Group 25

AML Project Deliverable 1

Background and context to the problem statement: **Image Captioning**

For this project, we are interested in the intersection of computer vision and natural language processing. Image captioning involves generating natural language descriptions for images, and it has a wide range of applications in the real world. For example, it can provide descriptions of visual content to visually impaired individuals. Image captioning is of critical importance to security. It can also be valuable in research when analysis for large corpus of data may be necessary. We could also extend our topic from image captioning to video captioning.

Dataset description:

- Flickr30k Image dataset: Link

It consists of 31,783 coloured images with their corresponding captions. Five reference sentences are provided as captions by human annotators, for each image. Therefore, there are a total of 158,915 records in the dataset, mainly of humans and animals performing daily activities.

- Fast-AI-COCO dataset: Link

COCO is a large-scale object detection, segmentation, and captioning dataset.
It contains approximately 118,287 training images, 5,000 validation Images and 40,670 test images. COCO includes a wide range of object categories, such as people, animals, vehicles, household items, and more. There are 80 object categories in total

Proposed ML techniques to solve our problem:

- The project will involve using two neural networks, one for extracting features from the raw images and another one for captioning the embeddings of images. We will use a Convolutional Neural Network (CNN) for feature extraction from our image dataset and a Transformed architecture or a Recurrent Neural Network (RNN) for training the model with the captions.
- The feature extraction with a CNN involves learning features from raw pixels using transformations in each layer of the network that provide more useful embeddings of the original image. The input layer slides through the image and by combining data learned through different filters moves from the simplest features in a photo to the most complex features.
- We could use Transformer models such as BERT for the text captioning. They capture long term dependencies between the image and text, which would help out with forming sentences.
- Alternatively, we could use a recurrent neural network which would be similarly powerful at capturing dependencies in an NLP context. We're planning to train with both and then compare the performance of the two models.