A Web-Based Service for Disturbing Image Detection

Markos Zampoglou[1(✉)], Symeon Papadopoulos[1], Yiannis Kompatsiaris[1], and Jochen Spangenberg[2]

[1] CERTH-ITI, Thessaloniki, Greece
{markzampoglou,papadop,ikom}@iti.gr
[2] Deutsche Welle, Berlin, Germany
jochen.spangenberg@dw.com

Abstract. As User Generated Content takes up an increasing share of the total Internet multimedia traffic, it becomes increasingly important to protect users (be they consumers or professionals, such as journalists) from potentially traumatizing content that is accessible on the web. In this demonstration, we present a web service that can identify disturbing or graphic content in images. The service can be used by platforms for filtering or to warn users prior to exposing them to such content. We evaluate the performance of the service and propose solutions towards extending the training dataset and thus further improving the performance of the service, while minimizing emotional distress to human annotators.

1 Introduction

With the proliferation of social media and capturing devices, Internet users are increasingly exposed to User Generated Content (UGC) uploaded by other users. In such environments, it is possible that a user may join a public platform and upload content that can be traumatizing to others, such as pornographic or violent imagery. While most platforms provide reporting mechanisms that allow users to request the removal of such content, users can only report content after they have been exposed to it. Furthermore, professionals such as journalists or police officers spend a lot of time browsing and collecting UGC in search of information. Due to the nature of their work, such content may often be violent or disturbing, and prolonged exposure to such content can be psychologically traumatizing. Automatically detecting such content and forewarning the user before displaying it could help protect both professional and casual users, not as a form of censorship but as a personal tool for preventing emotional trauma.

2 Background

We use the term *disturbing images* to refer to *depictions of humans or animals subjected to violence, harm, and suffering, in a manner that can cause trauma to the viewer*. Research suggests that users who are systematically exposed to such material are in danger of serious psychological harm [2][1]. In research literature,

[1] Also see http://dartcenter.org/.

L. Amsaleg et al. (Eds.): MMM 2017, Part II, LNCS 10133, pp. 438–441, 2017.
DOI: 10.1007/978-3-319-51814-5_37

Fig. 1. Sample *violent* images that do not qualify as *disturbing* for our task.

the term *violence detection* refers to a very active field, e.g. [1,7]. Yet such works usually concern video (exploiting audio and motion cues), and the content they aim to detect is quite different from the one discussed in this work, as their definitions of violence often include, e.g. individuals fighting or weapons being fired. A recent attempt at detecting violent images [8] is based on a dataset collected using Google image search with violence-related keywords, many of which *violent* images would not be considered *disturbing* given our definition (Fig. 1). Finally, another recent work aims at detecting *horror* images [5], which however often lack the realism of real-world *disturbing* UGC.

3 Disturbing Image Detection Service

3.1 Back-End

Dataset Collection and Organization. As, to our knowledge, this is the first effort in tackling this task, we had to build an annotated dataset in order to train and evaluate our service. The dataset consists primarily of images collected from the web. As the manual dataset generation process is psychologically demanding for annotators we used a two-step approach: An initial set was collected manually from the web, depicting war zones, accidents and other similar situations, with and without disturbing content. Adding around 100 non-disturbing images from the UCID dataset [6] and following manual annotation, we ended up with 990 images. We then used this initial set to build the rest of the dataset semi-automatically: we trained a classifier using the approach described below and employed automated scripts to download images from a number of websites specializing in disturbing and/or graphic pictures. Images were automatically classified as `disturbing` or `non-disturbing`, and the annotators only had to visually correct the results. This process proved a lot less stressful than manually annotating all images. The final dataset contains 5401 images, 2043 of which are labeled as `disturbing`. Due to their graphic nature we decided against demonstrating image samples, but will share the dataset upon request.

Classification Approach. With respect to classification, we decided to use Convolutional Neural Networks (CNN), as they have exhibited exceptional performance in many image classification challenges in recent years. However, given the relatively small size of our dataset, we cannot hope to train such a classifier from scratch. In such cases we can take a model that is pre-trained on

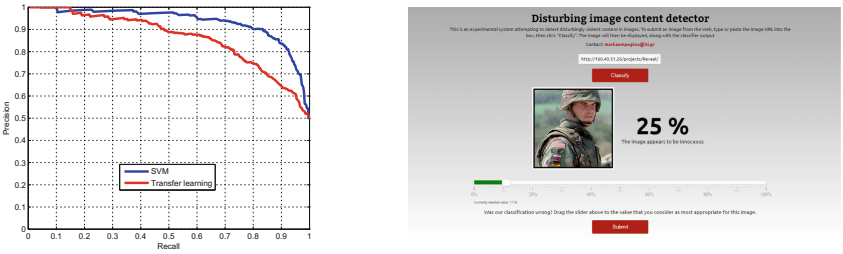

Fig. 2. Left: Results of classification using the two approaches. Right: A screenshot from the developed web User Interface.

