Combining CNNs for detecting pornography in the absence of labeled training data

Jongbin Jung
jongbin@stanford.edu

Rahul Makhijani
rahulmj@stanford.edu

Arthur Morlot
amorlot@stanford.edu

Abstract

We use deep learning to identify pornographic videos of different age groups, with the goal of applying our approach in detecting child pornography. We propose a two step hierarchical procedure — in the first step, we use an ensemble of CNNs for image classification to classify videos into porn or non-porn. In the second step we apply face detection and age classification to infer the age groups represented in different images. Given a video, we predict whether it portrays pornographic imagery of certain age groups by combining the output of the previous two steps. Our proposed method achieves 42% accuracy on a test set of four `porn/age group` categories — 1.7 times higher than a random guess.

1. Introduction

The detection of problematic content such as child pornography has implications and applications for both business and criminal justice. Not only would law enforcement agencies benefit from high precision detection of illegal content, but it would also be in the best interest of companies to detect, and if necessary, censor and report illegal activity on their platform. However, efforts to manually review potentially harmful content accompany many risks such as unwanted exposure and law suits [1]. While such a task seems ideal for the application of deep learning techniques, implementation is often hindered by the difficulty in obtaining legitimate training data, among a variety of other challenges facing child pornography detection [2]. In this project, we explore the possibility of devising a method to detect whether a video contains child pornography, without requiring any images, labeled or otherwise, of child pornography as training data.

Our approach is to divide the problem of classifying child pornography into two independent and more tractable subproblems for which training data exists. We divide the twp problem into a hierarchical classification problem:

- classifying whether an image is pornographic or not
- identifying whether the pornographic image contains children.

We exploit existing state-of-the-art image classification techniques for each of the two step process.

2. Related work

Our problem can be described in three modular parts:

1. classifying videos as either porn or not porn,
2. detection and age classification of people (faces), and
3. combining the two models of porn and age detection.

Most of the computer vision approaches for detecting pornographic images use the NPDI dataset provided by Avila, Thome, Cord, *et al.* [4]. The data was originally collected for training and testing a bag-of-words approach to porn detection, based on nudity-related features [4]. While there have been studies reporting 98% accuracy on detecting porn with CNNs [5], state-of-the-art performance with the NPDI dataset seems to be at 94.1% accuracy. [6] Note that since "porn" can be a rather ambiguous categorization, configuration of the test set could make higher performance trivial. For example, if all the negative examples in the test data were a random sample of videos from YouTube, it might be the case that a simple skin-color detection scheme could achieve very high performance. The NPDI dataset takes this into account, explicitly including "easy" non-porn videos, as well as "difficult" non-porn videos (see Figure 1 for some example frames). Given that it is unclear what data was used for the research reporting 98% accuracy [5], we will focus on comparing our implementation with the 94.1% state-of-the-art accuracy reported on the NPDI dataset.

The problem of age detection carries its own challenges, and the state-of-the-art leaves much to be desired [7]. For the age detection portion of our project, we plan to use data from the Face Image Project [8] for training (Figure 3). We will base our age-detection model on a CNN model developed by Levi and Hassner [9]. The model is trained on facial images that are labeled with broadly defined age categories, e.g., (0–7), (8–12), (12–20), (21–24), (25–30), etc.

