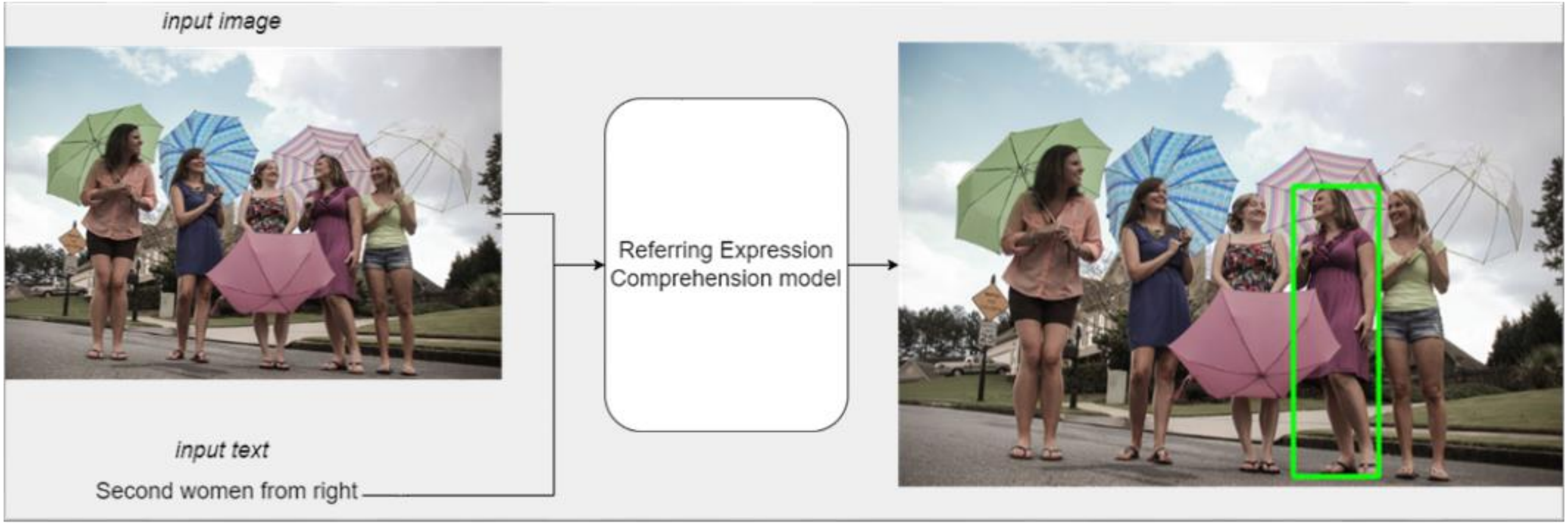
Referring Expression Comprehension with Audio Query

Muktan Patel

Sharveel Acharya

Nithin Pingili



Introduction

- Given an image and a set of text captions, localizing a target object in the image described by the referring expression phrased in natural language is called Referring Expression Comprehension (REC)

Motivation

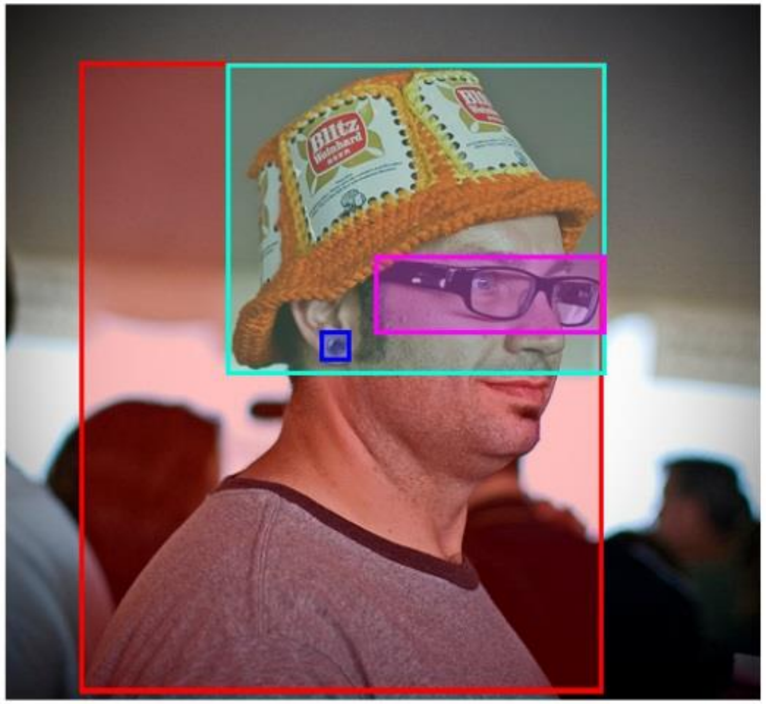
- In everyday life, language and vision are inextricably linked, we frequently employ references in our social and professional interactions, for example:
  - "please hand me the blue shirt on the table"
  - "the second to last apple looks rotten"
  - "that man who is bald and is wearing blue suit is our manager"
  - "pass me the yellow plate"
  - "the man holding the football is the captain"
  - "I guess the baby elephant on right is wounded"
  - "give soda can to the person sitting on the right side"
- Solving REC will also help creating new generation of artificially intelligent robots/systems

Object detection V/S REC

- Unlike object detection, REC can refer to objects that is not predetermined during the training using natural language expressions (NLE).
  - For example, "pass me the coke placed on left chair". If we consider object detection and coke is not one of the class that the model is trained on then it is not possible to predict it. While in REC if it has not seen coke earlier in the dataset, then it can use other information available in the query like "placed on left chair"

Visual Grounding V/S REC

- **Visual grounding** is the process of **identifying several object regions** in an image **that correlate to multiple noun phrases** from a sentence that describes the scene.
- The purpose of **REC** is to **locate** the **best matching region** for the given natural language expression.
- REC can be considered as the specialized version of Visual Grounding

