AI Powered Image Captioning App for Visually Impaired

Mohammed Mudassir Viquar[1], Prateek Sourav[2], Semparidhi. G[3], Dr. Divyashree B A[4]

Dept. of AI ML, BNMIT, Bangalore

[1]mmudassirv7@gmail.com, [2]prateeksourav@gmail.com, [3]semparidhig@gmail.com, [4]divyashreeba@bnmit.in

Abstract

We present a groundbreaking application designed to significantly enhance the daily lives of visually impaired users by providing a comprehensive understanding of their surroundings and facilitating easier access to information. This innovative app employs advanced AI technologies to recognize not only objects and scenes in images but also text in documents, generating detailed captions that are then read aloud to the user. In the field of assistive technology, this application can be emerged as a ray of hope for visually impaired people, offering innovative solutions to enhance their quality of life. Our method integrates state-of-the-art models such as Inception-v3 and Vision Transformers (ViT) encoder-decoder, enabling effective feature extraction and understanding of long-range dependencies within images. This model is trained with COCO 2017 dataset from tensor flow and undergoes rigorous evaluation, achieving an accuracy of 96% on classification tasks and demonstrating precision and recall metrics of 0.96 and 0.94, respectively. Confusion matrices visualize the model's classification performance, confirming its effectiveness in identifying and describing visual content.

Keywords: Inception-v3, Vision Transformers (ViT) encoder-decoder, COCO 2017 dataset

1. INTRODUCTION

Our objective is to implement sophisticated algorithms for artificial intelligence for real-time recognition of various objects, scenes, and textual content. By using technologies such as Vision Transformer, we strive to provide users with detailed and contextually relevant information. However, recent times have witnessed Transformer models become apparent as a major innovation in the field of deep learning, changing the how we approach problems such as language translation, text generation, and image annotation. Central to this paradigm shift is Vaswani et al.'s seminal paper "Attention is All You Need" [18], which introduces the Transformer architecture and mechanisms of attention.

Transformer's success has extended beyond NLP, demonstrating superior performance in generating descriptive annotations for images in applications across multiple domains, including image annotation. Using a self-aware mechanism, Transformer can grasp the global context and semantic relationships, resulting in more coherent and informative scripts. As a component of our work, the model has a lot of potential for improving our programs recognition capabilities.

By incorporating a self-awareness mechanism, we aim to improve the model's ability to recognize and describe objects, scenes, and is more accurate and context sensitivity using Transformer instead of using other recurrent networks such as convolution neural network (CNN), recurrent neural network (RNN) etc. We prioritize user-friendly accessibility by designing intuitive interfaces and integrating features such as haptic feedback and seamless smartphone integration. These enhancements make it easy to capture images from any screen position. Promoting independence and inclusiveness is central objective. Through auditory cues and detailed environmental information, users gain the confidence to navigate their surroundings autonomously. Improving text-to-speech functionality is another key goal, with the goal of clear and accurate translation of written content. This enhancement facilitates efficient access to documents such as books and newspapers.

2. RELATED WORK

The domain of image description generation is witnessing a surge of innovative techniques designed to create detailed and informative captions. Recent studies have suggested that various approaches, each offering a distinct perspective on tackling this challenge. An approach in 2022 mimicked human visual attention for image description tasks [2] by incorporating object attributes and an "importance coefficient" encoding scheme, this model prioritizes significant objects within an attention-based encoder-decoder framework. Some researchers introduced a method combining convolutional neural networks (CNNs) and recurrent neural networks (RNNs) [1,7,15] While CNNs excel at extracting visual features and RNNs translate them into natural language descriptions. [5] introduced CAPtor in 2021, an image description generator also combining CNN and RNN for capturing continuous context and spatial details. This approach struggles with capturing nuanced semantic and contextual information, occasionally leading to inaccurate captions, however, its performance heavily relies on training data quantity and quality, leading to potential overfitting issues.

