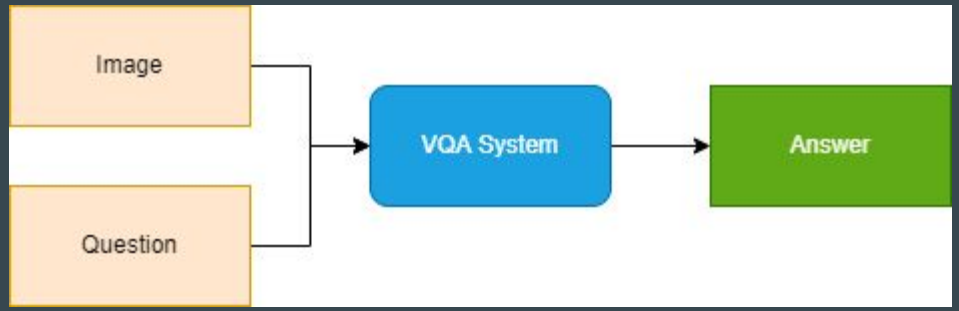
Visual Question Answering

Ajit Deshpande
A59003350

Parthasarathi Kumar
A59003519

Motivation and Problem Definition

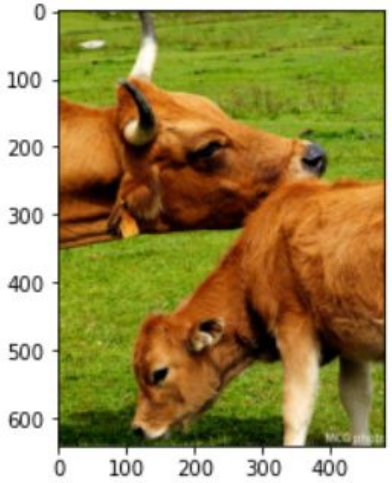
- Given a image and question pair, the model needs to predict a corresponding answer.
- Interesting problem at the intersection of vision and NLP.
- Has applications like system for the visually imapired.



