World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:18, No:07, 2024

Image Captioning with Vision-Language Models

Authors : Promise Ekpo Osaine, Daniel Melesse

Abstract : Image captioning is an active area of research in the multi-modal artificial intelligence (AI) community as it connects vision and language understanding, especially in settings where it is required that a model understands the content shown in an image and generates semantically and grammatically correct descriptions. In this project, we followed a standard approach to a deep learning-based image captioning model, injecting architecture for the encoder-decoder setup, where the encoder extracts image features, and the decoder generates a sequence of words that represents the image content. As such, we investigated image encoders, which are ResNet101, InceptionResNetV2, EfficientNetB7, EfficientNetV2M, and CLIP. As a caption generation structure, we explored long short-term memory (LSTM). The CLIP-LSTM model demonstrated superior performance compared to the encoder-decoder models, achieving a BLEU-1 score of 0.904 and a BLEU-4 score of 0.640. Additionally, among the CNN-LSTM models, EfficientNetV2M-LSTM exhibited the highest performance with a BLEU-1 score of 0.896 and a BLEU-4 score of 0.586 while using a single-layer LSTM.

Keywords : multi-modal AI systems, image captioning, encoder, decoder, BLUE score
Conference Title : ICEMNLP 2024 : International Conference on Empirical Methods in Natural Language Processing
Conference Location : Santorini, Greece
Conference Dates : July 11-12, 2024