effective image captioning but also holds the key to unlock the full potential of AI in applications, from social media content generation to accessibility support and beyond. Researchers in the field are actively working to refine these models, working on captions that are not only contextually relevant but also grammatically sound and coherent.

6. Dataset Description

The dataset collected from Kaggle by Aditya Jain.

The Flickr-8K dataset serves as the foundation for training image caption generators. The dataset's substantial size, ap- proximately 1GB in total.

This dataset has a repository of image names along with their corresponding captions. In total, there are 8,091 images stored within the 'Flicker8k-Dataset' folder. The textual descriptions for these images can be found in the text files within the 'Flickr- 8k-text' folder. The FLICKR 8K data set consists of 8,091 im- ages in which 6,050 images can be used for training the deep learning model and 1,041 images for development and 1000 images for testing the model. Flickr Text data set consists of five captions for each given image which describes the actions performed in the given images.

In Figure: 4, Simplifies visual representation of the Flicker8k dataset, plays an important element in our research for practical testing and experimentation.

Figure 5: Flickr8k-Dataset

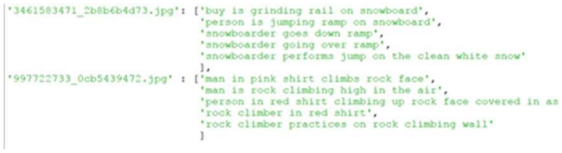


Figure 6: Flickr8k-Text

In Figure: 5, Show that All the image captions can be found in the "Flickr 8k.token" file within the "Flickr-8k-text" folder. If you examine this file closely, you'll notice a specific format. Each image and its corresponding caption are separated by a new line. Additionally, there are five captions for each image, numbered from 0 to 4.

7. Xception

Xception is an improved version of Inception that requires fewer computational resources. The ImageNet dataset is uti- lized to train the Convolutional Neural Network (CNN) model known as Xception, specifically designed for image feature ex- traction. Xception plays a pivotal role in capturing intricate patterns and characteristics from images.

7.1. Working process

1. Initializing the Xception model. (include-top=False) means that the top layer is removed, and the pooling layer is set to average, resulting in a single value for each chan- nel that summarizes the spatial information of the entire feature map.
2. Initializing a dictionary named features to store the ex- tracted features. Where each image filename will be as- sociated with its corresponding feature vector.
3. Resizing the image to 299x299 pixels. Pre-trained models frequently have input size limitations.

4. Adding a new dimension to the picture array. Because pre- trained models frequently employ batch input, this line en- sures that the image is regarded as a batch of size 1
5. Pre-processing the image and adjusting the input image to fit the model's expected input format.
6. Normalize the pixel values of the image 0 and 1s.
7. Using the Xception model to predict features from the pre- processed image and return the features.

8. CNN

CNNs are trained using large collections of diverse images. CNN can learn rich feature representations for a wide range of images. Convolutional Neural Networks (CNNs) are specialized deep learning models designed for processing data with a 2D matrix-like input shape, particularly well-suited for image representation.

8.1. Working process

1. The input layer is defined with a shape of (2048,), indicat- ing it takes an input with 2048 features. This represents the output of pre-trained image classification CNN (Xcep- tion).
2. A Dropout layer is added with a dropout rate of 0.5. Dropout is a regularization technique that randomly sets a fraction of input units to zero during training, which helps prevent overfitting.
3. The output of the Dropout layer is then passed through a fully connected Dense layer with 256 nodes and ReLU activation. This layer acts as a narrow passage, reducing the dimensionality of the features from the CNN.

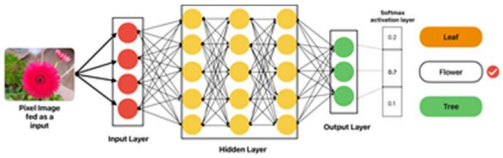


Figure 7: Convolution Neural Network Architecture for image classification

9. LSTM

Long Short-Term Memory networks (LSTMs) are a special- ized type of Recurrent Neural Networks (RNNs) designed to capture long-term dependencies in data. LSTMs excel at retain- ing information over extended periods, achieved through the use of "gates" that control their behavior.

