Visual Question Answering

Revathi Vijay, Jash Kakadia

Introduction

What is Visual Question Answering?

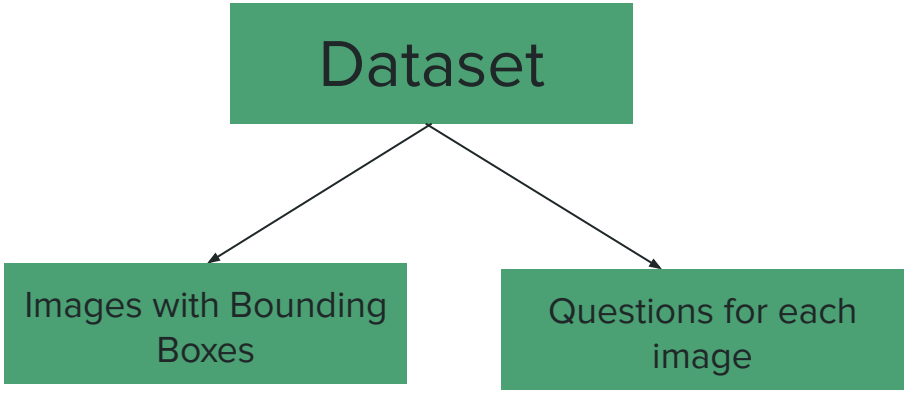
is there a banana?

yes

Image

Question

Result (Answer based on both the inputs)

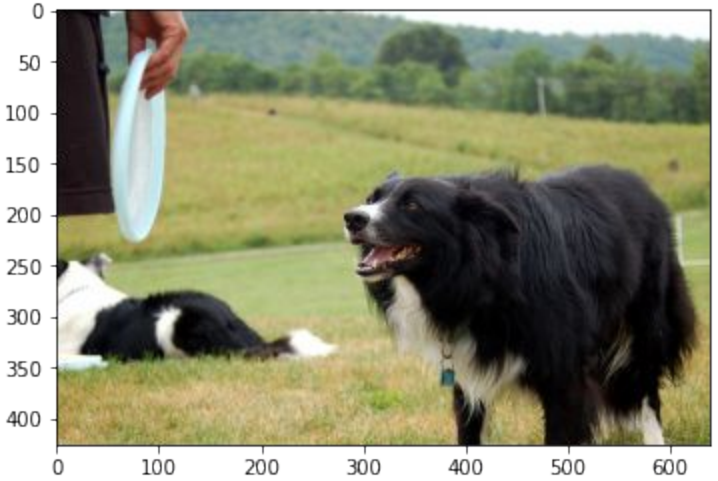


MS Coco:
Images with annotations

VQA (from their website):
≥ 3 Questions, 10 answers for every image in the MS Coco dataset.
Separated into Train, Val, Test

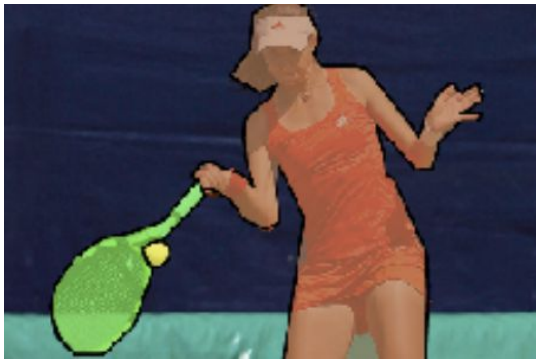
Dataset

- VQA Dataset Annotations
  - Question: ‘Is the dog looking at a tennis ball or frisbee?’
  - Question_id: 524291002
  - Answer type: ‘other’
  - Answers: [‘frisbee’]
  - Image id: 524291
  - Multiple Choice Answer: ‘frisbee’

