Visual Question Answering Using CLIP with Local Feature Enhancement

Sai Surya Vidul Chinthamaneni - 862466578
Yash Sanjay Kathe - 862464930
Tushar Baid - 862469499

Introduction

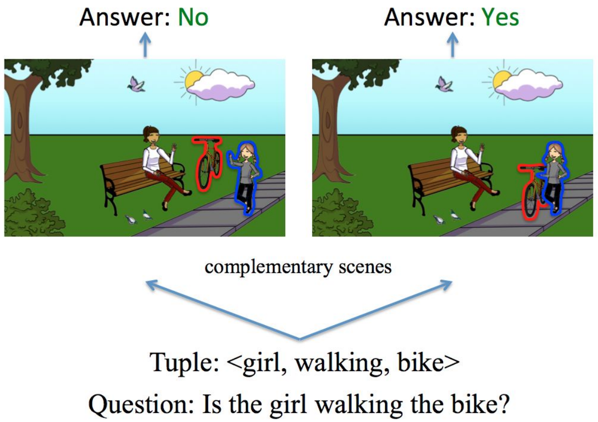
- What is VQA?
- Introduction to CLIP Model
- Project Objective

WHAT IS VQA (Visual Question Answering) ?

- VQA involves answering questions about the content within an image using both visual and textual understanding
- Models need to integrate object recognition, spatial reasoning, and language comprehension

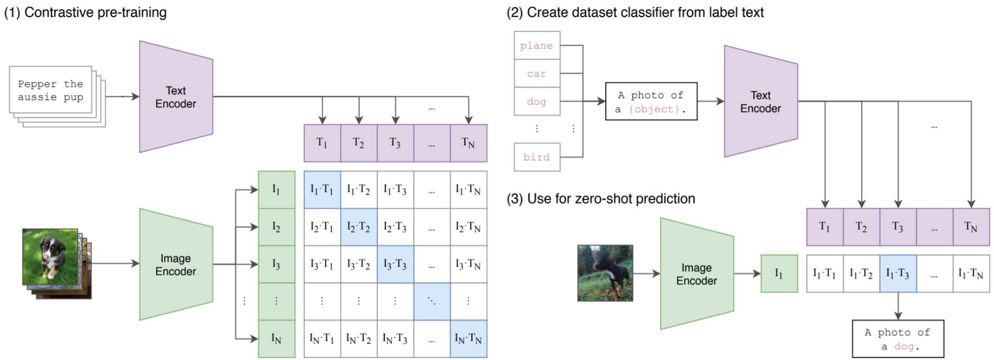
CHALLENGES IN VQA

- Recognizing fine-grained details like counting objects or identifying small elements.
- Bridging global context (whole image) and local details (specific regions).



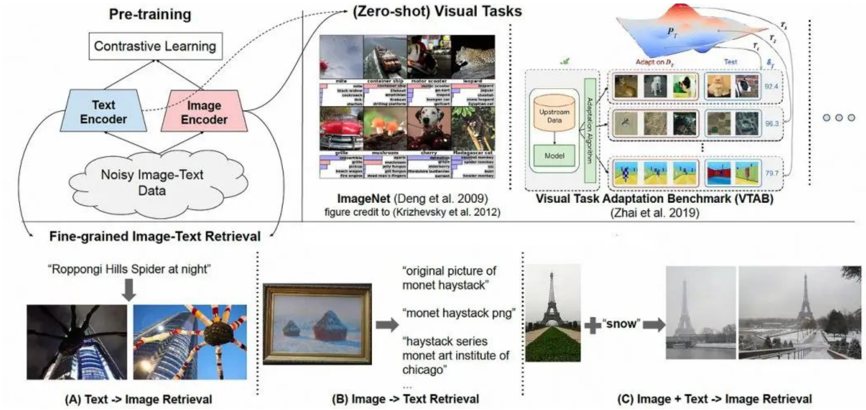
WHAT IS CLIP?

- *Contrastive Language-Image Pretraining* (CLIP) aligns images and text in a shared embedding space
- Requires large-scale image-text pairs for training



LIMITATIONS OF CLIP

- Relies heavily on global feature extraction and misses fine-grained local details
- Struggles with tasks requiring precise recognition and spatial reasoning, like Visual Question Answering (VQA)



OBJECTIVE

To Enhance CLIP’s architecture with focus on local features for better VQA performance.

