JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

SpeechSnap- Describing Images with voice and multi-language support features

[1]Prof. Anagha Chaudhari, [2]Mr. Aaditya Jadhav, [3]Mr. Amitesh Deshmukh, [4]Mr. Uddhav Patil, [5]Mr. Sujit Deore

[1]Professor, [2,3,4,5]Student
[1]Computer Engineering Department,
[1]Pimpri Chinchwad College of Engineering, Pune, India

Abstract: In this study, we have developed an Image Captioning model leveraging InceptionV3 for image classification and extracting relevant features. The model incorporates Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) networks, to generate descriptive captions in English for the images processed. Additionally, we have employed GloVe word embeddings to enhance the linguistic representation within the model. The system classifies images by identifying elements within them and generates meaningful statements to describe the fed images. The generated captions hold potential applications in various domains, such as social media posts, blog articles, accessibility for visually impaired individuals, e-commerce, advertising, surveillance, medical diagnosis, education, among others. The model is trained using Categorical Cross Entropy Loss, aiming to optimize the accuracy of caption predictions. As the model evolves, it is anticipated that it will provide more accurate descriptions for images of diverse formats and content, contributing to its versatility and utility across a wide array of applications.

Index Terms: Image Captioning, InceptionV3, Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), GloVe, Categorical Cross Entropy Loss, Gated Recurrent Unit(GRU), Encoder, Decoder, Image captioning.

I. INTRODUCTION

The process of creating natural language descriptions for images is known as image captioning. It is an interdisciplinary field where machine learning, natural language processing, and computer vision are combined. Due to its potential use in a number of industries, including e-commerce, social media, surveillance, and education, image captioning has been a focus of active research for a number of years.

Recurrent Neural Networks (LSTM) and InceptionV3 are two deep learning[A. Hani et al., 19] models that have recently demonstrated promising results in image captioning. These models are able to recognize objects, scenes, and relationships in pictures [Shetty et al., 23] and produce captions that accurately describe the image's subject matter [H. Inaguma et al., 19]. However, there are still a number of difficulties in image captioning, including dealing with uncommon words, and producing a variety of captions.

In this study, we suggest an image captioning model that addresses these issues and raises the precision and variety of the captions [C. S. Kanimozhiselvi et al., 22] that are produced.

II. Motivation

The goal of the study concerning picture captioning is to create algorithms that can provide textual descriptions of an image's content automatically and provide the features like multi language [Gallardo García et al., 21] support and voice support for generated captions.

Technology for image captioning may assist in communication between computers and people. It may be used, for instance, to create intelligent search engines that can locate photos based on textual descriptions supplied by users.

Automated Analysis: The automated analysis of visual data can be aided by image captioning [Ondeng et al., 23] technologies. It may be used, for instance, to monitor natural disasters or to analyze satellite photos to spot changes in land usage.

Personalization: In many applications, image captioning [Al-Malla et al., 22] technology may be utilized to customize the user experience. For instance, it may be used to produce captions for individual photo collections or to personalize the descriptions of photographs in a social media feed.

Scientific study: Captioning of images might benefit scientific research in disciplines including machine learning, natural language processing, and computer vision[Farhadi et al., 10].

III. Literature Review

In paper [Oriol Vinyals et al., 15**],"Show and Tell: A Neural Image Caption Generator":** It presented a deep learning strategy for picture captioning. The study suggested a model architecture that employs a recurrent neural network (RNN) to produce the caption and a convolutional neural network (CNN) to encode the picture. The article also established the BLEU score, a new assessment measure for picture captioning, and showed cutting-edge results on a number of benchmark datasets. The discipline of picture captioning still relies heavily on the application of the deep learning and transfer learning techniques discussed in this study.

In paper [H. Inaguma et al., 19**],"Multilingual end-to-end speech translation":** The paper presents a unique method to multilingual end-to-end speech translation (ST), where a universal sequence-to-sequence architecture is used to convert speech inputs in many source languages straight into destination languages. Although multilingual models have demonstrated potential in the fields of machine translation (MT) and automated speech recognition (ASR), our study is the first to apply them to the end-to-end ST problem. The paper shows the advantage of multilingual end-to-end ST models over their bilingual counterparts in both one-to-many and many-to-many translation scenarios through trials using publicly accessible data. The study also assesses how well multilingual training generalizes to low-resource language pairings through transfer learning. To encourage more study in this developing area of multilingual ST, the authors have made their databases and their codes available to the general public.

In paper [C**.** S. Kanimozhiselvi et al., 22**],"Image Captioning Using Deep Learning":** The image captioning problem is a rapidly expanding field of study that always needs better outcomes, and that is what this paper tries to address. To improve caption generation accuracy, the authors recommend a model that combines different Convolutional Neural Network (CNN) architecture configurations with Long Short-Term Memory (LSTM) networks. They investigate using three different CNN combinations (Inception-v3, Xception, and ResNet50) to extract features from images and combine them with LSTM to provide pertinent captions. The Flicker8k dataset is used for training, and the model's accuracy is used to determine which CNN-LSTM combination performs best overall. The work advances picture captioning algorithms and offers insights for future research in this subject by experimenting with various structures.