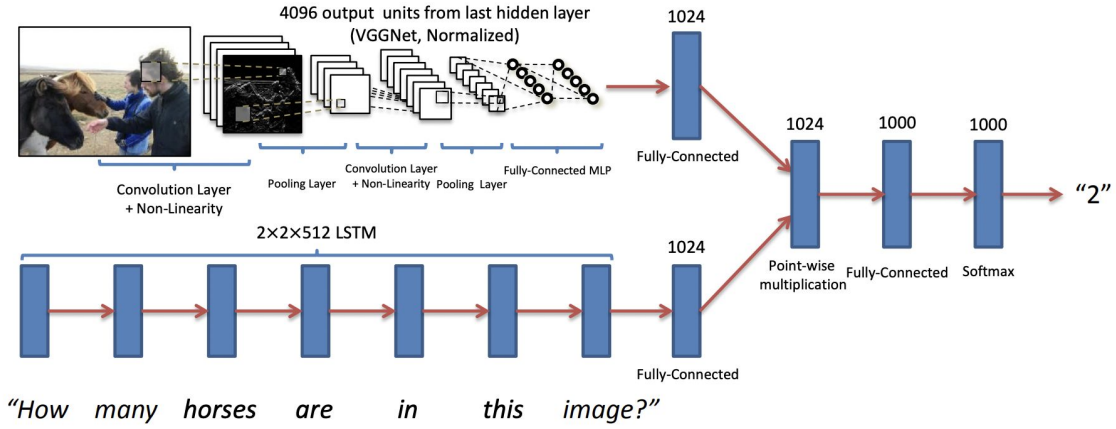


Dataset

- COCO Dataset Annotations
  - Each annotation corresponding to an image has segmentation and bounding box information.
  - Bbox: [223.38, 302.55, 139.18, 107.39]
  - Category_id: 18 ➜ corresponds to frisbee
  - Multiple annotations can be given to a given image
  - Only training data had bounding boxes



Baseline Method: LSTM + CNN



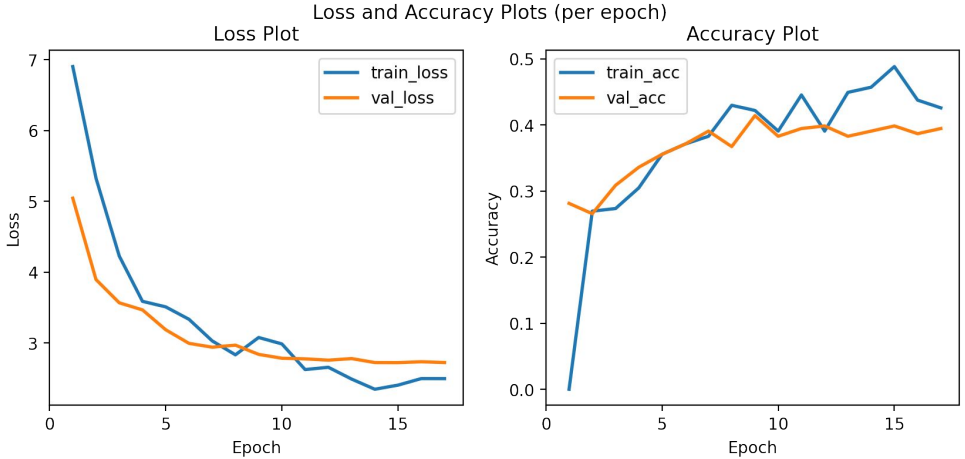
*The CNN used was a pretrained VGG-19 model.

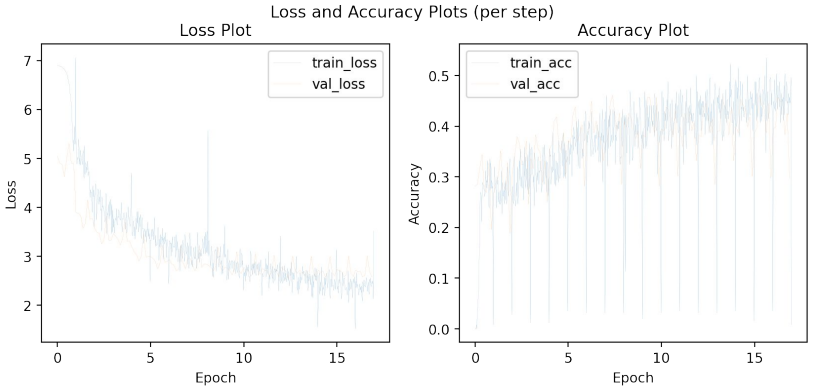
What this does:

1. Convert the questions into embeddings
2. Resize the image and pass it through the convolutional network
3. Flatten both the outputs and pointwise multiplication of both
4. Pass through one more FC layer and softmax to give the output

