Video Description using Deep Learning

Summer Undergraduate Research Award

Suyash Agrawal
2015CS10262
Computer Science
CGPA: 9.92
Mob: 9717060183
cs1150262@iitd.ac.in

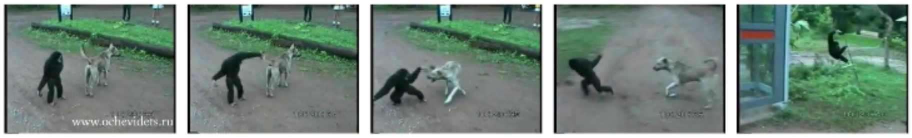
Madhur Singhal
2015CS10235
Computer Science
CGPA: 8.66
Mob: 9540972599
cs1150235@iitd.ac.in

Supervisor:-
Subhashis Banerjee
Professor
Department of CSE
suban@cse.iitd.ac.in
IIT Delhi

Prof. S. Arun Kumar
Head of Department
Department of CSE
sak@cse.iitd.ernet.in

1 Introduction

Video Description is the process of discovering knowledge, structures, patterns and events of interest in the video data and describing them in natural language. Video Description is an incredibly hard problem in computer vision and currently the only source of video description is manual labour.
Video Description has wide variety of applications. It can help visually impared people "see" the world by describing the scene around them. It has also use in automated survelliance by analysing the videos in real time and reporting criminal activites. Also, it can be used to efficiently index large video databases based upon their content for ease of accessibility.



Description: A monkey pulls a dog's tail and is chased by the dog.

Figure 1: Sample Video Description

Figure 1 shows a possible description of a sample video.

Most prior work has been done on generating natural language descriptions from images. Past two years have seen a lot of breakthroughs in image captioning problem but so far this has not been used for generating descriptions from videos.
Our aim is to use these new techniques of image captioning and use them to generate natural language descriptions from video data. In short, we will first encode video data using CNNs and then use RNNs to generate descriptions of these videos. Further, we will also explore using attention models to improve description of videos.

2 Objectives

Our main objective is to understand videos and generate natural language descriptions of them. This task can be further subdivided into following subtasks:

1. To ready all available datasets and organize them into training, validation and testing sets.

2. To construct CNNs with pretrained weights for image captioning.

3. To construct Long Short Term Memory (LSTM) Encoder Decoders.

4. To train the network using the test set data and use validation set data for setting hyper-parameters.

5. To benchmark results and improve using techniques like attention modelling.

3 Basic Concepts

3.1 Camera calibration

The camera parameters can further be subdivided into intrinsic and extrinsic parameters. **Camera intrinsic parameter** K is dependent on the focal length of the camera and principal point (which in most cases is the center of the image). The **camera extrinsic parameter** is composed of the rotation R and translation t between camera coordinate system and the world coordinate system. Together they form the camera projection matrix P, a 3×4 matrix which describes the mapping of a pinhole camera from 3D points in the world to 2D points in an image.

$$P = K[R|t] \tag{1}$$

3.2 Sparse 3D reconstruction

Given two different images of the same scene from different angles, the position of a 3D point can be found as the intersection of the two projection rays which is commonly referred to as **triangulation**. For this first point correspondences have to be established. Then using this point correspondences a Random Sampling Consensus (RANSAC) based voting framework is used to estimate the camera intrinsic and extrinsic parameters. Finally, a joint non-linear optimization is used to further refine the camera parameters and the 3D points in a **bundle adjustment** framework. This method is computationally very expensive and hence done only for very sparse set of points. This is known as sparse 3D reconstruction.

4 Mobile IMU sensors

IMU (Inertial Measurement Unit) sensors are on-chip devices embedded in most of the smart phones or hand-held devices today. It mainly consists of a series of motion sensors: accelerometer, gyroscope, magnetometer and gravitation sensor. The data from these sensors can be fused to obtain the orientation and the position of the device in the world coordinate system.

5 Conventional versus mobile 3D reconstruction

In the case of a smart-phone or any hand-held device having a camera and IMU sensors, we wish to use the IMU sensors to obtain extrinsic camera parameters in real time. This will help in reducing the load on conventional 3D reconstruction methods and get it in near real time.

