University of Rhode Island

DigitalCommons@URI

Open Access Master's Theses

2019

Deep Learning of Human Apparent Age for the Detection of Sexually Exploitative Imagery of Children

Jared Rondeau

University of Rhode Island, jaredrondeau1@gmail.com

Follow this and additional works at: <https://digitalcommons.uri.edu/theses>

Recommended Citation

Rondeau, Jared, "Deep Learning of Human Apparent Age for the Detection of Sexually Exploitative Imagery of Children" (2019). *Open Access Master's Theses*. Paper 1438.
<https://digitalcommons.uri.edu/theses/1438>

This Thesis is brought to you for free and open access by DigitalCommons@URI. It has been accepted for inclusion in Open Access Master's Theses by an authorized administrator of DigitalCommons@URI. For more information, please contact digitalcommons-group@uri.edu.

DEEP LEARNING OF HUMAN APPARENT AGE FOR THE DETECTION
OF SEXUALLY EXPLOITATIVE IMAGERY OF CHILDREN

BY

JARED RONDEAU

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
COMPUTER SCIENCE

UNIVERSITY OF RHODE ISLAND

2019

MASTER OF SCIENCE THESIS
OF
JARED RONDEAU

APPROVED:

Thesis Committee:

Major Professor Marco Alvarez

Noah Daniels

Haibo He

Nasser H. Zawia

DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND

2019

ABSTRACT

Over the last decade, advancements in deep learning and computer vision have led to a tremendous growth in performance at the tasks of automated human age estimation and nudity detection. Modern machine learning models can predict whether or not an image contains nudity or the presence of a minor with startling accuracy. When used in conjunction, these technological advancements can be used to identify new instances of child pornography without ever coming into contact with the illicit material during model training.

In this thesis, a label distribution learning framework for modeling human apparent age is proposed. Instead of directly modeling a person’s *biological age*, we use a probability distribution over a sample of humans guessing how old that person *looks like* as the ground truth. This allows us to better capture the subjective nature of a person’s age and advance state of the art performance at the task of apparent age estimation.

Next, we introduce a framework to automatically identify Sexually Exploitative Imagery of Children (SEIC) in both images and video. It is a synthesis of our original age estimation models and Yahoo!’s open sourced nudity detection model, OpenNSFW. Deep learning models are used to identify the presence of a minor or nudity in any given image or video. The performance of this approach is evaluated on several widely used age estimation and nudity detection datasets. Additionally, preliminary tests were conducted with the help of a local law enforcement agency on a private dataset of SEIC taken from real world cases.

ACKNOWLEDGMENTS

I first thank my adviser Dr. Marco Alvarez who pushed me to do work far beyond what I thought I was capable of.

I also thank the members of my committee, Noah Daniels and Haibo He, as well as other faculty members Gerard Baudet, Erin Denette, Lisa DiPippo, Jacob Fonseca, Lutz Hamel, Edmund Lamagna, Deborah Mathews, and Joan Peckham for their open doors, advice, and guidance throughout the years.

Finally, I thank my sister and my parents for their unconditional love and support during my journey.

This project was supported by Award No. 2016-MU-CX-K015 awarded by the National Institute of Justice, U.S. Department of Justice.

TABLE OF CONTENTS

ABSTRACT	ii
ACKNOWLEDGMENTS	iii
TABLE OF CONTENTS	iv
LIST OF FIGURES	vii
LIST OF TABLES	ix
CHAPTER	
1 Introduction	1
2 Related Work	7
2.1 Age Estimation	7
2.1.1 Biological Age Estimation	8
2.1.2 Apparent Age Estimation	9
2.2 Deep Learning	11
2.2.1 Convolutional Neural Networks	11
2.2.2 Label Distribution Learning	15
2.3 Pornography Detection	16
2.3.1 Automated Pornography Detection	17
2.3.2 Automated SEIC Detection	18
3 Apparent Age Estimation	21
3.1 Label Distribution Learning	21
3.2 Modeling Human Guesses	23

	Page
3.2.1 Histogram Distributions	24
3.2.2 Normal Distributions	24
3.2.3 Kernel Density Estimation	24
3.3 Evaluation Measure	24
3.4 Transfer Learning	25
3.5 Apparent Age Estimation Experiments	25
3.5.1 Datasets	26
3.5.2 Training Details	27
3.5.3 Protocols	29
3.5.4 Results and Analysis	30
4 Automated Detection of SEIC	36
4.1 Age Estimation and Label Distribution Learning	36
4.2 Integrating Deep Learning Models	36
4.3 Performing Inference	39
4.4 Datasets	40
4.5 Training Details	42
4.6 Robustness of NSFW and Age Models	45
4.6.1 Evaluation of Age Estimation and Nudity Detection Models on SFW and NSFW Images	45
4.6.2 Racial Bias	48
4.6.3 Challenging Images	49
4.6.4 Evaluation of Age Estimation Models on Videos	50
4.7 Classification of SEIC Videos and Images	52
4.7.1 SEIC Image Detection	52

	Page
4.7.2 NSFW Video Detection	53
4.7.3 SEIC vs. NPDI Videos	54
4.7.4 Interpretability (Sample Report/Qualitative Analysis) . .	56
4.7.5 Practical Limitations and Considerations	59
5 Conclusion	61
LIST OF REFERENCES	64
APPENDIX	
Additional Figures	70
A.1 Learned Label Distributions	70
BIBLIOGRAPHY	74

LIST OF FIGURES

Figure		Page
1	Examples of fitted kernel density estimation distributions on human guesses. For images with precise age labels, such as in the first image, the ground-truth curve is tighter. When the predictions are more scattered the curve may become skewed or bimodal such as in the last two. These distributions reflect the subjectivity of the age and the quality of the image as a training example.	23
2	Pipeline for model inference. In (a), face detection is performed. Next, in (b) each detected face is rotated to up-frontal position and cropped with a 40% margin before being passed through our DenseNet-161 model depicted in (c). Finally, in (d) inference is performed by taking the expected value of the predicted distribution.	28
3	Each box plot shows the variance among the 10 models trained for each protocol on the APPA-Real dataset. Models denoted with A were trained for 101-way classification while those with B, C, and D were trained for histogram, normal, and kernel density estimation label distributions respectively. The horizontal line is shown for comparison to the previous state-of-the-art model, Apparent DEX [1]	33
4	Examples of predicted label distributions using Kernel Density Estimation (KDE) and 101-way classification. Note the smoother output distributions from KDE. The classification model is generally very confident in a single age class resulting in sharp peaks in the output distribution while the label distribution models are much smoother. Both plots are from models trained with <i>P2</i>	34
5	Additional examples of predicted label distributions using Kernel Density Estimation (KDE) and 101-way classification. Note the smoother output distributions from KDE. The classification model is generally very confident in a single age class resulting in sharp peaks in the output distribution while the label distribution models are much smoother. Please refer to the Appendix for enlarged versions of these figures.	35

Figure		Page
6	Integration of deep learning models	37
7	Sample images from the newly introduced challenging images of adults and children dataset. The top row shows examples of adults and children in costume. In the middle, some with sunglasses are shown. Finally, in the bottom those with hats. Note that each image clearly contains either an adult or child as the focal point of the image although multiple people may be present in each image.	40
8	Confusion matrices for minor classification on the test fold of the APPA-Real dataset using cutoff ages 14 and 18. The model rarely mistakes an adult (-) for a child (+), but is prone to calling children adults.	45
9	Frame-by-frame analysis of age estimation model on controlled aging video. The video begins with a child of age 5 facing the camera with a monochromatic backdrop in an interview setting. Every few seconds, a new subject is introduced exactly 1 year older than the previous, up to age 75. For each frame of the video, age estimation was performed and plotted versus the ground truth age.	51
10	Sample frames drawn from the NPDI dataset. The top row displays frames from the non-porn hard subset of data. The bottom row are samples from some non-porn easy videos. Even though the content of the non-porn easy videos may be easy to distinguish from pornography, it may still be subject to challenges such as blur and poor quality.	54
11	Confusion matrices for SEIC (+) video classification, and the subset used for comparison (-) varies.	56
12	Plot of errors for RISP vs. all of NPDI classification. The proportion of faces where a minor ($\text{age} \leq 18$) is detected is plotted against the proportion of frames considered NSFW by OpenNSFW using a threshold of 0.402. Mistakes are marked by an 'X'.	57