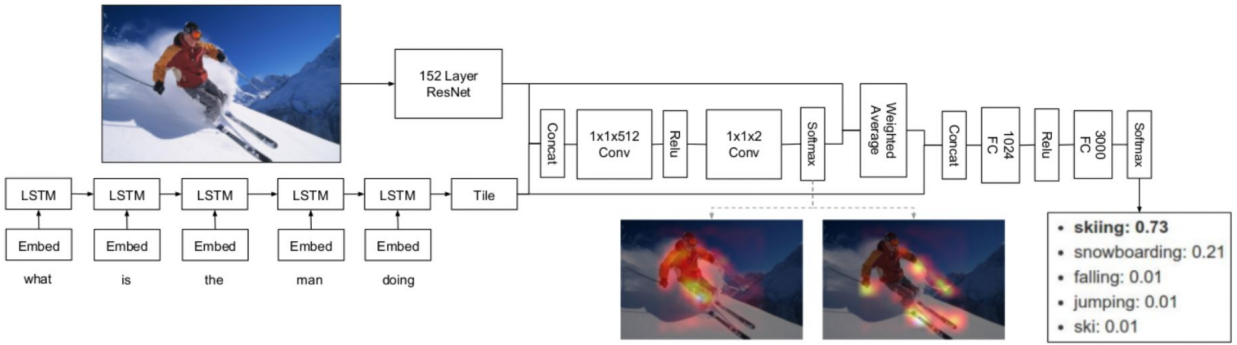
*Image & model from paper: https://arxiv.org/pdf/1505.00468.pdf

Results: Base LSTM + CNN Model





Baseline Methods: Attention Model

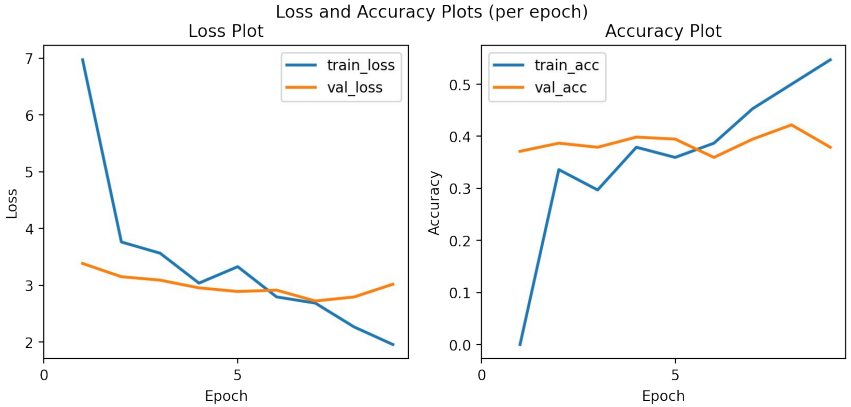


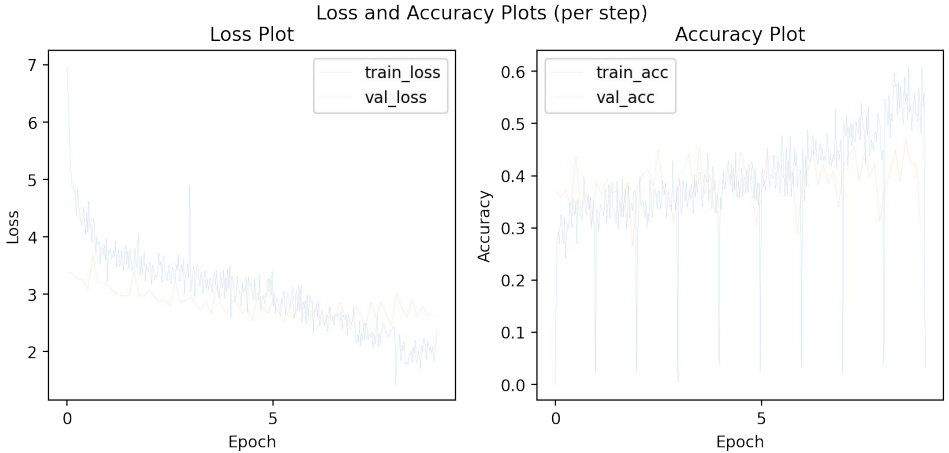
What this does:

1. Image embedding: Pretrained ResNet Architecture (CNN)
2. Convert the questions to embeddings and feed to an LSTM model
3. The concatenated image features and the final state of LSTMs are then used to compute multiple attention distributions over image features
4. The concatenated image feature glimpses and the state of the LSTM is fed to two fully connected layers two produce probabilities over answer classes

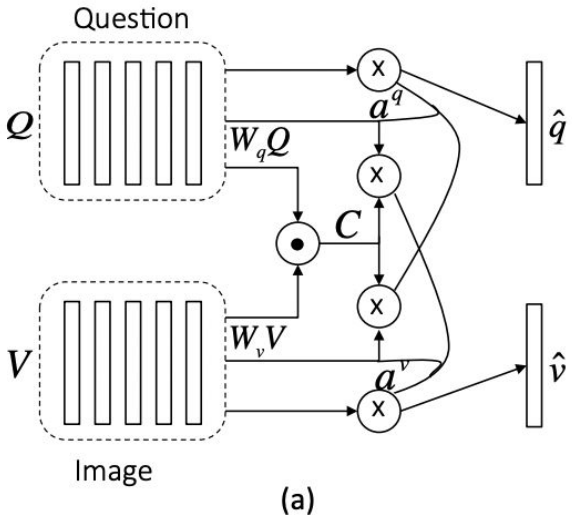
*Image & model from paper: https://arxiv.org/pdf/1704.03162.pdf