In 2022 [3] an image captioning model leveraging CNNs and Long Short-Term Memory (LSTM) networks was developed. Where some of the researches proposed [12] picture captioning using two-way semantic attention-based guidance long-short-term memory, efficiently capturing long-range dependencies and [4] proposed a semantic context-based explanation of videos using deep neural networks. Their encoder/decoder framework, combining a 2D-CNN model and layered LSTM, exhibits advanced performance in video subtitling.

Nonetheless, evaluating on larger and diverse datasets is necessary for generalizability. In the same year, [6] proposed a multi-gate attention network for image description, introducing gates to selectively focus on different parts of the input sequence. [8] introduced a dual attention mechanism for image captioning in 2021. Despite its performance improvement, the added computational complexity hinders real-time applications. [13] explored the vanishing gradient problem and the role of LSTM in addressing it in 2019. While addressing the challenge, their image caption generation model showed errors, indicating relative simplicity. In the same year, [14] worked on image captioning from Wikipedia for multilingual using deep learning models, facing challenges related to reliance on LSTM and potential conflicts with syntactic structures of different languages.

In 2020, [9] overviewed image caption generation methods emphasizing domain-specific captions using object information and semantic ontologies. The integration of x-linear attention blocks enhances caption accuracy and informativeness. In the same year [11] introduced Oscar, leveraging object-semantics aligned pre-training for perception tasks. While achieving state-of-the-art results, further analysis of Oscar's impact on different vision-language models is warranted. In 2018, [16] proposed multitask learning for cross-domain image captioning, achieving state-of-the-art performance with MLADIC. However, computational complexity and scalability limitations remain. [17] presented a fused GRU with semantic-temporal focus for video captioning , effectively combining visual and textual data. Yet, computational intensity hampers real-time applications.

Using camera [19] introduced a real-time caption generator with high accuracy to caption, image recognition, and object detection. However, challenges arise from the massive volume of training data necessary and CNNs' limitations. [20] proposed aligning where to see and what to tell in 2016, focusing on global scene-level information and local object-level details. Despite advanced results, computational costs and limitations in handling complex scenes remain and in some of the cases where [21] presented a deep neural network-based image caption generator, intending to generate semantically correct sentences. However, errors in object detection and reliance on human-labeled data pose challenges.

3. PROPOSED METHODOLOGY AND IMPLEMENTATION

A. Dataset used

We have utilized the functionality of TensorFlow datasets to access the COCO captions dataset. This particular edition encompasses images, bounding boxes, labels, and captions sourced from COCO 2014. The dataset is partitioned into subsets as delineated by Karpathy and Li (2015), and addresses certain data quality concerns present in the original dataset. For instance, it rectifies issues like some images lacking corresponding captions in the original dataset.

B. Model Architecture

The Figure 1: Model architecture represents flow of the project where image input which is fedto the model and further the following processing steps starts:

- Input Image: The process starts by introducing an input image. Regardless of size or format, images are usually resized to a standard format that works with neural networks processing.
- Inception-v3 step: Initially, Inception-v3 starts the feature extraction process from the input image. In this step, a feature vector is generated that depicts the important visual information extracted from the image.
- Feature Vector Representation: The output from the Inception-v3 network is implemented as a feature vector, which serves as a general representation of the visual content specific tothe input image.
- Vision Transformers encoder: The feature vectors are then processed through the Vision Transformers (ViT) encoder. This encoder uses self-aware mechanisms to identify long-range dependencies and interrelationships between different components in an image.
- let's try to understand the in detail working of the model with the help of the mathematical expressions. Given the input sequence $\mathbf{Y}1{:}m$ the transformer based on encoder-decoder uses conditional distribution of the target vectors $\mathbf{X}1{:}n$ such as

$p\theta\text{enc},\theta\text{dec}(\mathbf{X}1{:}n \mid \mathbf{Y}1{:}m)$

- Hence the mapping of transformer based on encoder-decoder encoding the input sequence $\mathbf{Y}1{:}m$ with the hidden states $\mathbf{Y}1{:}m$ can be expressed as

$f\theta\text{enc}: \mathbf{Y}1{:}m \rightarrow \mathbf{Y}1{:}m$

- At last the conditional probability distribution can be expressed as

$p\theta dec(\mathbf{X}1{:}n \mid \mathbf{Y}1{:}n)$

- where represents the target vector sequence $\mathbf{X}1{:}n$ and $\mathbf{Y}1{:}n$ represents the sequence of hidden state.

