WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Transforming Visual Understanding: A Comprehensive Survey of Image Captioning with Transformers

Singampalli Suma Latha[1], Bejjam Praveen[2], Dindi Shriya[3], Poddu Abhishek[4], Vannelaganti Sai Prabhu[5], R. Cristin[6]

[1]Department of Computer Science & Engineering, GMR Institute of Technology, Rajam, Vizianagaram, 20341A05H6@gmrit.edu.in
[2]Department of Computer Science & Engineering, GMR Institute of Technology, Rajam, Vizianagaram, 21345A0515@gmrit.edu.in
[3]Department of Computer Science & Engineering, GMR Institute of Technology, Rajam, Vizianagaram, 20341A05H3@gmrit.edu.in
[4]Department of Computer Science & Engineering, GMR Institute of Technology, Rajam, Vizianagaram, 20341A05E7@gmrit.edu.in
[5]Department of Computer Science & Engineering, GMR Institute of Technology, Rajam, Vizianagaram, 20341A05I4@gmrit.edu.in
[6]Department of Computer Science & Engineering, GMR Institute of Technology, Rajam, Vizianagaram, cristin.r@gmrit.edu.in

ABSTRACT

Image captioning plays a crucial role in automating the annotation and labeling of images, providing detailed descriptions of the objects, scenes, or activities depicted. This technology finds wide applications in image recognition, recommendation systems, and content classification, facilitating large-scale image analysis. Unlike traditional rule-based systems, which are time-consuming and struggle with capturing nuanced contextual details in complex images, the proposed method integrates two cutting-edge models: attention mechanism and Transformer. The attention mechanism stands out as a robust model for multimodal understanding, seamlessly merging the realms of vision and language processing. It harnesses insights from extensive datasets to grasp semantic relationships between images and text. By combining a multimodal understanding of attention mechanisms with the sequence modeling capabilities of Transformers, the resulting synergy ensures superior performance in image captioning tasks. Including additional Transformer layers further improves the model's ability to understand complex contextual dependencies contained in input headers. To evaluate the effectiveness of the proposed system, objective evaluation indicators including BLEU scores (1, 2, 3, 4) were used. These metrics provide a comprehensive assessment of image captioning capabilities by providing a quantitative assessment of the system's ability to generate captions that match human-perceived quality.

Keywords: Image description, Attention mechansim, Transformer, multimodal understanding, BLEU

I. INTRODUCTION

In recent years, there has been a transformative impact on natural language processing (NLP) and computer vision (CV) fields, thanks to the advent of multimodal deep learning approaches. Among these, the attention mechanism has emerged as a pivotal breakthrough, seamlessly integrating information from both images and textual context. This innovation enables a holistic comprehension of both vision and language, marking a significant leap forward. Another key player in this evolution is the Transformer architecture, renowned for its prowess in NLP tasks, leveraging attention-based mechanisms for effective sequence-to-sequence modeling. Capitalizing on the success of these advancements, we introduce a pioneering project aimed at elevating contextual image captioning. Our approach synergistically leverages the advantages of attention mechanism and converter architecture. Image captioning, which creates human text descriptions of images, requires a sophisticated understanding of visual content and related contextual information. Existing methods often fail by relying solely on visual features extracted from pre-trained convolutional neural networks (CNNs), missing important linguistic context that could potentially significantly improve the quality and relevance of generated captions.

II. LITERATURE SURVEY:

This research introduces a text-only model designed for visual question answering (VQA) tasks, specifically addressing the need for external knowledge. The experimentation utilizes the OK-VQA dataset, known for its knowledge-centric nature, necessitating information from unspecified external resources. The model employs a BERT-base transformer encoder, OSCAR (Object-Semantics Aligned Pre-training), to articulate image contents. Leveraging the implicit knowledge of the T5 language model, this approach surpasses the performance of pre-

trained multimodal models with a comparable number of parameters, achieving state-of-the-art results, particularly as the language model size increases. The method demonstrates superiority over pre-trained multimodal models when it comes to handling knowledge-intensive tasks. In particular, automatic captions that fail to convey relevant information in the image pose a challenge to the performance of CBM models. However, it should be recognized

that the proposed approach is less efficient in standard VQA tasks (VQA 2.0) compared to multimodal converters specifically designed to process both textual and visual information [1].

