Mid-Term Project Report

ASD Detection Project

Riddhi Chatterjee - IMT2020094
Siddharth Yedlapati - IMT2020013

GitHub link to our codebase: https://github.com/Riddhi-Chatterjee/ASD-Detection

Problem Statement

The goal of the project is to detect whether a person has Autism Spectrum Disorder or not by analysing their response to verbal instructions.

The subject is provided with a cluttered set of simple objects (say on a table). An automated system (an avatar) is supposed to select one of the objects and come up with a task for the subject to perform using that object. For this, at first, the avatar needs come up with a structural and spatial textual description of the selected object; followed by a textual description of the task to perform (for example: putting the selected object out of the scene; or maybe behind another object). If the proposed task involves other objects in the scene, then structural and spatial descriptions of those objects also needs to be curated. Next, all this information regarding the task to be performed needs to be verbally conveyed to the subject. Finally, the subject's response to the verbal instruction needs to be analysed, and based on the correctness of the task actually performed by the subject, a decision needs to be made on whether the subject is Autistic or not.

This leads to the following sub-problems for this project:

- **Multi-Object Tracking and Selection:**
  Involves the development of a system that allows the selection of an object among various other objects of similar or different type; and allows robust tracking of the selected object. The selected object might be moved from its initial position, or might get hidden temporarily. The system should be robust enough to handle these cases and continue to track the selected object. Currently the user interacts with this system directly. Later on, an avatar would be introduced which would perform the object selections.

  In short, when the user clicks on an object, it should get selected (Tap Selection) and the system should track this object continuously. When the user clicks on a different object, it should become the new selected object and should be the one that the system tracks from this point onwards.

- **Reversed Visual Grounding:**
  The task of visual grounding of images involves pointing out a portion of the image which corresponds to the given text prompt. Our task of coming up with a structural and spatial textual description of one particular object kept among multiple other objects in a cluttered scene is exactly the reverse of the visual grounding task. Thus there is a need to develop a system which accepts an image of a cluttered scene of objects, as well as a mask corresponding to one particular object in the scene, and outputs the structural and spatial textual description of that object.

- **Curation of a task to perform and verbal communication:**
  Involves the development of a module which automatically generates a task that the subject needs to perform using the selected object. The task might be to simply put the selected object out of the frame/scene, or it might be to put the selected object at some place relative to some other object in the scene.

- **Analysis of task performed by subject:**
  This module is responsible for analysing the subject's response to the avatar's verbal instruction regarding the task to be performed. Based on whether the subject performs the proposed task correctly or not, the system should determine whether the subject is Autistic or not.

Brief overview of earlier work done

Work on the Multi-Object Tracking and Selection module was completed and work on the Reversed Visual Grounding module was started.

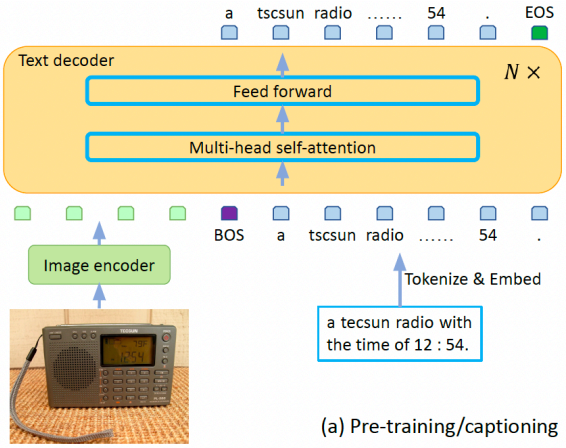
- **Progress with the Multi-Object Tracking and Selection module:** The module implements XMem — a robust real-time multi-object tracking solution, in our problem setting. XMem requires segmentation masks at the beginning of every tracking operation — and thus every time the user clicks on an object, we use the YOLO-v8 model to obtain the segmentation mask of the current scene, which is then passed onto the XMem model for it to start tracking the new selected object.
- **Initial work on the Reversed Visual Grounding module:** The Generative Image to Text (GIT) model by Microsoft was being experimented with. The original GIT model is a typical encoder-decoder based image captioning model. Thus designs for a modified GIT architecture were being developed which accepts both the image of the scene and the mask of an object to give the description of the object under consideration.