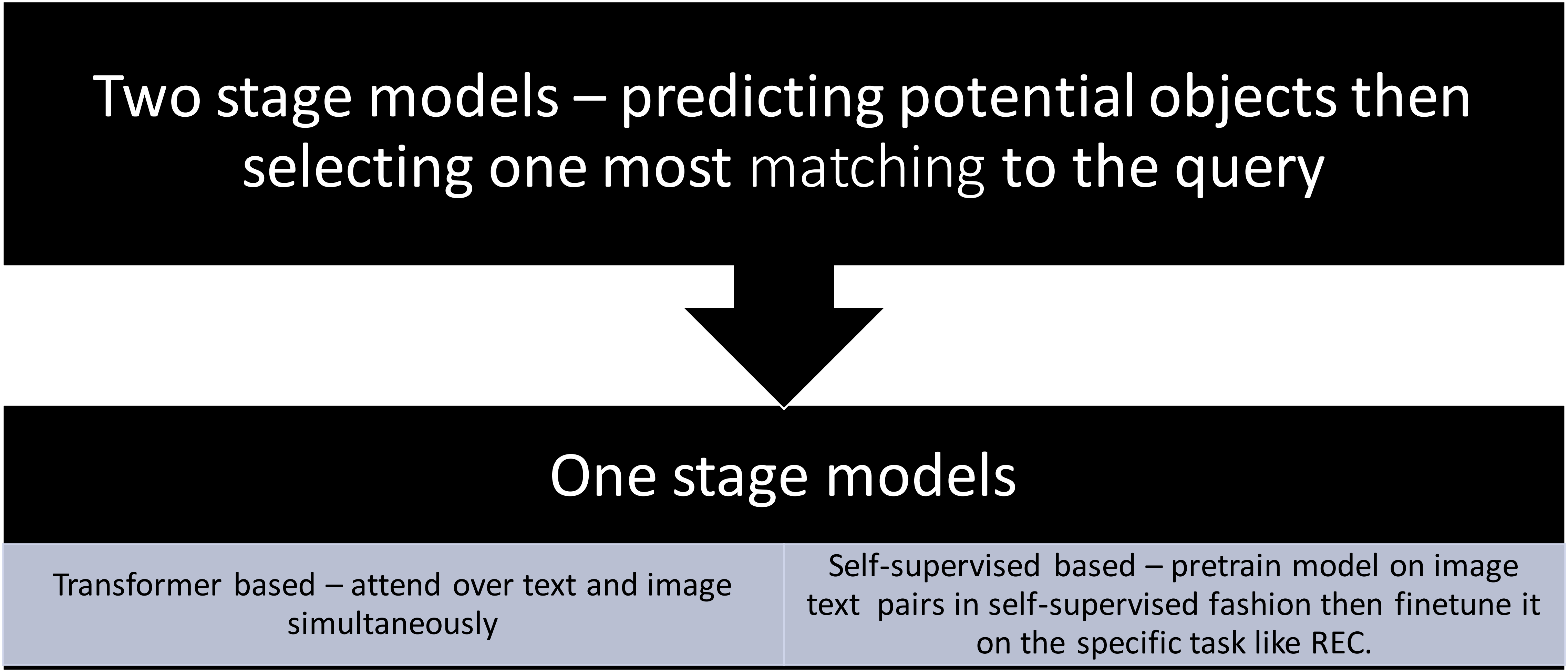


A man with **pierced ears** is wearing **glasses** and **an orange hat**.
A man with **glasses** is wearing **a beer can crotched hat**.
A man with **gauges** and **glasses** is wearing **a Blitz hat**.
A man in **an orange hat** starring at **something**.
A man wears **an orange hat** and **glasses**.

Datasets

- RefCOCO consists of 142,209 refer expressions for 50,000 objects in 19,994 images and split into:
  - train set with 120,624 expressions
  - validation set with 10,834 expressions
  - testA set with 5,657 expressions
  - testB set with 5,095 expressions
- The dataset provides the queries in text format only.
- Used Facebook’s FastSpeech 2 to generate the synthetic audio dataset for the RefCOCO dataset.

Past work at glance



Two stage models – predicting potential objects then selecting one most matching to the query

