Neurocomputing 000 (2017) 1–7

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Brief papers

Adult content detection in videos with convolutional and recurrent neural networks

Jônatas Wehrmann[a], Gabriel S. Simões[a], Rodrigo C. Barros[a,*], Victor F. Cavalcante[b]

[a] Pontifícia Universidade Católica do Rio Grande do Sul, Av. Ipiranga, 6681, Porto Alegre 90619-900, RS, Brazil
[b] Motorola Mobility, R&D-Brazil, Rodovia SP 340 Km 128.7, Jaguariuna 13820-000, SP, Brazil

ARTICLE INFO

Article history:
Received 6 September 2016
Revised 3 March 2017
Accepted 6 July 2017
Available online xxx

Communicated by Dr Xiang Xiang Bai

Keywords:
Adult content detection
Deep learning
Convolutional neural networks
Recurrent neural networks

ABSTRACT

The amount of adult content on the Internet grows daily. Much of the pornographic content is unconstrained and freely-available for all users, requiring parents to make use of parental control strategies for protecting their children. Current parental control devices depend on human intervention, and hence there is the need of computational approaches for automatically detecting and blocking pornographic content. Toward that goal, this paper proposes *ACORDE*, a novel deep learning architecture that comprises both convolutional neural networks and LSTM recurrent networks for adult content detection in videos. Experiments over the freely-available NPDI dataset show that *ACORDE* significantly outperforms the previous state-of-the-art approaches for this task, decreasing by half the number of false positives and by a third the number of false negatives.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

The automatic detection of adult (pornographic) content in images and videos is an important and challenging task, especially due to the huge amount of freely-available adult content on the web, whose spread has significantly increased with the massive adoption of mobile devices across the globe. A recent report[1] indicates that the Internet traffic to porn websites accounted for 8.5% of the total in the UK in June 2013, surpassing the traffic for shopping, news, business, and social networks.

Even though organizations such as MPAA[2] have developed rating systems to protect viewers from adult scenes in motion pictures, content available on the web is practically unconstrained and easy-to-access, motivating the development of computational approaches that are capable of automatically detecting pornography with the final goal of protecting sensitive populations (e.g., children under 18). The task of automatically identifying adult content, however, poses a greater challenge than other classification problems due to the degree of subjectivity and uncertainty surrounding the problem. For instance, it is hard even for human beings to properly assess degrees of sensuality in scenes where people wear swimsuits or underwear. Indeed, sometimes more than one image/frame is needed for contextualizing the scene in order to define whether it should be classified as adult content or not.

Earlier work on pornography identification focused on human skin detection [1–4], in which the idea is that greater amounts of detected skin would lead to higher probabilities of nudity within the image or video, hence characterizing the content as pornographic. Nevertheless, these approaches suffer with a high rate of false positives, especially in the context of beaches or practice of aquatic sports. More recent studies [5–8] approached the problem under the perspective of *Bag of Visual Words* (BoW) and similar models (e.g., BossaNova [8,9]) for aggregating (quantizing) sophisticated image descriptors.

For benchmarking the proposed approaches in the area in terms of both video and image detection, researchers have used the NPDI dataset [8]. The best results achieved in NPDI are described by [10], where the authors propose a video descriptor based on binary features (*BinBoost* [11]) which is used with the BoW/BossaNova representations. However, the very same approach reaches only 44.6% of mean average precision (mAP) in the well-known PASCAL VOC dataset [12], while recent deep learning approaches reach about 60% of mAP in that same dataset [13]. This is a clear indication that deep learning based approaches could be a good option for pornography detection in both images and videos.

Therefore, in this paper we propose a novel approach for adult content detection in videos, namely *ACORDE* (Adult Content Recognition with Deep Neural Networks). Its architecture makes use of a convolutional neural network (ConvNet) as a feature extrac-

* Corresponding author.
E-mail addresses: jonatas.wehrmann@acad.pucrs.br (J. Wehrmann), gabriel.simoes.001@acad.pucrs.br (G.S. Simões), rodrigo.barros@pucrs.br (R.C. Barros), victorfc@motorola.com (V.F. Cavalcante).

[1] http://goo.gl/nG0s7n.
[2] http://www.mpaa.org/.

http://dx.doi.org/10.1016/j.neucom.2017.07.012
0925-2312/© 2017 Elsevier B.V. All rights reserved.

(a) Easy non-adult class. (b) Hard non-adult class. (c) Adult class.

Fig. 1. Frames from the NPDI dataset.

tor and of a long short-term memory (LSTM) to perform the final video classification. *ACORDE* extracts feature vectors from the video *keyframes* of NPDI, building a sorted set of semantic descriptors. This set is used to feed the LSTM that is responsible for analyzing the video in an end-to-end fashion. The proposed approach does not require fine-tuning nor re-training the ConvNet. Results show that *ACORDE* comfortably establishes the new state-of-the-art for adult content detection in NPDI, reducing by half the number of false positives and by a third the number of false negatives.

