SCTE·ISBE CABLE-TEC EXPO 2019
NEW ORLEANS, LA
SEPT. 30-OCT. 3

2019 Fall Technical Forum
SCTE·ISBE • NCTA • CABLELABS

Machine Learning Applications in Cable TV Advertising – Usage and Challenges

A Technical Paper prepared for SCTE/ISBE by:

Srilal M Weerasinghe PhD
Principal Engineer
Charter Communications
8560 Upland Drive, Englewood, CO 80112-7138
720-699-5079
srilal.weera@charter.com

Table of Contents

Title **Page Number**

List of Figures

Title **Page Number**

List of Tables

Title **Page Number**

Introduction

Use of machine learning (ML) for image and video analyses would often include face recognition, personalization and recommendations. An emerging trend is the application of AI technology for TV advertising. In this paper, we present the unique challenges in applying machine learning to carrier-class video advertising. We focus the discussion on a specific use case that is common to all ad supported TV services.

The selected use case is Ad Ingest Quality Control (QC). In the United States, TV commercials are subjected to various rules and regulations. Ads containing specific content (e.g. Alcohol, firearms) are barred from airing during certain TV programs. Identifying these categories may pose a challenge, as off-the-shelf machine learning products are more oriented towards facial recognition. That is to be expected perhaps, as the video ML products were primarily intended for surveillance and sports applications. However, our research indicates that by judiciously combing metadata from multiple data streams, machine learning analysis results can be improved.

The intent of the paper is to outline the results and recommendations of a proof-of-concept study that will be helpful to the carrier-class video services community.

Content

1 TV Advertising – Quality Control at Ad Ingest

Multi-channel video programming distribution/ distributor (MVPD) is a highly regulated industry in the US. The term covers not only traditional cable companies, but any entity that provides TV service to consumers via fiber, coax, satellite, DSL and wireless. With the advent of internet-based TV service (also known as OTT), the moniker is modified as V-MVPD (virtual MVPD). In all cases, the content distributors could be responsible for the displayed video content, including advertisements [1]. This places the onus on the content distributor (also known as service provider/network operator), to prevent the 'non-compliant' content from reaching the TV audience.

In the context of the present discussion, there is a distinction between movie content and ads. While movies/episodes are originated from mainstream studios (and are properly vetted), the TV ads could originate from a multitude of sources. Therefore it is necessary to identify any non-compliant ads prior to airing at the Ad Ingest Quality Control (QC). Today, this is done manually by trained individuals. They examine tens of thousands of ads a month and quarantine the failed ones. The challenge is to automate that process with an AI/ML engine embedded into the workflow.

First we examine the basis for non-compliancy of ads. When a TV commercial is deemed non-compliant, the restriction usually stems from one of the three categories below.

a) Regulatory Compliance

The Regulatory constraints are primarily stipulated by FCC [1] but could also be under the purview of FTC, FEC and FDA [2] [3] and [4]. Listed below are some examples of regulatory requirements overseen by federal agencies. See the references cited above for full requirements.

- Ads related to alcohol, tobacco, firearms, gambling etc. must meet federal guidelines.
- A political ad is required to display a statement from the sponsor for at least 4 seconds.
- Truth-in-advertising – An ad may be deemed deceptive for misleading/missing information.
- Ads promoting certain lotteries, cigarettes or smokeless tobacco products are not allowed.
- Ads must comply with loudness mitigation requirements of CALM Act.

b) Contractual Compliance

Contractual constraints are imposed by content providers such as ESPN. An example would be the restriction on alcohol ads during ESPN Little League World Series program. For a complete list of applicable restrictions, see reference [5].

c) Business/Operational Compliance

These are generally operational guidelines and best practices established by the enterprise. Being sensitive to audience needs as well as delivering quality content could enhance a company credibility. One example is ‘frequency capping’ or limiting the display of the same ad multiple times.