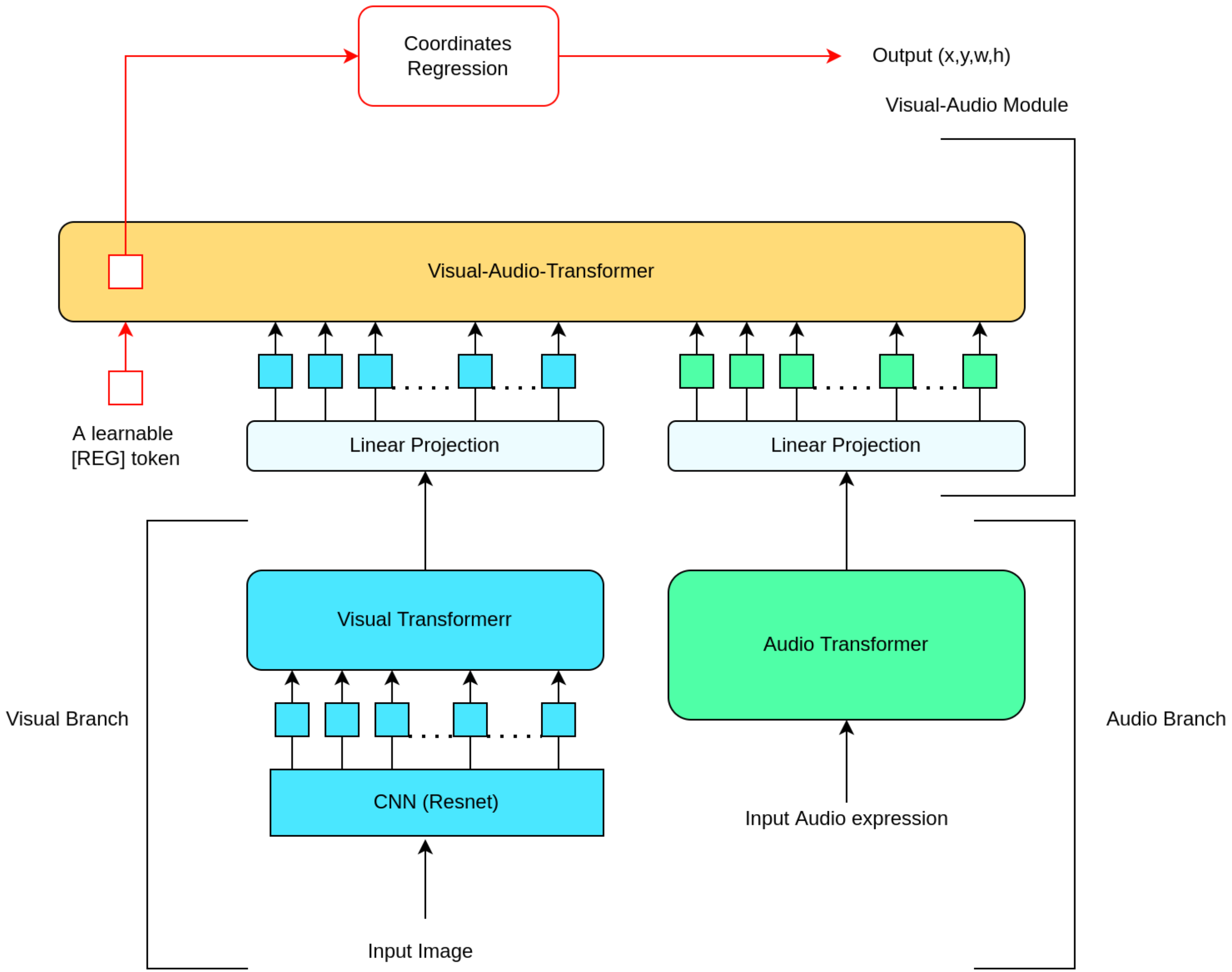
One stage models

Transformer based – attend over text and image simultaneously

Self-supervised based – pretrain model on image text pairs in self-supervised fashion then finetune it on the specific task like REC.

Our Work

- We have altered the TransVG architecture to allow audio queries for REC task.
- In TransVG, there are four basic components:
  - visual branch
  - linguistic branch
  - visual-linguistic fusion module
  - prediction head
- we swap the linguistic branch with audio branch



Deng, J., Yang, Z., Chen, T., Zhou, W., & Li, H. (2021). TransVG: End-to-End Visual Grounding with Transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1749-1759.

Experimental setup

- Accelerator (GPU) - 2 24GB A5000
- Max epochs – 85
- Batch size – 12
- Learning rate - 0.0001

Evaluation criteria

- The evaluation is performed by calculating the Intersection over Union (IoU) ratio between the true bounding box and the top predicted box for a referring expression
- If the IoU is larger than 0.5, the prediction is considered a true positive. Otherwise, we count it as a false positive. Finally, we calculate accuracy.
- The scores are then averaged over all referring expressions.

$$IOU = \frac{Area\,of\,\,overlap}{Area\,of\,\,union}$$

Comparing audio and textual models

- Audio query based model accuracy – 34%
- Text query based model accuracy – 73.7%
- *Note: If we consider the ASR module used to convert audio to text in text query based model then there can be error* propagation *of 10% so resulting REC accuracy can be around 60 – 65%*

- Audio query based model inference time – 6.5 sec
- Text query based model inference time – 4.7 sec

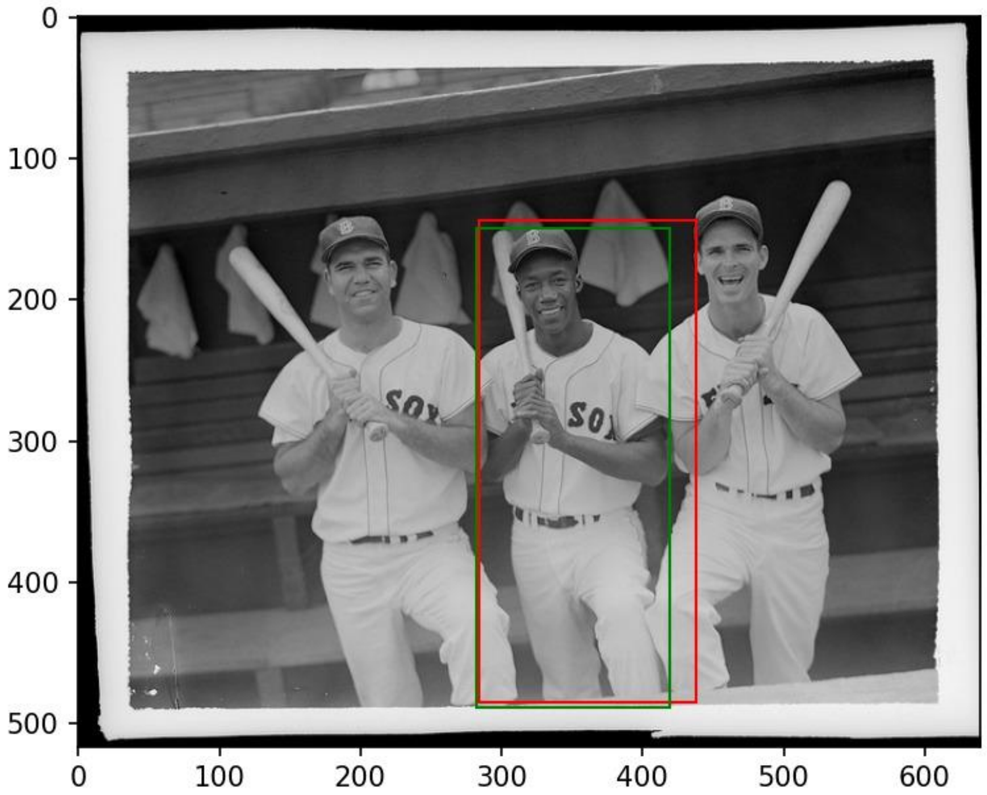
Note: If we consider the ASR module used to convert audio to text in text query based model then its latency will be added to the above time and will result around 6 sec for text query based systems too

Observations

- The audio model was underfitted, as it showed poor performance on both training and testing data. We can do following things to improve it
  - Create larger data. The text does not vary as speech does, for example the same query can be spoken in various accent, speed and audio intensity. So, when compared to text based queries we would require larger dataset.

