Combining CNNs for Hierarchical Classification in the Absence of Labeled Training Data

Jongbin Jung, Rahul Makhijani, Arthur Morlot

Stanford University

Problem

"Can we use CNNs to detect child pornography?"
(without any training examples)

Although image classification might seem like an elementary task, even the best methods rely heavily on context-specific training data. But for certain tasks — like identifying child pornography — it may be either infeasible or undesirable to generate/acquire such data.
Here, we explore the possibility of combining models trained on existing data to solve problems where no training data exists. Given that actually classifying child pornography would be difficult to validate (even without training data, we would require a dataset to test performance), we explore our ideas on a substitute problem: "can we detect different age groups of porn, without labeled training data of that domain?"

Data

We used the NPDI dataset provided by Avila, Thome, Cord, et al to train the porn classifier.

Table 1: Summary of the NPDI Pornography Database.

Class	Videos	Frames
Porn	400	6387
Non-porn (easy)	200	5170
Non-porn (difficult)	200	5170

For testing purposes, we manually collected frames of 144 videos which we organize into six categories, as shown below.

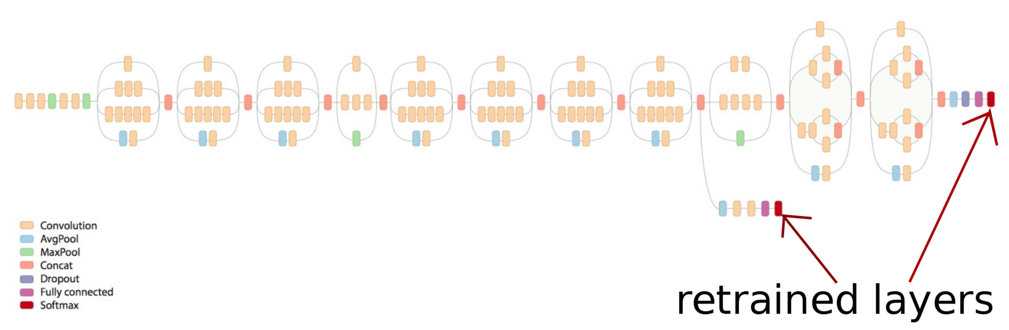
Table 2: Summary of video (frames) collected for testing.

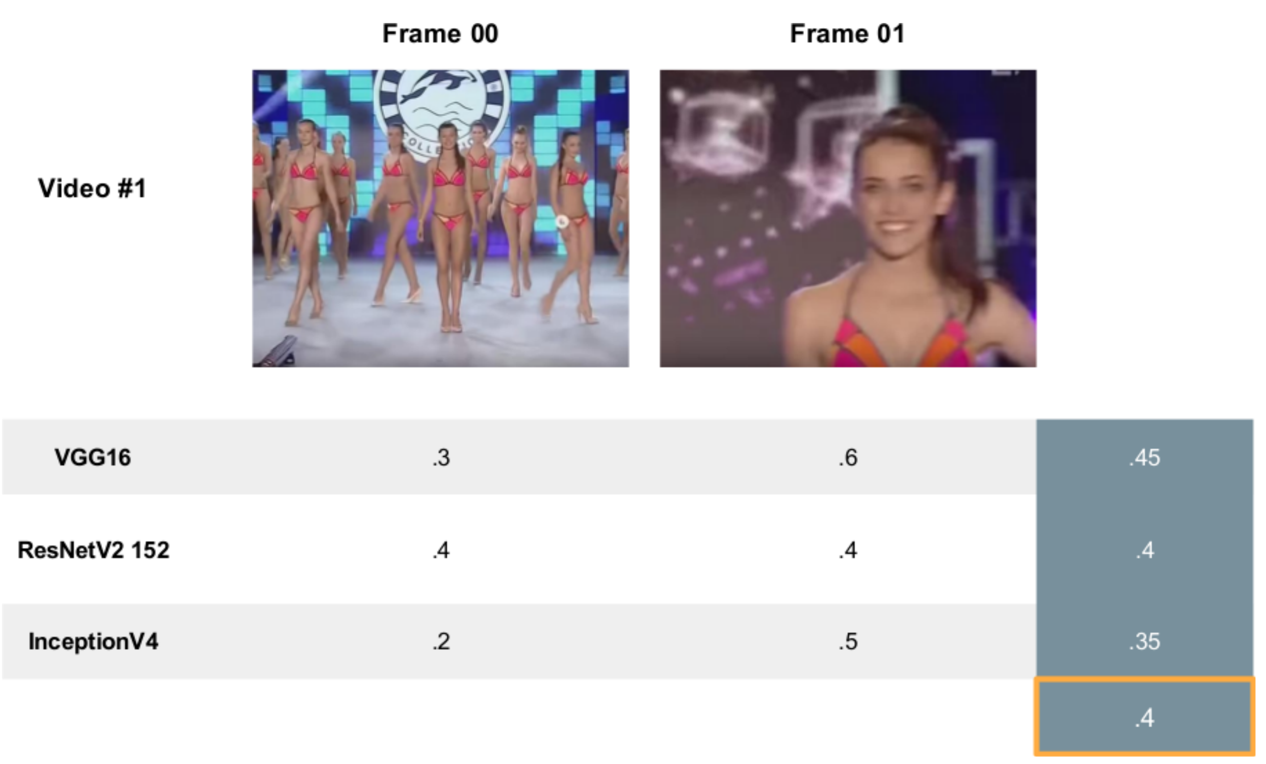
Class	Young	Old
Porn	35 (549)	24 (435)
Non-porn (easy)	20 (327)	18 (296)
Non-porn (difficult)	20 (327)	20 (330)

Methods

Hierarchical approach The method we propose is to divide an original "hard" problem in to smaller tasks for which training data is available. In this case, we separate to two tasks: (1) classifying images as pornographic and (2) detecting specific age groups in images.