9.1. Working process

1. The input layer is defined with a shape of (max-length,), indicating it takes sequences of integers as input. The max- length parameter is the length of the input caption
2. The input caption sequences are passed through an Em- bedding layer with a vocabulary size of vocab-size and an output dimension of 256.
3. Embedding layer converts integer indices into dense vec- tors of fixed size and is capable of handling variable-length sequences.
4. mask-zero=True parameter indicates that the model should mask the zero-valued entries in the input.

5. A Dropout layer is added with a dropout rate of 0.5. Dropout helps prevent overfitting by randomly setting a fraction of input units to zero during training.

6. The processed sequence data is then passed through an LSTM layer with 256 nodes. Capture long-term depen- dencies in sequences, making it suitable for generating captions based on the context of previous words.

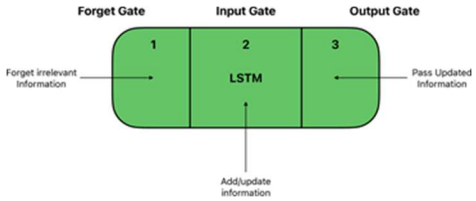


Figure 8: Long Short-Term Memory network architecture

10. API Key

An API key (Application Programming Interface key) is a code or token that is used to identify and authenticate appli- cations or users when making requests to an API (Application Programming Interface). APIs are used to enable communi- cation and interaction between different software systems, ser- vices, or platforms.

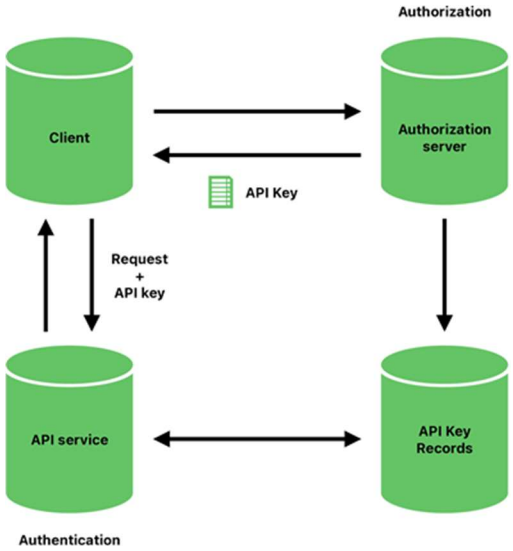


Figure 9: Authorization and Authentication API key

In this research paper, the API key used is from Ope- nAI (ChatGPT). After generating captions from the CNN and LSTM model, a request is sent to OpenAI through the API key provided by them. This is done by using the power of their service to make our captions more interesting and fluent.

11. Result Analysis

By observing testing Figures(9), (10), (11) these are the out- put caption generated by the CNN + LSTM and the fused re-

sult with the API key. The caption generated by the API key is more fluent and ethical compared to the caption generated by the CNN + LSTM model alone.

11.1. Testing

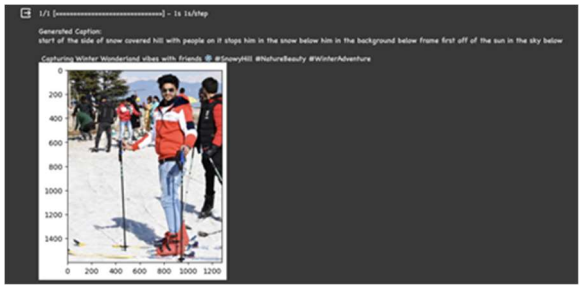


Figure 10: Caption for Test 1

S.No	Caption Without GenAI	Caption With GenAI
1	Start of the side snow covered hill with people on it stop him in the snow below	Capturing Winter Wonderland vibes with friends
2	Start of fan mascot leading on the ground with his arm around him and standing on the grass	Feeling the excitement in the air as the fan mascot leads the way to the grass
3	Start fan image of wine in sling and back bandanna on it is standing on the sidewalk with his flung out	Ready for the urban adventure

