



White Paper

Innovations in Machine Learning and AI for Media QC Operations

Shailesh Kumar
Interra Systems, Inc.

1. INTRODUCTION

The way that video providers analyze content is evolving. In the early days, automated quality control (QC) systems were used to check basic technical parameters like file format, resolution, bit rate, content structure, and simple perceptual problems like blockiness, blurriness, and black bars. Over time, perceptual issues grew and more sophisticated computer vision-based checks were added to the QC workflow, enabling operators to detect defective pixels, visual text, compression and ghosting artifacts, shot boundaries, loudness, and language.

Recent innovations in deep learning have brought QC to a whole new level. Today, operators can accurately identify violence and nudity in content, which was previously a manual task. This whitepaper will explore the next generation of QC checks, which involve in-depth semantic understanding of content by identifying and extracting objects, events, actions, scenes and spoken or visual words from the audio-visual content and using them for specific purposes like content compliance, content classification, indexing and retrieval, automatic generation of content description or captions.

2. OPTIMIZING CONTENT COMPLIANCE WITH DETECTION AND CLASSIFICATION

Today's video providers need to ensure that their content complies with local or regional requirements related to strong language, violence/disturbing images, nudity, sexual content, the presence of alcohol or smoking, parental guidelines, and more.

Content compliance can be achieved through low-level, fundamental detection and classification tasks. These tasks may include detecting objects inside frames, recognition of actions over several frames, classification of general scenery, detection of specific events in audio or video tracks, general classification of videos into specific activities or themes, conversion of speech to text, and detection and recognition of faces.

Performing these detection and classification tasks generates a huge amount of descriptive metadata and annotations in the content at frame and scene levels, which can be further analyzed for mapping to specific content compliance and regulatory needs. Let's take a look at how deep learning (DL) systems can be leveraged to enable smarter content classification and compliance. But before that, let's look at a bit of history behind the deep learning systems.

3. THE EVOLUTION OF DEEP LEARNING SYSTEMS

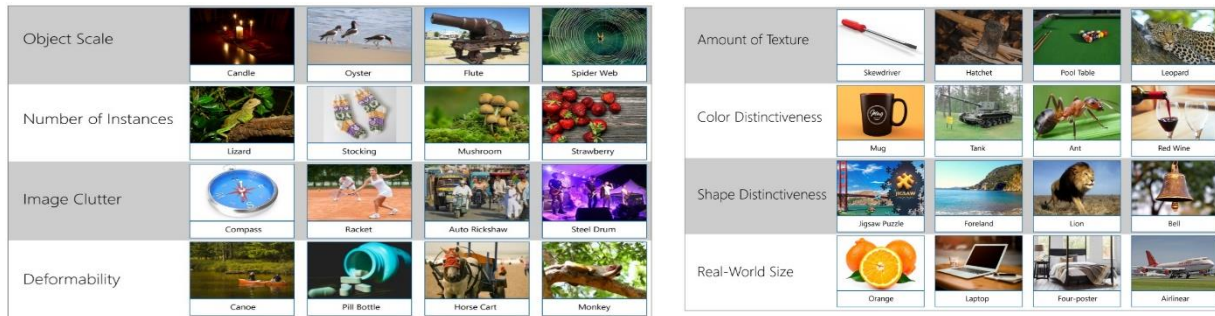
A machine learning (ML) system takes features describing an image as an input and outputs an inference (e.g., a class label in an image classification problem). In a traditional ML system, the features extracted from image are designed by humans themselves. Some of the typical features are Harris corners, SIFT feature vectors, histogram of gradients, and optical flow. The ML system is typically a linear or logistic regressor, an SVM, or a small neural network with a couple of layers. A substantial amount of effort is spent on feature engineering, which is the art of designing and extracting accurate features for a specific inference task. A Deep Learning (DL) system is essentially a more sophisticated machine learning system which takes an image directly as an input and has many neural network layers. Initial layers of the network work as feature extractors, which are directly learned from the data. This is in contrast to the traditional machine learning systems where feature extractors were designed by humans through feature engineering. Later layers are focused on the specific inference task (such as image classification, object detection or action classification etc.). The availability of high-performance Graphics Processing Units (GPUs) led to the rise of deep neural networks, since it became feasible to train these large networks over a couple of weeks on a GPU. Previously, running such large system would take years on a CPU and hence it was not practically feasible.

One of the biggest breakthroughs in Deep Learning happened in 2012 when AlexNet was designed by Alex Krizhevsky and published with Ilya Sutskever and Krizhevsky's PhD advisor Geoffrey Hinton. AlexNet is a convolutional neural network trained on 1.2 million real world images from the ImageNet dataset for the purpose of classifying them into 1000 different categories. With five layers and 60 million parameters, AlexNet was made

The process of adapting models for specific detection and classification needs is much more efficient than training a model from scratch, as this would require a huge dataset of images and take weeks of training time. If the dataset is small, then training a model from scratch is not worthwhile, as the model is quite prone to over fitting.