Sample Results from our model (Images)

Correct prediction

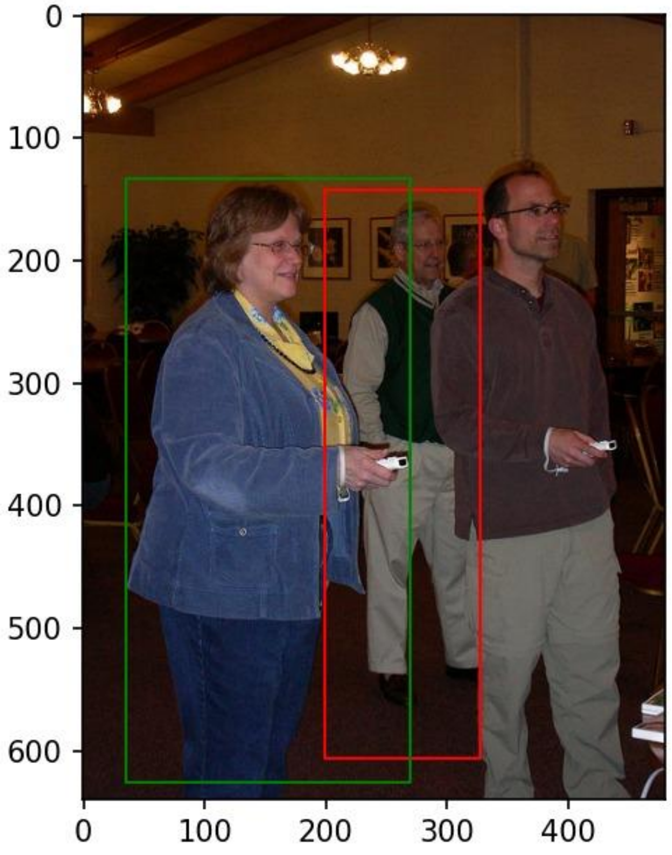


The man in the middle

Sample Results from our model (Images)

Incorrect Prediction

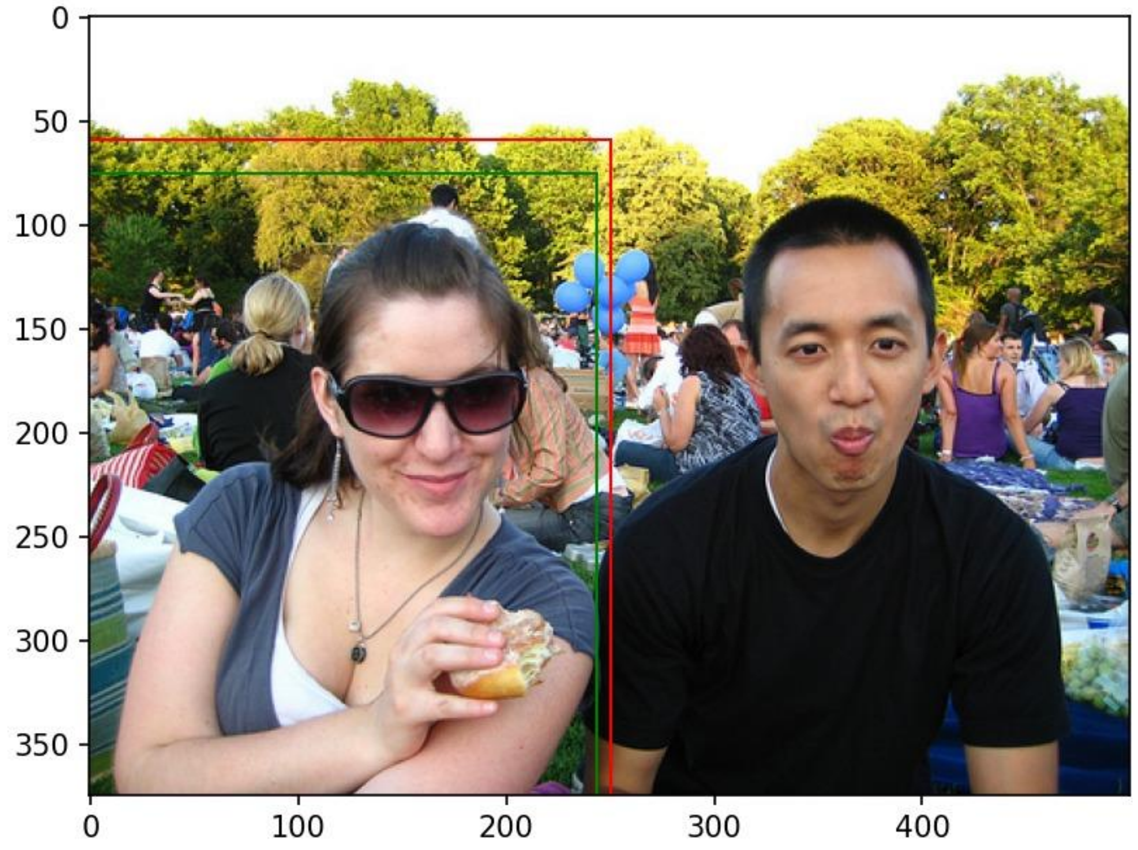
Observation: It at least predicted an object and didn't predict some random 4 points



Woman standing on left

Sample Results from our model (Images)