This paper introduces a novel model for affective image captioning in the context of visual artworks, leveraging emotion-based cross-attention mechanisms. The proposed model incorporates an affective visual encoder that seamlessly integrates emotion attributes and cross-modal joint features of images across all encoder blocks. Additionally, the inclusion of affective tokens fuses grid- and region-based image features, ensuring coverage of both contextual and object-level information. The evaluation of the model was conducted on the ArtEmis dataset, revealing its superior performance compared to baseline methods across all metrics in the emotion-conditioned task. Key methodologies employed in the paper encompass visual encoding, text generation, and specific training strategies. The algorithm at the core of this study, Affective Visual Encoder, has a high ability to integrate emotional attributes and cross-modal joint features and provides a comprehensive approach to emotional image captioning. Performance metrics including BLEU, ROUGE, METEOR, and CIDEr were used to evaluate the performance of the proposed model. However, it is worth pointing out that this paper has potential limitations in scope as it only focuses on emotion-driven tasks and does not allow users to specify arbitrary emotions [2].

This paper introduces a novel evaluation metric, ViLBERT Score, designed for assessing image captioning systems. Unlike existing metrics, ViLBERT Score utilizes both image and text information to generate image-conditioned embeddings for each token, employing ViLBERT for both generated and reference texts. By comparing contextual embeddings from each sentence-pair, the metric computes a similarity score that significantly outperforms all existing metrics in correlation with human judgments. The paper provides a comprehensive overview of widely used metrics for evaluating image captions, including n-gram similarity metrics, embedding-based metrics, and other task-specific measures. Notably, the limitations of n-gram similarity metrics are highlighted, particularly their inability to account for synonym matches in generated text. To solve this problem, this paper proposes embedding-based metrics such as word movement distance (WMD) and BERT Score. Empirical results on the Composite, Flickr8k, and PASCAL-50S datasets show that ViLBERT Score shows superior correlation with human judgment compared to previous measures [3].

This paper presents a new image captioning model that integrates a vision-enhanced encoder and a knowledge-based control module. This model aims to deeply explore internal connections within images and integrate visual observations with external knowledge. The evaluation of the proposed Vision Enhanced Control Module (VCT) model is performed on two well-known captioning datasets: MSCOCO and Flickr30K. Model performance is evaluated using popular metrics such as BLEU, METEOR, ROUGE, CIDEr, and SPICE. In particular, the proposed model outperforms state-of-the-art methods on the MSCOCO and Flickr30K datasets, achieving better results in all evaluation metrics. However, it is noteworthy that the performance of the proposed model relies heavily on external knowledge and consensus memory, and this dependency can cause problems in scenarios where that information is not available or applicable [4].

The paper introduces a novel image captioning method called Caption TLSTMs, leveraging a combination of transformer blocks and two LSTM modules to extract a more comprehensive multimodal intersection feature representation. This approach represents an enhancement over the ASG2Caption model, which utilizes abstract scene graphs for generating diverse and customized image descriptions. The evaluation of the proposed method is conducted on Visual Genome and MSCOCO datasets, demonstrating notable improvements in the overall quality of generated image captions. However, it's highlighted that the method's performance might be less optimal when applied to datasets with different characteristics than those used in the evaluation. Moreover, concerns are raised about scalability to larger datasets or more complex images, suggesting potential limitations in handling increased data volume or intricate visual content. The effectiveness of this method also depends on the availability of a significant amount of training data and sufficient computing resources, which represents a potentially resource-intensive requirement for optimal performance [5].

In this paper, we present an image captioning algorithm using a hybrid deep learning technique (CNN GRU). The technology includes three neural network modules: encoder, decoder, and semantic validator. The encoder extracts features from the input image and generates an attribute vector representing the possible high-level attributes of the caption dataset. The decoder uses the feature and attribute vectors to generate labels. It is important to note that a semantic validator reconstructs a representation that is semantically similar to the input image. During training, we use the Semantic Reconstructor reconstruction score combined with the likelihood to evaluate the quality of the generated signatures. The performance of the proposed model is compared with state-of-the-art methods using metrics such as BLEU@K, METEOR, ROUGE-L, and CIDER-D. The results show that the proposed model outperforms the LSTM-A5 model in terms of time complexity and image caption accuracy. However, they note that models can be limited when working with complex images containing large amounts of information and numerous objects [6].