Table 1: Result

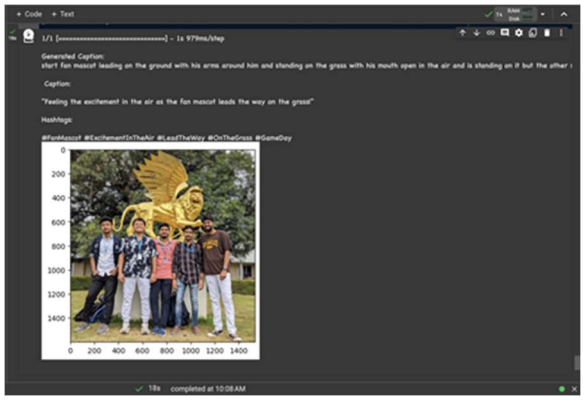
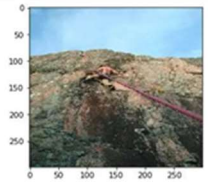


Figure 11: Caption for Test 2



```
**********************************************************
Predicted Caption :

 a man climbs a cliff face
**********************************************************
Referances :
 a man climbing a mountain
 a man climbs a mountain
 a man is climbing the side of a mountain
 a shirtless man climbs up a steep mountain
 a young white man is climbing a mountain with a rope as a guide
**********************************************************

BLEU Score :
unigram = 0.6666666666666666
bigram = 0.6
cumulative = 0.6324555320336759
```

Figure 13: Bleu Score

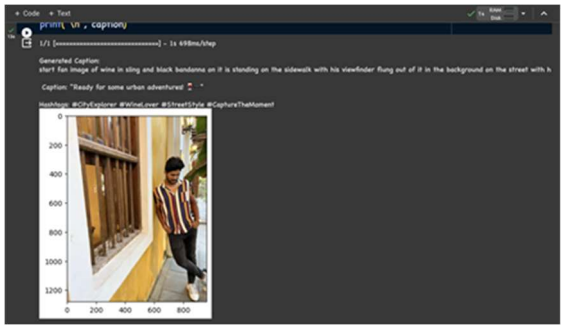


Figure 12: Caption for Test 3

11.2. Bleu score

The accuracy is calculated by employing the BLEU score; a metric widely used in natural language processing tasks. The BLEU score quantifies the similarity between the generated output and reference text, providing a numerical measure of the model's performance producing language accuracy and contex- tually relevant results.

$$\text{BLEU} = \text{BP}^* \exp\left(\sum_{n=1}^{N} w_n \, logp_n\right)$$

BP: Brevity Penalty

- pn: Precision for n-grams
- wn: Weights for different n-grams
- N: Maximum n-gram size considered

1. The performance of the model can be improved by training it on a larger dataset and hyperparameter tuning
2. Figure (13), shows the unigram bleu scores 66 percent pre- dictions.
3. Prediction is good if all the BLEU scores are high

This blue score determines whether or not the caption gen- erated by this model accurately represents the provided image.

However, with an OpenAI API Key and a higher blue score, this model can attain greater than 90% accuracy.

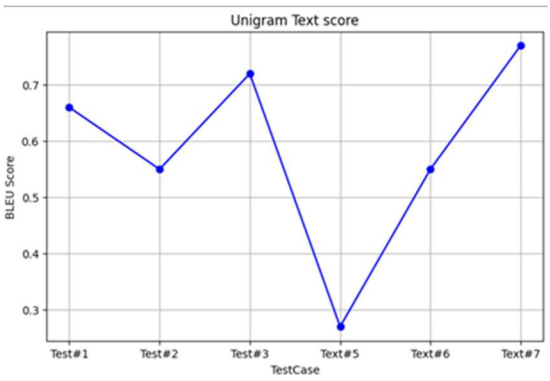


Figure 14: Unigram Score

The above figure (14). Represent the unigram score of 7 dis- tinct CNN and LSTM model test scores. The Bleu scores range from 30 to 79 percent. This test is critical for the LLM model to generate meaningful captions and hashtags.

