Auditory Aid for Understanding Images for Visually Impaired Students Using CNN and LSTM

Aashna Arun and Apurvanand Sahay

July 10, 2024

Auditory aid for understanding images for Visually Impaired Students using CNN and LSTM

Aashna Arun
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
bl.en.u4cse20003@bl.students.amrita.edu

Apurvanand Sahay
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
a_sahay@blr.amrita.edu

***Abstract*—The world is leading towards an equal experience in learning in all educational environments, it is also a world where visual content dominates the digital landscape. Hence it is our collective responsibility to ensure that no one is left in the shadows of the information age. By utilizing the power of image captioning, we are championing the cause of inclusion, enabling people with visual impairments to navigate the digital world with confidence and ease. This assistive technology is proposed to promote accessibility and inclusion in literature, allowing people with visual impairments to engage with a wide range of written works. We performed various evaluations to demonstrate the effectiveness of this proposed model, demonstrating its potential to improve the reading experience of blind people, thereby promoting independence and accessibility. Through this paper, we propose an innovative approach that aims to develop adaptive assistive technology to enable visually impaired people to access written and visual content. Ways to design a method of creating descriptive text descriptions of images and making it easier for blind people to understand visual content were explored. This research can be applied in promoting social inclusion of blind people in various fields and can be extended from the idea of blind students understanding images in textbooks to the blind public to have accessible E-learning in digital learning platforms to enrich their reading experience, in archives to understand historic artifacts such as images in historic documents, in Public Transportation to blind travelers to access important information in Transportation apps and digital displays, in social media platforms for blind users to understand shared images, in blind users being able to comprehend visual elements in their health reports and diagnostic images and overall empowering visually challenged individuals to traverse the digital world with ease and confidence.**

***Index Terms*—Assistive technology, CNN, LSTM, Image Caption**

I. Introduction

The accessibility for visually impaired students is crucial to give them a sense of independence while reading textbooks and for fostering inclusion. Visually impaired individuals often encounter barriers when accessing digital content, especially images in textbooks, which are typically not described in accordance to the blind students. The lack of image descriptions can severely limit the educational opportunities for visually impaired students, making it difficult for them to engage with the material on the same level as their sighted peers.

This paper proposes ways to incorporate image captioning to provide an auditory aid for visually impaired students in the attempts to bridge the accessibility gap. This paper explores various CNN models in attempts to determining which CNN model among ResNet-50, VGG-16, Capsule CNN, and Inception V3 offers the highest accuracy in feature extraction for the purpose of image captioning and then exploring the effectiveness of pairing these CNN models with Long Short Term Memory (LSTM) networks to generate human-like captions for images. This paper also proposes integrating the image captioning system with an auditory aid using Google Text-to-Speech (gTTs) API to provide a seamless user experience for visually impaired users and developing a user-friendly Graphical User Interface (GUI) that enhances the accessibility of educational materials for visually impaired students.

A thorough literature survey was conducted to find related works, a comparative analysis was conducted with the four CNN models paired with LSTM to identify the most accurate combination. The integrated model combining ResNet-50 with LSTM achieved an accuracy of 0.8884 on the Flickr 8k dataset. The proposed auditory aid uses the gTTs API in Python to translate textual images into narratives, ensuring that the beauty of the visual world is accessible to all. The paper explains the methodologies and its implementation, results, and analysis of the result followed by proposing the best model to improve accessibility for visually impaired students.

II. Related Works

In attempts to find a suitable model for a implementing image captioning thorough research was conducted. By highlighting linguistic collocations and contextual inference, the incorporation of language inductive bias into encoder-decoder frameworks improves the naturalness of output captions [1]. This approach could significantly improve the interpret ability of image captions for blind individuals as it would help them significantly combat their disability since they require human like descriptions for images. Additionally, the comparison between dynamic and static dictionaries in figure captioning [2] highlights the importance of handling multi-word units and out-of-vocabulary words, which could contribute to creating more descriptive and informative captions for images.

A comprehensive survey of automatic caption generation from images reveals the necessity for architectural diversity to convert visual data into coherent sentences [10]. It emphasizes bridging the semantic divide between high-level abstract notions and low-level visual elements., crucial for many real-world applications, especially for aiding the visually impaired in understanding images. Utilizing techniques such as deep learning-based image captioning models, including neural networks and LSTM architectures, could facilitate this process.

Further advancements in attention mechanisms demonstrate the potential to capture salient image features effectively.[4] However, there remains room for improvement, particularly in capturing complex visual relationships. Additionally, the utilization of separate networks for extracting image topics that could be understood from the paper titled ”Topic-Guided Attention for Image Captioning,”. It offers promising avenues for enhancing attention mechanisms, potentially leading to more accurate and contextually relevant image descriptions for the visually impaired.[5]

The paper titled ”Deep image captioning: A review of methods, trends and future challenges,” is authored by Liming Xu, Quan Tang et al.(2023), provides a thorough analysis of current practices, emerging issues, and deep image captioning techniques. It demonstrates how language production, visual encoding, and feature representation have evolved in image captioning systems. [6] Comprehending these developments is essential to building strong and useful models that help blind people understand visuals through written descriptions.

Furthermore, the paper titled ”From Show to Tell: A Survey on Deep Learning-based Image Captioning,” which is authored by Matteo et al.(2021) provides a survey on deep learning-based image captioning, emphasizing the essential role of connecting vision and language in generative intelligence. This paper underscores the extensive research efforts dedicated to describing images with meaningful sentences and highlights the potential benefits of leveraging deep learning techniques for image captioning tasks.[8]

ResNet-50 and its potential in providing a novel method to extract image features in a image captioning problem was discovered through the paper titled ”Human pose estimation via improved ResNet50,” X. Xiao et al.(2017). This paper introduces a model for human pose estimation via an improved ResNet-50 architecture, which is relevant to image captioning due to its multi-stage cascade approach. Understanding the advantages of such architectures can provide insights into building more effective image captioning models that accurately capture the content and context of images.

Several other papers discussed various CNN model architectures that could be employed in a image captioning problem. With this research four CNN models drew the most attention which are ResNet-50, VGG-16, Capsule CNN, and Inception V3.

III. Methodologies Used for Automatic Image Captioning

A. The proposed model

In the context of an image captioning project aimed at assisting blind students in comprehending images from textbooks, the choice of the ResNet-50 model would work the best offering several advantages. ResNet-50 is a deep convolutional neural network renowned for its exceptional performance in image recognition tasks. The flickr8k dataset consists of two kinds of data, 8092 images and text data that contains five captions for each of the 8092 images. The proposed model plans on using a merge-model architecture.

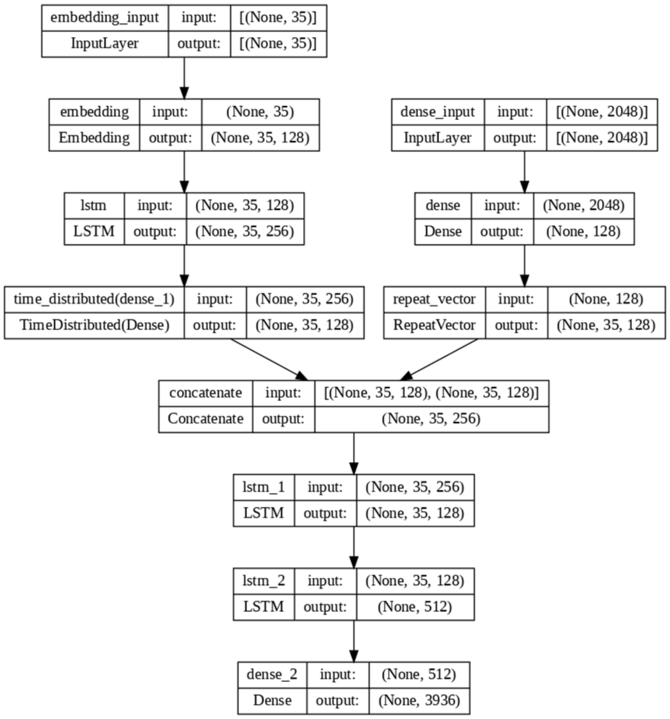

embedding_input | input: | [(None, 35)]
InputLayer | output: | [(None, 35)]
embedding | input: | (None, 35)
Embedding | output: | (None, 35, 128)
dense_input | input: | [(None, 2048)]
InputLayer | output: | [(None, 2048)]
lstm | input: | (None, 35, 128)
LSTM | output: | (None, 35, 256)
dense | input: | (None, 2048)
Dense | output: | (None, 128)
time_distributed(dense_1) | input: | (None, 35, 256)
TimeDistributed(Dense) | output: | (None, 35, 128)
repeat_vector | input: | (None, 128)
RepeatVector | output: | (None, 35, 128)
concatenate | input: | [(None, 35, 128), (None, 35, 128)]
Concatenate | output: | (None, 35, 256)
lstm_1 | input: | (None, 35, 256)
LSTM | output: | (None, 35, 128)
lstm_2 | input: | (None, 35, 128)
LSTM | output: | (None, 512)
dense_2 | input: | (None, 512)
Dense | output: | (None, 3936)

Fig. 1. Architecture of merge model using ResNet-50 and LSTM

After prepossessing the images, the images are used as input to the model, the pre-processing of images involves resizing all the images to 224x224 pixels, which is the input size expected by ResNet-50, then normalizing images by scaling the pixel values to match the distribution the ResNet-50 model and then feature extraction. As observed in Fig 1 in this model, the encoded features of an image are used along with the encoded text data to generate the next word in the caption.

The text is loaded from the captions.txt file which consists of five captions describing each image in the Images folder of the flickr 8k dataset and they are correctly mapped to their corresponding images using a dictionary in my case named captions$_{dict}$.

The methodology to implement this model involved tokenizing captions into numerical sequences and standardizing image sizes for uniformity. To make model training and evaluation easier, the dataset is then divided into training, validation, and test sets. A pre-trained ResNet50 model is used for image scaling and feature extraction to provide fixed-length informative vectors for each image, which helps with the creation of captions. Even if the photos in the dataset have different forms and sizes, scaling makes the images compatible with the 224x224x3 input size requirement of the ResNet50 model.

