Video Based Contextual Question Answering

Akash Ganesan
akaberto@umich.edu

Divyansh Pal
divpal@umich.edu

Karthik Muthuraman
mkarthik@umich.edu

Shubham Dash
shubhamd@umich.edu

CCS CONCEPTS

• **Computing methodologies** → **Object detection**; **Natural language generation**; **Scene understanding**; **Visual content-based indexing and retrieval**;

1 PROBLEM DEFINITION

The problem we are primarily looking to solve aims at building a contextual Question-Answering model for videos. The current methodologies provide a robust model for image-based Question-Answering, but we are trying to generalize this approach to be videos. The model should also be able to handle contextual queries across the whole video. For example, if a frame has an image of a man and a cat sitting, it should be able to handle queries like, where was the cat sitting with respect to the man? or ,what is the man holding in his hand?. It should be able to answer queries relating to temporal relationships.

Initially, we shall focus on handling only yes/no questions and then extend to complex questions like What?, How?, When? and Where?.

2 CHALLENGES

There are multiple challenges which we have to deal with in order to successfully build a model which can efficiently handle queries and answer them and also preserve the context embedded in the video. First challenge would be to build a contextual linking in between the scene graphs. Consider an example where a man is eating food in one frame which is considered as a key frame and in the second frame he is driving a car. So, the link from one scene graph and the second scene graph should be linked in such a way so that the graph when traversed for finding out the answers should be representative of how to video is evolving over time. Secondly, for gauging the efficiency of our proposed model, we need an evaluation metric. One more interesting challenge is scalability. Here, we use time taken to process the video and answer queries as the the metric.

3 PRIOR WORK

Previous related work can be found in the related fields of video summarization [7], image captioning [1, 3] and scene graph generation [4, 6]. We reviewed papers that propose methods to generate key-frames of interest from a long video. A major part of our project will draw from works relating to dense captioning of images and generating scene graphs. There is a wide body of recent literature which propose novel and optimized techniques for the aforementioned tasks. Most of these rely on Deep Learning methods for object detection and captioning and Machine Learning(ML) and optimization techniques to generate the scene graphs. Finally, in terms of evaluation processes, possible test datasets and possible metrics, there are papers and open datasets that are relevant to our project [5].

4 PROPOSED METHODOLOGY

We may ask questions like, When did the person get in the car?, from two adjacent frames that has a person outside the car and one has the person inside it. So, the relationship between the frames needs to be captured. Individual questions like where an object is in a frame may be answered by scene analysis. However, it is a challenge to ask for relationships between the frames that we plan to address.

In our methodology, we input the video data and identify key frames. For each key frame, we do semantic segmentation to get different localized objects, which will serve as nodes for the scene graph. YOLO and Faster R-CNN are usually used for semantic segmentation for their speed and accuracy. We then use a dense captioning algorithm to generate captions for each frame based on [1]. Now, we can use generated captions to form a scene graph. These scene graphs from captions are generated using the algorithm mentioned in [6]. Each node in the scene graph corresponds to an object in the current frame and each directed edge represents the relationship between the objects. The next step is to link the scene graphs. This step will establish a relation between the existing graph and the incoming scene graph from the current frame. We scan the existing graph to append any new nodes (objects, attributes) and/or edges (relationships) that come up in the new frame. To answer temporal queries, we add timestamp as an edge attribute. We propose to employ graph alignment techniques and explore possible graph similarity measures for detecting scene changes. We may use this to trigger new graph generation between distinct scenes.

5 DATASET DETAILS

The datasets which are considered for running our experiments are the Visual Genome and the Youtube-8M databases. The visual genome dataset has 108,077 images which has 75,729 unique objects, 40,480 unique objects and the number of unique relationships between those object, which form the nodes for the respective scene graphs as 40,513.