LIST OF TABLES

Table		Page
1	Relative quality of pre-training on IMDB-Wiki with biological age labels.	30
2	Comparison of MAE for both apparent age estimation and biological age estimation on APPA-Real. Our flagship model achieves a MAE of 3.688, surpassing the best performing result in [1].	31
3	Mean Absolute Error on APPA-Real for three different protocols $P0$, $P1$, and $P2$. For each protocol 4 models are trained: 101-way classification, histogram label distribution, normal label distribution, and KDE label distribution. Each model is trained 10 times with the same set of random seeds and the mean and standard deviation of the MAE is reported for both apparent and biological age estimation. We achieve a MAE of 3.688 with ($P2$) and Kernel Density Estimation, surpassing previous results for APPA-Real reported in [1].	32
4	Binary classification of faces extracted from NSFW images into adults and minors using a cutoff age of 14. The accuracy of the model degrades significantly here in comparison to APPA-Real, but still generalizes well to this far more uncontrolled environment.	46
5	Nudity detection on RedLight using the Yahoo Open NSFW model with a sensitivity level of 0.402. The model has almost perfect accuracy on the easiest examples in the dataset (solo nude male, female, and 3+) but makes some mistakes in the partial porn category.	48
6	OpenNSFW accurately identifies Caucasian pornographic images but makes a few mistakes for other ethnicities.	48
7	Binary classification of ethnic faces extracted from NSFW images into adults and minors using a cutoff age of 14.	49
8	Binary classification of extremely challenging images of children (+) and adults (-) using a cutoff age of 14.	50

Table		Page
9	Speculative results for SEIC image detection by combining nudity and minor detection. A threshold of 0.402 is used to consider an image NSFW and an age cutoff of 14 is used. The number of false positives from the RedLight dataset is also provided for “loose” comparison.	52
10	Performance on the task of pornographic video detection using the meta-classifier compared to other deep learning methods on the same dataset.	53
11	Breakdown of model performance varying the subsets of NPDI. For each experiment the SEIC dataset was used as the positive class label while the subset composing the benign label varied as indicated by the ‘X’.	55

CHAPTER 1

Introduction

The proliferation of digital media has led to the most frenzied growth in child pornography than at any other point in history. The Child Victim Identification Program from the National Center for Missing and Exploited Children (NCMEC) has reviewed more than 236 million images and videos and law enforcement has identified more than 14,500 child victims [2]. As of 2017, NCMEC has sent more than 209,000 notifications to service providers regarding publicly accessible websites (URLs) on which suspected child sexual abuse images appeared.

During the course of an investigation of suspected child pornography, a computer forensics specialist typically spends many hours looking at hundreds of thousands of images and videos. The seizure and further analysis of a suspect's computer and data is tedious, error prone, invades the privacy of the suspect, wears on the investigator, and demands time that the investigator could be using to address the backlog of cases he/she likely faces. Automating the process of searching images and videos on seized media would drastically reduce the amount of time that investigators have to spend looking at the images, reduce time spent looking through irrelevant non-pornographic photographs, and would allow investigators to concentrate on other aspects of the case.

We propose using recent computer vision and machine learning advances to automate the process of identifying Sexually Exploitative Imagery of Children (SEIC) in order to significantly decrease the amount of time law enforcement agents spend on child pornography investigations.

Traditional machine learning methods for this task rely on manual feature engineering and tend to generate many false positives, serving only as a coarse

filter for suspected material. When a large number of files are present on the suspect’s hard drive, the agent reviewing the case may become overwhelmed with imagery falsely flagged as being SEIC, especially when many pornographic images are present. This is because pornographic content is difficult to distinguish from SEIC using traditional techniques since the notion of age is not explicitly modeled. To address this issue, we propose fusing the predictions of more accurate deep learning models for nudity detection with our recent work in apparent age estimation to classify SEIC material as a synthesis of nudity and minor detection.

The development of computational methods for age estimation from human face images has been one of the most challenging problems within the field of facial analysis [3]. In addition to common difficulties in facial analysis, such as pose and illumination, age estimation proves a significant challenge due to the subjective nature of the problem. The process of aging is unique to every individual and is influenced by their genetics, diet, occupation, and hobbies [4]. This implies that two individuals with the same biological age can have quite different appearances. On the other hand, large annotated databases are difficult to collect, especially for individuals in the lower and upper ends of the spectrum of human ages.

The exact way in which we approach age estimation has been an active area of research. It may be posed as a classification task, where labels are discrete (age groups ranging several years or just a single year), as a regression task, where labels are continuous (in years), or as a hybrid task using both classification and regression methods.

Age estimation may be considered from either of two different representations: *biological age estimation* or *apparent age estimation*. In biological age estimation, the actual age of a human subject is predicted while in apparent age estimation the label is the aggregation of a group of guesses made by human labelers. This

aggregation is usually the arithmetic mean of the collection of guesses.

Apparent age estimation is a more recent topic, which has received increasing attention [1, 5, 6], as a result of the deep learning revolution and two apparent age estimation competitions run by ChaLearn in 2015 and 2016 [7]. State-of-the-art results are already beating the human reference. By focusing the attention on the apparent age of individuals, the hopes are to alleviate the subjectivity underlying biological age estimation tasks, since human guesses are expected to agree more on how old a subject *looks like*.

The goal of Chapter 3 is to approach the apparent age estimation problem under the framework of *label distribution learning* (LDL) [8]. Unlike classic single-label or multi-label classification, in which instances are assigned to a single or multiple labels, the aim of LDL is to assign instances to label distributions, i.e., vectors containing the probabilities of the instance having each label. Our motivation is essentially to find better ways to model the label ambiguity underlying the apparent age estimation problem. Furthermore, APPA-Real [1] has been recently made available. This dataset provides a large number of face images labeled with real and apparent age annotations. APPA-Real contains 7.6k face images with an associated number of nearly 300k human guesses.

We propose an end-to-end framework, based on convolutional neural networks, to learn distributions of apparent age labels. Given an input image of a human face x , we want the model to produce a discretized probability distribution vector where each value at index k represents the probability of x being k years old. In order to evaluate our framework, we conduct experiments with the APPA-Real dataset because it provides a number of human age guesses for each image.

Overall, the contributions of Chapter 3 include:

- A novel end-to-end framework, based on learning label distributions, that

leverages the availability of human guesses in the APPA-Real dataset for modeling the apparent age estimation problem;

- Better performance than state-of-the-art methods on apparent age estimation using the APPA-Real dataset. We improve the mean absolute error to 3.688 years;
- Empirical evidence that pre-training using label distributions yields higher performing models regardless of the target task.

Next, in Chapter 4, we introduce a framework for the automatic detection of SEIC videos and images using convolutional neural networks (CNNs). Given some seized hard drive or another source of digital media, all video and image files are located. The probability of each containing child pornography is estimated using our age estimation model [9]. Additionally, we calculate nudity scores using Yahoo!’s publicly available OpenNSFW nudity detection network [10].

Since we are explicitly incorporating the notion of age into our model, we are better able to capture the subtlety between pornography and SEIC. We can also provide more interpretable analysis for law enforcement agents. The number of faces found, the age of each person, and overall nudity detection score for each image will be presented to the user. Images may then be flagged as SEIC with arbitrary precision by tweaking the estimated age required for being a minor and nudity detection score.

