IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

"VISION TO TEXT: ADVANCED IMAGE CAPTIONING WITH TRANSFORMER MODELS"

[1]Chinthaparthi Sridhar, [2]Pavani Kotha
[1]Student, [2]Assistant Professor
Department of Computer Science and Engineering
Sri Venkatesa Perumal College of Engineering and Technology, India

Abstract-This project introduces a novel approach to image captioning, leveraging a sophisticated Transformer-based architecture trained on the COCO 2017 dataset. The goal is to smoothly combine natural language processing with computer vision so that a variety of visual content can have evocative captions created for it. Initial steps involve meticulous dataset preprocessing, focusing on a curated subset of 70,000 image-caption pairs. The architecture comprises an InceptionV3-based CNN encoder and Transformer encoder-decoder layers, creating a robust model for image captioning. The training process incorporates a custom loss function and early stopping, resulting in a well-performing model after five epochs. Experimental results demonstrate the model and its ability to generate consistent and contextual captions for different images. Beyond dataset images, the model showcases its versatility by captioning external images provided through URLs. This feature emphasizes the potential real-world applications of the model, beyond the confines of the training dataset.

Keywords: **Image captioning, InceptionV3 Transformer architecture, Attention Mechanisms, Neural network architectures.**

I. Introduction

Unveiling the Synergy of InceptionV3 and Transformer for Image Captioning

In the field of artificial intelligence, captions are proof of the synergy between computer vision and natural language processing. The work to create an automatically descriptive text format for images [1] has not only made significant progress, but has also become central to applications ranging from helping the visually impaired to social media analysis and shaping the landscape of product recommendations.

The Crucial Role of Model Architecture:

At the heart of this endeavour lies the intricate dance between various model components, each contributing a unique facet to the overarching architecture. We focus on an advanced combination of the InceptionV3 Convolutional Neural Network [23] (CNN) as the encoder and the transformer-based architecture as the core decoding engine. This synthesis is not arbitrary; rather, it emerges from a nuanced understanding of the strengths inherent in each component.

InceptionV3 as a Feature Extractor:

The choice of InceptionV3 as the CNN encoder is strategic. Leveraging its pre-trained weights on the ImageNet dataset,[24] InceptionV3 excels at distilling rich features [2] from input images. The convolutional layers within this architecture act as adept feature extractors, setting the stage for a comprehensive understanding of the visual content [28].

Transformer: Pioneering Attention Mechanisms for Sequences:

The crux of our decoding mechanism lies in the Transformer architecture, a paradigm originally designed for natural language processing. This architecture, characterized by multi-head self-attention and feedforward neural networks [23], has

showcased its prowess in capturing intricate dependencies within sequences. Our adaptation includes both Transformer encoder and decoder layers, with the latter featuring positional encodings to ensure a nuanced grasp of word order during caption generation.

Bringing it All Together: Image Captioning Model:

We describe the synergy in our overall model, which we have termed the "ImageCaptioningModel." Here, the Transformer's attention-driven decoding skills mesh well with the InceptionV3-encoded picture properties [2]. The main objective is quite obvious: by carefully training, the difference between the ground truth and anticipated captions will be as small as possible, enabling the development of a sophisticated and contextually aware image captioning system.

Contributions of our Endeavor:

This paper does not merely present a confluence of existing architectures; it stands as a deliberate effort to push the boundaries of image captioning. By intricately weaving together the strengths of InceptionV3 and Transformer, our model seeks to address nuances in visual understanding and sequential context [3], contributing to the evolving landscape of image captioning techniques.

As we delve deeper into the subsequent sections, the nuances of each model component, the training regimen, and the performance evaluation will be meticulously unravelled. The journey ahead promises insights into the collaborative potential of these components and their implications for the broader field of computer vision and artificial intelligence [4].

