Task 3: Image Captioning

Author: Sunny Kumar, B.Tech CSE 2nd Year

Introduction

Image Captioning is a challenging Artificial Intelligence (AI) task that combines Computer Vision and Natural Language Processing (NLP). The goal is to generate meaningful descriptions of images by first extracting features using pre-trained CNNs such as VGG, ResNet, or Inception, and then passing these features into a sequence model like RNNs, LSTMs, or Transformers to produce captions.

Methodology

1. **Feature Extraction**: Used InceptionV3/ResNet pre-trained on ImageNet to extract image features. 2. **Sequence Modeling**: Used an Encoder-Decoder model with LSTM to generate captions. 3. **Dataset**: Popular datasets include Flickr8k, Flickr30k, and MS COCO. 4. **Training**: The model is trained to minimize categorical cross-entropy loss. 5. **Output**: Once trained, the model generates captions for unseen images.

Sample Image and Caption

Generated Caption: 'A dog is standing on the grass and looking to the side.'

Conclusion

Image Captioning demonstrates the power of combining vision and language models. This task deepens understanding of AI applications in accessibility, autonomous systems, and human-computer interaction.

Author Declaration

I, Sunny Kumar (B.Tech CSE 2nd Year), hereby declare that this project was completed as part of my internship task on Image Captioning.