In paper [D. Harwath et al., 18**],"Vision as an Interlingua: Learning Multilingual Semantic Embeddings of Untranscribed Speech":** The process of developing neural network embeddings for voice waveforms and natural images while concentrating on content descriptions and avoiding the use of traditional speech recognition technologies or language transcriptions. This study broadens the investigation to include languages other than English, notably Hindi and English, in contrast to earlier research that was primarily done in English. The authors show off the adaptability of their method by successfully using a consistent model design for both languages. Moreover, they demonstrate that training a bilingual model simultaneously on Hindi and English yields better performance than training a monolingual model. In order to demonstrate the models' capacity for semantic cross-lingual speech-to-speech retrieval, the paper's conclusion highlights their potential for overcoming linguistic obstacles in the fields of language processing and vision.

In paper [Calgary, AB et al., 18**],"Multi-Language Support in TouchCORE":** The purpose of this paper is to provide TouchCORE 8, a multi-touch enabled modeling tool that makes it easier to create reusable and scalable models. TouchCORE 8 now supports several languages via language plug-ins and views, in contrast to earlier versions that were restricted to supporting class, sequence, and state diagrams. The method by which language designers incorporate modeling languages into the TouchCORE architecture is described in the paper, emphasizing the role that perspectives play in maintaining consistency between various models. The tool's general navigation features and split-view GUI support, which let users explore and compare many models at once while preserving inter-model consistency mappings, are also covered in the article. All things considered, TouchCORE 8 is a major step forward in terms of modeling tool features, providing better usability and flexibility for modeling activities.

In paper [M. Schiedermeier et al., 21**],"Image Caption Generation Using A Deep Architecture ":** The difficult task of creating a brief natural language description for a picture—known as image captioning—is discussed by the researchers. In order to extract features from photos and generate descriptive text based on these data, their suggested model combines convolutional neural networks (CNNs) and recurrent neural networks (RNNs) in a hybrid fashion. The process of creating captions is improved by the addition of an attention mechanism. The MSCOCO database was used for the model's evaluation, which produced encouraging and competitive outcomes. This work demonstrates how well it works to combine machine learning, natural language processing, and computer vision approaches to handle the challenging task of image captioning.ng activities.

IV. Model Methodology

The different algorithms used in the image captioning process are shown below – *Figure 1* [Shikha Gupta 23].

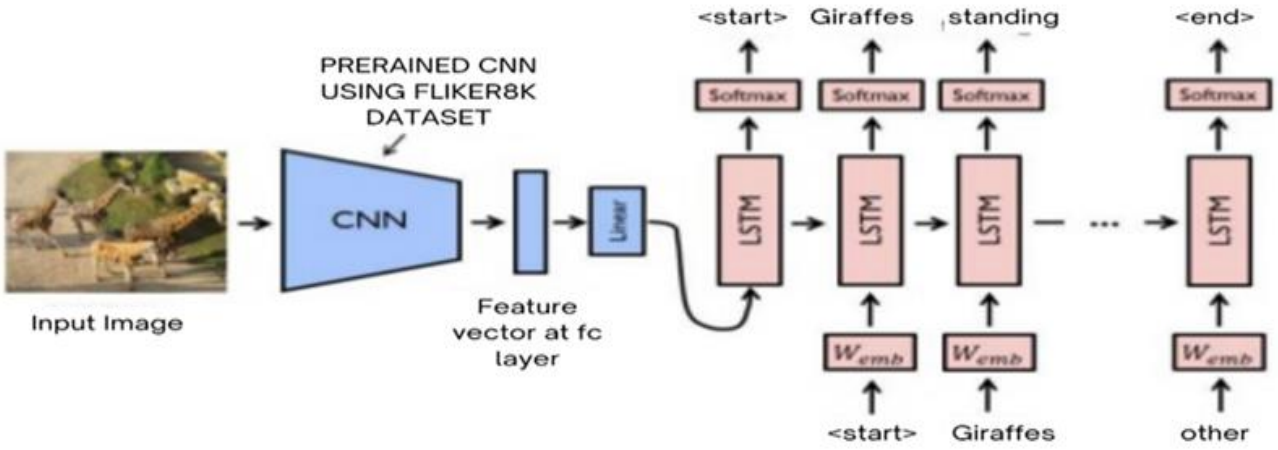

Figure 1: Model Working

The flow of *Figure 1* is described below:-

Data collection: Collect an image from the user for generating captions.

Data preprocessing: Preprocess the images by resizing them to a consistent size, applying data augmentation techniques to increase the dataset size and reduce overfitting, and tokenize the captions.

Feature extraction: A pre-trained InceptionV3 neural network, renowned for its image classification prowess, acts as the visual interpreter [Shetty et al., 23]. Just like a keen observer, it dissects the image, extracting its key features and characteristics. But instead of classifying what it sees, it captures these features as a numerical representation, ready to be combined with the words.

Text embedding: Each word in the caption holds meaning and connection, and GloVe, a word embedding technique, acts as the translator. It converts each word into a numerical vector, capturing its semantic relationships with other words. Imagine these vectors as colorful brushes, each representing a unique word-concept.

Model architecture: The InceptionV3 extracts visual features from the image, encoded into a 2048-dimensional vector [Shetty et al., 23]. This vector is fed into the GRU, which uses a "startseq" token to initiate caption generation. At each step, the GRU predicts the next word based on the encoded image and the previously generated caption, expanding the caption until it reaches an "endseq" token. Greedy search and beam search are used to explore different captioning paths, generating one or multiple final captions. Overall, the model architecture leverages deep learning for image understanding and language generation, offering an approach to image captioning.

