# Testing Image Understanding Through Question Answering

Vaibhav Singh, Sankalp Dayal, Kevin Tsai {vaibhvs, sankalpd, kevin259}@stanford.edu

## Visual Question Answering Problem

”How well do these networks really understand images?” We explore this question through through two approaches: first through the task of Visual Question Answering (VQA), where the model is trained on images and associated question answer pairs in natural language. Second, we measure effectiveness of transfer-learning to image recognition. We use a model pretrained on image-recognition task and retrain its weights during VQA training and then test it again on image recognition task.

## Methods & Algorithms

Architecture of VQA systems proposed has following modules

- Input Image Module: To extract features from the input imageInput Question Module: To learn some representation of input question
- Attention Module: For attention mechanism to identify the relevant regions of image (yang et. al, 2016, Xiong et. al., 2016) and question (Lu et. al., 2017),
- Output Module: For Softmax output

We considered Stacked attention network (SAN), Yang et. al. 2014, and Dynamic Memory Network+ (DMN+), Xiong et. al., 2016. .

In DMN+, for image input module, we feed the raw input image to a pre trained VGG-19 model. Each of these local regional vectors obtained from VGG's last pooling layer are multiplied by weights to give feature embeddings fi. The traversal to create embeddeing is done from left to right and row by row and given as input to a bidirectional GRU

$$\overrightarrow{f_i} = GRU_{fwd}(f_i, \overrightarrow{f_{i-1}}) \quad \overleftarrow{f_i} = GRU_{bwd}(f_i, \overleftarrow{f_{i+1}}) \quad \overleftrightarrow{f_i} = \overleftarrow{f_i} + \overrightarrow{f_i}$$

The output of this becomes input to the episodic memory module. In the episodic memory module the attention is implemented as follows

$$z_i^t = [\overleftrightarrow{f_i} o\, q; \overleftrightarrow{f_i} o\, m^{t-1}; |\overleftrightarrow{f_i} - q|; |\overleftrightarrow{f_i} - m^{t-1}|]$$

$$Z_i^t = W^{(2)} tanh(W^{(1)} z_i^t + b^{(1)}) + b^{(2)}$$

New episode memory state is obtained by $m^t = ReLU(W^t[m^{t-1}; c^t; q] + b)$

In SAN output of visual input module and question module are fed to Stacked attention network. the image feature vector is fed to a single layer neural network with softmax output to calculate the attention distribution.

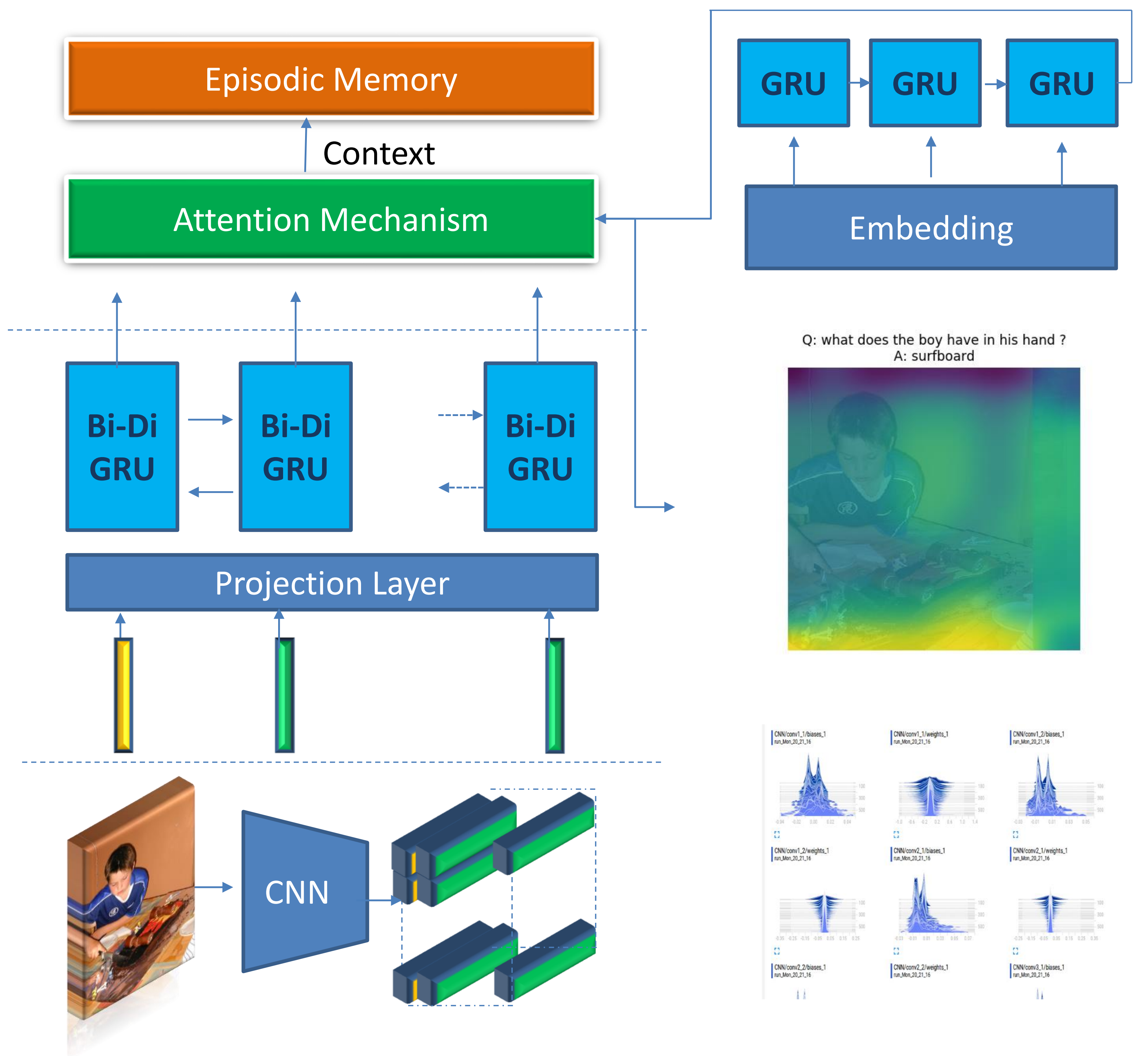
$$h_A = tanh(W_{I,A^{v_I}} \oplus (W_{Q,A^{v_Q}} + b_A)),$$