A vocabulary that includes all unique words from all cap-

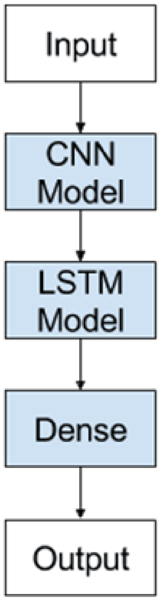


Fig. 2. Flowchart of the proposed model

tions is defined by indexing captions. The text processor uses Long Short-Term Memory (LSTM) units to encode text data, producing 128-length vectors. This preliminary stage establishes the foundation for training an image captioning model that is skilled at understanding visual content and producing evocative captions, promoting accessibility and comprehension for those with visual impairments. To address the identified challenges and leverage emerging technologies, using this methodology to help blind students overcome the challenge of understanding images and we are approaching this via the flowchart in Fig. 2. Features such as Image Caption Generation would utilize advanced neural networks to generate descriptive captions for images, enabling blind individuals to understand visual content through text. Text-to-Speech Conversion feature would convert generated captions into speech, providing auditory descriptions of images for blind individuals. Integration of Assistive Technologies explores the integration of CNNs with LSTM to enhance image understanding and improve the accuracy and coherence of generated captions for blind users. The creation of a pipeline helps develop a systematic pipeline for image processing, feature extraction, caption generation, and speech synthesis, ensuring efficient and seamless conversion of visual information to auditory output for blind users.

1) ResNet-50 Architecture for Feature Extraction: The methodology begins with the utilization of the ResNet-50 model, a deep convolutional neural network renowned for its efficiency in processing images and extracting detailed features. This architecture is pivotal for its ability to address the depth-related vanishing gradient problem through the incorporation of skip connections or shortcut connections. These connections allow the network to learn an identity function, ensuring effective representations are learned by deeper layers and are elaborated in Table I.

TABLE I
EXPLAINS THE ARCHITECTURE OF THE PROPOSED MODEL

Blocks	Layers	Input Size	Output Size
Image Feature Extractor (ResNet50)	Pre-trained CNN (ResNet50)	224x224x3	2048
Text Processor (LSTM)	Word Embedding	Vocab. Size	256
	LSTM	256	128
Output Predictor	Merge (Addition)	128 (Image) + 128 (Text)	128
	Dense Layer 1	128	128
	Dense Layer 2 (Softmax)	128	Vocab. Size

   - Input to ResNet-50: The input comprises images from the Flickr8k dataset, formatted to a uniform size to match the input requirements of the model (typically 224x224 pixels). Each image undergoes processing through ResNet-50's initial layers, which include a 7x7 convolutional layer and a 3x3 max pooling layer, setting the stage for deeper feature extraction.
   - Feature Extraction Process: The model consists of several bottleneck layers, each designed to process features at different granularities and scales. The output from ResNet-50 is a set of feature maps that encapsulate the essential visual information from the input image, which are then flattened or pooled to serve as input for the LSTM network.

2) LSTM Network for Caption Generation: Following feature extraction, the LSTM network takes over to handle the sequence prediction part of the captioning process. LSTM networks are specifically designed to manage sequence prediction problems by maintaining a memory of previous inputs using their internal state and gates, which allows them to produce a sequence of words as output.
   - Input to LSTM: The LSTM receives the processed feature vectors from the ResNet-50 model. These vectors represent a condensed version of the image's content, capturing various visual features important for generating the corresponding captions.
   - Caption Generation Process: Inside the LSTM, the input features undergo a series of transformations through LSTM cells, which decide what to retain in and omit from the memory. The LSTM cells contain mechanisms to modulate the flow of information using three types of gates: input gates, forget gates, and output gates. These gates collectively decide the next word in the caption sequence based on both the current input (the image features) and the previous

outputs (the words already generated).

- Integration and Dynamic Adjustments: To enhance the natural language output, our system incorporates attention mechanisms and dynamic dictionary adjustments. The attention mechanisms focus on different parts of the image when generating each word, allowing the model to create more contextually relevant and detailed captions. This is particularly useful in complex images where multiple objects or actions occur.
- Dynamic Dictionaries: The methodology also includes handling multi-word units and out-of-vocabulary words through dynamic dictionaries, which adjust based on the training data and the specific needs encountered during the caption generation. This flexibility ensures that the captions are not only accurate but also rich in vocabulary.

B. Comparative analysis of CNN models with LSTM

1) Comparing VGG-16+LSTM with ResNet-50+LSTM: The ResNet-50 + LSTM model is characterized by its depth, featuring a 50-layer deep convolutional network that is adept at extracting a wide range of features from images, from the most basic to the most abstract. This depth is important because it enables the model to identify complex patterns and subtleties in the visual input, which are necessary for the images to be accurately and fully described through the captions to be created.