2 Machine Learning in Carrier-Class Video Applications – Challenges

Identifying the above categories programmatically poses a challenge to ML tools, as off-the-shelf products are more oriented towards facial recognition. A familiar ML application is creating a ‘bounding box’ around a face and tracking it through a video-clip. Such applications are useful in sports and surveillance, however they are not directly applicable to MVPD market. The latter requires comprehensive ML analyses of multiple streams (video, audio and textual metadata).

In common usage, Machine Learning video products do a multi-pass analysis (each pass to identify faces, common objects, celebrities etc.). The results are presented as content descriptor metadata (labels). An accompanying ‘confidence level’ indicates the accuracy of prediction. Per our lab testing, off-the-shelf ML tools didn’t meet our needs right out of the box. It may be because the video content/Ads detection is still a nascent technology. Adapting such products for carrier-class video applications requires a certain amount of post-processing. Else, the results could be tainted with false positives or the tool may fail to identify content adequately (false negatives).

2.1 Technical Challenges

To train a neural network, a good selection of examples and counter-examples is needed. Else the machine learning model would be susceptible to 'overfitting'. That is, the model will fit the existing data well, but would fail when it encounters a new instance of the target data. While this is not an issue with common objects (e.g. cars) due to the abundance of examples, it is a challenge for objects with ambiguous signatures (such as fireworks or alcohol). Distinguishing 'fireworks' from similar signatures ('bright lights in a dark background'), is not an easy task. Similarly, an image classifier may find it hard to differentiate 'beer' from a similarly colored liquid in a bottle (e.g. olive oil).

The need for proper counter-examples becomes more acute as we move from image analysis to video activity identification. This is discussed in detail in the 'Issues Noted in Our Testing' section below.

Next we present a short overview of applicable deep learning algorithms.

2.2 Machine Learning Models for Image and Video Classification

General multi-perceptron based neural networks (ANN) are not able to meet carrier-class video classification requirements. Training time and accuracy would be hard to achieve. Convolutional neural networks (CNN) is the Deep learning based technology used for image classification. Most products use 'transfer learning' model; first training the model on a large public dataset such as ImageNet or Inception and then fine tuning it to meet the specific requirements. While image analysis has only spatial dependence, video analysis involves the temporal component.

For time series analysis, recurrent neural networks (RNN) deep learning model is the standard technique, due to its ability to store events happened in the past. However, it is well known that RNN, with many hidden layers, suffers from the vanishing gradient problem. This issue also manifests as the exploding gradient problem. (A simpler interpretation is that Tangent of the angles being very close to 0 or 90 degrees, respectively). The root cause is the exceedingly small derivatives of the 'loss function' (or error), during back propagation. A solution is to disregard certain intermediate steps to avoid extreme values of the gradients. A popular model for handling such sequence data is the Long Short Term Memory (LSTM) algorithm. LSTM discards certain data (via the 'forget gate') to reset the cell state thus keep the values getting extreme.

In the field of deep learning, new algorithms are routinely being developed (Fast R-CNN, Faster R-CNN etc.). These are mainly for improving the speed of analysis, as updating millions of parameters (weights and biases) associated with hidden states takes a lot of time.

2.3 Performance considerations

In our testing, the processing time as measured was not close to real-time. One reason could be the ML engines operate in multi-pass mode. This is necessary because at Ad-Ingest quality control, the ML engine works as a gate-keeper. On the other hand, if the intent is to find a single signature (e.g. either guns or alcohol), a single pass would be sufficient.

We have tested machine learning models in appliance mode as well as in the cloud. The cloud-based implementation is preferred if the data also resides on the same cloud. The appliances would be GPU-based (as opposed to CPU), due to the large number of cores which facilitates parallel computing. We tested with NVidia GTX and also plan to benchmark with NVidia DGX (with thousand TFLOPs of computing speed),

2.4 Limitations of Current Machine Learning Tools

To improve the detection accuracy, Machine Learning tools tend to use increasingly sophisticated algorithms. However, the algorithmic approach alone did not seem to produce expected results. Obtaining optimal results within a reasonable time is a challenge. Searching each video frame for a multitude of categories (alcohol, gambling, drugs, violence, trademarks, copyrighted content, explicit content, political content etc.) is time consuming. It could also be irrelevant (i.e. searching for all manners of firearms or medications would be wasteful, in the case of a beer ad).