Deployment: Deploy the model in a production environment, such as an application or web service, to generate captions for new images.

V. System Design

5.1 Proposed System Architecture / Block Diagram

The below *Figure 2* is an architecture diagram which shows the flow of our model that allows user to upload images and view caption and listen the generated captions in multiple languages.

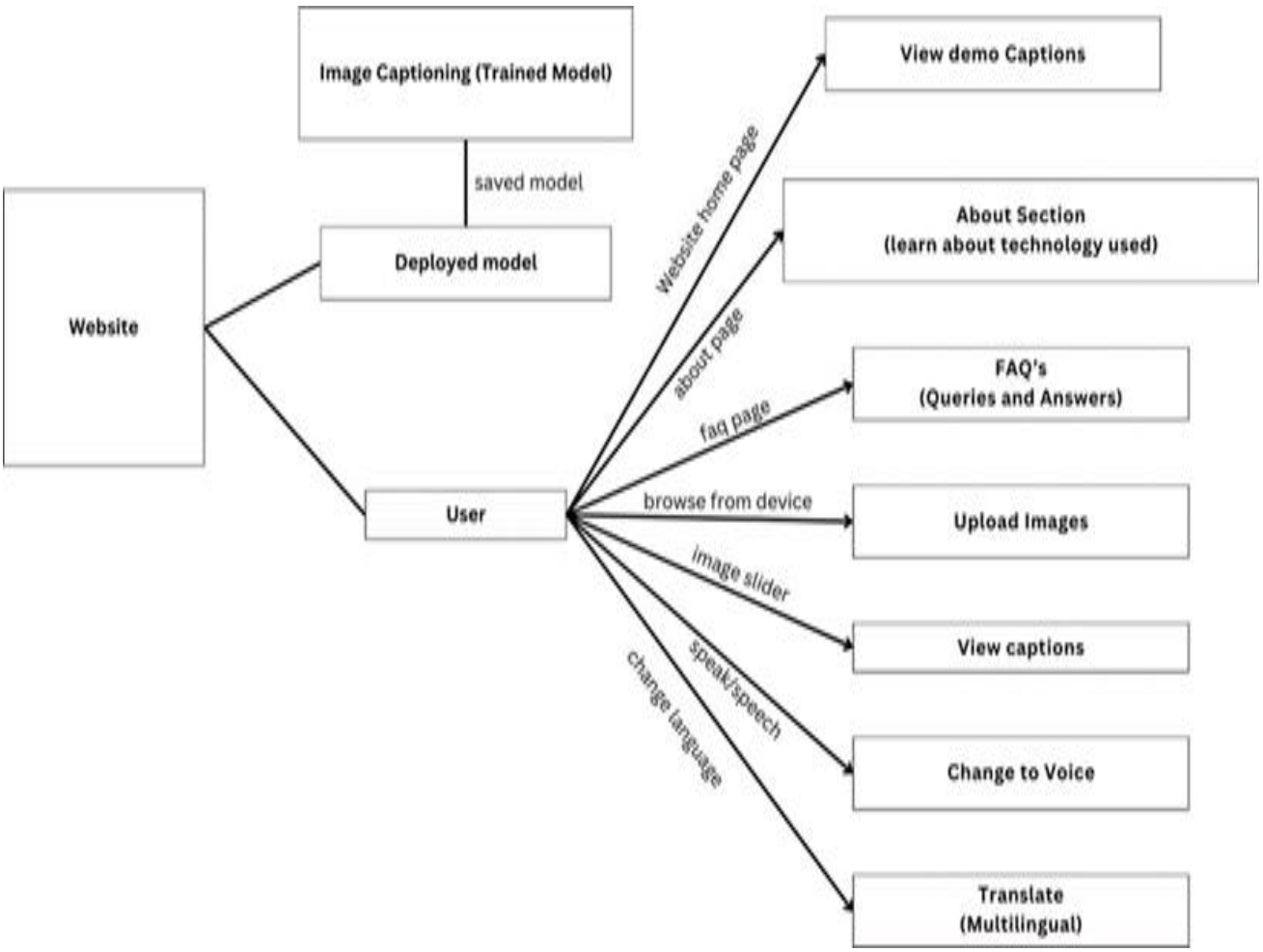

Figure 2: Proposed System Architecture

The *Figure 2* illustrates image captioning and converts it into multiple languages. Users upload images, and a pre-trained model generates captions. The system then displays the caption in text and offers additional features like voice playback and multilingual translation. FAQs and information about the technology are provided for user support. This user-friendly approach allows easy access to image descriptions.

VI. Dataset

The Flickr8k dataset is a popular benchmark for image captioning research. It consists of 8,000 images collected from Flickr, each paired with five human-written captions in English.

- Size: 8,000 images with 5 captions each, totaling 40,000 captions.
- Content: Images cover a wide range of everyday objects, scenes, and activities.
- Captions: Each caption describes the image content in natural language.
- Image format: JPEG format.
- Caption format: Text file.

