TEXT VISUAL QUESTION ANSWERING

Akshay Shukla (2022A7PS0087P)
PhD Student Mentor – Ms. Vijay Kumari

AIM

- Answering questions on an image by extracting information from the visual elements as well as the embedded text in the image.
- Direct applications includes helping the visually impaired by answering questions on images in daily life as well as helping them in navigation.
- Using the individual and multimodal attention mechanism for effectively joining all the modalities in a space to provide the exact answer.
- Implementation of the 2D-Attention and Variational Encoder for the already proposed Text-VQA model.
- Inclusion of multilingual functionality in the already proposed Text-VQA model. This can be designed to supplement the OCR, NLP and decoder functions.

ELEMENTS OF THE CURRENT MODEL

- BERT model (designed by Google Research) is used for the Natural Language Processing (NLP) task. This will provide us with an entry point to pose a question to the model.
- ResNet-152 model is used for reading the image provided to the model. All questions asked to the model will be answered with the help of this image.
- Faster R-CNN is used for Optical Character Recognition (OCR) from the visual images given to the model.
- Datasets used for training -
  - VQA 2.0 (https://visualqa.org/download.html)
  - TextVQA (https://textvqa.org/dataset/)
- Transformer based neural network with self attention mechanism is used for encoding the output vectors obtained by a combination of the above mentioned models.
- Long Short-Term Memory (LSTM) is used for decoding the vector and generating the output sequence of characters

EXISTING BASELINE MODELS

- LoRRA (27.63% accuracy) - A basic model using self and contextual attention over visual and textual features along with a module to copy OCR tokens directly to the answers.
- M4C (40.46% accuracy) - Different modalities taken through transformer embeddings and then a joint embedding is created and sent to a single-multimodal transformer.

- SSBaseline (45.66% accuracy) - OCR and Visual features are handled separately by vanilla attention blocks and then served to a transformer.
- The code for the SS Baseline model is - https://github.com/ZephyrZhuQi/ssbaseline and the literature is - https://arxiv.org/pdf/2012.05153.pdf

WORK TO BE DONE

- Work is to be done by keeping our current Text VQA model as the base and introducing multilingual capability in it. This should include the user providing the model a question in any one of multiple possible languages and the model processing the data and giving the answer in the same language as the question.
- This can further be extended to the model being able to read multiple languages using OCR on the image provided even if the text in the image is slightly rotated or distorted.