Videos are slightly more complex to classify. Videos are split up into a series of frames and treated as a set of images. Similar characteristics may be reported to the user at a per-frame resolution. A final machine learning model reports the probability of the video containing illicit content, and classification may still be performed to arbitrary precision by specifying some threshold required for video flagging. The agent reviewing the material will then be able to quantify exactly

how much nudity and how many children must be detected in a video before flagging it for review.

This novel approach results in an framework which may be as fine or coarse a filter as the agent specifies, in a way previous approaches cannot. It can even distinguish between challenging examples of pornographic and SEIC videos with 89% accuracy using the default thresholds. Since our approach relies on automatic representation learning through the use of convolutional neural networks, those involved in this work never had to be directly exposed to pornographic or SEIC content. Additionally, in contrast to most other works on SEIC content detection, we rigorously evaluate our models on a series of challenging datasets to analyze their performance before presenting results on data collected from real law enforcement cases.

Overall, the contributions of Chapter 4 include:

- A novel framework for the automatic detection of child pornography in videos and images
- A rigorous analysis of the performance of the nudity detection and age estimation models on ethnically diverse and challenging pornographic and non-pornographic images and videos
- Validation of our approach through empirical evidence that treating video classification as a per-frame image classification task with prediction aggregation achieves competitive results at pornography detection on the NPDI video dataset
- A rare evaluation of our framework for child pornography detection on a real world dataset collected by local law enforcement agents from 20 real world cases

This thesis concludes in Chapter 5 with a brief summary and a discussion of future work. Additional figures highlighting learned probability distributions are given in the Appendix.

CHAPTER 2

Related Work

In this chapter, the background literature for deep learning, age estimation, label distribution learning, and pornography detection, which forms the basis for our methodology, is summarized. Currently available techniques for the automatic detection of child pornography in images are also described.

2.1 Age Estimation

Lately the automatic estimation of age from facial images of humans has received great and increasing interest [11]. Age estimation can be posed in either of two ways. The question, “How old is the person in this photograph?” can be interpreted as either trying to determine the biological age of the subject, or as observing the “apparent” age. A computer vision algorithm can be created with the intention of satisfying either form of this question. Most of the early work utilizing feature engineering was focused on biological age.

Related work for the automatic estimation of biological age from facial images includes methods that employ hand-crafted features to represent age patterns, e.g. local binary patterns, histogram of oriented gradients (HOG), and biologically inspired features [12]. Given a set of hand-crafted features, the problem of facial age estimation can be modeled as a classification task for discrete age intervals [13], as a regression task for direct age estimation [11], or as a fusion of both tasks. A more complete review of existing age estimation methods is presented in Liu et al. [14].

2.1.1 Biological Age Estimation

The vast majority of existing computational methods focus on the prediction of biological age. This problem has been one of the most challenging problems in facial analysis. After relying mainly on human inspection of craniofacial features, later studies incorporated a plethora of computer vision and machine learning techniques, to first, extract features from images, and then pose the problem either as a classification or as a regression task. An extensive review on different approaches and datasets is presented in [15].

With recent developments in deep neural networks, where manual feature design is no longer required, current age estimation methods have shown impressive performances. In [3], a discussion of human accuracy on predicting the biological age on the FG-NET dataset is given. Through aggregation by outlier removal and the arithmetic averaging of ten votes from human labelers, a Mean Absolute Error (MAE)¹ of 4.7 is given as the human error on the dataset. It is interesting to contrast this with recent results from deep learning based models on this dataset from 2015 by DEX [6] which achieves a MAE of 3.09, surpassing human performance by this metric.

More recently, state-of-the-art results in biological age estimation have been obtained by casting age estimation as an ordinal classification problem. In [16], the authors proposed training a single Convolutional Neural Network (CNN) with many binary predictors. For l potential age classes, each output neuron $k \in \{1, 2, \dots, l - 1\}$ would predict the probability of example x being older than age k . The predictions are then aggregated together via Eq. 1 where $\hat{y}$ is the final predicted label, and $\hat{f}_k(x) \in \{0, 1\}$ is the output of the k -th classifying neuron given input image x . They jointly optimized k binary classifiers over the cross-

¹Absolute Error refers to the absolute value of the difference between the predicted and ground-truth label

entropy loss function.

$$\hat{y} = 1 + \sum_{k=1}^{l-1} \hat{f}_k(x) \quad (1)$$

The key idea here is recognizing that there is some ordinal relationship in the set of ages to predict. By jointly optimizing several binary classifiers, the authors force a network to extract a set of features to detect whether or not a person is older than age k . They then combine the predictions of each simple binary classifier to create a more accurate final prediction.

2.1.2 Apparent Age Estimation

Apparent age estimation can be considered a relatively new topic in facial analysis. Most age estimation datasets lack apparent age labels and are only suitable for biological age estimation. More recently, researchers have begun modeling the aging process by using the apparent age of a person to train their models. The motivation is that in reality, people have an apparent age that could differ perceptually from their chronological age. Since feed-forward neural networks learn by adjusting themselves based on self-error, punishing the model for a face that visually looks not in a certain age group may be counter productive. The quality of a machine learning model in this context depends on the availability of a large dataset with face images annotated with apparent age labels such as the one introduced in the ChaLearn Looking At People Apparent Age Estimation competitions run in recent years [17, 18].

An overview of the most popular methods for apparent age estimation can be found in the 2015 and 2016 ChaLearn LAP competitions [7]. Both competitions are relevant because they propelled research in apparent age estimation by providing the first dataset annotated with human guesses. Each image in these datasets is annotated with a mean age and a corresponding standard deviation of human

guesses.

One of the greatest difficulties in age estimation is how to pose the objective function. The winner of the ChaLearn LAP 2015 competition, whose model was named DeepEXpectation (DEX) [6], utilized a VGG-16 based CNN architecture and performed experiments with posing the problem as a regression or classification with groups of varying sizes. Treating age estimation as a 101-way classification and optimizing the network using the standard cross-entropy loss function (softmax) produced the best results. The key to their success was computing the softmax expected value. This formulation takes advantage of the assumption that when the model misclassifies an image, it is likely to predict an age closer to the ground-truth age. It is important to note that this implicit ordinal relationship is not exploited during the training phase. Additionally, they fine-tuned the network on a crawled dataset of 0.5 million celebrity pictures, collected from IMDB (Internet Movie Database) and Wikipedia. To the best of our knowledge, this is currently the largest annotated, and publicly available, dataset for biological age prediction (IMDB-WIKI).

The APPA-REAL dataset [19] contains a set of human face images with accompanying biological ages and human age guesses. The age guesses were collected via crowd sourcing with an average of 38 human guesses per image. With such a rich group of labels available for each image, more sophisticated techniques can be used for learning that exploit this idea of subjectivity and the ordered nature of ages. Such is the aim of label distribution learning (LDL), which seeks to assign input instances to entire label distributions, i.e., vectors where each element contains the probability of the instance having each label.

2.2 Deep Learning

Recent advances in machine learning algorithms, together with the availability of large online datasets and GPU (Graphics Processing Unit) technology have paved the way to tackling problems once considered impossible, particularly in the field of computer vision. Convolutional neural networks (CNNs) have achieved remarkable levels of accuracy for a variety of tasks, such as the automated detection of faces [20], nudity [21], and human age [16] as well as the traditional computer vision problem of image classification [22]. More spectacular examples of CNNs are those such as one trained for the task of automatically determining the location where a photo was taken just by processing its pixels, with data mined from geotagged images [23] or in artistic style transfer [24] from one image to another.

CNNs, a particular type of feedforward neural network, take advantage of the grid-like structure and spatial locality of images to accomplish these tasks. CNNs can not only perform traditional machine learning tasks such as classification or regression, but also learn complex feature hierarchies directly from raw pixels, eliminating the need for manual feature engineering. This ability to automatically learn rich features from the data as opposed to relying on hand-crafted feature design is key to their success. In practice, CNNs may prove difficult to train due to the massive number of parameters that must be learned. The large capacity of the network demands many annotated training samples, and even with the general abundance of labeled data found online, often times researchers have trouble locating large, specific datasets, although transfer learning may alleviate this issue [25].