The second dataset which we will consider for testing our contextual question-answering model on videos is the

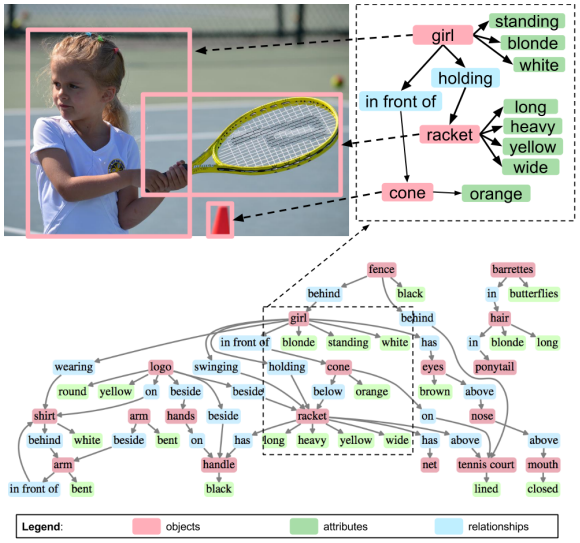

Figure 1: Example scene graph [2]

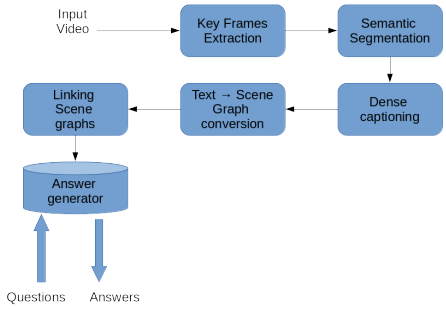

Figure 2: Model pipeline

Youtube-8M dataset. The dataset contains frame-by-frame annotations for eight million videos present.

We consider these datasets for evaluation of our captioning and scene graph generation algorithms.

6 EVALUATION CRITERIA

To evaluate the results, we will primarily rely on human evaluations as the baseline. This can be accomplished using Amazons mechanical Turk or in-class surveys. This provides a baseline to check the accuracy of the generated answers from our question-answering model. Once we have this, we have several methods existing for evaluating the performance of image based question-answering models such as Wu-Palmer Similarity(WUPS) [5]. WUPS is a similarity measure between words trained on WordNet. On thresholding it, we get different performances.

7 FUTURE WORK

Our work can be extended to incorporate speech content of video to generate more node edge combinations. This multimodal approach will make a denser graph but will store much more contextually rich information and can be used to answer much more in-depth questions. Once the graph is generated, a description text of the video can be generated. Other attributes of the object can be detected and incorporated to answer questions about emotion, expression, logic etc. Currently we focus mainly on actions and relationships but our work can be extended to emotion and inference based questions. Lastly, current video retrieval techniques rely heavily on video metadata such as video title/tags/description etc and less on the actual content/frames of the video. Extending our work, a video retrieval system can search on our representation of videos and hence the actual video content.

REFERENCES

[1] Justin Johnson, Andrej Karpathy, and Fei-Fei Li. 2015. DenseCap: Fully Convolutional Localization Networks for Dense Captioning. *CoRR* abs/1511.07571 (2015). arXiv:1511.07571 http://arxiv.org/abs/1511.07571

[2] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. 2015. Image Retrieval Using Scene Graphs. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[3] Andrej Karpathy and Fei-Fei Li. 2014. Deep Visual-Semantic Alignments for Generating Image Descriptions. *CoRR* abs/1412.2306 (2014). arXiv:1412.2306 http://arxiv.org/abs/1412.2306

[4] Alejandro Newell and Jia Deng. 2017. Pixels to Graphs by Associative Embedding. *CoRR* abs/1706.07365 (2017). arXiv:1706.07365 http://arxiv.org/abs/1706.07365

[5] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. 2016. Image Question Answering Using Convolutional Neural Network with Dynamic Parameter Prediction. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 30–38. https://doi.org/10.1109/CVPR.2016.11

[6] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D. Manning. 2015. Generating Semantically Precise Scene Graphs from Textual Descriptions for Improved Image Retrieval. In *Workshop on Vision and Language (VL15)*. Association for Computational Linguistics, Lisbon, Portugal.

[7] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. 2016. Video Summarization with Long Short-term Memory. *CoRR* abs/1605.08110 (2016). arXiv:1605.08110 http://arxiv.org/abs/1605.08110