iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** XII **Month of publication:** December 2023

DOI: https://doi.org/10.22214/ijraset.2023.57766

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue XII Dec 2023- Available at www.ijraset.com

Visual Image Captioning through Transformer

Muneeb Nabi[1], Rohit Pachauri[2], Shouaib Ahmad[3], Kanishk Varshney[4], Prachi Goel[5], Apurva Jain[6]
[1, 2, 3, 4]*CSE Department, Dr. Akhilesh Das Gupta Institute of Professional Studies, New Delhi, India*
[5, 6]*Asst. Prof. of CSE Department, Dr. Akhilesh Das Gupta Institute of Professional Studies, New Delhi, India*

Abstract: ***The convergence of computer vision and natural language processing in Artificial Intelligence has sparked significant interest in recent years, largely propelled by the advancements in deep learning. One notable application born from this synergy is the automatic description of images in English. Image captioning involves the computer's ability to interpret visual information from an image and translate it into one or more descriptive phrases. Generating meaningful descriptions requires understanding the state, properties, and relationships between the depicted objects, demanding a grasp of high-level picture semantics. Automatically captioning images is a complex task that intertwines image analysis with text generation. Central to this process is the concept of attention, determining what to describe and in what sequence. While transformer architectures have shown success in text analysis and translation, adapting them for image captioning presents unique challenges due to structural differences between semantic units in images (usually identified regions from object detection models) and sentences (composed of individual words). Little effort has been devoted to tailoring transformer architectures to suit images' structural characteristics. In this study, we introduce the Image Transformer, a novel architecture comprising a modified encoding transformer and an implicit decoding transformer. Our approach involves expanding the inner architecture of the original transformer layer to better accommodate the structural nuances of images. By utilizing only region features as inputs, our model achieves state-of-the-art performance on the MSCOCO Dataset. This research employing CNN-Transformer architectural models for image captioning aims to detect objects within images and convey information through textual messages. The envisioned application of this method extends to aiding individuals with visual impairments, using text-to-speech messages to facilitate their access to information and nurture their cognitive abilities. This paper meticulously explores fundamental concepts in image captioning and its standardized procedures, introducing a generative CNN-Transformer model as a significant advancement in this field.***

Keywords: MSCOCO, Transformer, CNN, Attention Mechanism,InceptionV3

I. INTRODUCTION

The combination of computer vision and natural language processing has led to significant strides in comprehending visual content, with image captioning emerging as a pivotal application. This fusion harnesses the interpretative prowess of both disciplines, enabling machines to generate descriptive narratives for visual content. This interdisciplinary effort aims not only to perceive intricate image details but also to articulate their essence in human-like language—a complex task demanding a deep understanding of visual semantics and linguistic structure. Image captioning, which automatically generates English descriptions based on image content, plays a crucial role in interpreting scenes by amalgamating computer vision data with language processing. Future advancements in image retrieval, particularly through contextual cues like image captions, promise solutions to this challenge. In our daily lives inundated with visual stimuli on social media and in the news, humans possess the innate ability to recognize photographs without captions, a skill yet to be fully replicated in machines. For machines to comprehend images, they must be taught. The encoder-decoder architecture of Image Caption Generator models utilizes input vectors to produce coherent and relevant captions, bridging the realms of natural language processing and computer vision. Recognizing and contextualizing images precedes the generation of descriptive language, such as English, elucidating the contents captured within the visual frame.

II. RELATED WORK

We categorize existing attention-based image captioning models into distinct categories: single-stage attention models, two-stage attention models and transformer-based models.

A. Single-Stage Attention Based Image Captioning

Single-stage attention-based image captioning models implement attention during the decoding phase, allowing the decoder to focus on the most relevant region [15] within the image when generating each corresponding word.

©IJRASET: All Rights are Reserved

The availability of extensive annotated datasets [16,17] facilitated the training of deep learning models for image captioning. Vinyals et al. [18] introduced the initial deep model for this task, employing a pre-trained CNN from ImageNet [16] to encode the image. Subsequently, they utilized an LSTM-based language model [19] for decoding the image features into a word sequence. Xu et al. [1] innovatively incorporated an attention mechanism during word generation in their image captioning model. Their attention module produced a matrix that weighted each receptive field in the encoded feature map, incorporating this weighted feature map and the previous word to generate subsequent words. Differing from merely focusing on receptive fields, their approach involved re-weighting each feature channel for word generation. Considering not all words in a sentence directly correlate with visual elements in the image, Lu et al. [20] proposed an adaptive attention strategy. Their model integrated a visual sentinel, enabling adaptive decisions on when and where to rely on visual information. While the single-stage attention model demonstrates computational efficiency, it may lack precision in pinpointing informative regions within the original image.

