Visual Question-Answering

GROUP 3

PARTH SHAH (20759634)

PRIYA PATEL (20787652)

PRATEETI DEB CHAUDHURI (20802005)

UNIVERSITY OF WATERLOO

Road Map

Introduction

Literature review

Methodology

Experimental Result

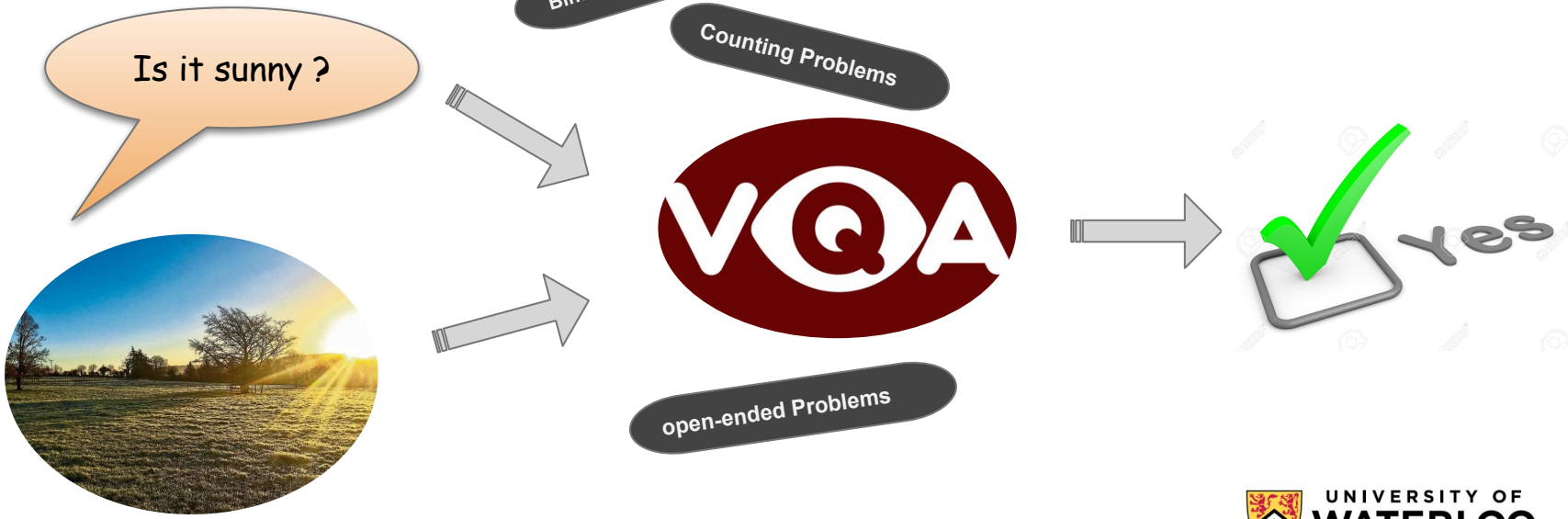
Summary of work

References

UNIVERSITY OF WATERLOO

What is VQA ?

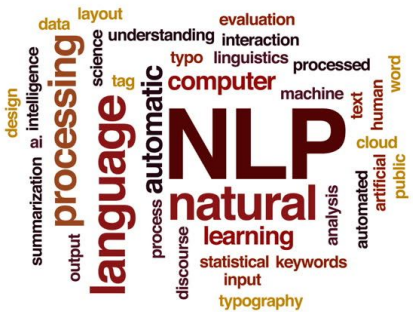
Predict the answer of given question related to an image.

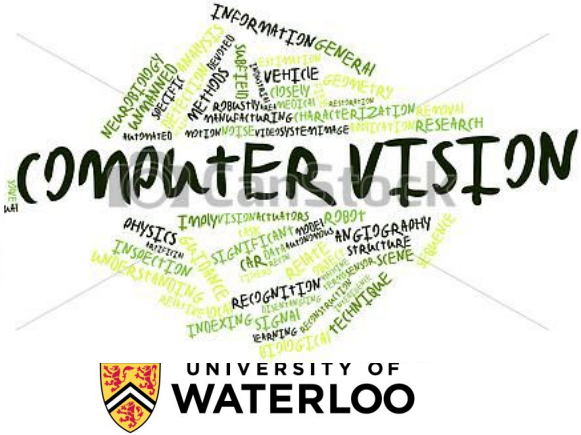


UNIVERSITY OF WATERLOO

Why VQA?