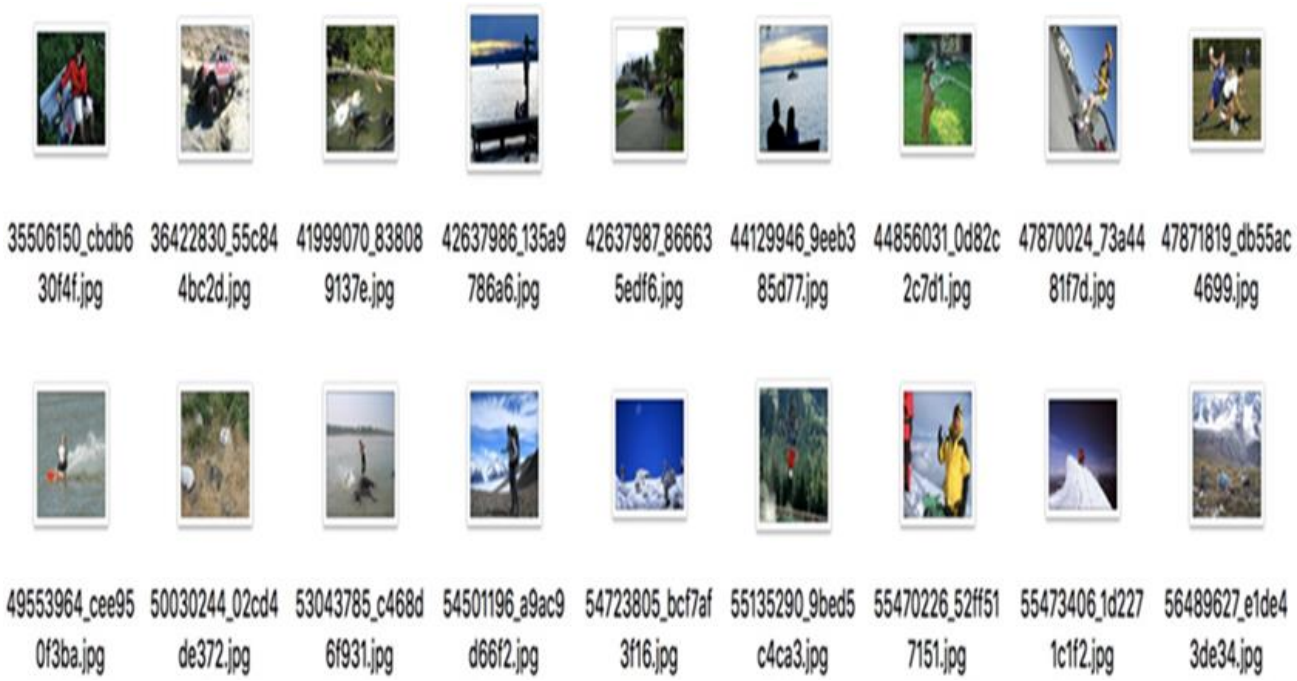

Figure 3: Dataset

The above *Figure 3* represents the example of images present in flickr8k dataset. This dataset was downloaded from Kaggle [Kaggle-ADITYAJN105 2].

VII. Working Of Model

The below *Figure 4* is working on a model which shows image captioning system and allows user to upload image and view and listen the three generated captions of that image.

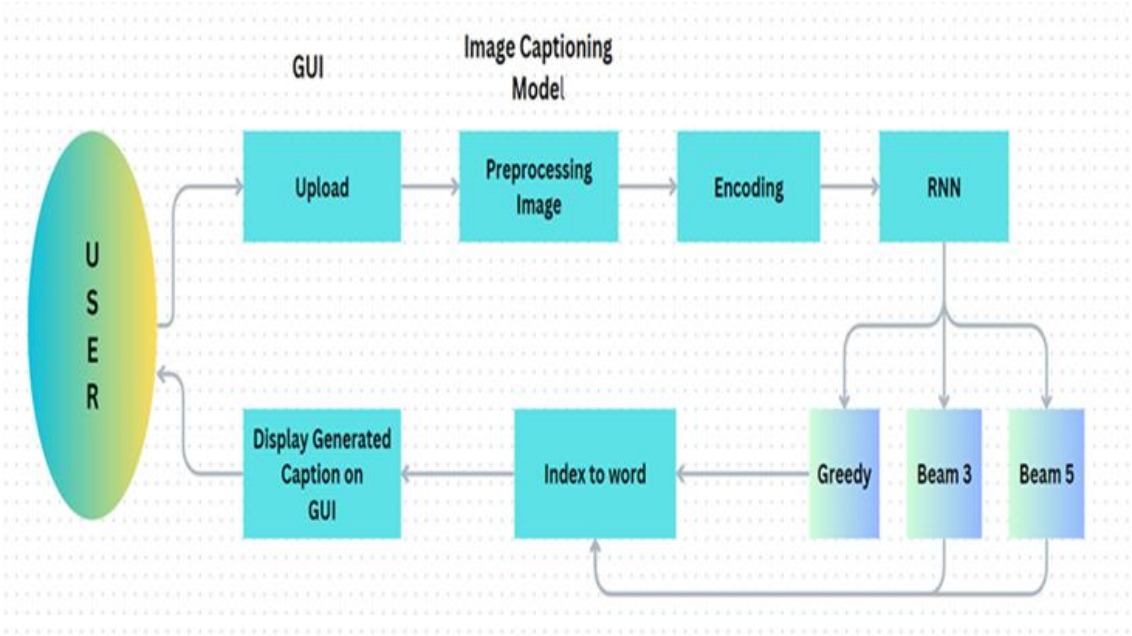

Figure 4: Working Of Model

Deep Dive into the Image Captioning Model (from gui.py)

The provided code delves into the fascinating world of image captioning, utilizing a deep learning model to generate textual descriptions of uploaded images. Here, we embark on a journey to unravel the intricate workings of this model, understanding the key steps involved in transforming visual information into a captivating narrative.

