A Comparative Study of Image Captioning Models

[1]Akhilesh Dixit, [2]Dr. Kirti Wanjale, [3]Narendra Jadhav

[1,2,3]Department of Computer Engineering, Vishwakarma Institute of Information Technology, Pune, India. [1]akhilesh.22010131@viit.ac.in, [2]kirti.wanjale@viit.ac.in, [3]narendra.22010817@viit.ac.in

***Abstract*— Image captioning, the art of generating descriptive textual explanations for visual content, stands at the intersection of computer vision and natural language understanding. In this study, we embark on a meticulous comparative analysis of two distinct image captioning models, each distinguished by a unique Convolutional Neural Network (CNN) architecture. One model seamlessly integrates the venerable VGG-16 with a traditional Recurrent Neural Network (RNN), while the other expertly combines the state-of-the-art InceptionV3 with a transformer-based sequence generator. This study explores the comparative efficiencies and innate strengths of these models in the domain of image labeling. Our work presents thorough methodologies, exhaustive experimental findings, and meticulous analysis of results. As the merging of computer sight and natural language comprehension continues advancing, our examination contributes understandings with potential to enhance human-machine interaction involving visual material. In addition to propelling discussion in computer vision and natural language comprehension, our research brings a practical aspect to image labeling. Beyond the generation of descriptive textual explanations for visual content, we enhance user accessibility by implementing a text-to-speech conversion feature using the Google Text-to-Speech (gTTS) library in Python. This augmentation aims to empower individuals with varying needs, facilitating a more inclusive and user-friendly experience in interacting with the generated image captions.**

Keywords*—*Image Captioning, Machine Learning, VGG16, ImceptionV3, RNN, Transformer

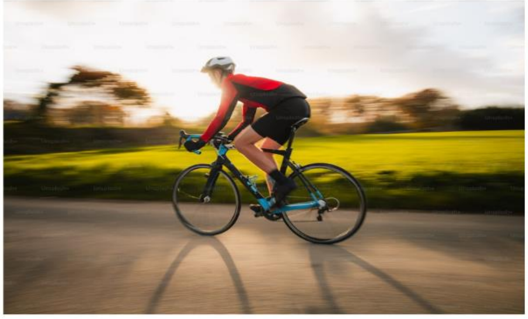
I. INTRODUCTION

In modern times, pictures have become an omnipresent means for sharing information. This paper discusses the growing importance of crafting captions for visual content that resemble those made by people, a job at the meeting point of computer sight and computer language. The advancement of picture captioning, aimed toward spontaneously generating detailed and situationally applicable written depictions for images, has been noteworthy.

This study concentrates on a comparative analysis of two prominent approaches to image captioning: one leveraging the conventional Recurrent Neural Networks (RNNs) and the other harnessing the transformative capabilities of the transformer architecture. The evaluation of both models centers on their proficiency in generating high-quality image captions, considering a spectrum of challenges. Beyond merely comprehending visual content, the challenges encompass a deeper understanding of intricate linguistic structures. This underscores the necessity for nuanced approaches to address the inherent complexity of the problem at hand.

This research engages in a comparative analysis of two image captioning models, aiming to determine their respective efficacy through parameter calculations. The models under scrutiny are Recurrent Neural Networks (RNNs) and transformers. Through an exploration of the synergy between these models, the study seeks to enrich the broader comprehension of how advancements in computer vision and natural language processing can elevate the production of pertinent captions for visual content.

Our work also extends a bit beyond the traditional image captioning paradigms by integrating a text-to-speech conversion component. This augmentation, facilitated by the gTTS library in Python, enables the generated image captions to be dynamically translated into spoken words. This development holds promise for enhancing user engagement and accessibility in diverse applications, ranging from assistive technologies to interactive visual content experiences.