B. Two-Stages Attention Based Image Captioning

Two-stage attention models encompass both bottom-up and top-down attention mechanisms. The bottom-up attention initially employs object detection models to identify multiple informative regions within the image. Subsequently, the top-down attention mechanism focuses on these detected regions while generating each word. In contrast to single-stage attention models relying on coarse receptive fields as informative regions, Anderson et al. [3] conducted training of detection models using the Visual Genome dataset [21]. These trained models can detect a range of 10 to 100 informative regions within an image. Their approach involves a two-layer LSTM network for decoding. The first layer generates a state vector based on the embedded word vector and the mean feature extracted from the detected regions. The second layer utilizes the previous layer's state vector to compute weights for each detected region. These weighted features are aggregated to form a context vector for predicting the subsequent word. Similarly, Lu et al. [4] devised a comparable network, employing a detection model trained on the MSCOCO dataset [22], which, being smaller than the Visual Genome, yields fewer informative regions. The performance of two-stage attention-based image captioning models surpasses that of single-stage attention models. However, each detected region remains isolated from others, lacking direct relationships with other regions.

C. Transformer Based Image Captioning

Transformer-based image captioning models employ the dot-product attention mechanism to implicitly relate informative regions. Following the inception of the original transformer model [11], subsequent advancements focused on tailored architectures for machine translation, considering sentence structure and natural characteristics [19,22]. In the domain of image captioning, AoANet [12] adopts the original internal transformer layer architecture, enhancing it with a gated linear layer [20] atop the multi-head attention. The object relation network [14] introduces relative spatial attention into the dot-product attention mechanism. Notably, Herdade et al. [14] discovered that the simple position encoding, akin to the original transformer, did not notably enhance image captioning performance. The entangled transformer model [13] integrates a dual parallel transformer to encode and refine both visual and semantic information, fused via a gated bilateral controller. Unlike scene graph-based models requiring auxiliary models to detect and construct the scene graph initially, transformer-based models prove more computationally efficient in this regard. However, current transformer-based models adhere to the inner architecture of the original transformer, designed primarily for text. Each transformer layer contains a single multi-head dot-product attention refining module, limiting the model's ability to fully capture complex relations between image regions. Therefore, our proposition involves modifying the inner architecture of the transformer layer to better suit image data. We expand the transformer layer, enabling multiple refining modules within each layer to address diverse aspects of image regions across both encoding and decoding stages.

III. PROCEDURE &DESIGN

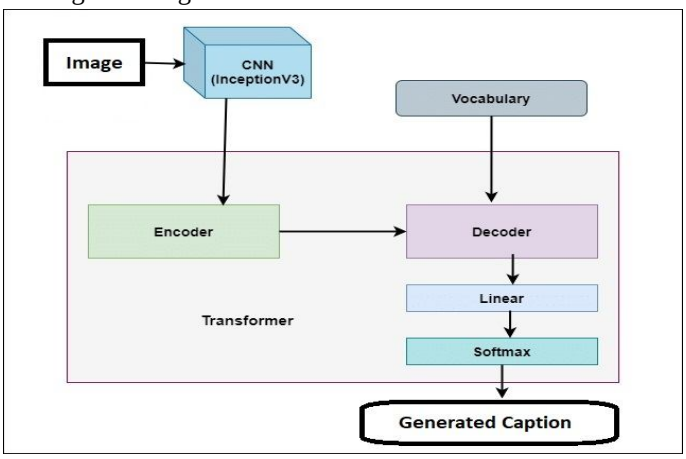
Convolutional Neural Networks (CNNs) play a crucial role in processing data structured in specific formats, such as 2D grids, making them particularly adept for image-related tasks. These networks analyze images systematically, scanning from left to right and corner to corner, extracting key features and amalgamating them to interpret the complete image. Their versatility extends to interpreting rotated, scaled, or modified images, allowing for robust image analysis. Operating as a deep learning algorithm, CNNs process input images, assigning significance to distinct elements within them, thereby distinguishing one image from another.

