UE20CS302: Machine Intelligence - Mini Project

Multi-modal Textbook answering

Team Number: 4

Team Guide: Prof.Bhaskarjyoti Das

Team members:

Shreyas S - PES1UG20C408

Shishira Bhat O - PESUG20CS397

Shashank Varma - PES1UG20CS395

Siddharth Kumar - PES1UG20CS424

Problem Statement

- Textbook Question Answering (TQA) is a task to choose the most proper answers by reading a multi-modal context of abundant essays and images.
- Since Textbooks consist of images and long texts, solving the problem of textbook answering requires multi-modal contexts in complex input data.
- TQA therefore connects computer vision and natural language processing and pushes forward boundaries of both fields.
- The questions, which vary in complexity are not simple lookup problems. These generally require in depth parsing and reasoning, which makes it necessary to have multi modal models which can analyze both the text and image based dataset given.

Application & Uses

1) The model can be widely applied in the field of academia to generate simple QA's from complex paragraphs of text along with image inputs.

2) In scenarios where the need for better understanding of the data is needed, simple parsing and elementary methods fail to provide accurate results. In these scenarios, the model presented can be applied.

3) While previous practices may help us solve true/false or simple mcq questions. Complex subjective questions with images adding context to the scenarios, these models fail to deliver. This calls the need for complex multi modal models.

Dataset

Dataset Source:
Textbook Question Answering(TQA) from Allen Institute for AI

Dataset size:

- The dataset is split into a training, validation and test set at the lesson level.
- The training set consists of 666 lessons and 15,154 questions,
- The validation set consists of 200 lessons and 5,309 questions.
- The test set consists of 210 lessons and 5,797 questions.

Dataset

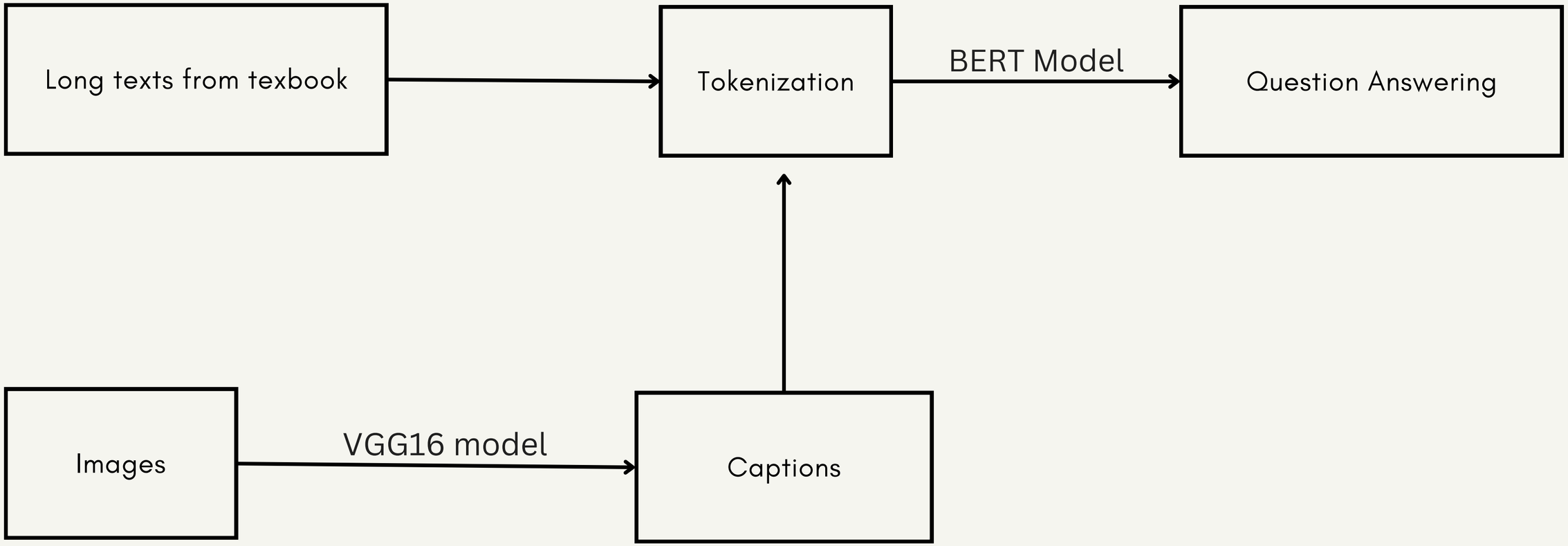
Dataset Source:
Flickr30k Image Dataset

Dataset size:

- This is used for training VGG16 model to generate captions for our images.
- It consists of 8091 images along with 5 captions associated with each image.

a little girl in a pink dress going into a wooden cabin .
a little girl climbing the stairs to her playhouse .
a little girl climbing into a wooden playhouse .
a girl going into a wooden building .
a child in a pink dress is climbing up a set of stairs in an entry way .

Architecture



Approach

Since the dataset consists of questions with and without images,we need to be sure to handle questions with Images differently from the ones which have images.

Questions without images:

- Extract the description of the topic the question is related to,from the JSON file.
- Pass this description to the BERT model.
- The BERT model will then be able to answer the questions by referring to the given description.