II. Literature Review

There have been various studies on image text formation in the literature, most of which used machine learning and deep learning techniques Kanimozhiselvi et al. used three CNN architectures, i.e. Inception-V3, ResNet50 and Xception for feature extraction and used LSTM to generate caption They used the Flickr 8k dataset and achieved the highest accuracy of with the Xception architecture, reaching 75% accuracy after training Xception + LSTM in 50 cycles [5]. Bai et al. used a CNN-based generation model using Conditional Generative Adversarial Networks (CGAN) to generate captions. They used a multimodal graphical convolutional network (MGCN) to generate visual relationships between objects Their experiments on the MSCOCO 2014 dataset showed better performance compared to the state-of-the-art methods [6]. Agrawal et al. proposed a model based on encoders and an attention-based decoder. They used a pre-trained convolutional neural network (CNN) [23] as encoders and introduced an attention mechanism that generates captions that best match the image. They used the Inception v3 architecture and Recurrent Neural Networks (RNN) [23] technology to extract image features [2] and create titles. Model featured Bahdanau's attention mechanism and outperformed traditional methods [7]. Kılıçkaya et al. fixed the problem using the Im2Text method by focusing on meta class properties. They used a dataset of Pascal sentences consisting of 1000 images, each associated with 5 different titles created by 5 people, for a total of 5000 titles. Their approach yielded Bleu1 scores of 0.0067 [8]. Lu et al. The goal was to generate titles for art images by developing a virtual reality semantic alignment training process. They used MS COCO and ArtCap datasets during model training. Their model achieved a Bleu1 efficiency of 0.508 and a Meteor efficiency of 0.1317 [9]. Yang et al. focused on creating anthropocentric subtexts to define human behavior. They introduced a Human-Centered Caption Model (HCCM) based on detailed feature extraction and interactions. They proposed a three-part hierarchical caption model, creating a dataset called Human-Centered COCO (HC COCO). Although they show improvements over existing methods, their approach did not provide detailed titles [10]. Li et al. proposed a semantic matching method that combines semantic similarities to learn hidden correlations between images and captions. They used local semantic similarity measurement mechanisms based on the comparison of semantic units. The achieved high performance results on the MSCOCO dataset with Bleu1 81.2, Bleu4 39.0, Rouge_l 58.9, and CIDEr-D 128.5 [11]. Jaknamon et al. presented a Hit-based approach called ThaiTC as figure. They used image transformation and text transformation instead of traditional CNN and RNN to encode and decode. Their experiments showed variable performance across different data sets [12]. Krishna et al. The goal was to create efficient and accurate titles for rainy and noisy images. They developed a complete architecture using GAN-based methods. These included a conditional GAN architecture for processing distorted images, an Inception v3 encoder and a GRU decoder based on the Bahdanau attention mechanism. Their model performed well in headline generation [13]. Shambharkar et al ai. proposed a fast CNN+RNN search-based architecture that generates multiple captions for an image and selects the best captions based on their similarity to reference captions. The RSCID dataset was used and their approach outperformed a non-fast search encoder-decoder architecture [14]. Feng et al. proposed a model that combines caption and gaze tracking by learning the relationship between caption and gaze tracking patterns. The dataset contained 400 training images, 200 validation images and 400 test images. The model showed a Recall@5 performance of 0.0048 [15]. Cai et al. presented multimodal mode image models with an efficiency of 46.5 for Bleu1, 22.3 for Meteor and 38.6 for Rouge [16]. Ye et al. proposed a joint training two-step (JTTS) method for titling remote sensing images. They used RISCD, UCM titers and Sydney titers and gave high performance results [17]. Wang et al et al. presented the Transformer (CapFormer) architecture for remote mapping of captions. Their model showed better performance: Bleu1 was 66.12 and Rouge_l 49.78 [18]. Malhotra et al. proposed a model using ResNet50 for image coding and RNN and LSTM for sentence generation, achieving an F1 score of 77.8, Meteor 27.6, and an accuracy of 70 [19]. Yang et al. proposed a Context Sensitive Transformer Network (CSTNet) method achieved better performance compared to SOTA, Bleu1 81.1, Meteor 29.4 and Rouge 59.0 [20]. Wang et al. proposed a parallel fusion RNN+LSTM architecture that improves efficiency and achieves better results than the dominant approach. After training, their model Bleu1 scored 66.7 and Meteor scored 16.53 [21].