To improve the results, we propose adding a software engine to the workflow to perform additional analyses.

3 Lab Evaluation

Our findings are presented below in a vendor agnostic manner.

3.1 Image Analysis

Content descriptors (Labels) need to be sufficiently descriptive for effective contextual analysis, i.e. instead of generic labels such as ‘person/human’, the ML tool needs to identify whether a person is young/old, male/female, mood etc.

3.2 Video Analysis

Activity identification is a challenge for current ML tools. This is a burgeoning field of research at premier AI/ML research institutions [6]. For the MVPD space, ‘activity identification’ would open up new applications. E.g. identifying a car chase from a video (as opposed to cars in a still image) would offer new ad opportunities. Table-1 below depicts sample activities that are relevant to contextual advertising.

Table 1 – Machine Learning Identification of Activities for Ads*

Dominant Activity	**Suggested Ad Types**
Cooking	Kitchen Appliances & Utensils, Cooking Classes
Car chase	New Cars, Auto Repairs, Auto Insurance
Shopping	Retail Stores
Eating	Food, Restaurants
Dancing	Clothing , Personal Care, Alcohol
Drinking	Alcohol
Social gathering	Clothing, Jewelry
Kids playing	Toys, Food and Drinks, Medicines, Clothing
Sports activities	Sports Related Products
Anxiety, Arguing	Pain Medications, Lawyers

(*examples only)

3.3 Types of Errors

3.3.1 False Positives

In this example, the tool misidentifies the bright light in the dark background as ‘fireworks’ (with a high confidence level).

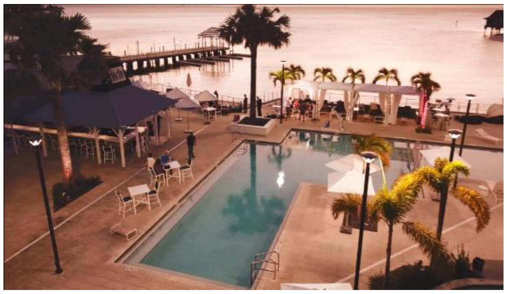


Figure 1 – False Positive - Fireworks

Table 2 – Machine Learning Detection and Error Mitigation

Detected Category	**Initial Confidence Level**	**New Confidence Level**
Fireworks has been detected from 00:00:02 to 00:00:03	90%	< 30%

In the “Proposed Solution’ section below we present a methodology to mitigate this issue.

3.3.2 False Negatives

In this example, the tool fails to identify the alcoholic beverages in the image analysis. However, the term ‘Cocktails’ is noted in the audio transcript as depicted in the JSON file (Figure 3).

In the “Proposed Solution’ section below we present a methodology to mitigate the false negative impact.

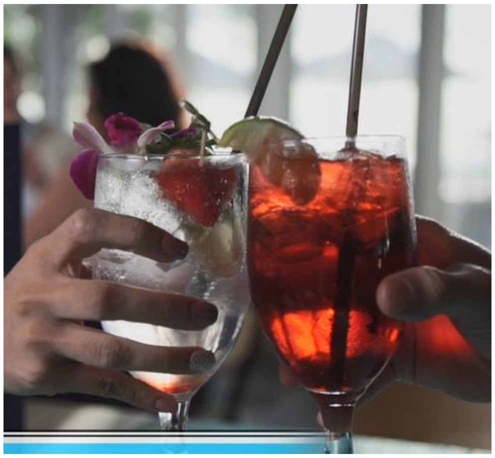
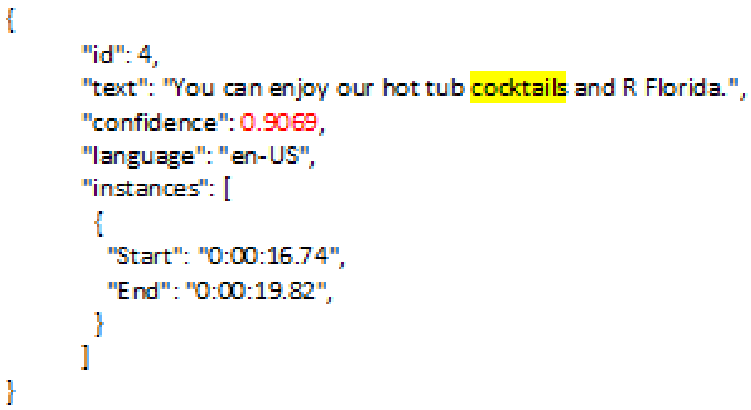