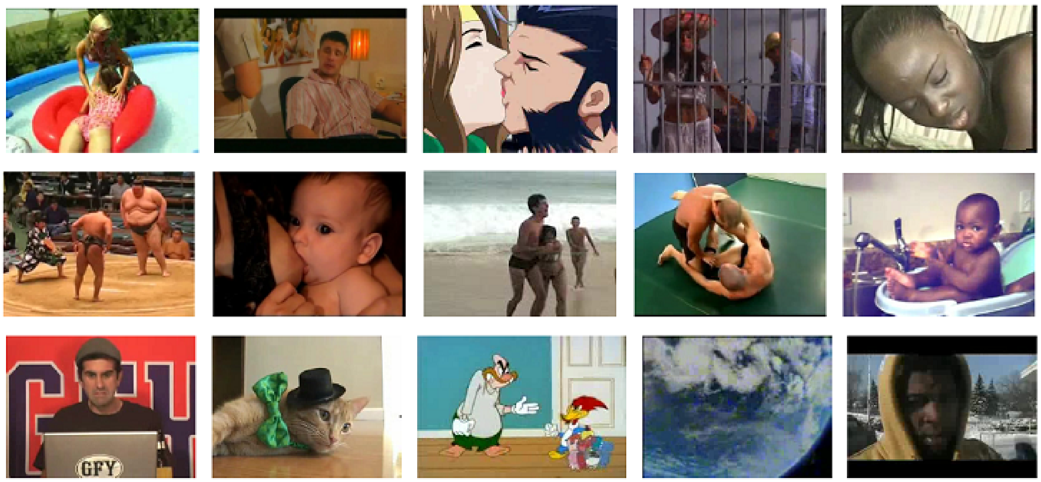

Figure 1. Some example frames from the NPDI Pornography Database[3] illustrating the diversity of pornographic videos (top row), challenges of "difficult" non-pornographic videos (middle row), and some "easy" non-pornographic videos.

While we realize that detecting age groups via facial detection might not be optimal or feasible in the context of porn — where it is unclear how often a face is visible — we have decided to implement a face-based age detection scheme in the interest of time.

While there been qualitatively similar approaches to hierarchically modeling a prediction problem to exploit structure in the target labels [10], we have yet to find prior work that apply the idea to mitigating lack of labeled training data.

3. Methods

The overall steps of our proposed method are outlined below:

1. Given a video, extract key frames as static images
2. Use standard image classification CNNs to classify each frame as either porn or non-porn
3. Take the majority vote of each individual frame to determine whether a video is porn or not
4. With the frames extracted in Step 1, detect faces — if any — in each frame
5. Use an age classification CNN [9] to determine whether certain age groups of interest (e.g., pre-teen, teen, adult) are present in each frame
6. Combine the predictions from Steps 3 and 5 to determine whether a video falls into the class of interest, i.e., porn depicting children.

For the purpose of this project, with realistic constraints in mind, we assume employing standard tools and techniques for Step 1 of the proposed method. This is operationalized by using the frames that have already been extracted for the existing porn dataset, and constructing the test dataset as a collection of frames[1].

In each of the classification steps, in lieu of building our own network from scratch, we extend existing state-of-the-art methods in each problem category (i.e., image classification and age detection) via transfer learning. Specifically for age detection (Steps 4 and 5), we employ an off-the-shelf implementation of `YOLO` [11], combined with transfer learning with labeled facial data.

4. Experiment setup

Here, we provide details on how each of the proposed steps were implemented. When training models, we consistently use an `Adam` solver, starting with learning rate `1e-3`, and exponential decay with decay rate of 0.94, every two epochs. In Section 4.3, we discuss some challenges of testing our proposed method in the context of child pornography, and how we address them.

4.1. Porn detection

For the first part of our proposed method, detecting porn, we using the NPDI dataset provided by Avila, Thome, Cord, *et al.* [4]. While there have been studies reporting 98% accuracy on detecting porn with CNNs [5], state-of-the-art performance with the NPDI dataset seems to be at 95% accuracy. [6] Given that "porn" can be a rather ambiguous categorization, we note that configuration of the test set could make higher performance trivial. To account for this, the NPDI dataset explicitly divides the non-porn category into "easy" and "difficult". Some example frames of each cat-

[1] We provide more detail on manual test data collection in Section 4.3

egory in the NPDI dataset is provided in Figure 1. Since it is unclear what data was used for the research reporting 98% accuracy [5], we focus on comparing our implementation with the 95% state-of-the-art accuracy reported on the NPDI dataset.

Table 1. Summary of the NPDI Pornography Database.

Class	Videos	Hours	Shots/Video
Porn	400	57	15.6
Non-porn (easy)	200	11.5	33.8
Non-porn (difficult)	200	8.5	17.5

4.1.1 Porn classifier training

We use an ensemble of transfer learning with pretrained state-of-the-art image classification models [12]–[15]. The models we include, and corresponding layers that were retrained for porn classification are summarized in Table 2. In implementing and adapting the pretrained models, to the specific porn detection problem, we used the data generation scripts provided with the pretrained models as boiler plates for processing the porn classification training data appropriately for each architecture.

Table 2. Pretrained image classification models.