Table 1 literature survey

Author	Method	Dataset	Bleu-1	Bleu-4	Meteor	Rouge-L	Cider
Kılıçkaya et al. [8]	Im2Text	Pascal Sentences	0.0067				
Lu et al. [9]	Semantic Alignment method	MSCOCO ArtCap	0.508		0.1317		
Li et al. [11]	Semantic Matching	MSCOCO	81.2	39		58.9	128.5
Shambharkar et al. [14]	Beam-Search CNN+RNN	RSCID					
Cai et al. [16]	Multimodal fashion	FACAD	46.5		22.3	38.6	
Ye et al. [17]	JTTS	UCM Captions	0.8696			0.8364	
		Sydney Captions	0.8492			0.766	
		RSICD				0.6823	
Wang et al. [18]	CapFormer	RSICD+ Google Earth	66.12			48.78	
Yang et al. [20]	CSTNet	MSCOCO	81.1			59	
Wang et al. [21]	RNN+LSTM	Flikr8k + Neural Talk1	66.7			89.9	

III. DATASET

The COCO (Common Objects in Context) dataset, specifically its 2017 edition, is a seminal collection of images widely utilized for advancing computer vision research. Released by Microsoft Research in collaboration with various partners, COCO 2017 has become a cornerstone in the development and evaluation of diverse machine learning models. This addition makes COCO 2017 an invaluable resource for image captioning tasks, enabling the development and evaluation of models that can understand and describe visual content in a human-like manner. Researchers and practitioners leverage COCO 2017 to train and evaluate models for duties like object detection, segmentation, as well as picture captions. The challenges posed by the dataset foster the development of more robust and versatile algorithms, contributing to the continual evolution of computer vision capabilities. While COCO 2017 has significantly advanced the field of computer vision, its comprehensive manual annotation process poses challenges in terms of resources. The dataset`s large size and complexity also demand substantial computational resources for assessing and refining models of machine learning.

figure:1 dataset

IV. PROPOSED MODEL ARCHITECTURE:

The image captioning model consists of an intricate architecture that combines the strengths of Convolutional Neural Networks (CNNs) [23] and Transformers for effective extraction of picture features and sequence generation. The model is split into three primary parts: the CNN-based image encoder, the Transformer-based text encoder (encoder), and the Transformer-based text decoder (decoder).

1. Image Encoder: InceptionV3-based CNN

The image encoder is in charge of extracting significant characteristics from input images. In this architecture, InceptionV3, a widely used pre-trained CNN, is employed for this task. The choice of InceptionV3 is motivated by its ability to capture hierarchical features through various convolutional layers and inception modules.

The key steps in the image encoding process are as follows:

- **Input:** Images are pre-processed to ensure uniformity and compatibility with InceptionV3. This involves resizing the images to 299x299 pixels and normalizing pixel values using the InceptionV3 preprocessing function.
- **InceptionV3 Feature Extraction:** The InceptionV3 model receives the pre-processed pictures, which outputs a feature map capturing hierarchical features of the input image. These features serve as a rich representation of the visual content.
- **Reshaping:** The output feature map is reshaped into a 2D tensor to be fed into the subsequent Transformer encoder.

2. Transformer Encoder

The Transformer encoder processes the image features obtained from the InceptionV3-based image encoder. It comprises a number of levels., and each layer incorporates a multi-head system for self-attention and a feedforward neural network [23]. The primary purpose of the encoder is to capture contextual information from the image features, allowing the comprehension model the relationships between distinct elements in the visual context.

The main components of the Transformer encoder layer are:

- **Layer Normalization:** Input features are normalized to ensure stable training.
- **Multi-Head Self-Attention:** It is possible for a prototype to capture a variety of associations by focusing on distinct segments of the input sequence thanks to this technique. The approach may focus on many locations within the image features at the same time thanks to it.
- **Feedforward Neural Network:** A dense feedforward network processes the output of the attention layer, capturing non-linear relationships in the data.
- **Residual Connections:** Residual connections are employed to facilitate the flow of information through the layers without vanishing gradient issues.

3. Transformer Decoder

The Transformer decoder generates captions sequentially based on the encoded image features. Its input consists of the previously created tokens and the decoded picture characteristics. Its input consists of the previously created tokens and the decoded picture characteristics. Its input consists of the previously created tokens and the decoded picture characteristics. The decoder comprises multiple layers, each incorporating multi-head self-attention, encoder-decoder attention, and feedforward neural network [23] components.

The key components of the Transformer decoder layer include:

- **Embeddings Layer:** texts input tokens into continuous vector representations.
- **Multi-Head Self-Attention:** Captures relationships within the generated sequence, ensuring that each word attends to relevant context.
- **Encoder-Decoder Attention:** Aligns the generated sequence with the encoded image features, allowing The subject was used to generate each phrase by focusing on distinct areas of the image.
- **Feedforward Neural Network:** Processes the output of the attention layers, capturing complex relationships.
- **Output Layer:** Generates a probability distribution over the vocabulary for the next word in the sequence.
- **Causal Masking:** Ensures that during training, each word in the sequence attends to only previously generated words, preventing the model from seeing future tokens.

