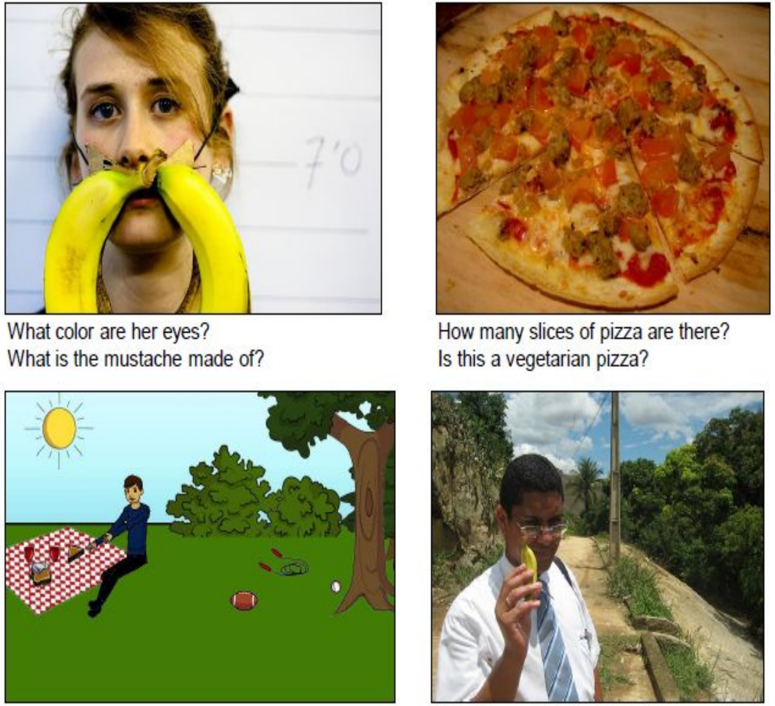
Visual Question Answering

..because it caught our attention

Motivation

1. Visually impaired users
2. Information retrieval for a chain of tasks



What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Is this person expecting company?
What is just under the tree?

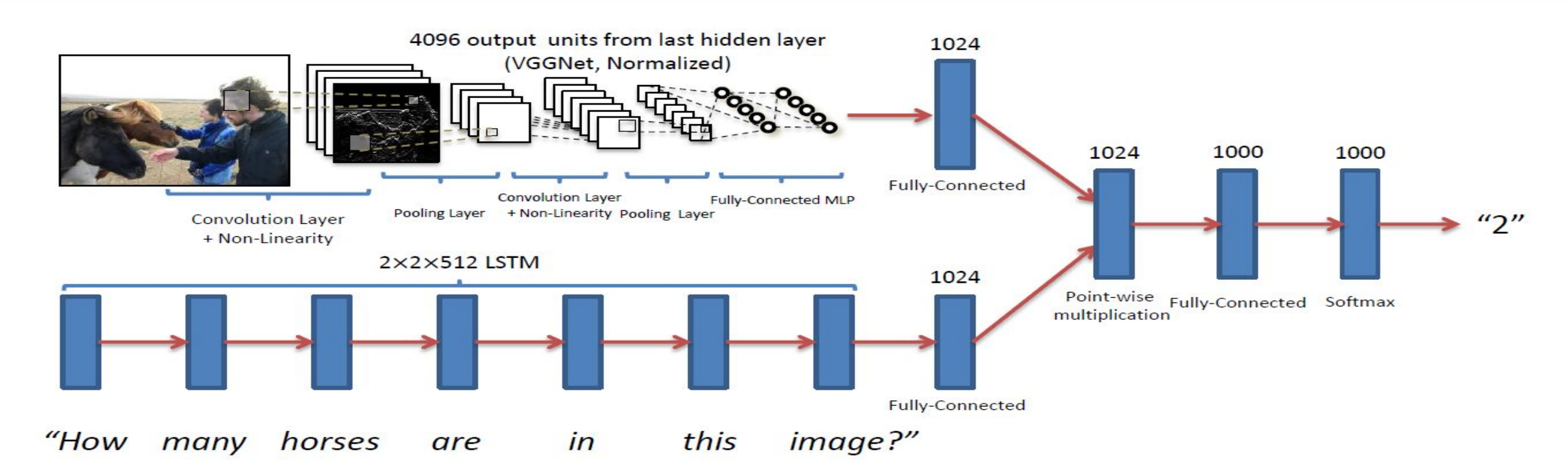
Does it appear to be rainy?
Does this person have 20/20 vision?