Caption: Man Riding a Bicycle

Fig 1. Example of Image Captioning

DOI : 10.35291/2454-9150.2024.0056

A. Motivation and Significance

Image captioning has a wide array of applications, including assisting visually impaired individuals in understanding visual content, enhancing image search engines, and contributing to an enriched user experience in social media platforms. Accurate image captioning not only necessitates the comprehension of image contents but also requires the ability to construct coherent and contextually relevant sentences, making it a multi-modal task that draws on both computer vision and natural language processing techniques.

B. Research Objectives

The primary objective of this research is to rigorously evaluate and compare two prominent image captioning approaches. The first approach leverages a traditional RNN-based model, where the output of a VGG16(CNN) is used as input for generating captions. The second approach, in contrast, employs a transformer architecture, a paradigm-shifting model in natural language processing, where output of InceptionV3 is given as input for the transformer.

II. LITERATURE SURVEY

The development of image captioning technology has a rich history and is not a recent endeavour. Corporations, research communities, and open-source initiatives have long been engaged in the extraction and classification of features from images. While image classification is a well-established field, our endeavour is primarily focused on improving and advancing image captioning technology.

In the era of deep neural networks, the race for enhanced performance has intensified, leading to the creation of numerous neural networks, each surpassing its predecessor in terms of capabilities and features.

For instance, Krizhevsky et al. [1] introduced a neural network empowered by GPU training procedures, which effectively reduced input dimensions to produce (None, 1000) outputs without overfitting. They achieved this through a model incorporating five convolution layers, Maxpooling layers, and dropout regularization.

Karpathy and FeiFei [2][8] made significant strides by processing image datasets alongside their associated sentence descriptions. This enabled computers to generate image descriptions. They introduced a Multimodal Recurrent Neural Network (m-RNN) that capitalized on the co-linear arrangement of features to accomplish this task.

Vinyals et al. [3] took image captioning to new heights by developing a generative model consisting of an RNN. This innovation contributed to machine translation and computer vision, ensuring a higher probability of generating accurate sentence descriptions for target images.

Xu et al. [4][9] pioneered the development of an attention-based model capable of automatically learning and describing images. In paper Relation-aware Transformer for Image Captioning [5] authors introduce a transformer-based architecture specifically designed for image captioning. It goes beyond object detection and focuses on capturing relationships between objects in the image. Diverse Beam Search for Image Captioning with Transformers [6] tackles the challenge of generating diverse captions for a single image. It leverages a beam search approach within the transformer architecture.

III. METHODOLOGY

In this section, we present the methodologies applied to develop and evaluate the two image captioning models: one based on the VGG-16 architecture and traditional RNN (VGG-16 RNN), and the other based on the InceptionV3 architecture with a transformer (InceptionV3 Transformer).

A. Dataset Selection and Preprocessing:

We detail the dataset utilized for our image captioning research, the Flickr8k dataset, and cover the preprocessing steps undertaken to make the dataset compatible with our models.

Flickr8k Dataset:

The Flickr8k dataset, a well-established collection of images with associated textual descriptions, was selected as the primary dataset for our research. This dataset comprises a total of 8,091 images, each paired with five descriptive captions, resulting in a rich source of visual and textual data for image captioning experiments.

Data Split:

To facilitate training and evaluation, the Flickr8k dataset was divided into two subsets: the training set and the validation set. The training set was allocated most of the images, ensuring an ample supply of data for model training. The validation set was designated for model performance monitoring during training. Out of 40,455 data pairs, 32,650 were assigned for training, and the rest 8,095 were used for validation.

Data Preprocessing:

Data preprocessing was a vital aspect of our methodology. To ensure data uniformity and compatibility for deep neural network models, we implemented several preprocessing steps. For images, resizing and normalization were performed to meet the specific input requirements of our chosen models. Image captions underwent tokenization, lowercasing, and the removal of punctuation and non-alphanumeric characters. These steps were essential to standardize the text data, making it more suitable for the image captioning task.

A. VGG-16 RNN Model

VGG16 is a renowned convolutional neural network architecture, consisting of 16 layers. Its simplicity,

DOI : 10.35291/2454-9150.2024.0056

employing 3x3 convolutional filters and max-pooling layers, alongside its impressive performance in tasks like image classification and object detection, have made it widely adopted in computer vision research [7].