For this research, Inception-v3, a member of the Inception family, is the chosen Convolutional Neural Network architecture. It incorporates several enhancements, including Label Smoothing, Factorized 7 x 7 convolutions, and the integration of an auxiliary classifier to convey label information throughout the network, along with the application of batch normalization to layers in the sidehead.

The architecture of a transformer consists of stacked multi-head dot-product attention-based refining layers. Each layer processes an input A in R^N x D, comprising N entries of D dimensions. In natural language processing, these entries might represent embedded word features within a sentence. In computer vision, specifically image captioning, these entries describe features of regions within an image. The core function of the transformer is refining each entry by attending to others using multi-head dot-product attention. Residual connections are embedded within each module of the transformer layer, with the final output achieved by adding the transformer layer's output to its input. Furthermore, a feed-forward network introduces non-linearity for additional processing. Each refining layer uses the output of its preceding layer as input, with the initial layer taking the original input.

The decoding phase is also composed of a stack of transformer refining layers, utilizing the output from the encoding phase as well as embedded features from the previously predicted word.

We aim to outline two distinct architectures that collectively form an automated image caption generation model known as the CNN-Transformer model. In this framework, both architectures work collaboratively to generate captions for input photographs. The CNN architecture is employed to extract key features from the input image. For this purpose, we have utilized InceptionV3, a pre-trained model, to effectively extract essential image features. Meanwhile, the Transformer plays a pivotal role in storing, analyzing, and leveraging the data or features extracted by the CNN model. Its primary function is to contribute to the generation of accurate and descriptive captions for the given image.



Working of the Model

1) An image is processed through CNN to identify the objects.
2) CNN scans images left-right, and top-bottom and extracts important image features. By applying Fully Connected layer , and thus using the activation function, we successfully extracted features of every image.
3) Feature Vectors from the image then goes to the encoder layer of the transformer which then encode them and passes on to the next part of the transformer that is decoder.
4) Decoder of the transformer uses the encoded feature vectors to genenrate the possible caption for the given image.

The model underwent training on the MSCOCO image captioning dataset, employing Karpathy's splits. This dataset comprises 330,000 images segregated into training and validation sets. Evaluation metrics, including Bleu, were utilized to gauge accuracy. The proposed model underwent training for 50 epochs, employing cross-entropy loss optimization. The training process utilized a batch size of 32 to optimize the model's performance.

IV. RESULTS AND DISCUSSION

We conducted a comparative analysis of our model's performance against various published image captioning models. These models encompass a range of architectures, including the top-performing single-stage attention model, Att2all [22], as well as two-stage attention-based models like n-babytalk [4] and up-down [23].

Our evaluation also includes models such as GCN-LSTM [23], AUTO-ENC [21], ALV [24], and transformer-based models like Entangle-T [13], AoA, and others in the transformer-based category.

Table 1- Accuracy

Models	BLEU1(c5)	BLEU4(c5)
Single Attention Based		
GCN-LSTM	80.8	38.7
ALV	79.9	37.4
AUTO-ENC	-	38.5
Transformer Based		
Entangle-T	81.2	38.9
AoA	81	39.4
Ours	80	39.5