Dataset and Evaluation Metrics

1. Images — COCO dataset
2. Questions — 3 unique individuals
3. Answers — 10 unique individuals
4. Accuracy = Minimum (correct answers/3, 1)

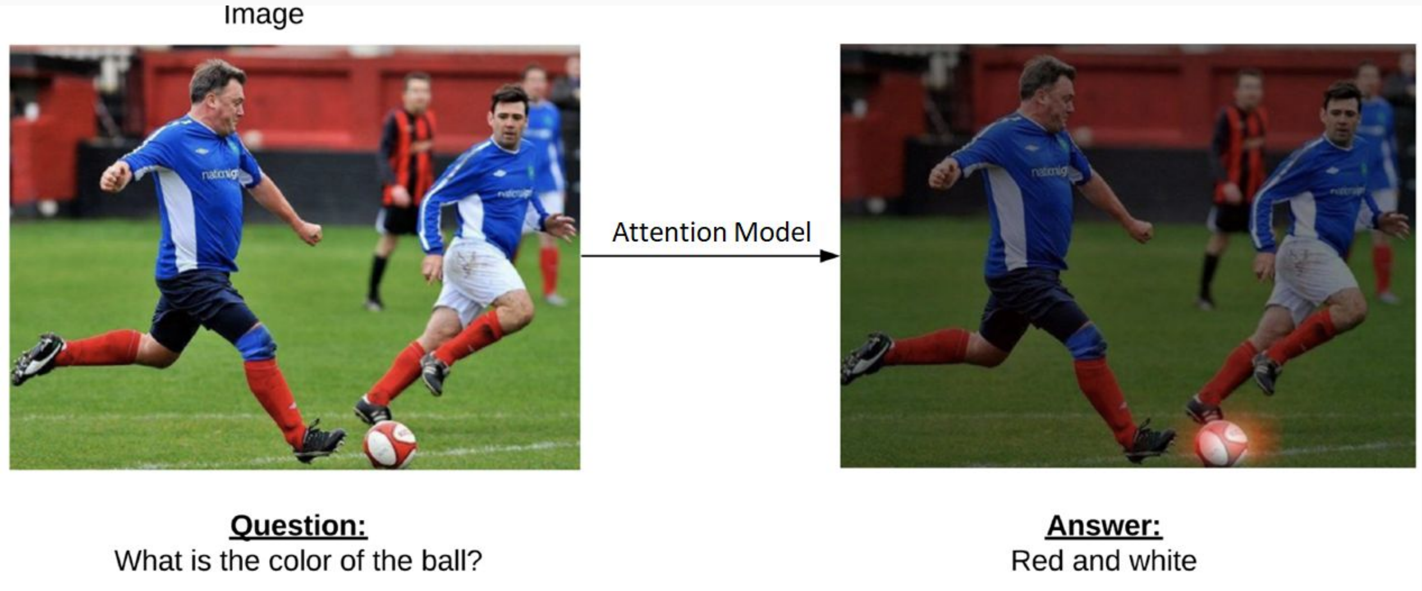
Baseline

1. Image Encoding: VGG19
2. Question Encoding: LSTM
3. Combined the inputs using element wise multiplication
4. Passed through two fully connected layers before finally taking a softmax



Attention

1. Attention models extract more information
2. Higher weights are put on the visual regions that are more relevant to the question.