Setting the Stage: Preprocessing and Encoding

The journey begins with an image uploaded by the user. This raw visual data undergoes preprocessing to ensure compatibility with the model. The preprocess_img` function resizes the image to 299x299 pixels, a standard size often preferred by deep learning models like InceptionV3. This prepares the image for efficient analysis by the network.

Next, the `encode` function extracts visual features from the preprocessed image. It relies on the pre-trained InceptionV3 model [Shetty et al., 23], which has already learned to identify patterns and relationships within images. This powerful encoder converts the image into a 2048-dimensional vector, capturing the essence of the visual content in a more abstract and meaningful form.

Bridging the Gap: From Features to Words

With the image's essence encoded, the task now shifts to translating it into human language. This crucial step is handled by a recurrent neural network (RNN), specifically a Gated Recurrent Unit (GRU) model. However, before feeding the encoded vector directly to the RNN, we need some context.

The model leverages a special "startseq" token, signifying the beginning of the caption. This token is incorporated into the encoded vector, providing valuable information about the caption's intended starting point.

Now, the concatenation or integration between image and language truly begins. The RNN receives the encoded vector with the "startseq" token at each time step. Based on this information, it predicts the next word in the caption, continuously updating its internal state to capture the emerging context. This iterative process unfolds until the model predicts the "endseq" token, indicating the completion of the caption.

Greedy Search: This method takes a simpler approach, selecting the word with the highest probability at each time step. While fast and efficient, it can sometimes get stuck in local optima, generating captions that may not capture the full complexity of the image.

Beam Search: This method explores a wider range of possibilities by considering multiple caption candidates simultaneously. At each time step, it retains the top "beam index" most likely caption continuations, fostering diversity and potentially leading to more creative and informative captions.

Decoding the Message: From Vector to Text

Once the caption generation process reaches the "endseq" token, the predicted word sequence needs to be transformed back into human-readable language. This is achieved by mapping the predicted word indices back to their corresponding words in the vocabulary using the `ixtoword` dictionary.

The Final Chapter: Presenting the Results

With the caption generated, the user is presented with the textual description alongside the uploaded image. This allows them to compare the model's interpretation with their own understanding of the scene, sparking a dialogue between human and machine perception.

VIII. Algorithms Used

8.1 Greedy Search

The method in *Figure 5* takes a simpler approach, selecting the word with the highest probability at each time step [S. K. Satti et al., 23]. While fast and efficient, it can sometimes get stuck in local optima, generating captions that may not capture the full complexity of the image.

```
def greedy_search(pic):
    start = 'startseq'
    for i in range(max_length):
        seq = [wordtoix[word] for word in start.split() if word in wordtoix]
        seq = pad_sequences([seq], maxlen=max_length)
        yhat = model.predict([pic, seq])
        yhat = np.argmax(yhat)
        word = ixtoword[yhat]
        start += ' ' + word
        if word == 'endseq':
            break
    final = start.split()
    final = final[1:-1]
    final = ' '.join(final)
    return final
```

Figure 5: Greedy Algorithm

- The provided Python code defines a function called greedy_search for generating a textual description of an image using a pre-trained neural network model.
- The function takes an image representation pic as input and initializes a starting sequence with the token 'startseq'.
- It then iteratively predicts the next word in the sequence using the model, incorporating the generated word into the sequence.
- This process continues until the model predicts the 'endseq' token or until the maximum sequence length (max_length) is reached.
- The final generated sequence is then processed to remove the 'startseq' and 'endseq' tokens, converting it into a coherent textual description.
- The code utilizes a vocabulary mapping (wordtoix and ixtoword) and employs padding (pad_sequences) to ensure compatibility with the model's input requirements.
- The generated description is returned as the output of the function.

8.2 Beam Search

The method in *Figure 6* shows a wider range of possibilities (3 different possibilities) by considering multiple caption candidates simultaneously. At each time step, it retains the top "beam_index" most likely caption continuations, fostering diversity [X. Wang et al., 22] and potentially leading to more creative and informative captions.

```
def beam_search(image, beam_index=3):
    start = [wordtoix["startseq"]]
    start_word = [[start, 0.0]]

    while len(start_word[0][0]) < max_length:
        temp = []
        for s in start_word:
            par_caps = pad_sequences([s[0]], maxlen=max_length)
            e = image
            preds = model.predict([e, np.array(par_caps)])

            word_preds = np.argsort(preds[0])[-beam_index:]

            for w in word_preds:
                next_cap, prob = s[0][:], s[1]
                next_cap.append(w)
                prob += preds[0][w]
                temp.append([next_cap, prob])

        start_word = temp
        start_word = sorted(start_word, reverse=False, key=lambda l: l[1])
        start_word = start_word[-beam_index:]

    start_word = start_word[-1][0]
    intermediate_caption = [ixtoword[i] for i in start_word]
    final_caption = []

    for i in intermediate_caption:
        if i != 'endseq':
            final_caption.append(i)
        else:
            break

    final_caption = ' '.join(final_caption[1:])
    return final_caption
```

Figure 6: Beam Algorithm

- The provided Python code implements a beam search algorithm for generating image captions using a pre-trained neural network model.
- The function beam_search takes an image representation image and an optional parameter beam_index.
- It initializes a starting sequence with the 'startseq' token and its corresponding probability. The algorithm then iteratively expands the sequences by predicting the next word using the model, considering multiple possibilities (beam search).
- The sequences are ranked based on their accumulated probabilities, and only the top beam_index sequences are retained for further exploration. This process continues until the maximum sequence length (max_length) is reached.
- The code efficiently manages the sequences by sorting them based on their probabilities, ensuring that only the most likely sequences are considered at each step.
- After the completion of the beam search, the final caption is extracted by excluding the 'startseq' and 'endseq' tokens.

