**Towered Visual Question Answering in Bangla Language**

**Motivation**

Large-scale vision and language representation learning has shown promising improvements on various vision-language tasks. As computers are becoming efficient at understanding visual information and transforming it into a written representation, research interest in tasks like automatic image captioning has seen a significant leap over the last few years. Visual question answering is a combination of vision and language reasoning and a challenging task under the zero-shot setting. Recent years have witnessed unprecedented performance gains on many natural language reasoning tasks, especially in zero-shot and few-shot settings.

Recent advancements in computer vision and deep learning research have enabled enormous progress in many computer vision tasks, such as image classification, object detection, activity recognition. As humans, we can identify the objects in an image, understand the spatial positions of their objects, infer their attributes and relationships to each other, and also about the purpose of each object given in the surrounding context. We can ask arbitrary questions about images and also communicate the information gleaned from them.

VQA is a computer vision task where a system is given a text-based question about an image, and it must infer the answer. Question can encompass many subproblems in computer vision such as:

Object Recognition - What is in the image?

Object Detection - Are there any humans in the image?

Attribute Classification - What is the color of the Car?

Counting - How many dogs are there in the image?

Our System will be capable of answering input questions from uploaded image data and it can help in educational sectors.

The Report is outlined as follows: Chapter 2: Background Study: where we have discussed available methods and algorithms that were used for VQA like {names of the models}. Chapter 3: Literature Review which is required for developing and increasing the performance of our chosen model. Chapter 4: Dataset, where we briefly described the dataset. Chapter 5: Proposed System, will contain the information of our proposed model and how it will be implemented. Chapter 6: Budget and Project Timelining shows the total cost and time estimation for the project. At last Chapter 7 gives a conclusion to our work.

1. Madical
2. Education

The project will be helpful for easy learning, our system doesn't match up with environmental issues

\section{Natural Language Processing}

Natural Language Processing(NLP) refers to the part of Artificial Intelligence-AI that is concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.

NLP combines computational linguistics—rule-based modeling of human language—with statistical, machine learning, and deep learning models. It drives computer programs that translate text from one language to another, respond to spoken commands, and summarize large volumes of text rapidly—even in real time. Several NLP tasks break down human text and voice data in ways that help the computer make sense of what it's ingesting. Some of these tasks includes :

\textbf{a. Speech recognition} which is the task of reliably converting voice data into text data. Speech recognition is required for any application that follows voice commands or answers spoken questions.

\textbf{b. Word sense disambiguation} which is the selection of the meaning of a word with multiple meanings through a process of semantic analysis that determines the word that makes the most sense in the given context.

\textbf{c. Sentiment analysis} that attempts to extract subjective qualities—attitudes, emotions, sarcasm, confusion, suspicion—from text.

\textbfd. {Natural Language generation} is sometimes described as the opposite of speech recognition or speech-to-text; it's the task of putting structured information into human language. etc

NLP can be used for solving many real life problems such as

\textbf{Spam Detection} - technologies that use NLP's text classification capabilities to scan emails for language that often indicates spam or phishing. These indicators can include overuse of financial terms, characteristic bad grammar, threatening language, inappropriate urgency, misspelled company names, and more.

\textbf{Machine Translation} - Google Translate is an example of widely available NLP technology at work. Truly useful machine translation involves more than replacing words in one language with words of another.

\textbf{ Virtual agents and chatbots} - such as Apple's Siri and Amazon's Alexa use speech recognition to recognize patterns in voice commands and natural language generation to respond with appropriate action or helpful comments. Chatbots perform the same magic in response to typed text entries.

\textbf{ Sentiment Analysis} - NLP has become an essential business tool for uncovering hidden data insights from social media channels. Sentiment analysis can analyze language used in social media posts, responses, reviews, and more to extract attitudes and emotions in response to products, promotions, and events–information companies can use in product designs, advertising campaigns, and more.

\textbf{ Text Summarization} - that uses NLP techniques to digest huge volumes of digital text and create summaries and synopses for indexes, research databases, or busy readers who don't have time to read full text.

\section{Convolutional Neural Network}

CNN is a type of deep learning model for processing data that has a grid pattern, such as images, which is inspired by the organization of animal visual cortex and designed to automatically and adaptively learn spatial hierarchies of features, from low- to high-level patterns. CNN is a mathematical construct that is typically composed of three types of layers: \textbf{convolution, pooling & fully connected layers}. The first two, convolution and pooling layers, perform feature extraction, whereas the third, a fully connected layer, maps the extracted features into final output, such as classification. A convolution layer plays a key role in CNN, which is composed of a stack of mathematical operations, such as convolution, a specialized type of linear operation. One layer feeds its output into the next layer, extracted features can hierarchically and progressively become more complex. The process of optimizing parameters such as kernels is called training, which is performed so as to minimize the difference between outputs and ground truth labels through an optimization algorithm called backpropagation and gradient descent, among others.

but sorry if my actions ever hurts you