The paper introduces a Text-Guided Generation and Refinement (TGGAR) model for image captioning, designed to enhance caption quality using guide text. This model adopts an encoder-decoder architecture, featuring a Text-Guided Relation Encoder (TGRE) along with two submodules: a Generator for primary sentence generation and a Refiner for sentence refinement. Evaluation of the proposed TGGAR model is performed using two performance metrics: B@4 score and CIDEr score. The experiments are conducted on the MS COCO captioning dataset, comprising 123,287 images with five human-annotated descriptions for each image. The results suggest the model's efficacy in improving caption quality based on the specified metrics. The paper also recommends testing the TGGAR model on other datasets to evaluate its performance and generalizability beyond the MS COCO dataset [7].

A unique method for fine-grained image-to-language synthesis is presented in this paper: the Context-Aware Visual Policy network (CAVP). It is designed to especially target picture paragraph and sentence captioning. The CAVP is compared with conventional image captioning techniques like Google NIC, Hard Attention, Adaptive Attention, and LSTM-A, as well as reinforcement learning-based techniques like PG-SPIDEr-TAG, SCST, Embedding-Reward, and Actor-Critic, in order to evaluate its efficacy. BLEU, ROUGE, METEOR, CIDEr, and SPIDEr are just a few of the performance metrics used in the evaluation. Future research directions proposed in the paper include investigating more complex visual features, increasing the suggested model's

applicability to other image captioning tasks, incorporating outside knowledge sources, and creating training algorithms that are more effective. This suggests a prospective approach to expanding the functionalities and adaptability of the Context-Aware Visual Policy network [8].

The research aims to enhance Unpaired Image Captioning (UIC) performance through the use of a prompt-based learning strategy that makes use of semantic/metric prompts and Vision-Language Pre-Trained Models (VL-PTMs).presents a Prompt-based Learning method using semantic and metric prompts for Unpaired Image Captioning (PL-UIC). To improve caption prediction accuracy, it makes use of Vision-Language Pre-Trained Models (VL-PTMs) for adversarial learning, cross-domain cue inference, and iterative refinement of pseudo image-caption samples. assesses the suggested PL-UIC model using common evaluation metrics including BLEU, METEOR, ROUGE-L, and CIDEr on the COCO and Flickr30K datasets. The quantity and quality of the training data, as well as the efficiency of the VL-PTMs employed for cross-domain cue inference, have a significant impact on the model's performance. [9]

The approach it proposes uses a combination of language models and diffusion models to handle the problem of creating images with different contexts for text-only image captioning. It selects several statements that describe a scenario using a language model, and then condenses them into a single sentence having several contexts. Diffusion models are also used to create both basic and sophisticated visuals. METEOR, ROUGE-L, CIDEr, BLEU-1, BLEU-2, BLEU-3, and BLEU-4 are some of the evaluation metrics that are employed. The incapacity of LLMs to choose captions from the full corpus and the domain gap between artificial and natural photos are two of the constraints of the suggested method. To achieve state-of-the-art performance on widely-used datasets, the MCDG framework combines diffusion models and LLMs to provide efficient multi-context training data for text-only picture captioning. The use of synthetic data improves image captioning tasks significantly.[10].

III. METHODOLOGY

1. Inception V3 Model:

Convolutional neural networks (CNNs) like Google's InceptionV3 are state-of-the-art in image processing and are specifically made for picture categorization and, most importantly, image captioning. The utilization of Inception modules, precisely designed building blocks that make use of different filter sizes (1x1, 3x3, 5x5) and pooling operations to collect characteristics at different scales, is what gives it its architectural novelty. InceptionV3 dives deeply into visual representations with an amazing 48 layers, making it possible to extract minute patterns and information. One unique aspect of InceptionV3 is that completely linked layers are skipped in favour of global average pooling at the network's end. By reducing the number of parameters, this design decision improves computing performance and mitigates overfitting.