- Useful for helping blind or visually impaired persons
- Multidisciplinary task
- A Visual Turing Test





Literature Review

- **VQA: Visual Question Answering(Aishwarya Agrawal, Jiasen Lu, Et al)**
  - Presented the open-ended Visual Question Answering dataset consisting MS COCO images, open ended questions and human annotated answers.
  - Baseline model for the task

Literature Review (cont.)

- **Visual Question Answering: Datasets, Algorithms and Future Challenges(Author: Kushal Kafle and Christopher Kannan)**
  - Compared the datasets available for VQA like DAQUAR, Visual 7W, FM -IQA, VQA, etc.
  - Reported all the different types of techniques (i.e. Baseline models, attention modes, bayesian models, etc.) used to solve the VQA and compared them.

Literature Review (cont.)

- **Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge (Damien Teney,Peter Anderson, Et al.)**
  - Used image features from bottom-up attention that provide region-specific features
  - massive exploration of architectures and hyperparameters
  - sigmoid output

Road Map

Introduction

Literature review

Methodology

Experimental Result

Summary of work

References

UNIVERSITY OF WATERLOO

Dataset used

- COCO-VQA dataset
- ~83K training images,
- ~41K validation images,
- ~81K testing images
- Open-ended task.
- 1 Image * 3 question * 10 answer

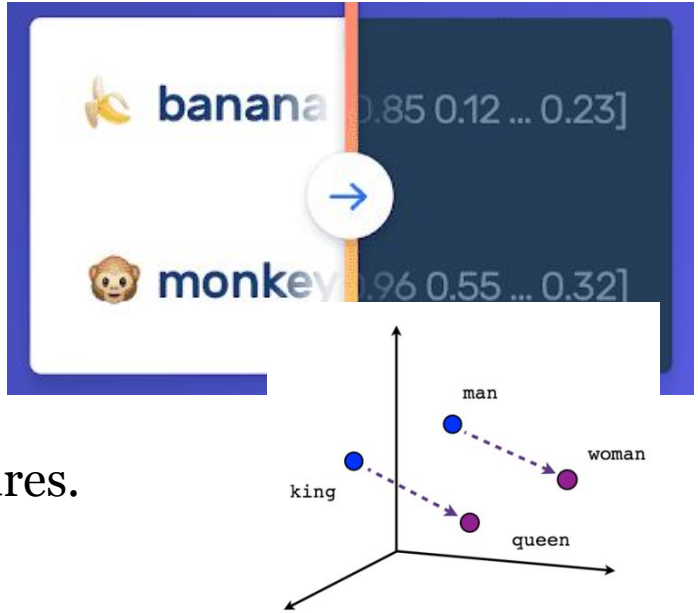
Evaluation Metric

$$\text{Acc}(ans) = \min\left\{\frac{\#\text{humans that said } ans}{3}, 1\right\}$$

- The annotators are generated for the VQA dataset as ten answers per question.
- Where the total number of annotators is denoted by # which has the same answer as that of the algorithm.

UNIVERSITY OF WATERLOO

Pre-trained models used



VGG16 Pretrained weights

- very versatile, simple and relatively small.
- Remove the last 2 layers to extract the features.