2.2.1 Convolutional Neural Networks

Let $\mathcal{X}$ denote some feature space and $\mathcal{Y}$ denote a set of labels. Given some dataset $\mathcal{D} = (x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(n)}, y^{(n)})$ where $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and each

$(x^{(i)}, y^{(i)})$ pair form a single labeled training example, the goal of supervised learning is to find some function g which best approximates the mapping $\mathcal{X} \rightarrow \mathcal{Y}$. This mapping is typically learned by minimizing some loss function L with respect to the parameters of g using *stochastic gradient descent*. A loss function typically takes a form given as Eqn. 2 where N is the number of training examples and θ are the parameters of the model.

$$L = \frac{1}{N} \sum_i^N L_i(g(x_i, \theta), y_i) \quad (2)$$

One of the most popular loss functions used for classification tasks is the cross-entropy loss function, given as Eqn. 3 where the notation g_j denotes the j th element of the vector of class scores produced by our model g and g_{y_i} denotes the score of the vector element corresponding to the ground truth class.

$$L_i = -\log\left(\frac{e^{g_{y_i}}}{\sum_j e^{g_j}}\right) \quad (3)$$

When computed, the value of a loss function is a scalar representing how happy we are with the model's predictions for the training set. The loss function must be bounded below, typically at 0, which represents a perfect matching of training examples to their ground truth labels. If our model g were to make mistakes in mapping input instances x to their corresponding labels y , the output of our loss function would increase; preferably this increase would be proportional to the severity of the mistake, but this is not always the case with generic loss functions as we will see in Section 3.1.

A wide range of machine learning models exist. Some types of models are better at particular tasks than others. Over the last decade, the increasing computational power offered by Graphical Processing Units (GPUs) has given researchers the ability to train classes of models with millions of parameters in a reasonable

amount of time. This has led to the dominant performance of *convolutional neural networks*, a specific type of feedforward neural network, at various computer vision tasks such as image classification. In this thesis, we are concerned with such tasks - therefore, we may further restrict the input space $\mathcal{X} = \mathbb{R}^{W \times H \times C}$. Each instance of $x \in \mathcal{X}$ is an image represented by a tensor of pixel values and is of dimension W, H, C - the width, height, and number of color channels in the image respectively.

In convolutional neural networks (CNNs) [26], the prediction function g can be thought of as the composition of a linear sequence of mathematical operations organized into a computational graph. Each operation is more commonly referred to as a layer in the neural network. There are many types of layers; the convolutional layer being the hallmark of the CNN. In a convolutional layer a $N \times N \times K$ “filter” is slid across the $L \times M \times K$ input to the layer. At each spatial location a dot product is performed between the tensor of weights which make up the filter and the portion of the input matrix the filter is currently sliding over. The weights in this filter are a learnable parameter of the network. Convolutional layers have two key properties that make them especially good at tasks where the input is arranged with grid-like structure like an image: sparse interaction and parameter sharing [27].

Sparse interaction: In typical feedforward neural networks, the output of each layer is connected to each hidden node in the next layer. This results in a matrix multiplication with a runtime of $O(M \times N)$ compared to the $O(K \times N)$ complexity of the convolution operation since every input and output neuron pair are not connected by a unique weight. In most applications K is typically orders of magnitude smaller than M , so the savings in computational cost, memory, and model capacity gained by using convolutional layers are significant.

Parameter Sharing: Each convolutional filter is slid across all spatial locations of the input layer. For each of these positions, the same matrix of weights is used. This results in a set of weights that are tied together; different elements of the input do not get unique weights. They share the same value. This results in a network that needs to learn the set of weights corresponding to a filter only once for it to be applied to the entire input. If we wanted the same function that is represented by these weights in a traditional neural network the same weights would have to be learned many times over in the much larger parameter space, and we would end up with many redundant representations of the same function.

More intuitively, the role of the convolutional layer is to learn some intermediary feature representation of the input. For a given image, a convolutional filter will learn to detect visual patterns that are useful for the given task. The first few layers of convolutional filters typically learn simple edge detectors and gabor filters while those deeper in the network learn more complex, abstract representations of the input interpreted as linear combinations of the preceding layer. For a CNN trained to classify images as containing cats or not, some convolutional filters may learn that combinations of lines making up circles are useful features. A filter deeper in the network may learn that particular orientations of circles are useful for determining if eyes are present in an image. The network in its entirety will combine the “detections” of all such filters to make some final prediction, interpreted by the human as whether or not the image is classified as “cat”.

Other types of layers present in CNNs include fully connected layers, non-linearities, pooling, batch normalization, and dropout.

A fully connected layer is simply the matrix multiplication of every input element with its corresponding hidden weight. If there are N input neurons connected to a hidden layer of size M , there will be $N \times M$ weights. Non-linearity layers

are the same as in typical fully connected networks, with $\text{ReLU} = \max(0, x)$ being the activation function of choice [28]. Non-linearities are necessary to prevent a network from devolving into a single matrix multiplication because without them a series of fully connected layers could be viewed as a single linear transformation.

Pooling layers replace the output of a convolutional layer with a summary statistic of the activations close together spatially. Pooling helps to make intermediary representations invariant to translation, and reduce the dimensionality of the input. Batch normalization [29] forces the input to a layer to be unit gaussian by learning scaling and shifting parameters α and β . This operation helps alleviate difficulties with properly initializing neural networks, and was instrumental in training the first deep CNNs.

Dropout [30] layers are used to randomly “drop” the input value from one layer to another with some probability by setting it equal to zero. This forces networks to learn redundant representations of the input and prevents two layers from becoming co-dependent on one another. It can also be viewed as both a form of regularization and a form of ensembling with each possible bit-mask representing a unique CNN when performing inference.

We refer the reader to the text Deep Learning ² by Ian Goodfellow for a more rigorous explanation of deep learning and convolutional neural networks or the Stanford course CS231n ³, both of which may be found online for free.

2.2.2 Label Distribution Learning

One of the early works in learning from label distributions, in the context of facial age prediction, is presented in [31]. Their work is inspired by the observation that faces at close ages look quite similar. Thus, learning from an input image of age a will also improve the model’s performance on ages close to a . The authors

²<https://www.deeplearningbook.org/>

³<http://cs231n.stanford.edu/syllabus.html>

define a label distribution as a vector of real numbers, in which values $P(y) \in [0, 1]$ represent the *degree* that the corresponding label y describes an instance. All values in this vector sum to 1, i.e., a distribution over the set of labels. A number of methods have been proposed to address this task. In [31, 8, 32], proposed LDL methods are based on the maximum entropy model [33]. As distributions from the exponential family arise as natural solutions to the maximum entropy problem, the generality of the solution is restricted. A different group of LDL approaches aim to extend existing machine learning algorithms to deal with learning distributions. Extensions have been proposed for support vector machines [34], boosting [35], and decision trees [36].

More recently, in [37], CNNs are proposed as an end-to-end learning framework that minimizes the Kullback-Leibler divergence between the predicted and ground-truth discrete label distributions. As ground-truth label distributions are not available in most existing datasets, the authors generated discrete label distributions under proper assumptions. For example, in the age estimation context, the authors labeled each image with a label distribution generated from a normal probability density function. This density function is a natural choice given that, for each image, a mean μ and standard deviation σ are available in the training dataset. When the standard deviation is not available, the value of $\sigma = 2$ is arbitrarily chosen. Our work does not assume that human guesses for apparent age estimation are normally distributed, but rather, shows that non-parametric distributions outperform the assumption of a normal distribution.

2.3 Pornography Detection

In this section, we outline recent work in both the automated detection of pornography and SEIC.