Model name	Retrained layers
`VGG16`	`vgg_16/fc8`
`ResNET152`	`resnet_v2_152/logits`

`InceptionV4`	`InceptionV4/logits,`
	`InceptionV4/AuxLogits`

4.1.2 Aggregate porn prediction

After computing predictions for whether each *frame* of a video is porn or not, the final prediction of whether a *video* is porn or not is computed by averaging predicted probabilities over all classifiers and frames. To be precise, denote video i as V_i with N_i static frames extracted, and let p_{ij}^m be the prediction from model m, on probability scale[2], of whether frame j of video i is porn or not, where $0 \geq p_{ij}^m \geq 1$. Then, we compute

$$\Pr(V_i = \texttt{porn}) = (N_i M)^{-1} \sum_m \sum_j p_{ij}^m,$$

where $M = 3$ is the number of models in the ensemble. An example of this process is illustrated in Figure 2.

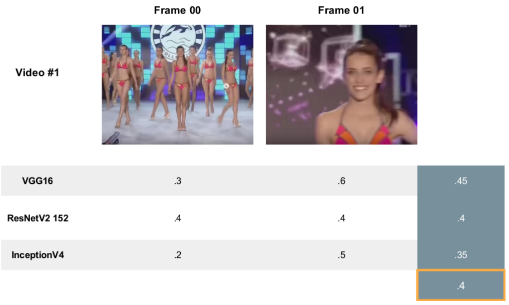

Figure 2. Example of video classification.

4.2. Age Detection

The problem of age detection carries its own challenges, and the state-of-the-art leaves much to be desired [7]. For the age detection portion of our project, use data from the Face Image Project [8] for training (Figure 3). We base our age-detection model on a CNN model developed by Levi and Hassner [9] and the open-source implementation for Tensorflow [16]. Actual implementation of the model involves a facial detection network, coupled with a separate age detection network, since the age detection network assumes that there is a face within the presented image. While we realize that detecting age groups via facial detection might not be optimal or feasible in the context of porn — where it is unclear how often a face is visible — we have decided to implement a face-based age detection scheme in the interest of time, but it could be improved to take into account other features in the image, leading to more accurate age detection.

4.2.1 Face detection

To first detect faces from any given image, we used a state of the art implementation of YOLO ("You Only Look Once") which evaluates all possible crops in a single pass. We adopt an open-source implementation of YOLO [16]. This allows us efficiently process each image, while maintaining model accuracy—in terms of facial detection—as illustrated in the Figure 4.

4.2.2 Age detection

For the purpose of age detection, we used once again transfer learning. To satisfy time constraints, we limit the model two a single `InceptionV4` network. The face detection project [8] provides a pretrained checkpoint for this purpose. We fine tuned this checkpoint using the data from the face detection project and used this as our final age detection model.

[2] Given logit scores from any model, we use a inverse-logit transform to compute scores in probability scale.

Figure 3. Some example images from the age database [8] that highlight the challenges of age captioning such as lighting, multiple faces, angle, and occlusion

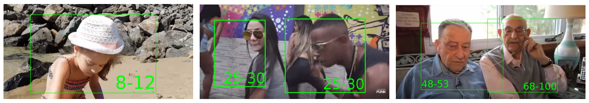

Figure 4. Example of age detection with YOLO on sample test frames.

4.2.3 Aggregate age prediction

After computing predictions for the age group of faces detected in each *frame* of a video, the final prediction for the age group of a *video* as a whole is computed by aggregating predicted probabilities over all faces detected in all frames. To be precise, denote video i as V_i with F_i faces detected across all frames, and let q_{ik} be the prediction for face k, on probability scale, of whether face k of video i is considered `young` or not, where $0 \geq q_{ik} \geq 1$. Then, we compute

$$\Pr(V_i = \texttt{young}) = (F_i)^{-1} \sum_k q_{ik}.$$

4.3. Inference

Although one of the core contributions of our project would be to eliminate the need for acquiring possibly illegal data for training, ideally, actual data of child pornography would be necessary to evaluate the true performance of our final result, i.e., we would need labeled test data. Despite our efforts to contact law enforcement agencies in order to acquire legitimate access to test data, we were not be able to acquire such data by the end of the project deadline. In lieu of labeled child pornography test data, we manually collected a dataset of frames labeled `porn-young`, `porn-old`, `nonporn-young`, `nonporn-old`. As with the NPDI pornography dataset, in order to account for the ambiguity and challenge of accurately identifying pornographic content with *intent*, we explicitly collect easy and difficult sub-categories of `nonporn` labels.