- Using Bayes' theorem this distribution can be factorized as

$p\theta dec(\mathbf{X}1{:}n|\mathbf{Y}1{:}n)=i=1\prod np\theta\text{dec}(\mathbf{y}i|\mathbf{X}1{:}n,\mathbf{Y}0{:}i-1)$

- Vision Transformer receiver: After the encoding process, the output goes to the Vision Transformer receiver. Here, the agency is modified and prepared for further work.
- Complex decoder output: The output of the Vision Transformer receiver serves as an accurate representation of the processed information extracted from the input image and interprets the image as "a dog eyeballing something cooking in the oven". This refined output serves as input for subsequent translation modules.
- Functions of the translation module: The translation module according to the input converting the text descriptions produced by the decoder into the target language. Specialized machine translation models able can be used in facilitate this conversion.

- Text-to-speech: After translation, the description of the translated text is subjected to text-to-speech (TTS) synthesis. This synthesis process converts text input into human-like speech.
- Audio Output: Finally, the synthesized audio forms the final output of the entire process. This audio output can be presented to the user in an audio format that reflects the an explanation of the translated text obtained from the input image.

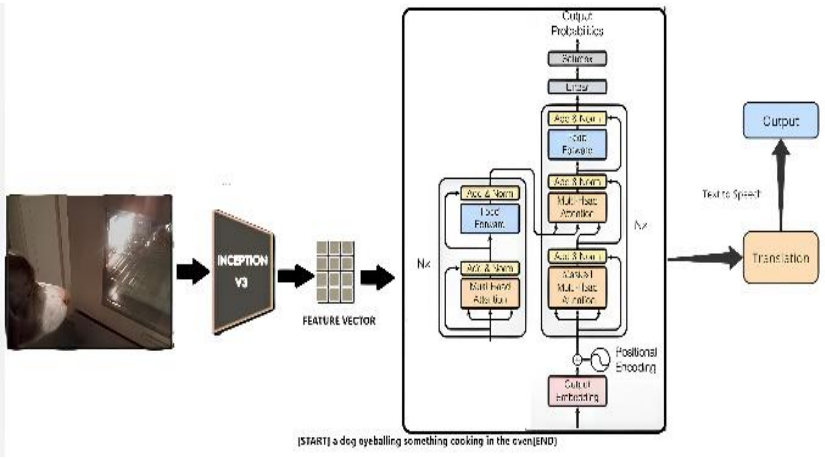

Figure 1: Model Architecture

C. Training

We used TensorFlow to retrieve the COCO captions dataset for training purposes. This version of the dataset contains photos, bounding boxes, labels, and captions. Notably, attempts were undertaken to address situations in which certain photographs lacked accompanying subtitles. The dataset is quite large, with a download size of 37.57 GiB and a dataset size of 37.35 GiB. The dataset contains a range of samples, from 40,504 to 82,783, across various splits such as 'test', 'test2015', 'train', and 'validation'.

Our training approach includes grouping sentence pairs from the COCO captions dataset. Sentences were encoded using byte-pair encoding, with a shared source-target vocabulary of around 37,000 tokens. Each training batch consisted of a set of sentence pairs including around 25,000 source tokens and 25,000 target tokens. This batching strategy was essential for efficient training and handling large datasets.

D. Softmax function

The softmax function takes a vector of real numbers as input and normalizes it into a probability distribution. Each element in the output vector represents the probability of the corresponding class.

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad \text{for } i = 1, \ldots, K \text{ and } \mathbf{z} = (z_1, \ldots, z_K) \in \mathbb{R}^K.$$

Above equation is softmax mathematical expression explained as,

- α_ts: Attention weight for source vector (encoder output) s at time step t in the decoder.
- h_s: Source vector (hidden state) at time step s from the encoder output.
- h_t: Decoder state (hidden state) at time step t.

score: Scoring function that calculates a compatibility score between the source vector and the decoder state.