2.3.1 Automated Pornography Detection

Currently many popular websites such as YouTube and Facebook augment their automated content flagging systems with human laborers to review and moderate User Generated Content [38]. The human laborers they employ often quit within months of being hired and are usually not trained or prepared to deal with the trauma of seeing so many deplorable videos [39]. There are clear incentives for tech companies to produce an accurate automated content moderator but, so far, major web companies have not come up with an effective algorithm to automatically remove unwanted content without human intervention - at least not one that has been disclosed publicly.

Most of the current research on automatic pornography detection considers the application of machine learning techniques to still images [40]. Typically, a classifier is trained with hand-crafted features extracted from a database of labeled images. Most common features include some encoding of human skin color; if the image contains a large portion of skin-tone pixels, then it will likely be flagged as positive by a manually crafted classifier. As the sole presence of skin is not enough for reliably detecting pornography, researchers have been experimenting with a plethora of different features that include shape and edge features, local image descriptors, histograms of gradients, visual words, higher level facial features, among others. Recently, an enriched Bag-of-Words (BoW) framework [41], based on HueSIFT descriptors, outperformed other standard BoW models. Additionally, as described in [42], most existing works can roughly be characterized as standard BoW models exploring different types of static features.

In 2010, the Digital Forensics and Cyber Security Center (DFCSC) at the University of Rhode Island conducted research on the automatic identification of pornographic images to assist law enforcement agents in SEIC cases. The software

has been released as RedLight. The application used a combination of low-level and high-level features, such as skin tone percentages, haar classifiers, and facial proportions, to train machine learning models for pornography classification [43]. This approach worked well for detecting adult pornography, where many instances of video are studio grade but failed to generalize to lower resolution, and improperly filmed videos.

Not all digital media containing large amounts of body exposure is considered pornography. Pictures and videos of beach scenes, sports games, and people in revealing clothing fall into a difficult to categorize group of images where a large amount of skin is exposed in a non-sexual context. Traditional computer vision algorithms for pornography detection that use skin detection features to detect nudity often fail to capture this nuance [44, 45, 43]. Ulges et al. improve upon this by using by using color visual words instead of skin detection to identify this content [46]. More recently, convolutional neural networks have been shown to achieve great success at pornographic image and video classification [47, 48, 49, 21], particularly on the NPDI video pornography dataset [50]. This dataset consists of three categories (pornography, non-porn easy and non-porn hard) and serves as a challenging, modern benchmark for pornography detection.

2.3.2 Automated SEIC Detection

Microsoft’s PhotoDNA technology [51], a resizing resistant image hashing algorithm, is available as a free service and helps stop child exploitation images from being shared online. The algorithm works by comparing the hash value of a suspected image to the hash values contained in a database of identified SEIC images. When an image is flagged as containing SEIC, the service provides the capability to report the illegal content to the NCMEC and appropriate law enforcement agencies. Unfortunately, it is limited to detecting images already cataloged

in their database and cannot detect SEIC video at all; once a new instance of illicit material is discovered it takes time to be verified and make it into the database.

Significant effort has been made into automating the detection of new instances of SEIC. In [52], traditional visual and audio features were used to classify SEIC images and videos while [53] used CNNs. Both train their models and present results on real SEIC data through collaboration with local law enforcement agents. The work of Sae-Bae, which does not make use of CNNs, takes a different approach similar to the one proposed in this paper and poses SEIC detection as a hybrid task of nudity detection followed by age estimation [44], but relied on manual feature engineering.

A rigorous study on the difficulty of SEIC image classification was given in [54]. Five law enforcement agents were tasked with identifying illicit material with the intention of analyzing the challenges faced when categorizing these images. In order to be categorized as SEIC, the image had to be identified as containing a minor and indecency.

The agents reported it difficult to identify the age of the victims in the photographs when there was a discrepancy between bodily and facial features, the victim was “staged” to alter their appearance (makeup or jewelery was applied that a child does not typically wear), there was an absence of secondary sex characteristics, and finally because of the natural variation in sexual development and variability across ethnic groups. Children in the developmental stage of early and late childhood were easy for law enforcement to identify (≤ 10 years old).

Indecency was hard to identify when the offender was absent from the image, the child had positive facial expressions, the context of the image was ambiguous, and the image was taken in a public area. Indecency was easy to identify when there was evident sexual activity between an adult and a child, the victim was in

obvious distress, the image was taken in a sexual context, or when the background of the image was suspicious.

This analysis motivates our approach for a hybrid approach to SEIC content detection. To be successful at detecting SEIC content our models must therefore learn how to detect both age and indecency. Models trained for pornography detection that try to generalize to SEIC classification have no concept of age and will likely fail to distinguish normal pornography from SEIC [52]. In contrast to other approaches, we designed a series of experiments inspired by the recent study of agents tasked with identifying SEIC content to analyze exactly how robust our framework is in situations humans have difficulty in.

CHAPTER 3

Apparent Age Estimation

In this chapter we define the Label Distribution Learning (LDL) problem [8], our proposed solution, and an overview of our network architecture and training.

3.1 Label Distribution Learning

Let $\mathcal{X} = \mathbb{R}^{w \times h \times c}$ denote the input space of images, where w , h and c are the width, height, and number of channels of the input instance. Let $\mathcal{Y} = \{y_1, y_2, \dots, y_l\}$ denote the ordered set of labels. A LDL problem is defined as learning the mapping function $f : x \mapsto d$ between an input instance $x \in \mathcal{X}$ and its corresponding label distribution $d = [d_1, d_2, \dots, d_l]^T \in \mathbb{R}^l$. Each element d_i in the label distribution expresses the probability of an instance x having the label y_i . A label distribution d is also constrained to $d_i \in [0, 1]$ and $\sum_{i=1}^l d_i = 1$. Finally, a training dataset is denoted by $\mathcal{D} = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(n)}, y^{(n)})\}$

Moreover, considering the age estimation problem, given an input image x with a discrete label distribution d , we interpret each value d_i as the probability of x being i years old. In the context of this work, the total number of labels is $l = 101$ with ages ranging from 0 to 100.

When training LDL models with neural networks, specifically CNNs, the Kullback-Leibler (KL) divergence is used as the loss function. The KL divergence can be seen as a similarity measure between the ground-truth discrete label distribution and the predicted label distribution. We seek to optimize the loss function given in Eq. 4 where x is an input image, $\hat{f}(x)$ is the log-normalized probability vector the model produces, and d is the ground-truth label distribution. When performing inference, we follow [6] in taking the expected value of the output distribution to get our final prediction $\hat{y} = \mathbb{E}[\exp(\hat{f}(x))]$ over the ages

$[0, 1, \dots, 100]$.

$$L(\hat{f}(x), d) = \sum_{i=0}^{100} d_i (\ln d_i - \hat{f}_i(x)) \quad (4)$$

Consider the context of training CNNs for single label classification using the cross-entropy loss, where each class is an age in years. It is worth to note that iterative gradient updates will induce changes in the model’s weights, ignorant of the ordinal relationship between classes (labels). For example, if this hypothetical classifier produces the 101-dimensional normalized probability vector $[0.15, 0.60, \dots, 0.00]$ for some image x whose true class is 0 years old, the gradient update of the cross-entropy loss with respect to the output layer will be evaluated to $[-0.85, 0.60, \dots, 0.00]$, drastically penalizing the output of the first unit, even when the probability for the neighboring class (1 years old) is acceptable. When using the LDL model the goal is to learn the probability of each label involved in the description of an input instance, not only the label of the correct class.

Furthermore, as we can see in sample images from the APPA-Real dataset shown in Figure 1, the age of some people can be extremely difficult to predict. Consider the middle image in the same figure - the apparent age guesses vary wildly from 14 to 29 years old. If we limit ourselves to training an apparent age classifier on this image for the subject’s mean apparent age of 20, we will cause the model to make more drastic changes in its weights for a “misclassification” into the adjacent age of 19, even though by visual inspection this error seems perfectly reasonable. When we train on label distributions, we alter the ground-truth label to include some adjacent classes.