6 Approach to the project

First, we shall be implementing image captioning in order to be able to encode individual video frames in fixed-length vector represenation of it. Then, we will be using our constructed CNN to encode video frames sampled at a fixed interval. We will then use RNNs to translate this representation of video into natural language domain. This translation will be achieved by using a set of encoder and decoder RNNs. Since we are working with variable length input and output, we will specifically be using LSTMs for encoding and decoding pursposes as they have been proven to be excellent in machine translation and generalize very well on long data input.

1. Image Captioning
   (a) Construct a VGG/Inception V3 with pre-trained weigths on object classification.
   (b) Make an LSTM with CNN translated image as input that will be used to translate the image representation into natural language text.
   (c) Train the model on datasets like MS COCO.
2. Video Representation
   (a) We will first sample video at a fixed rate and convert each individual frame to a fixed length vector representation using the CNN trained in image captioning part.
   (b) We will then try out different approaches of video representation like:
      - Averaging over all video frame encodings obtained to get a fixed vector representation of video.
      - Using RNNs to encode this variable length video representation.
3. Natural Language Conversion
   (a) Based on how we choose to encode our video, we will have to select appropriate archietecture of LSTM to be used to decode the representation

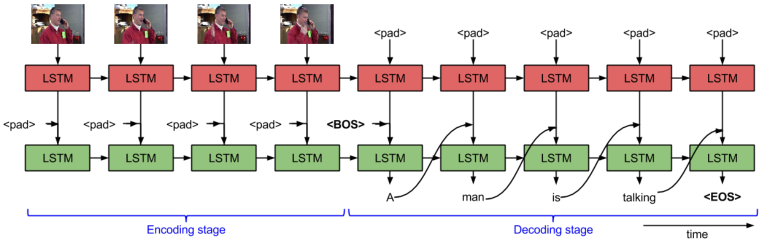


Figure 2: Video description model with 2 LSTM levels

   (b) One popular choice(Figure 2) that we will try out first will be to use a two level LSTM model that will do a sequence to sequence mapping from variable length video representation to variable length natural language sentence.

   (c) Then we will train our model on the training data we have obtained and plot the learning curves.

   (d) We will also have to check for over-fitting and under-fitting during our training process and finetune our hyper parameters according to it.

4. Further Possibilities

   (a) We will look for techniques of data augumentation and transfer learning to compensate for the limited amount of training data for video description.

   (b) We will also look into some recent techniques of using optical flow for attention modelling which has shown in some cases to improve the results of action recognition.

   (c) We will also consider using more efficient vocabulary representation than one hot encoding as vocabulary in this case would be very large.

   (d) Try to make forward propagation fast and more memory efficient.

   (e) We will also try to develop an end user application that will speak out description of a video that a person shoots.

7 Datasets

Though the source of video description datasets, we have found some viable options that can be worked upon. Even though these are not very large datasets (like MSCOCO) but some research has been done using these and we are confident that we can work using these datasets. Some of them are:

1. **Microsoft Video Description (MSVD)**: The dataset contains 41.2 hours and 200K clip-sentence pairs in total, covering the most comprehensive categories and diverse visual content, and representing the largest dataset in terms of sentence and vocabulary.

2. **MPII Movie Description Dataset (MPII-MD)**: MPII-MD contains around 68,000 video clips extracted from 94 Hollywood movies. Each clip is accompanied with a single sentence description which is sourced from movie scripts and audio description (AD) data.

3. **Montreal Video Annotation Dataset (M-VAD)**: The M-VAD movie description corpus is another recent collection of about 49,000 short video clips from 92 movies. It is similar to MPII-MD, but only contains AD data and only provides automatic alignment.

8 Uses and applications

- Assisting visually impared people to get description of their surrounding, thus enabling them to ”see”.
- Very useful for automated survelliance and theft detection but being able to analyze large amounts of data that is not viewable by humans.
- Allowing content based video retrieval by describing the contents of video in textual format which is indexable by web crawlers.
- This can also be used to detect catastrophic events through security cameras like fire breakout, murder etc.
- This project can also be applied in helping robotic vision as this project basically allows one to understand what is happening in the video and thus robots will be able to get a ”true” sense of their surroundings.

9 Budget and duration

9.1 Budget

No budget is required.

9.2 Duration

We will try to complete this project by the end of the summer break i.e. the end of July, 2017.