GloVe word embedding

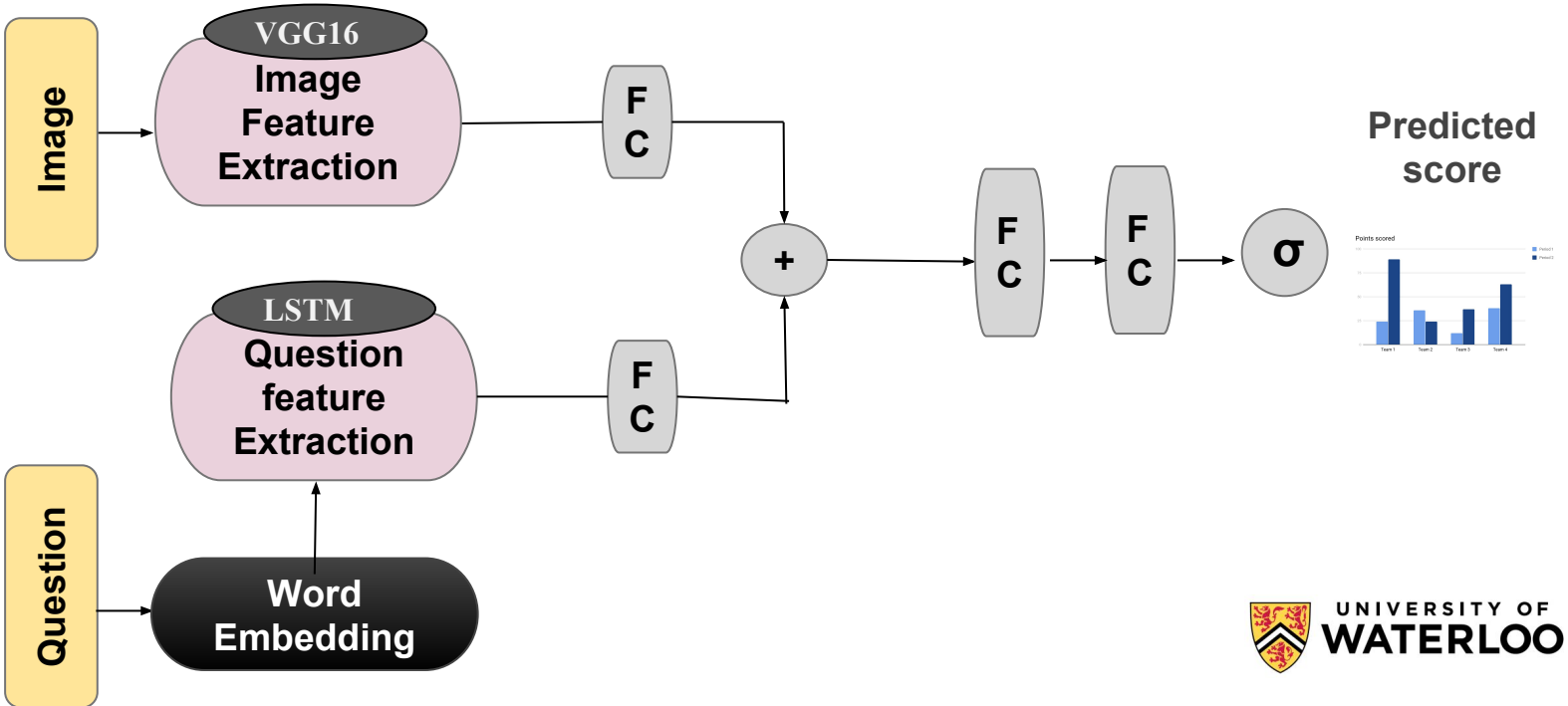
- reduces a given token into a 300 dimensional representation.
- co-occurrence matrix (words X context) that basically count how frequently a word appears in a context.

UNIVERSITY OF WATERLOO

Let's start with Basic Model

- Built a basic model using
  - **Convolutional Neural Network (CNN)** for image recognition and,
  - **Long Short Term Memory (LSTM)** for natural language processing
  - Then, combining the results to deliver the final answer for VQA.

Basic Model for VQA



UNIVERSITY OF WATERLOO

Road Map

Introduction

Literature review

Goal and Methodology

Experimental Result

Summary of work

References

UNIVERSITY OF WATERLOO

VQA DEMO

Question

Which animals are there?

Output

```
64.17 % ['birds']!
5.91 % ['soccer']!
4.09 % ['eagle']!
2.45 % ['bears']!
2.13 % ['cows']!
```

UNIVERSITY OF WATERLOO

VQA DEMO

Question

Is it raining?

Output

```
80.14 % ['no']!
19.85 % ['yes']!
0.0 % ['summer']!
0.0 % ['in air']!
0.0 % ['green']!
```

UNIVERSITY OF WATERLOO

VQA DEMO

Question

Is there a baby?

Output

```
90.58 % ['yes']!
9.41 % ['no']!
0.0 % ['boy']!
0.0 % ['christmas']!
0.0 % ['girl']!
```

UNIVERSITY OF WATERLOO

VQA DEMO

Question

What is the color of the car?

Output

```
63.77 % ['red']!
12.73 % ['silver']!
6.82 % ['red and white']!
4.19 % ['blue']!
3.86 % ['white']!
```

UNIVERSITY OF WATERLOO

VQA DEMO

Question

How many children are there?

