Pictophrases: teaching machines to describe images

Darshan Rajopadhye, Saumya Gupta, Kevin Heleodoro, Poornima Jaykumar Dharamdasani

Proposal :
In this project, our primary objective is to create an image captioning model capable of generating rich and descriptive captions for images. Our model is structured into two distinct stages: The initial stage involves extracting intricate image features through the utilization of Convolutional Neural Networks (CNNs). In the subsequent stage, we harness these extracted features to generate captions using advanced text generation techniques.[1]

For the text generation component, we explore and evaluate two prominent methods: Bi-directional Long Short-Term Memory (LSTM) models enhanced by the integration of attention mechanisms and the more recent Transformer architecture. Through this comparative analysis, we aim to discern the strengths, weaknesses, and overall performance of each method to determine which is better suited for the task of image captioning in terms of accuracy, fluency, and overall quality.

Proposed Architecture :
Our image captioning model employs a comprehensive three-part architecture, featuring an initial image feature extraction stage utilizing a CNN (such as Xception, ResNet50, or Faster RCNN). To generate image captions, we meticulously assess and compare two leading methods: a Bi-directional LSTM [2] equipped with attention mechanisms to enhance the quality and relevance of the generated captions, and self-attention-based Transformers [3], ensuring a thorough evaluation of both techniques for optimal captioning performance.

Dataset :
The Microsoft Common Objects in Context (MS-COCO) dataset [4]: It is a vast dataset for object detection, image segmentation, and image captioning. This dataset contains many features, such as image segmentation, 328,000 images, 91 object classes, and five captions for each image. This will be a valuable resource to train and evaluate our model effectively.

Evaluation Metrics :
We will assess the model's performance using:
- METEOR (Metric for Evaluation of Translation with Explicit Ordering) evaluates generated text using a robust method that is more akin to human judgments.
- BLEU (Bilingual Evaluation Understudy) evaluates generated text against reference text in terms of sequences of words.
- CIDEr (Consensus-based Image Description Evaluation) evaluates generated text by computing the similarity to the reference text based on a consensus of human annotations.

References :

1. Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. 2023. Deep Learning Approaches on Image Captioning: A Review. ACM Comput. Surv. 56, 3, Article 62 (March 2024), 39 pages. https://doi.org/10.1145/3617592.

2. Cheng Wang, Haojin Yang, and Christoph Meinel. 2018. Image Captioning with Deep Bidirectional LSTMs and Multi-Task Learning. ACM Trans. Multimedia Comput. Commun. Appl. 14, 2s, Article 40 (April 2018), 20 pages. https://doi.org/10.1145/3115432.

3. Hashem Parvin, Ahmad Reza Naghsh-Nilchi, and Hossein Mahvash Mohammadi. 2023. Transformer-based local-global guidance for image captioning. Expert Syst. Appl. 223, C (Aug 2023). https://doi.org/10.1016/j.eswa.2023.119774

4. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, Computer Vision –ECCV 2014, pages 740–755, Manhattan, New York, USA, 2015. Springer, Springer International Publishing.ISBN 978-3-319-10602-1