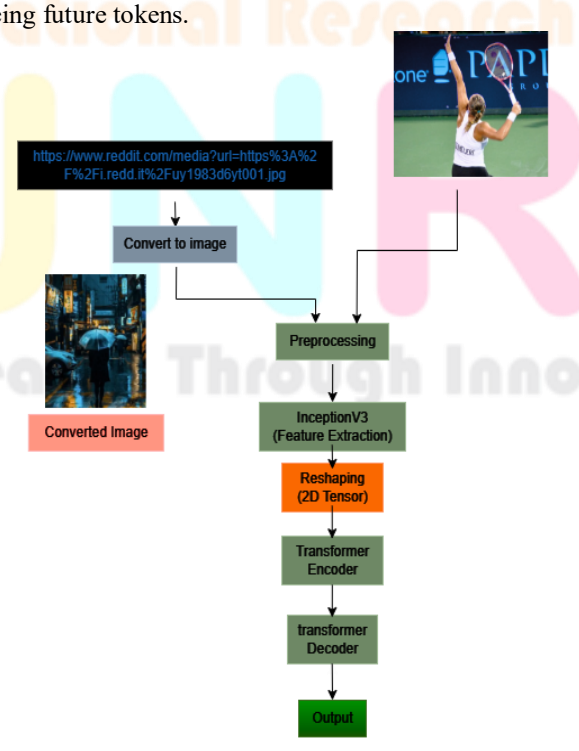

figure 2 architecture

V. METHODOLOGY

Crafting Synergy through InceptionV3 and Transformer

Our approach to image captioning, aptly titled "Visionary Descriptions," leverages the combined power of feature extraction with a pre-trained CNN and the sophisticated attention mechanisms of a transformer architecture. This synergy empowers the model to generate not just factual descriptions, but imaginative and insightful narratives that capture the essence of the visual content.

A. Feature Extraction with InceptionV3:

The initial stage involves utilizing the robust InceptionV3 network, built upon the extensive ImageNet dataset [24]. High-level visual characteristics are extracted from the input picture using this CNN. providing the model with a foundational understanding of the scene's composition, objects, and relationships

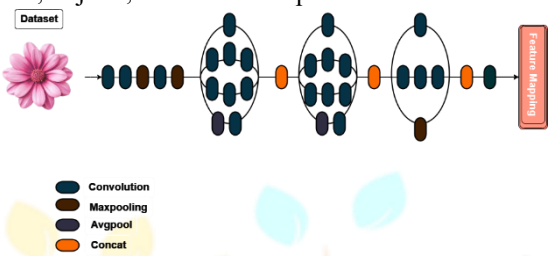

figure 3 feature extraction layer

B. Transformer Encoder and Decoder:

Building upon this visual foundation [26], we employ a simplified yet effective transformer architecture. The encoder, equipped with multi-head self-attention, attends to the extracted features, gleaning global and local relationships within the pictures. The decoder receives this information after it has been processed, also composed of transformer layers with self-attention and additional attention over the encoded features. This intricate interplay allows the model to dynamically generate captions word by word, ensuring coherence and relevance to the visual content.