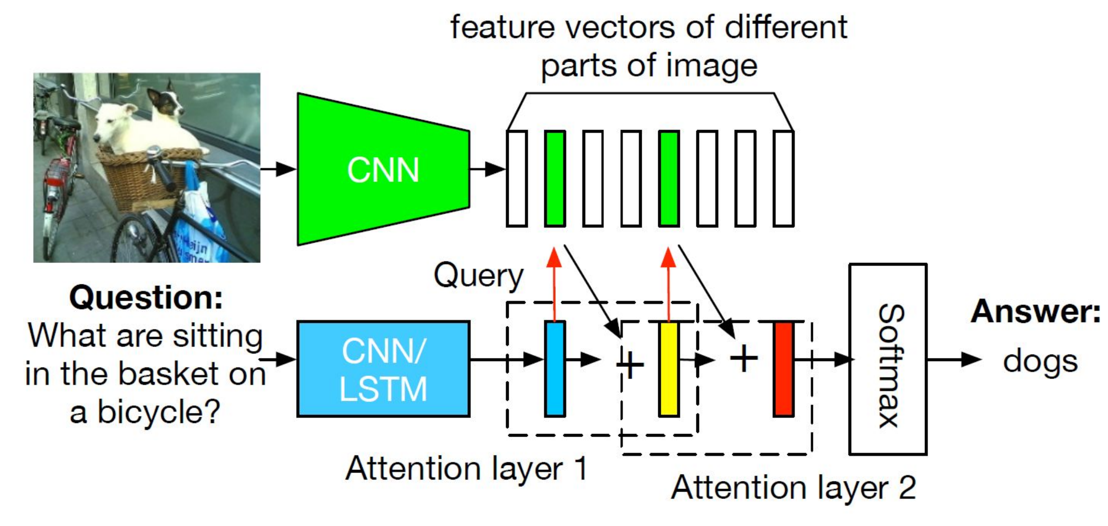


Question:
What is the color of the ball?

Answer:
Red and white

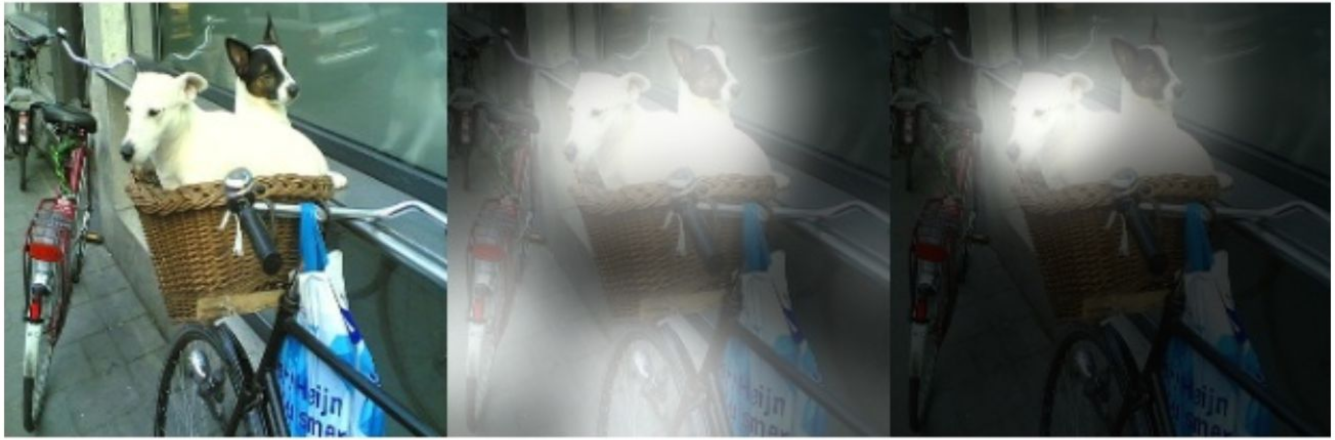
Stacked Attention

1. Iterate the above query-attention process using multiple attention layers
2. Extracts more fine-grained visual attention information for answer prediction


feature vectors of different parts of image
CNN
Query
Question:
What are sitting in the basket on a bicycle?
CNN/ LSTM
+
+
Softmax
Answer:
dogs
Attention layer 1
Attention layer 2

Visualizing Stacked Attention

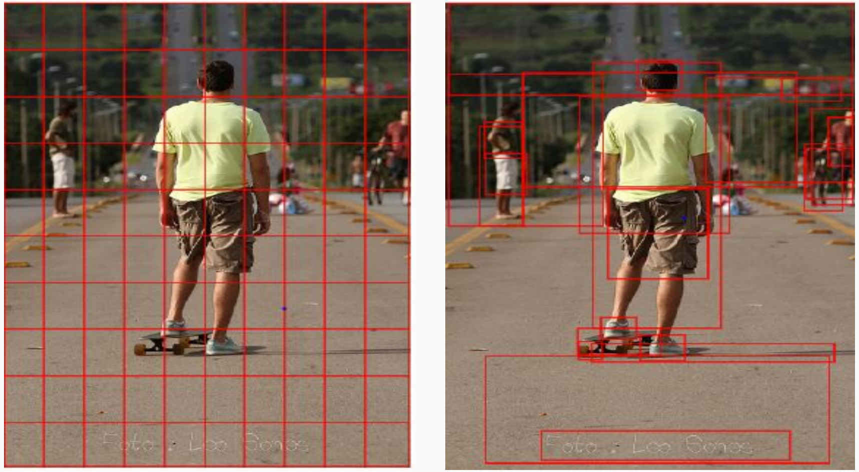
Question: What are sitting in the basket on a bicycle?