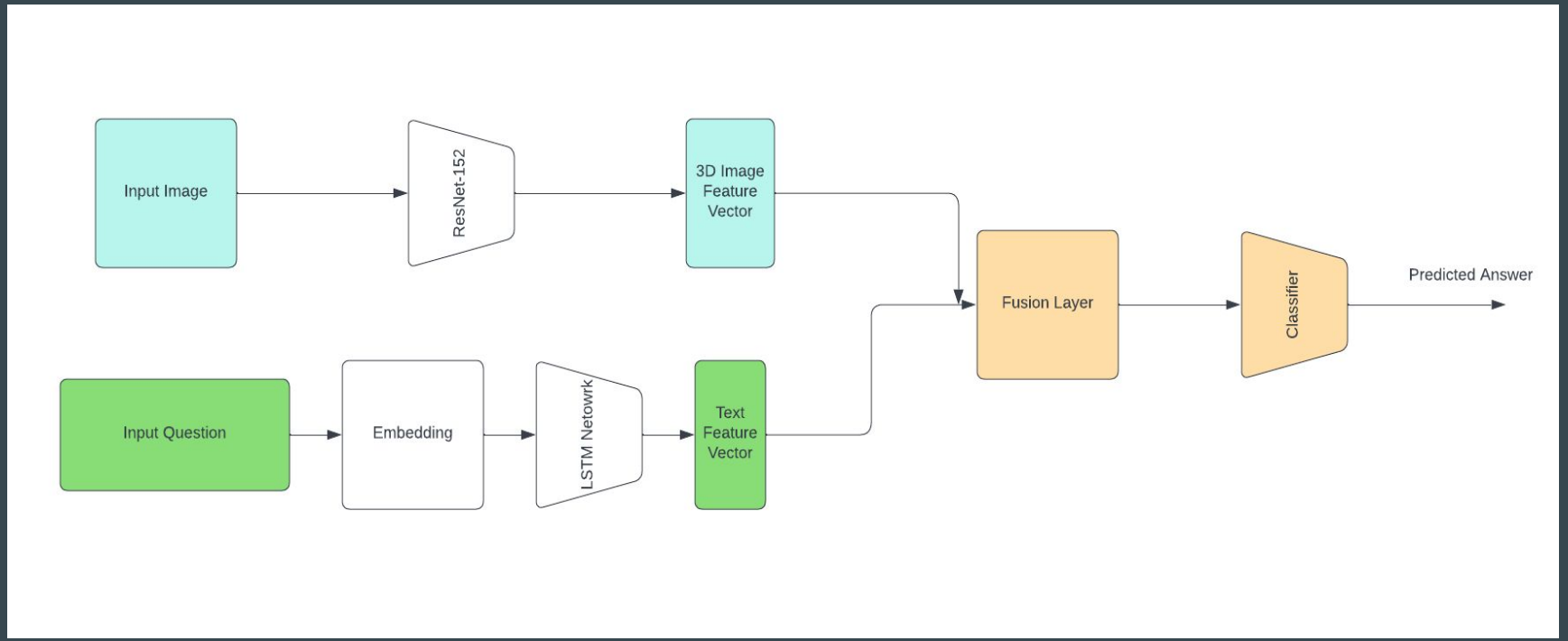
Dataset Overview

- VQA Dataset
  - Training Set has 82,783 images from the MSCOCO dataset
  - Each image is 640X480
  - Each image has 3 associated questions
  - Each question has 10 answers from 10 annotators
  - Validation Set - 40,504 images and Test Set - 81,434 images

```
Question : How many horns does the animal on the left have?
Answers : ['2', '2', '2', '2', '2', '1', '2', '2', '2', '2']
Most Common Answer : 2
----------------------Image--------------------
```

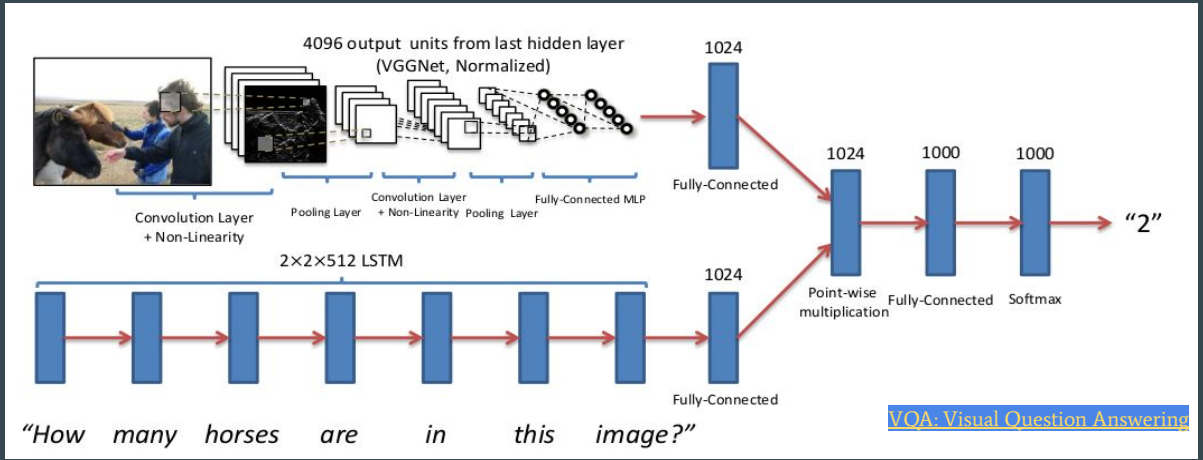


Method Overview



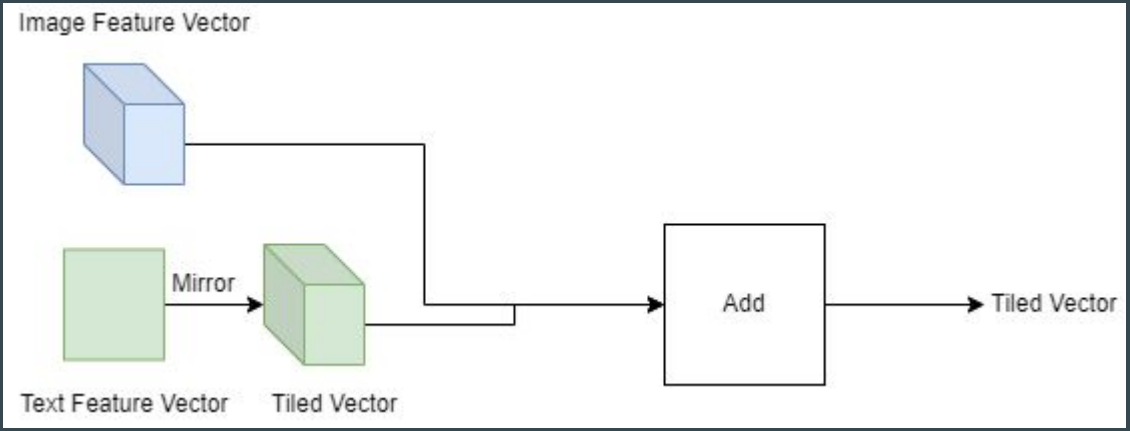
Method Details - Text and Image Encoding