Output

```
66.53 % ['2']!
9.49 % ['3']!
6.78 % ['4']!
5.47 % ['5']!
3.87 % ['1']!
```

UNIVERSITY OF WATERLOO

Road Map

Introduction

Literature review

Goal and Methodology

Experimental Result

Summary of work

References

UNIVERSITY OF WATERLOO

Results

- Achieved 52.87% accuracy on the validation set after 18th epoch
- Training Time: 10-11 mins/epoch on single NVIDIA K80

Moving Ahead

- Models built using only question/answer pair have shown 45.06% of accuracy.
- Using global features alone may obscure task-relevant regions of the input space.
- certain visual regions in an image and certain words in a question are more informative than others for answering a given question.

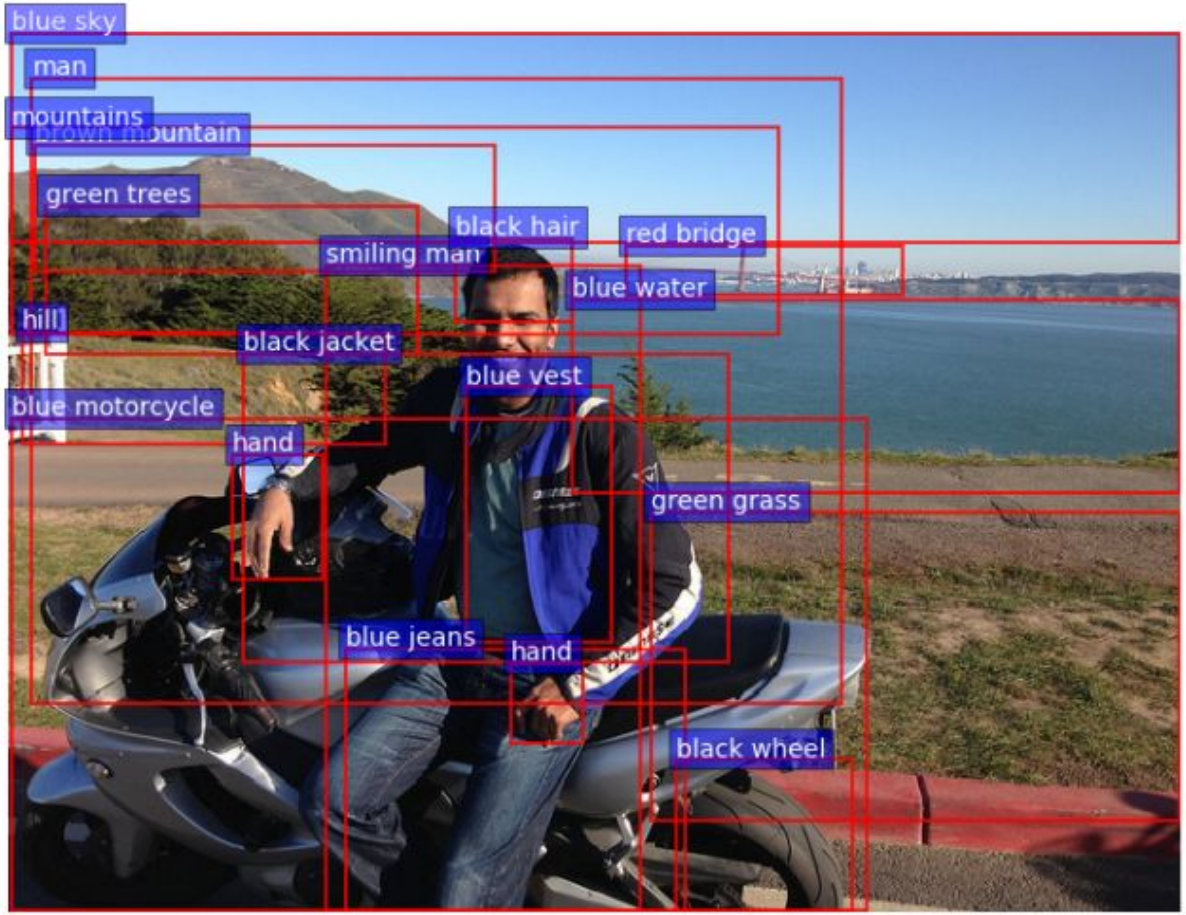
Attention Network

- Helps the algorithm decide which parts of the image to focus on
- It uses an output from one model to focus on specific sections of another neural network.