Another critical factor that led to the success of DL is the availability of huge well-labeled datasets. In ImageNet, over 14 million images have been hand-annotated to indicate the objects present in the image. Thanks to these well annotated datasets, people have been able to train networks to a degree where visual recognition has outperformed human accuracy (in specific test settings).

4. OBJECT DETECTION CHALLENGES



There are a number of challenges that an object detection system has to grapple with in correctly classifying an object. Typical attributes of an object like its scale, quantity, texture, color, shape can vary a lot in different object categories in real life images. See examples in the pictures above. Objects can vary in scale. The number of instances that the same object is used can vary significantly, from one to hundreds. In some clean images, there is no background clutter, while in others an object (e.g., a drum) is surrounded by hundreds of other objects. Moreover, the shape of an object can deform in various ways (e.g., a horse cart or a monkey). The amount of texture on the surface of an object can widely vary. While some objects like red wine may have a distinctive color, other objects like a mug may come in any possible color. The shape of a bell is quite rigid and standard, while a jigsaw puzzle may come in all possible shapes. The size of an object in an image is completely uncorrelated with its real life size.

Essentially, no particular object attribution is comprehensive enough in distinguishing 1000s of object categories in large scale. Different attributes interact in myriad ways. Learning the distinctive combinations of these attributes is incredibly hard in a large scale classification. A deep learning system which learns the distinguishing features from the training data automatically helps obviate this challenge.

5. APPLICATIONS FOR ML/DL IN THE QC WORKFLOW

ML/DL can be used in a variety of ways for quality and compliance purposes. This section of the article will review some of the applications for ML in media QC workflows.

Activity recognition

Activity recognition aims to identify the actions and goals of one or more agents from a series of video frames. In the past, techniques like optical flow, Kalman filtering, and Hidden Markov models were used to address specific activity recognition problems. In 2014, Karpathy et. al. showed how a deep convolutional network could be used for large-scale action classification by fusing information from multiple frames in a sliding time window. They showed that CNN architectures are capable of learning powerful features from weakly labeled datasets that far surpass traditional feature-based methods in classification performance. In Figure 2, the blue rows indicate the ground truth labels and the bars below show model predictions sorted in decreasing confidence. The green labels are the correct, while the red ones are incorrect predications. As shown in the figure, the model has identified the right category in top five predictions, but the right category is not always the first one. It's also noticeable in the figure that the categories the network gets confused about are often similar. For example, ultra-marathon, half marathon, and running are related activities. Even ordinary humans would require sufficient training to distinguish between these categories. In 2015, a 3D convolutional network was proposed, which works on video clips that are 16 frames each in size. This relatively simple architecture can produce learned features that are generic, compact, efficient to compute, and simple to implement. Furthermore, this network can model appearance and motion information simultaneously.



Figure 2. Activity recognition enabled by deep convolutional network. Source: Karpathy, 2014

Visual text recognition

Recognizing onscreen visual text can be useful for various tasks, such as caption alignment, language detection, and advertisement classification. Using a combination of computer vision, machine learning, and natural language processing techniques video providers can improve onscreen text recognition. Vision is used for separating out the background and identifying the regions in the image that contain text. Once the text regions are extracted, a ML-based trained model can be used for recognizing the characters. The sequence of characters is then fed into an NLP pipeline, which can form words and sentences from them. Then the extracted text can be used for various tasks.

Audio events

Events in audio can be helpful for detection purposes. For example, while a gunshot looks amorphous in video, it can be clearly recognized in audio. Screaming sounds are useful indicators of violent or horror activities. It is easy to adapt the deep convolutional networks for audio event recognition, too. Audio can be broken into fixed sized frames, and then 128 dimensional mel spectra can be constructed for each frame. It's possible to combine 128 consecutive frames to form an image of size 128x128 and then feed it into the CNN. A challenge is localizing an exact event to specific frames. What's more, human annotation can be costly. This can be addressed by using just weak labels, which characterize whether a particular clip has an event or not. The network uses smart tricks to localize an audio event, even from these weak labels.

Caption alignment

A unique application of ML is caption alignment. A voice activity detector can identify the time codes where the dialog is happening. The speech-to-text conversion system can then extract actual words being spoken. It is then easy to match the actual text in audio with the text in transcript. While there may be some mistakes in the speech-to-text system, it is enough for identifying the matching captions. Then the differences in time codes in audio and captions can be estimated and corrected easily.