Pornography classification For pornography classification, given that this is a basic image classification task, we take full advantage of existing pre-trained models. Specifically, we re-train just the last layers of vgg, resnet, and inception classifiers [3]. 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.





Age detection We use a pre-trained deep neural network, YOLO, to detect and draw bounding boxes around faces, and train the last layer of an inception network to predict the age of each detected face. As with porn, labeled data is readily available for facial age group detection [2]. The steps to compute age group class scores for each video are: (1) divide predicted age into groups of "young" ($\leq$ 20), "old" ($\geq$ 30), and "undetermined", (2) discard all "undetermined" predictions, (3) average predictions across faces detected in each frame, and (4) average those values across frames.

Results

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 [1] 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[a]. The age classification model achieved 70% AUC on the test set, in classifying videos as either "young" or "old" categories.

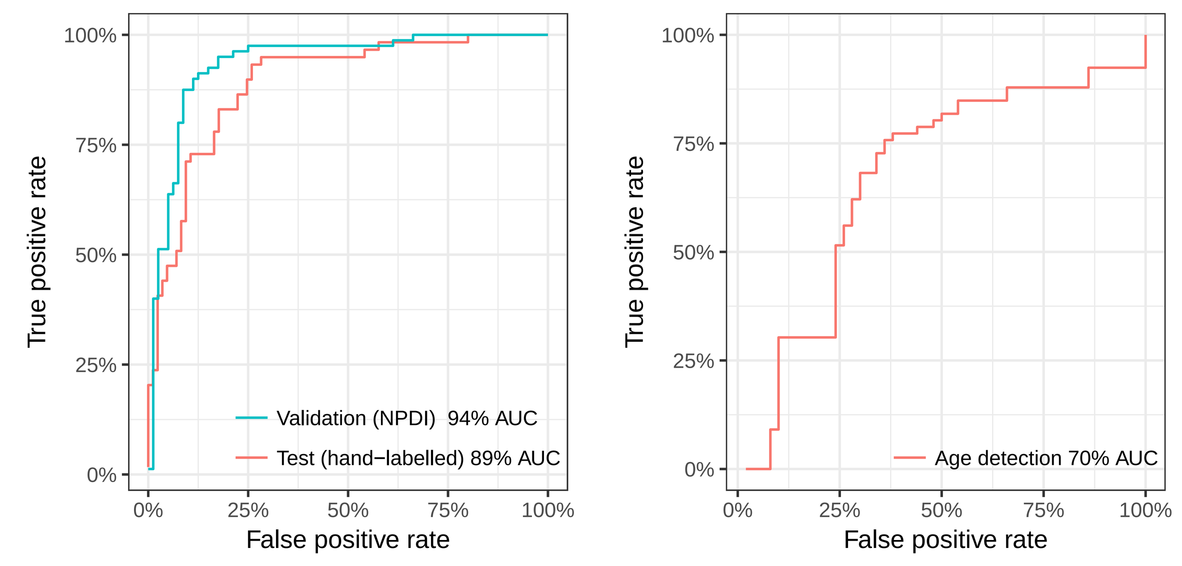


Figure 1: ROC curve for porn (left) and age (right) detection.

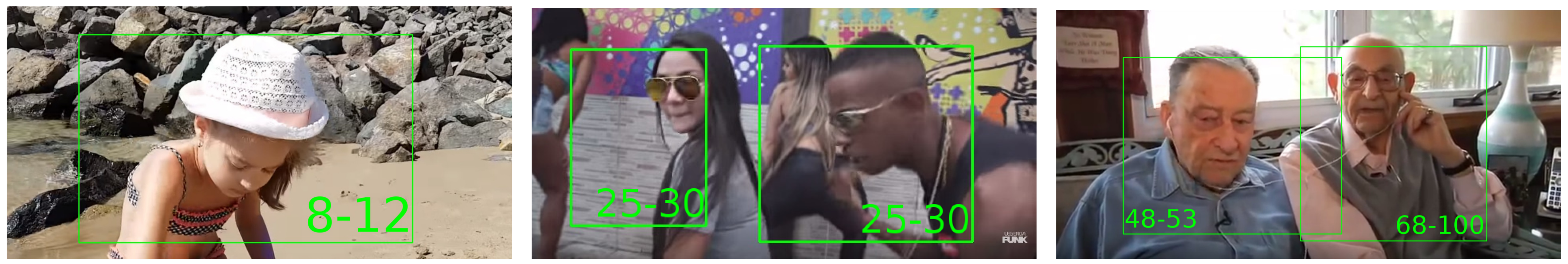


Figure 2: Example of age detection with YOLO on sample test frames.

Final predictions Given two probabilities p and q, the predicted probability that a video 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, pq, $p(1-q)$, $(1-p)q$, and $(1-p)(1-q)$. Using this method, we are able to achieve an accuracy of 42.36%, about 1.7 times higher than a random guess.

Table 3: 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%

[a] While 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.

Future work

- Improve age detection via non-facial features
- Clarify definitions of `porn/age group`
- Fortify prediction aggregation scheme
- Exploit sequential nature of videos
- Implement ensembles for age detection
- Re-train more layers during transfer learning

References

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

[2] 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.

[3] Pretrained image classification models: `https://github.com/tensorflow/models/tree/master/slim#pre-trained-models`