Capstone Project Proposal: Machine Learning Tools for Film Editing Automation

José Marcelo Sandoval-Castañeda

Project Summary

Abstract

The purpose of this project is to develop tools that use machine learning principles and their capabilities to aid in the process of film editing. More specifically, it will address the subprocesses of tagging metadata, writing descriptions, and identifying main themes and keywords within media content in both audio and video. These tasks are conventionally done by humans at the beginning of film editing to facilitate the rest of the editing process through searching and classifying media. This will hopefully be achieved through using state-of-the-art tools tailored and modified for the purposes of filmmaking, such as Fast R-CNN for object detection and Latent Dirichlet Allocation (LDA) for topic summarisation. If the set of tools to be developed is able to do these tasks reliably, it could potentially save hundreds of man-hours in the film editing process, proving really useful for independent filmmakers as well as more established studios.

Introduction, Specific Aims, and Background

Object recognition in video is a field that has seen significant development over the past few years. There are plenty of systems that offer this kind of service to significant accuracy. However, most of these systems are not aimed at the film editing process, and present significant limitations when applied to this area. In particular, most of these tools tag every element in the frame and do not have a way to prioritise the most important objects within a video, which is particularly detrimental to the tagging process in film editing.

On the other hand, speech recognition is an area that has been thoroughly researched in the past, and especially the topic summarisation from speech is a key piece of this project. Through Latent Dirichlet Allocation one can reliably summarise the main topics from text, for which now a substep must be added before LDA for a reliable speech-to-text model. For this, state-of-the-arts tools such as recurrent neural networks for language modelling (RNNLM).

Again, although these are all pre-existing tools, there are several key components that are either missing or not optimised for filmmaking, which would be what this project is about. In particular, cross-referencing the tags generated from topic summarisation and the ones in object recognition could potentially not only save time and money in film editing, but also provide creative suggestions on how to edit a particular film piece.

Goals and Potential Impact

In the film industry it is not uncommon to work with large amounts of media which will eventually be cut down during the editing process. The most significant example of it is the movie Apocalypse Now, shot before digital film existed, which consisted of 1.5 million feet of 35mm film, which is roughly translated to 300 hours of footage for a movie that ultimately lasted about 2 hours and 33 minutes. This is not just the case of big budget movies, especially with the popularisation of digital media. The documentary The Color of the Current (by Marcelo Sandoval-Castañeda) had about 60 hours of raw footage before being edited down to its 10 minutes final cut. A rule of thumb in fiction filmmaking is to have an hour of raw footage per minute of the final cut, which makes the editing process extremely important in narrowing down the useful footage.

All of this footage has to be tagged and thoroughly described for the editing to be somewhat efficient. This obviously requires a person or group of people to spend time watching and writing these tags and descriptions, which would at least take as long as the amount of existing raw footage, assuming only one watch was needed to describe a video appropriately. In reality, videos have to be seen multiple times before tagging and describing audio and video, taking up significant man-hours in the editing process. Furthermore, as the size of the project increases, and the amount of people working in metadata tagging and describing, imprecisions start to arise, as some people may deem one aspect important where others would not.

The most important impact this project could have is to reduce these to a minimum, where this tagging and describing is mostly automated and not reliant on human input. This is particularly impactful for independent filmmakers that cannot necessarily afford to pay that many people and that much time in a studio, for instance. However, larger studios may also find this tool important in making their film production process much more efficient.

Methodology

The media is first separated between audio and video through a library called MoviePy. Then, each of these undergoes a completely different process.

For audio, the newly created audio file will go through an RNNLM that converts the relevant parts that are speech into text. Afterwards, LDA is run to tag the text into its main topics that will eventually become the tags.

For video, the file will go through a Fast R-CNN that will be able to identify all the possible objects across the video. After that, depending on the number of appearances of an object and the amount of space it occupies on the screen, the most important tags are conserved while the rest are discarded.

The hardest aspect of this problem is probably accessing the large amounts of data that will be required for these different neural networks to train, which will be partially obtained in collaboration with the NYUAD Film department and the footage they have used for real projects, and by using a number of open and paid libraries containing immense amounts of footage that is already tagged.

Moreover, the relevant tags generated through this process have then to be outputted in a format that can at least be understood by the top three non-linear editing software in the market: Adobe Premiere, Avid, and Da Vinci Resolve.

References

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. “Latent Dirichlet Allocation.” UC Berkeley, 2003.

Han, Kyu J., Akshay Chandrashekaran, Jungsuk Kim, and Ian Lane. “The CAPIO 2017 Conversational Speech Recognition System.” Cornell University, 2017.

Hardesty, Larry. “Computer Learns to Recognize Sounds by Watching Video.” *MIT News*. MIT News Office, 2016.

Miranda, Eduardo Reck. “An Artificial Intelligence Approach to Sound Design.” *Computer Music Journal*, 19:2. Pp 59-75. Massachusetts Institute of Technology, 1995.

Mobahi, Hossein, Ronan Collobert, and Jason Weston. “Deep Learning from Temporal Coherence in Video.” *Proceedings of the 26th Annual International Conference on Machine Learning*. Pp 737-44. University of Illinois at Urbana-Champaign, 2008.

Pathak, Ajeet Ram, Manjusha Pandey, and Siddarth Rautaray. “Application of Deep Learning for Object Detection.” *ScienceDirect*. Pp 1707-17. International Conference on Computational Intelligence and Data Science, 2018.

Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.” Cornell University, 2016.

Tao, Qingyi, Hao Yang, and Jianfei Cai. “Zero-Annotation Object Detection with Web Knowledge Transfer.” *European Conference on Computer Vision*. Pp 1-16. Computer Vision Foundation, 2018.