The ResNet-50 and LSTM merge model was able to employ two LSTM layers, which work in tandem to process the sequence of word embeddings. The temporal relationships and subtleties within the text are particularly well captured by this dual-layer method, which results in more logical and contextually relevant captions as shown in Fig 1. In contrast, the VGG-16 + LSTM model, while also effective, has a simpler architecture as shown in Fig 2. It uses a single LSTM layer to process the sequence of word embeddings. While this may be sufficient for basic captioning tasks, it might not capture the depth of temporal relationships as effectively as the dual LSTM layers in the ResNet-50 model. Additionally, VGG-16 includes dropout layers, which are instrumental in preventing overfitting—a common challenge in machine learning. These layers randomly deactivate certain neurons during training, which helps the model generalize better to new, unseen images. However, this technique can also result in the loss of some potentially useful information.

Another key difference lies in how each model combines the features extracted from the image with the processed text data. The ResNet-50 model utilizes a concatenation operation, which preserves all the information from both the image features and the LSTM output. This ensures that the final caption generation process has access to the full spectrum of data. On the other hand, the VGG-16 model employs an additional operation, which, while still effective, may lead to some loss of information as it merges the features.

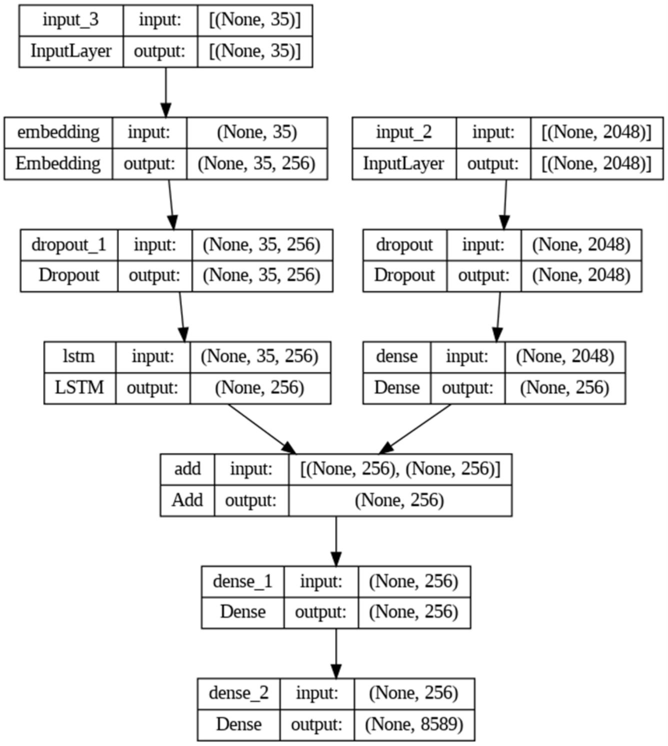

input_3 | input: | [(None, 35)]
InputLayer | output: | [(None, 35)]
embedding | input: | (None, 35)
Embedding | output: | (None, 35, 256)
input_2 | input: | [(None, 2048)]
InputLayer | output: | [(None, 2048)]
dropout_1 | input: | (None, 35, 256)
Dropout | output: | (None, 35, 256)
dropout | input: | (None, 2048)
Dropout | output: | (None, 2048)
lstm | input: | (None, 35, 256)
LSTM | output: | (None, 256)
dense | input: | (None, 2048)
Dense | output: | (None, 256)
add | input: | [(None, 256), (None, 256)]
Add | output: | (None, 256)
dense_1 | input: | (None, 256)
Dense | output: | (None, 256)
dense_2 | input: | (None, 256)
Dense | output: | (None, 8589)

Fig. 3. Architecture of merge model using VGG-16 and LSTM

2) Comparing Capsule CNN+LSTM with ResNet-50+LSTM: As a more recent development, capsule CNNs provide a special method for deciphering spatial hierarchy in images. The model was believed to potentially lead to better performance in image captioning trained and tested upon flickr8k dataset.

However, the actual effectiveness of Capsule CNNs, particularly in image captioning, is still an area ripe for research, and their performance may significantly depend on the dataset and the nature of the task at hand. Therefore, while ResNet-50 is a reliable choice based on its track record, Capsule CNNs warrants further investigation to fully understand their capabilities for image captioning as observed from Table.II the accuracy represent how the model performed with the test images.

The routing mechanism and dynamic routing algorithm are central to the functioning of Capsule Networks, as they determine the hierarchical relationships between capsules. The complex routing process leads to issues such as vanishing gradients, which is a common problem in training deep neural networks. This occurs when the gradients become too small for the network to learn from, effectively halting the learning process. This explains the poor accuracy scores and why the other models performed better in comparison.

