Differential Attention for Visual Question Answering

Paper: https://arxiv.org/abs/1804.00298

Team - 13:

- Chirag Parikh - 2022900005
- Neeraj Veerla - 2021121008
- Adhiraj Deshmukh - 2021121012
- Shreya Patil - 2021121009

Contents

- VQA
- Attention for VQA
- Differential Attention for VQA
- Reference Models
- DAN Network
  - Finding Exemplars
  - Triplet Loss
  - Visualizing Attention maps
  - Benchmarking on different subsets of VQA dataset
- Experiments & Datasets to be used.
- DCN Network

Overview
Visual Question Answering (The Problem)

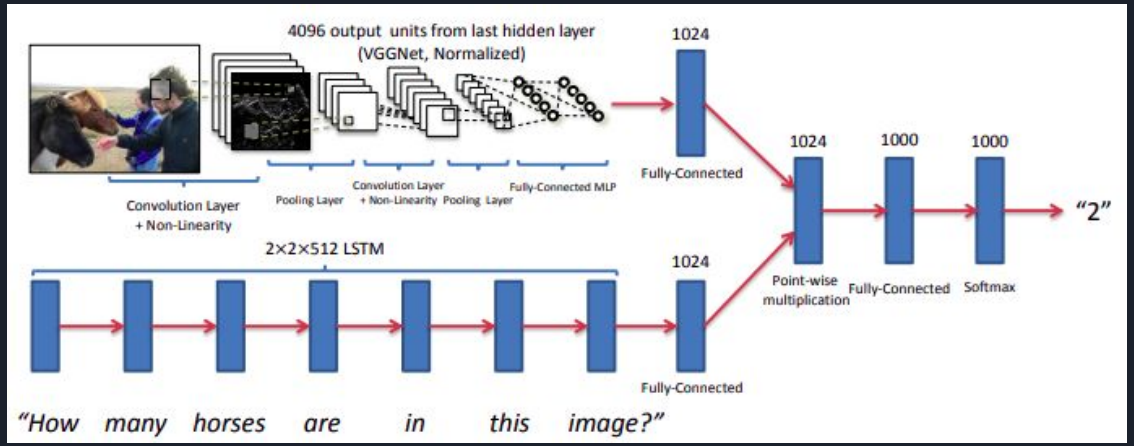
Visual Question Answering (VQA) is a task in Machine Learning that aims to develop models capable of answering questions based on visual information.

The goal is to train a model capable of extracting features from both the image and the question, then combine them to generate an answer.

VQA is a challenging task as it requires a combination of computer vision and natural language processing techniques, since the model needs to be able to:

- Understand the content of the image
- Understand the meaning of the question
- Combine the visual and textual information
- Output an appropriate answer

Basic Model for VQA



An initial approach towards solving this problem, in the deep-learning era, was by looking at the problem as a classification problem using encoded embeddings.

Generally, a soft-max classification was used over an image embedding (obtained by a CNN) and a question embedding (obtained using an LSTM).

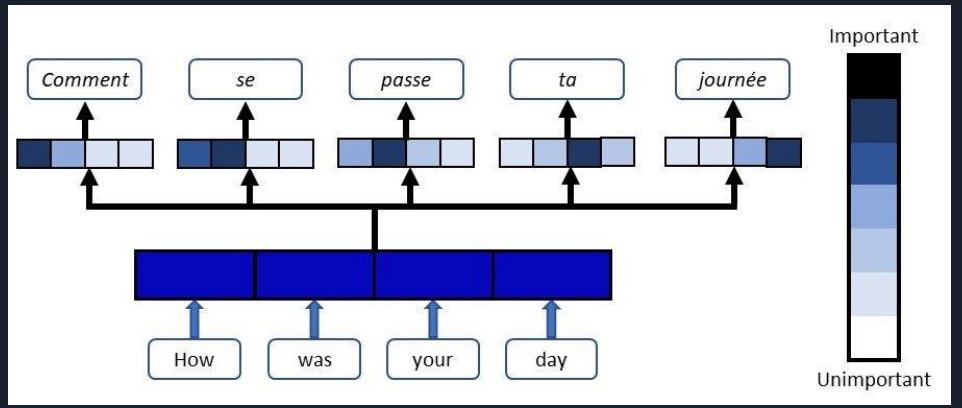
However, most of these methods are not attention based. Use of attention enables us to focus on specific parts of an image or question that are pertinent for an instance and also offer valuable insight into the performance of the system.