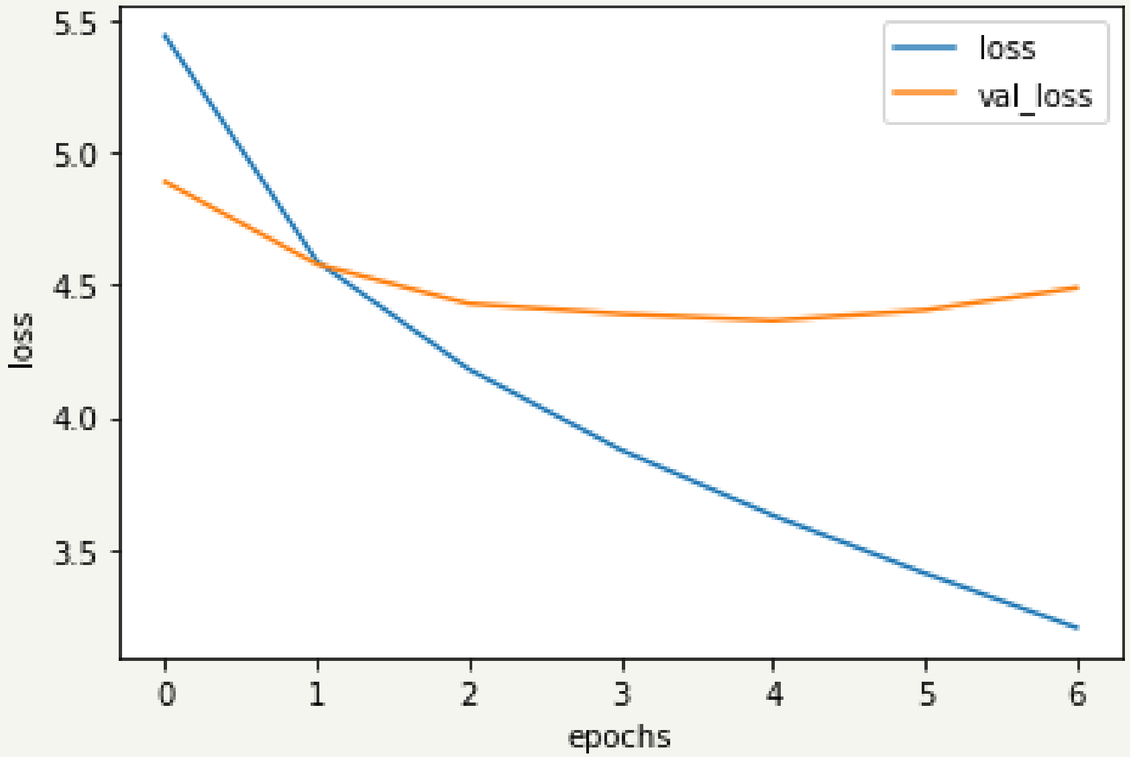
Approach

An image conveys a lot of information regarding the question or sometimes it is associated with the question itself.In such cases its important that we ensure to process the image properly.

If the question has images:

- Pass the image through a caption generation model (VGG16) to get the caption of the image.
- Pass the generated caption along with the given topic description as the final description to the BERT model.
- The BERT model will then be able to answer the questions after referring to the description that has been passed to it.

Results



THE LOSS VS EPOCHS PLOT FOR THE VGG-16 MODEL

Results

The final output we achieved was a model which could answer any question asked from the textbook.

```
Question:
Layer of the sun that surrounds the radiative zone

Answer:
The core , the ra ##dia ##tive zone and the convection zone ..
```

We also built a function which could answer questions based on a paragraph of text given as reference to it.

```
Please enter your text:
Virat Kohli is an Indian international cricketer and former captain of the India national cricket team. Widely regarded as one of the greatest batsmen of all time, Kohli plays as a rig

Please enter your question:
What team does he play for?

Predicted answer:
Royal challengers bangalore
```

References

1. Are You Smarter Than A Sixth Grader? Textbook Question Answering for Multimodal Machine Comprehension by Aniruddha Kembhavi,Minjoon Seo, Dustin Schwenk,Jonghyun Choi,Ali Farhadi,Hannaneh Hajishirzi,Allen Institute for Artificial Intelligence, University of Washington

2. Textbook Question Answering with Multi-modal Context Graph Understanding and Self-supervised Open-set Comprehension by Daesik Kim - Seoul National University, Seonhoon Kim - V.DO Inc.,Nojun Kwak - Search&Clova, Naver Corp.

References

3. A. S. Salim, M. B. Abdulkareem, Y. E. Fadhel, A. B. Abdulkarem, A. M. Shantaf and A. B. Abdulkareem, "Novel Image Caption System Using Deep Convolutional Neural Networks (VGG16)," 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)

4. Alloatti, Francesca & Di Caro, Luigi & Sportelli, Gianpiero. (2019). Real Life Application of a Question Answering System Using BERT Language Model. 250-253. 10.18653/v1/W19-5930.

5. Yuwen Zhang & Zhaozhuo Xu . BERT for Question Answering on SQuAD 2.0.

References

6. MD. ZAKIR HOSSAIN, FERDOUS SOHEL, MOHD FAIRUZ SHIRATUDDIN & HAMID LAGA.(2018). A Comprehensive Survey of Deep Learning for Image Captioning

7. Yuwen Zhang & Zhaozhuo Xu . BERT for Question Answering on SQuAD 2.0.

8. Shah Nawaz, Alessandro Calefati, Moreno Caraffini, Nicola Landro, Ignazio Gallo ·.Are These Birds Similar: Learning Branched Networks for Fine-grained Representations

9. Kuan Liu, Yanen Li, Ning Xu, Prem Natarajan. (2018).Learn to Combine Modalities in Multimodal Deep Learning

Thank you!