3) Comparing InceptionV3+LSTM with ResNet-50+LSTM: Another well-known convolutional neural network design that was compared in this study is InceptionV3, which uses a unique method for feature extraction and representation. The inception modules, which are made up of parallel convo-

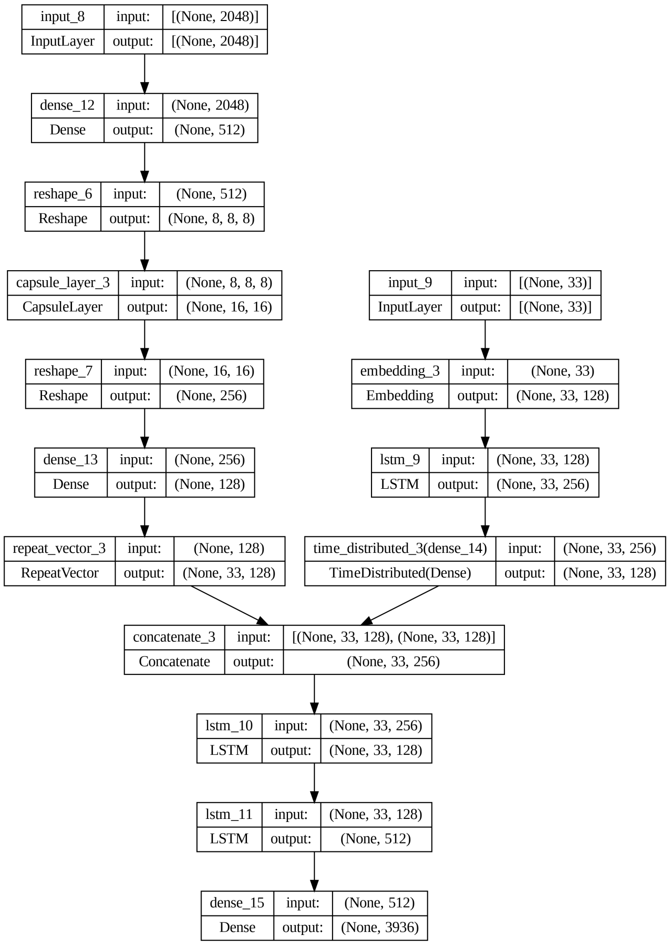

input_8 input: [(None, 2048)]
InputLayer output: [(None, 2048)]
dense_12 input: (None, 2048)
Dense output: (None, 512)
reshape_6 input: (None, 512)
Reshape output: (None, 8, 8, 8)
capsule_layer_3 input: (None, 8, 8, 8)
CapsuleLayer output: (None, 16, 16)
reshape_7 input: (None, 16, 16)
Reshape output: (None, 256)
dense_13 input: (None, 256)
Dense output: (None, 128)
repeat_vector_3 input: (None, 128)
RepeatVector output: (None, 33, 128)
input_9 input: [(None, 33)]
InputLayer output: [(None, 33)]
embedding_3 input: (None, 33)
Embedding output: (None, 33, 128)
lstm_9 input: (None, 33, 128)
LSTM output: (None, 33, 256)
time_distributed_3(dense_14) input: (None, 33, 256)
TimeDistributed(Dense) output: (None, 33, 128)
concatenate_3 input: [(None, 33, 128), (None, 33, 128)]
Concatenate output: (None, 33, 256)
lstm_10 input: (None, 33, 256)
LSTM output: (None, 33, 128)
lstm_11 input: (None, 33, 128)
LSTM output: (None, 512)
dense_15 input: (None, 512)
Dense output: (None, 3936)

Fig. 4. Architecture of merge model using Capsule CNN and LSTM

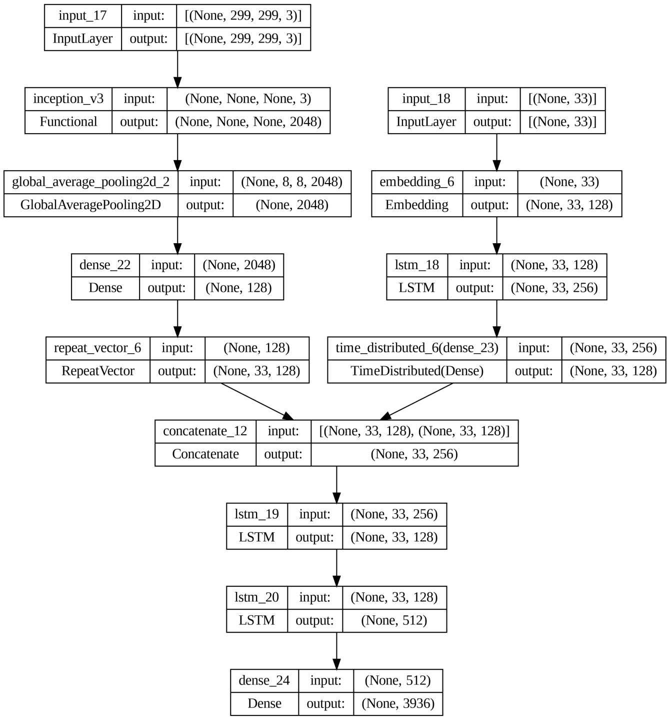

input_17 input: [(None, 299, 299, 3)]
InputLayer output: [(None, 299, 299, 3)]
inception_v3 input: (None, None, None, 3)
Functional output: (None, None, None, 2048)
global_average_pooling2d_2 input: (None, 8, 8, 2048)
GlobalAveragePooling2D output: (None, 2048)
dense_22 input: (None, 2048)
Dense output: (None, 128)
repeat_vector_6 input: (None, 128)
RepeatVector output: (None, 33, 128)
input_18 input: [(None, 33)]
InputLayer output: [(None, 33)]
embedding_6 input: (None, 33)
Embedding output: (None, 33, 128)
lstm_18 input: (None, 33, 128)
LSTM output: (None, 33, 256)
time_distributed_6(dense_23) input: (None, 33, 256)
TimeDistributed(Dense) output: (None, 33, 128)
concatenate_12 input: [(None, 33, 128), (None, 33, 128)]
Concatenate output: (None, 33, 256)
lstm_19 input: (None, 33, 256)
LSTM output: (None, 33, 128)
lstm_20 input: (None, 33, 128)
LSTM output: (None, 512)
dense_24 input: (None, 512)
Dense output: (None, 3936)