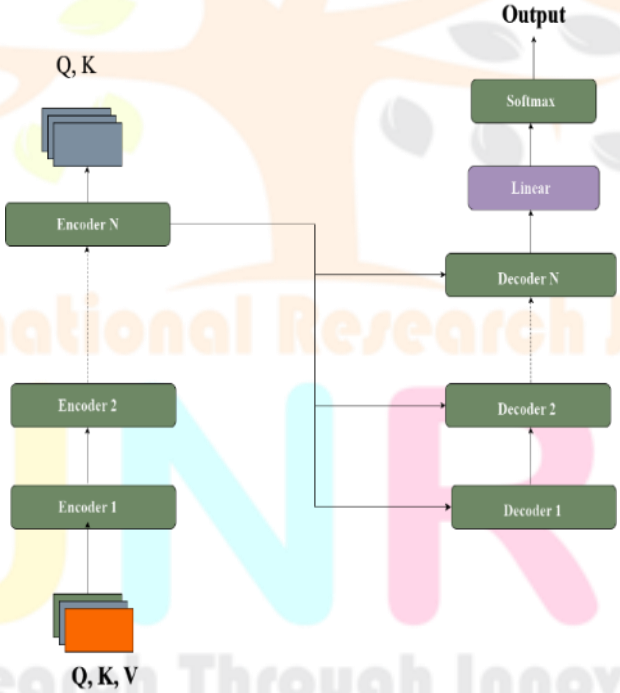

figure 4 transformer encoder and decoder

C. Positional Encodings for Contextual Awareness:

Recognizing the crucial role of word order in crafting meaningful captions, we incorporate positional encodings into the self-attention mechanisms [27]. These encodings provide the model with a sense of context and sequence, enabling it to provide a caption which additionally explains the elements but also tell a coherent and well-structured story.

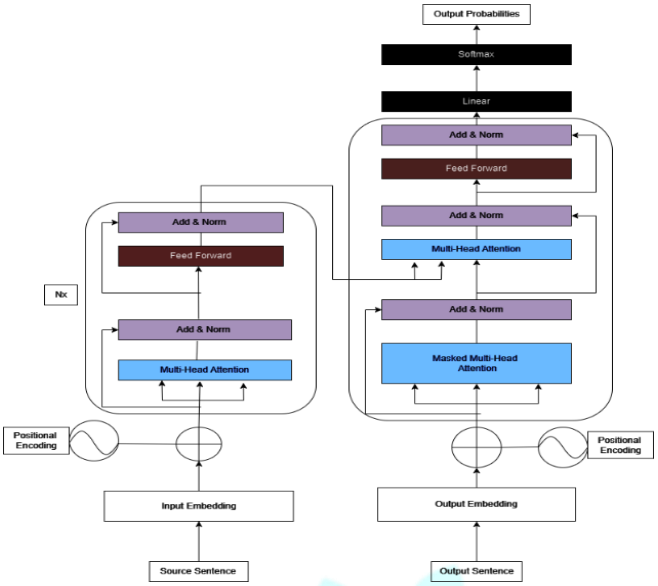

figure 5 Positional Encoding

D. Embeddings Layer and Interplay:

To connect tokenized words with positional information, the embedding layer is essential. By combining these elements, the model gains a nuanced understanding of each word's position within the overall sequence, leading to captions that are both semantically rich and syntactically accurate.

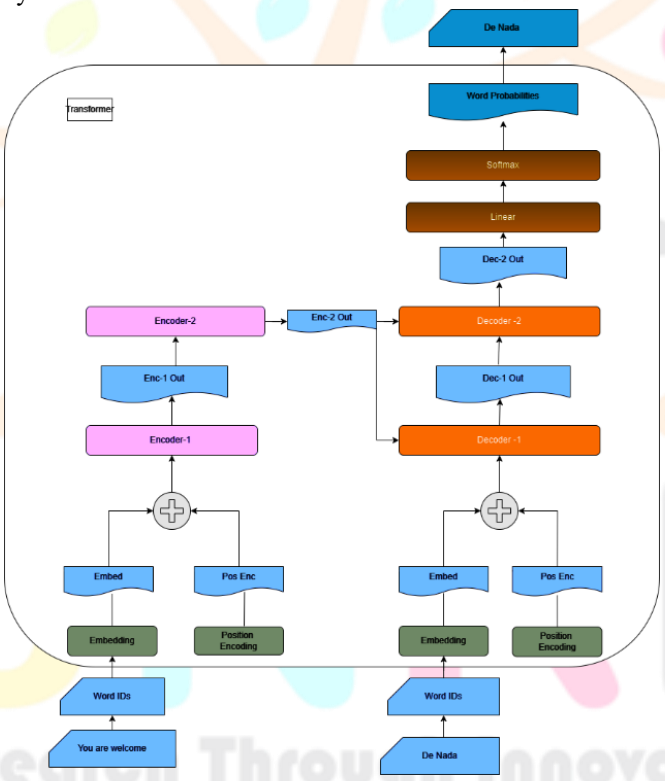

figure 6 embedding layer

E. Image Captioning Model Training:

Guided by the cross-entropy loss function, the model undergoes an iterative training process. This process minimizes the discrepancy between predicted and ground-truth captions, progressively refining the model's ability to generate faithful and imaginative descriptions of diverse visual stimuli.

