ECE-GY 7123
Deep Learning
Spring 2023

Image Captioning using Vision Transformer

Abirami S(as16288), Nagharjun M(nm4074), Rakshana BS(rb5118)

2023-04-12

1 Problem Statement

Our project aims to tackle the critical challenge of generating sentences from images, at the intersection of computer vision and natural language processing. This involves not only identifying objects in an image but also describing their relationships in a way that sounds natural in human language. With over 12 million visually impaired individuals [1] above the age of 40 in the United States alone, our goal is to provide accurate image captioning to establish the relationship between sequential images, thereby enhancing the mobility and independence of such individuals. Our solution involves training a deep learning model, specifically a Transformer-based neural network, to encode and decode visual features and linguistic structures in parallel, enabling the generation of highly accurate and semantically meaningful captions. Furthermore, our Transformer-based model integrates image captioning into one stage, enabling a comprehensive and powerful approach to handle the vast amounts of unstructured image data that dominate the digital landscape.

1.1 Literature Survey

Over the past few years, Transformer-based architectures have gained significant attention for computer vision tasks. Dosovitskiy et al. introduced the Vision Transformer (ViT), which encodes images into vector features using the Transformer[2] architecture without any image-specific biases. Inspired by ViT, He, Sen, et al. proposed the Image Transformer, a modified Transformer architecture designed for automatic image captioning, which exploits the spatial relationships[3] between image regions to achieve state-of-the-art performance on the MSCOCO dataset. In parallel, another group of researchers proposed a pure Transformer-based model for image captioning that replaces Faster R-CNN with SwinTransformer[4] as the backbone encoder for end-to-end training. They introduced a refining encoder to capture intra-relationships between grid features and a pre-fusion process of refined global features and generated words in the decoder to enhance multi-modal interaction. The recent advancements in Transformer-based architectures highlight their ability to improve various computer vision tasks and pave the way for future research in this direction.

1.2 Dataset

Our project will use the MSCOCO dataset, which is a comprehensive dataset for object detection, segmentation, key-point detection, and captioning. With a total of 328K images, including 82783 for training and 40504 for validation, and 40,775 testing images, the dataset offers a rich and diverse set of examples for training and testing our image caption generation model.

1.3 Goals

The goal of our project proposal is to develop an end-to-end transformer network for image captioning that can generate descriptions of images using only raw pixel data. This model will include a transformer encoder to extract image features and a transformer decoder to generate the caption word by word, with learned features fused before passing to the rest of the decoder. The proposed model will be trained on the MSCOCO dataset, comprising of 330,000 images with corresponding captions, and evaluated using both automated metrics and human evaluations to assess its performance.

2 References

1. Fast facts of common eye disorders. Centers for Disease Control and Prevention. (2022, December 19)
2. Dosovitskiy, Alexey, et al. Än image is worth 16x16 words: Transformers for image recognition at scale.ärXiv preprint arXiv:2010.11929 (2020).
3. He, Sen, et al. Ïmage captioning through image transformer.”Proceedings of the Asian conference on computer vision. 2020.
4. Wang, Yiyu, Jungang Xu, and Yingfei Sun. Ënd-to-end transformer based model for image captioning.”Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. No. 3. 2022.