This paper is organized as follows. Section 2 briefly introduces the NPDI dataset as well as recent methods for pornographic classification of videos. Section 3 describes our proposed approach in detail. Section 4 presents how the experimental setup was organized for performing the empirical analysis, which is presented in Section 5. Finally, in Section 6 we detail our conclusions and future work directions.

2. Background

This section discusses earlier work that performs adult content detection, and also describes the NPDI dataset, which will be used to validate our novel approach.

2.1. NPDI dataset

Currently, the largest publicly-available pornographic dataset is NPDI [9], which comprises nearly 80 h from 802 videos (half of them with adult content), all downloaded from the Internet. The non-adult class is further sub-divided in 201 easy-to-classify videos and 200 hard-to-classify videos. The latter were selected based on textual search queries like *beach, wrestling*, and *swimming*, in order to verify the ability of the proposed classifiers in scenarios of high skin-exposure. The adult class comprises 401 videos selected from adult content web sites. Fig. 1 shows a sample of frames from the easy non-adult, hard non-adult, and adult classes.

As described in [9], a scene segmentation algorithm was employed to extract keyframes from the videos, resulting in a total of 16,727 images. Each video may contain 1–320 keyframes. The average amount of keyframes per class are: 15.6 for adult videos; 33.8 for easy non-adult videos; and 17.5 for hard non-adult videos. NPDI has a wide ethnic diversity with asian, black, white, and multi-ethnic videos. Issues like one-keyframe videos and *anime*-style content considerably increase the challenge of NPDI.

2.2. Related work

In the work of [9] and [10], the authors make use of both low and mid-level visual features extracted from the NPDI dataset. They use such features to build a final movie representation. The method is based in a low-complexity alternative for feature extraction using binary descriptors and a combination of mid-level representations. They aggregate the descriptors via the BoW model [14], generating the *BoW video descriptor* (BoW-VD). Also, they use the BossaNova method [15], which is an improved extension of the BoW model, generating the *BossaNova video descriptor* (BNVD). BNVD is a video descriptor that represents the median distance for each visual word of a given *codeword*[3] for a *codebook*.[4]

The work of [16] is the first to use deep neural networks to address the pornography detection problem. That work proposes a method that requires fine-tuning two distinct ConvNets, namely *AlexNet* [17] and *GoogLeNet* [18]. The author performs the training phase by reusing models pre-trained over the ImageNet dataset [19] and fine-tunes them over NPDI. That approach requires the training of ten distinct models: one model per training fold (5) and per network (2). Keyframes were rescaled to 256×256 to allow the data augmentation process with crops of the size 224×224 randomly sampled from each image in order to avoid overfitting. To normalize the data, the author subtracted the mean image from all instances.

Unfortunately, several methodological aspects are not clearly detailed in the paper, such as: (i) the stopping criteria adopted, (ii) the usage of a validation set, (iii) values of important hyper-parameters like learning rate, momentum, and regularization; and (iv) the updated layers in each model. Note that the absence of a proper validation set may compromise the reliability of the results. For the test phase, each network predicts *benign* (non-adult) and *adult* probabilities for each keyframe. The probabilities from both models are averaged, and a video is classified as adult (benign) when most of its keyframes are predicted as belonging to the *adult* (*benign*) class.

3. ACORDE

In this paper we propose a novel method for adult content detection in images and videos, namely *ACORDE* (Adult Content Recognition with Deep Neural Networks). The architecture of *ACORDE* comprises a convolutional neural network (ConvNet) [20] for feature extraction and a long short-term memory network (LSTM) for sequence learning [21]. ConvNets are the current state-of-the-art for many computer vision tasks such as image classification [22], object detection [13], video analysis [23–26] and image segmentation [27]. LSTMs are well suited to learn representations of sequences such as videos and texts. The conjoint use of both algorithms has been used to solve problems in video analysis [28] and scene captioning [29].

A ConvNet is a deep learning strategy that combines three ideas to ensure some degree of shifting, scale, and distortion invariance regarding the image content: local receptive fields (filters), shared weights, and spatial (or temporal) pooling [30]. The convolution operator is applied in order to replace fully-connected matrix multiplications, granting the two first mentioned ideas and considerably reducing the amount of parameters within a network. Convolutional filters are learned using the well-known backpropagation algorithm [31]. This process can be seen as *representation learning*, i.e., the network acting as a feature extractor. Learning representations from images is vital for the success of the computer vision task at hand. Eq. (1) defines a convolution, where (x, y) is a position on the jth feature map from the ith network layer; m indexes

[3] A codeword is the centroid of a given visual words cluster.
[4] A codebook is a set of codewords.

Please cite this article as: J. Wehrmann et al., Adult content detection in videos with convolutional and recurrent neural networks, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2017.07.012

Download English Version:

https://daneshyari.com/en/article/6865306

Download Persian Version:

https://daneshyari.com/article/6865306

Daneshyari.com