APPROACH

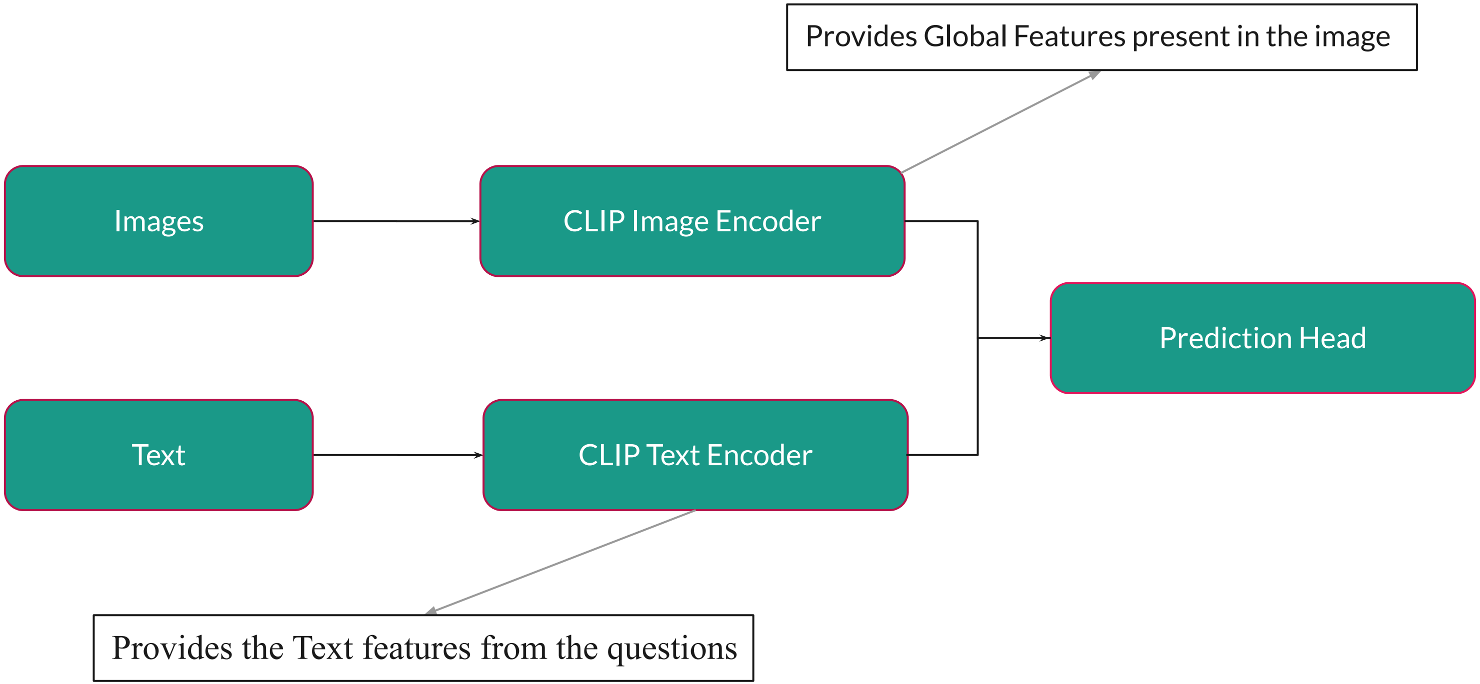
Dual Encoder Architecture

1. CLIP + VIT with *Concatenation based Fusion Layer*
2. CLIP + VIT with *Attention based Fusion Layer*

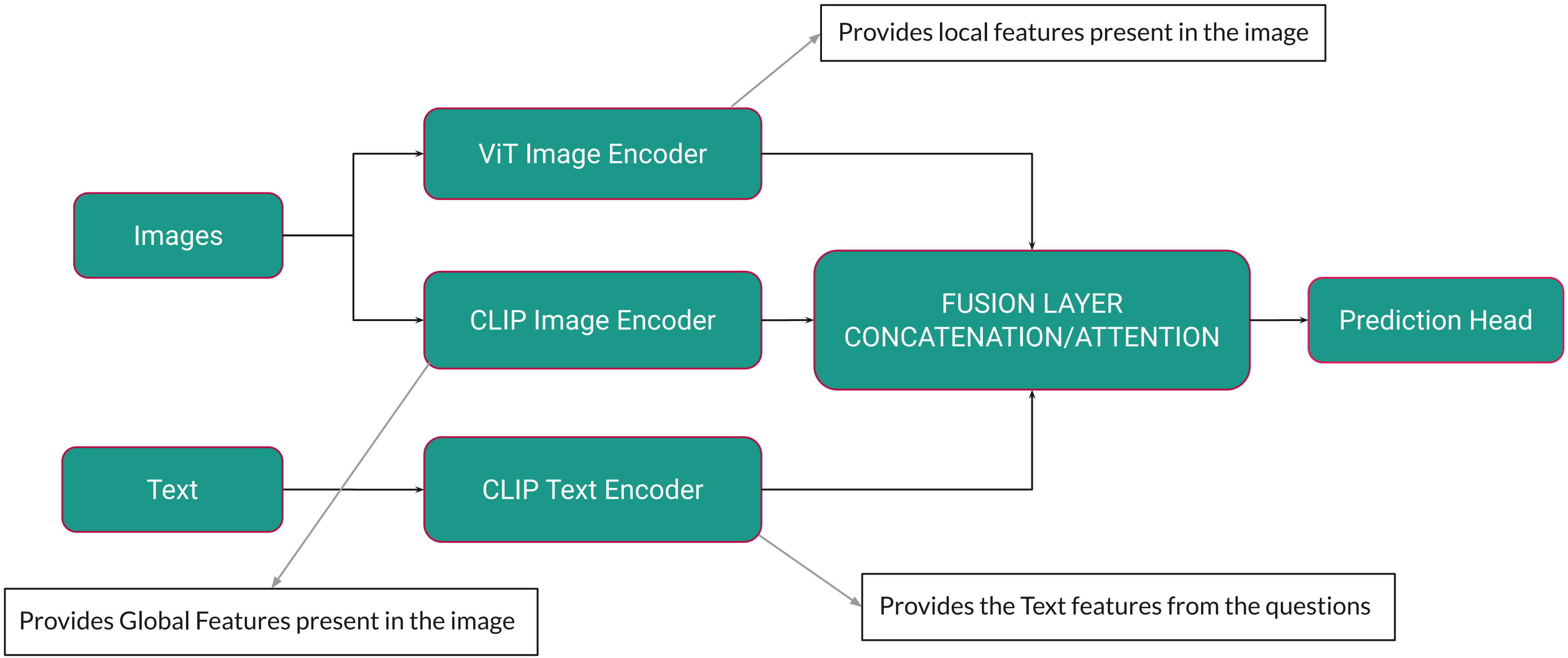
Methodology and Training

