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

## Problem Statement and Study Rationale

The area of the research that is focused on this study is the development of an image description or the captioning model for the images that can work in autonomous manner. In general, humans tend to describe the images in their daily life activities where, describing the given images is a very simple task for the humans (Wang et al, 2020). However, making the computers to describe the given images tend to be highly complex and time taking process. The main reason for this is due to the need of computers performing different tasks like object recognition, relate the object with other elements in the image and apply the natural language processing to express the descriptions of the image (Sharma et al, 2020). Even through this task of generating descriptions/captions for the images is complex, it is considered as highly important because of its application across medical field to describe the x-rays and support blind people understand the images through computer generated automated captions (Hossain et al, 2019). The activity of addressing this complex problem of generating the image captions come under the research area of computer vision, which involves in combining the image processing along with the pattern recognition for having better understanding of images (Tov et al, 2021). Even though computer vision functionality is similar to that of the human eye functionality, the level of intelligence, functionality and performance of this computer vision tend to be less than the human eye (Vanneste et al, 2021). For instance, human brain even in the subconscious state can undertake all activities like image feature recognition and image description, but for the computers all these images tend to be displayed as numerical values, so importance tends to be given for identifying patterns so that they can be recognised for describing the images (Herdade et al, 2019). However, computer vision techniques can only support image recognition but automatically describing the image and its elements tend to be highly challenging in nature.

To deal with these complex activities, deep learning techniques were used in the literature where convolutional neural networks and the recurrent neural networks were employed (Hossain et al, 2019). The application of the deep learning method resulted in categorizing the image caption methods as either top-down or the bottom-up approaches, which tend to differ in terms of the top-down approach generating the semantic representation followed by decoding to caption, whole bottom-up being involved in generating items and integrating them for forming the captions (He et al, 2019). Here, top-down approaches require extra overhears of learning the language modelling process, while the bottom-up approach suffer from the problem of overfitting. The evidence from Vanneste et al (2021) indicated that due to this overfitting, there is a change of the model developing the caption for more than two images. This shows that even through there are automatic models for generating descriptions for the images, there is significant room for the improvement of these models. Also, due to the need of automatically generating descriptions for the images across the healthcare applications, there is still a significant room for improvement in terms of developing a model that can automatically generate the captions with high accuracy level for the images (Ghandi et al, 2022). As a result, this research will be involved in developing a transformer-based model that will be trained and testing using MS-COCO dataset.

**Aim and Objectives of Project**

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

Research Objectives

* To analyse the importance of developing a model that can perform image captioning activity in autonomous manner.
* To develop autonomous image captioning model using transformer architecture and machine learning algorithms.
* To train the developed model to perform the autonomous image captioning using MS Coco and Flickr8k datasets.
* To test the developed model using MS CoCo and Flickr 8k dataset for autonomous image captioning.

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

**Methodology**

The methodological approach that will be used in this project for developing the project artefact is Rapid Application Development (RAD) methodology. The reason for using this rapid application development methodology is mainly because of the methodology supported by agile practices, which helps in doing the project activities on iteration and also because of helping the project to complete on iterative manner (Hirschberg, 2018). By applying this rapid application development methodology, this project will be carried out in structured manner by using various project steps like problem statement analysis, model development using transformer architecture, preprocessing the data set, training the model, testing the Image captioning model and evaluating it.

**Significance of the Study**

The number of images that are generated and currently available in the internet has extensively increased over the past few years. Understanding the contents of the images tend to be easy for humans but for computers it is a complex activity. The task of automatically generating captions for the images has been gaining high level of importance in different application areas like supporting the blind people and understanding the medical images in automatic manner (He et al, 2019). The process of providing description for the images in an accurate manner is greatly based on having a dedicated image description model which can understanding image and provide description for it in the natural language (Ghandi et al, 2022). The importance of achieving high accuracy in providing the descriptions for the images as mentioned in Bernardi et al (2016) resulted in carrying out this research activity to develop an image description model. The contribution of this research activity is not just associated with developing the model but is also involved in identifying whether Flickr-8K and MS-COCO datasets can help in achieving high accuracy level of the predictions for the given images.

**References**

* Ghandi, T., Pourreza, H., & Mahyar, H. (2022). Deep Learning Approaches on Image Captioning: A Review. *arXiv preprint arXiv:2201.12944*.
* He, S., Tavakoli, H. R., Borji, A., & Pugeault, N. (2019). Human attention in image captioning: Dataset and analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 8529-8538).
* Herdade, S., Kappeler, A., Boakye, K., & Soares, J. (2019). Image captioning: Transforming objects into words. *Advances in neural information processing systems*, *32*.
* 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.
* Sharma, H., Agrahari, M., Singh, S. K., Firoj, M., & Mishra, R. K. (2020, February). Image captioning: a comprehensive survey. In *2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC)* (pp. 325-328). IEEE.
* Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., & Cohen-Or, D. (2021). Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, *40*(4), 1-14.
* Vanneste, P. ET AL (2021). Computer Vision and Huamn behavoour, Emotion and congnitoon, detection. Mathematics, 9(5), 298
* Wang, Z., Feng, B., Narasimhan, K., & Russakovsky, O. (2020). Towards unique and informative captioning of images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16* (pp. 629-644). Springer International Publishing.