**Model Development for Autonomous Image Captioning Using Transformer Architecture**

**1. Project Background and Topic Summary**

Image captioning specifies a dedicated research area gaining significant level of importance as it aims at generating descriptions in natural language for the visual content present in stationary images (Deorukhkar and Ket, 2022). This research focuses on the research problem of autonomous image captioning, which consists of numerous capabilities, involving significant work across wide range of application domains like captioning medical images, analysis of traffic data and facilitating interaction between human and computer (Ghandi et al, 2023). For human beings, describing visual content of images and providing captions to it is a quick daily activity which doesn’t need any kind of complex activities. A machine that autonomously generate image captions is required to carry out a number of complex activities like categorizing image objects, abstracting attributes of the image, classification of relation between objects and its attributes and scene deducing in accurate manner for generation of caption briefing (Deorukhkar and Ket, 2021). In other words, the complexity involved in autonomous image captioning is because of the machines expected to carry out many activities like object recognition from visual content of the image, extracting the information about the object, knowing about image context besides presenting the captions of the image using natural language. Despite the existence of such complex activities, it is still important to carryout them because of its usage across wide application areas like medical image captioning, X-ray description, assisting visually impaired and content related image retrieval in biomedicine (Hossain et al, 2019).

**Project Rationale**

The problem of the autonomous image captioning was addressed in the existing literature using deep learning techniques (Ghandi et al, 2023; Stefanini et al, 2023). Some of these existing deep learning techniques applied for image captioning are attention method, graph based techniques, reinforcement learning techniques and the convolutional networks (Ghandi et al, 2023). Also, evaluation of these existing techniques showed the presence of the limitations within the image captioning process. Some of these limitations are related to concrete object description, absence of considering the context of visual content and dependency of each word generated on the previous words generated in the sentence (Stefanini et al, 2022). This shows the existence of significant area of improvements and potential future research directions in this area of image captioning, which include development of high quality captions for the images, enhancing the caption generation through incorporation of vision-language methods. The rationale of undertaking this research is to apply transformer architecture to develop an autonomous image captioning model. MS COCO and Flickr 8k datasets are considered in this research to train and test the autonomous image captioning model as part of this research work. This project is also involved in applying the transformer architecture and comparing it with recurrent neural networks and long short term memory models. Transformer architecture is an architectural model of deep learning, which used self-attention method for input data processing and handling a large sequence of the data in parallelized manner (Min et al, 2022). Recurrent Neural Networks (RNN) come under a kind of the artificial neural network involved for processing the sequential data (Salem, 2022). Long Short Term Memory (LSTM) is a type of the recurrent neural network, which was developed to address the gradient problem faced by the recurrent neural networks. LSTM working is based on applying memory cell along with system of the fates used for controlling information flow (Min et al, 2022).

**Aim and Objectives of Project**

This research is aimed at developing an autonomous image captioning model using transformer architecture.

Research Objectives

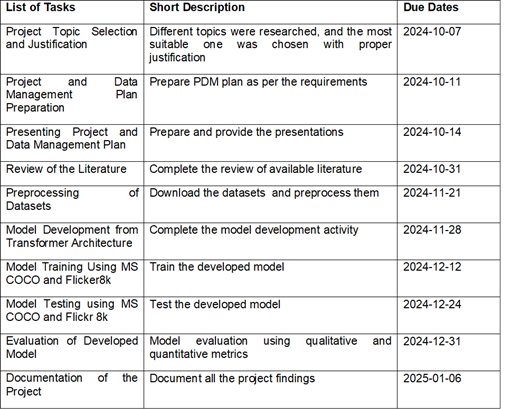
* To analyse the importance of developing a model that can perform image captioning activity in an autonomous manner.
* To develop an autonomous image captioning model using transformer architecture and machine learning algorithms.
* To train the developed model to perform autonomous image captioning using MS COCO and Flickr 8k datasets.
* To test the developed model using MS COCO and Flickr 8k dataset for autonomous image captioning.
* To evaluate the developed transformer model by comparing with Recurrent Neural Networks (RNNs) and Long Short Term Memory (LSTM) models.

**Research Question**

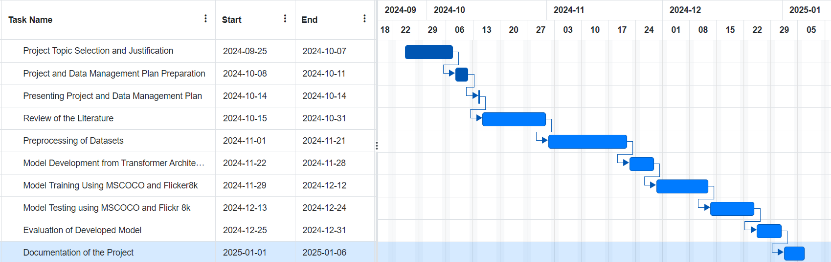
How does a transformer architecture support the model development for autonomous image captioning?

**2. Project Plan: Task List and Project Timeline**

**Task List**



**Timeline**

****

**3. Data Management Plan**

**Overview of Dataset**

MS COCO dataset was collected by Microsoft Corporation by selecting images about the everyday scenes and the objects. Flickr 8k dataset was collected by selecting images from the Flickr image sharing platform. Both MS COCO and Flick 8K datasets contain images about the everyday events of the people. Yes, the image dataset is greatly anonymised.

**Data Collection**

**Data set1: MS COCO**

<https://cocodataset.org/#home>

**Data set 2: Flickr 8K**

<https://www.kaggle.com/datasets/adityajn105/flickr8k>

**Meta Data**

MS COCO and Flick 8k datasets are both images. MS COCO consists of nearly 330,000 images where 200,000 images are labelled. Flick 8k consists of 8000 images.

**Document Control**

**Readme File:** This file will contain list of steps that are to be followed to run the developed model and generate autonomous image caption.

**Security and Storage**: Project files will be backed up once in a week into Google Drive of my email that is password protected. Project files and documents will be shared to the supervisor through GitHub and OneDrive depending on the requirements.

**Ethical Requirements:** The chosen datasets: MS COCO and Flicker8k doesn’t come under GDPR requirements as it doesn’t consist of any personal information of the people. Also, this project meets UH ethical policies due to the use of secondary data for model development. There exists ethical permission and license to use MS COCO and Flickr datasets for this project work.

**References**

* Deorukhkar, K., & Ket, S. (2022). A detailed review of prevailing image captioning methods using deep learning techniques. *Multimedia Tools and Applications*, *81*(1), 1313-1336.
* Geetha, G., Kirthigadevi, T., Ponsam, G. G., Karthik, T., & Safa, M. (2020, December). Image captioning using deep convolutional neural networks (CNNs). IOP publishing.
* Ghandi, T., Pourreza, H., & Mahyar, H. (2023). Deep learning approaches on image captioning: A review. *ACM Computing Surveys*, *56*(3), 1-39.
* Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, *51*(6), 1-36.
* Min et al (2022). Transformer Model: An overview from architecture perspective a*rXiv:2202.08455*.