Figure 2 – False Negative - Alcoholic Beverage



```
{
        "id": 4,
        "text": "You can enjoy our hot tub cocktails and R Florida.",
        "confidence": 0.9069,
        "language": "en-US",
        "instances": [
         {
           "Start": "0:00:16.74",
           "End": "0:00:19.82",
         }
        ]
}
```

Figure 3 – JSON file of audio script of the parsed ad

The JSON file in Figure 3 indicates the word ‘cocktails’ as parsed from the audio transcript. This data is available even though the image analysis failed to recognize alcoholic beverage in the video.

3.3.3 Machine Learning Tool Performance

Another issue with some ML products is the excessive time taken for video analysis. In our studies, a 30-second ad would take 2-3 minutes for a multi-pass analysis. This can be improved substantially with faster GPU processors.

4 Proposed Solution

Current Machine learning products treat metadata of each stream separately; e.g. video/image analysis is separate from audio or text analysis, albeit each may use neural networks based classification algorithms. We believe that interrelating the video, audio and text data could enhance the accuracy of predictions. For example, a gambling ad for a casino may have telltale signs on video-audio-text streams. These accompanying signatures (supplementary/auxiliary data on multiple streams) are utilized by the software decision module introduced below.

The proposed solution consists of multiple stages. A modified workflow is introduced with an embedded decision module to accommodate heuristic analysis. ‘**Heuristic’** in the present context would mean an educated guess based on supplementary data. It is not a rigorous deterministic algorithm, but yields results in a reasonable time. Note that a signature-based deterministic approach is not guaranteed to work in the selected use cases. For example, some beer ads do not use the term ‘beer’ in the audio stream. In such cases, auxiliary signs in other streams (images of joy, relax, young people, bottles/cans, OCR data) could be strong clues.

4.1 Steps Summary

1. First pass is a general ML analysis to derive ‘content descriptors’.
2. Next, a heuristic analysis is performed using auxiliary data to assess the initial results.
3. The ‘confidence levels’ are reassessed and revised based on rules set.
4. Option to perform a more refined 2nd stage ML analysis for content classification.