Fig. 5. Architecture of merge model using Inception V3 and LSTM

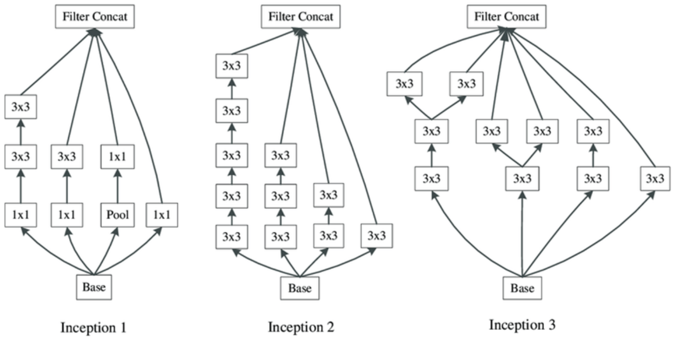


Fig. 6. Down sampling in InceptionV3

lutional branches with various filter sizes, are the central components of InceptionV3 which is observed in Fig.5. These branches enable the efficient extraction of multi-scale features by capturing features at several resolutions and scales. The architecture of InceptionV3 contains the stacked inception modules interleaved with max-pooling layers for downsampling as observed in Fig. 6., culminating in global average pooling and fully connected layers for classification. By requiring fewer parameters and calculations than previous designs such as VGG-16, InceptionV3 highlights computational efficiency. Smaller convolutions and dimensionality reduction strategies are used within the inception modules to achieve this efficiency. Compared to ResNet-50, InceptionV3 usually has fewer layers even though it is still deep. In other words, InceptionV3's reliance on inception modules may not offer the same level of capability in capturing nuanced visual information in comparison with ResNet-50 but performs better than VGG-16. Due to its shallower depth and focus on efficiency, InceptionV3 may struggle to fully grasp the complexities of images, particularly like in our case when tasked with understanding long-range dependencies and subtle features which is very crucial for tasks such as image captioning especially for blind students.

IV. RESULT AND ANALYSIS

In the analysis of outcomes, the focus is on evaluating the performance of the image caption generation system developed within the scope of this individual project. After conducting the evaluation to find the performance of three well-known convolutional neural network (CNN) architectures—ResNet-50, InceptionV3, and VGG-16—in the context of image captioning, when paired with a Long Short-Term Memory (LSTM) recurrent neural network with the goal of developing an assistive technology for visually impaired individuals. The primary objective was to determine which model architecture and which one could generate the most descriptive textual descriptions of images, thereby facilitating the understanding of visual content by blind individuals. In this system, ResNet-

50 CNN serves as a powerful feature extractor, encoding important visual information from input images. Through fine-tuning a comprehensive dataset, the pre-trained ResNet-50 model becomes proficient in capturing the semantic nature of images, ensuring accurate interpretation of visual scenes. The encoded features are then passed to the LSTM network, which acts as a language model to build text descriptions. This LSTM model generates word-by-word annotations, gradually building coherent and descriptive stories that match the visual content. The evaluation encompasses metrics such as accuracy, precision-recall and F-1 score. The chosen dataset served as a foundational resource for training and validating the image captioning model, tailored specifically to the educational context and it was found that ResNet50 paired with LSTM generated the most accurate captions for images.

Models	Accuracy	Loss
ResNet50 + LSTM	0.8884	0.4702
VGG-16 + LSTM	0.7648	0.5845
Capsule CNN + LSTM	0.3750	1.0984
InceptionV3 + LSTM	0.7834	0.4889

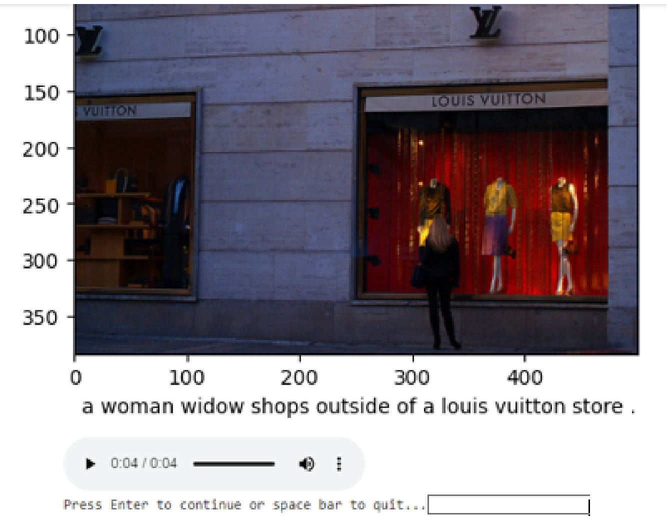
TABLE II
MODEL ACCURACY AND LOSS COMPARISON

The ability to comprehend visual content through text is a crucial trait that changes the way blind people acquire information. In addition, the generated captions can be converted into voice using Text-to-voice Conversion technology, which meets the auditory learning preferences of blind users by providing an audio representation of the visuals. Through this, a blind student could hear what the image is trying to portray through the click of a button.

