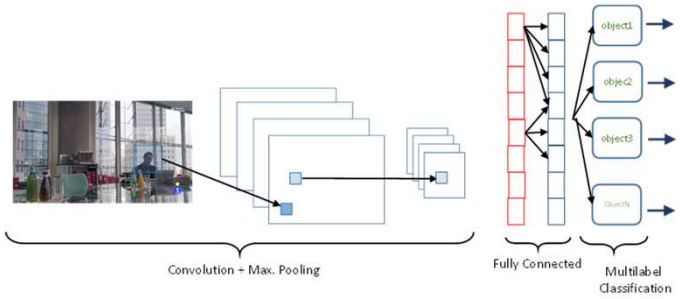
UNDERSTADING WHAT'S INSIDE VIDEO IN TV SERVICE PROVIDERS

Miguel Jose Esteve Brotons

1.- INTRODUCTION

Currently, in the context of pay TV services, there is fierce competition between traditional TV content providers, and new incumbents offering similar services through streaming video platforms, in Over The Top (OTT) mode. Traditional TV service providers are those who have made use of means of transport such as Digital Terrestrial Television (DTT), satellite, cable or Internet Protocol Television (IPTV). The competition between "traditional" providers and the new Over The Top streaming platforms is based on several axes, the most important being: content offer, offer package and / or bundling.

TV service providers have access to a large amount of linear and non-linear content. It is time to offer additional capabilities based on video analytics and smart metadata extraction. Deep learning has emerged as a new field able to provide new advanced capabilities to extract relevant metadata from the multimodal analysis of video, speech and text.



2. MULTIMODALITY

Extracting meaningful features from a raw video signal is difficult due to the high dimensionality. Every frame is represented by a tensor of (width x height x 3), where the last dimension represents the three RGB channels. For video content this adds up quickly: if we use common image recognition models like Res Net or VGG19 with an input size of 224 x 224, this already gives enough amount of features for a one minute video segment at 25 fps.

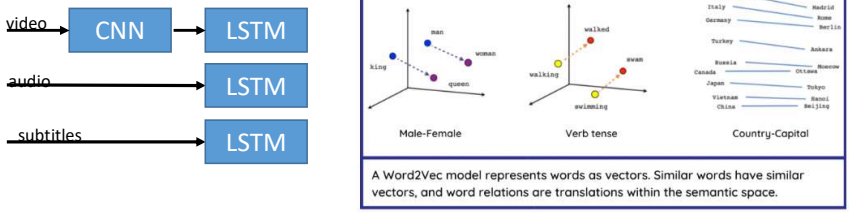
Detecting the changes in the events in the stream flow is a key part of video understanding. The very changing semantic context on TV programs does not allow to determine with precision the points at which a given program start ends.

In addition to video classification, some authors propose to apply text transcripts on top of video streams. The easiest and cleanest option may be to use subtitle files as a text transcript, but these are clearly not available for all videos. In this case there is a need to use a speech recognition engine to extract those text transcripts.

3.- PROPOSAL

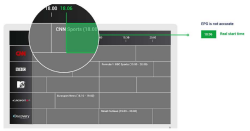
Analysis of TV service provide video sources to extract meaningful information based on multimodal video analysis, encompassing object recognition, logo recognition, action/topic recognition and speech to text analysis. We'll set focus specifically in detection of start/end event, in order to fix with accuracy the start/end time of each program, and action recognition. For that, we'll look for some key patters in the video frame sequence, comprising TV logo detection, and speech/subtitle context switching.

We propose to analyze combined CDNN and LSTM architecture able to cope with the objective to extract rich information from TV streams as to improve traditional service provider offerings. In addition, NLP techniques as Word2Vec will also be included for the speech to text part involved in the multimodal analysis.



4. CONCLUSIONS

We expect that combination of Video and NLP techniques lead to a quick classification of features inside video able to support new features in the TV service provider echosystems.



5. REFERENCES

Kok Meng Pua, John M Gauch, Susan E GauchJ, Jedrzej Zmiadowicz. ***Real time repeated video sequence identification***, Computer Vision and Image Understanding, March 2004, Volume 93, Issue 3, pp.310-327

Shervin Minaee, Imed Bouazizi, Prakash Kolan, Hossein Najafzadeh***. Ad-Net: Audio-Visual Convolutional Neural Network for Advertisement Detection In Videos***. 2018 . arXiv: 1806.08612v1.

Stromatias, Karthik Yadati, Martin Prins & Joost de Wit***, Using Machine Learning to create personalized snackable content***. Media Distillery, The Netherlands, ibc 2018. Available at: https://www.ibc.org/download?ac=6531