- Architecture of BaseLine Clip and Dual Encoder Approach
- Training the Models

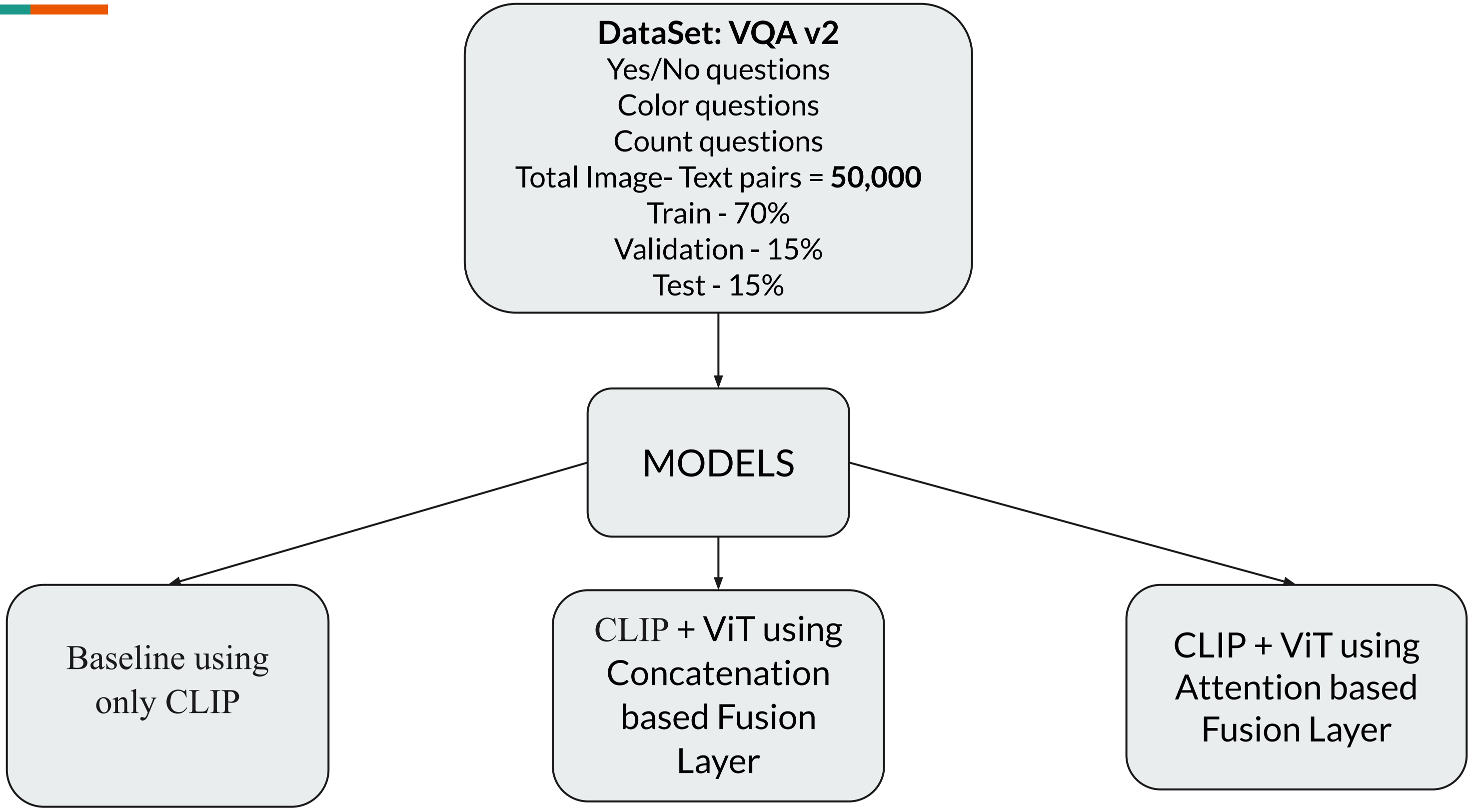
METHODOLOGY - BASELINE CLIP



METHODOLOGY - DUAL ENCODER APPROACH



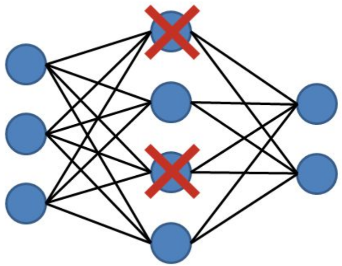
TRAINING


DataSet: VQA v2
Yes/No questions
Color questions
Count questions
Total Image- Text pairs = **50,000**
Train - 70%
Validation - 15%
Test - 15%
MODELS
Baseline using only CLIP
CLIP + ViT using Concatenation based Fusion Layer
CLIP + ViT using Attention based Fusion Layer