Original Image First Attention Layer Second Attention Layer

Bottom-Up and Top -Down Attention

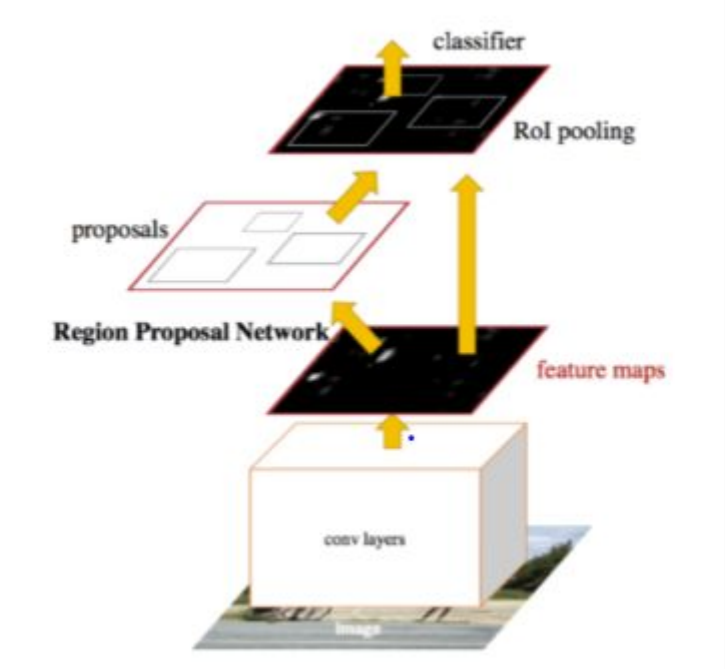
- All the previous approaches on attention can be summarized as top-down approaches
- They make the model focus on particular region of the image based on the state of the current task (e.g. looking for something).
- However, in this approach all the image features are divided into uniform grid which are then weighted by the top-down attention.
- Dividing image into regions by interesting objects is a better idea.



In the left image the image regions are divided into uniform grid which is weighted by attention. In the right,weight image by interesting regions.

Architecture

- Hence, the authors propose a bottom-up attention model. This visual encoder provides the encoded image features as well as the objects of interest.
- This is achieved by using a **Faster R-CNN**.
- The first stage predicts object proposals and the second stage extracts a feature map for each box proposal.