Current work on Reversed Visual Grounding

Work on the Reversed Visual Grounding module is completed (with some finishing touches yet to be given).

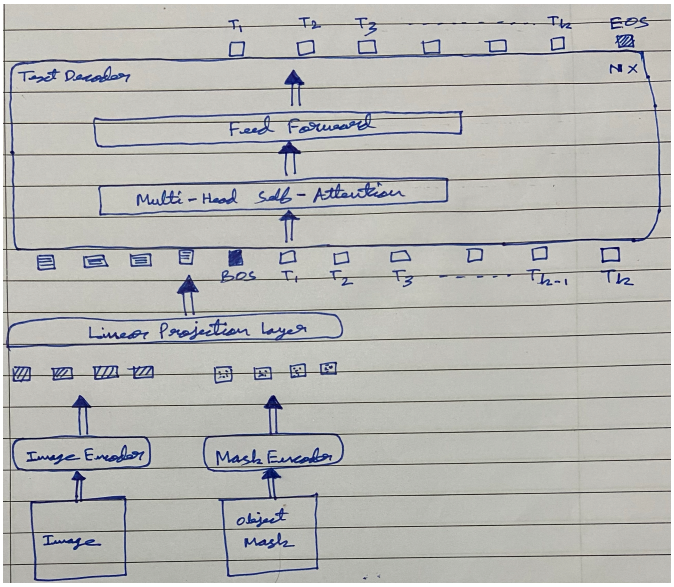
Development of the modified Microsoft GIT model

- **Architecture of the original Generative Image to Text model:**
  The Generative Image-to-Text (GIT) model for image captioning consists of two key components: an image encoder and a text decoder.
  - **Image Encoder:** Based on a Swin-like vision transformer, the encoder transforms raw images into compact 2D feature maps. These are projected into a fixed dimension and passed to the text decoder. This eliminates the need for external object detectors or OCR modules, making the architecture simpler.
  - **Text Decoder:** A transformer network processes the image features along with tokenised text. Starting with a [BOS] token, the decoder generates captions in an auto-regressive manner until an [EOS] token or the maximum length is reached. The attention mechanism ensures the decoder leverages both image features and preceding text tokens.
  - **Training:** The model is pre-trained on large-scale image-text pairs, using a language modelling task to map images to their corresponding captions. This unified approach allows GIT to achieve state-of-the-art performance in image captioning benchmarks without the need for task-specific modules.



(a) Pre-training/captioning

- **Architecture of the modified Generative Image to Text model:**
  The modified Generative Image to Text model uses an image encoder for encoding the image of the scene of objects and another image encoder for encoding the object mask. The image and mask features are then merged using a linear layer and projected into the same dimension as the inputs of the text decoder. The text decoder is then responsible for generating the structural and spatial textual description of the object.

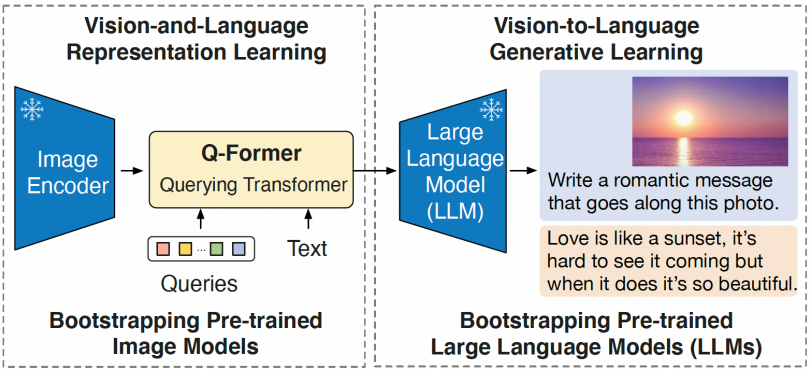


Inspite of making several architectural improvements and improvements to the training process, the modified GIT model failed to train properly. Thus we focussed on the BLIP-2 model, which is a much more recent work and uses frozen image encoders and LLMs to prevent catastrophic forgetting.