12. Conclusion

In conclusion, the image undergoes processing via a blend of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. The refined output is then for- warded to the OpenAI API key, improving the generated cap- tions for ethical, engaging content for a broad social media au- dience. This collaborative combination of neural networks and advanced language models establishes a robust image classifi- cation model, consistently delivering ethically crafted captions adaptable to diverse contexts.

13. Reference

1. Min, Kyungbok ; Dang, Minh ; Moon, Hyeonjoon. “Deep Learning-Based Short Story Generation for an Image Us- ing the Encoder-Decoder Structure” published on 2021, Vol. 9. pp. 113550-113557.
2. Megha J Panicker, Vikas Upadhayay, Gunjan Sethi, and Vrinda Mathur ”Image Caption Generator” published in January 2020.
3. Shikha Gupta “Step by Step Guide to Build Image Caption Generator using Deep Learning” published on 2021.
4. Aishwarya Maroju, Sneha Sri Doma, and Lahari Chandar- lapati “Image Caption Generating Deep Learning Model”, published on September 20, 2021
5. M. Sailaja; K. Harika; B. Sridhar; Rajan Singh; V. Charitha; Koppula Srinivas Rao Image Caption Genera- tor using Deep Learning - Published On 2020 2nd Interna- tional Conference on Advances in Computing, Communi- cation Control and Networking
6. Sandra Kublik, Shubham Saboo “GPT-3: Building In- novative NLP Products Using Large Language Models”- O’Reilly Media, Year: 2022
7. MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shi- ratuddin, Hamid Laga “A Comprehensive Survey of Deep Learning for Image Captioning” - date publication on 2019
8. Risto Miikkulainen, Jason Liang, Elliot Meyerson, Aditya Rawal, Daniel Fink, Olivier Francon, Bala Raju, Hormoz Shahrzad, Arshak Navruzyan, Nigel Duffy, Babak Hodjat. “Evolving Deep Neural Networks” - date of publication on 2018.
9. N. Indumathi; R.J. Divyalakshmi; J. Stalin; V. Ramachan- dran; P. Rajaram. “Apply Deep Learning-based CNN and LSTM for Visual Image Caption Generator” - IEEE Ac- cess Published: 2023
10. Oriol Vinyals et al., “Show and tell: A neural image cap- tion generator”, Computer Vision and Pattern Recognition (CVPR) 2015 IEEE Conference on, 2015.
11. Xu Jia, Efstratios Gavves, Basura Fernando, Tinne Tuyte- laars, “Guiding the Long-Short Term Memory Model for Image Caption Generation” Publisher: IEEE on 2016.

12. O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and tell: A neural image caption generator", Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 3156-3164, Jun. 2015.
13. C. Wang, H. Yang and C. Meinel, "Image captioning with deep bidirectional LSTMs and multi-task learning", ACM Trans. Multimedia Comput. Commun. Appl., vol. 14, no. 2, pp. 1-20, May 2018.
14. S. Gaur, "Generation of a short narrative caption for an image using the suggested hashtag", Proc. IEEE 35th Int. Conf. Data Eng. Workshops (ICDEW), pp. 331-337, Apr. 2019.
15. Ayan Ghosh, Debarati Dutta, and Tiyasa Moitra. "A Neural Network Framework to Generate Caption from Images". Springer Nature Singapore Pte Ltd., pages 171–180, 2020.
16. Yuting Su, Yuqian Li, Ning Xu, and An-An Liu. "Hierar- chical deep neural network for image captioning". Neural Processing Letters, 52(2):1057–1067, 2020.
17. Akash Verma, Arun Kumar Yadav, Mohit Kumar, Divakar Yadav "Automatic Image Caption Generation Using Deep Learning". Published Research gate in January 2022.
18. Chaithra V, Charitra Rao, Deeksha K, Shreya. "Image cap- tion generator using deep learning". Published IJEAST in June 2022.
19. Ch. Sneha, B. Premanvitha, B.Shanmukh, Kavitha Chadu- vula. "Image caption generator using deep learning" pub- lished Research gate in January 2020.
20. T Swarnim, S Ravi. "Image Caption Generator Using CNN and LSTM". Published IRJET in August 2021