Alternatively, if a model is trained for regression and the Mean Squared Error is used as the loss function, an insight can be developed for a different type of error. Suppose for some image x the model predicts $\hat{y} = 4$ when in reality $y = 0$

Figure 1: Examples of fitted kernel density estimation distributions on human guesses. For images with precise age labels, such as in the first image, the ground-truth curve is tighter. When the predictions are more scattered the curve may become skewed or bimodal such as in the last two. These distributions reflect the subjectivity of the age and the quality of the image as a training example.

and the image contains an infant. For another image, we may get $\hat{y} = 35$ when $y = 31$. By using the MSE we would get a loss of 16 for both examples even though intuitively predicting a newborn baby as being 4 years old is inherently a worse error than predicting a 31 year old adult as being 35. By transforming the ground-truth into label distributions, we can expect the apparent age curve for an infant to be much narrower than the curve for a thirty-five year old. Empirically, in [6], the authors show that training a model for age regression performs much worse than classification and suggest the large gradients produced by errors prevents the network from converging.

3.2 Modeling Human Guesses

One of the advantages of APPA-Real is that each face image in the dataset is labeled with a number of human guesses (approximately 38 per image). This implies that the ground-truth label distribution d , although not directly available, can be easily generated under certain assumptions. For each image in the training set, a ground-truth label distribution is created from all its corresponding apparent age guesses, using the following approaches:

3.2.1 Histogram Distributions

The label distribution is associated with the frequency of counts for all guesses normalized by the total number of guesses.

3.2.2 Normal Distributions

The mean apparent age and standard deviation are calculated from the apparent age guesses. Then, a normal probability density function parametrized by the mean and standard deviation is used to generate the label distribution, at intervals $\{0, 1, \dots, 100\}$. This type of distribution is most similar to other work done in the field [31, 37], and assumes the “subjectivity” in the apparent age of a person is normally distributed.

3.2.3 Kernel Density Estimation

A probability density function of the apparent age guesses is estimated using kernel density estimation in conjunction with a gaussian kernel for smoothing. In this approach, bandwidth selection is critical and contributes to the interpretability of the final model. By selecting a small bandwidth, the resulting density is “tighter” and using a larger bandwidth allows the density function to be stretched out over the age guesses. Figure 1 shows examples of images with densities estimated with different bandwidths. A reasonable bandwidth h for a random variable X of length n can be calculated using the formula described in [55], given as Eq. 5, where Var denotes the variance and IQR denotes the interquartile range.

$$h = \frac{0.9m}{n^{\frac{1}{5}}} \text{ with } m = \min \left(\sqrt{\text{Var}(X)}, \frac{\text{IQR}(X)}{1.349} \right) \quad (5)$$

3.3 Evaluation Measure

To evaluate the performance of all models, the Mean Absolute Error (MAE) is used. It is the most commonly used performance metric in the age estimation

literature, and is defined for N images in Eq. 6 as the average difference between the *predicted age* $\hat{y}_i$ and the ground truth label y_i . In this work, results are reported on both the apparent and biological MAE. The former uses the arithmetic mean of the apparent age guesses on an image as the ground truth label, while the latter uses the biological age. As the predicted output of our model is a label distribution, we define the predicted age as the expected value of the predicted label distribution.

$$\frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i| \quad (6)$$

3.4 Transfer Learning

In [56], the authors explored transferability of existing CNN models for age and gender classification. Experimental results show significant gains when comparing transferred CNNs with a baseline model. Furthermore, in [57], state-of-the-art performance is achieved for facial expression recognition by using transfer learning from the popular ImageNet dataset.

As large high quality datasets for apparent age estimation are scarce, in this work we conduct a thorough exploration of the impact of transfer learning on learning label distributions. Our motivation is to transfer low-level feature representations from large datasets, such as ImageNet and IMDB-Wiki and explore how using a consistent loss function across pre-training and fine-tuning may benefit the performance of the network on the final task.

3.5 Apparent Age Estimation Experiments

In this section we first provide a detailed description of our experiments, including the datasets, network architecture, transfer learning procedures, hyperparameter selection, and learning rate schedules. Then, a summary of the results

and findings is presented followed by discussion and analysis.

3.5.1 Datasets

To the best of our knowledge, IMDB-Wiki [6] is the largest publicly available labeled aging database, with 523,051 images of 20,284 individuals. It consists of face images of actors and actresses collected from the popular Internet Movie Database¹ and Wikipedia². Each image is labeled by its corresponding biological age. In all experiments, we use only a subset of IMDB-Wiki. At the time of our experiments, the creators of the dataset reported issues with images from the Wikipedia portion of the dataset. All Wikipedia images were excluded from our local copy of the dataset. Images containing multiple faces were also excluded since the identity of the human associated with the age is not easily inferred. Likewise, all images where the face of the subject could not be located were also removed. This pre-processing was done with the meta-information provided by the maintainers of the dataset. We used 90% of the remaining images for training and 10% for validation, that is, 165,970 and 18,442 images respectively.

In [1], the APPA-Real dataset was introduced with the objective of gathering a large, robust set of human apparent age predictions on images of people “in the wild”³. The authors relied on crowd-sourcing to label each of the 7,591 images. The dataset is divided into three folds, containing respectively 4,113 images for training, 1,500 for validation, and 1,978 for testing. The pictures collected belong to about 7,000 different individuals taken in varying conditions of lighting and image quality, which makes the dataset more representative of the real world. Experimental results are presented supporting the benefits of training age estimation models on a “wide” dataset [56]. In APPA-Real, each image is labeled with

¹<http://www.imdb.com>

²<https://www.wikipedia.org/>

³“In the wild” refers to pictures taken in uncontrolled conditions.

both biological and apparent age labels. On average, each image contains apparent age guesses given by 38 different people. A biological MAE, which refers to the average difference between the biological age and the predicted age, of 4.573 was aggregated from the human labelers. This means that when we put together the predictions of all the crowd sourced respondents, the “wisdom of the crowd” is wrong by 4.573 years on average. APPA-Real is a unique dataset that provides both real and apparent age labels required for LDL.

3.5.2 Training Details

For all pre-training and fine-tuning tasks, we implement the same pre-processing procedure described in [6], which includes a face detector that performs face rotation to up-frontal position, and performs a crop of the face with 40% margin to maintain background context in the image. The IMDB-Wiki and APPA-Real datasets provide pre-processed images of this form online.

As for the network architecture, we used the implementation of dense convolutional networks (DenseNets) [58] provided by PyTorch. We chose a DenseNet based architecture because of its reduced number of parameters, computational efficiency, and training times. We specifically chose the 161-layer architecture because of its superior performance at the ImageNet challenge. Figure 2 shows the steps taken for performing inference on an input image, which involve pre-processing the image followed by a feedforward pass on the trained DenseNet.

All of our models are trained with stochastic gradient descent, a nesterov momentum of 0.9, a batch size of 64, and weight decay of $1e-5$. These parameters remained consistent throughout all pre-training and fine-tuning procedures. The learning rate schedule varied depending on the loss function being used.

When pre-training on IMDB-Wiki, the initial learning rate is set to 0.01 and 1.0 for the cross-entropy and KL-divergence losses respectively. In both cases, the

Figure 2: Pipeline for model inference. In (a), face detection is performed. Next, in (b) each detected face is rotated to up-frontal position and cropped with a 40% margin before being passed through our DenseNet-161 model depicted in (c). Finally, in (d) inference is performed by taking the expected value of the predicted distribution.

learning rate is reduced by a factor of $\gamma = 0.25$ every 5 epochs and training is terminated when the model begins to overfit the training set.