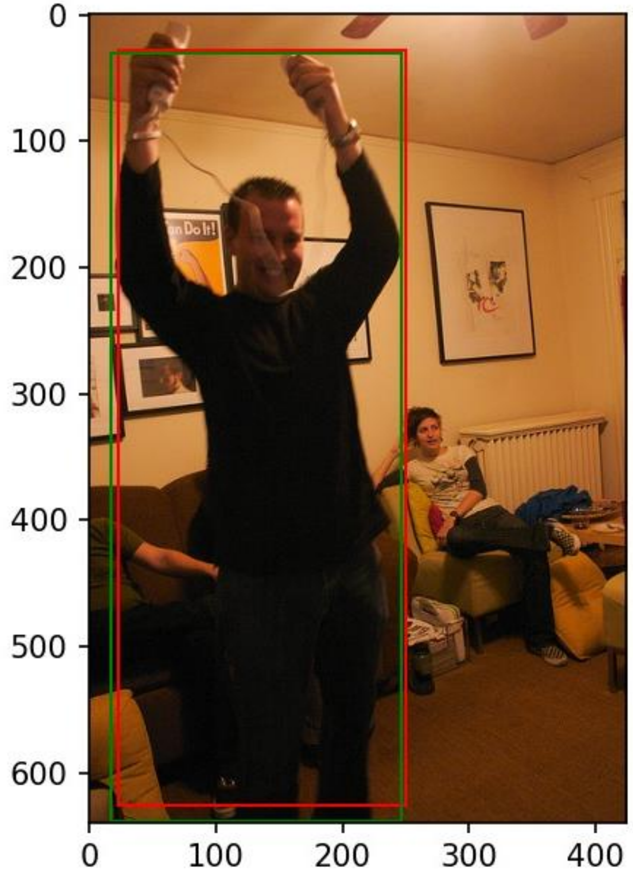
Correct Prediction



Woman in left with food

Sample Results from our model (Images)

Correct Prediction

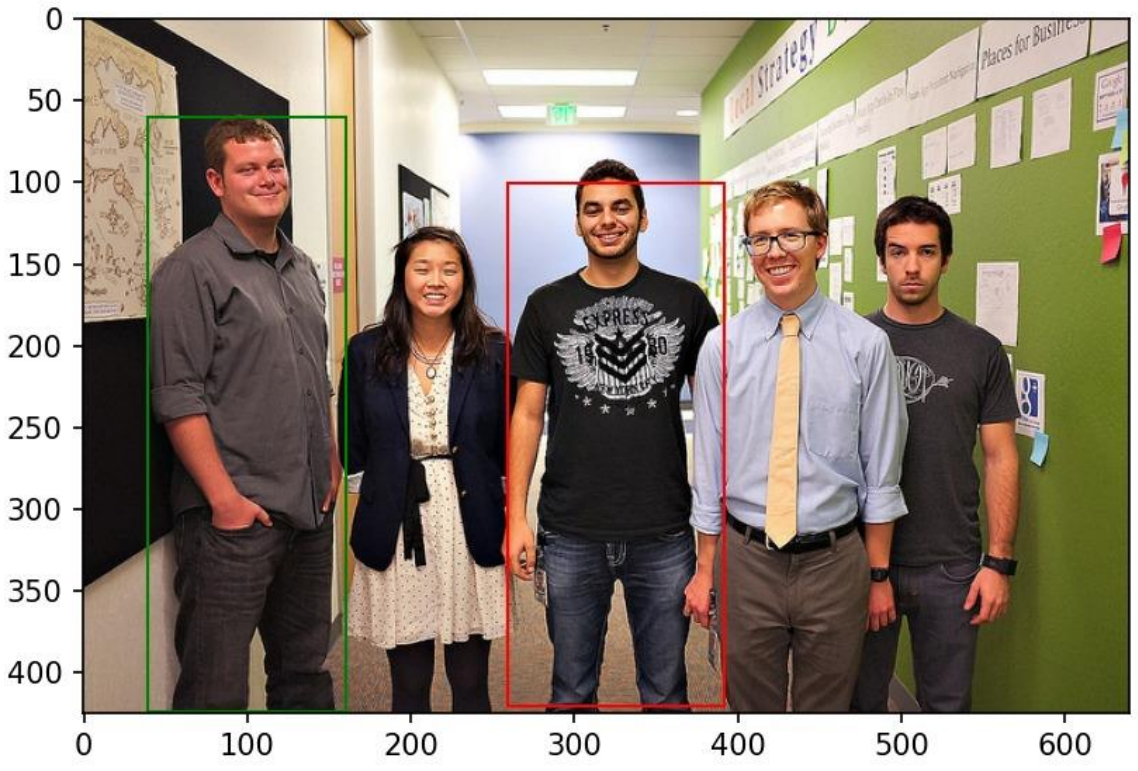


Man in full black standing

Sample Results from our model (Images)