a general set of concepts (e.g. publicly available networks trained on the ImageNet dataset) and either use the output of the second-to-top layer as an image descriptor for a standard classifier (e.g. SVM), or re-train only the upper layers of the network using our data, in an approach known as *transfer learning*. We attempted both approaches using the *BVLC Reference CaffeNet* network[2] [4] from the Caffe framework [3]. Figure 2 (left) shows that, given our current dataset, SVM classification with CNN features worked best, reaching 0.864 Precision at 0.868 Recall, versus 0.769 Precision and 0.763 Recall by the transfer learning-based CNN model. Thus we use the SVM approach for our demo.

Web Service. We used Caffe to make the classification model accessible via a REST service. The back-end is built in Python (PyCaffe) and Java. It accepts an image URL, downloads and runs the image through the classifier, and returns the prediction as JSON. If the incoming load overcomes the system limits, a queuing mechanism allows the service to respond asynchronously to meet the demands of batch classifying images in large collections. The API also offers a feedback mechanism, which accepts an image URL and the correct annotation for an image. This feature can be used both to improve classifier performance, but also as a first step towards a personalized classifier that can take subjective preferences into account. It should be noted that in its current form the system accuracy drops significantly when classifying thumbnails (<200 pixels in either dimension), thus it is currently aimed at images above that size.

3.2 Front-End

The web front-end provides a simple interface for the classification and feedback services[3]. A user submits an image URL and the service visualizes the classifier prediction as a percentage on a bar, indicating the probability that the image has disturbing content. The bar allows the user to submit their own perception on what the prediction should have been, in case it differs from the one provided.

[2] Downloaded from http://caffe.berkeleyvision.org/model_zoo.html.
[3] The demo is available at http://reveal-mklab.iti.gr/reveal/disturbing/.

In that case the image is stored alongside the value to be used in refining the classifier. Figure 2 (right) shows the classification result for a web image.

4 Conclusions and Future Steps

We presented a public service for the analysis and potential filtering of disturbing images using state-of-the-art technologies on a novel dataset, with feedback mechanisms to further improve future results by incorporating user annotations. In order to further refine our dataset, given the problems of manually annotating large numbers of disturbing images, we are trying to incorporate semi-automatic and crowd-sourcing methods to disperse the emotional burden of the task. Given the -partly- subjective nature of the task, we are also considering the potential for personalized classification. We also believe there is significant application potential if the interface could take the form of a browser plug-in that would be able to analyze all images to be displayed on the browser. Finally, we are considering the possibility of a localization system that can isolate only the offending part of the image -however, manually annotating such a dataset for training and evaluation would incur considerable emotional toll on annotators, and how to reduce this toll through semi-automatic means is an open issue.

Acknowledgements. This work is supported by the REVEAL and InVID projects, partially funded by the European Commission under contract numbers 610928 and 687786. In addition, we would like to acknowledge the support that NVIDIA provided us through the GPU Grant Program.

References

1. Déniz, O., Serrano, I., et al.: Fast violence detection in video. In: Battiato, S., Braz, J. (eds.) VISAPP 2014, vol. 2, pp. 478–485. SciTePress, Setúbal (2014)
2. Dubberley, S., Griffin, E., Bal, H.M.: Making secondary trauma a primary issue: a study of eyewitness media and vicarious trauma on the digital frontline (2015). http://eyewitnessmediahub.com/uploads/browser/files/Trauma%20Report.pdf
3. Jia, Y., Shelhamer, E., et al.: Caffe: convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
4. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS 2012, pp. 1097–1105 (2012)
5. Li, B., Xiong, W., et al.: Horror image recognition based on context-aware multi-instance learning. IEEE Trans. Image Process. **24**, 5193–5205 (2015)
6. Schaefer, G., Stich, M.: UCID: an uncompressed color image database. In: Storage and Retrieval Methods and Applications for Multimedia, vol. 5307, pp. 472–480. SPIE (2004)
7. Sjberg, M., Ionescu, B., et al.: The MediaEval 2014 affect task: violent scenes detection. In: MediaEval (2014)
8. Wang, D., Zhang, Z., Wang, W., Wang, L., Tan, T.: Baseline results for violence detection in still images. In: AVSS, pp. 54–57 (2012)