1. VGG16 Model:

The VGG-16 convolutional neural network was employed as the feature extractor. Notably, the classification layer of the VGG-16 model, which was the last layer responsible for class predictions, was deliberately removed. This restructured model allowed us to extract image features directly from the output of the last convolutional layer.

2. Image Feature Extraction:

After feature extraction with VGG-16, the extracted image features were saved to a file named "features.pkl." These features served as the input data for the captioning component of the VGG16-RNN model.

3. Text Tokenization:

In the text preprocessing phase, we performed caption tokenization using the Tokenizer class. This process involved breaking down captions into individual words, allowing them to be utilized for model training.

4. Tokenizer Saving:

To ensure consistent tokenization in both training and inference, we saved the tokenizer as "tokenizer.pkl" using Pickle. This saved tokenizer can be readily loaded for future use.

5. RNN Architecture:

For the captioning component, we employed a Recurrent Neural Network (RNN). The RNN architecture was configured with Long Short-Term Memory (LSTM) units. LSTM networks are well-suited for sequential data tasks like language generation due to their ability to capture long-range dependencies in the data. In our model, LSTMs facilitated the generation of descriptive captions for images.

6. Evaluation Metrics:

To assess the quality of captions generated by the VGG16-RNN model, we employed standard evaluation metrics, consistent with our InceptionV3 model. These metrics included BLEU (Bilingual Evaluation Understudy).

7. Hardware and Software:

Experiments were conducted on a system equipped with high-performance GPUs to facilitate efficient model training. We used Python as our primary programming language and leveraged deep learning frameworks such as TensorFlow or PyTorch for model implementation.

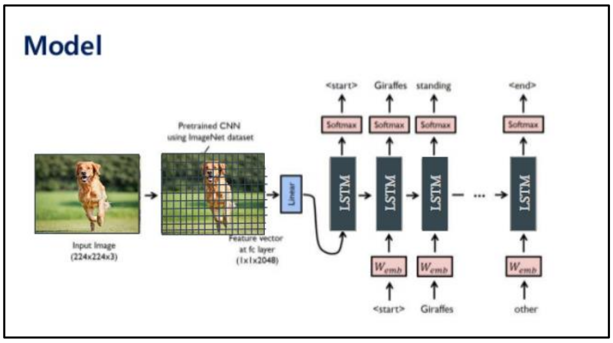


Fig 2. Structure of Model

Description (Fig 2): This figure presents a graphical representation of the image captioning process, illustrating the sequential steps involved from data preprocessing to caption generation.

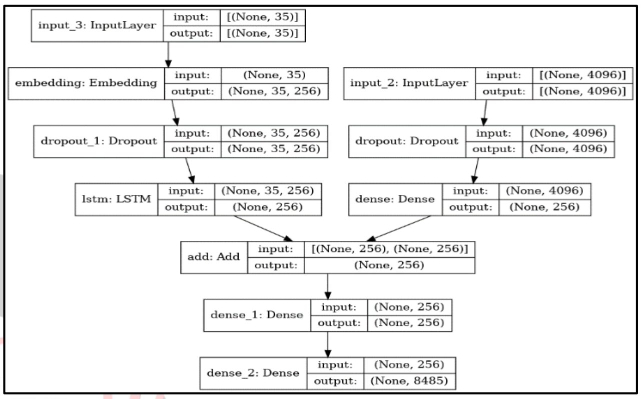


Fig 3. Pictorial representation of the process

B. *InceptionV3 – Transformer Model*

Transformers, a groundbreaking deep learning architecture, have revolutionized NLP and sequential data tasks by utilizing self-attention mechanisms for parallelization and improved long-range dependency modelling, surpassing traditional RNNs [11]. With encoder-decoder structures and multi-head attention mechanisms, transformers excel in machine translation, text generation, and sentiment analysis, becoming a cornerstone in modern deep learning research, driving progress across diverse domains beyond NLP [15].