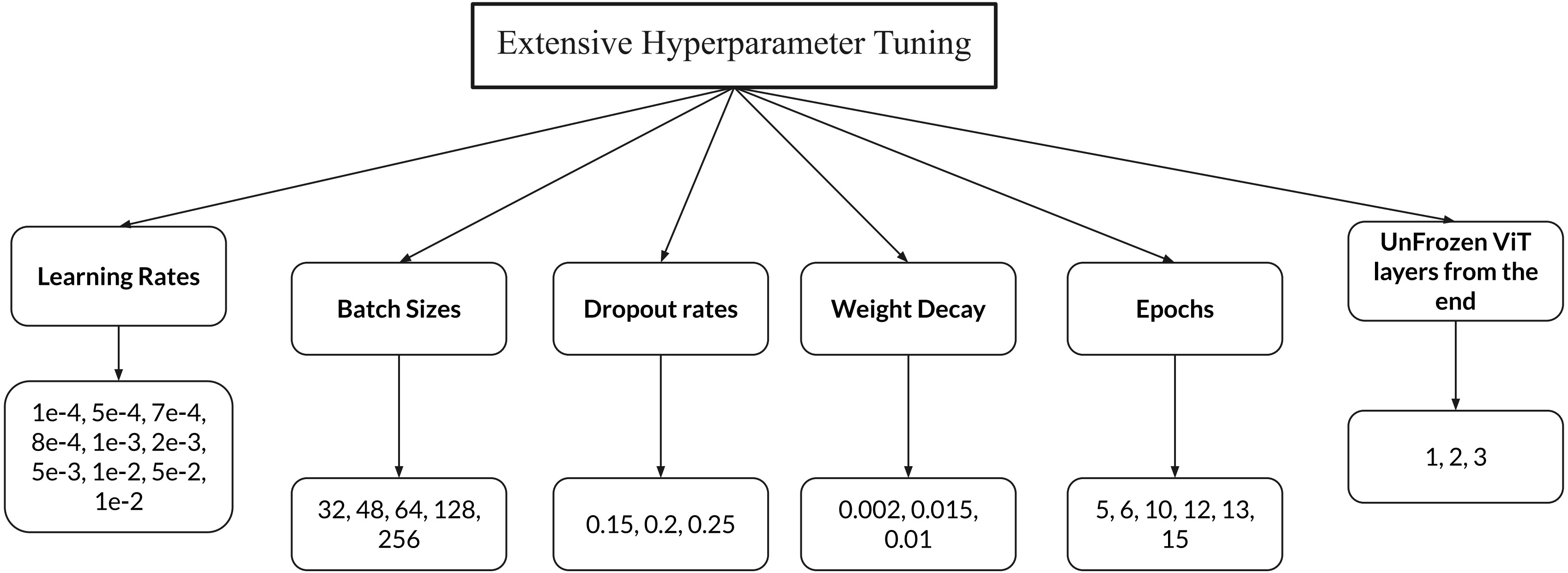
TRAINING

$$H = -\sum p(x) \log p(x)$$

- Used Cross Entropy loss and AdamW optimizer
- Implemented mixed precision training
- Dropout is used reduce overfitting.



TRAINING


Extensive Hyperparameter Tuning
Learning Rates
1e-4, 5e-4, 7e-4, 8e-4, 1e-3, 2e-3, 5e-3, 1e-2, 5e-2, 1e-2
Batch Sizes
32, 48, 64, 128, 256
Dropout rates
0.15, 0.2, 0.25
Weight Decay
0.002, 0.015, 0.01
Epochs
5, 6, 10, 12, 13, 15
UnFrozen ViT layers from the end
1, 2, 3