Attention

Attention is a mechanism in machine learning that allows a model to focus on specific parts of input data that are most relevant to the task at hand. In the context of VQA, attention can be used to help the model decide which par of the image and the question are most important for generating an answer.

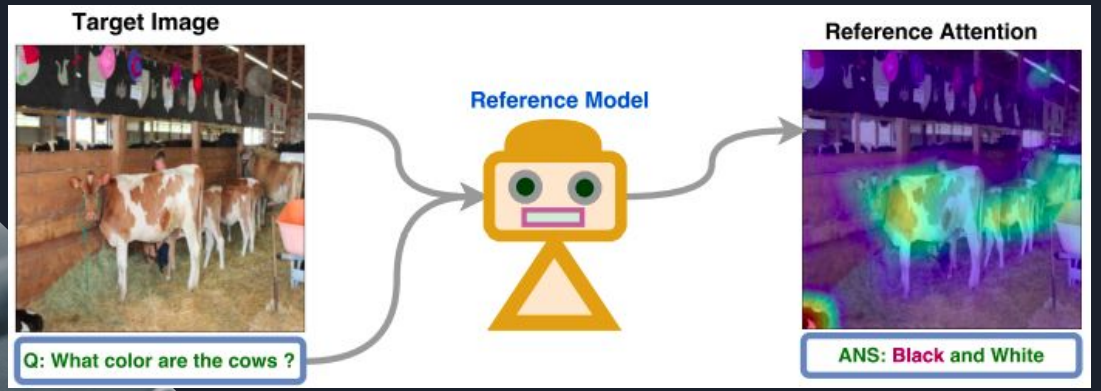
Attention can be thought of as a set of weights that determine
the importance of different features of the input data.

For example:
In translation models, the Attention component of the network map each word in the output sentence to the important and relevant words from the input sentence and assign higher weights to these words, enhancing the accuracy of the output prediction.

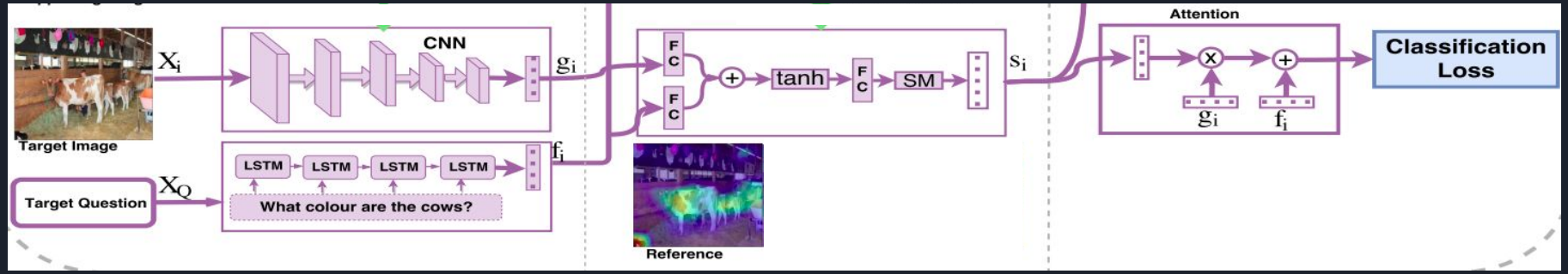


Attention for VQA

- When answering visual questions there are often specific regions a human focusses on while answering the question.
- Improve performance by use of attention, figuring out “where to look” and explicitly incorporating this information into the mode



Basic Attention Model



This is a basic attention model which uses the Image and Question embeddings to compute a single attention map over the image, and based on that we use soft-max classifier to get a prediction for the Answer.

Use of Exemplar Theory for computing Differential Attention

Exemplar theory is a concept from cognitive psychology that suggests that we (humans) categorize objects by comparing them to specific exemplars, or examples, that we have encountered before.

The model can accomplish this by comparing the objects in the image to specific examples, or exemplars, that it has encountered during training.