When fine-tuning on APPA-Real, the initial learning rate is 0.005 and 0.5 for the cross-entropy and KL-divergence losses respectively. We found the unconventionally large learning rates necessary for minimizing the KL-divergence due to the relatively small initial loss values observed. Fine-tuning all models on APPA-Real is performed for 40 epochs, and in all cases the validation set is used for model selection.

All images are normalized using the mean pixel values and standard deviations given by ImageNet, and then scaled down to 256×256 pixels. Data augmentation is performed by randomly cropping images to 224×224 resolution during training as well as randomly flipping each image horizontally. During test time, we take the center crop of the image. These data augmentation procedures help combat overfitting. Interestingly, using dropout did not appear to improve the performance.

All experiments are repeated 10 times per task using the same seeds appropriately. During pre-training we use uniform seeds for weight initialization, estimating the order of training examples, and data augmentation.

The hardware configuration for the experiments includes 4 Pascal Titan X GPUs. Pre-training on IMDB-Wiki from ImageNet weights took about 1 day while fine-tuning on APPA-Real took 1 hour per each model. Inference is performed at 126 images/second using all four GPUs.

3.5.3 Protocols

In order to organize our experiments we define the following protocols:

P0 : Model is pre-trained on ImageNet and fine-tuned using the APPA-Real dataset.

P1 : Model is pre-trained on ImageNet, fine-tuned on IMDB-Wiki by performing 101-way classification with the cross-entropy loss, and finally fine-tuned on APPA-Real.

P2 : Model is pre-trained on ImageNet, fine-tuned on IMDB-Wiki by performing LDL with Normal Label Distributions, and finally fine-tuned on APPA-Real. The IMDB-WIKI step assumed normal distributions parametrized by the biological age of each image and a fixed standard deviation of 3.

For each of the protocols described above, we fine-tuned the models with APPA-Real using 4 different target tasks:

- 101-way classification with cross-entropy loss;
- Histogram label distribution with KL-divergence loss;
- Normal label distribution with KL-divergence loss; and
- Kernel density estimation label distribution with KL-divergence loss.

3.5.4 Results and Analysis

Performance on IMDB-Wiki is not commonly reported in the literature, as it is a dataset used for pre-training only. The biological MAE on the validation set is reported in Table 1 with the sole purpose of comparing the relative quality of the pre-training for each loss function. The resulting mean absolute errors are fairly similar, with 101-way classification outperforming the model trained on normal distributions with an assumed standard deviation of 3.

This label distribution model was not trained on apparent age using a sample of guesses, but on a normal distribution parameterized by the biological age and a fixed standard deviation of 3, since IMDB-Wiki does not have the required human guesses to form a proper label distribution. The validation set from our pre-training on IMDB-Wiki was used for model selection and the best performing model was used as a basis for further fine-tuning onto APPA-Real in the subsequent experiments.

Even though the performance of the label distribution model is worse in terms of MAE, it is suspected that the model has learned a better feature representation from which more accurate models may be fine-tuned because of the explicit ordinal relationship introduced when training for label distributions.

Table 1: Relative quality of pre-training on IMDB-Wiki with biological age labels.

Task	Bio MAE
P1: 101-way Classification	5.268
P2: Normal L. D.	5.346

It is important to note that as of now only one other work has been published using the APPA-Real dataset, which we compare our results to in Table 2. Our best performing model, a label distribution learning model trained on kernel density estimation distributions, surpasses the previous state-of-the-art apparent age

estimation model reported in [1]. Interestingly, our model trained for apparent age estimation matches the performance of the Real DEX model trained for biological age estimation at predicting the biological age.

Table 2: Comparison of MAE for both apparent age estimation and biological age estimation on APPA-Real. Our flagship model achieves a MAE of 3.688, surpassing the best performing result in [1].

Model	App MAE	Bio MAE
P2: KDE L. D.	3.688	5.434
Apparent DEX [1]	4.082	5.729
Real DEX [1]	4.513	5.468
Real + Residual DEX [1]	4.450	5.352

In Figure 3 and Table 3, we report our final results on the APPA-Real dataset. Figure 3 shows the tight variance in our model’s performance at apparent age estimations while Table 3 also includes the standard deviation and biological MAE. The model trained using protocol (P2) and Kernel Density Estimation label distributions achieves a new state-of-the-art MAE of 3.688 ± 0.017 . This is an improvement of at least 9.5% over [1]’s MAE of 4.082, which performed 101-way classification with a posterior expected value refinement.

Models fine-tuned from *P2* consistently outperform their counterparts in *P1*. The boost in performance for the 101-way classification model is especially surprising, since a different loss function is used. This may happen because when a model is trained for label distribution learning, the ordinal relationship among the classes is exploited during the optimization process. When the loss function is changed back to the cross-entropy loss, the model may have an easier time preserving this relationship as it does not have to learn it implicitly in order to be a good estimator of the MAE. The implicit ordinal relationship can be seen in Figure 4, as the peaks of the predicted distribution happen to be close together even though nothing in

Table 3: Mean Absolute Error on APPA-Real for three different protocols $P0$, $P1$, and $P2$. For each protocol 4 models are trained: 101-way classification, histogram label distribution, normal label distribution, and KDE label distribution. Each model is trained 10 times with the same set of random seeds and the mean and standard deviation of the MAE is reported for both apparent and biological age estimation. We achieve a MAE of 3.688 with (P2) and Kernel Density Estimation, surpassing previous results for APPA-Real reported in [1].

Protocol	Task	App MAE	Bio MAE
P0:	101-way	5.530 ± 0.100	7.362 ± 0.089
ImageNet	Histogram L. D.	5.203 ± 0.116	7.085 ± 0.105
	Normal L. D.	4.594 ± 0.052	6.507 ± 0.057
	KDE L. D.	4.558 ± 0.061	6.516 ± 0.076
P1:	101-way	4.362 ± 0.069	6.222 ± 0.053
ImageNet + 101-way on IMDB-Wiki	Histogram L. D.	4.019 ± 0.074	5.712 ± 0.067
	Normal L. D.	3.867 ± 0.063	5.722 ± 0.054
	KDE L. D.	3.860 ± 0.037	5.672 ± 0.048
P2:	101-way	3.973 ± 0.091	5.677 ± 0.087
ImageNet + Normal L. D. on IMDB-Wiki	Histogram L. D.	3.896 ± 0.028	5.508 ± 0.037
	Normal L. D.	3.706 ± 0.032	5.475 ± 0.033
	KDE L. D.	3.688 ± 0.017	5.434 ± 0.025

Figure 3: Each box plot shows the variance among the 10 models trained for each protocol on the APPA-Real dataset. Models denoted with A were trained for 101-way classification while those with B, C, and D were trained for histogram, normal, and kernel density estimation label distributions respectively. The horizontal line is shown for comparison to the previous state-of-the-art model, Apparent DEX [1]

the optimization process explicitly encourages the model to behave this way. Better feature representations are learned when the ordinal relationship of the classes is exploited during pre-training, even if the exact relationship is not known and

Figure 4: Examples of predicted label distributions using Kernel Density Estimation (KDE) and 101-way classification. Note the smoother output distributions from KDE. The classification model is generally very confident in a single age class resulting in sharp peaks in the output distribution while the label distribution models are much smoother. Both plots are from models trained with $P2$.

we must assume some normal distribution with fixed standard deviation.

The significantly improved performance of label distribution models in $P0$ and $P1$ over 101-way classification suggests that it is not just pre-training with the same loss function that is responsible for the improved performance of the models; those fine-tuned using label distribution learning perform better in all cases.

In order to get additional insights, Figures 4 and 5 highlight some accurately predicted label distributions as well as a few mistakes. All of these images come from the test set. We plot the predicted label distribution for KDE as well as the classification model. Both models have similar apparent age estimates for these examples, but the 101-way classification model has most of its probability mass over a single age class.

