Technical research & algorithm design

30th Mar. 2021, Team Cinefly

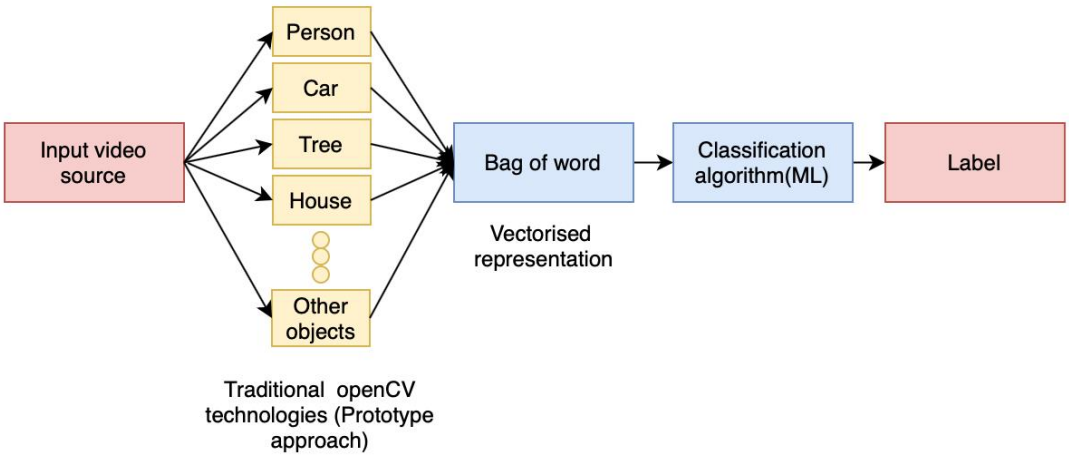
1. Research

1.1 Traditional video classification algorithms

Before the rise of deep learning, most video classifications were based on hand-designed feature-extracting and typical machine learning methods.

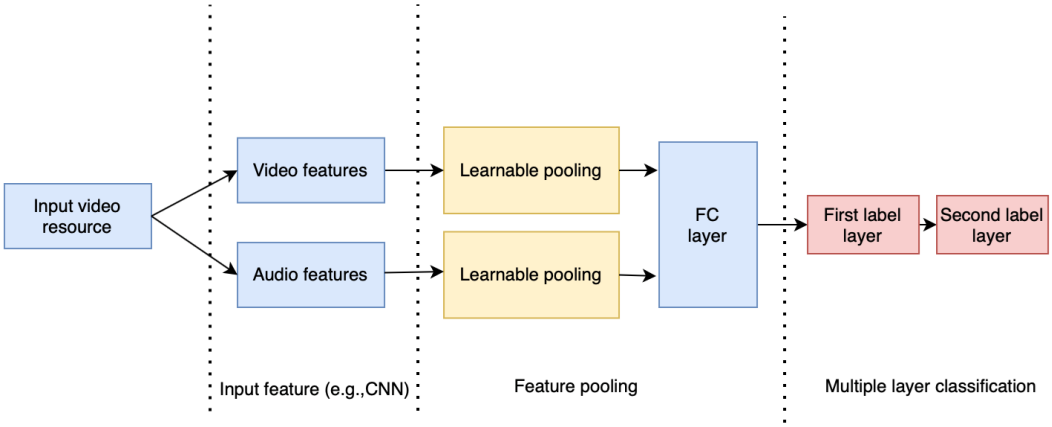
What does it mean?

For example, in traditional methods, we can segment video into many component objects based on their color, pixel coordinates and timeline locations. There are many approaches to do that, for example, the HOG、HOF、MBH algorithm. Then we can use a vector embedding approach to transform these objects into vectors. Let's say if we have extracted 100 vectors from a single video, then we can use "Bag of words" algorithms to generate one big vector to represent the whole video. After getting the video vectors, we can perform traditional machine learning classification methods like K-means and logistic regression to get the final labels.



1.2 Video classification algorithms based on deep learning.

In practice, the efficiency of traditional methods is not high, and CNN, a typical algorithm in deep learning, is very extraordinary in image recognition, segmentation, detection, and retrieval. When using the CNN method, there is no need to manually specify the video features that need to be extracted. That is to say, we only need to take the entire video as input instead of manually extracting specific objects. This processing method improves accuracy. For example, a feasible algorithm is NetVLAD.



2. Algorithm design

This is basically the algorithm design framework of our project. This framework is based on some existing video processing algorithms which can be used to classify limited types of videos. As shown in the figures, there are two ways to implement our project goal, one is using a traditional approach and the second is to use deep learning algorithms. In this semester, we plan to at least deliver a quick prototype using a traditional approach. If everything goes well, we will look at the deep learning algorithm to improve the performance.

Next, I will introduce the traditional approach we are exploring now. As you can see from the picture, firstly, we will extract objects from the given videos using OpenCV technologies. Then we will vectorize these objects using some existing embedding technologies. After that, the bag of words algorithm will be used to generate a single vector representation of the whole video. This single vector can be viewed as a representation of the theme of the video. We will further apply some classification methods to determine the actual theme from the vector representation. The actual theme will function as the label we want. Currently we can extract objects from videos in real time. This will be demonstrated later.

References:

Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C]//Thirty-First AAAI Conference on Artificial Intelligence. 2017.

Arandjelovic R, Gronat P, Torii A, et al. NetVLAD: CNN architecture for weakly supervised place recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 5297-5307.

Arandjelovic R, Gronat P, Torii A, et al. NetVLAD: CNN architecture for weakly supervised place recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 5297-5307.