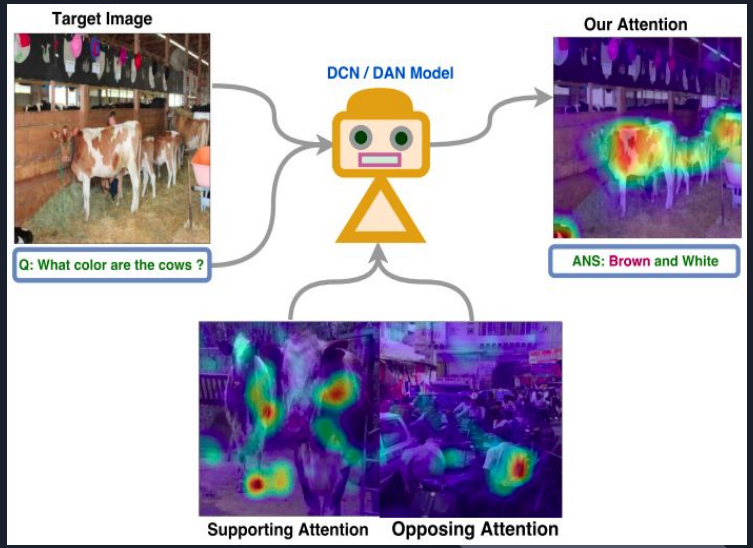
Based on this theory, the paper proposes an exemplar model which computes the differential attention over some image region by computing the difference between a nearest semantic exemplar and a far semantic exemplar.

Further, the paper claims that the obtained attention regions are more correlated to human attention regions as compared to the baseline models, and also obtains better accuracy when tested on VQA-1, VQA-2 and HAT datasets.

Differential Attention for VQA

- Since the differential attention is based on the cognitive exemplar, the theory is that it obtained attention maps should be similar to human attention relation and thus performing better for these tasks.

- And then the paper shows that this differential attention mechanism helps to obtain significant improvement in solving the visual question answering task as seen in the image.



Finding Supporting and Opposing Exemplars

- The first task of the proposed model is to find these exemplars based on feature embeddings.
- The feature embeddings based on Image level similarity does not suffice as the nearest neighbour can be very misleading as they don't have any context for the activity. Ex- children playing, visually similar to children not playing.
- In this model, **joint image-question level embedding** using CNNs and LSTMs is used to relate meaningful exemplars.
- Semantic nearest neighbours are then found using k-NN in the eucledian space.
- For obtaining the supporting and opposing exemplars, some thresholds are defined for selecting the nearest and farthest neighbours respectively.

Triplet Loss

- A distance metric which ensures that the distance between the attention weighted regions of near examples is less and the distance between attention weighted far examples is more.

$$T(s_i, s_i^+, s_i^-) = \max(0, ||t(s_i) - t(s_i^+)||_2^2 + \alpha - ||t(s_i) - t(s_i^-)||_2^2)$$