- Typical VQA systems have an image encoding component and question/text encoding component.
- CNN based feature encoder - pretrained ResNet model
- Questions are padded and made of same length.
- String mapped to embeddings for passing to a LSTM based system



Fusion Strategy - Tiled Attention Model

- Create tiles of the text feature vector
- Repeat the text feature vector to make it the same size as image feature vector
- Add the two feature vectors

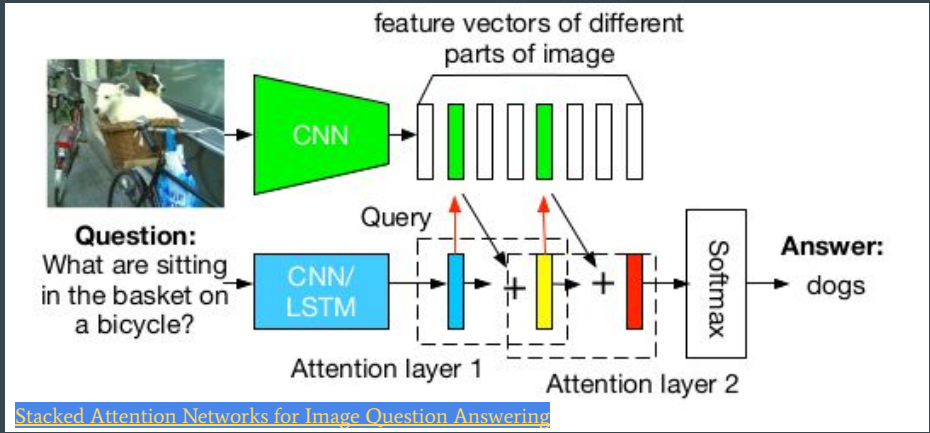


Fusion Strategy - Stacked Attention Model

Based on attention where the question embedding used to retrieve the image feature.

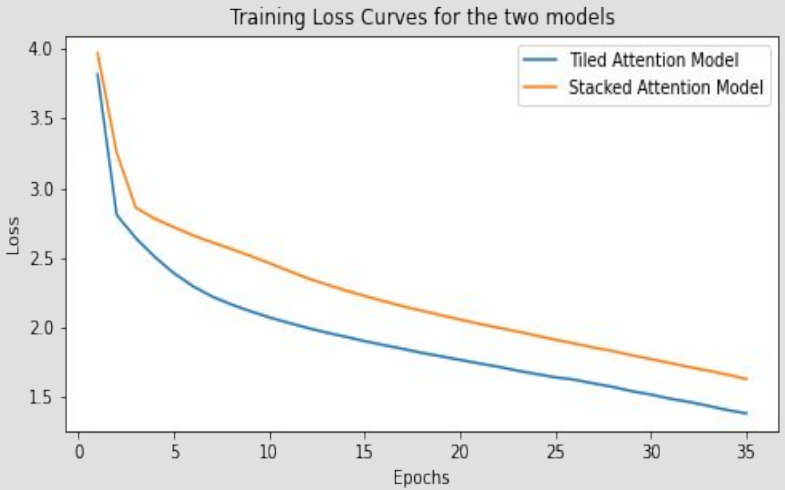
Combined with question embedding to create a refined embedding.

