Classification of Not Suitable for Work Images: A Deep Learning Approach for Arquivo.pt

Daniel Bicho[a][b]
daniel.bicho@gmail.com

Artur Ferreira[b][c]
artur.ferreira@isel.pt

Nuno Datia[b][d]
nuno.datia@isel.pt

[a] Arquivo.pt, and
[b] ISEL, Instituto Superior de Engenharia de Lisboa
Instituto Politécnico de Lisboa
[c] Instituto de Telecomunicações

[d] NovaLincs, FCT, Universidade Nova de Lisboa

Abstract

Arquivo.pt is a Web Archiving initiative, storing contents preserved from the .pt Web Pages. Among these contents, there are many image files. Some of these images explicitly nudity and pornography, which are offensive for the users, and thus are Not Suitable For Work (NSFW) images. In this paper, we propose a solution to classify NSFW images on Arquivo.pt, using deep learning approaches. We set up a dataset of images with Arquivo.pt data and the ResNet and SqueezeNet models, are evaluated and fine tuned for the NSFW classification task. These models reported an accuracy of 93% and 72%, respectively. After a fine tuning stage, the accuracy of these models improved to 94% and 89%, respectively. This solution is available at `https://arquivo.pt/images.jsp`.

1 Introduction

The collection of portions of the World Wide Web (WWW) to preserve information is named as Web Archiving (WA). This preservation keeps old and historical information available for future use by the general public. Typically, Web Archives resort to Web crawlers, such as Heritrix [10], to collect the web contents. These contents include many resource types, and among them we often find images and videos. There are different WA initiatives, such as the European Commission Historical Archives, `https://ec.europa.eu/historical_archives`, the national top-level domain UK Web Archive, `https://www.webarchive.org.uk/ukwa`, or the Internet Archive, `https://archive.org`. Arquivo.pt, `https://arquivo.pt`, is a WA initiative to preserve the Portuguese *.pt* top-level domain. It provides a research infrastructure, making its contents searchable and publicly available in open access. Arquivo.pt provides a full-text search system and an Image Search Service (ISS) to browse to all its data. This service enables image retrieval capabilities to Arquivo.pt , with an interface in which users can perform queries in natural language and the service retrieves images related to the user query.

One portion of the images stored at Arquivo.pt are Not Suitable For Work (NSFW) for most users, because they contain offensive or explicit images (such as naked persons, violence, and pornography). This may be caused, for instance, by a website that got hacked for Web spam before it was crawled and its contents retrieved. Thus, we need to avoid the exposure to these types of contents, mainly for children and young persons. An example of this NSFW contents retrieved, using the ISS is depicted in Figure 1, in which the query term *angela* was fulfilled on the ISS, and some retrieved results can be considered offensive. In this paper, we propose an approach to filter out nudity/pornography content from the Arquivo.pt resources, through a binary classification task. The remainder of this paper is organized as follows. Section 2 briefly reviews related work. Section 3 presents our approach. The experimental results and some concluding remarks are reported in Section 4.

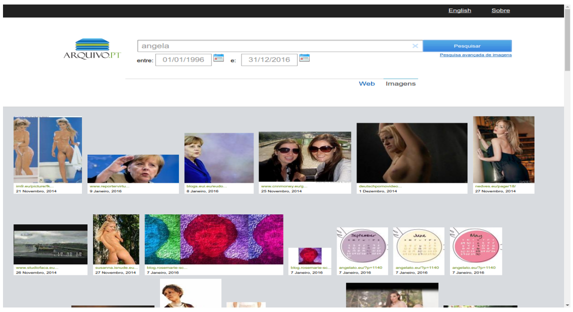

Figure 1: Example of Arquivo.pt problematic content, retrieved with the ISS using the query term *angela*.

2 Not Suitable For Work Image Classification

There are many approaches to the problem of identify NSFW content from images [3, 4, 9, 14]. Among these methods, we can find the first techniques based on skin detection. Another type of techniques are based on Bag-of-Visual-Words (BoVW) and more recently Neural Networks (NN) and Deep Learning (DL) techniques have been proposed.

An automatic system to detect human nudes was present in an image was proposed [4]. It resorts to methods to mark skin-like pixels combined with color and texture properties. These marked regions are then analyzed by a specialized grouper, to group a human figure using geometric constraints on the human structure, followed by classification. The POESIA filter [2], is an open source implementation of a skin-color-based filter. These methods present high false positive rates in images related to beach as well as sports activities.