In recurrent neural networks, the softmax function is often used in the output layer to convert the network's activations into a probability distribution over different classes. This is particularly useful in tasks like sentiment analysis, where the network needs to classify text into categories like positive, negative, or neutral.

4. Results

A. Attention Visualization

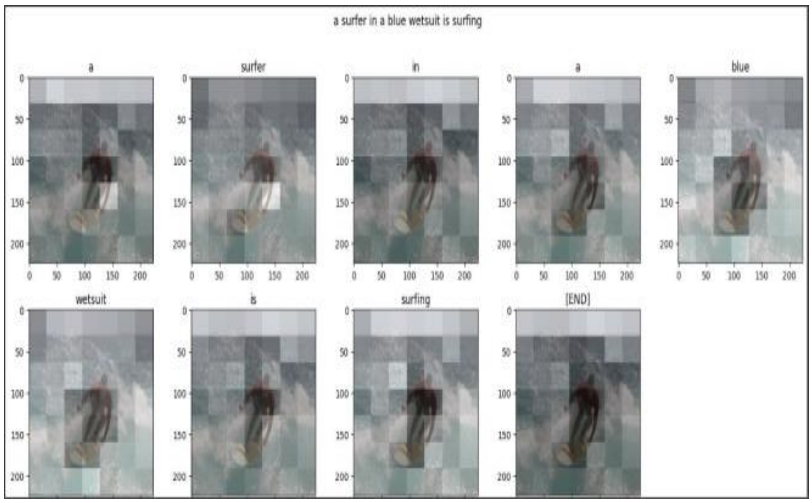

Figure 2: Attention visualization of an image

At each step of caption generation (for each word), the model calculates an attention score for each region of the image as shown in figure 2. This score indicates how consistent that image field is with the word being generated in the caption.

Imagine that the model is generating the caption word "surfing" (highlighted in yellow). Brighter areas in the heatmap around the feet and boards in the image indicate that the model is paying more attention to these regions. This makes sense because these image parts are crucial to understanding the act of surfing. Dim areas in the heatmap background indicate that the model is giving less weight to those areas when predicting the word "surfing".

B. Graphical Visualization

The training_loss appears to be steadily decreasing throughout the era, which is a positive sign this is represented in figure 3, validation_loss also appears to be decreasing, but at a slower rate than training loss. This indicates that the model is generalizing reasonably well to missing data.

Both accuracy and recognition accuracy curves is increase over training epochs, and validation_accuracy curves is as ideally staying close as we can interpret this from figure 4.