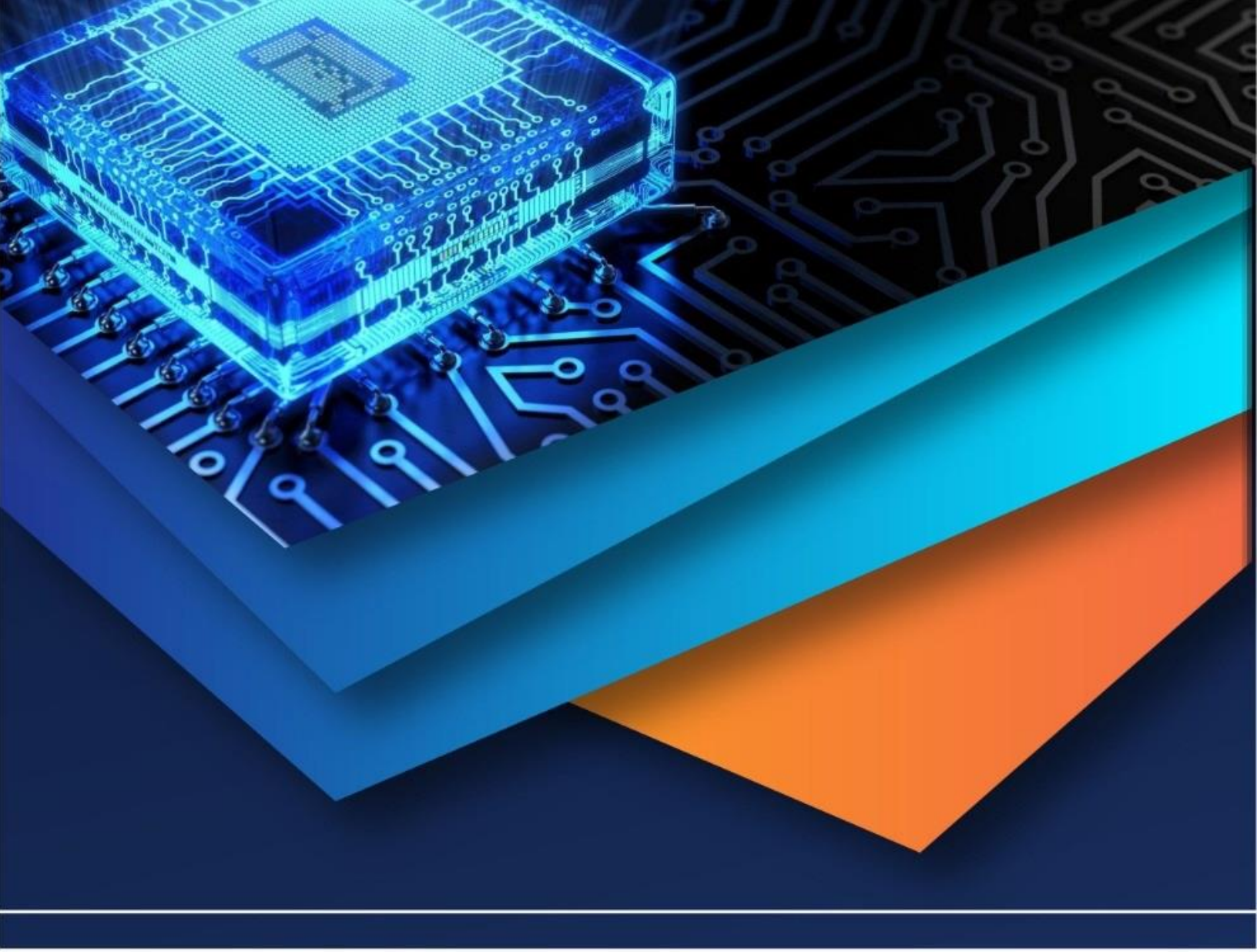
BLEU(c5) is used to measure the accuracy for the result of the developed model.
BLEU score turnsout to be excellent comparing to the other model that are being worked on.
The research relied on the "COCO Dataset," a collection featuring 12 main categories and 80 potential sub-categories within each. This dataset comprised images grouped into various categories, with five captions provided for each image. To assess system performance, a general confusion matrix was employed. Results and projections for multiple models were documented, encompassing a total of 130 iterations across a 30-iteration span.

V. CONCLUSION AND FUTURE SCOPE

We presented the image transformer architecture, which extends the original transformer layer from machine translation to suit the characteristics of images. Our aim is to inspire the development of more sophisticated transformer-based architectures. These can not only improve image captioning but also cater to other computer vision tasks requiring intricate relational attention. This adaptable concept holds promise across a wide array of applications. By merging insights from CNN models and Transformers, we've overcome previous constraints in graphical image captioning. Our CNN-Transformer model efficiently scans input images, extracts crucial details, and translates them into concise, natural-language English sentences, effectively bridging the gap between visual input and linguistic output.

REFERENCES

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object caption- ing at scale. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 8948–8957, 2019.
[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. arXiv preprint arXiv:1612.00576, 2016.
[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In European conference on computer vision, pages 382–398. Springer, 2016. 6
[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6077–6086, 2018. 1, 3, 4, 5
[5] David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. Paint by word. arXiv preprint arXiv:2103.10951, 2021. 2
[6] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5659–5667, 2017. 3
[7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C Lawrence Zitnick. ´ Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325, 2015. 2, 3, 4



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)