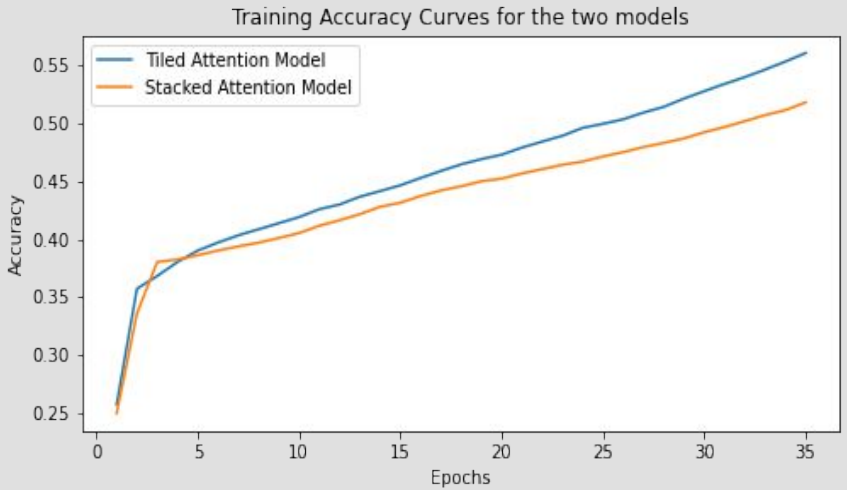
Process repeated twice to mimic effective attention over the image features based on question.



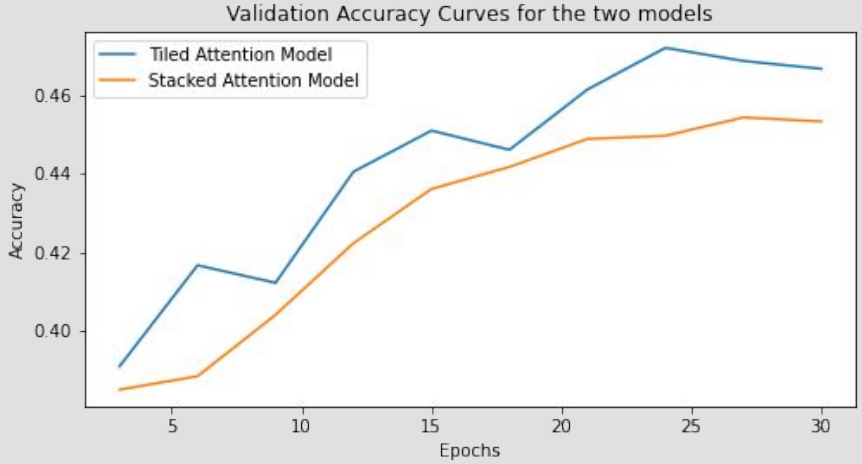
Stacked Attention Networks for Image Question Answering

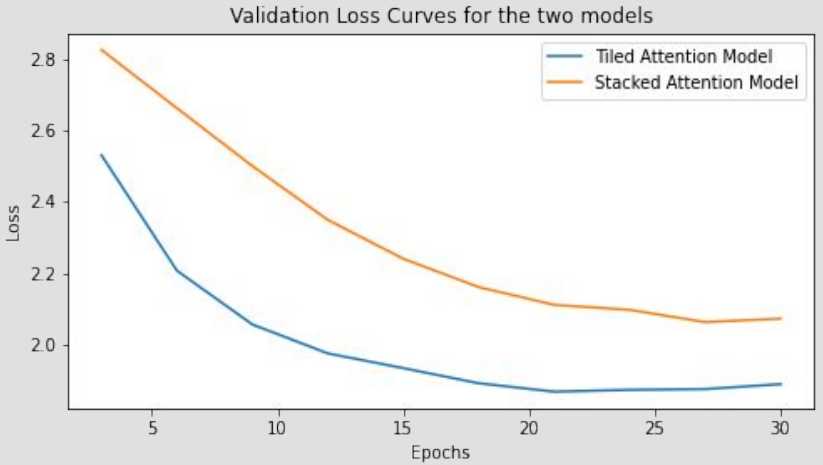
Quantitative Results - Training Curves





Quantitative Results - Validation Curves





Model	Training Accuracy	Validation Accuracy
Tiled Attention Model	0.672	0.472
Stacked Attention Model	0.518	0.455

Qualitative Results

System performs well on questions that have a strong relation with the image, “What” type questions?