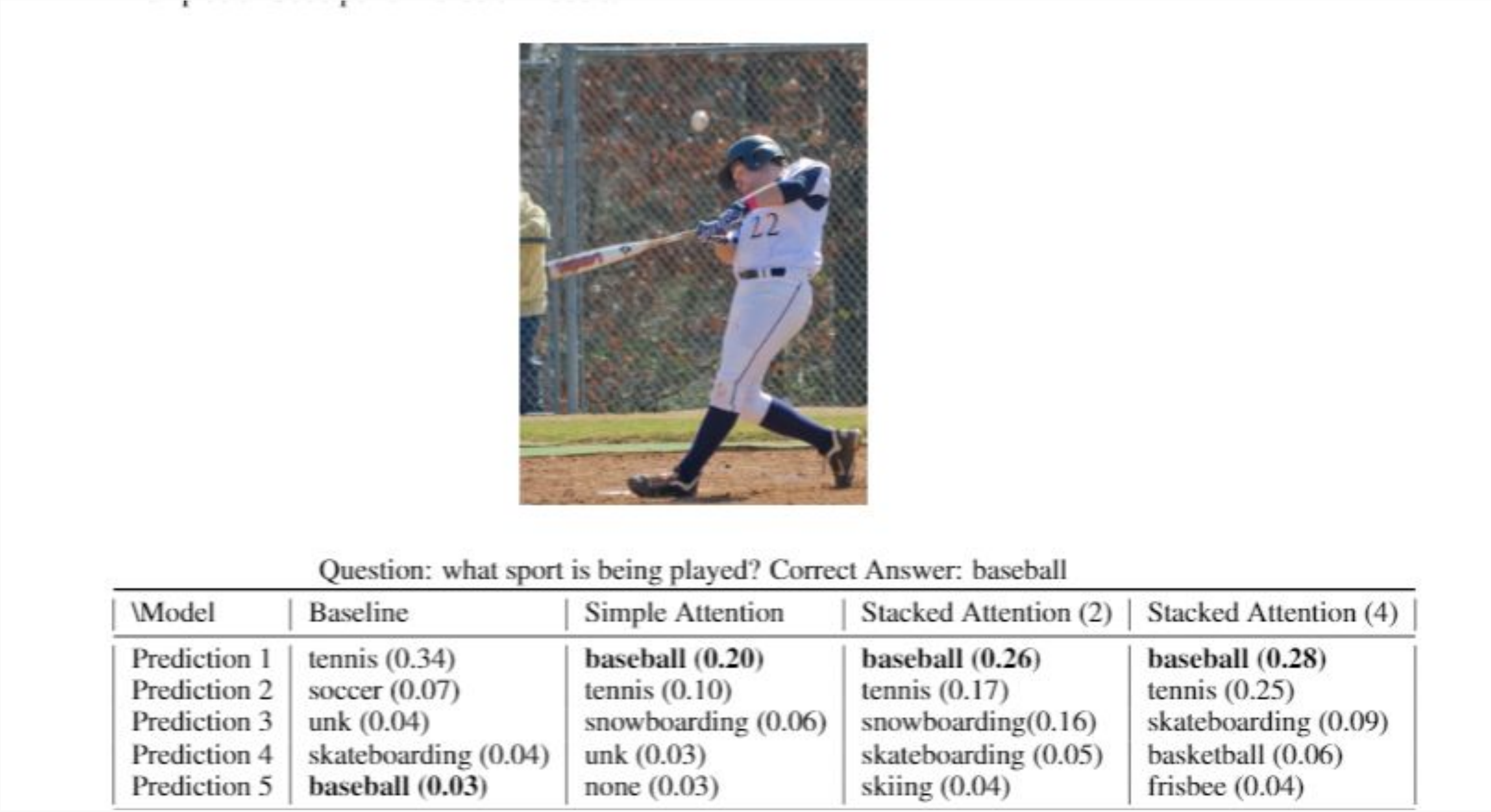


Results

Results on Validation Data Set

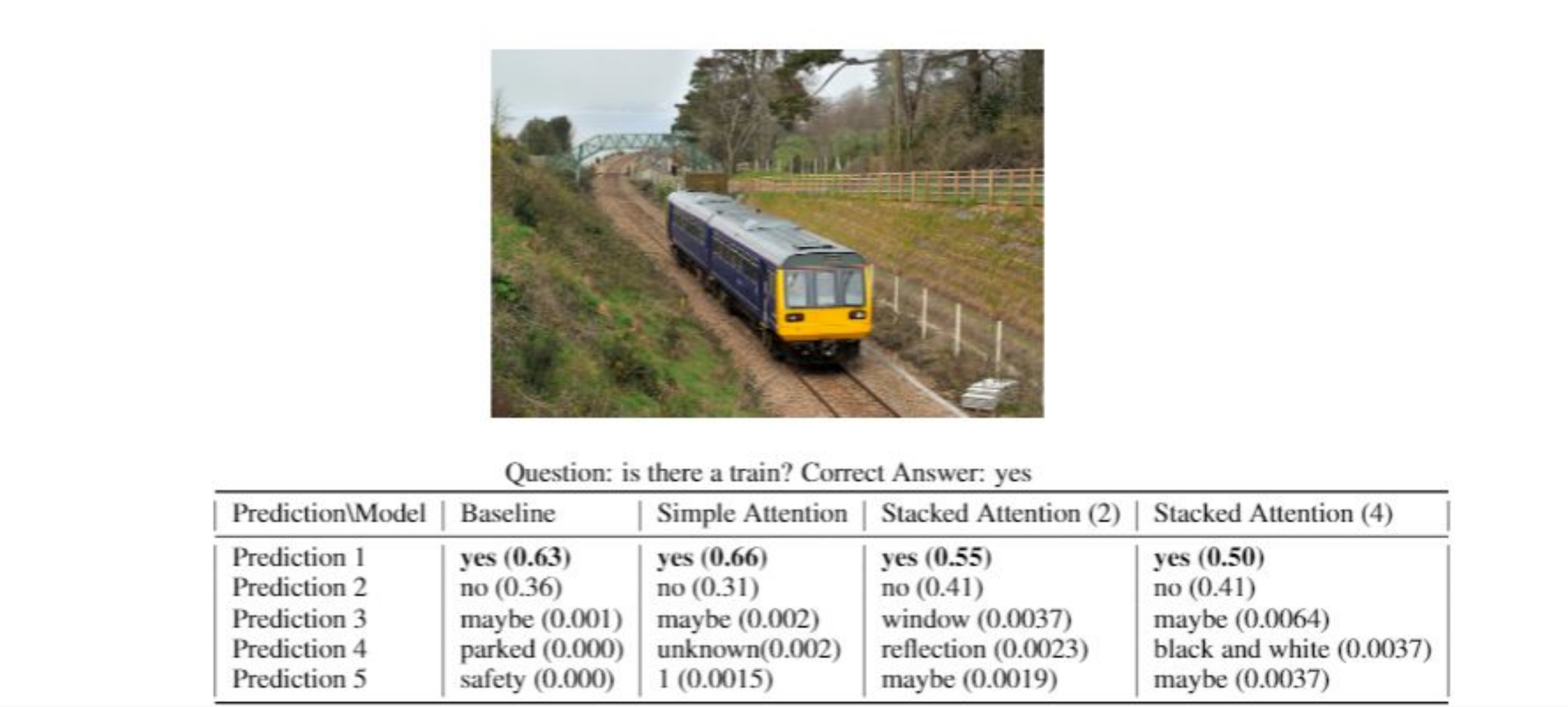
Model	**Validation Accuracy**
Baseline	**57.50%**
Simple Attention	**59.01%**
Two Layered Stacked Attention	**59.98%**
Four Layered Stacked Attention	**60.48%**
Bottom-up and Top-Down Attention	**62.71%**

Demo Images



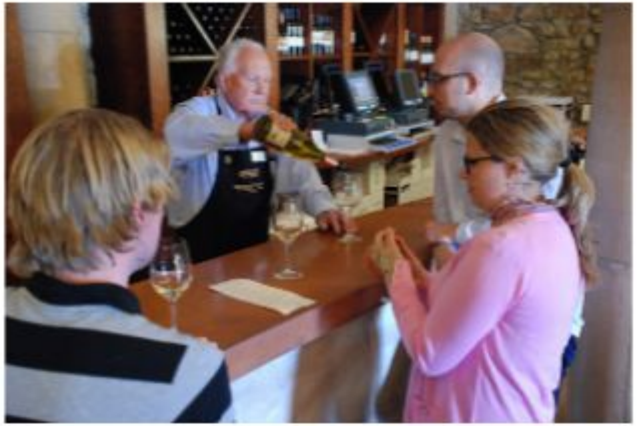
Question: what sport is being played? Correct Answer: baseball

\Model	Baseline	Simple Attention	Stacked Attention (2)	Stacked Attention (4)
Prediction 1	tennis (0.34)	**baseball (0.20)**	**baseball (0.26)**	**baseball (0.28)**
Prediction 2	soccer (0.07)	tennis (0.10)	tennis (0.17)	tennis (0.25)
Prediction 3	unk (0.04)	snowboarding (0.06)	snowboarding(0.16)	skateboarding (0.09)
Prediction 4	skateboarding (0.04)	unk (0.03)	skateboarding (0.05)	basketball (0.06)
Prediction 5	**baseball (0.03)**	none (0.03)	skiing (0.04)	frisbee (0.04)



Question: is there a train? Correct Answer: yes

Prediction\Model	Baseline	Simple Attention	Stacked Attention (2)	Stacked Attention (4)
Prediction 1	**yes (0.63)**	**yes (0.66)**	**yes (0.55)**	**yes (0.50)**
Prediction 2	no (0.36)	no (0.31)	no (0.41)	no (0.41)
Prediction 3	maybe (0.001)	maybe (0.002)	window (0.0037)	maybe (0.0064)
Prediction 4	parked (0.000)	unknown(0.002)	reflection (0.0023)	black and white (0.0037)
Prediction 5	safety (0.000)	1 (0.0015)	maybe (0.0019)	maybe (0.0037)



