USING DEEP LEARNING TECHNOLOGIES TO PROVIDE DESCRIPTIVE METADATA FOR LIVE VIDEO CONTENTS

A. Maraga, V. Zamboni

Metaliquid, Italy

ABSTRACT

As the consumption of video contents rises and consumers' behaviour changes, video broadcasters must find new strategies to engage with their viewers. In this context, metadata are used to provide customers with personalized, relevant contents.

More recently, content-descriptive metadata have enhanced the personalization of services and have offered providers new monetization opportunities thanks to, e.g., more effective content recommendations and contextual advertising. So far, descriptive metadata have been applied to pre-recorded contents, while they still represent a huge amount of unexploited valuable information for live streams.

This paper highlights how deep learning technologies can add value to information contained in each video frame of both live and non-live contents, automatically providing detailed descriptive metadata.

This technology opens up the opportunity for broadcasters to offer new services and an innovative way of video consumption.

INTRODUCTION

Over the last years, the access to video contents has rapidly grown and customer behaviour in consuming these contents has changed accordingly. To give an idea of the growth of digital media, according to Cisco, by 2021, 82% of global Internet consumption will be video content (Cisco (1)). PwC's reports that the total worldwide subscription spending on Netflix and other over-the-top (OTT) subscription video-on-demand (SVOD) services grew by 33.8% in 2014 and 32.3% in 2015, that is 77% in two years (PwC (2)). One of the challenges broadcasters must face is how to use the benefits offered by these new models to engage with their viewers and generate new revenue opportunities from the expanding customer base. Metadata are the key to achieve these goals.

Content attributes together with viewer information are currently used to offer a higher quality viewing experience, providing customers with personalized, relevant contents. More recently, metadata describing the video content itself have been used to enrich video contents by providing external links and relevant, detailed information. Furthermore, content-descriptive metadata offer broadcasters new monetization opportunities, enabling

more effective recommendations, contextual advertising, and new content-management tools.

So far, however, descriptive metadata have been used for VOD and non-live contents, since current metadata extraction from video relies heavily on human tagging and annotation; even if this procedure is highly precise, it can be unfeasible for large volumes of content. On the other hand, descriptive metadata still represent a huge amount of unexploited valuable information for live stream, which needs a flexible and efficient feature-extraction method able to react in real-time.

The increase in the number of available video contents and the unexplored opportunities offered by live streams require automatic tools to analyse and understand video contents. In this context, deep learning technologies can add value to information contained in each frame of both live and non-live contents. This paper highlights how deep learning techniques combined with motion and sound analysis enable to identify in real-time a broad set of concepts (e.g., recognition of people, locations, sensitive contents, languages) in a video and create a semantic relationship between them, automatically providing detailed content-descriptive metadata.

HOW TO USE CONTENT-DESCRIPTIVE METADATA?

Higher-level information provided by content-descriptive metadata is used in various scenarios. On the one hand, it can be used to offer an interactive, premium customer experience, enriching contents with clickable links, on-screen specific information, and innovative services. On the other hand, it gives broadcasters the opportunity to increase their revenue streams.

By understanding a content as the viewer perceives it provides media companies with higher-quality data which can improve the technology that powers content-discovery features. Companies like Netflix and HBO keep on adding newer shows to their repertoire, bringing content libraries to gigantic proportions. To combat that, viewer-specific recommendations become one of the best ways to keep users engaged with the type of content they prefer. To improve customer experience, market players need to collect quality metadata to match viewers' choices and leverage the power of content recommendation algorithms that predict what users may like.

Contextual advertising is another service which will be empowered by content metadata which are the key to meet viewers' interests, achieve the best brand positioning, and maximize the opportunity to show an advertisement. Research institutes agree that the evolving landscape needs to gather deep data in order to offer context relevant ads which goal is to be related both to contextual sentiment and content and to viewers' interests and taste.

Digital asset management will also gain a competitive advantage for market players who will adopt metadata enrichment. With the rise of digital video, the huge, constantly-growing catalogue of contents represents broadcasters' most important asset; identifying the key concepts associated to each scene enables the development of new content-management tools able to search and organize video contents according to specific needs and

categories. This will allow both a longer lifecycle of content thus making also grow the monetization opportunities of each content.

The way users consume video contents affects how metadata are produced and used and which ones are relevant for the video stream. The main difference is between non-live and live events.

Video on demand

Video on demand is changing viewers' role, which are not anymore only an audience but active users, that wish to have greater control than ever before over what they watch, when they watch, and how they watch.

