Design

* There are mainly 3 section in our VQA framework.
* Feature Extraction, feature fusion and classification
* Now feature extraction part is further subdivided into 2: Image feature extraction and question feature extraction.

Image Feature Extraction:

* Image features are extracted using a VGG16 model, which is a pretrained CNN model
* VGG16 model take images of fixed size input, 224x224x3, 3 being the number of channels
* We have used a VGG16 model here that had been trained on the Imagenet dataset.
* The full VGG16 model consists of 3 dense layers and 1 softmax at the end which outputs a 1000 valued vector
* But For our model, we do not include the final Dense and Softmax layers, and hence VGG16 outputs the extracted features of the image in a 7x7x512 array
* Now this 7x7x512 array is flattened into 1x25088 array
* This 1x25088 array contains the features of one image
* Since we can see that the size of each image feature is extremely large, we reduce the dimensions of the image.
* For this purpose principal component analysis is performed and the size of each image feature is further reduced to 1x512 from 1x25088
* So for each image we get a 512 dimension vector which is to be passed for feature fusion

Question Feature Extraction:

* For Question feature extraction, we are using GLoVe word embedding and LSTMs
* GloVe word embeddings provide a vector representation for each word in the question and LSTMs have a feature of persistent memory, which enables them to learn the relevant features necessary and more accurately compared to simple RNNs
* Now each question encoded using a tokenizer and padded to a maximum length size of 16. 16 being the maximum length of the questions in training set.
* After that we embed the question features using GLoVe word embeddings to obtain a 16x100 shaped vector,
* This is then passed onto the LSTM model to obtain a 1x800 dimensional feature vector
* Here also we perform Principal Component Analysis to reduce the dimension from 800 to 512, similar to the images.

Feature Fusion:

* Coming to the feature fusion part, we have performed the simplest feature fusion method that is concatenating both the Image and Question Features together to form a 1x1024 vector.
* This vector is then passed onto the Classification Model

Classification:

* The classification Model consists of two dense layers and one softmax layer.
* The softmax layer has 133 units, 133 being the total number of unique answers across the dataset
* This includes answers such as yes or no, organ names, disease names etc
* And finally the answer the maximum probability is chosen

Design Changes:

* Earlier, we planned on using Question conditioned reasoning and Task Conditioned Reasoning, MAML and MMQ modules.
* We have removed all these for the time being

Suggestion from Panel: