JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Image Caption Generator UsingTransformers

[1] Mr.P.Muthyalu, [2] K.Charan Reddy,

[1]Professor, [2] Student
[1]Dept.of computer science, Narayana Engineering College, Gudur, India
[2] Dept.of computer science, Narayana Engineering College, Gudur, India

Abstract : Image Captioning is the task of translating an input image into a textual description. As such, it connects Vision and Language in a generative fashion, with applications that range from multi-modal search engines to help visually impaired people. Although recent years have witnessed an increase in accuracy in such models, this has also brought increasing complexity and the challenges in interpretability and visualization. In this work, we focus on the Transformer-based image captioning models and provide qualitative and quantitative tools to increase interpretability and assess the grounding and temporal alignment capabilities of such models.
Firstly, we employ attribution methods to visualize what the model concentrates on in the input image, at each step ofthe generation. Further, we propose metrics evaluate the temporal alignment between model predictions and attribution scores, which allows measuring the grounding capabilities of the model. Experiments are conducted on three different Transformer- based architectures, employing both traditional and Vision Transformer-based visual features.

I. INTRODUCTION

In recent years, the intersection of computer vision and natural language processing has witnessed remarkable advancements, enabling machines to comprehend visual content and describe it in human-like language. One of the significant breakthroughs in this domain is the development of image caption generators using transformer-based models. Transformative architectures like BERT, GPT, and T5 have revolutionized language understanding and generation tasks, and when adapted to image captioning, they offer unprecedented capabilities in generating rich and contextually relevant descriptions for visual content.

Traditionally, image captioning relied heavily on handcrafted features and sequential models, which often struggled to capture intricate relationships between visual elements and the corresponding textual descriptions. However, transformer-based models, with their attention mechanisms and hierarchical representations, excel in capturing long-range dependencies and contextual nuances, making them ideal candidates for the task of image captioning.

We consider different encoder-decoder architectures for image captioning and evaluate their interpretability by developing solutions for visualizing what the model concentrates on the input image at each step of the generation.

We propose an alignment and grounding metric which evaluates the temporal alignment between model predictions and attribution scores, thus identifying defects in the grounding capabilities of the model.

We conduct extensive experiments on the COCO dataset and the ACVR Robotic Vision Challenge dataset, considering different model architectures, visual encoding strategies, and attribution methods.

II. RELATED WORK*:*

Before the advent of deep learning, traditional image captioning approaches were based on the generation of simple template sentences, which were later filled by the output of an object detector or an attribute predictor. With the surge of deep neural networks, captioning has started to employ RNNs as language models and the output of one or more layers of a CNN was employed to encode visual information and to condition the generation of language. On the training side, initial methods were based on a time-wise cross-entropy training. A notable achievement has then been made with the introduction of the REINFORCE algorithm, which enabled the use of non-differentiable caption metrics as optimization objectives. On the image encoding side, instead, additive attention mechanisms have been adopted to incorporate spatial knowledge, initially from a grid of CNN features, and then using image regions extracted with an object detector. To further improve the encoding of objects and their relationships, graph convolution neural networks have been employed as well, to integrate semantic andspatial relationships between objects or to encode scene graphs.

Explainability and visualization in vision-and-language :

While the performance of captioning algorithms has been increasing in the last few years, and while these models are approaching the level of quality required to be run in production, providing effective visualizations of what the model is doing, and explanations to why it fails, is still under-investigated. It shall be noted, in this regard, that early captioning models based on additive attention were easy to be visualized – as their attentive distribution was a single-layer weighted summation of visual features. In the case of modern captioning models, instead, each head of each encoder/decoder layer takes an attentive distribution, thus making visualization less intuitive and straightforward. A solution that is becoming quite popular is that of employing an attribution method, which allows attributing predictions to visual features even in presence of significant non-linearities.

