Enhanced Image Captioning for Social Media using Inception V3and Transformer Networks

Maheshwaran. T[1], Ragul. K[2], Monish Coumar. S[3], Narendheran. A[4]

Sri Manakula Vinayagar Engineering College, Puducherry, India smvec@smvec.ac.in

Abstract:

In this digital age, social media is flooded with tons of visual content, automatic image captioning is a must for better content accessibility and engagement. This project is an advanced approach to image captioning using Inception V3 for feature extraction and Transformer for natural language generation. Inception V3 is known for its object detection capabilities, it serves as the backbone to capture the fine details of the image while Transformer with its attention mechanism generates contextually rich and coherent captions. Our system aims to provide more accurate and context sensitive descriptions for social media images, to improve user experience and content discoverability. We used Flickr8k dataset for training and testing, we show the effectiveness of this hybrid model in handling complex scenes, multiple objects and varying context. The proposed solution is a scalable and efficient way to generatecaptions that will enhance social media interaction and accessibility.

Keywords: Image Captioning, Inception V3, Transfomer Network.

1 Introduction

With the rapid growth of social media, vast amounts of visual content are shared daily, making it challenging to provide accessible and contextually meaningful descriptions for images. Manually generating captions for such a large volume of content is labor-intensive, inconsistent, and often fails to capture the intricate details of images. This creates a need for automated image captioning systems that can generate accurate and descriptive captions without human intervention. However, traditional models have struggled with capturing both visual features and contextual understanding, often resulting in vague or irrelevant captions. To address this challenge, we propose a hybrid approach that combines the strengths of Inception V3 and Transformer networks. Inception V3, a powerful convolutional neural network, is employed for robust feature extraction, capturing detailed visual elements from images. Meanwhile, the Transformer network, known for its efficient attention mechanisms, processes these features to generate contextually accurate captions in natural language. By leveraging these two models, we aim to improve the quality and relevance of image captions, especially in complex scenarios with multiple objects or subtle contexts. This approach offers a scalable and efficient solution for automatically captioning images on social media platforms, enhancing both content accessibility and user engagement while reducing the need for manual intervention.

2 Related Work

1 **Roshni Padate. [1]** This paper presents a novel model for automatic image captioning that combines low-level features like contrast and color with high-level features such as motion and facial expressions. Using an optimized CNN fine-tuned with the Spider Monkey Optimization algorithm (SMO-SCME), the model improves caption accuracy. It bridges computer vision and natural language processing, outperforming existing methods in generating rich, meaningfuldescriptions.

2 **Damsara Ranasinghe. [2]** This study introduces a deep learning model for generating image captions in Sinhala, utilizing an RNN with InceptionV3 for feature extraction and LSTM for language processing. Evaluating on the Flickr8K and MS COCO datasets, it finds human- entered captions outperform Google-translated ones. Future work will focus on improving accuracy by using larger datasets and exploring alternative architectures.

3 **Brandon Birmingham. [3]** The paper presents KENGIC, an innovative approach for image captioning that relies on keyword-driven, n-gram graph techniques to improve both the quality and explainability of captions without needing paired datasets. By testing various keyword sets and evaluation metrics, KENGIC proves competitive with top models in the field. The study also addresses some limitations in current evaluation metrics and tackles challenges like hallucination and synonym issues in the generated captions.

4 **Madhuri Bhalekar. [4]** The paper introduces D-CNN, a new image captioning model designed to assist visually challenged individuals by generating detailed captions and extracting textual information from images. Combining Convolutional Neural Networks (CNNs) for feature extraction with Long Short-Term Memory (LSTM) networks for generating captions, D-CNN outperforms existing models and has the potential to be implemented as a mobile application.

5 **M. Poongodi. [5]** The paper presents a novel approach for automated image and audio captioning using deep learning, combining computer vision with natural language processing. It achieves a Top 5 accuracy of 67% and Top 1 accuracy of 53%, with the goal of aiding visually impaired individuals. The research addresses challenges in integrating image and sound, proposing a joint model with promising applications in social networks and content moderation.

6 **Rita Ramos. [6]** The paper introduces S MALL C AP, a lightweight image captioning model that boosts performance through retrieval augmentation. Using a pre-trained CLIP encoder and GPT-2 decoder, the model enables quick training and adaptability across various domains without needing retraining. It shows competitive results on the COCO dataset and performs exceptionally well in out-of-domain evaluations by utilizing diverse data sources for enhancedcaptioning.