Incorrect Prediction

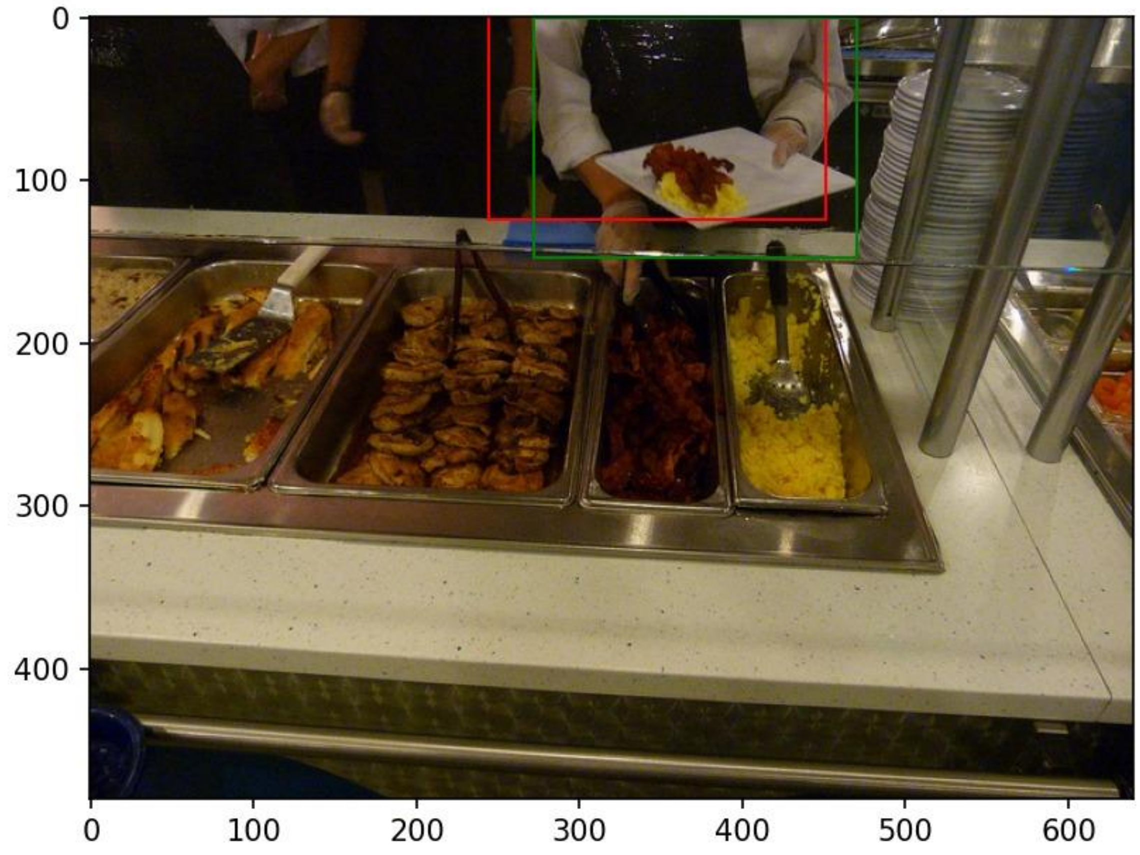
Observation: Even though such queries are often seen in training data still it made such mistake. But it at least predicted an object and didn't predict some random 4 points



leftmost person

Sample Results from our model (Images)

Correct Prediction

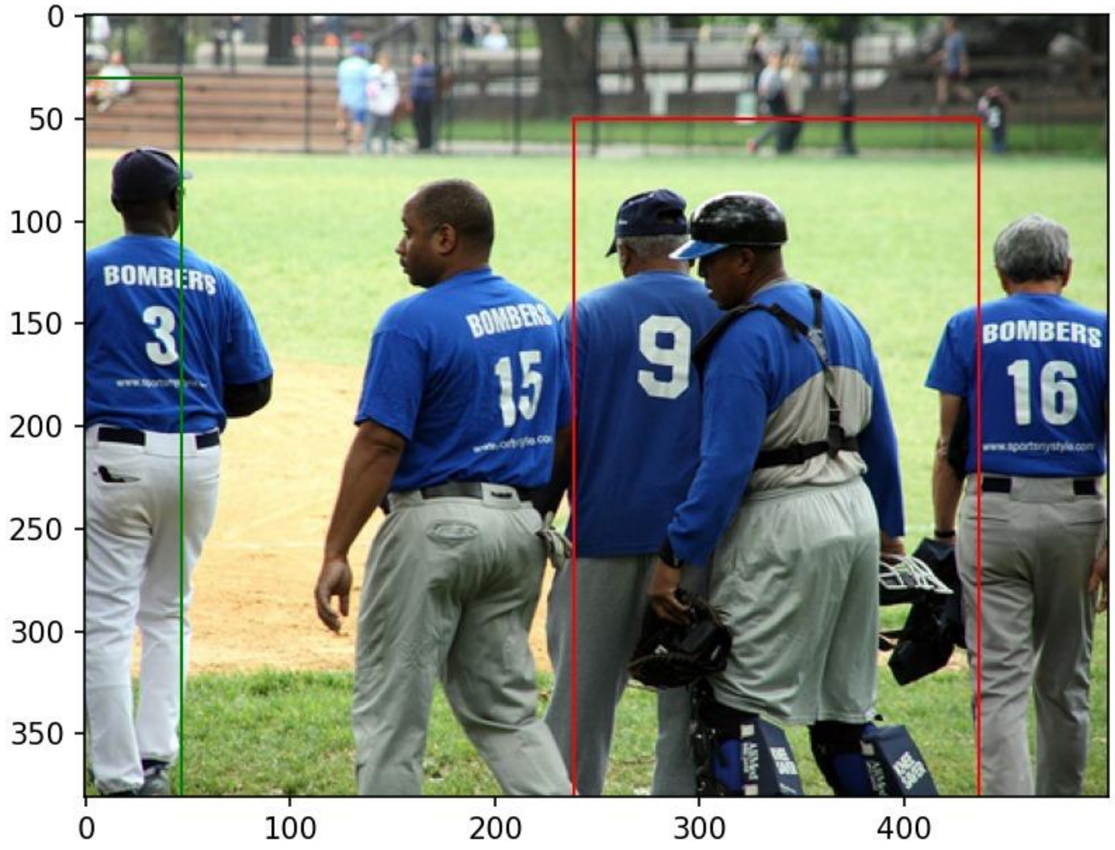


Person taking food

Sample Results from our model (Images)