Another approach is the Bag-of-Visual-Words (BoVW) [3], which extracts, a set of visual features represented as words, setting up a vocabulary vector with the number of occurrences of these visual words representing local image features. Those features usually are derived from detecting keypoints or local descriptors variations. A classifier that uses these representations is then trained to classify the image content.

Recently, Deep Neural Networks (DNN) and more specifically Convolutional Neural Networks (CNN) showed state of the art results, for image recognition tasks. For instance, CNN have been widely used on image recognition tasks [8, 12], for NSFW image classification [13, 15]. Many different CNN architectures have been published with improved accuracy on the standard ImageNet classification challenge [11].

3 Proposed Approach

3.1 Building a dataset

We started by building a dataset of NSFW images and its opposite with 17 655 images, manually labelled from Arquivo.pt with 8 273 labelled as NSFW and 9 382 as SFW, as described in Table 1.

Table 1: Evaluation Dataset.

	SFW	NSFW	Total
Labelled Dataset	9 382	8 273	17 655
Non-Labelled Dataset	-	-	18 626

These images were acquired from Arquivo.pt using two methods. The first method was through the existent Beta Images Indexes, which are Lucene indexes provided by the Solr platform[1]. The second method was through Arquivo.pt Text Search API [1], querying the API to retrieve Web pages, and from those Web pages the images contained were extracted to be manually labelled. On Arquivo.pt the total of images that belong to the NSFW class is much less than the images from the SFW class. The main difficulty at this task is to find enough images from the NSFW class, in order to build a dataset with a significant number of images and with both classes balanced in terms of the number of images. A large number of noisy images were being returned, such as image banners and icons.

[1] `http://lucene.apache.org/solr/`

To reduce the amount of this type of images, only images with resolution above 150x150 pixels were considered in the experimental results reported in Section 4.

3.2 Proposed solution and integration on Arquivo.pt

After building the dataset, we have considered two different topologies of DNN, namely ResNet [7] and SqueezeNet networks [5]. We have addressed the problem as a binary classification task, and thus we used the Cross-Entropy as the loss function [6] and the Stochastic Gradient Descent (SGD) algorithm as the optimizer, using transfer learning. The developed solution to classify image contents as NSFW is integrated in the Arquivo.pt ISS indexing workflow, to extract images and related metadata. The integration is modular, and can be extended by changing the underlying model. It also supports a real time classification, exposed as a Web Service. Figure 2 shows an example of the solution integration, showing a case of a misclassified image by the NSFW classifier (hidden by a gray rectangular box). Figure 3 shows the outcome of the same query, without using the NSFW classifier.

4 Experimental Evaluation and Discussion

The hardware used to evaluate these models is a common laptop with 8 GB RAM, a GeForce GTX 860M as GPU and a Intel(R) Core(TM) i7-4710HQ CPU @ 2.50 GHz. The models were also tested using server class hardware available at Arquivo.pt infrastructure. The server is a Dell PowerEdge R730xd model with 256 GB RAM and an Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.4 GHz. Table 2 reports the best experimental results (accuracy and loss) on the NSFWSqueezeNet model, while Table 3 does the same for the OpenNSFW model.

Table 2: NSFWSqueezeNet fine-tuning accuracy and loss.

Model	4-Fold Accuracy	4-Fold Loss
NSFWSqueezeNet 1 Ep. Aug. 10K	0.88 ± 0.002	0.28 ± 0.004
NSFWSqueezeNet 1 Ep. Aug. 10K	0.89 ± 0.002	0.27 ± 0.006

There is a significant accuracy improvement, from the initial model accuracy of 72% to 89%, after a fine tuning stage in which all the network layers are retrained, using as starting point the network weights from a pre-trained model. The OpenNSFW model is computationally more expensive to train. With the limited hardware available and time constraints, an attempt to improve it was made, freezing all the network layers and retraining only the last fully-connected and the softmax layers. The solver used was also the SGD with the same learning parameters as the model above. The OpenNSFW model provides better results than the SqueezeNet model.

Table 3: OpenNSFW fine-tuning accuracy and loss.