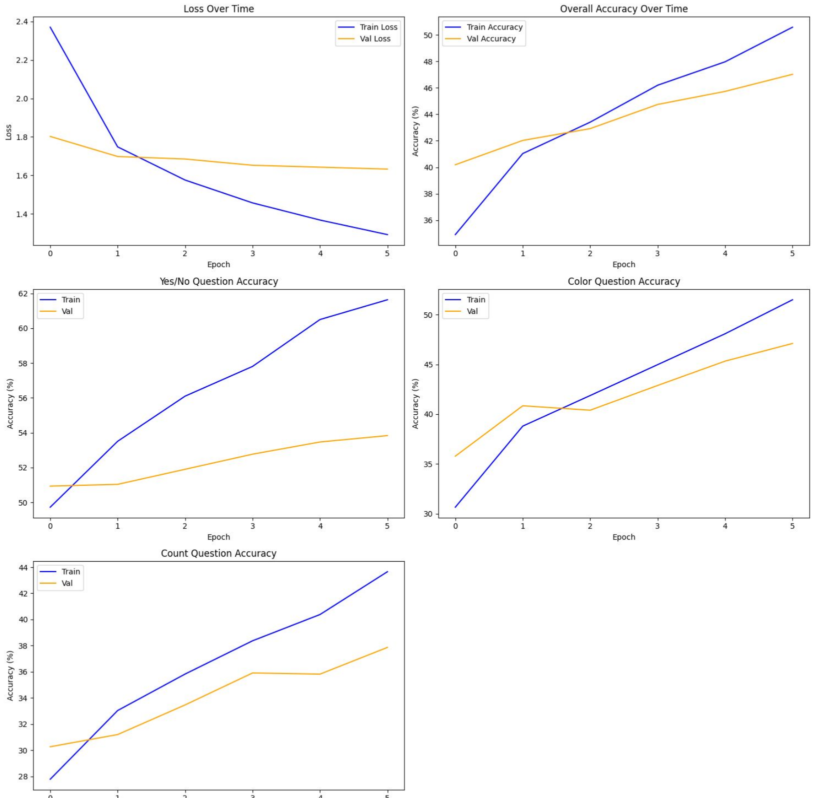
Results and Conclusion

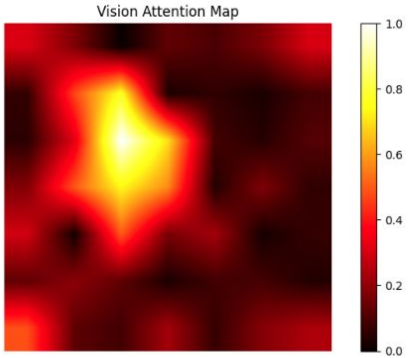
- Performance of all the Models
- Challenges Faced
- Future Scope
- Contribution

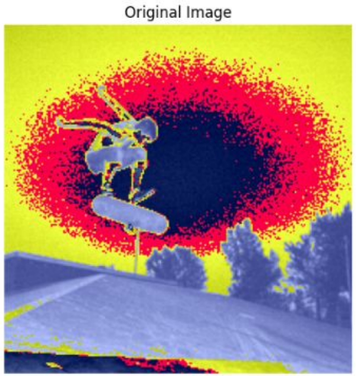
RESULTS

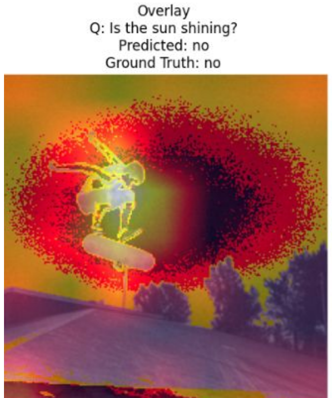
MODEL	BLEU-1	METEOR
BASELINE CLIP	0.4672	0.2382
DUAL ENCODER + CONCATENATION	0.4634	0.2372
DUAL ENCODER + ATTENTION	0.4326	0.2214

