Pictophrases: teaching machines to describe images

Team Members :

1. **Darshan Rajopadhye**
2. **Saumya Gupta**
3. **Kevin Heleodoro**
4. **Poornima Jaykumar Dharamdasani**

Abstract :

Image captioning models generate appropriate captions to describe images. These models describe the content of images - the objects present in the image, their colors, and generally the relationship between the objects. These models have valuable applications across various domains, particularly in enhancing accessibility for individuals with visual impairments. They play a significant role in tasks such as content indexing and retrieval, benefiting fields like e-commerce and digital asset management, where automatic tagging and categorization of images are essential. Integration of image captioning into assistive technologies, such as screen readers, enriches web content by providing detailed image descriptions on websites and in documents. In education and healthcare, these models extract pertinent information from images, facilitating diagnoses and saving valuable time. Lastly, image captioning models contribute to content moderation on social media platforms by identifying and filtering inappropriate or harmful images as part of content moderation efforts.

Many image captioning models have shown promise, but they struggle to generalize to out-of-domain images with novel scenes or objects. These models typically generate captions solely based on visual information, lacking a deep understanding of context, relationships between objects, fine-grained details, and the complexity of images, which may have multiple interpretations.

We aim to build an image captioning model using different combinations of Convolutional Neural Network architecture along with Long short-term memory (LSTM) and incorporate attention mechanisms based on the previous works [1] done on the topic of image captioning by making the model focus on relevant features of the image. We hope to generate better and more relevant captions for complex images by the end of this project than any other traditional image captioning model. By quantifying metrics such as BLEU, METEOR, and CIDEr, we measure the quality and diversity of generated captions. Moreover, this project examines the computational demands of these models, considering their suitability for real-time applications. Currently, we are planning to use the MS-COCO dataset which contains a large collection of around 328,000 images. These images cover a wide range of scenes, objects, and concepts, making them diverse and representative of real-world scenarios. We are also exploring the possibility of using other datasets widely known to be used for image captioning like Flickr30k, Visual Genome, and SBU Captioned Photo Dataset.

Ultimately, the results of this project will offer valuable insights into the state of the art in image caption generation. It will guide practitioners and researchers in selecting the most appropriate model for their specific use cases and inspire innovations in the field of computer vision and natural language understanding.