UNIVERSITY OF WATERLOO

Pretrained Feature

Example of features extracted from the image using attention model



Attention Model Architecture

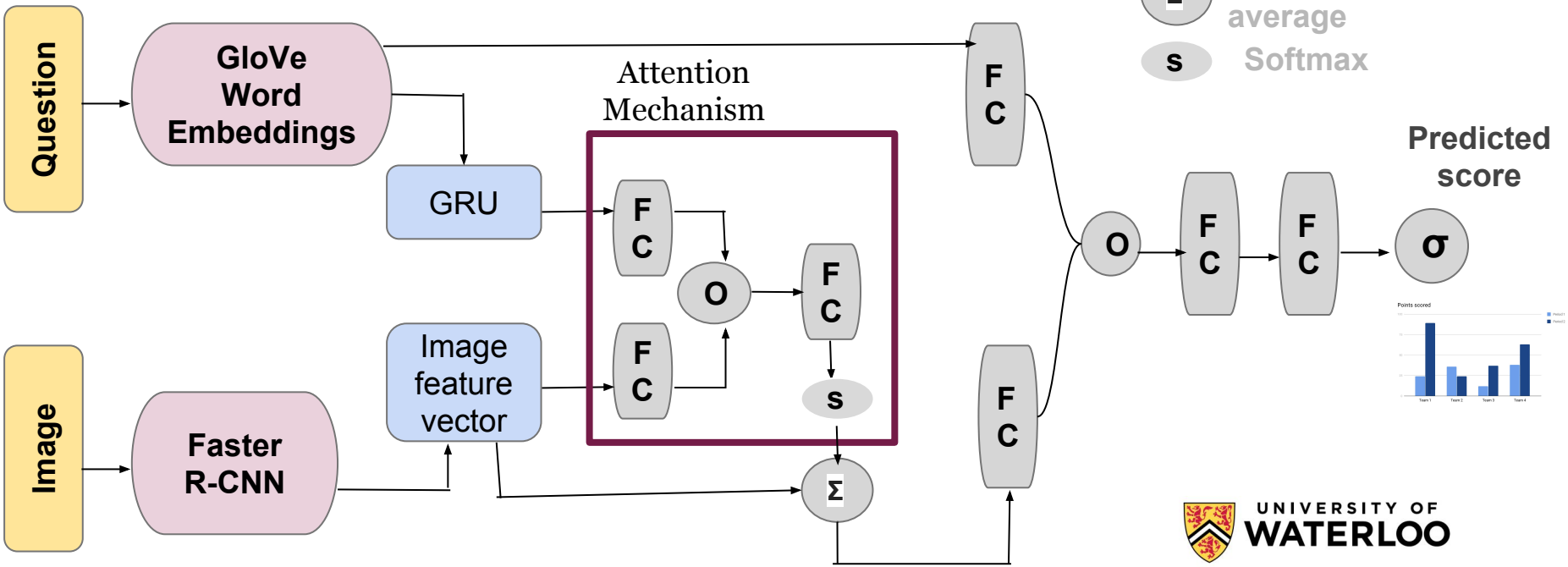
FC: Fully Connected

O Element wise multiplication

σ Sigmoid

Σ Weighted sum of average

s Softmax



Road Map

Introduction

Literature review

Goal and Methodology

Experimental Result

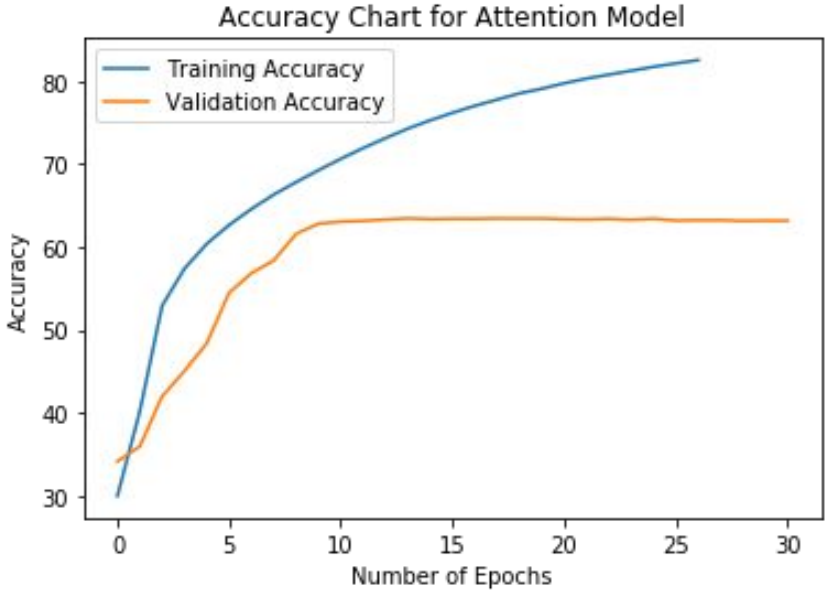
Summary of work

References

UNIVERSITY OF WATERLOO

Accuracy Charts

- Validation Accuracy:63.46 after 12 Epoch
- Training Time: Approx 8-9 mins/epoch on Single NVIDIA Tesla K80



UNIVERSITY OF WATERLOO

Road Map

Introduction

Literature review

Goal and Methodology

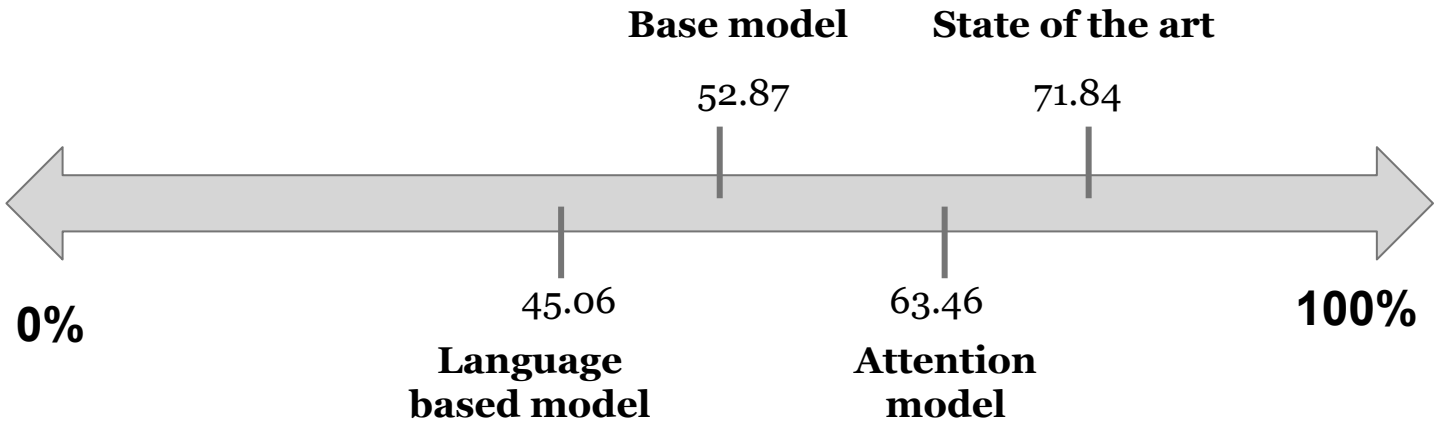
Experimental Result

Summary of work

References

UNIVERSITY OF WATERLOO

Comparison of Results



UNIVERSITY OF WATERLOO

Interesting Observation

- Difficult to predict effects of hyperparameters.
- Fancy tweaks may just add more capacity to network May be redundant with other improvements.
- By only using a fixed number of objects per image (K=36) and We don't use extra data from Visual Genome(Another dataset for VQA), we still getting similar results compared to the paper.

Challenges

- Performance ≈ (# Ideas) * (# GPUs) / (Training time)
- Experimenting with reduced training data does not translate into improved performance while training on the whole data set using same architecture and parameters
- Model tend to learn language biases.
- Model is inefficient (~39% accuracy in base model and ~49.7% accuracy) when it is asked to count something.