III. METHODS AND MATERIAL

The use of Image caption generator involves several steps of image processing to confirm and document individuals' captions. Here's a breakdown of the steps outlined:

3.1 Feature Extraction:

- Image Upload
- Extraction
- Model Architecture
- Sentence generation

3.1.1 : Image Upload:

This initial step involves gathering physical or communication tests in predefined situations and over a specified period of time. It encompasses image upload as input .

3.1.2 : Extraction:

After uploading images , the next step is to extract relevant data from these images to create templates for caption generation. This process involves analyzing and encoding image features such as the objects and actions, and other distinctive characteristics.

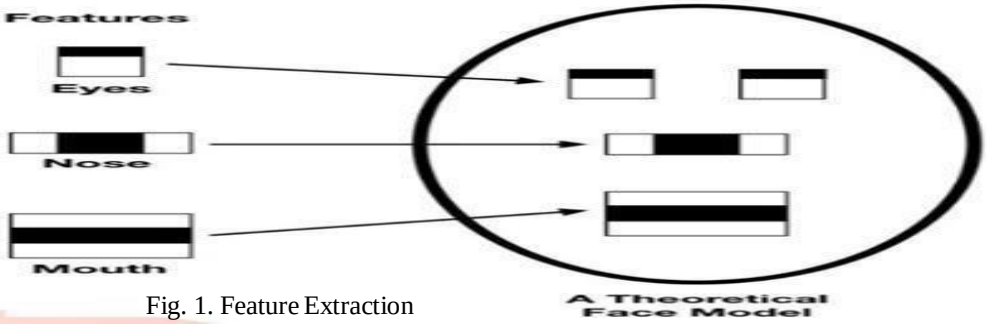

Fig. 1. Feature Extraction

3.1.3 Model Architecture :

Design the architecture to integrate the transformer model with a pre-trained CNN for feature extraction. This can involve using a fusion mechanism to combine visual features with the textual context provided by the transformer. Implement attention within the transformer to focus on relevant parts of the image when generating captions.

3.1.4 : Sentence Generator :

Given a new image, extract its features using the pre-trained CNN and pass them through the fine-tuned transformer model. Generate captions by employing techniques like beam search or sampling to decode the output probabilities intosentences.

3.2 Methodology :

The face recognition module is developed using OpenCV and TensorFlow. It involves the following steps:

- Data Collection and Preprocessing
- Feature Extraction
- Pre-trained Transformer Model Selection
- Fine-tuning
- Model Architecture Design
- Training
- Caption Generation

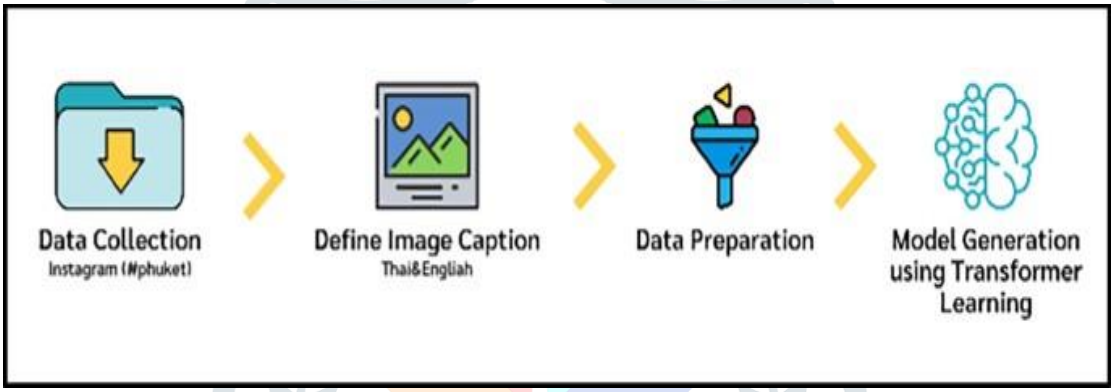

Fig. 2. Methodology

3.3 :Methodology Description:

3.3.1 : Data Collection and PreProcessing:

Preprocess the images by resizing them to a uniform size and normalizing pixel values. Tokenize the captions into words or subwords for input to the transformer model.