$$p = softmax(W_h + b)$$

## VQA Dataset And Experiments

In dataset for each image, three questions and for each question there are ten answers annotated by human annotators. There are two versions of this dataset. Version 2 has balanced real images as compared to Version1. Both versions have 204,721 COCO images. Of which 82,783 training images and 81,434 test images. It has 4,437,570 annotations about these images and 443,757 training questions. For test, there are no annotations and 447,793 testing questions

## Architecture

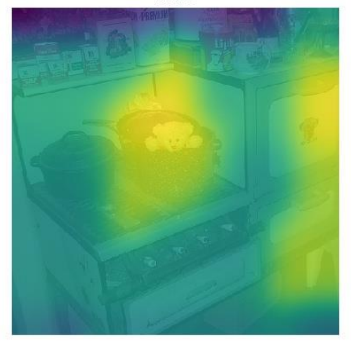

Episodic Memory
Context
Attention Mechanism
GRU
GRU
GRU
Embedding
Bi-Di GRU
Bi-Di GRU
Bi-Di GRU
Projection Layer
CNN

## Results

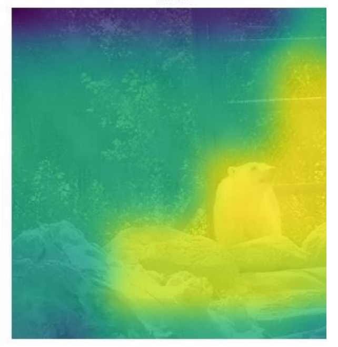
Performed analysis on SAN and different settings of DMN+. Following table shows the overall accuracy and for different answer types

| Method | Dataset | All | Yes/no | Number | Other |
|---|---|---|---|---|---|
| SAN(2, CNN) | VQA 1.0 | 52.3 | 79.3 | 36.6 | 46.1 |
| DMN+ (glove initialization) | VQA 1.0 | 52.64 | 77.3 | 29.34 | 31.37 |
| DMN+ (random initialization) | VQA 1.0 | 54.27 | 78.33 | 38.24 | 31.46 |

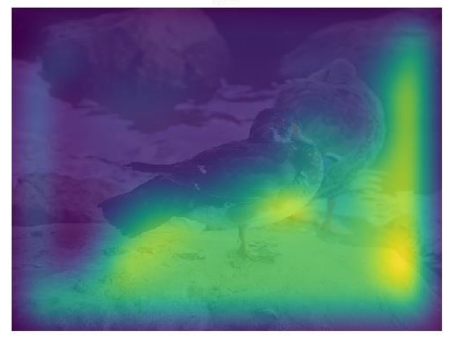
Q: does this stove look updated ?
A: no



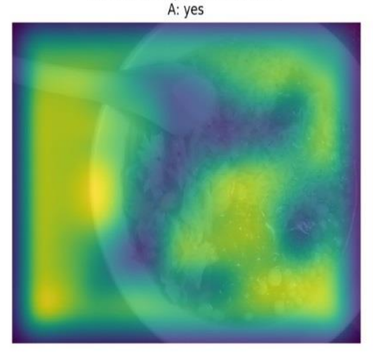
: is this bear's fur different than other bear's
A: no