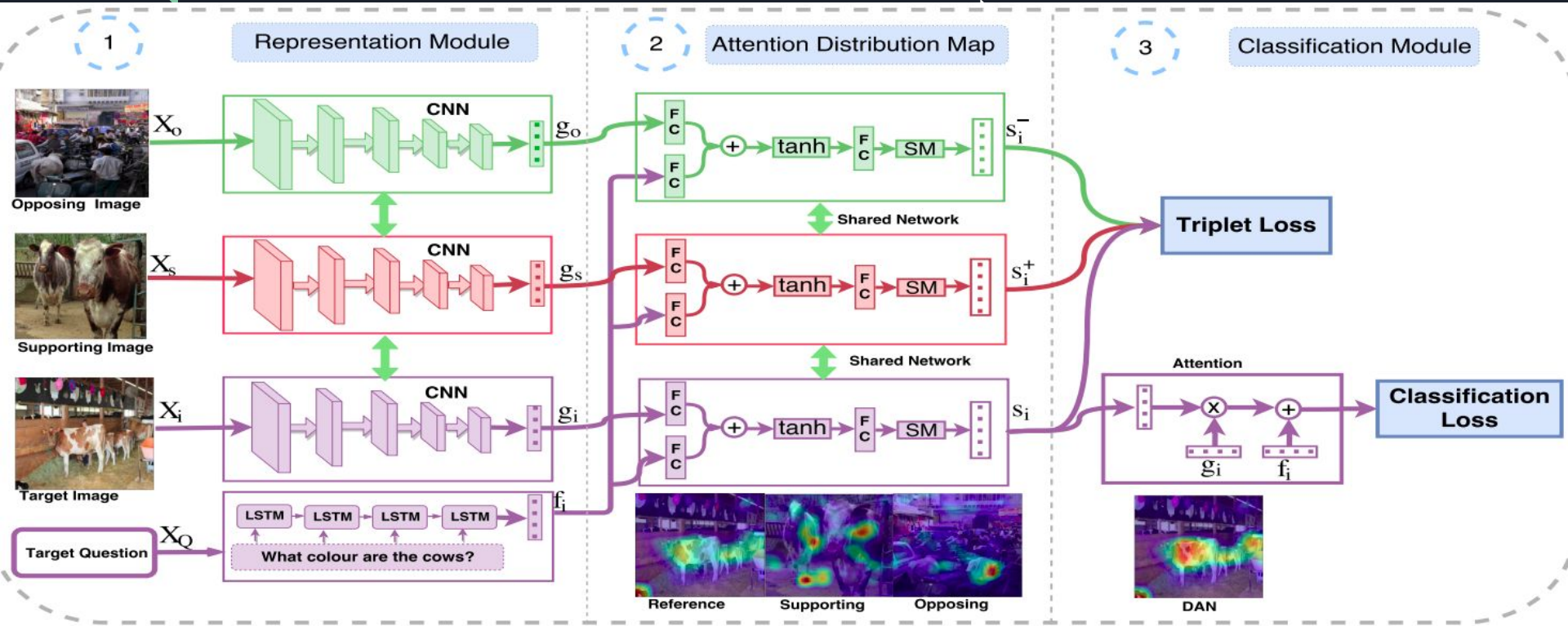
Cross-Entropy Loss

- Softmax Loss is **a Softmax Activation plus a Cross-Entropy Loss**. Softmax is an activation function that outputs the probability for each class. Cross Entropy loss is just the sum of the negative logarithm of the probabilities.

$$L(\mathbf{s}, \mathbf{y}, \theta) = \frac{1}{N} \sum_{i=1}^{N} (L_{cross}(\mathbf{s}, \mathbf{y}) + \nu T(s_i, s_i^+, s_i^-))$$

$$L_{cross}(\mathbf{s}, \mathbf{y}) = -\frac{1}{C} \sum_{j=1}^{C} y_j \log p(c_j|\mathbf{s})$$

Differential Attention Model (DAN)



VQA-v2 Dataset

- Contains open-ended questions about images which require an understanding of vision, language and common sense knowledge to answer.
- 265,016 images (COCO and abstract scenes)
- At least 3 questions (5.4 questions on average) per image
- 10 ground truth answers per question
- Total dataset size: 14GB

Experiments and Timelines

1. Implement a basic VQA model [Status: **Done**]
2. Add Single Attention layer to the basic VQA model [Status: **Ongoing**, ETA: **April 1st Week**]
3. Implement Differential Attention Network (DAN) [Status: **Ongoing**, ETA: **April 3rd Week**]
    a. Finding Supporting & Opposing Exemplars using K-nearest neighbours
        i. Experiments with different values of K
    b. Implementing point (a.) training pipeline of Reference model
    c. Implement Triplet Loss for the Attention features of the input images and exemplars.
    d. Implementing different backbones (Resnet152 and Bi-directional LSTMs for Image and Question features)
    e. Implement Triplet Loss directly for Image features
    f. Experimenting with different ways of combining the image and question features (like dot product, etc.)
    g. Experimenting with the hyper-parameters of the entire model (no. of layers, hidden layers, feature dimensions, etc.)
4. Stacked Attention Network for VQA (n=2) [Status: **Yet-to-Start**, ETA: **April 2nd Week**]
5. Visualize and Compare Attention maps of above models [Status: **Yet-to-Start**, ETA: **April 4th Week**]
6. Benchmarking performance of above models on different subsets of the VQA v2.0 dataset [Status: **Yet-to-Start**, ETA: **April 4th Week**]

