Project 1

Visual Question Answering

**Problem Statement/Applications**

Visual Question Answering (VQA) system is an algorithm that takes an image and a natural language question about it as input and produces an accurate answer as the output. VQA represents not a single narrowly-defined problem but rather a rich spectrum of semantic scene understanding problems and associated research directions.

Visual Question Answering provides user a more convenient and effective way to analyze any given image during work, study, data analysis, research and life. For instance, users can easily get the number of bullets on the battle field, count numbers of people between various races, capture those key information that the user wants to quickly obtain on the picture, quickly establish data pool based on given images and even set up human facial recognition system. Therefore, VQA system can save a huge amount of time and human resources.

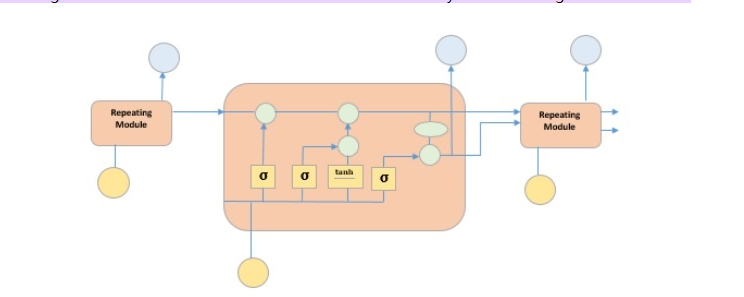
**Current Research**

There are two main VQA Models

1. LSTM(Long short-term Model) + CNN Model
2. Hierarchical Co-Attention Model

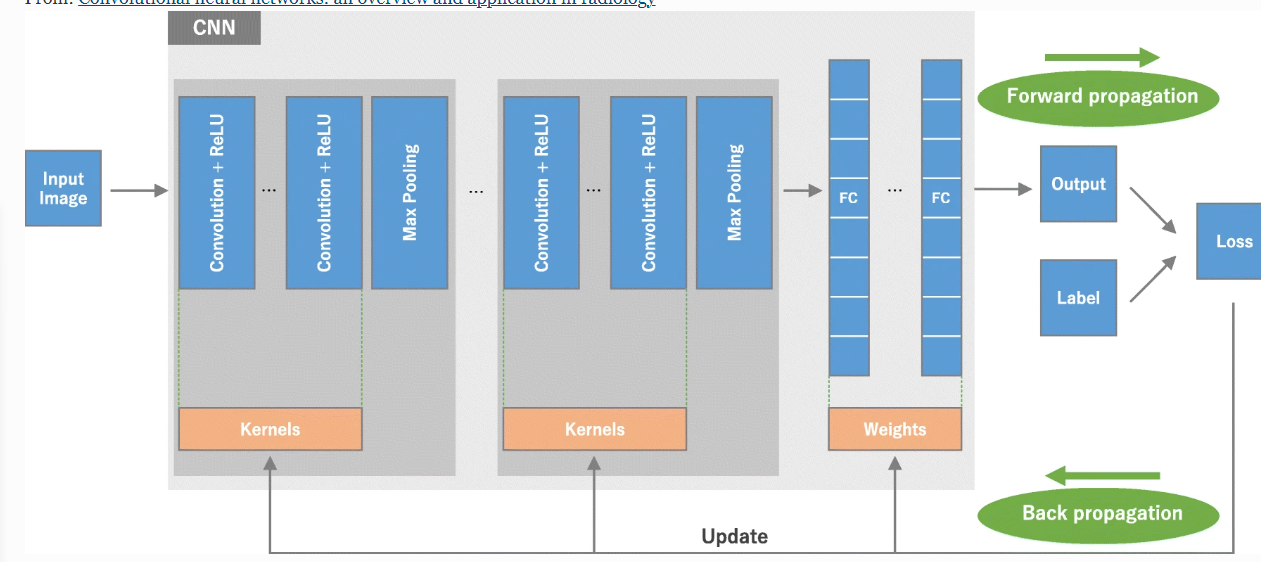
LSTM is defined as long short-term model, which is a special type of recurrent neural network that could learn long and short term dependencies in data.

The below picture shows that the LSTM is consisting of four neural network layers which are interacting with each other.



CNN Model is defined as convolutional neural network, which is a type of artificial neural network that specifically designed to process pixel data during image recognition and processing.

CNN Model is a combination of convolution layers, pooling layers fully connected layers



**Current problem**

Evaluating a VQA model correctly is a difficult problem. Testing sets always share the same distribution as the training sets. Therefore they always share the same biases or shortcuts. So currently, the mainstream of solution is to create out-of-distribution datasets. The OOD datasets are assumed that a biased model would fail to answer in rare cases, because it learned superficial correlations, while a correct model would manage to answer even in contexts that are different from its training set.

However, the correct evaluation with the out-of-distribution setup is still challenging because it does not match the usual assumptions made in machine learning.

Also, the data-set creation is a challenging problem. It is hard to gather enough out-of-distribution data that are different from the training set.

RESOURCE LINKS:

1. <https://sanghani.cs.vt.edu/research-project/visual-question-answering-vqa/>
2. <https://cdancette.fr/2020/11/21/overview-bias-reductions-vqa/>
3. <https://www.tutorialspoint.com/time_series/time_series_lstm_model.htm#:~:text=It%20is%20special%20kind%20of,layers%20interacting%20with%20each%20other.>
4. <https://insightsimaging.springeropen.com/articles/10.1007/s13244-018-0639-9/figures/>