7 **Sampath Boopath. [7]** The paper reviews deep learning techniques for automatic sentence generation and sentiment analysis, highlighting models such as RNNs, LSTMs, and GRUs. It tackles challenges in natural language processing, including ambiguity and named entity recognition, and underscores the importance of human evaluation for assessing text quality. Future research will focus on improving the efficiency of text generation systems.

8 **Teng Wang. [8]** The paper introduces "Caption Anything" (CAT), a framework for controllable image captioning that blends visual and language controls to produce personalized image descriptions. Leveraging pre-trained models, CAT enables interactive user engagement through visual prompts and language styles. With components like a segmenter, captioner, and text refiner, CAT effectively generates accurate and contextually relevant captions for various applications.

9 **JianjieLuo. [9]** The paper introduces Semantic-Conditional Diffusion Networks (SCD-Net), a groundbreaking non-autoregressive model for image captioning that improves visual-language alignment through a diffusion process and semantic priors. By utilizing guided self-critical sequence training, SCD-Net achieves superior performance on the COCO dataset compared to existing methods, highlighting the effectiveness of diffusion models in producing high-qualityimage captions.

10 **Pang-Jo Chun. [10]** The paper presents a deep learning model that integrates CNN and GRU with an attention mechanism to automatically generate explanatory texts for bridge damage from images. Trained on a large dataset, the model accurately describes damage, supporting maintenance engineers and improving bridge inspection efficiency. The study highlights its practical applications in civil engineering, showcasing its potential to enhance bridge maintenance processes.

3 **Table 1 : Analysis Table**

S.No	**Paper Title -Author Name**	**Dataset**	**Merits**	**Demerits**
1.	Image Captioning using Wavelet transform-based CNN with Attention Mechanisms and LSTM Networks -Reshmi Sasibhooshan	Flickr8k dataset	Enhancedfeature extraction with attention mechanisms	Object detection errors and complexity
2.	A deep learning-based image captioning method to automatically generate explanatory texts for bridge damage - **JianjieLuo**.	Photos from 3118 bridges inspected by Kanto Regional Development Bureau, Japan.	High accuracy in generating bridge damage explanations, aiding maintenance and improving inspection efficiency.	Existing datasets lack civil engineering-specific explanations, causing inaccuracies with standardmodels.

3.	Semantic-Conditional Diffusion Networks for Image Captioning - **Pang-Jo Chun**	COCO (82,783 training, 40,504 validation, 40,775 testing images)	SCD-Net outperforms existing non-autoregressive methods and matches ensemble autoregressive methods, highlighting the effectiveness of semantic conditioning in image captioning diffusion models.	Non-autoregressive methods like SCD-Net can suffer from issues such as word repetition or omissions, inherent to the approach.
4.	Deep Learning-Techniques Applied for Automatic - Sentence Generation - **Samapth Boopathi**	Data is news articles, the generated text should also read like news articles	The quality of generated text is assessed using metrics like coherence, diversity, grammar, syntax, style, and through human evaluation.	No single metric fully captures text quality, making it challenging to evaluate text generation methods.
5.	Caption Anything: InteractiveImage Description with Diverse Multimodal Controls - **Teng Wang**	Segmentation dataset with 1 billion masks on 11 million images	The Caption AnyThing (CAT) framework offers flexible visual controls and aligns with user intent by integrating pre-trained image captioners and an instruction- tuned LLM.	Existing controllableimage captioning models often depend on human-annotated (image, text, control signal) tuples, limiting their flexibility and control signal comprehension due to the small dataset scale.

4 Conclusion

In conclusion, this project aims to enhance the automatic image captioning process for social media platforms by leveraging a hybrid approach combining Inception V3 and Transformer networks. By integrating Inception V3 for detailed image feature extraction with Transformer networks for generating rich, contextually accurate captions, the project addresses key challenges in image captioning, such as complex scenes and multiple objects. The use of the Flickr8k dataset ensures that the model is trained on a diverse range of images, improving its ability to handle varying contexts and enhance content discoverability. This approach not only boosts the quality of image descriptions but also offers a scalable solution for improving user engagement and interaction on social media. The successful implementation of this model represents a significant step forward in creating more accessible and engaging visual content, contributing to advancements in the field ofautomatic image captioning.