From Algorithms to Applications

For next-generation QC solutions it's possible to construct a system architecture that combines all of the deep learning technologies for specific applications for content understanding, compliance, and quality control purposes. (See Figure 3.) Under this scenario, the system is divided into three layers: a content layer, algorithms layer, and applications layer. The algorithms are general purpose and extract all of the relevant information from the audio-visual content like objects, scenes, events, activities, topics/themes, faces, visual text and spoken dialogs laid out on the content timeline. Applications map it to the specific requirements of content compliance.

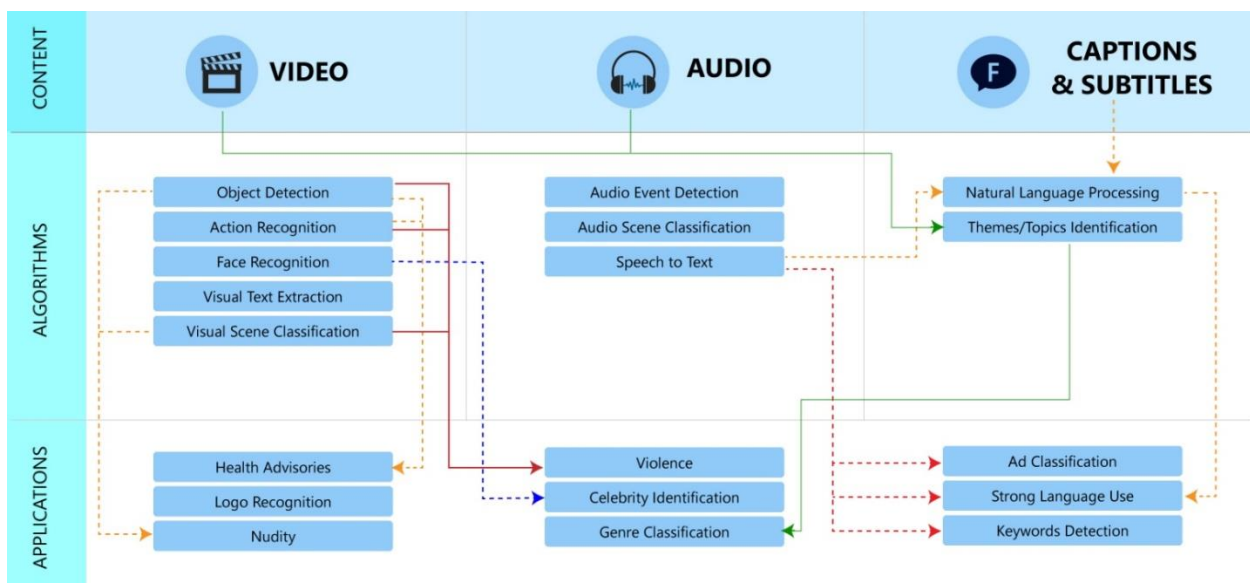


Figure 3. DL architecture applied to video, audio, and caption and subtitles.

There are a few different applications for which this architecture is beneficial. One use case is explicit content. Explicit content usually maps to three different classes. Nudity or minimal covering can be discovered by use of object detection and visual scene classification. Mild sexual situations depend on activity recognition, audio events, and dialog/caption analysis. Identifying explicit sexual situations involves the usage of object detection, scene classification, audio events, dialog, and activity classification. It's important to note that the explicit content detection is inherently a multi-modal inference problem and both audio and visual cues are critical in solving it.

Similarly, object detection can be used to detect violence, in particular the presence of various firearms and cold arms. Activity classification is used for identifying actions like killing, car crashes, and gun shots. Audio event detection can be used for identifying events like gunshots and screams.

Moreover, some regions ban the presence of alcohol and smoking in video content. Alcohol can be primarily identified using object detection. Smoking involves a combination of object detection for cigarettes, cigars and other vaping devices and activity recognition for the act of smoking itself.

In a typical video file, the number of scenes that actually create a compliance issue (e.g., a violent scene) is few. A metric should be defined around only the relevant frames. Useful metrics from this perspective are precision and recall. Recall refers to the ratio of number of correctly detected scenes with the number of relevant scenes. Precision is the ratio of correctly detected scenes vs. the total number of detected scenes. A high recall means few false negatives. A high precision means few false positives. From the QC perspective it is important to ensure that the system has a very high recall. Some precision can be sacrificed to achieve this, as the cost of false negatives is much higher than false positives.

Conclusion

Over the years, as technology has improved, automated QC has become more efficient and accurate for media operations. What seemed impossible a few years ago without human eyeballs, now is quite achievable. Thanks to recent advancements in ML and AI, software-based QC solutions can carry out compliance and quality related tasks with about the same level of accuracy as manual workflows. As leaders in the media QC software space, Interra Systems is helping to bring these technological advances to video providers in a way that seamlessly integrates with existing media QC workflows.

ENDS