1. InceptionV3 Architecture:

The InceptionV3 convolutional neural network served as the feature extractor for our image captioning model [12]. We configured the model to remove the classification layer, thus enabling the extraction of image features from the final convolutional layers. This architecture excels at capturing intricate visual details from images, which are essential for generating descriptive captions.

2. Image Feature Extraction:

The InceptionV3 model was employed to extract high-level image features. After removing the classification layer, the model produced a rich representation of visual information from images. These extracted features were

DOI : 10.35291/2454-9150.2024.0056

then utilized as input for the subsequent Transformer-based captioning component. This integration allowed us to leverage the power of the InceptionV3 model for comprehensive image feature extraction, serving as the foundation for the caption generation process.

3. Transformer Architecture:

Our image captioning model utilized a Transformer architecture for the caption generation component. Transformers have demonstrated exceptional capabilities in sequence-to-sequence tasks, making them a compelling choice for natural language generation. We configured the Transformer architecture to process the high-level image features extracted by the InceptionV3 model and generate descriptive captions.

4. Image Feature Integration:

The extracted image features from InceptionV3 were seamlessly integrated into the Transformer model's architecture. These features provided the model with a comprehensive visual context, which was essential for generating detailed and contextually relevant captions. This integration allowed the model to effectively fuse visual and textual information, enhancing the quality of generated descriptions.

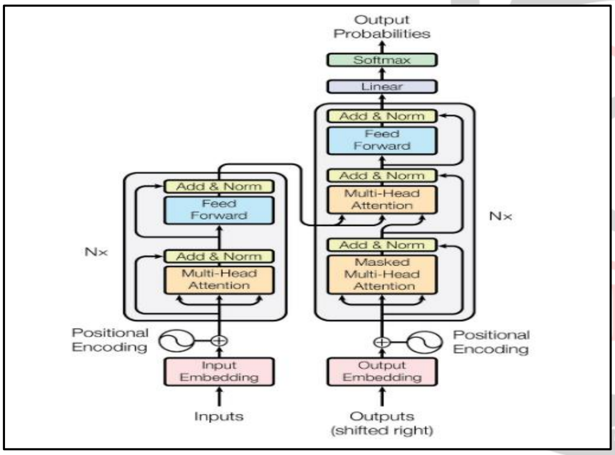


Fig 4. Design of a Transformer [10]

Description (Fig. 4): This figure illustrates the architectural structure of the Transformer model used in our image captioning framework.

C. Text–To-Speech Integration:

1. Text-To-Speech Conversion:

Leveraging the gTTS library in Python, the generated textual caption is dynamically converted into audible speech. This conversion significantly enhances the accessibility of our system, allowing users to engage with image captions through both visual and auditory modalities.

2. Audio File Saving:

The resulting audio file is saved in MP3 format, ensuring compatibility and ease of use. The file is stored in a designated directory to facilitate seamless access and retrieval.

3. Caption Display:

Simultaneously, the generated textual caption is displayed to users, providing a visual representation of the image description.

4. Audio Offering:

Our system allows users to play and pause the generated audio in real-time within the interface. Furthermore, users have the option to download the audio representation of the image caption in MP3 format. These features enhance user engagement, providing both immediate interaction and the flexibility to access audio content offline.

IV. ENVIORMENTAL SETUP

Our experiments were conducted on the Kaggle platform, utilizing an NVIDIA Tesla P100 GPU. This high-performance GPU significantly accelerated the model training process. The deep learning frameworks TensorFlow [13] and PyTorch [13] were employed for model implementation. Training parameters included stochastic gradient descent (SGD) as the optimization algorithm, with a chosen learning rate of $10.e^{-7}$. We used batch sizes of 32 for efficient training. The training process involved 20 epochs on a dataset comprising 32,650 training samples, and model performance was monitored on a validation set consisting of 8091 samples. To prevent overfitting, early stopping was applied. The best-performing models were saved for further evaluation and analysis.

