Project 1

Visual Question Answering

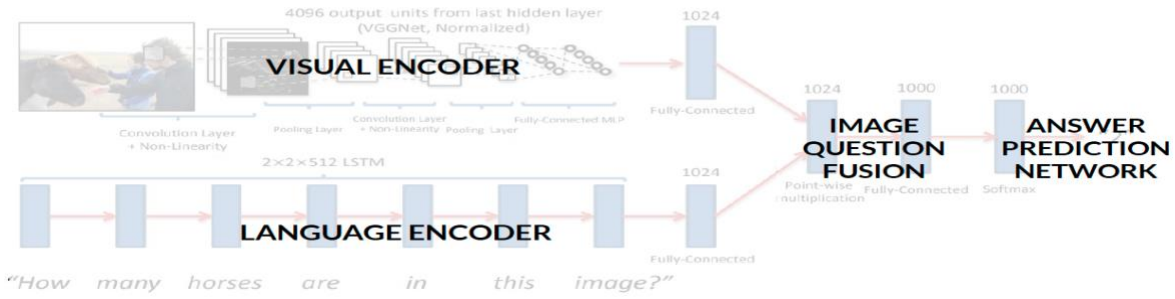
Mahmoud Khalil

(A) What does the topic cover?

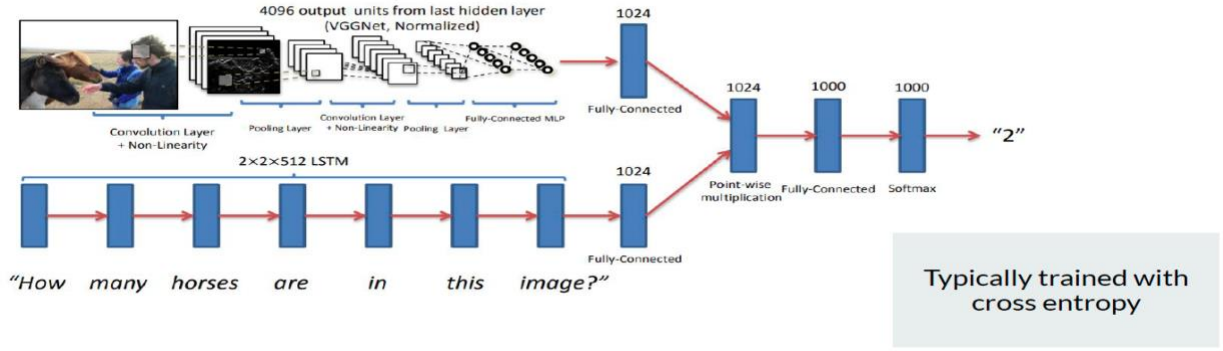
The VQA system is considered an algorithm that takes as input an image and a natural language question about the image and generates a natural language answer as the output. The broader idea of this problem is to design systems that can understand the contents of an image similar to how humans and communicate effectively about the image in natural language. The system should be dynamic enough to accept any arbitrary question. Visual question answering entails many subproblems, these include: object recognition, object detection, attribute classification, scene classification, activity recognition, knowledge-based reasoning, etc.

A VQA system that solves the above problems consists of four main parts: Image featurization for converting images into their feature representation, question featurization for converting natural language into their embeddings, joint feature representation to combine both the image and question features, and lastly, answer generation to generate the correct answer.

Overview of a VQA System:



Example of VQA process:



(B) Why it is important?

Pictures are everywhere and words are how we communicate, therefore if we can find a way of connecting these two then we have a product that can be applied to many interesting projects. Going through images is a daunting task that is required across various applications. This system can help replace the manual labor and make it easier for the user to retrieve the information needed in a shorter time frame.

(C) What are the applications of the topic? What is the societal significance of the research?

Visual question answering has many applications that include:

- Better image retrieval which in turn can helps increase the accuracy of search engines found on amazon, google, Pinterest … etc.

- Interact with, organize, and navigate unstructured visual data

- Aid visually impaired users

- Summarize visual data for analysts

(D) Challenges:

1- The primary challenge for visual question answering is the linguistics, answering a question that is grounded in an image is a crucial ability that requires understanding both the question and visual content as well as their interactions at many linguistic levels. There has been recent progress in this area which encompass the development of datasets, models, and frameworks.

https://onlinelibrary.wiley.com/doi/full/10.1111/lnc3.12417

2- A second issue is data. If you feed a model poorly, then you can only expect poor results. This can manifest itself in two ways. Lack of data and lack of good data.

Lack of data: VQA requires large amount of training to data to give useful results. However, large amount of training data requires a lot of money and manual labor. Data augmentation is useful to some extent but there are limitations.

Lack of good data: The same way that having a lack of a good features can cause an algorithm to perform poorly, having a lack of good ground truth can also limit the capabilities of the model. In the area of VQA, one major challenge is having a data set that is more representative of the world. For example, in the case of animals, there seems to be more labelled images for giraffes than any other animal.