Q: how many birds ?
A: 2



Q: is there a spoon in the bowl ?
A: yes



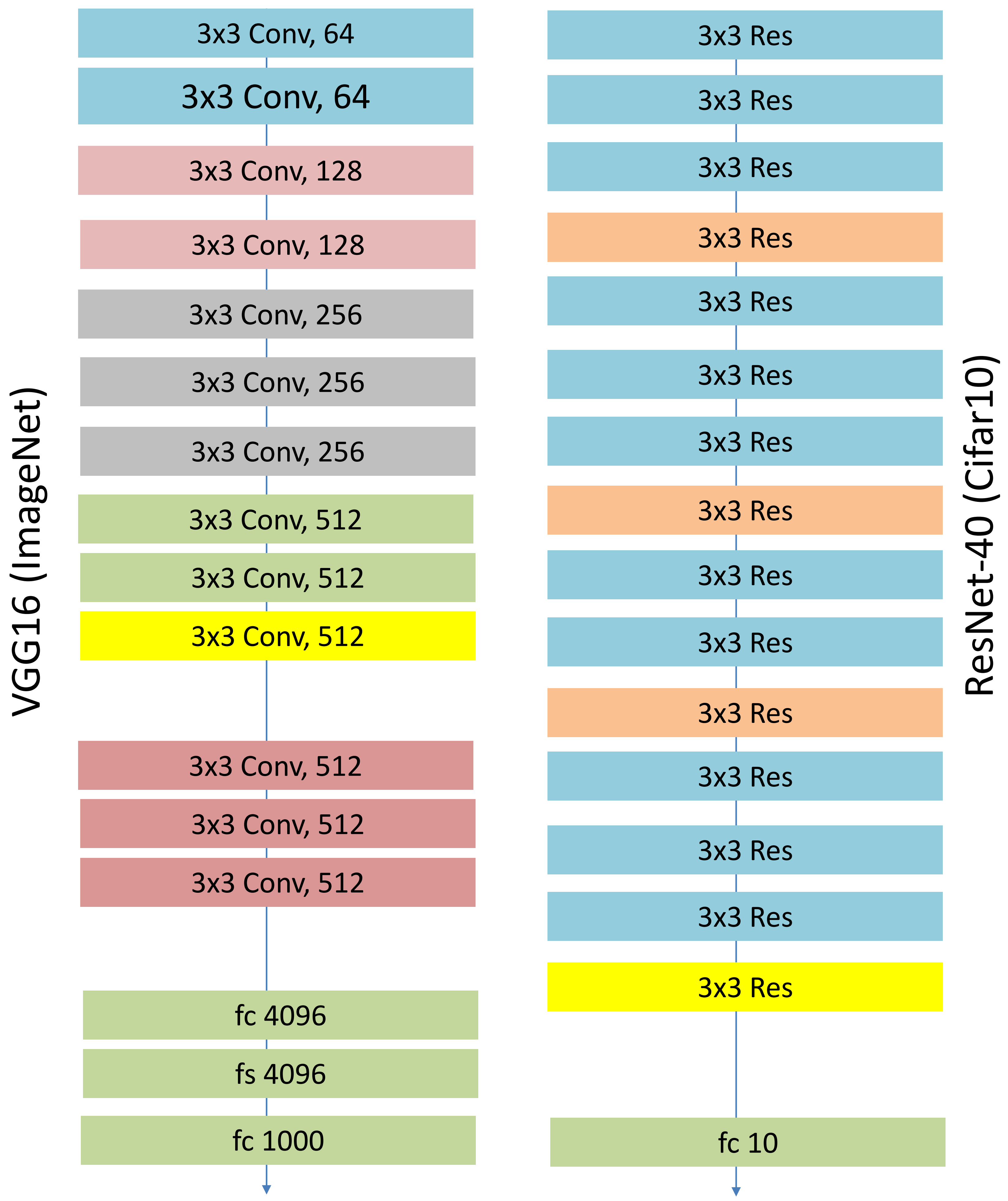
Q: is this a beach ?
A: no

Q: how many trees are in this image ?
A: 2

Q: what is the lady wearing around her neck ?
A: nothing

Q: what is the woman standing on ?
A: surfboard

## Feature Extraction


VGG16 (ImageNet)
3x3 Conv, 64
3x3 Conv, 64
3x3 Conv, 128
3x3 Conv, 128
3x3 Conv, 256
3x3 Conv, 256
3x3 Conv, 256
3x3 Conv, 512
3x3 Conv, 512
3x3 Conv, 512
3x3 Conv, 512
3x3 Conv, 512
3x3 Conv, 512
fc 4096
fs 4096
fc 1000

ResNet-40 (Cifar10)
3x3 Res
3x3 Res
3x3 Res
3x3 Res
3x3 Res
3x3 Res
3x3 Res
3x3 Res
3x3 Res
3x3 Res
3x3 Res
3x3 Res
3x3 Res
3x3 Res
3x3 Res
fc 10

## Conclusions and Ongoing Work

- SAN are DMN+ show very similar performance
- Random initialization of DMN+ doesn't affect the performance
- Visualizations on attention demonstrate the model is able to identify regions of interest well
- Ongoing performance evaluation on VQA 2.0
- Ongoing using trained Resnet on CIFAR 10

## References

[1] Stacked Attention Netowrks for Image Question Answering", Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng and Alex Smola. To appear in CVPR 2016.

[2] Dynamic Memory Networks for Visual and Textual Question Answering Caiming Xiong, Stephen Merity, Richard Socher

[3]Hierarchical Question-Image Co-Attention for Visual Question Answering, Jiasen Lu, Jianwei Yang, Dhruv Batra, Devi Parikh