Technical Disclosure Commons

Defensive Publications Series

September 12, 2018

SYSTEMS AND METHODS FOR DEEP LEARNING TELEVISION (TV) ADVERTISEMENT DETECTION

Devraj Mehta

Ravi Solanki

SYSTEMS AND METHODS FOR DEEP LEARNING TELEVISION (TV) ADVERTISEMENT DETECTION

Various systems and methods can identify whether an advertisement or a program is playing on a display device. For example, the systems and methods can distinguish between television advertisements and television programs displayed on a television. The systems and methods can maintain a library of videos of all known television advertisements. For example, the library may be a video database or videos of a video sharing network uploaded to the video sharing network by an advertiser. The systems and methods can index the videos to "memorize" an ordered sample of frames of the video. For a video being displayed on a television, the systems and methods can compare samples of the video to the memorized sample frames to determine whether the video is a program or an advertisement.

The systems and methods that rely on a video library may be limited in performance by the quality and quantity of the videos of the library. In order to improve the quality of the library of videos, systems and methods for an advertisement identifier can be used to identify advertisement videos and update the library with identified advertisement videos. The advertisement identifier can scan television broadcasts to identify repeating advertisement content that is a specific duration (e.g., a 30 second duration, a 60 second duration, etc.) that is aired on multiple channels during programs of different genres. The advertisement identifier can present the identified advertisement to a human labeler who can confirm whether the video is an advertisement and can add the video to the library.

Fundamentally, comparing pixels between multiple videos or against a video library is problematic since it is limited in scale. Using a library requires continuous updates while advertisement identification based on advertisement comparison requires for the advertisement identifier to view the advertisements over potentially lengthy periods of time. In this regard, deep learning systems and methods can be implemented to determine whether a video is an advertisement without requiring a library or requiring the comparison of multiple videos against each other.

Human beings can intuitively identify a video as an advertisement even if they have never viewed the video before. In this regard, computer vision systems can be configured to mimic this human intuition to identify videos based only on the current video and not on a comparison between multiple videos or a video library. The computer vision systems can, given pixels of a video frame, or a sequence of frames, and no other information, determine whether the frame is from an advertisement. As discussed herein, the computer vision systems and methods succeed in identifying advertisements based on nothing other than the pixels of the video frame.

The computer vision system discussed herein was tested and was found to be able to classify single frames from videos including television advertisements and television programs (non-advertisements) more accurately than human classifiers. Computer vision can be implemented with a single frame, "a single-frame classifier," that may disregard the time dimension. Then the single-frame classifier can be used as a building block feeding into higher-level classification models that may incorporate the time dimension. The higher-level models,

discussed later herein, can build upon the single-frame classifier to analyze sequences of frames. This "sequence classifier" can classify a video as an advertisement or a program by taking into account the time dimension.

Training, validation, and test data sets can be collected and used to train the single-frame classifier. The sets may each include labelled television frames sampled equally in number from advertisements and non-advertisements. A binary classifier e.g., a modified Inception V3 model convolutional neural net (CNN), can be trained to perform image classification. Based on this trained binary classifier, it was found that the single-frame classifier was superhuman in video classification.

Single-Frame Classifier

A single-frame classifier, as shown in Figure 1 below, can be implemented based on the Inception V3 pre-trained model. The pre-trained model may be modified in its topology to perform binary classification (i.e., classification between an advertisement and a non-advertisement). The single-frame classifier can be implemented by retraining the modified pre-trained model based on ground-truth data. The ground-truth data may be multiple advertisements and non-advertisements appropriately labeled as advertisement or non-advertisement. To perform advertisement classification, the pre-trained model can be modified by removing a classification head and replacing the classification head with one that has a custom classifier.