1. Image Preprocessing:

The process starts with an image input, which is then resized to a standard size and converted into a numerical format suitable for the model.

2. Feature Extraction:

An InceptionV3 model, pre-trained on a massive image dataset, extracts visual features from the preprocessed image [Shetty et al., 23]. This step encodes the image's content into a 2048-dimensional vector.

- Caption Generation: A Recurrent Neural Network (RNN), specifically a Gated Recurrent Unit (GRU), uses the extracted features to generate a textual description.
- It starts with a special "startseq" token and predicts the next word in the caption at each step.
- The prediction considers both the encoded image features and the previously generated words.
- This process continues until an "endseq" token is predicted, indicating the caption's completion.

3. Greedy vs. Beam Search: The diagram presents two caption generation strategies:

- Greedy search: Chooses the most likely word at each step, resulting in a faster but potentially less diverse caption.
- Beam search: Considers a beam of multiple candidate captions simultaneously, exploring different possibilities and potentially leading to more creative and informative captions.

4. Output:

The final outcome is a generated caption that describes the content of the input image.

IX. Technology Used

9.1 Model Development Technology

- **Convolutional Neural Networks (CNNs):** Utilized for image feature extraction in image captioning.
- **InceptionV3 Architecture:** Employed for extracting high-level features from images [Szegedy et al., 16].
- **Pre-trained Weights:** Leveraged weights from ImageNet for efficient feature extraction.
- **Recurrent Neural Networks (RNNs):** Utilized Long Short-Term Memory (LSTM) for sequential modeling of image captions.
- **Long Short-Term Memory (LSTM):** Implemented for sequential modeling of image captions.

9.2 Word Embedding

- **GloVe Vectors:** Utilized pre-trained GloVe embeddings to represent words in a continuous vector space.
- **Embedding Layer:** Integrated into the model to enhance the semantic understanding of captions [K. Nguyen et al., 22].

9.3 Advanced Features Integration

- **Multi Language Support:** Incorporated through API endpoints to dynamically serve content in different languages.
- **Voice-Enabled Captions:** Integrated API-based voice synthesis for a more immersive user experience.

9.4 Development Environment

- **Anaconda, Jupyter Notebook and VS Code:** Used for its seamless performance and variety of features support for development of Machine Learning models.

X. Result and Analysis

10.1 Home Page

Welcome to SpeechSnap! The below *Figure 7* offers easy image uploads, instant caption generation, text-to-voice conversion, and multilingual translation. Simply upload an image, generate a caption, convert text to speech, or translate it into multiple languages. Experience seamless communication and accessibility with Speech Snap!

Figure 7: Home Page of Speech Snap Website.

10.2 Upload Features

The below *Figure 8* homepage welcomes you with a clear call to action: "Upload Your Image." Simply drag and drop your photo or use the dedicated upload button. Once uploaded, unleash the power of AI on your image. Generate witty captions, giving your picture a voice in seconds with text-to-speech conversion, and even translate your captions into multiple languages, letting your photos speak volumes to a global audience.

Figure 8: Uploading the image to generate the captions.

The Speech Snap homepage welcomes you with a clear call to action: "Upload Your Image." Simply drag and drop your photo or use the dedicated upload button. Once uploaded, unleash the power of AI on your image. Generate witty captions, giving your picture a voice in seconds with text-to-speech conversion, and even translate your captions into multiple languages, letting your photos speak volumes to a global audience.

10.3 Generated Captions

Below *Figure 9* we employ three distinct algorithms—Greedy, Beam-3, and Beam-5 to generate image descriptions. Additionally, we offer the capability to seamlessly convert these captions into Hindi, providing a multilingual experience tailored to your preferences.