One shortcoming and challenge of this manual test data collection was in appropriately determining the somewhat subjective labels. For consistency, we limited the sources of videos to two websites: YouTube and PornHub, and strictly labeled any videos from YouTube as `nonporn`, and any videos from PornHub as `porn`. However, while collecting "difficult" `nonporn` examples, we noticed some content from YouTube which we would subjectively have labeled as `porn` (see Figure 5 for an example). Hence, there were cases in which our "ground truth label", as determined by the source of the video, was in conflict with the "sensible" label we would expect a human to apply, and we notice that such ambiguity seemed to contribute to a drop in performance of porn classification on the test data, i.e., our porn classification and personal opinion on a `nonporn` video would agree that the video should be categorized as `porn`.

We further realize that the manually collected database does not exactly represent the problem which we wish to solve — given the lack of an explicit `Porn/Child` category — but we believe it would serve as an informative proxy in the absence of legally obtainable data that would

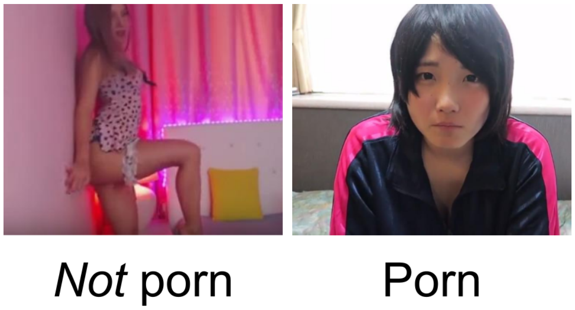

Figure 5. An example of the difficulty in test data collection. The image on the left is a frame collected from YouTube, thus categorized as *not* porn, while the image on the right is a frame collected from PornHub, and thus categorized as porn.

allow us to test the effectiveness of our proposed approach.

Table 3. Summary of video (frames) collected for testing.

Class	Young	Old
Porn	35 (549)	24 (435)
Non-porn (easy)	20 (327)	18 (296)
Non-porn (difficult)	20 (327)	20 (330)

5. Results

5.1. Initial model performance

First we evaluate performance of the porn classification ensemble on (1) a holdout test set from the NPDI data and (2) the manually collected test data. Our porn classifier achieved 94% AUC on the NPDI data — close to state-of-the-art reported [6] for that dataset (94.1%) — and 89% AUC on our custom test data. Note, given that we manually collected the test data, there were ambiguities involved in determining whether some porn-like videos should be labeled porn or not. Our ultimate rule was to limit the "porn" category to videos that were collected from known porn websites, but such ambiguity might explain lower performance on the test data[3]. The age classification model achieved 70% AUC on the test set, in classifying videos as either "young" or "old" categories.

5.2. Final prediction performance

Given two probabilities p_i and q_i, the predicted probability that video i is `porn` and in the `young` age group, respectively, we compute the probability of each image belonging to one of the four porn/age group categories by taking the maximum of the four products, $p_i q_i$, $p_i(1 - q_i)$, $(1 - p_i)q_i$, and $(1 - p_i)(1 - q_i)$. Using this method, we are able to achieve an accuracy of 42.36%, about 1.7 times higher than a random guess. We provide category specific metrics in Table 4.

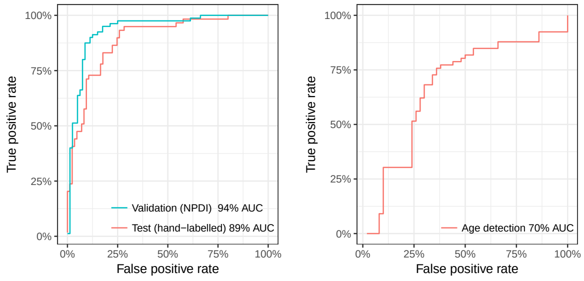

Figure 6. ROC curve for porn (left) and age (right) detection.

6. Conclusion and Future Work

6.1. Conclusion

In this project, we've shown that a problem in which no labeled training data is directly available (e.g., child pornography), we can still construct a meaningful classification tool by exploiting structure of the original problem, and readily available data that address these subproblems. While we weren't able to test performance of our approach on the original problem of detecting child pornography, due to lack of available *test* data, we believe that the proxy problem that we have constructed of detecting different age groups in porn/nonporn contexts provides a meaningful proof-of-concept.

Perhaps one non-obvious benefit of our approach is that once we've broken down the original problem into smaller subproblems, performance of the final model is easier to diagnose. For example, from this project, it is clear that the overall detection of age groups in pornographic images could benefit greatly by improving the age detection, while pornography detection seems to be quite accurate. Noticing that this project has been limited to the time and resources available to us within the contex of this course, we further address some shortcomings which we are aware of, and would be good starting points for future expansion of this work.