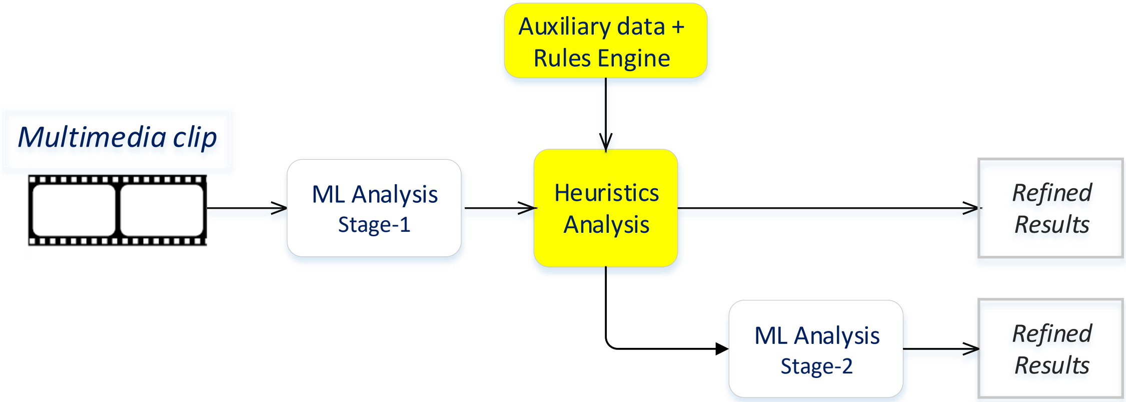


Figure 4 – Summary of Steps of the Proposed Solution

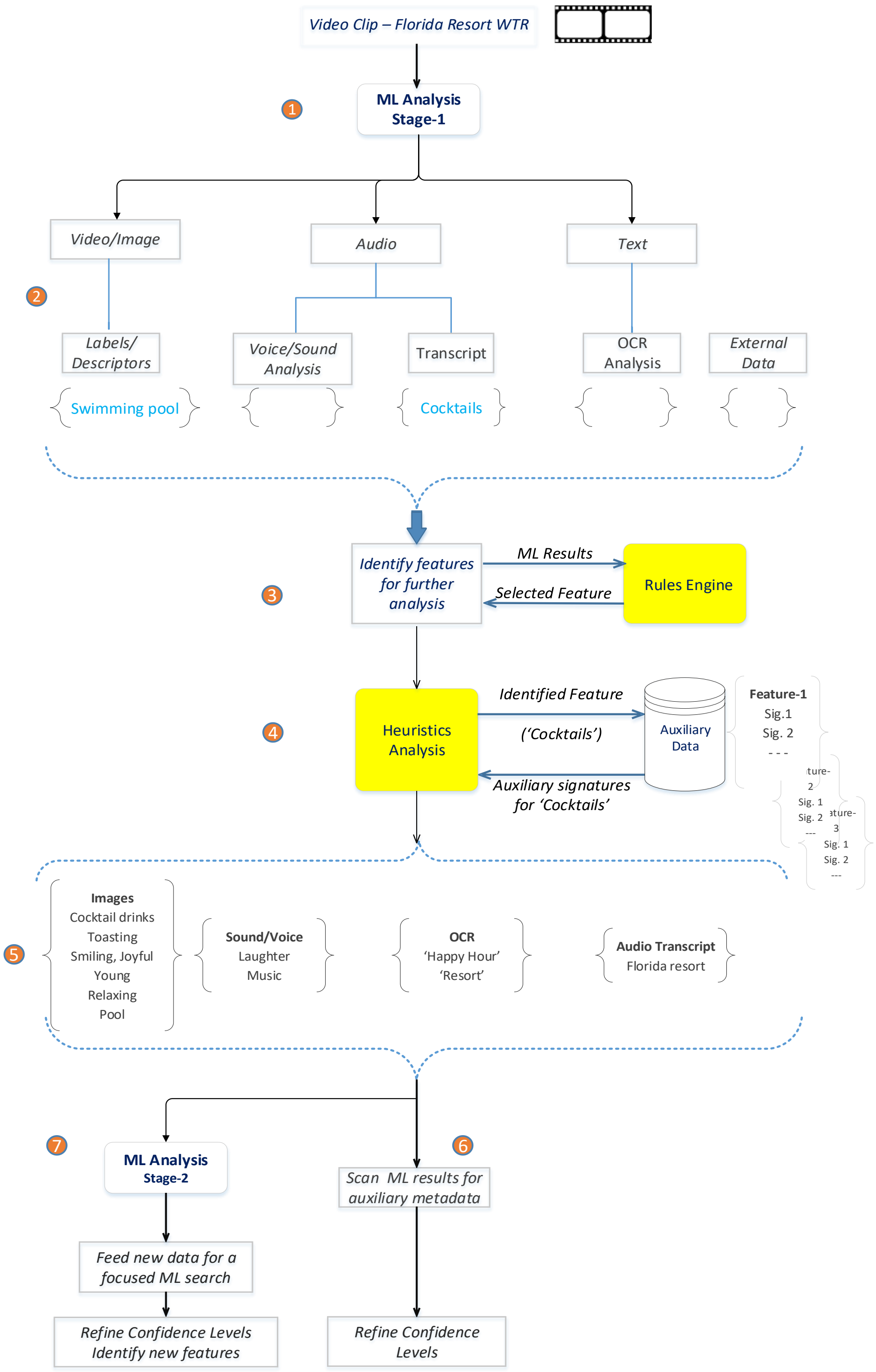


Figure 5 – Proposed Solution for Mitigating ML Results

4.2 Solution Details

(Numbering below corresponds to Figure 5)

1. Ad creatives/video content are fed into the ML engine (these could be 30-second Ads, long-form ads, TV episodes, movies or other multimedia content)

2. ML Engine conducts machine learning based analysis. The results are the identified content descriptors and confidence levels.