V. RESULTS

In our study, we employed BLEU (Bilingual Evaluation Understudy) scores as an objective metric to assess the quality and precision of the image captions generated by our models. BLEU is a widely accepted reference-based metric, commonly used in natural language processing and machine translation tasks [14]. Its purpose is to quantify the similarity between machine-generated text and reference texts, providing a quantitative measure of how well the generated text aligns with what is expected or human-generated [17].

BLEU scores are computed at different n-gram levels, specifically ranging from 1 to 4. N-grams refer to contiguous sequences of n words. The BLEU-1 score evaluates the precision of unigrams (single words), BLEU-2 considers the precision of bigrams (two consecutive words), BLEU-3 assesses trigrams (three consecutive words), and BLEU-4 evaluates the precision of four consecutive words. By calculating these scores, we gain insights into the performance of our models at various levels of linguistic granularity.

In natural language processing and machine translation, BLEU scores serve as valuable tools for evaluating the quality of generated text, whether it be machine-generated

DOI : 10.35291/2454-9150.2024.0056

translations or, in our case, image captions. Each BLEU score measures how well the generated text aligns with the reference text, providing a nuanced understanding of the model's ability to capture not only individual words but also the coherence and structure of the language used in the expected or human-generated captions. This multi-level evaluation allows for a comprehensive analysis of the effectiveness of our models in generating image captions with varying degrees of linguistic complexity.

Here's the significance of each BLEU score:

1. BLEU-1 (Unigram Precision):

   i. BLEU-1 measures the precision of individual words (unigrams) in the generated text.

   ii. It helps evaluate how well the model is at generating individual words that match those in the reference text.

2. BLEU-2 (Bigram Precision):

   i. BLEU-2 measures the precision of word pairs (bigrams) in the generated text.

   ii. It evaluates how well the model captures the relationships and ordering of words in pairs.

3. BLEU-3 (Trigram Precision):

   i. BLEU-3 measures the precision of sequences of three words (trigrams) in the generated text.

   ii. It evaluates how well the model generates longer sequences and assesses the fluency and correctness of generated text.

4. BLEU-4 (Fourgram Precision):

   i. BLEU-4 measures the precision of four-word sequences (fourgrams) in the generated text.

The significance of these BLEU scores lies in their ability to assess different aspects of text generation, from individual words to longer sequences [18].

Table 1. Values For VGG16-RNN Model

Name	Weights	Value
BLEU-1	(1.0, 0, 0, 0)	0.521346
BLEU-2	(0.5, 0.5, 0, 0)	0.298672
BLEU-3	(0.3, 0.3, 0.3, 0)	0.215856
BLEU-4	(0.25, 0.25, 0.25, 0.25)	0.109987

Table 2. Values For InceptionV3-Transformer Model

Name	Weights	Value
BLEU-1	(1.0, 0, 0, 0)	0.755042
BLEU-2	(0.5, 0.5, 0, 0)	0.570088
BLEU-3	(0.3, 0.3, 0.3, 0)	0.713781
BLEU-4	(0.25, 0.25, 0.25, 0.25)	0.325000

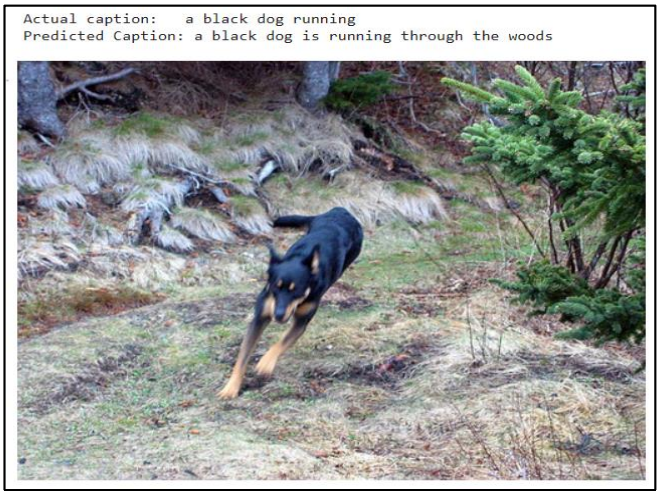
Description (Table 1, Table 2): This table compares the performance of the VGG-16 RNN and InceptionV3 Transformer models using various evaluation metrics, including BLEU score