(E) Literature Review:

Prior to starting research on VQA, the one task that peopled have looked at in the space of vision and language is image captioning, the aim of the system is, when given an image the system would provide an image caption. The issue with the setup is that image captioning is considered quite generic, one image caption could be generalized to a lot of the training data making it difficult to measure the accuracy of the model.

Therefore, in the case of image captioning:

- Image captions tend to be generic
- Coarse understanding of image + simple language model can suffice
- Passive

The next step was to actively seek the information needed from an image and that is the idea behind VQA.

Image featurization: Most of VQA literature use CNNs for this step. These include networks like: ResNet, VGGNet, AlexNet, EfficentNet, ZFNet, etc. The most popular CNN used is VGGNet across all papers published under the umbrella of VQA, however, most recently, ResNet has been the preferred CNN.

Question featurization: Older approaches include TF-IDF vectorization, count vectorization, bag of words, etc. However, the most popular approach to question featurization in VQA is LSTMs. Which is a neural network that was built to fix the vanishing and exploding gradient problem found in most RNNs. LSTM fix this issue by adding gates that manage the importance of each of the previous elements of the sequence.

https://arxiv.org/pdf/1606.00061.pdf, uses a hierarchal architecture for his model, that co-attends the question and image at three level, word level, phrase level, and question level.

Also, many research focused on a particular type of question asked:

https://arxiv.org/pdf/1511.07394.pdf, focused on multiple choice questions in relation to VQA, while https://arxiv.org/pdf/1511.05099.pdf focused on binary (yes or no) questions

Joint feature Representation:

There are different ways in handling the different feature vector coming from both the language and visual encoders, these include:

- Element wise addition and multiplication (if feature vectors are of same length) https://arxiv.org/abs/1605.02697

- Concatenating the two and letting later layers find the right weights for each

- Dot product https://arxiv.org/pdf/1511.07394.pdf

- Canonical Correlation analysis, which is a method that finds the correlation between two sets of vectors, in this case the two feature vectors. https://arxiv.org/pdf/1611.00393.pdf

- Co-attention mechanism, either by parallel co-attention or alternating co-attention. https://arxiv.org/pdf/1606.00061.pdf

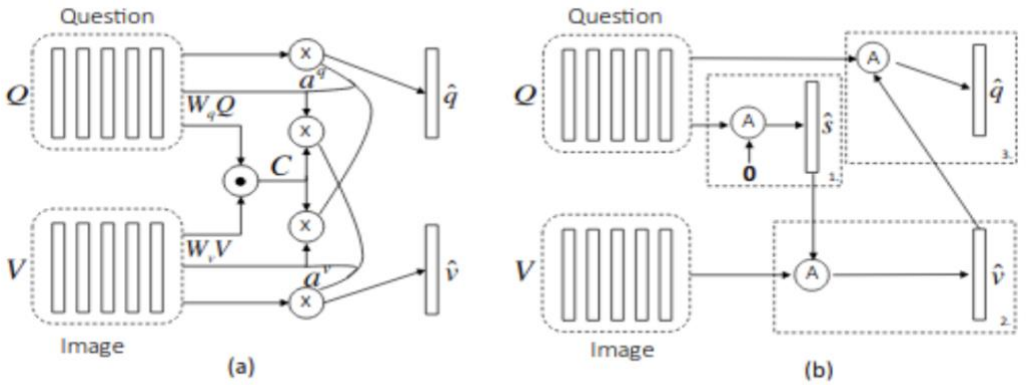


Figure 2: (a) Parallel co-attention mechanism; (b) Alternating co-attention mechanism.

Answer Generation:

Questions in VQA can be of two types, each requiring its own method for answer generation:

- Open-ended questions, the joint features are transformed into answering using a RNN such as an LSTM

- Binary questions or Multiple-choice questions, the joint feature is usually passed through a neuron layer that acts as a classification layer. Example of such layer is a sigmoid for binary and SoftMax for multi-class.

Evaluation Metrics:

Evaluation metrics used depend on the types of questions, for multiple choice or binary question a simple accuracy would be able to evaluate performance. However, in the case of open-ended questions, it is bit more complicated. Examples of metrics developed are BLUE: Bilingual Evaluation Understudy and METEOR: Metric for evaluation of translation with explicit ordering.

(F) Area of Focus:

An area that I feel VQA will add a lot of value is criminal investigations. Having a system that is able to go through thousands or millions of images and provide you with the one that have the criteria you are looking for can be very useful. It will save hours or days that people spend searching through images.

(G) Open-Source Research and Testing:

I have begun researching open-source research models that will be useful for this project. An example of one is the following: **https://github.com/jnhwkim/ban-vqa.** We will have a better idea on what open-source research we would need once we decide on what area will be our focus.

Also, CloudCv has a cloud based VQA testing environment that allows you to demo their VQA model, that I used to test a few images. **http://vqa.cloudcv.org/**