Model	4-Fold Accuracy	4-Fold Loss
OpenNSFW 1 Ep. Aug. 10K	0.92 ± 0.003	0.20 ± 0.006
OpenNSFW 5 Ep. Aug. 10K	0.94 ± 0.004	0.16 ± 0.007

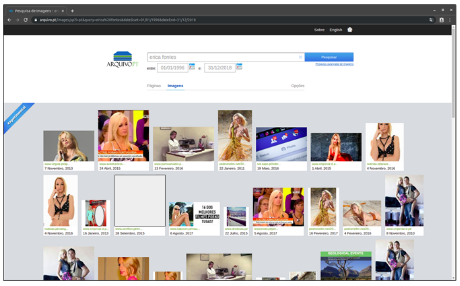

Figure 2: Image filtering interface integration, with query term 'Erica Fontes', with the NSFW classifier. The gray rectangular box highlights NSFW contents which were misclassified.

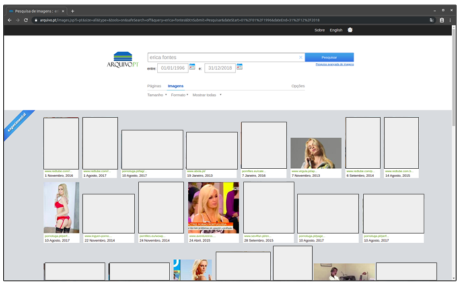

Figure 3: Image filtering interface integration, with query term 'Erica Fontes', without the NSFW classifier. The gray rectangular boxes correspond to NSFW contents (hidden for proper display).

In summary, in this paper we have briefly described a solution that automatically identifies not suitable for work images, which was integrated into Arquivo.pt infrastructure. The solution uses a convolutional neural network to identify this content type and provides the classification result which is used to hide not suitable for work contents, from the retrieved results. The proposed solution is currently available at `https://arquivo.pt/image.jsp`. As future work, the model's accuracy can be improved by building a larger dataset and considering the categorization with more classes.

References

[1] Arquivo.pt. Arquivo.pt API v.0.2 (beta version), https://github.com/arquivo/pwa-technologies/wiki/Arquivo.pt-API-v.0.2-(beta-version), March 2018.

[2] M. Daoudi. POESIA - Filtering Software @ONLINE, January 2018. URL `http://www.poesia-filter.org:80/`.

[3] T. Deselaers, L. Pimenidis, and H. Ney. Bag-of-visual-words models for adult image classification and filtering. In *2008 19th International Conference on Pattern Recognition*, pages 1–4, Dec 2008. doi: 10.1109/ICPR.2008.4761366.

[4] D. Forsyth and M. Fleck. Automatic detection of human nudes. *International Journal of Computer Vision*, 32(1):63–77, Aug 1999. ISSN 1573-1405. doi: 10.1023/A:1008145029462. URL `https://doi.org/10.1023/A:1008145029462`.

[5] F. Iandola, M. Moskewicz, K. Ashraf, S. Hang, W. Dally, and K. Keutzer. SqueezeNet. *arXiv, 1602.07360*, 2016. ISSN 0302-9743. doi: 10.1007/978-3-319-24553-9.

[6] K. Janocha and W. Czarnecki. On loss functions for deep neural networks in classification. *CoRR*, abs/1702.05659, 2017.

[7] H. Kaiming, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.90.

[8] A. Krizhevsky, I. Sulskever, and G. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information and Processing Systems (NIPS)*, pages 1–9, 2012.

[9] T. Lindeberg. Scale Invariant Feature Transform. *Scholarpedia*, 7 (5):10491, 2012. doi: 10.4249/scholarpedia.10491.

[10] G. Mohr, M. Stack, I. Ranitovic, D. Avery, and M. Kimpton. Introduction to Heritrix, an archival quality web crawler. In *4th International Web Archiving Workshop (IWAW04)*, Bath, UK, 2004.

[11] Stanford University. ImageNet, http://www.image-net.org/, January 2018.

[12] Y. Sun, B. Xue, M. Zhang, and G. G. Yen. Evolving deep convolutional neural networks for image classification. *IEEE Transactions on Evolutionary Computation*, 24(2):394–407, 2020.

[13] D. Zhelonkin and N. Karpov. Training effective model for real-time detection of nsfw photos and drawings. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 301–312. Springer, 2019.

[14] H. Zheng, M. Daoudi, and B. Jedynak. Blocking Adult Images Based on Statistical Skin Detection. *Electronic Letters on Computer Vision and Image Analysis*, 4(2):1–14, 2004. ISSN 1577-5097. doi: 10.5565/rev/elcvia.78.

[15] R. Zhu, X. Wu, B. Zhu, and L. Song. Application of pornographic images recognition based on depth learning. In *Proceedings of the 2018 International Conference on Information Science and System*, pages 152–155, 2018.