Our evaluation of the image captioning models revealed a notable trend in the performance metrics, particularly with respect to BLEU scores. The InceptionV3-Transformer model consistently demonstrated higher BLEU scores across all n-gram levels (1-4) when compared to the VGG16-RNN model. This signifies the model's proficiency in generating captions that exhibit a closer alignment with the reference captions. The superior BLEU scores of the InceptionV3-Transformer model suggest a higher level of precision and recall, indicating the model's capability to produce more accurate and contextually relevant descriptions of images.

The Transformer architecture, with its attention mechanism, is designed to capture long-range dependencies and effectively process sequences. In the context of image captioning, this proved to be a strategic advantage. The InceptionV3-Transformer model seamlessly integrated high-level image features with textual context, producing coherent and contextually relevant captions. This integration allowed it to leverage both visual and textual information, resulting in improved caption quality [16].

While the VGG16-RNN model showcased commendable performance, particularly in capturing intricate image details, the notable difference in BLEU scores underscores the importance of the Transformer architecture's effectiveness in sequence-to-sequence tasks. These results suggest that the InceptionV3-Transformer model excels in generating fluent and accurate descriptions, positioning it as a strong candidate for image captioning applications.

These findings shed light on the strengths of the InceptionV3-Transformer model and underscore the importance of considering the specific architecture when implementing image captioning systems. Future research in this domain may explore further enhancements to capitalize on the Transformer's capabilities and continue to improve image captioning quality.



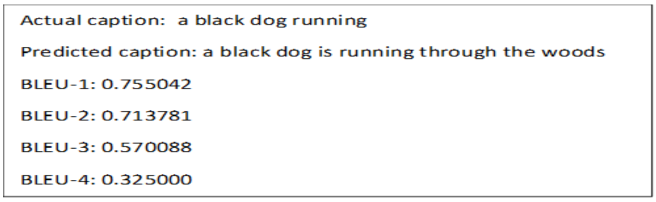
```
Actual caption:   a black dog running
Predicted Caption: a black dog is running through the woods
```

Fig 5. Example of Output generated by

InceptionV3+Transformer model.

DOI : 10.35291/2454-9150.2024.0056

Description (Fig 5): This figure presents a screenshot featuring the input image and the corresponding predicted caption generated by the INceptionV3+ Transformer model.



```
Actual caption:  a black dog running
Predicted caption: a black dog is running through the woods
BLEU-1: 0.755042
BLEU-2: 0.713781
BLEU-3: 0.570088
BLEU-4: 0.325000
```

Fig 6. Output of BLEU values generated by InceptionV3+Transformer model.

Description (Fig 6): This figure presents the BLEU scores associated with the input image and the corresponding predicted captions generated by the image captioning model.

A. *Tests And Results:*

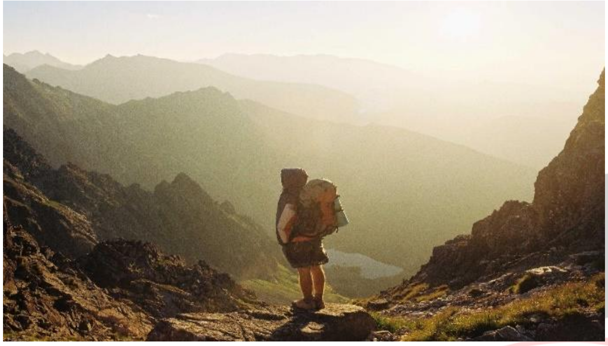


Fig 7. Test Image

Description (fig. 7): This figure showcases a selection of test image with captions generated mentioned below.