Incorrect Prediction

Observation: this was a tough query

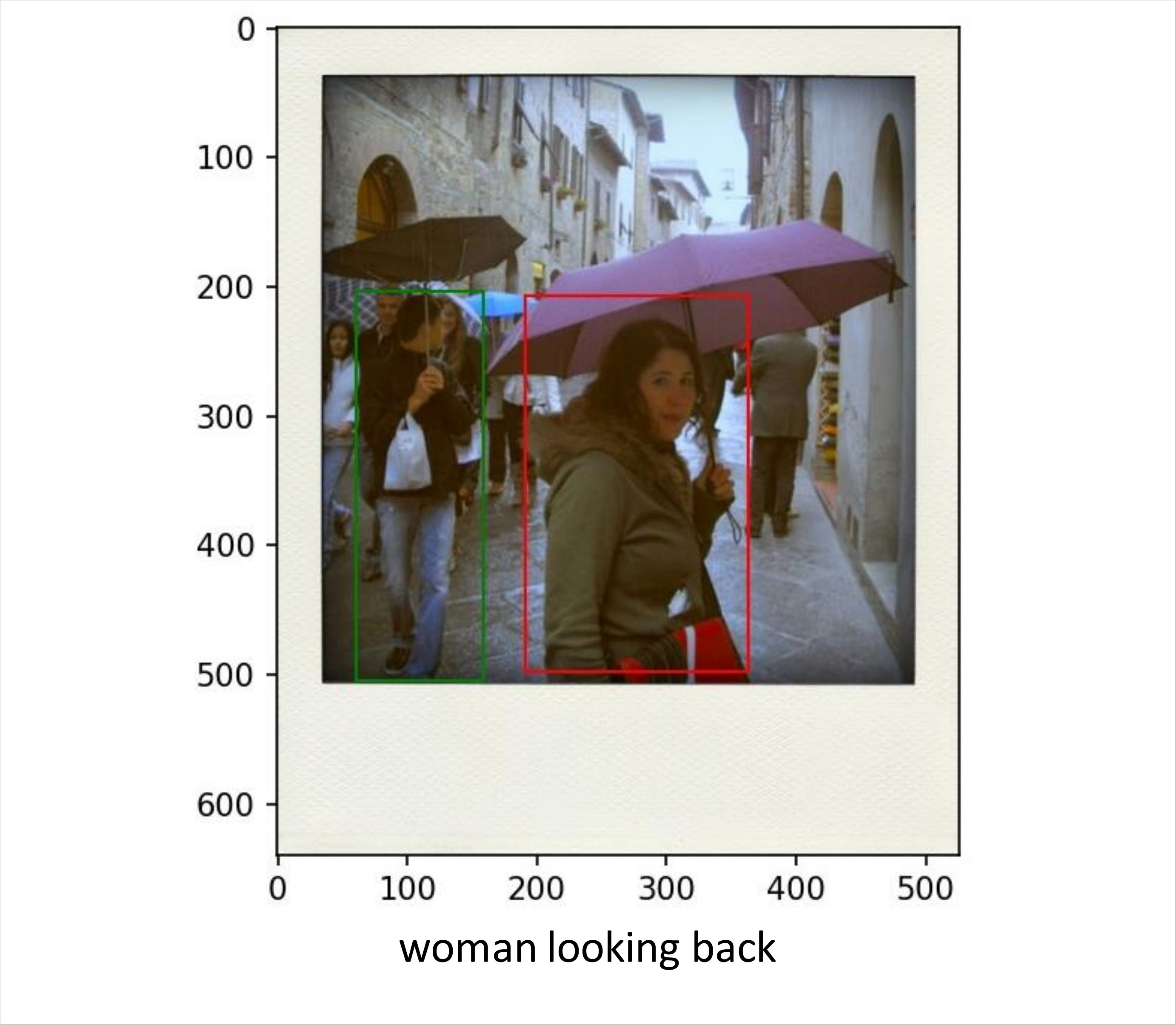


"3"

Sample Results from our model (Images)

Incorrect Prediction

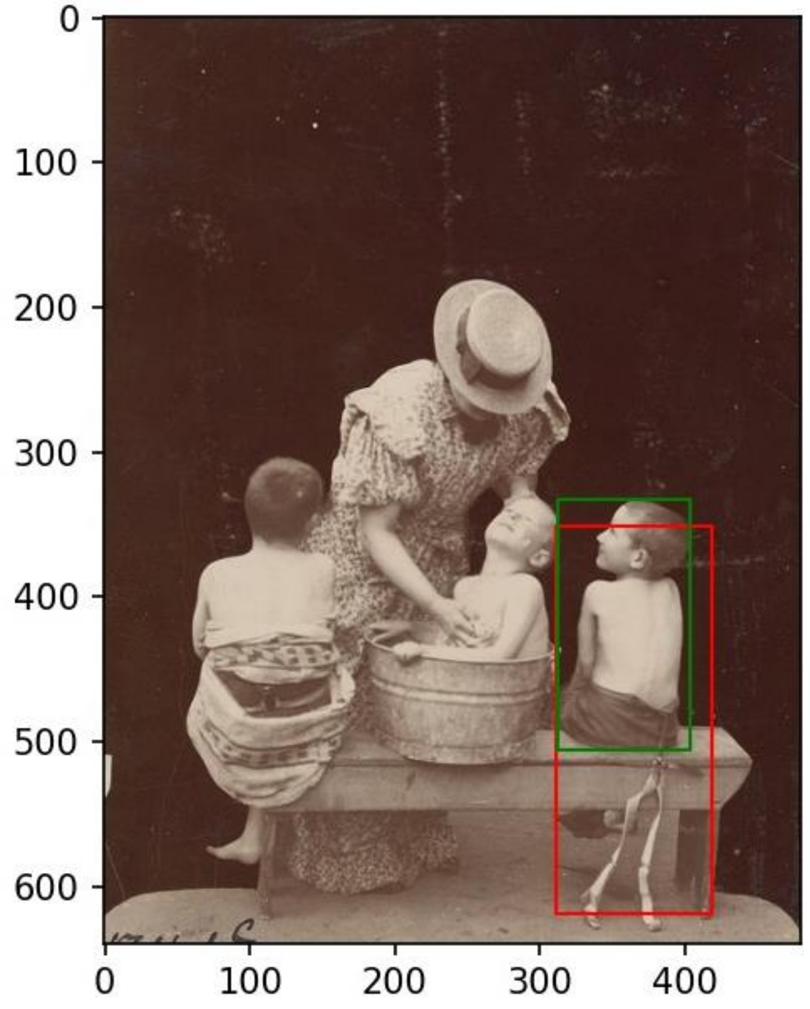
Observation: It got confused with another person who was looking back (not towards camera)



