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. 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.

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]. Ex-

tracting 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.

1.3 Objectives