Captions Generated:

Model 1 (VGG16-RNN):

I think, It is a man stands on cliff overlooking the water

Model 2 (InceptionV3-Transformer):

I think, it is a man standing on a rock with a mountain

B. *Outputs:*

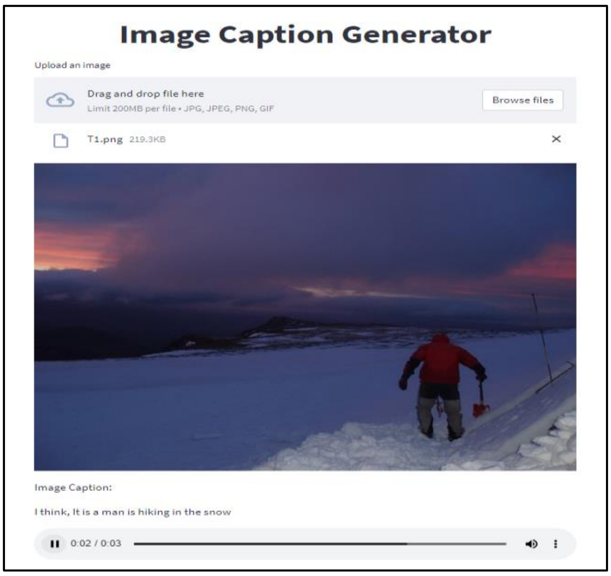


Image Caption Generator

Upload an image

Drag and drop file here

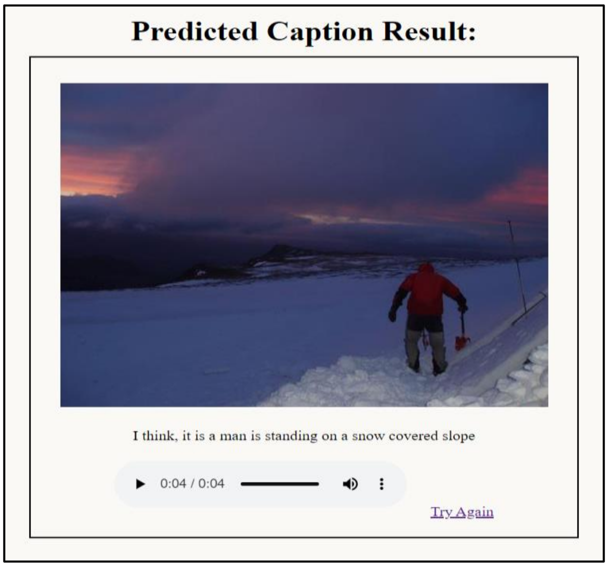
Limit 200MB per file • JPG, JPEG, PNG, GIF

T1.png 219.3KB

Image Caption:

I think, It is a man is hiking in the snow

Fig 8. Output of VGG16 and RNN Model



Predicted Caption Result:

I think, it is a man is standing on a snow covered slope

Fig 9. Output of InceptionV3 and Transformer Model

Description (Fig 8, Fig. 9): These figures depict the interface of the webpages showcasing the generated image, its corresponding caption, and an option to play the audio representation of the caption.

VI. CONCLUSION

In this study, our investigation into the VGG16-RNN and InceptionV3-Transformer models for image captioning yielded valuable insights into their performance and capabilities. The outcome of our research revealed distinct strengths inherent in each model. The VGG16-RNN model demonstrated proficiency in producing detailed and contextually relevant captions by effectively capturing intricate image features. Conversely, the InceptionV3-Transformer model exhibited remarkable prowess in generating fluent and coherent descriptions, showcasing its suitability for the image captioning task.

Looking ahead, there are several avenues for future exploration and enhancement in the field of image captioning. One potential direction is the development of hybrid architectures that combine the strengths of multiple models to further improve caption quality and versatility. Additionally, integrating external knowledge sources, such as contextual information or domain-specific knowledge, could enhance the descriptive capabilities of image captioning systems.