By integrating this technology, we have facilitated an inclusive educational environment where blind students can experience a level of autonomy in their learning process. As observed in Fig.8 we can see that the GUI's simplicity ensures ease of use, while its efficiency in delivering accurate image descriptions enhances the learning experience. The blind student has to only click 2 buttons which are ”Enter” to continue to the next image in the textbook or the next written content, then ”Spacebar” to quit the session or in other words stop reading the contents in the book. This accomplishment not only signifies a step forward in assistive technology but also reflects our commitment to creating equal learning opportunities for all students. The successful implementation of this GUI stands as a testament to the potential of innovative solutions in overcoming educational barriers.

V. CONCLUSION

This paper presents a auditory aid based on deep-learning model which automatically generates human like captions for images. The proposed model is based on a CNN that encodes an image into a compact representation evaluating image features, followed by an LSTM model that generates corresponding sentences based on the learned image features. The model was trained to maximize the accuracy of the caption generated given the image. After comparing four CNN models the ResNet-50 and LSTM merge model had the best accuracy among the others. The various reasons that could lead to ResNet-50 and LSTM's merge model to work over other CNN-based merge models with LSTM were analyzed.



a woman widow shops outside of a louis vuitton store .

```
Press Enter to continue or space bar to quit...
```

Fig. 7. Proposed GUI

The model worked quite well when it was tested on several images from within and outside the training dataset. The captions it generated for the images were quite accurate. The source of the input image also played an important role in feature extraction and hence caption generation hence the model can be further trained with diverse datasets.

Moreover, continuous refinement and optimization of the model architecture and training procedures can lead to incremental improvements in caption quality and generation speed. Experimenting with different pre-trained CNN architectures and LSTM variations, as well as exploring alternative text generation models, can help uncover more effective strategies for image captioning.

Its applications extends beyond vlind students understanding textbook images to facilitating accessible e-learning platforms, interpreting historic artifacts in archives, providing critical information in transportation apps, and allowing blind users to understand health reports and social media images.

In conclusion, while the image captioning system developed in this project demonstrates promising capabilities, there is still significant room for refinement and expansion and help blind students understand images from textbooks. This innovation aims to impact the field of computer vision research and be able to empower blind individuals in various domains involving digital or photographic information.

VI. FUTURE SCOPE

The future scope of this research includes several key areas of development aimed at enhancing the capabilities and accessibility of image annotation systems. Architecture testing will involve exploring different pre-trained CNN architectures and LSTM variants to design sub-strategies topics more effectively. Contextual relevance will be prioritized to ensure that subtitles fit seamlessly with accompanying academic materials, promoting a cohesive educational experience.

Interactive features with appropriate electronic hardware should be developed such that visually impaired students to actively interact with visual content. A hardware that allows visually impaired students to ask questions and receive detailed descriptions of specific visual elements.

In order to enhance comprehension and supplement visual captions, multi sensory integration—such as using tactile graphics or audio descriptions—will be crucial. Using a user-centered design approach will entail working closely with visually impaired students to modify the system in a way that best suits their individual requirements and preferences. Furthermore, tight cooperation with educators will be maintained to guarantee conformity with curricular requirements and instructional strategies.

Finally, the commitment to comply with and contribute to the development of new accessibility standards for educational materials will be maintained, ensuring that image annotation systems remain inclusive and accessible for all users. Through ongoing research and experimentation, coupled with advancements in deep learning and computer vision techniques, we aim to further enhance the system's performance and usability.

References

[1] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-Encoding Scene Graphs for Image Captioning," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 10677-10686, doi: 10.1109/CVPR.2019.01094.

[2] X. Qian, E. Koh, F. Du, S. Kim, J. Chan, R. Rossi, S. Malik, and T. Y. Lee, "Generating Accurate Caption Units for Figure Captioning," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 17, no. 2, Art. no. 34, Apr. 2021, pp. 2792-2804, doi: 10.1145/3442381.3449923.

[3] V. Volobuev and P. Y. Afonichkina, "Generating Photo Captions for Instagram," in *2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*, St. Petersburg, Moscow, Russia, 2021, pp. 735-738, doi: 10.1109/ElConRus51938.2021.9396441.

[4] V. Agrawal, S. Dhekane, N. Tuniya, and V. Vyas, "Image Caption Generator Using Attention Mechanism," in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, 2021, pp. 1-6, doi: 10.1109/ICCCNT51525.2021.9579967.