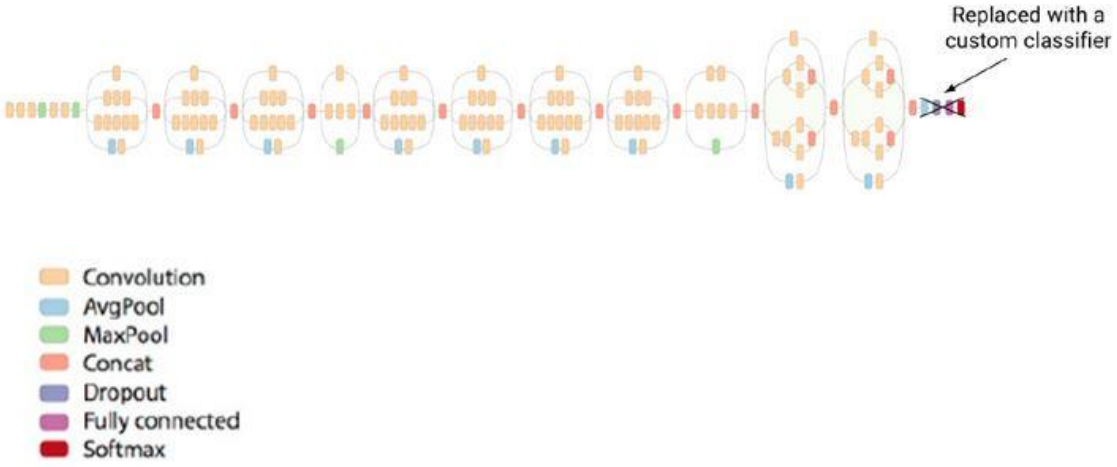

Figure 1 Single-Frame Classifier

First, a training dataset that contains television frames labelled as advertisements or non-advertisements can be collected. It may be important that the dataset has good coverage of a variety of program genres and television channels, although it does not need to require every advertisement creative. The dataset can be split between advertisements and non-advertisements and used to train the single-frame classifier. The output of the trained classifier may be a probability that an input television frame is not part of an advertisement.

Sequence Classifier

The single-frame classifier may be a building block for a more sophisticated sequence classifier configured to analyze sequences of frames and classify the sequences of frames as either an advertisement or a non-advertisement. The sequence classifier can be implemented as a "CNN-RNN," a hybrid convolutional neural network (CNN) and recurrent neural network

(RNN). Between the CNN and the RNN, image embedding of the input sequence of frames can be used.

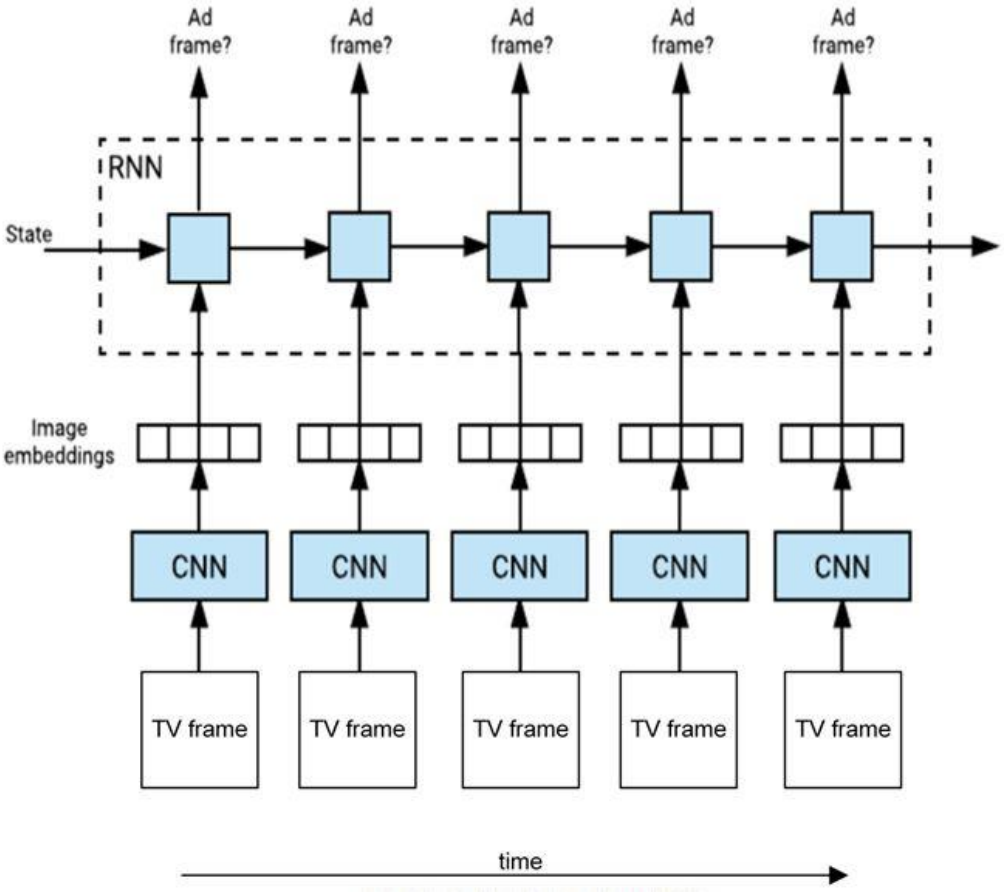

Figure 2 Sequence Classifier

The architecture shown in Figure 2 enables the sequence classifier to carry state through time. A modified version of the single-frame classifier can be the basis for the embedding model. Furthermore, based on the computer vision systems and methods described herein, clustering images based on visual characteristics can identify certain types of programming, e.g., news broadcasts.

Television Metering Accuracy

Audience measurement panels can be configured to measure viewership of television. Such a panel can measure viewership based on audio matching, where captured audio signals are matched against a reference database of television audio signals. One significant complication in some countries (e.g., the United States) is that television channels are often carried by multiple cable or satellite companies and have multiple local versions. One subsequent effect is that many of the advertisements in the original television channel feed are replaced by networks further down the broadcast chain with local advertisements. This is commonly referred to as local advertisement insertion.