Future Scope

- Training on task-specific datasets may help enable practical VQA applications.
- Trying different embedding For questions
  - FastText embedding
  - ELMo (Deep Contextualized word representations)
- Bilinear Attention Networks

Road Map

Introduction

Literature review

Goal and Methodology

Experimental Result

Summary of work

References

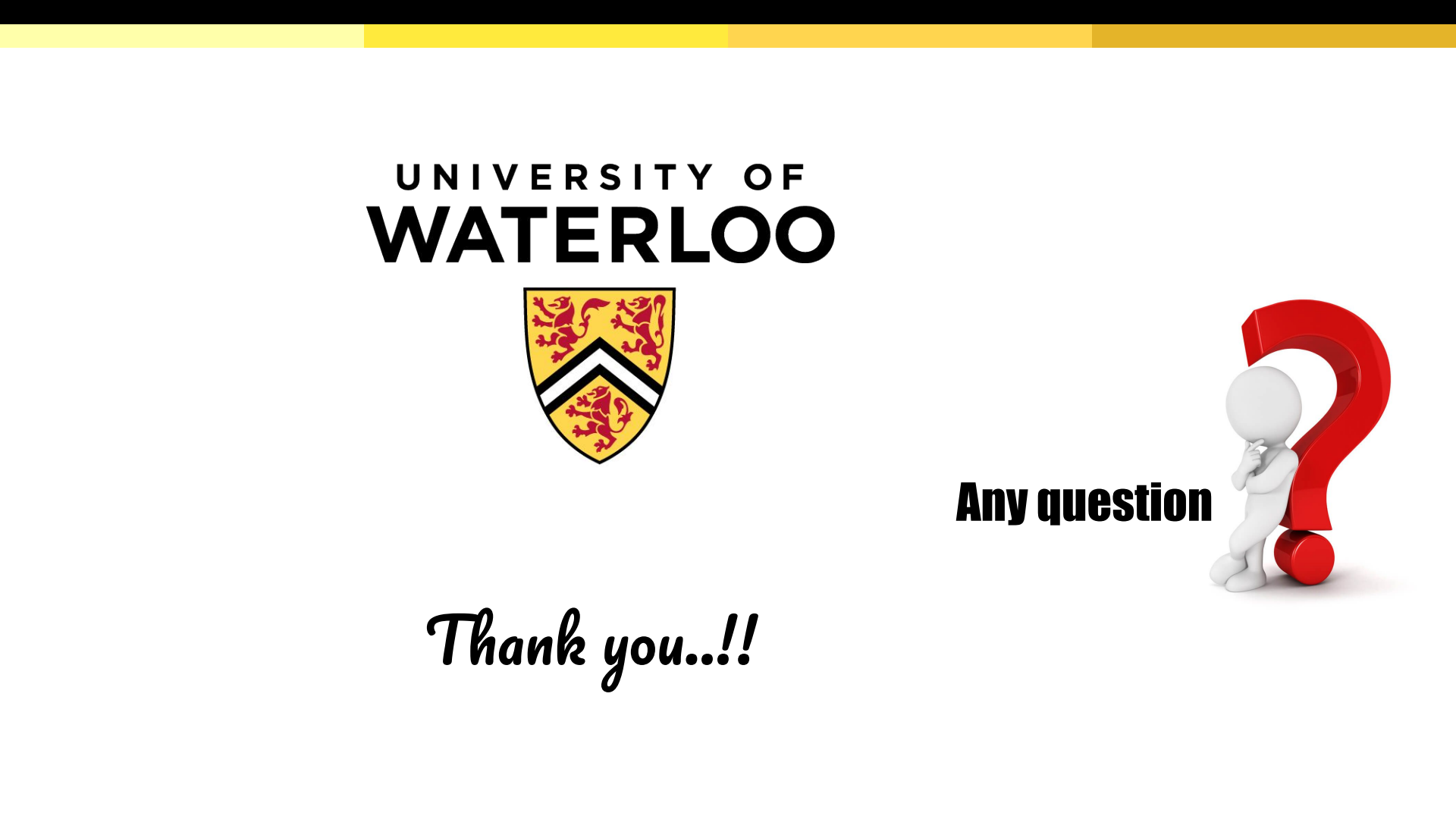
UNIVERSITY OF WATERLOO

References

- https://visualqa.org/evaluation.html
- Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge, Damien Teney, Peter Anderson, Xiaodong He, Anton van den Hengel
- VQA: Visual Question Answering, Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol.
- An Analysis of Visual Question Answering Algorithms, Kushal Kafle Christopher Kanan
- https://github.com/GT-Vision-Lab/VQA_LSTM_CNN
- https://github.com/peteanderson80/bottom-up-attention

UNIVERSITY OF WATERLOO

UNIVERSITY OF
WATERLOO



Thank you..!!