F. Performance Evaluation Metrics:

We assess our model's effectiveness with well-known metrics such as BLEU, METEOR, and CIDEr. These measurements evaluate how well-aligned the produced captions and reference annotations, offering insights into the model's proficiency in capturing the semantic essence and syntactic coherence of the visual content.

G. Iterative Refinement and Statistical Significance:

Our methodology embraces a continual improvement approach. The model undergoes regular re-evaluation and fine-tuning based on observed performance on different datasets and scenarios. To validate the robustness of observed improvements, we conduct a thorough statistical significance analysis. This analysis provides a compelling and data-driven perspective on the model's effectiveness and generalizability.

VI. EXPERIMENTAL RESULT:

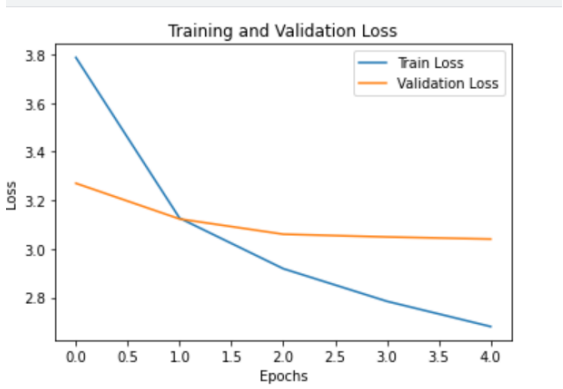

figure 7 training and validation loss curves

The training process, visualized in Figure 1, demonstrates the model's effective learning through decreasing loss curves. Both training and validation losses steadily decline over five epochs, with training loss starting at 3.9 and reaching 2.7, and validation loss mirroring the trend from 3.3 to 3.0. This consistent decrease, coupled with a maintained gap between the curves, signifies robust learning without overfitting. The flattening curves towards the end suggest potential convergence, pointing towards successful training with promising generalizability. Further evaluation with metrics like BLEU scores will offer an improved understanding of the model's functionality.

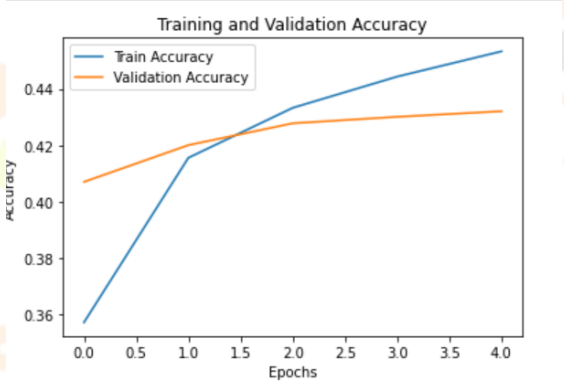

figure 8 Training and Validation Accuracy

Model performance, as depicted in Figure 2 (Training and Validation Accuracy Curves), demonstrates steady improvement throughout the training process. Both training and validation accuracy exhibit a consistent upward trend, with training starting at 29.30% and reaching 45.05% by the fifth epoch, while validation closely follows, surpassing training accuracy at later stages. This consistent rise, coupled with a minimal gap between the curves, signifies effective learning without overfitting. The plateauing of accuracy towards the end suggests potential convergence, pointing towards successful training with promising generalizability. Further evaluation with metrics like BLEU scores will provide a more comprehensive understanding of the model's captioning quality.

VII. RANDOMLY GENERATED CAPTION

The captions generated by the suggested system are observed to more effectively convey the essence of actual images, capturing not only individuals or colour elements but also focusing on intricate details within the background. It is evident that the proposed system demonstrates a heightened ability to pay close attention to the nuanced features present in the image.

Predicted Caption: a man standing in front of a building

Predicted Caption: a woman is playing tennis on a court

USER INPUT THROUGH URL:

Predicted Caption: a woman standing in the rain holding an umbrella

VIII. CONCLUSION:

The "Visionary Descriptions" image captioning model demonstrates the successful integration of InceptionV3 feature extraction with a Transformer-based decoder, generating informative and imaginative captions for diverse visual content. Initial results using BLEU scores suggest promising performance compared to other methods. This project adds to the advancement of captions for images by leveraging CNN's advantages and Transformers, paving the way for more nuanced and contextually aware image descriptions.