Furthermore, with massive datasets such as ImageNet, pre-trained weights are frequently advantageous for InceptionV3. The model gains a thorough understanding of generic features from this pre-training, which may be tailored for particular tasks like captioning images. As an image encoder, InceptionV3 is essential to the captioning of images. In order to produce rich representations that are fed into sequence generation models like transformers or recurrent neural networks to produce informative captions, it extracts high-level visual information from input photos. This two-step procedure guarantees that the output captions are detailed and contextually relevant, with InceptionV3 managing feature extraction.

The strength of InceptionV3 is in its deep architectural complexity and effective feature extraction capabilities. InceptionV3, a mainstay in the field of computer vision, is widely used due to its capacity to comprehend and interpret complex visual information. It provides a reliable solution for a range of image-related tasks, most notably picture captioning.

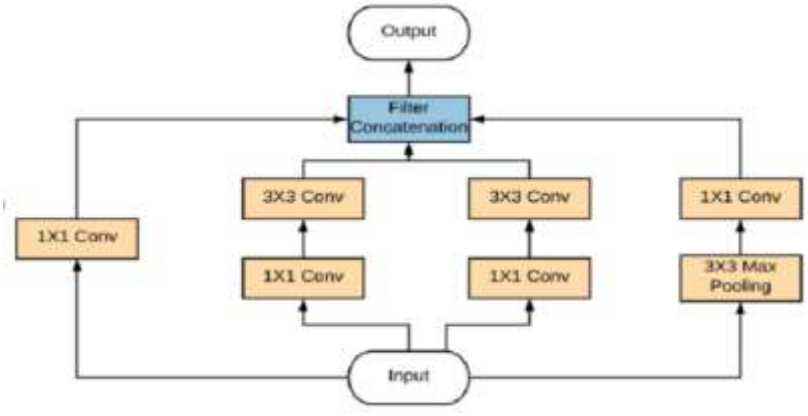

Fig.1. Inception V3 Model

2. Transformer Model:

Originally intended for use in natural language processing, the Transformer model has proven to be effective in handling complex linkages and long-range dependencies, leading to its use in image captioning. A pre-trained convolutional neural network (CNN), such as InceptionV3 or ResNet, is used by an encoder in the adapted architecture for image captioning to extract visual features. A decoder with Transformer layers then generates captions, allowing attention mechanisms to focus on pertinent portions of the visual features.

Transformers' self-attention mechanism is essential for allowing the model to assess the relative importance of various input sequence segments, which is necessary for identifying contextual relationships in image data. Positional encoding is used to address the sequential order issue that Transformers provide, maintaining spatial links in the visual elements necessary for comprehending an image. The model integrates visual elements and positional encodings in the decoder to automatically handle multimodal inputs and produce captions that capture spatial relationships in addition to describing image content.

Using datasets containing photos and their related captions, the model is trained end-to-end. Transformer models have the capacity to capture long-range correlations between visual data and words in captions, as well as effective parallelization for speedier training. Nevertheless, there are difficulties, including as computational complexity (particularly for larger models) and the requirement for a large amount of labelled data for efficient training. To sum up, the Transformer model that has been modified for the purpose of captioning images uses its parallelization and self-attention mechanisms to fully capture complex relationships seen in visual data, producing accurate and contextually rich image captions.