RESULTS - BASE CLIP

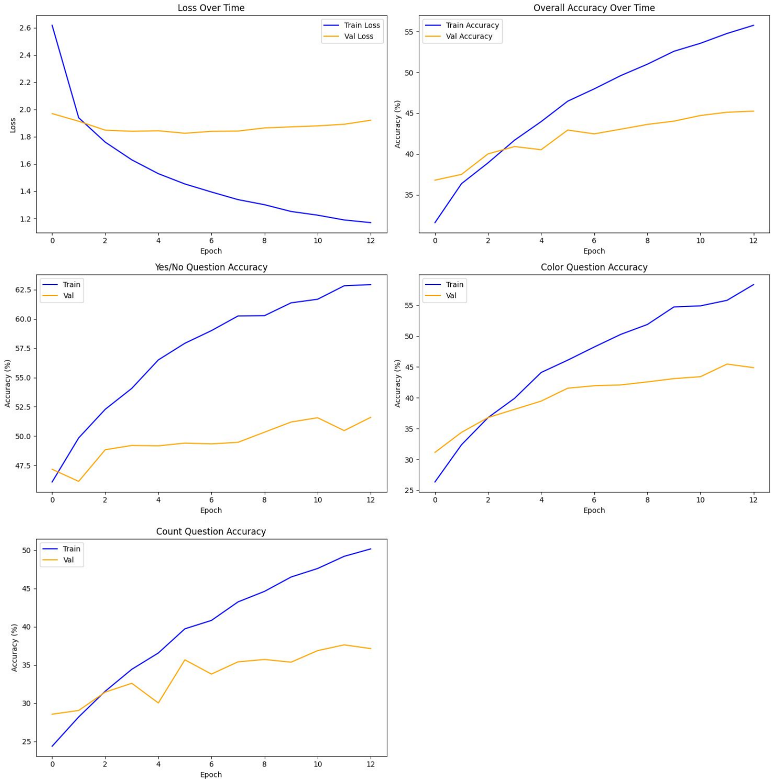


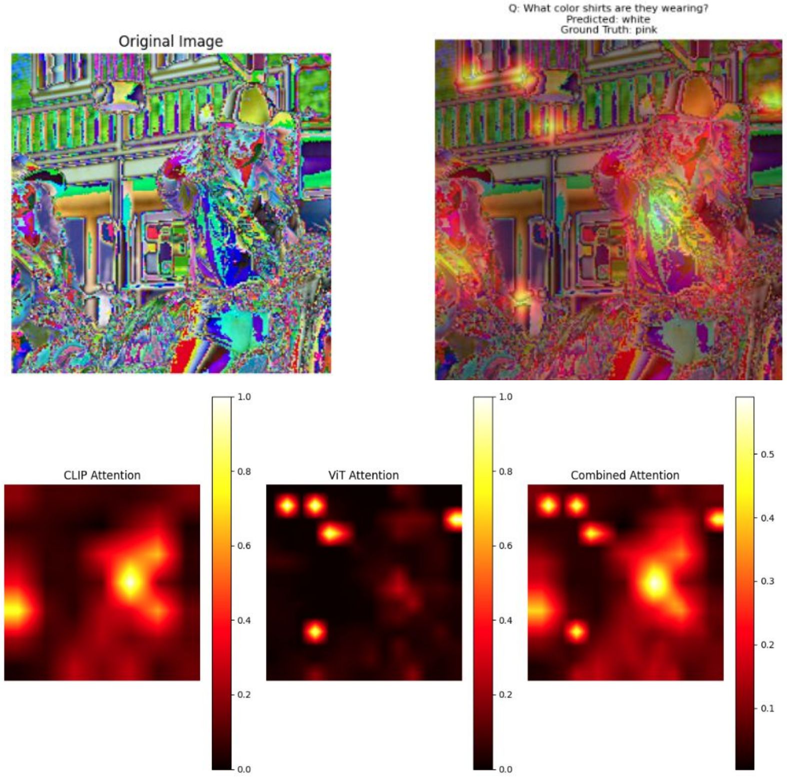




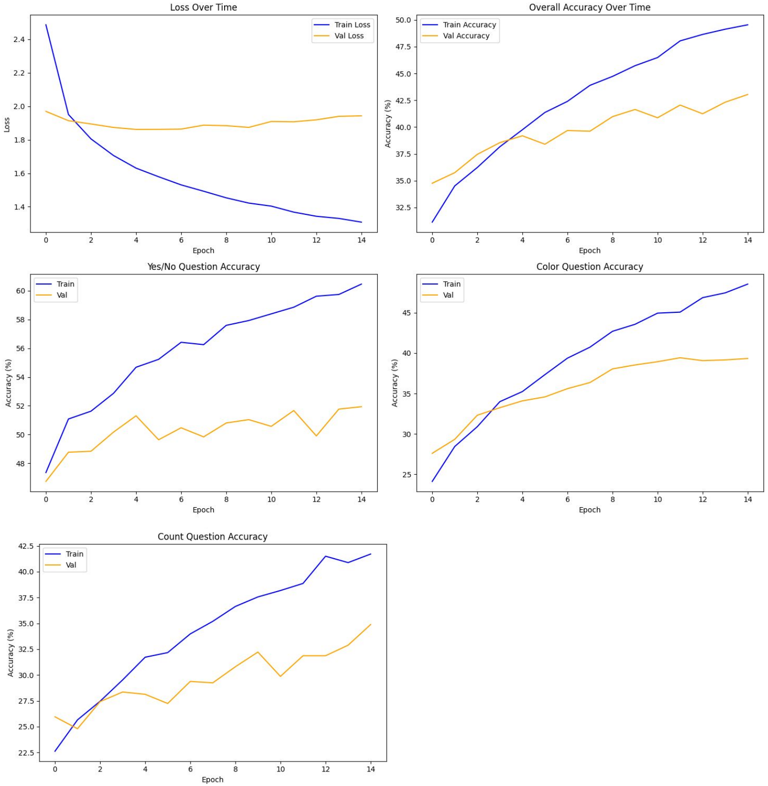


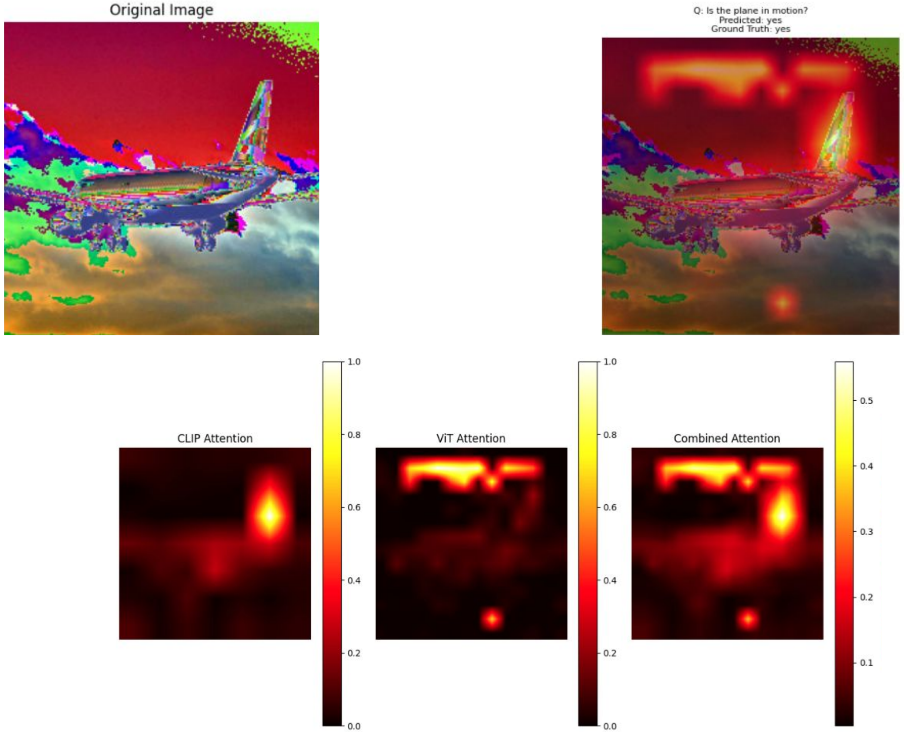
RESULTS - DUAL ENCODER WITH CONCATENATION



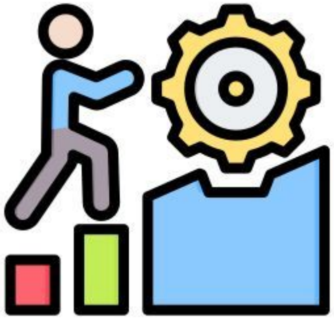


RESULTS - DUAL ENCODER WITH ATTENTION





CHALLENGES



- Computationally very intensive - ran the experiments on a mobile 4070 GPU
- Hyperparameter tuning took a lot of time
  - Base clip took about ~3 hours per run
  - Dual encoder approaches took about ~5 hours per run
- Overfitting was a major problem due to the size of the dataset and the models complexity

FUTURE SCOPE

- Can use a Cosine similarity based prediction head
- Try scheduling learning rate and weight decay
- Implement ViT with smaller patch sizes
- Implement LoRa
- Consider a bigger dataset with more complex image - text pairs to show the true power of the dual encoder approach (to capture the local features which will help with highly contextual questions as well)

Contribution

Team Member	Tasks Done
Sai Surya Vidul	Data Preprocessing - 15%, Baseline CLIP - 15%, Dual Encoder with Concatenation Fusion - 70% , Dual Encoder with Attention Fusion - 70%, Visualization and Testing - 15%, Presentation -15%
Tushar	Data Preprocessing - 70%, Baseline CLIP - 15%, Dual Encoder with Concatenation Fusion - 15% , Dual Encoder with Attention Fusion - 15%, Visualization and Testing - 15%, Presentation - 70%
Yash	Data Preprocessing - 15%, Baseline CLIP - 70%, Dual Encoder with Concatenation Fusion - 15% , Dual Encoder with Attention Fusion - 15%, Visualization and Testing - 70%, Presentation - 15%

THANK YOU!