Development of the modified Salesforce BLIP-2 model

- **Architecture of the original BLIP2 model:**
  BLIP-2 (Bootstrapping Language-Image Pre-training) introduces a more efficient vision-language model by using frozen image encoders and frozen large language models (LLMs). The key to bridging the gap between these two frozen modules is the Querying Transformer (Q-Former). The Q-Former acts as an interface between the visual and language domains, extracting relevant visual features from the image encoder and preparing them for interpretation by the language model.

  The Q-Former consists of a set of trainable query embeddings that interact with frozen image features. These learned queries serve as information bottlenecks, pulling the most salient visual features from the image encoder, which are then used by the language model for various downstream tasks. By employing cross-attention mechanisms, the Q-Former links the visual representations to the text representations, making it easier for the language model to process the visual information without needing direct access to the image data.



- **Two-stage pre-training procedure of the Q-former:**
  1. **Vision-Language Representation Learning:** In the first pre-training stage, the Q-Former is connected to a frozen image encoder. The goal is to teach the Q-Former to learn visual representations that are most informative for the corresponding text. This is achieved through three objectives: Image-Text Contrastive Learning (ITC), Image-Text Matching (ITM), and Image-Grounded Text Generation (ITG). These objectives ensure that the Q-Former extracts relevant features from images that align with the text. During this stage, no interaction occurs between the text and image tokens directly; the trainable queries handle the alignment.

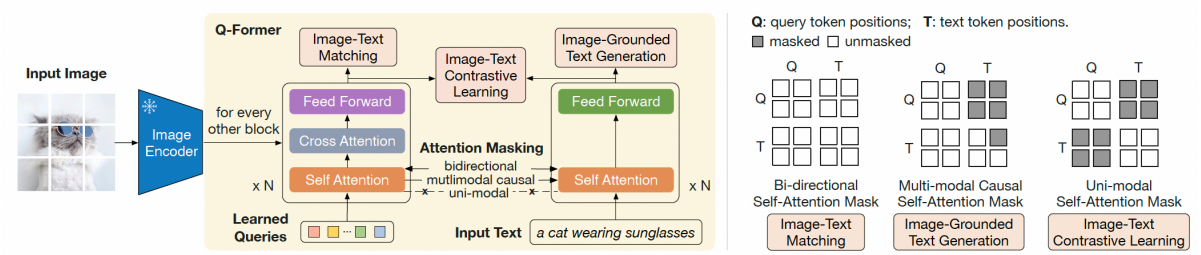


Figure 2. (**Left**) Model architecture of Q-Former and BLIP-2's first-stage vision-language representation learning objectives. We jointly optimize three objectives which enforce the queries (a set of learnable embeddings) to extract visual representation most relevant to the text. (**Right**) The self-attention masking strategy for each objective to control query-text interaction.

  2. **Vision-to-Language Generative Learning:** In the second stage, the output from the Q-Former is connected to a frozen LLM. The Q-Former, already trained to extract text-informative visual features, passes these features through a fully connected layer to the LLM. This setup enables vision-to-language generative tasks, where the LLM generates text conditioned on visual inputs. The frozen LLM can now generate coherent text based on visual features, enabling tasks like image captioning or visual question answering. Depending on whether the LLM is a decoder-based or an encoder-decoder-based model, different training strategies like language modeling or prefix language modeling are used.