Question: how many people are there? Correct Answer: 4

Prediction\Model	Baseline	Simple Attention	Stacked Attention (2)	Stacked Attention (4)
Prediction 1	2 (0.23)	1 (0.22)	1 (0.10)	1 (0.19)
Prediction 2	1 (0.22)	2 (0.21)	2 (0.07)	2 (0.11)
Prediction 3	3 (0.14)	3 (0.06)	0 (0.06)	**4 (0.08)**
Prediction 4	**4 (0.10)**	**4 (0.06)**	3 (0.05)	0 (0.08)
Prediction 5	0 (0.08)	many (0.04)	10 (0.05)	3 (0.06)

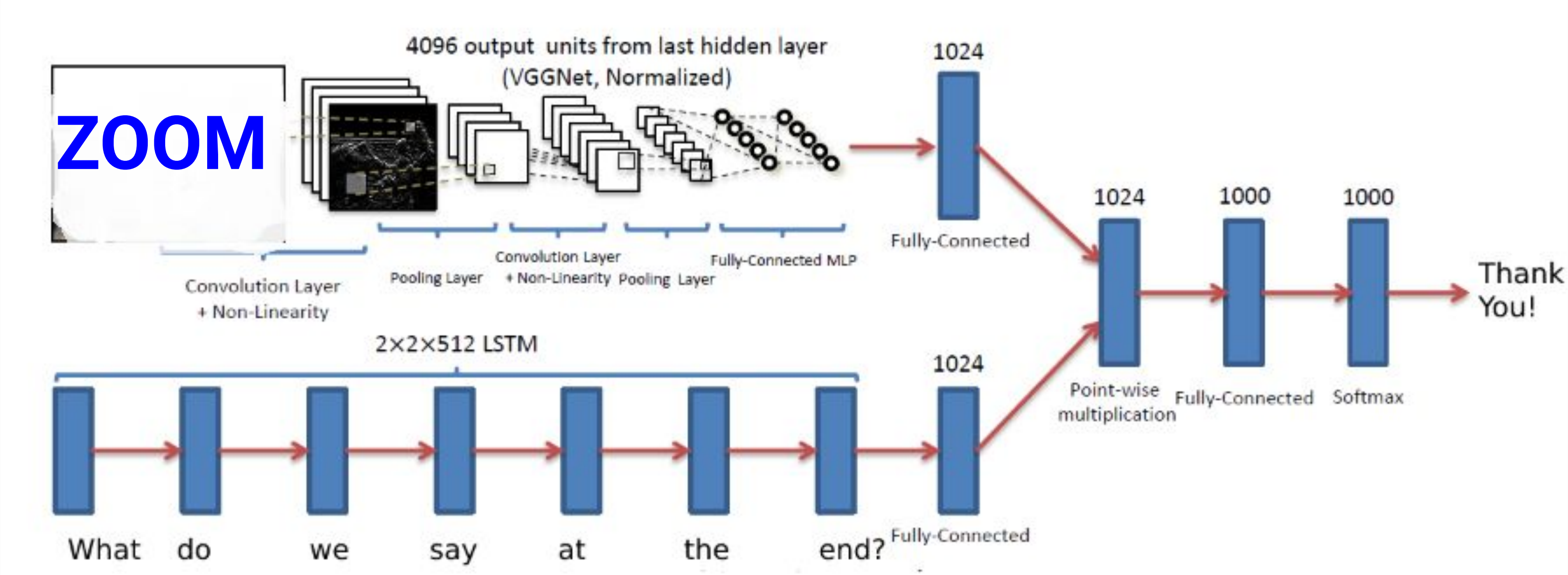
Next Steps

1. Combine lower, middle, and final layers of CNN
2. Bilinear Attention Models
3. Mask R-CNNs

Summary

1. Various models for VQA
2. Improved performance with the use of attention
3. Plausible answers obtained on new images

Questions?


ZOOM
4096 output units from last hidden layer
(VGGNet, Normalized)
Convolution Layer
+ Non-Linearity
Pooling Layer
Convolution Layer
+ Non-Linearity
Pooling Layer
Fully-Connected MLP
1024
Fully-Connected
2×2×512 LSTM
What
do
we
say
at
the
end?
1024
Fully-Connected
1024
Point-wise
multiplication
1000
Fully-Connected
1000
Softmax
Thank
You!