Easy accessibility to entire seasons of TV shows has led customers to change their consumption behaviour and binge watching has become the norm (Netflix (3)). Detecting the title sequence and the closing credits enables broadcasters to suggest the right content, i.e., the next episode, at the right time, enabling high efficiency in the management of binge watching functionality.

Video-feature extraction, such as the presence of objects and people, their actions or audio classification, represents a huge and valuable source of information since every single frame of the video is described using detailed metadata. This kind of annotations is an effective way to improve consumers' interaction and engagement: it can be used to offer users the possibility to discover additional information related to a certain content, providing, for instance, external links to relevant websites. Furthermore, the monetization of contents can be increased including purchasing links to the items which are displayed in the video.

Moreover, combining viewer information and content-descriptive metadata, both providers and advertisers can maximize their revenue opportunities. Indeed, the effectiveness of the standard approach to deliver an advertisement is limited by its passive viewing by consumers and its limited relevance to the viewing context, which can lead people to skip ads. Exploiting metadata, however, personalized ads are shown in a relevant context, without impairing the viewing experience, and at the right time (Kancherla et al (4)).

The presence of sensitive content, such as nudity or violence, in a video can reduce the potential audience and, consequently, the potential revenue for broadcasters and advertisers. Detecting the scene showing inappropriate content allows broadcasters to inform viewers and enables both providers and users themselves to skip or block it, if necessary.

Live content

Together with the remarks made for the pre-recorded contents, live events pose an additional big challenge: video contents must be analysed and metadata produced in real-time.

A significant and valuable example are live sport events, which are among the most-watched broadcasts. Users who are not able to view the live stream or start watching some minutes late require accessing the most relevant moments of the event, such as goals during a football match or crashes in a car race. Broadcasters must be able to

provide such a content in real-time, otherwise customers will probably look for it on other platforms, and an important monetization opportunity will be lost. To achieve this goal, a detailed annotation of the key moments of the event must be produced in real-time. Furthermore, these metadata are useful also in the post-production stage, making contents easily searchable.

MACHINE LEARNING APPROACH

As the volume of available video sources constantly grows, too much effort is required for the manual tagging and annotation of contents and an automatic feature-extraction method is becoming a need.

Video search is currently an active field of research and neural networks have been demonstrating as effective models able to analyse and understand video contents.

Overview on neural networks

Neural networks are systems that can learn from examples. They are modelled as collections of simple units called neurons, usually organized into distinct layers; the most common layer type is the fully-connected layer in which neurons of two subsequent layers are fully pairwise connected, but neurons within a single layer do not share any connections. Figure 1 shows a simple example of neural network: each circle represents a neuron and arrows are connections between them. Each neuron takes in a set of inputs, computes an output value, which in turn becomes the input to other neurons; chaining together a lot of these simple units enables to model complex functions.

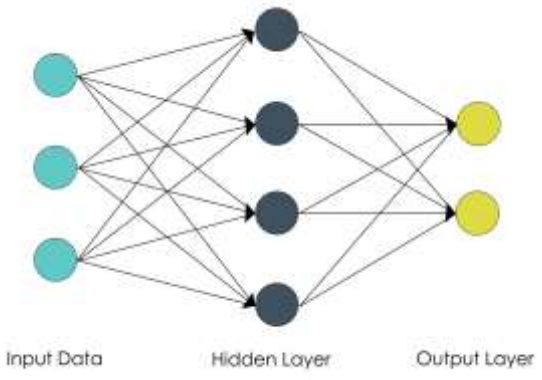

Figure 1 – Neural network

Neural networks have been used to solve many different tasks. Image and video recognition are classification problems: given a set of categories, one wants to predict to which class a new example belongs. To this end, a neural network is trained on a set of examples labelled according to the classes taken into account: the algorithm analyses the training data repeatedly, trying to learn general features. Once the classifier is trained, its

quality is evaluated on a test set of unseen examples: the output layer of the neural network returns the probability of belonging to each considered class, and the one with higher probability is considered as the predicted label, which can be compared with the ground truth.

Neural networks described so far, however, do not scale well to high resolution images. Recently, a new architecture has been designed specifically to deal with images: convolutional neural networks (CNNs). The main difference with respect to regular neural networks is that the layers of a CNN have neurons arranged in three-dimensional volumes. The most important building block in this kind of architectures is the convolutional layer, in which, unlike fully connected layers, each neuron is connected only to a small region of the input volume. Furthermore, the depth of the volume corresponds to different filters which learn to recognise different features in the images, such as edges or colours. CNNs have been shown to be effective in solving tasks such as image detection and recognition. Many different models based on CNN architecture have been proposed in literature (Krizhevsky et al (5), Simonyan and Zisserman (6), Szegedy et al (7), He et al (8)). However, the optimal choice strongly depends on the task and on the available dataset.