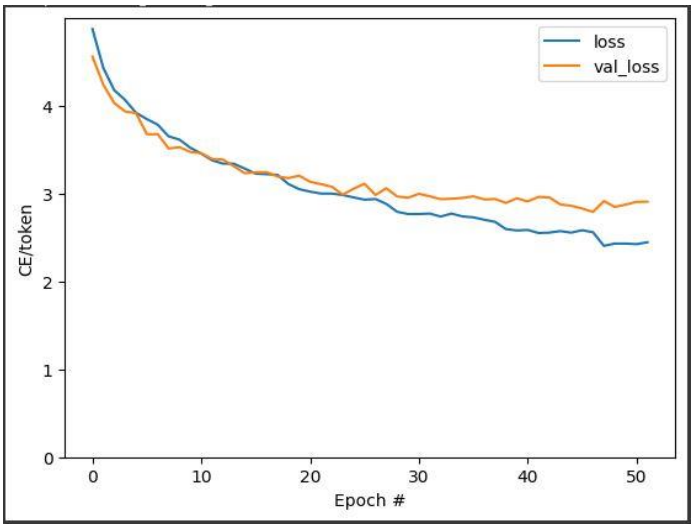

Figure 3: Train and validation loss graph

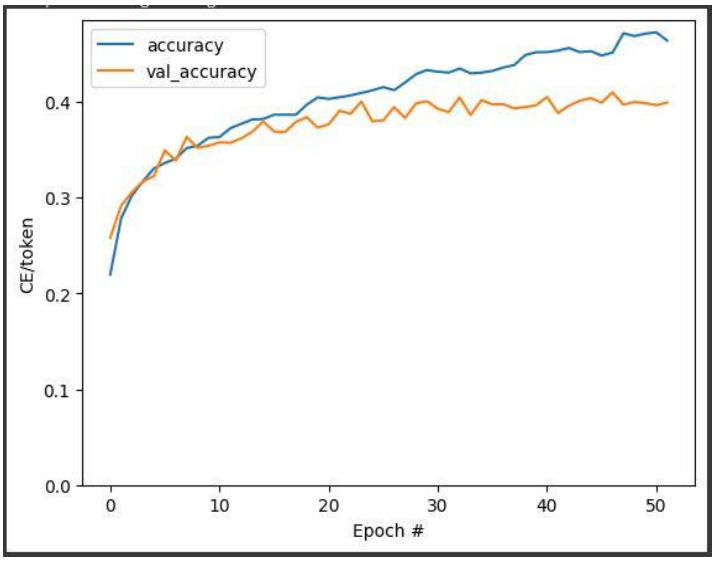

Figure 4: Accuracy and val_accuracy graph

C. Metrics

Precision (0.96): This metric indicates the proportion of predicted positive labels that were actually correct. A precision of 0.96 means that out of all the instances the model classified as positive, 96% were truly positive. Recall (0.94): This metric represents the proportion of actual positive labels that were correctly identified by the model. A recall of 0.94 means that out of all the actual positive cases in the data, the model identified 94% of them correctly.

Final results obtained from the methodology and after successful implementation of transformer model. The figure 5 has 3 results which shows how the visually impaired person can just with one click can get the description of the environment, lets analyze the results, the image depicts a stuffed animal resembling a cat. It lies on its back atop a patterned pillow, with a light-colored background resembling bed sheets or a rug. The caption generated for the image is "a cat laying on a blanket on a rug." In similar way the other images has the description such as " a newspaper with a picture of a person lying on it" and another image description is " a laptop keyboard with a monitor on top of it", this is the description of the third image but as you can see that it's a laptop hence is not perfect because it inaccurately describes the image. The system learns to identify patterns in the images and associates them with the corresponding words in the captions. In this case, the system likely identified the shape and texture of the laptop and incorrectly matched it with a similar looking computer keyboard in its database.

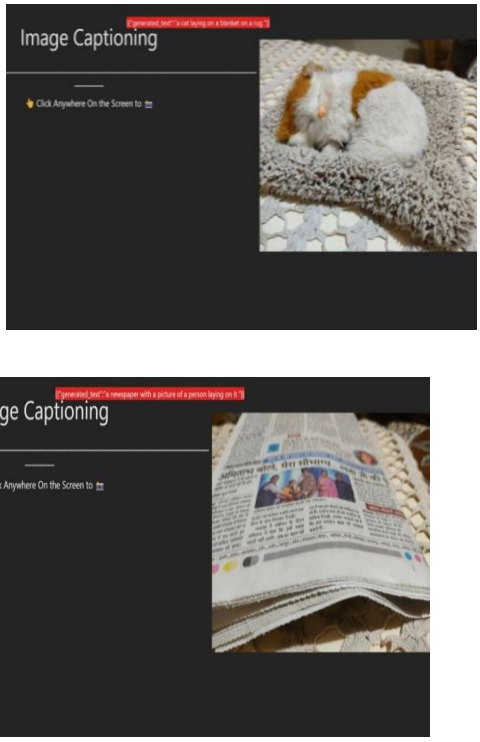

Figure 5: Results predicted by the model

5. CONCLUSION

Consequently, this paper reviews the current condition of innovation in image description generation and focuses on various methodologies and developments in model architectures. We discovered their importance in our proposed methodology through a careful study of transformer-based encoder/decoder models, inspired by the initial piece "Presence is all you need".

The model architecture presented in this paper uses transformers for feature extraction and context understanding and represents a significant advance in assistive technology, adopting a transformer-based encoder/decoder model for captioning applications offers tangible benefits. This includes improving cognitive capabilities, user-friendly accessibility, promoting independence, and enriching experiences for those who are blind or visually impaired users Using the power of transformers, we are poised to develop creative ideas that allow that for those who are blind or visually impaired to more effectively navigate and participate in the visual world.

