**Video Question Answering in Cooking Environment**

**Problem Statement:**

If a question is asked about a video (like what vegetables is he cutting?), then the system must be able to scan the video and generate legitimate answer to the question correctly.

**Related Work in the Past:**

Recently lot of work has been done in QA in images (Visual Question Answering); the area being dominated by Dhruv Batra and Devi Parikh’s research group.  
  
The Visual QA started with baseline of applying LSTM over textual features and CNN over image features. Currently Visual QA involves artificial intelligence and reinforcement learning approaches, namely executing specific program according to the context of the question (example, if question is who is sitting on top of the car, then the system finds the car, then looks above it and then answers the question; which would take 3 programs to implement).

Video domain has been explored comparatively in smaller scale. The most relevant work has been *MovieQA* by Antonio Torralba, and *TGIF-QA* by Yale Song.  
The MovieQA paper have used unique joint embedding representation of text and video for Question Answering.

Our problem statement involves Question Answering about actions and objects. Action Recognition is more relevant than object detection, since the structure of action recognition architectures would detect the presence of objects.

**Relevant Papers:**

Visual QA:  
1. [VQA – Visual Question Answering](https://arxiv.org/pdf/1505.00468.pdf)

2. [Inferring and Executing Programs for Visual Reasoning](https://arxiv.org/pdf/1705.03633.pdf)  
  
Video QA:  
1. [MovieQA](https://arxiv.org/pdf/1512.02902.pdf)

2. [TGIF-QA](https://arxiv.org/pdf/1704.04497.pdf)

3. [TVQA](https://arxiv.org/pdf/1809.01696.pdf)

4. [Video QA by Hierarchical Spatio Temporal Attention Networks](https://www.ijcai.org/proceedings/2017/0492.pdf)

Action Recognition:  
1. [Two Stream CNN for Action Recognition](https://arxiv.org/pdf/1406.2199.pdf) (most imp)

2. LRCN, TSN, ActionCLAD and C3D, all based on above paper.

Joint Embedding:  
1. [Unifying Visual Semantic Embedding for Images](https://arxiv.org/pdf/1411.2539.pdf)

2. [Aligning Books and Movies](https://arxiv.org/pdf/1506.06724.pdf)

**What Video QA People have done:**

The most explored datasets for Video QA have been **MovieQA** and **TGIF-QA** datasets.  
  
MovieQA has been all about movies, questions are story based, and **MCQs**. The dataset has movie clips and textual summary. The most notable element of this work is the use of **Joint Embedding Space**, which has been designed by applying transformations into a feature space, where pairwise ranking loss is minimum, or in other words, the textual features are similar to video features.

TGIF-QA has GIFs data, with max about 30 seconds, and GIFs consisting of whole scene at one go.  
The questions are open ended and multiple choice, with open ended being count of actions and about objects, while multiple choice questions being about actions and state transitions of actions.  
Resnet features for GIFs, and Glove embeddings for questions have been used.

**What we are doing different:**

We are going to work on Egocentric dataset, longer than 5 minutes of length (typically 10 to 20 minutes) and we are going to address questions about actions and interactions with objects.  
We are going to work on generating open ended answers for all questions unlike the other papers, making the problem tougher.  
The length of the video also makes the problem tougher, because we will have to localize the appropriate frames / time most relevant to the questions.  
  
Types of Questions  
*Basic Questions:*1. What vegetables am I cutting? (Finding **nouns**)  
2. What am I doing with banana? (Finding unique action **verbs** in correct order / sequence)  
3. Am I using avocado anywhere? / Did I take out anything from fridge? (**Presence** of object or action)  
4. How many times am I cutting something? (**Count** of action / verb instances on macro level without tracking / segmenting each object / noun).  
5. What did I take **out of** fridge? (**Interaction** questions between nouns).  
  
*Medium Questions:*1. What did I do before / after opening fridge? (**State transition**, with action, noun and verb)  
  
*Advanced Questions:*How many apples did I cut? (**Tracking objects** throughout video)

We are planning to tackle Basic and Medium questions initially.

**Dataset details:**

We will be using **Epic kitchens** dataset, since people have worked baseline of recognizing verbs and nouns from the dataset, and it has annotations of objects as well.

<https://data.bris.ac.uk/data/dataset/3h91syskeag572hl6tvuovwv4d>

**Procedure sketch:**

1. Generate Question answer pair for each video from action / verb annotations for all types of questions.
2. Design networks to compute Action recognition features and textual features (popular state of the art networks like TSN and skip thoughts / sentence 2 vec).
3. Go through joint embedding space code and design embedding space between action recognition features and textual features, as done in MovieQA.
4. Train to generate answers from this joint embedding space as base model.
5. Advanced model involving attention layers on Questions as well as Video content.
6. Advanced model for Advanced type of questions.

Write WACV paper, without results. (By 3rd May)