3.3.2 : Feature Extraction:

- Extract Image Features: During inference, pass the image through the encoder to get its feature representation.
- Caption Generation: Use the transformer decoder to generate captions. Start with a start token and iteratively predict the next word until an end token is generated or the maximum caption length is reached.
- Decoding Strategies: Implement strategies like greedy decoding or beam search to improve the quality of the generated captions. Greedy decoding selects the most probable word at each step, while beam search explores multiple paths to find the best sequence.

3.3.3 : Pre-trained Transformers Model Selection:

Choose a suitable pre-trained transformer model such as BERT, GPT, T5, or similar architectures. Consider factors like model size, computational resources, and task-specific performance.

3.3.4 Fine-Tuning :

Fine-tune the selected transformer model on the image-caption dataset. This involves feeding the image features (extracted using a pre-trained CNN) along with the caption tokens into the transformer model and optimizing it for the caption generation task.

3.3.5 : Model Architecture Design :

Design the architecture to integrate the transformer model with a pre-trained CNN for feature extraction. This can involve using a fusion mechanism to combine visual features with the textual context provided by the transformer. Implement attention mechanisms within the transformer to focus on relevant parts of the image when generating captions.

3.3.6 : Training :

Train the combined model on the image-caption dataset using supervised learning techniques. Use techniques like teacher forcing or scheduled sampling during training to improve caption generation performance. Matching the features extracted from the new image with those in the database.

- Similarity Scoring: Using algorithms to score the similarity between the captured features and stored templates. If the score exceeds a predefined threshold, the system identifies a match.

3.3.7 : Caption Generator :

During inference, input the image features into the trained transformer model. Utilize beam search or other decoding strategies to generate captions probabilistically. Generate captions word by word, conditioning each word's generation on the previously generated words and the image features.

IV. ADVANTAGES

Image caption generators using transformers offer several advantages over traditional methods. These advantages stem from the unique properties and capabilities of transformer architectures, particularly in handling complex sequences and contextual information. Here are some key advantages:

Improved Contextual Understanding

Transformers, especially models like the Vision Transformer (ViT) and Vision-Language Transformers (like CLIP), excel at capturing the contextual relationships between different elements in an image and the corresponding text. This results in more accurate and coherent captions.

Handling Long-Range Dependencies

Transformers are designed to handle long-range dependencies more effectively than recurrent neural networks (RNNs) or convolutional neural networks (CNNs). This allows for better interpretation of images that contain multiple objects and intricate scenes.

Parallel Processing

Transformers process all elements of a sequence simultaneously, rather than sequentially as RNNs do. This parallel processing capability significantly speeds up training and inference times, making the caption generation process more efficient.

Scalability

Transformers can be scaled up in terms of the number of layers and parameters, which can lead to improvements in performance. This scalability is crucial for handling large datasets and complex tasks, leading to better generalization and more sophisticated captioning capabilities.

Consistency and Fluency in Captions

Transformers generate captions that are more consistent and fluent. The self-attention mechanism in transformers allows the model to focus on different parts of the image and the previously generated text, resulting in more coherent and contextually relevant captions.

IV. RESULT AND DUSCUSSION :

An image caption generator typically produces a brief, descriptive text that summarizes the content and context of an image.The generated captions aim to be coherent, relevant, and reflective of the visual elements and context of the image.

Result :

Image:

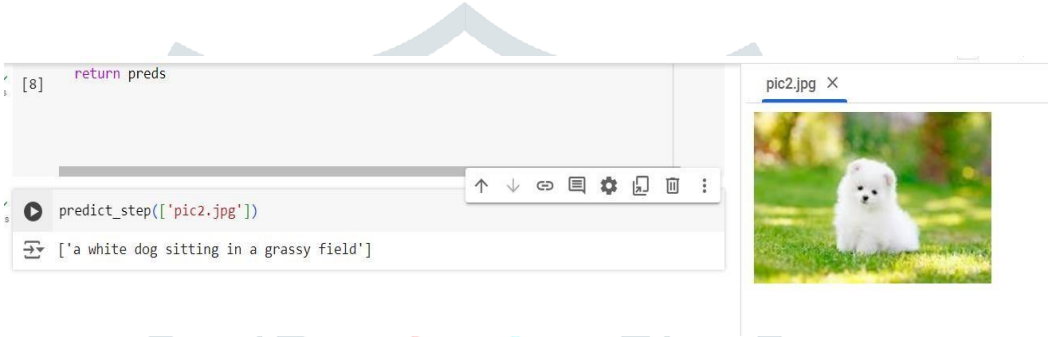

Figure 7.2 Final output for the image captioning generatorGenerated Caption: " 'a

white dog sitting in a grassy field'."

By focusing on these elements, an image caption generator aims to create informative and engaging captions that enhancethe understanding and appreciation of the image.

1. **Objects and Subjects:** The caption will identify the primary objects or subjects in the image, such as people, animals, or inanimate objects.

Example: "A young girl holding a red balloon."

2. **Actions:** It will describe any actions being performed by the subjects in the image.

Example: "A dog running through a field."

3. **Context and Settings:** The caption may include information about the setting or environment where the image was taken.

Example: "A family enjoying a picnic in a park."

4. **Attributes and Details:** Specific attributes or details about the subjects or objects, such as colors, sizes, or conditions, might be included.

Example: "A blue car parked next to a brick building."

V. CONCLUSION:

In conclusion, building an image caption generator using transformers presents a promising avenue for advancing the field of computer vision and natural language processing. Overall, an image caption generator using transformers represents a powerful tool for bridging the gap between visual content and natural language understanding, opening up new possibilities for communication, creativity, and storytelling in the digital age. By continuing to innovate and address emerging challenges, we can unlock the full potential of this technology and create meaningful impact across diverse domains and industries.

Future Scope:

The future scope of an image caption generator project is promising, with opportunities for further advancements in technology, research, and applications. Here are some potential areas of future development:

1. Improving Caption Quality :

Enhancing the language model's understanding of context, semantics, and storytelling to generate more accurate, diverse, and contextually relevant captions. Exploring advanced natural language processing techniques, including pre-training strategies, transfer learning, and fine-tuning on domain-specific data, to improve caption quality.

2. Fine-Grained Image Understanding :

Incorporating fine-grained image understanding capabilities, including object recognition, scene understanding, and visual relationships, to generate captions with detailed and nuanced descriptions. Leveraging techniques from object detection, semantic segmentation, and visual question answering to enhance the model's perception of visual content and generate more informative captions.

3. Adapting to User Preferences :

Personalizing caption generation based on user preferences, interests, and demographics to deliver tailored and relevant captions that resonate with individual users.Developing interactive interfaces and feedback mechanisms to allow users to provide input and guide the caption generation process according to their preferences.

4. Ethical and Inclusive Captioning :

Ensuring that image caption generation systems are sensitive to ethical considerations, cultural diversity, and inclusivity, avoiding biases, stereotypes, and offensive language in generated captions. Incorporating mechanisms for bias detection and mitigation, fairness assessment, and diversity promotion to promote responsible and inclusive captioning practices.

REFERENCES

1. Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems, pages 8024– 8035.

2. Aung, San & Pa, Win & nwe, tin. (2020). "Automatic Myanmar Image Captioning using CNN and LSTM-Based Language Model." Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), pages 139–143. Print.

3. Ali Furkan Biten, Lluis Gomez, Marc¸al Rusinol, and Dimosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 12466–12

4. J. Song, Y. Guo, L. Gao, X. Li, A. Hanjalic, and H. Shen, ―From deterministic to generative: multi-modal stochastic RNNS for video captioning,‖ IEEE Transaction on Neural Networks and Learning System, vol. 30, no. 10, pp. 3047– 3058, 2018.

5. J Vaishnavi; V Narmatha. ―Video Captioning based on Image Captioning as Subsidiary Content‖ Image caption generation and video captioning. 2022 IEEE, 2022.