Question -> What is he doing?
****
Stacked Answer -> surfing
Tiled Answer -> surfing
****
Ground truth Answer -> surfing

Question -> What color is the tennis court?
****
Stacked Answer -> blue
Tiled Answer -> blue
****
Ground truth Answer -> blue

Question -> How many levels is the bus?
****
Stacked Answer -> 2
Tiled Answer -> 2
****
Ground truth Answer -> 2

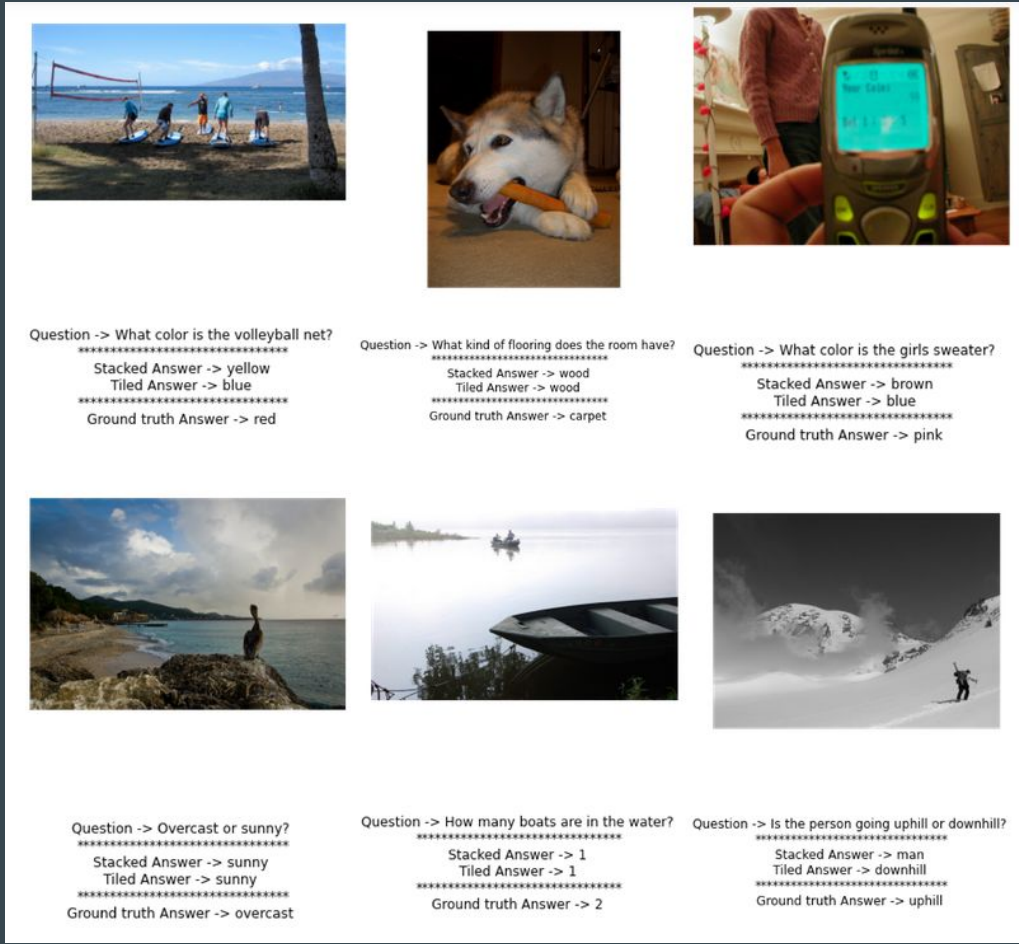
Question -> What room are they in?
****
Stacked Answer -> kitchen
Tiled Answer -> kitchen
****
Ground truth Answer -> kitchen

Question -> What game are they playing?
****
Stacked Answer -> baseball
Tiled Answer -> baseball
****
Ground truth Answer -> baseball

Question -> What game system are they playing?
****
Stacked Answer -> wii
Tiled Answer -> wii
****
Ground truth Answer -> wii

Qualitative Results

Failure cases involve questions that are subjective / ambiguous or associated with image features that are difficult to observe



Conclusion

- We explored the problem of VQA with 2 different attention mechanisms.

- Extension of this work would be to analyse the intermediate attention maps to understand where the images focus to predict the answer the question
- Explore more sophisticated attention based architectures like transformers.
- Explore the literature of visual grounding to improve the VQA performance by intermediate supervision on the attention maps generated by the fusion component.

Thank You!