3. Identify features for further analysis. The criteria is based on rules built previously, such as a list of keywords.

4. Heuristic Analysis – Retrieve 'auxiliary data' for the feature identified above. These signify plausible signatures that may appear in other streams for a given feature.

5. This step depicts sample auxiliary signatures for the term 'Cocktails'. These are pre-populated in the database.

6. In this step, previous ML results (from the first pass) are fed into the software module programmatically. It searches for the presence of auxiliary data in other streams. Based on the analysis, 'confidence level' is adjusted. (i.e. If the metadata terms 'drinks', 'toasting' appear in the video analysis, the confidence level for 'Cocktails' is increased. Conversely, if there are no supporting auxiliary data, the confidence level is lowered.

7. Optionally, a second stage ML analysis is supported for a more refined search. Using Auxiliary data for the classification algorithm would enable a focused and accurate search.

Using the above process, the false positive/negative impacts are mitigated. Column 3 of Table-2 shows the results of applying heuristic analysis. In the case of fireworks, if there are no auxiliary signs on other streams (such as 'noise'), then the Confidence Level is lowered by a factor. The Rules Engine contains the pre-set value of the multiplier (e.g. 0.75)

Note that in the case of real 'fireworks', an image frame taken a second later would have the lights diminished. That heuristic signature could be used to differentiate fireworks from other lights and reduce false positives.

In the same fashion, if the auxiliary signatures are present in other streams (as in Figure 5), then the original confidence level is multiplied by a factor (e.g. 1.25) which would increase the final confidence level value.

Artificial Neural Networks loosely mimic the functioning of biological neurons. Extending the analogy a step further; when the human brain receives a plausible signature from one of the streams (visual, aural, olfactory, gustatory or haptic/tactile), the normal behavior of the brain is to seek supplementary evidence, i.e. auxiliary data from other streams, to validate its initial detection.

The proposed solution posits a similar functionality based on multi-stream analysis.

5 Conclusions

Based on our testing, visual analysis alone is not sufficient to make meaningful recommendations for carrier-class video (unlike surveillance or sports use cases). A multi-stream analysis of Video, Audio and Text (OCR) streams would provide a better contextual interpretation.

Machine learning applications to carrier-class video services is still a nascent field. We outlined some of the unique challenges. A multi-stream heuristic method was proposed to complement the current algorithmic approach. The MVPD space is a fertile field for AI/ML applications, and much work still needs to be done.

Abbreviations

AI/ML	Artificial Intelligence/Machine Learning
DL	Deep Learning
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short Term Memory
R-CNN	Region based Convolutional Neural Network
SSD	Single Shot Detector
TFLOP	Trillion floating-point operations per second
JSON	Java Script Object Notation

Bibliography & References

[1] FCC Guidelines for Ads - https://www.fcc.gov/consumers/guides/complaints-about-broadcast-advertising

[2] FTC Guidelines for Ads - https://www.ftc.gov/tips-advice/business-center/guidance/ftcs-endorsement-guides-what-people-are-asking

[3] FEC Guidelines for Ads - *https://www.fec.gov/help-candidates-and-committees/making-disbursements/advertising/*

[4] FDA Guidelines for Ads - https://www.fda.gov/media/82590/download

[5] ESPN Advertising Guidelines - http://www.espn.com/adspecs/guidelines/en/ESPN_AdStandardsGuidelines.pdf

[6] MIT AI Lab research - http://moments.csail.mit.edu/explore.html