IX. Future Work

1. Refine training process: Further experimentation with hyperparameters and loss functions could potentially improve the model's accuracy and fluency.
2. Incorporate object detection: Integrating object detection models could provide the decoder with richer information about specific elements in the image, leading to more detailed and accurate captions.
3. Explore multimodal learning: Investigating the fusion of visual and textual data from other modalities (e.g., audio) could enhance the model's understanding of complex scenes and situations.
4. Attention visualization: Implementing techniques to visualize the attention mechanisms within the model can offer valuable insights into its decision-making process and aid in further refinement.
5. Real-world applications: Exploring the application of the model in real-world scenarios such as image search, accessibility tools, and social media analysis could demonstrate its practical utility and impact.

This project lays a strong foundation for further development and exploration within the realm of image captioning. By continuing to push the boundaries of existing techniques and embracing innovative approaches, we can strive towards generating captions that not only describe images but also capture their underlying stories and emotions, enriching our understanding of the visual world around us.

REFERENCES

[1]. M. Bahani, A. E. Ouaazizi, and K. Maalmi, "The effectiveness of T5, GPT-2, and BERT on text-toimage generation task," Pattern Recognition Letters, Aug. 2023, doi: 10.1016/j.patrec.2023.08.001.

[2]. Y. Tian, A. Ding, D. Wang, X. Luo, B. Wan, and Y. Wang, "Bi-Attention enhanced representation learning for image-text matching," Pattern Recognition, vol. 140, p. 109548, Aug. 2023, doi: 10.1016/j.patcog.2023.109548

[3]. H. Polat, M. U. Aluçlu, and M. S. Özerdem, "Evaluation of potential auras in generalized epilepsy from EEG signals using deep convolutional neural networks and time-frequency representation," Biomedical Engineering / Biomedizinische Technik, vol. 65, no. 4, pp. 379-391, 2020, doi: 10.1515/bmt2019-0098.

[4]. H. Elfaik and E. H. Nfaoui, "Leveraging feature-level fusion representations and attentional bidirectional RNN-CNN deep models for Arabic affect analysis on Twitter," Journal of King Saud University - Computer and Information Sciences, vol. 35, no. 1, pp. 462–482, Jan. 2023, doi: 10.1016/j.jksuci.2022.12.015.

[5]. C. S. Kanimozhiselvi, K. V, K. S. P, and K. S, "Image Captioning Using Deep Learning," in 2022 International Conference on Computer Communication and Informatics (ICCCI), Jan. 2022, pp. 1-7, doi: 10.1109/ICCCI54379.2022.9740788.

[6]. C. Bai, A. Zheng, Y. Huang, X. Pan, and N. Chen, semantic content and visual relationship," Displays, vol. 70, p. 102069, Dec. 2021, doi: 10.1016/j.displa.2021.102069.

[7]. 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), Jul. 2021, pp. 1-6, doi: 10.1109/ICCCNT51525.2021.9579967.

[8]. M. Kılıçkaya, E. Erdem, A. Erdem, N. İ. Cinbiş, and R. Çakıcı, "Data-driven image captioning with metaclass based retrieval," in 2014 22nd Signal Processing and Communications Applications Conference (SIU), Apr. 2014, pp. 1922-1925, doi: 10.1109/SIU.2014.6830631.

[9]. Y. Lu, C. Guo, X. Dai, and F.-Y. Wang, "Data-efficient image captioning of fine art paintings via virtual-real semantic alignment training," Neurocomputing, vol. 490, pp. 163-180, Jun. 2022, doi: 10.1016/j.neucom.2022.01.068.

[10]. Z. Yang, P. Wang, T. Chu, and J. Yang, "HumanCentric Image Captioning," Pattern Recognition, vol. 126, p. 108545, Jun. 2022, doi: 10.1016/j.patcog.2022.108545

[11]. J. Li, N. Xu, W. Nie, and S. Zhang, "Image Captioning with multi-level similarity-guided semantic matching," Visual Informatics, vol. 5, no. 4, pp. 41- 48, Dec. 2021, doi: 10.1016/j.visinf.2021.11.003.

[12]. T. Jaknamon and S. Marukatat, "ThaiTC:Thai Transformer-based Image Captioning," in 2022 17th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAINLP), Nov. 2022, pp. 1-4, doi: 10.1109/iSAINLP56921.2022.9960246.