When dealing with video recognition, also the temporal dimension must be taken into account. In order to process arbitrary-long sequences of inputs, recurrent neural network (RNN) models have been introduced. In particular, RNNs which contain the so-called long short-term memory (LSTM) units have been demonstrated to be capable of long-range learning. Combining the ability of CNNs in extracting features from still images (i.e., video frames) with LSTM models, it is possible to perform activity recognition, image captioning, and video description (Donahue et al (9)).

The importance of datasets

One of the crucial elements to take into account when dealing with deep learning is datasets.

Since neural networks learn from examples, they need a good training set to work properly. As a consequence, it is important to select data in such a way that the neural network can extrapolate significant features.

Furthermore, deep learning is particularly dependent on the availability of large quantities of high-quality training data in order to learn an underlying behaviour that can be generalized to unseen, test data and can yield reliable prediction on any example from the problem domain. The importance of having a proper dataset is strictly related to the problem of overfitting. Overfitting refers to a model that fits the noise in the data instead of their underlying relationship, overreacting to minor fluctuations in the training examples. In other words, overfitting occurs when a model begins to memorize training data instead of learning how to generalize. As a consequence, such a model will have good performance on training examples and poor predictive performance on data it has never seen. One of the solutions commonly used in machine learning when considerably large datasets are not available is data augmentation, that is, the application of one or more deformations to a set of examples, which results in new, additional training data. For instance, when dealing with images, typical perturbations considered are flipping, random crops, colour jittering, and lighting noise.

Methods

CNN architectures that have been trained are inspired to those proposed in (6). They are built by stacking blocks of convolutional layers characterized by small convolutional filters. RNN architectures, instead, combine a CNN for feature extraction and a single LSTM layer.

For each task, a specific dataset has been created. It has been split into two parts, one for training the model and one for testing. During training, the hyperparameters of the network have been properly tuned in order to obtain the best result.

RESULTS

The deep learning approach described in the previous section has been applied to a variety of tasks, both on still images and on video, whose results are summarized below. For each task, the optimal model has been built, tuned and tested. For training neural networks, custom, tailored datasets have been created.

Audio classification

Interestingly, the classification of the audio-signal of a video can be done representing it as an image, called spectrogram. A spectrogram is the representation of the spectrum of frequency (on the vertical axis) in time domain (horizontal axis); the colour of each pixel in the image indicates the amplitude of a certain frequency at a certain time (Figure 2).

For building the dataset, more than 200 hours of audio has been sampled in short windows of ~2.5 s, obtaining ~300K images which have been analysed using a CNN. The algorithm is able to distinguish music, noise, and dialogue in Italian, English, French, and Spanish.

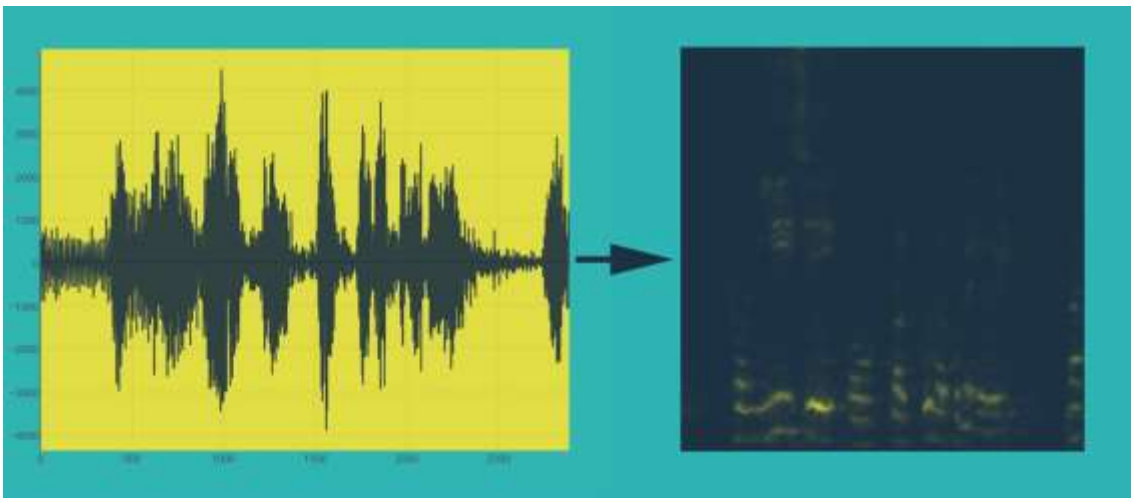

Figure 2 – Spectrogram

Detection

The first problem to solve when analysing a video content is the detection of relevant entities. In particular, faces, sensitive content, and objects related to sport events have been considered. To perform this task, each video frame has been processed with a CNN at different magnitude scales, in order to generate a heatmap associated to each class, which indicates the presence and the position of the corresponding entity (Figure 3-4).

