Image Captioning

It’s a computer vision task that generates descriptions about an image and creates human readable captions using concepts of NLP.

The initial task involves feature extraction. The goal is to extract high-level visual features from the image that can be used as input to the captioning model. CNN pre-trained models like VGG, mobileNet are a good suit to achieve this.

Before feeding images to these models we need to pre-process the images in specific formats which the models are expecting. This involves tasks like resizing, normalizing pixel values to a specific range (e.g., [0, 1] or [-1, 1]).

Of the features extracted we need to apply pooling to reduce the spatial dimensionality while retaining the most important features. The extracted features are then passed to a transformer model which act as a language model.

The language model generates a sequence of words or tokens, usually starting with an initial token like "start" or "<SOS>" (start of sequence). At each step, it predicts the next word based on the current word and the context (which includes the image features and the previous words in the sequence).

The caption generation process continues until an end token like "end" or "<EOS>" (end of sequence) is generated or a predefined maximum sequence length is reached.

After generating the sequence of words, the model may perform post-processing to improve the quality of the caption. This can include removing duplicate words, capitalizing the first letter, and ensuring proper sentence structure.