Furthermore, future research could focus on optimizing model training procedures and exploring novel techniques for improving efficiency and scalability. This could involve investigating advanced training algorithms, data augmentation strategies, or model compression techniques to reduce computational overhead and enhance deployment feasibility in real-world applications.

Moreover, as the field of image captioning continues to evolve, there is a growing need for benchmark datasets and standardized evaluation metrics to facilitate fair comparisons and reproducibility across different studies. Collaborative efforts to establish common benchmarks and evaluation

DOI : 10.35291/2454-9150.2024.0056

protocols could foster advancements and drive innovation in image captioning research.

Overall, our study contributes to the ongoing discourse in image captioning and lays the groundwork for future research endeavours aimed at advancing the state-of-the-art in this exciting and rapidly evolving field. By addressing these future research directions, we can further enhance the capabilities and applicability of image captioning systems, ultimately benefiting a wide range of domains and applications.

REFERENCES

[1] Krizhevsky A., Sutskever I., and Geoffrey E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, Communication of the ACM, May 2017 Vol. 60 No. 6

[2] Karpathy A., Fei-Fei L., Deep Visual Semantic Alignments for Generating Image Descriptions, IEEE Transactions on Pattern Analysis and Machine Learning, 1 April 2017, Page(s): 664 -676 Electronic ISSN: 1939-3539 Vol. 39 Issue 4

[3] Vinyals O., Toshev A., Bengio S., Erhan D., Show and Tell: A Neural Image Caption Generator, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 15 October 2015, Print ISSN:1063-6919

[4] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations (ICLR).

[5] He, J., Xu, X., Cao, J., & Lin, J. (2022). Relation-aware Transformer for Image Captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 15322-15331).

[6] Min, J., Kim, H., Kang, J., & Yoo, S. (2023). Diverse Beam Search for Image Captioning with Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1433-1442).

[7] Hossain MD. Z., Sohel F., Shiratuddin M. F., Laga H., “A Comprehensive Survey of Deep Learning for Image Captioning”, “ACM Computing Surveys”, February, 2019, vol. 51, Issue 6, 118.

[8] Mao J., Xu W., Yang Y., Wang J., Huang Z., and Yuille A., Deep Captioning with Multimodal Recurrent Neural Networks (m- RNN). In International Conference on Learning Representations (ICLR).

[9] Yang Z., Zhang Y. J., Rehman S., Huang Y., “Image Captioning with Object Detection and Localization”, International Confernece on Image and Graphics (ICIG), 2017.

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.

[11] Liu W, Chen S, Guo L, Zhu X, Liu J (2021) CPTR: full transformer network for image captioning. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2101.10804

[12] Wang Y, Xu J, Sun Y (2022) End-to-end transformer based model for image captioning. Proceedings of the AAAI Conference on Artificial Intelligence36(3):2585 2594. https://doi.org/10.1609/aaai.v36i3.20160

[13] https://keras.io/examples/vision/image_captioning/

[14] Phukan, B.B. and Panda, A.R., 2021. An efficient technique for image captioning using deep neural network. In Cognitive Informatics and Soft Computing: Proceeding of CISC 2020 (pp. 481-491). Springer Singapore.

[15] Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y. and Liu, W., 2019. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 603-612).

[16] H. Maru, T. Chandana and D. Naik, "Comparison of Image Encoder Architectures for Image Captioning," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 740-744, doi: 10.1109/ICCMC51019.2021.9418234.

[17] Staniūtė, R. and Šešok, D., 2019. A systematic literature review on image captioning. Applied Sciences, 9(10), p.2024.

[18] Ehud Reiter; A Structured Review of the Validity of BLEU. Computational Linguistics 2018; 44 (3): 393–401. doi: https://doi.org/10.1162/coli_a_00322

[19] Paigude, S., Pangarkar, S.C., Hundekari, S., Mali, M., Wanjale, K. and Dongre, Y., Potential of Artificial Intelligence in Boosting Employee Retention in the Human Resource Indust

DOI : 10.35291/2454-9150.2024.0056