The consequence of local advertisement insertion is that the audio matching systems are less accurate during advertisement breaks since the reference database may not include audio data for the local advertisements. To improve the audience measurement panel, a signal can be fed into the post-processing steps of the television audio matching system generated based on a television advertisement classifier (e.g., the single-frame classifier or the sequence classifier) to improve the accuracy of the audio matching system.

Television audience measurement panels can compare audio fingerprints against a library of reference television fingerprints. These matches can include both false positive and false negatives, and so a post-processing step, EDIT and BRIDGING rules, can be applied. Each television measurement company may develop their own set of post-processing rules (e.g., EDIT and BRIDGING rules) based on the expected performance of their own matching algorithm.

The system described herein is described with reference to EDIT and BRIDGING rules but can be applied to any type of post-processing rules.

Every fingerprint match can occur during a program or during an advertisement break. The systems and methods can assign different (lower) EDIT and BRIDGING weights to advertisements since classification of an advertisement is more susceptible to errors due to local advertisement insertion.

By using the classifier, the EDIT and/or BRIDGING rules can be adjusted such that there is a low probability that the advertisement panel spuriously decides that a user has switched channels during an advertisement break. However, the advertisement panel is still able to determine whether a user has actually changed channels during the break. A process for adjusting EDIT and BRIDGING rules is as follows:

1. Perform fingerprint matching (e.g., audio fingerprint matching) for an incoming frame against a database of reference frames (and/or audio fingerprints for the frames) to determine whether the incoming frame is a frame of an advertisement or a program.

2. For each of the references frames of the database, classify the reference frames as either an advertisement or a program with a classifier (e.g., a single-frame classifier and/or a sequence classifier).

3. If the incoming frame matches a reference frame classified as a program by the classifier, apply normal EDIT and/or BRIDGING rules. The applied EDIT

and/or BRIDGING rules giving strong weight to the incoming frame being a true positive match to the reference frame.

4. If the incoming frame matches a reference frame classified as an advertisement by the classifier, apply low EDIT and/or BRIDGING rules. The applied low EDIT and/or BRIDGING rules giving a low weight to the incoming frame being a true positive match to the reference frame.

5. If the incoming frame does not match any of the reference frames and previous frames are matched as advertisements, apply very low EDIT and/or BRIDGING rules since it is anticipated that it is currently an advertisement break and there is a higher chance of false matching during the advertisement break due to potential local advertisement insertion.

Local Advertisement Insertion False Positive Reduction System

Many applications rely on knowing all the advertisements on television, i.e., when the advertisements appear and on what channels. These applications can be configured to measure the effectiveness of the advertisements or can be configured to trigger other events to happen in a synchronized manner (for example enabling an on-line ad campaign or customizing a website).

In some embodiments, advertisement identification systems can match audio signatures and/or matching video signatures against a reference database of audio or video signatures. Unfortunately both of these systems suffer from false positives where some television content looks like an advertisement (e.g., are falsely classified by the fingerprint matching), even though the television content is not actually an advertisement. Common examples which trigger false

positives in video matching systems are when there is a head-and-shoulders view of a person in front of a plain background, which commonly happens in news and documentary programs.

To correct this, the video fingerprint of the program content can match similar advertisements for multiple seconds (which also commonly have a person talking to camera in front of a plain background). This invention aims to reduce the false positives by applying an additional signal generated based on an image classifier (e.g., the single-frame classifier and/or the sequence classifier) which specifies if the currently broadcast program is likely to be program content or an advertisement.

The system can be configured to classify multiple seconds of video to be as either an advertisement or a program in order to add hysteresis into the turning on and off of the ad detection system to cope with the program/ad detection system having some noise in its output. The system can perform a process with the single-frame classifier and/or the sequence classifier as described elsewhere herein. The process can include the steps:

1. Classify each incoming frame as a program or advertisement with a classifier (e.g., the single-frame classifier and/or the sequence classifier).

2. If the frame is classified as a program return to step 1.

3. Determine how many seconds have been classified as an advertisement. If the number of seconds is less than three seconds, return to step 1 and store the classified advertisement frames in a buffer.

4. Continue performing steps 1-3 by feeding the buffered frames and subsequent live frames into the single-frame classifier and/or the sequence

classifier until 3 seconds of frames have been classified as an advertisement. In response to three seconds of frames being classified as an advertisement, classify the video as an advertisement.

ABSTRACT

The method detailed herein relates to a classifier for classifying a video as an advertisement or program without using a reference database of frames. The method can include classifying videos with a single-frame classifier that analyzes a single frame to determine whether the frame is an advertisement or a program. Furthermore, the method includes classifying videos with a sequence classifier that analyzes sequences of frames to determine whether the frame is an advertisement or a program. In some embodiments, the method relates to classifying predefined amounts of frames as advertisements or programs to determine whether a video is an advertisement or a program. In some embodiments, the method relates to classifying frames of a reference database of an advertisement matching system in order to adjust various analysis weights so that the performance of the advertisement matching system does not fall due to locally inserted advertisements that may not be part of the reference database.