Name: Fahad Ahmed Pranto

Id: 16122607042

Paper: **Autonomous Image Caption Generator**

Link: //scholarworks.gvsu.edu/cgi/viewcontent.cgi?article=1005&context=mcnair\_manuscripts

**Abstract:** This research goal is to create an automated device able to create image descriptions that are semantically specific. Two are included in this issue Tasks: recognizing and grammatically explaining an image. Their approach uses state-of-the-art machine learning: using a deep learning approach   a convolution neural network that "learns" to recognize images and retrieve their influential characteristics, and a recurrent neural network that learns to create meaningful, standardized words. This network functions are integrated in order to generate a single platform that arbitrary images are used as input and related captions are output. Using different language metrics, such as the Bilingual Test Understudy intended to evaluate language translation model, the model's accuracy is analyzed.

**Introduction:** With the growth of the internet and the development of social media, online images are embedded in our everyday lives. Many of these pictures, sadly, lack corresponding explanations. Deep learning is a computer science area that is a machine learning subcategory, and is a subfield of artificial intelligence itself. The use of artificial neural networks, computational algorithms focused loosely on how the brain functions, centers on deep learning. Those networks operate by taking in a vector input, which is converted by the weights producing the next layer's values. This process is replicated across the network layers, allowing for several length scales before the value of the output is produced. The generated results are compared to the given ground truth output while the network is practicing, enabling the model to change its weights and "learn" by using backpropagation.

**Literature review:** Automated image captioning is a concern that stands at the intersection of two main fields of research in deep learning, computer vision (CV) and natural language processing (NLP). Two specialized deep neural network architectures combine their approach to the image captioning problem. To derive the image's characteristics, a convolution neural network architecture is used, and a recurrent neural network architecture is used to create the picture definition.

**Methods:** The encoder-decoder architecture used in machine translation and the textural architecture are the two key technologies used in image captioning. A CNN is also used to retrieve the various features of the image in a traditional encoder-decoder method. These characteristics are then transferred to the word embedding, which produces terms that are incorporated to create a caption using the feature representations. A CNN is used to remove image attributes in a traditional compositional system, these characteristics are then converted to the language model that produces several possible captions. Then these descriptions are listed using a deep model of multisensory comparison, and the best quality title is picked. Examples of the architecture of the encoder-decoder can be found and examples of the structure of the composition are given in.

**Datasets:** For the purposes of examining and addressing the topic of image captioning, a variety of separate databases have been released that represent image caption sets. The most popular datasets are Flickr8k [16], containing a total of 8,000 images, Flickr30k [66], containing 30,000 images, and MSCOCO [32], containing more than 300,000 images. This research study used the abstract caption dataset of Google containing 3.3 million images, each with just one relevant caption.

**Results**: One way to get a feel for the success of a model and whether or not it is improving is to analyze a few instances of training examples, noticing how through the training period their captions evolved. For the whole training package, the training phase can be split into runs, known as epochs**.** They can see that by examining these cases, while the captions are

Not great, the standard at epoch 30 seems to drop. They can see in the illustrations the captions developing as the train sets. This means the model is beginning to learn how to generate captions correctly.

**Conclusion:** After testing on a portion of Google's computational caption dataset, the existing system is capable of producing captions of certain consistency. Work continues to develop both the capacity of the model to understand and to generalize, prompting more experimenting with different design choices and a closer analysis of the details. These extra studies, owing to the large scale of the dataset, It is possible that large quantities of time will be needed**.**