Sample Results from our model (Images)

Correct Prediction

Observation: When predicting the kid, it bounded his clothes

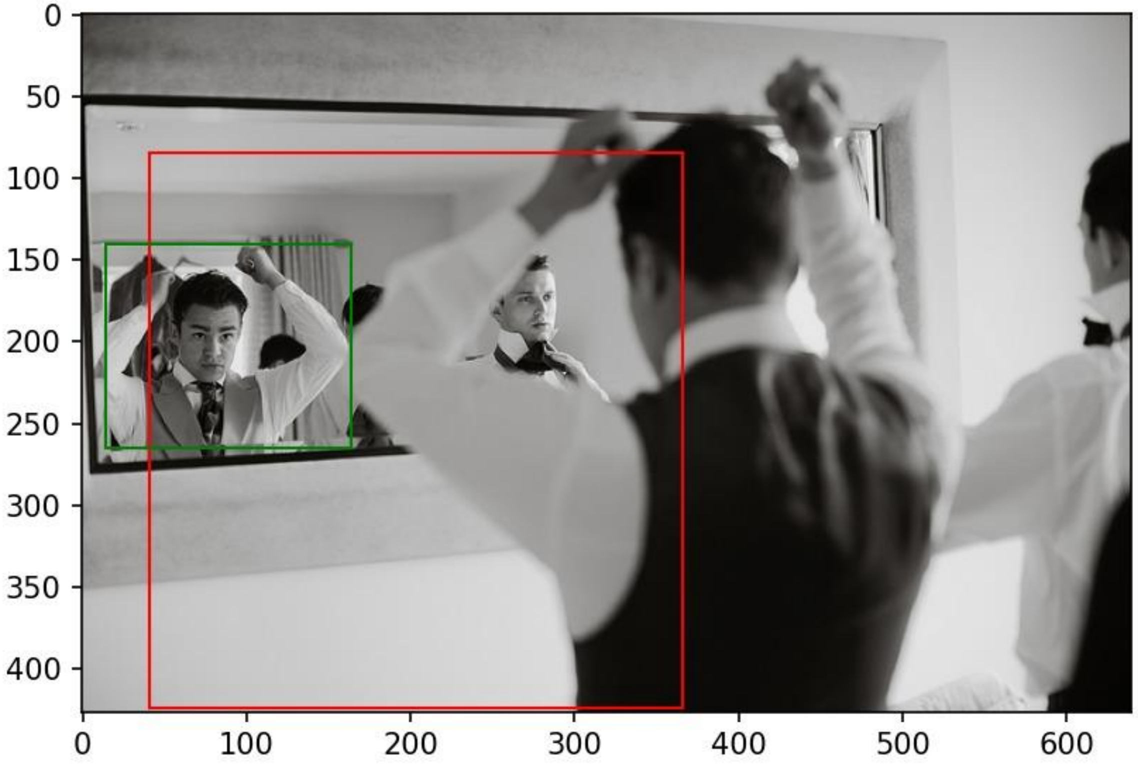


kid sitting on right

Sample Results from our model (Images)

Incorrect Prediction

Observation: It also covered the reflection of the person on right. However, it bounded some of the regions out of the mirror too. Trying to make ambiguous predictions

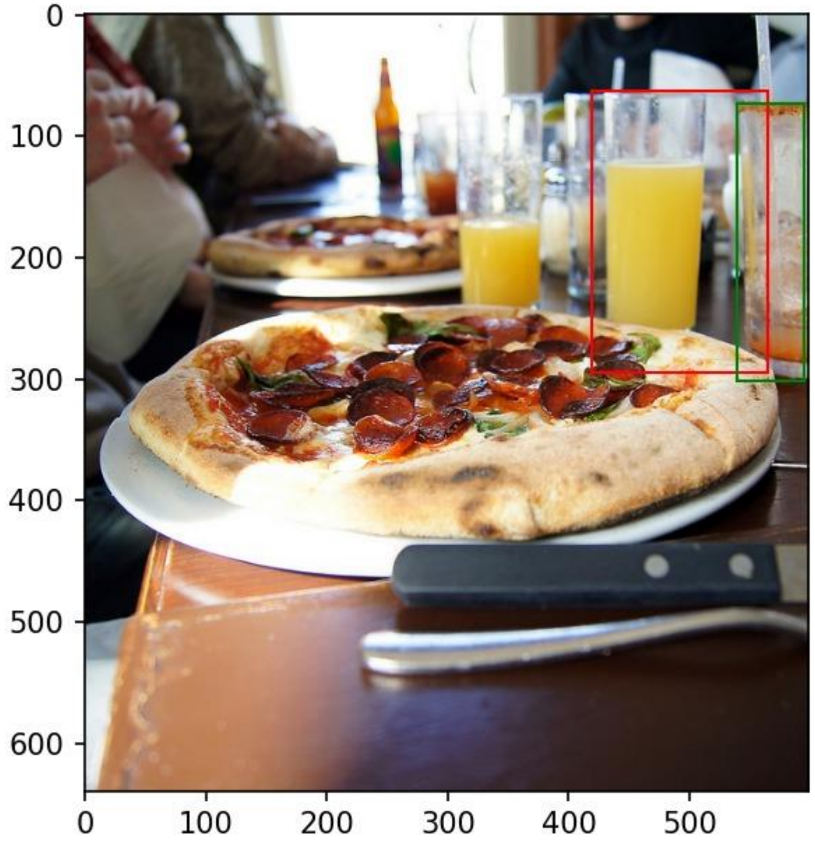


reflection of person on left

Sample Results from our model (Images)

Incorrect Prediction

Observation: It seems like it was not sure about the rightmost glass so it covered it half including second rightmost glass. Trying to make ambiguous predictions



Rightmost glass

Future Work

Creating

Creating larger dataset to bring more variety of speaker audio

Performing

Performing self-supervised learning on Image-audio dataset

Thank You everyone and Special thanks to Professor for providing GPU for experimentation

Any questions?