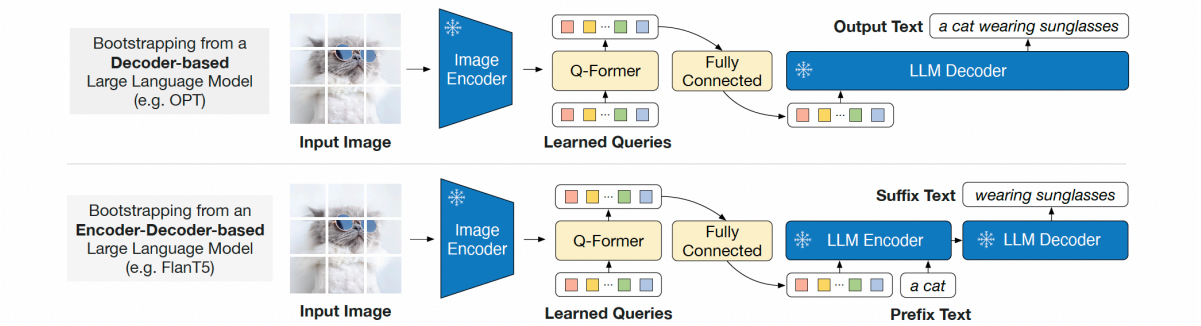


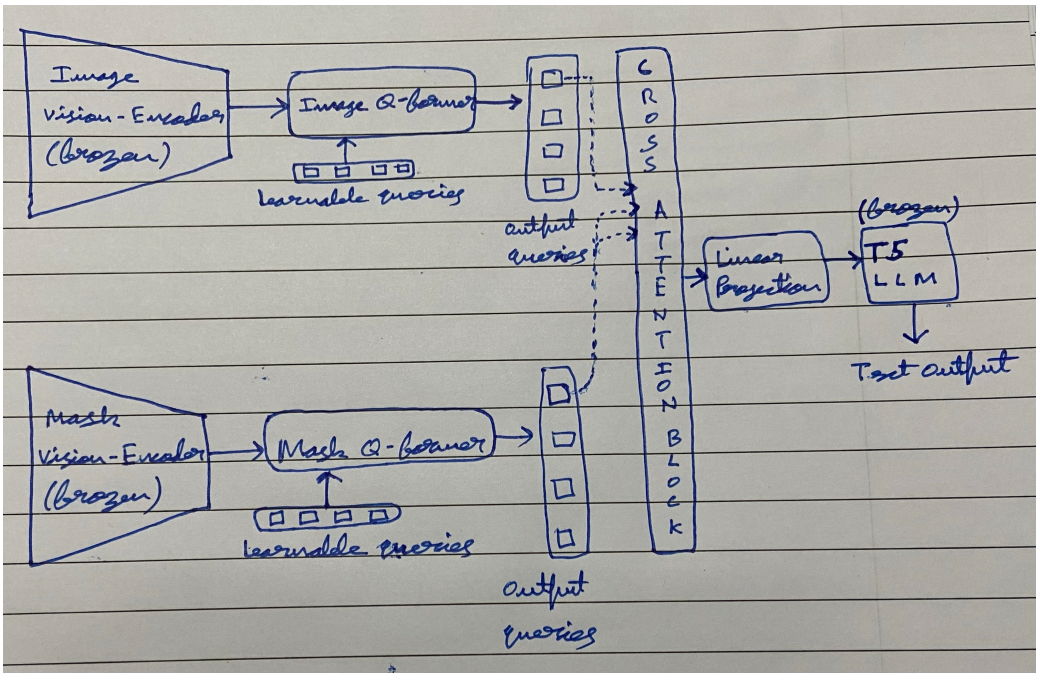
Figure 3. BLIP-2's second-stage vision-to-language generative pre-training, which bootstraps from frozen large language models (LLMs). (**Top**) Bootstrapping a decoder-based LLM (e.g. OPT). (**Bottom**) Bootstrapping an encoder-decoder-based LLM (e.g. FlanT5). The fully-connected layer adapts from the output dimension of the Q-Former to the input dimension of the chosen LLM.