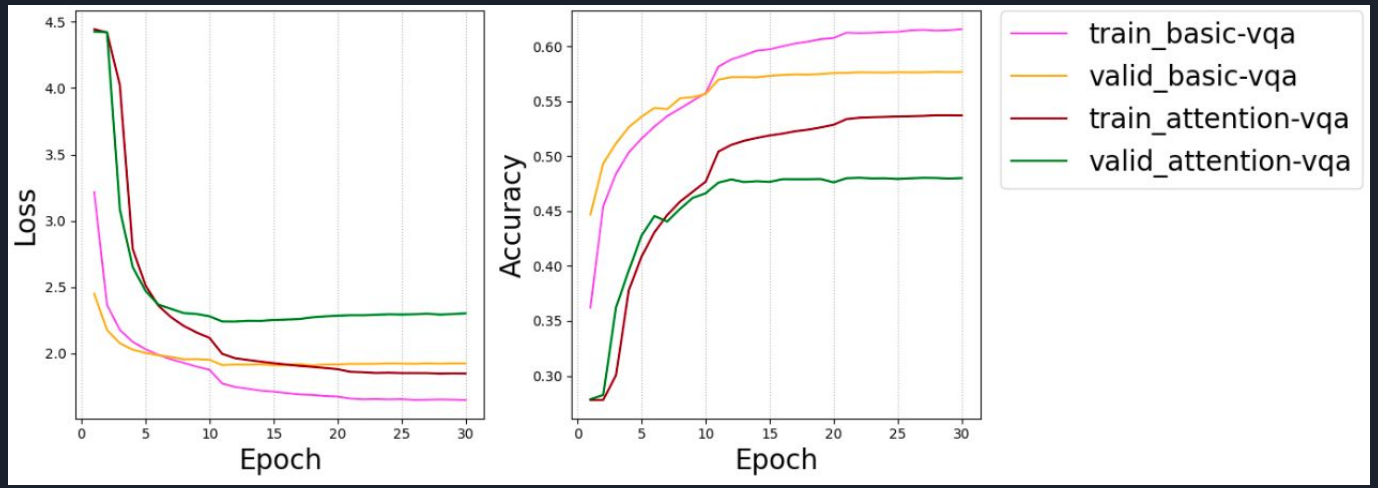
Configurations of the Implemented Models

- Basic VQA model
  a. VGG-19 backbone model pre-trained on Imagenet for Image features
  b. 2 layer LSTM with 512 hidden layer
- Add Single Attention layer to the basic VQA model (Buggy implementation)
  a. Element-wise summation of Image and Question features

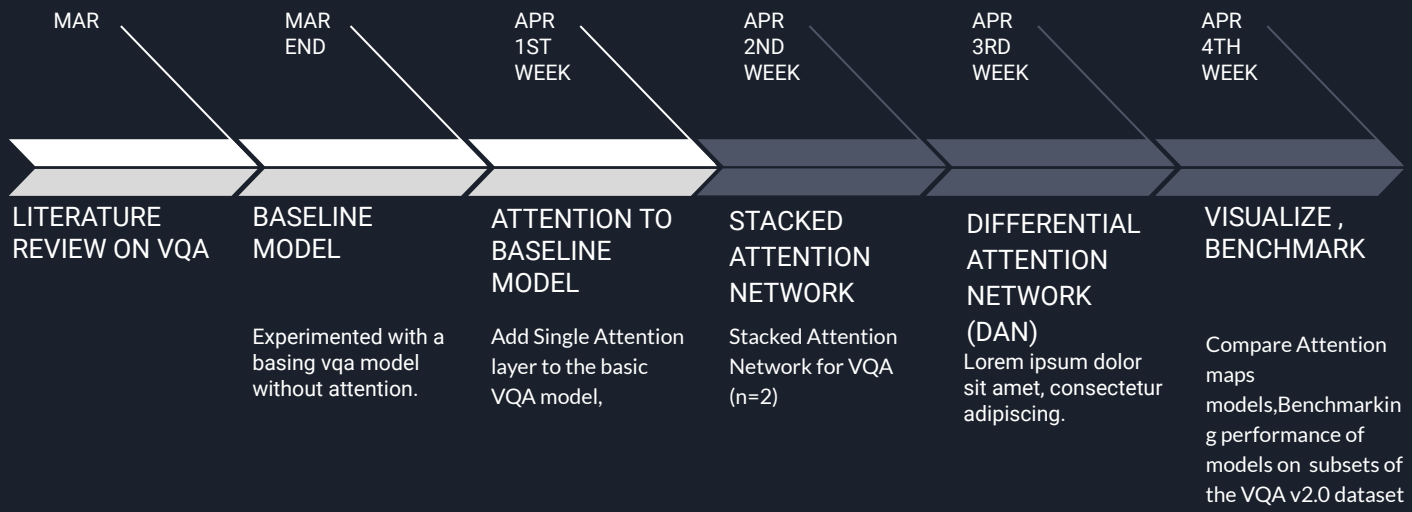
Model0	Val-accuracy
Basic model	57.55%
Single Attention model	48.50%

Compute Resources:

* We used 4 GPUs (12GB memory) for training the above models.
* 1 Training experiment took around 1.5 days to complete.



Project timeline



Thank you!

Prof - Avinash Sharma

TA’s -

- Amogh
- Pratyush