1 Introduction

Writing algorithms which can answer questions on pictures or videos with a high level of accuracy and generality has been a goal of researchers in the AI community for many years. Recently, a lot of progress has been made in this area; with advances in neural network models and the production of larger datasets allowing reseachers to significantly improve accuracy on question answering models.

Formally, Visual Question Answering (VQA)[2]. is a task where, given an image and a question posed in natural language about the image, a model is required to produce an open-ended answer to the question. Video Question Answering (VideoQA) is a related task where a model is given a video (mulitple images in sequence) and a question. These questions can be related to a single frame of the video, effectiively making VideoQA a superset of the VQA task.

This project will attempt to produce a hybrid model for VideoQA - one which makes use of both neural networks and knowledge representation and reasoning methods based on first-order logic.

1.1 Motivation

VQA and VideoQA tasks attract attention because of their difficulty; both problems are considered "AI-complete" - they require knowledge from multiple modalities beyond a single sub-domain[1]. Building systems which have a deep understanding of the world would be a significant achievement for AI research; allowing many tasks which require significant human time and effort to be automated. Solving the VideoQA problem, which require image understanding, natural language understanding and commensense reasoning to be deployed, could be a major step towards this.

Furthermore, subproblems of VideoQA have already been shown to have applications in real-world tasks, for example event recognition has been used for identifier attacks on computer networks[4], detecting credit card fraud[7] and recognising cardiac arrhythmias[6].

Finally, the use of a hydrid model for VideoQA brings with it a nmuber of advantages. Firstly, representing knowledge in logical form allows the injection of commensense or background knowledge which can significantly improve accuracy in question answering tasks[5]. Secondly, there has recently been a significant increase in research related to explainable AI methods - machine learning techniques that enable human users to understand, trust and manage emerging artificially intelligent partners[3]. Extracting the knowledge from a neural network into logical form could be an important step toward explaining and understanding their behaviour.

1.2 Problem Description

As mentioned above, the VideoQA problem can be defined as building a model which, when presented with a short video and an open-ended, natural language question about the video, can produce a natural language answer to the given question. In our case we are looking to design a hyrid model for VideoQA. More specifically, convolutional neural networks will be used to extract knowledge from each frame of the video. This knowledge, along with the question to be answered, will be represented in a fashion that is amenable to searching for the answer to the question. This framing of the problem leads us to outline the following sub-problems, which will need to solved in order to produce a satisfactory VideoQA model.

1. **Object Detection**. Given a frame, we need a model which can produce a rough estimate (a bounding box, for example) of the location of an object in the frame. We will also need a model which can classify each detected object into a set of predefined classes.

2. **Property Extraction**. Given an object, which is the output of the 'object detection' model above, we need an algorithm which can produce a set of values for that object for some set of predefined properties. For example, we might need to give a value for the colour, size or shape of an object.

3. **Event Detection**. Given two sequential frames, we need a model which can classify the event(s) which occurred between the two frames into a set of predefined classes (possibly including a catch-all 'no event' class). This model will also be required to list objects involved in the event and what their role in the event was. This will require some level of object tracking so that it is clear how objects are related between frames.

4. **Question and Knowledge Representation**. Given the outputs of the models above and the natural language question, we need a way of representing the background knowledge, the knowledge contained in the frames of the video and the question. These must be represented in a way that allows an answer to the question to be found.

1.3 Objectives