6.2. Future work

Improving age detection. The strength of our approach is that by breaking the original problem down, we are able to diagnose our final model. Currently, the weak link of our pipeline is the age detection part, as it's accuracy is significantly lower than the detection of pornographic content. We have several options that we have been considering for improving on this benchmark :

- Use more than 1 model to make a decision: as with de-

[3] we've decided not to display explicit images, we're quite convinced that many people would have a hard time classifying some of the "difficult" non-porn images collected from YouTube.

Table 4. Summary of metrics for final porn/age classification.

Metric	Porn/Young	Porn/Old	Not Porn/Young	Not Porn/Old
Precision	40.43%	25.42%	76.19%	64.71%
Recall	54.29%	62.50%	35.56%	27.50%

tecting pornographic content, we could use `ResNET`, `VGG` and `Inception` to vote on each frame and make a decision, which should improve our results.

- Use non-facial features: age can sometimes be inferred from non-facial features (like wrinkles on hands). While the state-of-the-art in age detection still seems to be based on facial features [17], we strongly believe that incorporating other bodily features will enhance the accuracy of age detection—especially in the context of pornographic images where a subject's body is often more visible than their face.
- Use of the video aspect: Videos have an internal coherence. Therefore, if all the frames of the video except one says that the age of the protagonists is 25 and only one frame rates 82, it is most probably a detection error. By using the information of frames that are close in time, we should be able to improve the accuracy of the network.

Implement temporal features. In the current iteration of this project, we have limited the input to static frames extracted from each video. We do not take into account that the frames within a single video are correlated with the target label. While it is unclear whether using temporal information will improve porn classification — given that we are on par with state-of-the-art methods that do indeed consider temporal information — this approach could further help the age detection model.

Implement a parallel pipeline for real-time analysis. Currently, each frame of the image takes around 5 seconds to be processed by the network. Although this is pretty fast, this does not enable real time analysis of the data. We would like to implement a parallel reader, in which several networks will run at the same time. Each of them will evaluate a different frame, allowing it to act much closer to real time for the analysis of videos.

References

[1] Telegraph. (Jan. 12, 2017). Two former microsoft employees 'developed PTSD after watching child porn as part of their job', [Online]. Available: `http://www.telegraph.co.uk/news/2017/01/12/two-former-microsoft-employees-developed-ptsd-watching-child/`.

[2] G. Thompson, "Automatic detection of child pornography," 2009.

[3] S. Avila, E. Valle, and A. D. A. AraúJo. (May 15, 2017). NPDI pornography database, [Online]. Available: `https://sites.google.com/site/pornographydatabase/`.

[4] S. Avila, N. Thome, M. Cord, E. Valle, and A. D. A. AraúJo, "Pooling in image representation: The visual codeword point of view," *Computer Vision and Image Understanding*, vol. 117, no. 5, pp. 453–465, 2013.

[5] F. Nian, T. Li, Y. Wang, M. Xu, and J. Wu, "Pornographic image detection utilizing deep convolutional neural networks," *Neurocomputing*, vol. 210, pp. 283 –293, 2016.

[6] M. Moustafa, "Applying deep learning to classify pornographic images and videos," *ArXiv preprint arXiv:1511.08899*, 2015.

[7] S. Thakur and L. Verma, "Identification of face age range group using neural network," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 5, pp. 250–254, 2012.

[8] T. Hassner. (May 15, 2017). Face image project, [Online]. Available: `http://www.openu.ac.il/home/hassner/Adience/data.html`.

[9] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 34–42.

[10] R. Brant, "Assessing proportionality in the proportional odds model for ordinal logistic regression," *Biometrics*, pp. 1171–1178, 1990.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[12] Github. (Jun. 1, 2017). Pretrained image classification models, [Online]. Available: `https://github.com/tensorflow/models/tree/master/slim#pre-trained-models`.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ArXiv preprint arXiv:1409.1556*, 2014.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[15] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, inception-resnet and the impact of residual connections on learning," *CoRR*, vol. abs/1602.07261, 2016. [Online]. Available: `http://arxiv.org/abs/1602.07261`.

[16] Github. (Jun. 1, 2017). Age/gender detection in tensorflow, [Online]. Available: `https://github.com/dpressel/rude-carnie`.

[17] A. Dehghan, E. G. Ortiz, G. Shu, and S. Z. Masood, "Dager: Deep age, gender and emotion recognition using convolutional neural networks," *ArXiv preprint arXiv:1702.04280*, 2017.