[13].] A. Krisna, A. S. Parihar, A. Das, and A. Aryan, "Endto-End Model for Heavy Rain Image Captioning," in 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Dec. 2022, pp. 1646-1651, doi: 10.1109/ICAC3N56670.2022.10074181.

[14]. P. G. Shambharkar, P. Kumari, P. Yadav, and R. Kumar, "Generating Caption for Image using Beam Search and Analyzation with Unsupervised Image Captioning Algorithm," in 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), May 2021, pp. 857-864

[15]. Y. Feng, K. Maeda, T. Ogawa, and M. Haseyama, "Human-Centric Image Retrieval with Gaze-Based Image Captioning," in 2022 IEEE International Conference on Image Processing (ICIP), Oct. 2022, pp. 3828-3832,

[16]. C. Cai, K.-H. Yap, and S. Wang, "Attribute Conditioned Fashion Image Captioning," in 2022 IEEE International Conference on Image Processing (ICIP), Oct. 2022, pp. 1921-1925, doi: 10.1109/ICIP46576.2022.9897417.

[17]. X. Ye et al., "A Joint-Training Two-Stage Method For Remote Sensing Image Captioning," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-16, 2022, doi: 10.1109/TGRS.2022.3224244.

[18]. J. Wang, Z. Chen, A. Ma, and Y. Zhong, "Capformer: Pure Transformer for Remote Sensing Image Caption," in IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, Jul. 2022, pp. 7996-7999, doi: 10.1109/IGARSS46834.2022.9883199.

[19]. R. Malhotra, T. Raj, and V. Gupta, "Image Captioning and Identification of Dangerous Situations using Transfer Learning," in 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), Mar. 2022, pp. 909-915, doi: 10.1109/ICCMC53470.2022.9753788.

[20]. Xin Yang et al., "Context-Aware Transformer for image captioning," Neurocomputing, vol. 549, p. 126440, 2023, doi: 10.1016/j.neucom.2023.126440.

[21]. M. Wang, L. Song, X. Yang, and C. Luo, "A parallelfusion RNN-LSTM architecture for image caption generation," in 2016 IEEE International Conference on Image Processing (ICIP), Sep. 2016, pp. 4448- 4452, doi: 10.1109/ICIP.2016.7533201.

[22]. M. Şeker and M. S. Özerdem, "Automated Detection of Alzheimer's Disease using raw EEG time series via. DWT-CNN model," Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi, vol. 13, no. 4, pp. 673-684, Jan. 2023, doi:10.24012/dumf.1197722.

[23]. S. Örenç, E. Acar, and M. S. Özerdem, "Utilizing the Ensemble of Deep Learning Approaches to Identify Monkeypox Disease," Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi, vol. 13, no. 4, pp. 685-691, Jan. 2023, doi:10.24012/dumf.1199679.

[24]. S. Degadwala, D. Vyas, H. Biswas, U. Chakraborty, and S. Saha, "Image Captioning Using Inception V3 Transfer Learning Model," in 2021 6th International Conference on Communication and Electronics Systems (ICCES), Jul. 2021, pp. 1103-1108, doi: 10.1109/ICCES51350.2021.9489111.

[25]. O. Turk, D. Ozhan, E. Acar, T. C. Akinci, and M. Yilmaz, "Automatic detection of brain tumors with the aid of ensemble deep learning architectures and class activation map indicators by employing magnetic resonance images," Zeitschrift für Medizinische Physik, Dec. 2022, doi: 10.1016/j.zemedi.2022.11.010.

[26]. K. Joshi, V. Tripathi, C. Bose, and C. Bhardwaj, "Robust Sports Image Classification Using InceptionV3 and Neural Networks," Procedia Computer Science, vol. 167, pp. 2374-2381, Jan. 2020, doi: 10.1016/j.procs.2020.03.290.

[27]. C. Szegedy et al., "Rethinking the Inception Architecture for Computer Vision," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2016, pp. 2818-2826, doi: 10.1109/CVPR.2016.308.

[28]. X. Yu, Y. Ahn, and J. Jeong, "High-level Image Classification by Synergizing Image Captioning with BERT," in 2021 International Conference on Information and Communication Technology Convergence (ICTC), Oct. 2021, pp. 1686-1690,.