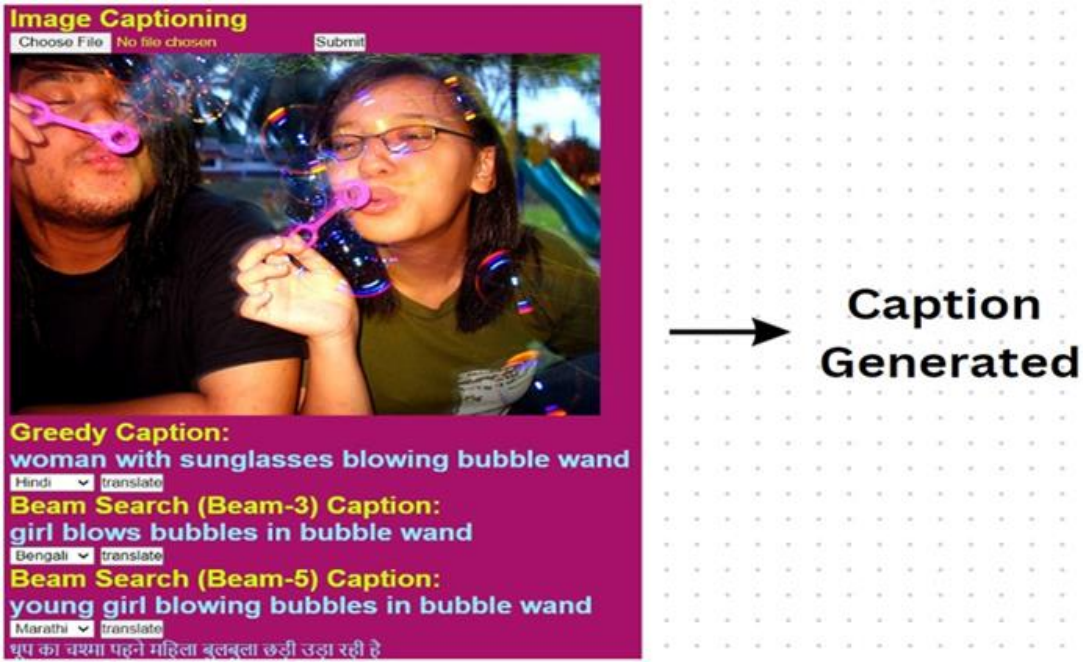

Figure 9: Output captions are generated for an input image

10.4 Translate Feature

The below *Figure 10* we have added speaks to the platform's functionality by allowing users to listen to generated text in multiple languages. Now, alongside the existing capabilities for generating captions in various languages, users can also experience the convenience of auditory communication. This enhancement not only broadens accessibility but also enhances the overall user experience.

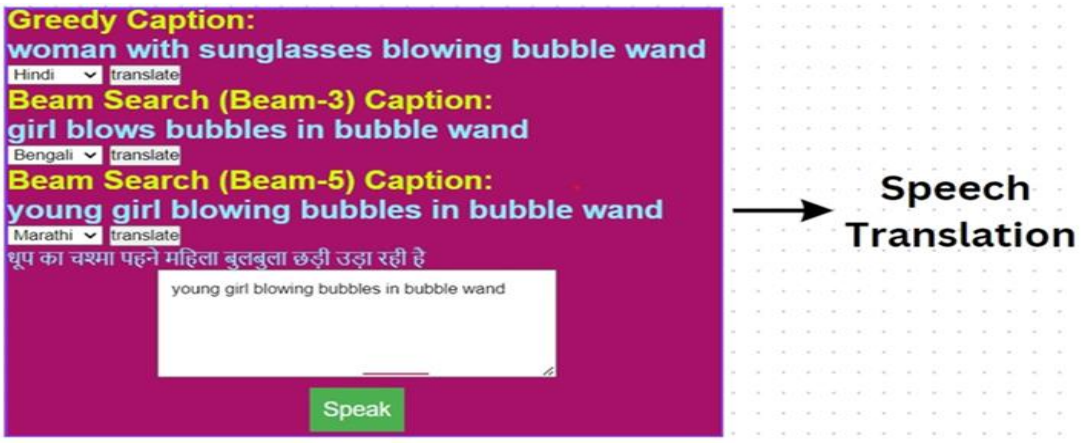

Figure 10: Selecting the language to listen the generated captions

The below *Figure 11* we can observe that the caption of Beam-5 is generated in German language.

Figure 11: Translated Caption

XI. Conclusion

In summary, this research paper introduces an innovative image captioning model that addresses challenges in generating diverse and accurate captions for images. Motivated by the potential applications in various industries, the study emphasizes the importance of technology in facilitating communication between computers and humans, automated analysis of visual data, and personalization in applications. The literature review provides insights from key studies in the field, showcasing diverse approaches to image captioning.

The methodology outlines a systematic approach to model development, highlighting the significance of diverse datasets and advanced technologies. The working of the model is elucidated, showcasing its dual-input architecture and detailed processes from training to deployment. The technology stack incorporates key tools and frameworks, emphasizing the efficiency and versatility of the proposed model.

In conclusion, this research contributes not only to image captioning but also to broader applications in technology and scientific research. The findings and methodologies presented pave the way for future advancements in the evolving field of image captioning.

REFERENCES

[1] [Oriol Vinyals et al., 15] Computer Vision and Pattern Recognition (cs.CV) arXiv:1411.4555 [cs.CV].

[2] [H. Inaguma et al., 19] H. Inaguma, K. Duh, T. Kawahara and S. Watanabe, "Multilingual End-to-End Speech Translation," 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 2019, pp. 570-577, doi: 10.1109/ASRU46091.2019.9003832. keywords: {Training; Task analysis; Decoding; Pipelines; Speech processing; Speech recognition; Training data; Speech translation; multilingual end-to-end speech translation; attention-basedsequence-to-sequence;transfer learning}.

[3] [C. S. Kanimozhiselvi et al., 22] C. S. Kanimozhiselvi, K. V, K. S. P and K. S, "Image Captioning Using Deep Learning," 2022 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2022, pp. 1-7, doi: 10.1109/ICCCI54379.2022.9740788. keywords: {Deep learning; Computational modeling; Computer architecture; Feature extraction; Convolutional neural networks; Informatics; Long short term memory; Convolutional Neural Network; Xception; Inception v3;ResNet 50;Long Short Term Memory}.