Results: Attention Model





Baseline Method: Co-Attention Model

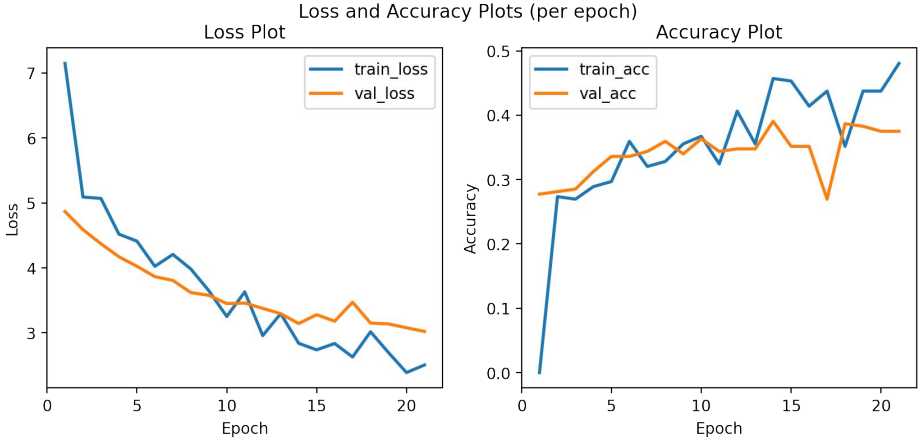


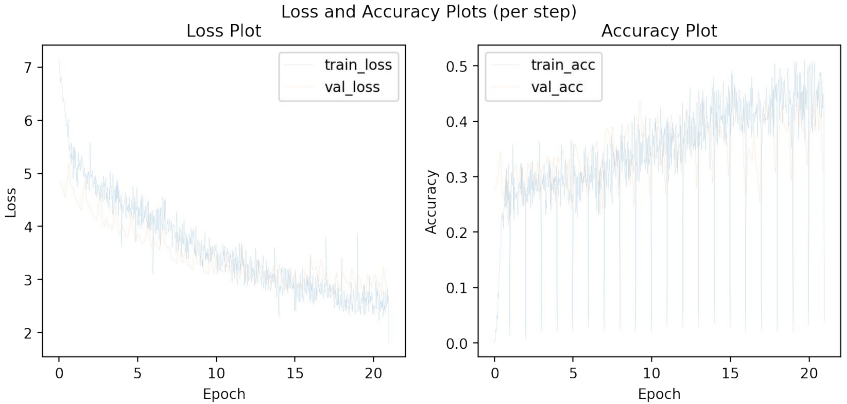
What this does:

1. Parallel Co-Attention: connect the image and question by calculating the similarity between image and question features

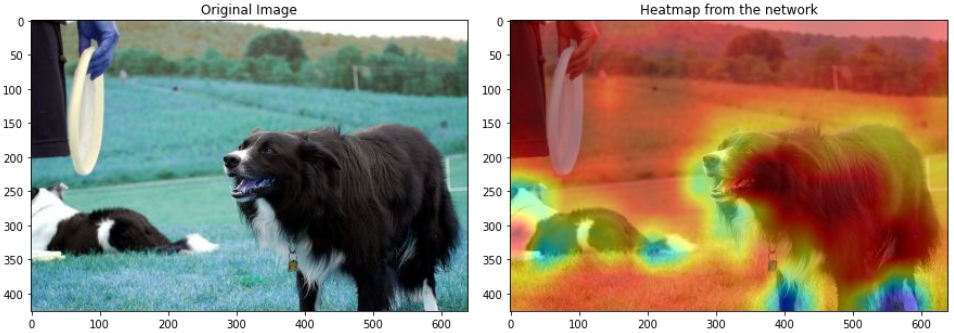
*Image & model from paper: https://arxiv.org/pdf/1606.00061.pdf

Results: Co-Attention Model





HeatMap of Image and Question Visualization



Pending

- Implement a multi-modal approach that uses bounding boxes to train a third model and analyze effect on performance and heatmap visualization.

- Focus on yes/no questions and try to analyze effect of changing question structure and/or inverting questions.

Thank you!