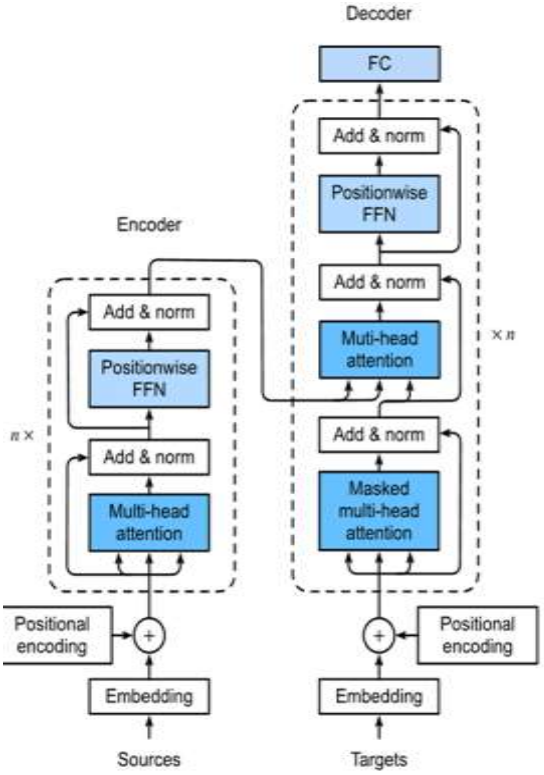

Fig2. Transformer Model Architecture

IV. RESULTS

The evaluation of the model's performance in image captioning is conducted through the use of BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores, as presented in Table 1. These scores serve as quantitative metrics to assess the quality and accuracy of the generated captions. A higher BLEU score indicates a closer alignment between the model's outputs and the reference captions. Additionally, insights into the image captioning results are provided in Figure 4, offering a visual representation of the model's effectiveness. Analysing Figure 4 allows for a qualitative understanding of how well the generated captions align with the actual content of the images. Overall, the combination of quantitative BLEU scores and visual insights from Figure 4 provides a comprehensive evaluation of the transformer model's proficiency in image captioning, shedding light on both the quantitative accuracy and the perceptual quality of the generated captions.

Evaluation	Scores
BLEU-1	33.333
BLEU-2	57.735
BLEU-3	71.922
BLEU-4	75.983

Table1. Comparative Analysis: BLEU Scores

Fig3. Comparison between the real & predicted captions using BLEU

V. DISCUSSION

In summary, the transformer-based image captioning model shows promising results in image caption generation. The BLEU score indicates the quality of the subtitles produced, and more testing and fine-tuning can yield better results. When evaluating the performance of such models, it is important to consider both quantitative indicators and qualitative aspects.

VI. CONCLUSION

The methods and code for an image captioning model that uses the Transformer architecture are described in this work. The BLEU score, which denotes the calibre of the generated subtitles, is used to assess the model's capacity to produce subtitles for images. The provided code serves as a starting point for further research and experimentation in image captioning tasks. For more information and code explanations, see the provided code comments and the comments in the individual codes section.

VII. REFERENCES

[1] Cao, S., An, G., Zheng, Z., & Wang, Z. (2022). Vision-Enhanced and Consensus-Aware Transformer for Image Captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10), 7005-7018.

[2] Ishikawa, S., & Sugiura, K. (2023). Affective Image Captioning for Visual Artworks Using Emotion-Based Cross-Attention Mechanisms. *IEEE Access*, *11*, 24527-24534.

[3] Lee, H., Yoon, S., Dernoncourt, F., Kim, D. S., Bui, T., & Jung, K. (2020, November). Vilbertscore: Evaluating image caption using vision-and-language bert. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems* (pp. 34-39).

[4] Yan, J., Xie, Y., Luan, X., Guo, Y., Gong, Q., & Feng, S. (2022). Caption TLSTMs: combining transformer with LSTMs for image captioning. *International Journal of Multimedia Information Retrieval*, *11*(2), 111-121.

[5] Ahmad, R. A., Azhar, M., & Sattar, H. (2022, December). An Image captioning algorithm based on the Hybrid Deep Learning Technique (CNN+ GRU). In *2022 International Conference on Frontiers of Information Technology (FIT)* (pp. 124-129). IEEE..

[6] Wang, D., Hu, Z., Zhou, Y., Hong, R., & Wang, M. (2022). A text-guided generation and refinement model for image captioning. *IEEE Transactions on Multimedia.*

[7]Zha, Z. J., Liu, D., Zhang, H., Zhang, Y., & Wu, F. (2019). Context-aware visual policy network for fine-grained image captioning. IEEE transactions on pattern analysis and machine intelligence, 44(2), 710-722.

[8]Zhu, P., Wang, X., Zhu, L., Sun, Z., Zheng, W. S., Wang, Y., & Chen, C. (2023). Prompt-based learning for unpaired image captioning. IEEE Transactions on Multimedia.

[9] Lian, Z., Zhang, Y., Li, H., Wang, R., & Hu, X. (2023). Cross modification attention-based deliberation model for image captioning. Applied Intelligence, 53(5), 5910-5933.

[10] Ma, F., Zhou, Y., Rao, F., Zhang, Y., & Sun, X. (2023). Text-Only Image Captioning with Multi-Context Data Generation. arXiv preprint arXiv:2305.18072