- **The modified BLIP2 model architecture and training approach:**

  - **Model architecture:**
    In our modified BLIP-2 architecture for reversed visual grounding, the task involves processing an image of a scene and a mask of a specific object to generate a textual description that captures the object's structural and spatial characteristics within the scene. The architecture consists of several key modifications to the original BLIP-2 model to adapt it for this task.

    1. **Inputs:** The model processes an image and a mask, where the image contains the full scene, and the mask highlights the object of interest.

2. **Dual Encoder Setup:** The model employs a frozen BLIP-2 image encoder to process both the image and the mask independently. Corresponding frozen layer normalization layers are applied to each input, ensuring consistency in the visual features extracted.
3. **Independent Q-Formers for Image and Mask:** Two Q-Formers are used — one for the image and one for the mask. These Q-Formers extract relevant information from the visual inputs using trainable queries. Initially, both Q-Formers use queries from the pretrained original BLIP-2 model, but the queries are later adapted to the specific task.
4. **Cross-Attention Mechanism:** The output queries from the image and mask Q-Formers are combined using a trainable cross-attention module. This step helps capture relational and spatial attributes between the object in the mask and the rest of the scene in the image.
5. **T5 for Text Generation:** The integrated queries from the cross-attention module are passed through a T5 projection layer to align the output with the input space of the T5 model. The frozen T5 model is responsible for generating textual descriptions, with no prompts provided. This setup allows the model to generate the desired text output directly based on the visual inputs.



- **Model workflow:**
  The image and mask are first processed through independent frozen image encoders and layer normalisation layers to extract visual features. These features are then passed through two independent Q-Formers (one for the image and one for the mask), generating queries that represent key information from both the image and the mask. A cross-attention module is then used to combine these queries, enabling the model to capture the spatial relationships between the object in the mask and the rest of the scene. The combined queries are projected into the input space of the T5 model using a linear projection layer, after which the T5 model generates a textual description of the masked object, without requiring any prompt.

- **Three-stage training approach:**
  - **Step 1 – Initialisations from the pretrained original BLIP2 model:**
    In this stage, the pretrained image vision-encoder, mask vision-encoder, and Q-Formers (for both the image and mask) are frozen, while the cross-attention module and T5 projection layer are trained. The T5 model remains frozen during this phase, and the focus is on establishing effective cross-modal interactions between the image and mask queries.

- **Step 2 – Additionally fine-tuning the Mask Q-Former:**
  Building on the model trained in Step 1, the mask Q-Former and the corresponding queries are made trainable, allowing it to fine-tune its ability to extract features related to the masked object's structural and spatial properties, such as position, shape, and size. The system learns to adapt the mask queries to the reversed visual grounding task.
- **Step 3 – Additionally fine-tuning the Image Q-Former:**
  In this final stage, the image Q-Former and the corresponding queries are also made trainable, allowing the entire model to be fine-tuned for optimal performance. The interaction between the image and mask features is further refined, ensuring the model generates highly accurate textual descriptions that capture the object's characteristics in the scene.

Due to resource constraints we have only been able to train our modified BLIP2 model on less data and with batch_size=1. In the experiments we conducted, the variation of training and validation loss values are quite satisfactory for the modified BLIP2 model, unlike the modified Generative Image to Text model that was developed earlier.

Future work

- Regarding the finishing touches to the work on the Reversed Visual Grounding task, further training and evaluation experiments using different evaluation metrics should be done to understand the full potential of the modified BLIP2 model that was developed.
- The remaining two sub-problems — **Curation of a task to perform and verbal communication** and **Analysis of task performed by subject** need to be worked on.
- Integration of all the modules and the incorporation of an avatar needs to be done to come up with the final end-to-end solution for ASD detection.