[4] [D. Harwath et al., 18] D. Harwath, G. Chuang and J. Glass, "Vision as an Interlingua: Learning Multilingual Semantic Embeddings of Untranscribed Speech," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[5] [Calgary, AB et al., 18] Calgary, AB, Canada, 2018, pp. 4969-4973, doi:10.1109/ICASSP.2018.8462396. keywords:{Training;Visualization; Acoustics; Linguistics;Speech processing;Semantics;Context modeling;Vision and language;unsupervised speech processing;cross-lingual speech retrieval}.

[6] [M. Schiedermeier et al., 21] "Multi-Language Support in TouchCORE," 2021 ACM/IEEE International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C), Fukuoka, Japan, 2021, pp. 625-629, doi: 10.1109/MODELS-C53483.2021.00096. keywords:{Navigation; Model driven engineering; Graphical user interfaces; multi-viewmodelling; model consistency; inter-model navigation; perspective}.

[7] [A. Hani et al., 19] A. Hani, N. Tagougui and M. Kherallah, "Image Caption Generation Using A Deep Architecture," 2019 International Arab Conference on Information Technology (ACIT), Al Ain, United Arab Emirates, 2019, pp.246-251, doi:10.1109/ACIT47987.2019.8990998. keywords: {image captioning; convolutional neural networks; recurrent neural networks; attention mechanism}.

[8] [Al-Malla et al., 22] Al-Malla, M.A., Jafar, A. & Ghneim, N. Image captioning model using attention and object features to mimic human image understanding. J Big Data 9, 20 (2022). https://doi.org/10.1186/s40537-022-00571-w.

[9] [Farhadi et al., 10] Farhadi, A. et al. (2010). Every Picture Tells a Story: Generating Sentences from Images. In: Daniilidis, K., Maragos, P., Paragios, N. (eds) Computer Vision – ECCV 2010. ECCV 2010. Lecture Notes in Computer Science, vol 6314. Springer, Berlin, Heidelberg. https://doi.org/10.10.

[10] [Ondeng et al., 23] Ondeng, O.; Ouma, H.; Akuon, P. A Review of Transformer-Based Approaches for Image Captioning. Appl. Sci. 2023, 13, 11103. https://doi.org/10.3390/app131911103.

[11] [K. Nguyen et al., 22] K. Nguyen, D. C. Bui, T. Trinh and N. D. Vo, "EAES: Effective Augmented Embedding Spaces for Text-Based Image Captioning," in IEEE Access, vol. 10, pp. 32443-32452, 2022, doi:10.1109/ACCESS.2022.3158763.

[12] [Cho et al., 23] Cho, S.; Oh, H. Generalized Image Captioning for Multilingual Support. Appl. Sci.2023, 13, 2446. https://doi.org/10.3390/app13042446.

[13] [Gallardo García et al., 21] Gallardo García, R., Beltrán Martínez, B., Hernández Gracidas, C., Vilariño Ayala, D. (2021). Towards Multilingual Image Captioning Models that Can Read. In: Batyrshin, I., Gelbukh, A., Sidorov, G. (eds) Advances in Soft Computing. MICAI 2021. Lecture Notes in Computer Science(), vol 13068.Springer,Cham. https://doi.org/10.1007/978-3-030-89820-5_ 2.

[14] [Hu et al., 21] Hu, X., Yin, X., Lin, K., Zhang, L., Gao, J., Wang, L., & Liu, Z. (2021). VIVO: Visual Vocabulary Pre-Training for Novel Object Captioning. Proceedings of the AAAI Conference on Artificial Intelligence, 35(2),1575-1583. https://doi.org/10.1609/aaai.v35i2.16249.

[15] [S. K. Satti et al., 23] S. K. Satti, G. N. V. Rajareddy, P. Maddula and N. V. Vishnumurthy Ravipati, "Image Caption Generation using ResNET-50 and LSTM," 2023 IEEE Silchar Subsection Conference (SILCON), Silchar, India, 2023, pp. 1-6, doi: 10.1109/SILCON59133.2023.10404600.

[16] [X. Wang et al., 22] X. Wang, Z. Chen, B. Jiang, J. Tang, B. Luo and D. Tao, "Beyond Greedy Search: Tracking by Multi-Agent Reinforcement Learning-Based Beam Search," in IEEE Transactions on Image Processing, vol. 31, pp. 6239-6254, 2022, doi: 10.1109/TIP.2022.3208437.keywords:{Target tracking;Tracking; Visualization;Search problems; Reinforcement learning;Trajectory; Decision making;Visual tracking;multi-agent reinforcement learning;beam search;local and global search;greedy search}.

[17] [Shetty et al., 23] Shetty, A., Kale, Y., Patil, Y. et al. Optimal transformers based image captioning using beam search. Multimed Tools Appl (2023). https://doi.org/10.1007/s11042-023-17359-6

[18] [Szegedy et al., 16] Szegedy, Christian & Vanhoucke, Vincent & Ioffe, Sergey & Shlens, Jon & Wojna, ZB. (2016). Rethinking the Inception Architecture for Computer Vision. 10.1109/CVPR.2016.308.

[19] [Kaggle-ADITYAJN105 20] https://www.kaggle.com/datasets/adityajn105/flickr8k

[20] [Shikha Gupta 23]https://i.stack.imgur.com/XygNZ.png