Interpretation: Based on the accuracy and precision values, the model seems to be good at making correct positive predictions (classifying positive cases correctly).
The recall of 0.94 is slightly lower than precision, indicating that the model might have missed a small number of actual positive cases.

Future Works that can help update our assistive app if we include real-Time Processing, Implementing real-time image processing capabilities to enable users to receive instant feedback on their surroundings, allowing for more immediate and dynamic interaction with the visual environment, Introducing support for additional languages to cater to a more diverse user base, ensuring that individuals from various linguistic backgrounds can benefit from the app. Introducing customizable settings that allow users to personalize their experience based on preferences such as speech rate, volume, language, and verbosity level.

References

[1] R. Sasibhooshan, S. Kumaraswamy, and S. Sasidharan, "Image caption generation using Visual Attention Prediction and Contextual Spatial Relation Extraction," IEEE Access, 2023.

[2] M. A. Al Malla, A. Jafar, and N. Ghneim, "Image captioning model using attention and object features to mimic human image Understanding," Springer Nature, 2022.

[3] J. Basnet, S. Kumari, M. Rathore, and Dipanshu, "Image caption generator using CNN and LSTM," IJARIIT, 2022.

[4] D. Naik and C. D. Jaidhar, "Semantic context driven language descriptions of videos using deep neural Network," Springer Nature, 2022.

[5] K. Singh, N. Kumar, K. Gautam, and R. Yeolkar, "CAPtor - Neural Image Caption Generator," IRJET, 2021.

[6] W. Jiang, X. Li, H. Hu, Q. Lu, and B. Liu, "Multi-Gate Attention Network for Image Captioning," IEEE, 2021.
[7] A. Sai Bhargav and C. BV, "Image Caption Generator using Machine Learning," IRJET, 2021.
[8] L. Yu, J. Zhang, and Q. Wu, "Dual Attention on Pyramid Feature Maps for Image Captioning," IEEE, 2021.
[9] et al., "An overview of image caption generation methods," IEEE, 2020.
[10] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-Linear Attention Networks for Image Captioning," IEEE, 2020.
[11] X. Li et al., "Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks," European Conference on Computer Vision, 2020.
[12] P. Cao, Z. Yang, L. Sun, Y. Liang, M. Q. Yang, and R. Guan, "Image Captioning with Bidirectional Semantic Attention-Based Guiding of Long Short-Term Memory," Springer, 2019.
[13] J. Basnet, S. Kumari, M. Rathore, and Dipanshu, "Image Caption Generator Using CNN and LSTM," International Journal of Creative Research Thoughts, 2019.
[14] A. Garlapati, N. Malisetty, and Amrita Vishwa Vidyapeetham, "Image Captioning from Wikipedia for MultiLanguage using Deep Learning Models," Asian Journal of Convergence in Technology, 2019.
[15] P. Yadav, V. Vishwakarma, A. Tiwari, and K. Champanerkar, "Detection and Recognition of Object for Image Captioning," Asian Journal of Convergence in Technology, 2019.
[16] M. Yang, W. Zhao, W. Xu, Y. Feng, Z. Zhao, X. Chen, and K. Lei, "Multitask Learning for Cross-domain Image Captioning," IEEE, 2018.
[17] L. Gao, X. Wang, J. Song, and Y. Liu, "Fused GRU with semantic-temporal attention for video captioning," Science Direct, 2018.
[18] A. Vaswani et al., "Attention is all you need," in Advances in neural information processing systems, 2017, pp. 5998-6008.
[19] et al., "Camera to caption: A real time caption generator," in International conference on Computational Intelligence in data science IEEE, 2017.
[20] K. Fu, J. Jin, R. Cui, F. Sha, and C. Zhang, "Aligning Where to See and What to Tell: Image Captioning with Region-based Attention and Scene-specific Contexts," IEEE, 2016.
[21] et al., "Image caption Generator based on deep Neural Networks," IEEE, 2014.