[5] Z. Zhu, Z. Xue, and Z. Yuan, "Topic-Guided Attention for Image Captioning," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Athens, Greece, 2018, pp. 2615-2619, doi: 10.1109/ICIP.2018.8451083.

[6] X. Xiao and W. Wan, "Human pose estimation via improved ResNet50," in *4th International Conference on Smart and Sustainable City (ICSSC 2017)*, Shanghai, China, 2017, pp. 1-5, doi: 10.1049/cp.2017.0126.

[7] L. Xu, Q. Tang, J. Lv, B. Zheng, X. Zeng, and W. Li, "Deep Image Captioning: A Review of Methods, Trends and Future Challenges," *Neurocomputing*, vol. 546, pp. 126-287, 2023, doi: 10.1016/j.neucom.2023.126287.

[8] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, "From Show to Tell: A Survey on Deep Learning-Based Image Captioning," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 539-559, 1 Jan. 2023, doi: 10.1109/TPAMI.2022.3148210.

[9] A. Jamil et al., "Deep Learning Approaches for Image Captioning: Opportunities, Challenges and Future Potential," in *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3365528.

[10] Ms. Yugandhara, A. Thakare, and Prof. K. H. Walse, "Automatic Caption Generation from Image: A Comprehensive Survey," *ACI@ISIC*, 2022.

[11] M. Hodosh, P. Young, and J. Hockenmaier, "Framing Image Description as a Ranking Task: Data, Models, and Evaluation Metrics," *J. Artif. Intell. Res.*, vol. 47, 2013, pp. 853-899.

[12] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *arXiv preprint arXiv:1502.03167*, 2015.

[13] A. Karpathy, A. Joulin, and L. Fei-Fei, "Deep Fragment Embeddings for Bidirectional Image Sentence Mapping," in *Proc. NIPS*, 2014.

[14] R. Kiros, and R. Z. R. Salakhutdinov, "Multimodal Neural Language Models," in *Proc. NIPS Deep Learning Workshop*, 2013.

[15] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Baby Talk: Understanding and Generating Simple Image Descriptions," in *Proc. CVPR*, 2011.

[16] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi, "Collective Generation of Natural Image Descriptions," in *Proc. ACL*, 2012.

[17] Y. Chu, X. Yue, L. Yu, S. Mikhailov, and Z. Wang, "Automatic Image Captioning Based on ResNet50 and LSTM with Soft Attention," *Wireless Communications and Mobile Computing*, vol. 2020, Article ID 8909458, 2020, https://doi.org/10.1155/2020/8909458.

[18] Y. Liu, K. Xu, and J. Xu, "Periodic Surface Defect Detection in Steel Plates Based on Deep Learning," *Applied Sciences*, vol. 9, p. 3127, 2019, doi: 10.3390/app9153127.

[19] Y. S. Jain, T. Dhopeshwar, S. K. Chadha, and V. Pagire, "Image Captioning using Deep Learning," in *2021 International Conference on Computational Performance Evaluation (ComPE)*, Shillong, India, 2021, pp. 040-044, doi: 10.1109/ComPE53109.2021.9751818.

[20] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[21] Sahay, Apurvanand & Joseph, Amudha. (2020). Integration of Prophet Model and Convolution Neural Network on Wikipedia Trend Data. *Journal of Computational and Theoretical Nanoscience*. 17. 260-266. doi: 10.1166/jctn.2020.8660.

[22] S. Staby and M. R, "Spatial-Temporal Analysis for Traffic Incident Detection Using Deep Learning," in *2023 Innovations in Power and Advanced Computing Technologies (i-PACT)*, Kuala Lumpur, Malaysia, 2023, pp. 1-6, doi: 10.1109/i-PACT58649.2023.10434683. keywords: Deep learning;Training;Adaptation models;Surveillance;Roads;Videos;Accidents;Transportation;traffic incident detection;multiclass video classification;LRCN network;deep learning approach,

[23] K. Serath Chandra, R. Poda, A. Vinod, and R. A. Arun, "Pixels to Phrases: Bridging the Gap with Computationally Effective Deep Learning models in Image Captioning," in *2023 12th International Conference on Advanced Computing (ICoAC)*, Chennai, India, 2023, pp. 1-7, doi: 10.1109/ICoAC59537.2023.10249809.

[24] S. Sarath and J. Amudha, "Visual question answering models Evaluation," in *2020 International Conference for Emerging Technology (INCET)*, Belgaum, India, 2020, pp. 1-5, doi: 10.1109/INCET49848.2020.9154094. keywords: Visualization;Computational modeling;Predictive models;Knowledge discovery;Feature extraction;Computer science;Context modeling;VQA;Visual Question answering;Pythia;CNN;LSTM;NLP,

[25] A. Anil and S. Santhanalakshmi, "Caption Generation for Images with Deep Neural Networks," in *2022 2nd International Conference on Intelligent Technologies (CONIT)*, Hubli, India, 2022, pp. 1-9, doi: 10.1109/CONIT55038.2022.9847841. keywords: Deep learning;Image recognition;Neural networks;Feature extraction;Convolutional neural networks;Task analysis;Convolutional Neural Networks;Recurrent Neural Networks;Long Short Term Memory;CNN-RNN,