In order to train the detector a dataset of 500K images has been used.

Face detection

The trained CNN is able to detect both frontal faces, profiles, and in-plane-rotated faces at different magnitude scales.

The detector finds also a bounding-box that provides the image for face recognition.

Sensitive content detection

The algorithm is able to detect several sensitive contents, such as nudity, weapons, blood, and explosions.

Figure 3 – Face and sensitive content heatmaps

Object detection

The detector has been trained to recognise several entities related to sport events. It is currently able to distinguish cars and motorbikes during races, the football pitch, the basket court, and the crowd watching the sport event.

Figure 4 – Crowd, football pitch and motorbike heatmaps

Face recognition

For this task, a CNN has been trained on a dataset of ~800K images of ~500 international celebrities, that has been specifically created. The algorithm is able to recognise both frontal faces and profiles. The dataset can be easily updated and the network rapidly retrained to respond to specific needs.

Setting recognition

A tailored dataset based on the settings that characterized movies and TV-series has been built. A CNN has been trained to distinguish internal and external setting; in particular, among the external settings, there are urban and rural areas, sea, snow, and desert.

Action recognition from video

In contrast to the case of still images, the analysis on video has been performed using RNNs. The main focus of the study was sport events.

In a first stage, a RNN has been trained in order to recognise the sport being played, such as football, rugby, American football, hockey, Formula One and motorcycle races.

In a second stage, the recognition of particular actions in motorcycle races has been investigated. The algorithm is currently able to distinguish crashes and overtakes, and also to recognise the drivers involved in these actions.

REAL-TIME

A big challenge posed by live events is metadata extraction in real-time. The technology presented in this paper can be used to perform all the analysis described in the previous section in real-time on a HD video content in 25 fps. Such performance can be achieved analysing the content with a distributed framework on a cluster composed by three nodes with NVIDIA Tesla K80 GPU, 36 cores and 60 GB memory in total.

CONCLUSIONS

As the available choices of video contents grow, broadcasters must deal with huge catalogues, which, in turn, represent their most important asset. On the one hand, they must find new strategies to engage with their viewers and provide more and more personalized contents. On the other hand, mastering content discovery, contextual advertising, and asset management has become even more critical.

In this context, automatic feature-extraction methods have become a need. Neural networks have demonstrated to be a flexible and efficient tool able to understand video contents as viewers would do. This paper shows that by designing a proper network architecture and using a good dataset for training, it is possible to automatically provide detailed content-descriptive metadata for video contents in real-time. In particular, CNNs have been used for audio classification, object detection, and face and setting recognition, while action recognition in sport videos has been performed with RNNs. Tailored datasets

for training have been created; they can be easily customized and the networks rapidly retrained in order to respond to specific needs.

This technology opens up the opportunity for broadcasters to offer new services and an innovative way of video consumption.

REFERENCES

1. Cisco, 2017. Cisco Visual Networking Index: Forecast and Methodology, 2016–2021. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html.

2. PwC, 2016. Global Entertainment & Media Outlook 2016-2020. https://www.pwc.com/gx/en/entertainment-media/publications/assets/tv-and-video-outlook-article-december-2016.pdf.

3. Netflix, 2013. Netflix Declares Binge Watching is the New Normal. https://media.netflix.com/en/press-releases/netflix-declares-binge-watching-is-the-new-normal-migration-1.

4. S. Kancherla, R. Warey, and S. Ramki, 2016. Using metadata to maximize yield and expand inventory in TV - contextual advertising. IBC 2016 Conference. p. 7.

5. A. Krizhevsky, I. Sutskever, and G. E.Hinton, 2012. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems 25. pp. 1097-1105.

6. K. Simonyan, and A. Zisserman, 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. ICLR 2015.

7. C. Szegedy, W. Liu, and Y. Jia, 2015. Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1-9.

8. K. He, X. Zhang, S. Ren and J. Sun, 2016. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770-778.

9. J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, 2015. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2625-2634.