Chapter 68
Neural Semantic Video Analysis

Hamid Mohammadi
https://orcid.org/0000-0002-8854-9342
University of Alberta, Canada

Tahereh Firoozi
University of Alberta, Canada

Mark Gierl
University of Alberta, Canada

ABSTRACT

Videos are a rich form of data intended for capturing, storing, and communicating information. The availability of inexpensive and accessible video-capturing sensors in smartphones, handheld cameras, and consumer security cameras has exponentially increased global video footage generation over the past decade. Since video is a popular form of widely consumed and produced data, it is essential to develop automated systems to analyze and identify relevant information within the large body of video material. This chapter demonstrates how the emergence of neural networks, including CNNs and transformers, has revolutionized semantic video analysis. Through convolutional filters, spatial patterns can be captured at the pixel level through this type of neural network. The learning capability of CNN-based models has been exceeded more recently by self-attention-based models. Both CNN-based and transformer-based semantic video analysis models take advantage of transfer learning, self-supervised learning, and more to compensate for the lack of large, supervised video datasets.

INTRODUCTION

Semantic Video Analysis (SVA) is the process of extracting conceptual information from raw visual data. In any video, whether it is a production-level movie, a smartphone video clip, or drone footage taken at high altitudes, information is not explicitly and concisely represented in the video pixels. The information can range from the event or action in the video to landmarks, objects, or people appearing in the video. Obtaining such information requires an understanding of the semantics behind video pixels' simple color

DOI: 10.4018/978-1-6684-7366-5.ch068

representations. In this chapter, we review the challenges of SVA and the methods and techniques commonly used to overcome them. The general utility and density of video information make it an effective data format for storing and communicating a variety of information. Since video data formats have this valuable property, they have been adapted to a wide range of information communication and storage mediums. Whether it's for video calls, social media, or security systems, videos play a crucial role in our daily lives. Thus, a massive amount of video is produced every minute around the world. Considering the usefulness and wide application of videos, SVA is an important topic in science and engineering.

Despite all the benefits of SVA, extracting useful information from videos remains challenging and ongoing. Videos can be long and noisy, as well as sparse. For example, finding particular information in low-quality recordings of an hour-long classroom video can be challenging and tedious. This is because a SVA system must find a match for a particular pattern of information among thousands if not millions of meaningful patterns to find the intended information. Pattern matching is made more difficult and time-consuming when the data is noisy and of low quality. Naive and simplistic approaches to SVA, therefore, result in inapplicable or ineffective results. An intelligent approach that is based on visual semantics is required to comprehend the volume, complexity, and sparsity of videos.

Traditional SVA relies on human intelligence to search for and extract meaningful information from videos. We have evolved to recognize and search for useful visual information. In order to incorporate useful visual information into their intelligent behaviors, humans are proficient at conditionally searching for that information. However, human intelligence is limited in terms of accuracy and scalability. For example, reviewing YouTube videos for possible violations of content policies might take an individual 156 million hours[1]. This is equivalent to 17,000 years of intensive work. This assumes the full concentration of the human agent for maximum accuracy. On top of that, content creators upload 720,000 hours of video on the YouTube platform every day[2]. Performing this task using human intelligence is time and cost inefficient.

METHOD

The core promise of machine learning is the automation and scaling of human intelligence. Neural networks, as the most powerful machine learning tool, are utilized to provide viable solutions for the challenges of SVA. Learning spatiotemporal patterns from sample videos enables neural SVA to recognize complex spatial and temporal patterns in noisy and long videos. Modern neural SVA architectures are capable of human-level performance in this regard. Furthermore, these models can be efficiently computed using modern graphical processing units (GPUs), so scalability is limited only by the available GPUs. Hence, neural networks are the preferred solution for SVA due to their accuracy and scalability.

For machine learning models to function properly, it is critical to understand how images are represented and stored on computers. RGB, short for red, green, and blue, is the most common format for representing images on computers. Typically, RGB images are stored as a matrix of pixel values in a 2D plane (or 3D, if RGB values are considered as dimensions). Each image is composed of pixels arranged in rows (across its height) and columns (across its width). The intensity of red, green, and blue colors is indicated by three values per pixel at each X (horizontal position) and Y (vertical position). Different combinations of red, green, and blue values produce a variety of colors. Essentially, videos are just extended versions of images. So, videos can be regarded as a succession of images across time. Video adds a third dimension to visual data by adding a temporal dimension. The added Z dimension indicates