Improving Video Search with Automated Tagging

Project Proposal CSC - 680 Graduate Capstone in Computer Science

Andrii Sherman

Supervisor: Professor Saleh Aliyari

Department of Mathematics & Computer Science

Abstract

Search has been an important part of accessing information on the web since the internet has existed. Starting from simple catalogues and relevancy term searches to the ranking methods which display the most relevant information to the search user. The problem for many of those situations has been that it was only suitable for searching and indexing of text data. As more information is getting shared in video form, searching for that becomes harder. The main source of that information has been text information: video title, description, tags. With the rise of computer vision it is now possible to extract information from images, specifically video frames, and use that information for search instead. In this project, I make use of YouTube-8M, a video classification benchmark, to get appropriate video labels, and use them to construct a video search engine. I will compare how well the search performs using only the user-entered tags, the autogenerated labels, and both together, and propose whether automated tagging system could be used to improve video search in general.

Background

Automated tagging has rapidly improved image search, with massive collections such as ImageNet being attributed to the increase in quality of automated tagging. These sorts of systems have improved the ability for people to search for images, as automated tagging greatly improves the effectiveness, because without them the only information that could actually be searched is the metadata associated with the image. Automated image tagging therefore increases the amount of available metadata by describing what is in the image.

While such improvements have been happening for images for quite some time, it has been a longer time for such collections to come for videos. Like for images, automated tagging of videos only started to take off after large-sclae datasets were made available. The dataset used in this paper, YouTube-8M contains ~8 million videos (500k hours of video) that is annotated with 4800 visual entities. This dataset has been the target of 3 challenges, hosted by Google, that has provided increased understanding of how to effectively work automated video tagging.

I've done previous work on this exact topic back in 2016, before Youtube-8M was released. The biggest problems with my work from that time had been with a lack of any publically available tagged video dataset to be used for evaluating my model, and as such most of the time back then had to be devoted to gathering and tagging video files and keyframes. Because YouTube-8M now exists there is now no need to go through the same process, and my work can continue from evaluating whether the use of automated tagging.

Project Description

The goal of this project is evaluating whether the use of the use of automated video tagging can be used to imrove video search. This can be separated into 2 distinct parts:

1. Efficient tagging of videos based on frame-level information

   Youtube-8M provides 2 different levels of models - a frame-level one, and a video-level one. The frame-level model is created based on feature extraction from images, in this case frames of a video. By using this frame information efficient labels can be created for the whole video. This can then be evaluated against the video-level model also provided by YouTube-8M.

2. Determine if search is improved by the automatic tagging

   After generating a video-level tag I can evaluate whether the use of such a model will improve the ability to search for this video. The ways I'll evaluate will be covered in the testing section down below.

Project Stages

Hardware

The hardware used will be limited to what I have readily available, which means that the model will have to be trained in a reasonable timeframe on a single machine, as I have no access to multi-machine ML models. This does have real-world advantages though, because the resulting model also has to be small in practice for timely delivery to the end-user, so the advantages of multi-machine training will likely not come into play as much.

Dataset Gathering

As already mentioned, the dataset used will be the YouTube-8M dataset. The biggest problems for using this information is just the amount of size it has. Because this contains metadata for both frame and video-level models for 8 million videos, the amount of information is on the order of 1.6 Terabytes. Just downloading that information has proven to be a challenge.

Software packages

The main package to be used in loading the information, training and evaluating the models will be TensorFlow. Other python packages will be included on a need basis, to aid in processing the tags and evaluating the search improvements.

Software development

The development of this project can be separated into 3 distinct parts:

1. Developing a video-level model
   Using available frame-level information from videos, generate a set of tags best describing the video. Compare those tags with the tags in the video-level model provided by YouTube-8M to evaluate its performance.

   Subgoals:

   1. Use a starter model provided by Google in order to unserstand how to work with this dataset
   2. Read through some of the publications from the 3 challenges in order to get an idea of what sorts of models could be used in making these models
   3. Implement some of the model architectures, train and evaluate them on smaller subsections of the dataset
   4. Choose the best perfoming model, and train and evaluate it on the whole dataset

2. Use the model to automatically tag videos

3. Test whether the use of automated tagging will result in better video search
   3 Competing sets of tags will be used to evaluate search performance:

   1. Just the automated tags

   2. Just the user-provided tags
      Many people who upload videos provide them with a good set of metadata, from the description, to the title, to tags. As these are the features that are used for video search today, this is the benchmark to compete against

   3. The user-provided tags augmented with the automated tags. This will