Figure 5: Additional examples of predicted label distributions using Kernel Density Estimation (KDE) and 101-way classification. Note the smoother output distributions from KDE. The classification model is generally very confident in a single age class resulting in sharp peaks in the output distribution while the label distribution models are much smoother. Please refer to the Appendix for enlarged versions of these figures.

CHAPTER 4

Automated Detection of SEIC

In this chapter a framework for SEIC detection is introduced, as are the techniques used for image and video processing and inference. The various deep learning models used are described in detail.

4.1 Age Estimation and Label Distribution Learning

By taking advantage of the per-image collection of human guesses available in the APPA-REAL dataset, a unique normal distribution may be fit to model the age of each image as described in [9]. The model will then be trained to predict each image’s unique discretized distribution using the Kullback-Leibler (KL) divergence. Optimizing our model against a distribution of age labels explicitly takes advantage of the ordered nature of aging in a way that the cross-entropy loss cannot. Since a unique distribution can be specifically fit to each image, the subjectivity or difficulty of the image is captured as the width of the image’s label distribution. The intuition we are hoping to achieve from modeling these distributions is that the tighter the distribution gets the easier an age is to guess.

For more details on label distribution learning, see Section 3.1

4.2 Integrating Deep Learning Models

The framework has been developed with the goal of a unified data pipeline in mind. A basic diagram of the approach is provided as Fig. 6. Highlighted in gray are the applications in which deep learning is employed to achieve the specified task. Tasks that are vertically aligned can be performed in parallel. In the current implementation, each task is run sequentially (to optimize GPU memory availability) and log files containing relevant information move between

Figure 6: Integration of deep learning models

each module in Fig. 6. In future iterations, the efficiency could be improved by implementing the entire framework as more of a literal pipeline, where we load all of the programs in memory and pass smaller batches of images through each step. By moving batches of images through the pipeline, the application would be able to process media in real-time and work with live video.

Each part of the application is independent of the others' operation. Processing starts with the file crawler, which given a base directory, identifies every image and video file on the system. Next, videos are sent to the frame extractor to separate each video into a series of discrete frames to be processed as images. Currently each video is sampled at an arbitrary rate of 1 frame per second and saved to the file system as a set of images. While greater accuracy could likely be achieved with a higher frame rate or by taking advantage of spatio-temporal features present in videos, such analysis is excluded due to concerns for the increased computational cost of such techniques and to preserve the simplicity of our models. A more intelligent frame extraction technique which makes multiple passes on a video to collect more frames from difficult to analyze portions of a video could easily be fit into the architecture of the framework and would likely yield improvements in performance as well.

Additionally, our framework relies on nudity, face detection and age estimation to identify illicit SEIC material. Large scale datasets for facial age estimation in

videos do not yet exist and is an open area of research.

Yahoo’s OpenNSFW CNN serves as the base of the pornography detection module [10]. The model achieves satisfactory performance at identifying pornography in images containing SEIC without any fine-tuning, which we evaluate using the publicly available weights. It accepts images as input and outputs the probability of the image containing NSFW content. Preprocessing for this CNN includes converting to an RGB color format, and resizing the image to be 256x256 using a bilinear interpolation method.

For face detection, the publicly available S^3FD : Single Shot Scale-invariant Face Detector [20] is used. The performance of this face detector is impressive, yielding few false positives and tight, high quality bounding boxes around human faces of all sizes and resolutions. After detecting all the human faces in an image, each face is cropped and saved to file with an extended margin of 40% of the height and width of the face. This follows the work of [6] and gives the age estimator “context” in the region around the face. Only faces with a confidence score of 0.80 or higher and a width and height greater than 25 are sent to the age estimator in the hopes of filtering out smaller and lower quality faces. Input images for the S^3FD : Single Shot Scale-invariant Face Detector are converted to an RGB color format, resized to 640x640 using a bilinear interpolation method, and each color channel has the mean color value of image net subtracted from itself.

The age estimation module is based on our previous work in [9]. As input it accepts human facial images, and outputs a probability distribution across ages. In order to determine if an image contains a minor, we simply take the expected value of the distribution and check to see if the predicted value is less than the threshold for adulthood. To prepare input images, they are converted to an RGB color format, resized to 256x256 with a bilinear interpolation, center cropped to a

size of 224x224, and each channel is standardized to image net’s mean and standard deviation.

4.3 Performing Inference

When performing inference on images, the predictions of the age estimator with that of the NSFW detector are combined. If a minor is detected and there is NSFW content in an image, then it is flagged as suspected child pornography. The ages of each person detected in the image and the NSFW score is logged to file. For videos, inference is performed on each extracted frame.

Results for each frame are then aggregated for final video classification. A video is not flagged as illicit material if only a single frame contains contraband because the application may produce many false positives. As the number of frames sampled in a video grows, so would the likelihood of a false positive using this criteria. We therefore propose training a support vector machine (SVM) for classifying a video as containing child pornography.

This meta-classifier is trained on a normalized histogram feature set consisting of the output predictions of both the nudity detection and age estimation models. In addition to these discretized age and NSFW group bins, the number of faces, children, frames with nudity, total frames, proportion of frames with nudity, and proportion of faces containing a child are used as features. Two categorical features were also added to the model based on whether or not 30 NSFW frames or children were found in the video. The value 30 was chosen because, due to our frame sampling rate, it would imply that at least 30 seconds of nudity or children appeared in the video.

Another benefit to posing the final video classification this way is that the SVM will be able to provide a confidence value in the classification of the video. This value may be tuned to the users preference of whether it is better to identify

more examples of illicit material at the expensive of increased false positives or serve as a more restrictive filter, only presenting the most probable instances to the law enforcement agent using the tool.

4.4 Datasets

Figure 7: Sample images from the newly introduced challenging images of adults and children dataset. The top row shows examples of adults and children in costume. In the middle, some with sunglasses are shown. Finally, in the bottom those with hats. Note that each image clearly contains either an adult or child as the focal point of the image although multiple people may be present in each image.

In [1], the APPA-Real dataset was introduced with the objective of gathering a large, robust set of human apparent age predictions on images of people “in the wild”¹. The authors relied on crowd-sourcing to label each of the 7,591 images. The pictures collected are of about 7,000 different individuals taken in varying conditions of lighting and image quality, which makes the dataset more representative of real world photo capture. Each image is labeled with both biological and an average of 38 apparent age guesses.

The dataset of challenging images of adults and children was collected from

¹referring to pictures taken in uncontrolled conditions.

Flickr² in the fall of 2017. It consists of approximately 100 images in each of the following categories: adults in hats, sunglasses, and costumes and children in hats, sunglasses, and costumes. These images were collected through queries such as “adult hat” and “child Halloween”. Each image is high resolution and may contain multiple faces, although most images contain solely children or solely adults. In total, 693 images were collected. Examples from this dataset are provided as Fig. 7.

The RedLight dataset [59] contains 12,180 images, manually separated into several categories describing the type of pornography. It also contains 15,401 images which are non-pornographic to help evaluate models testing on this dataset. This dataset was collected by the Digital Forensics and Cyber Security Center at the University of Rhode Island in 2010 to assist in the development of the RedLight pornography scanner. One issue prevalent in skin tone based pornography detection tools is the inability to generalize to all races. RedLight provides several hundred NSFW images categorized into various ethnicities, which we view as important to report results on.

NPDI [50] is another pornographic dataset containing only videos, separated into three groups: non-porn easy, non-porn difficult, and porn. These contain 200, 200, and 400 videos respectively. The non-porn hard subset of the dataset was intended to be extremely challenging and contains many tricky videos, such as those containing people at the beach, wrestling, swimming, and even infants breastfeeding. The non-porn easy videos were randomly selected from YouTube³. The group of pornography videos were selected from websites which exclusively host content of pornographic nature, and includes examples of animated pornography. In total, there is 77 hours of footage in the dataset. The pornographic video group is also ethnically diverse; 46% of the videos contain Caucasians, 16%

²<https://www.flickr.com/>

³<https://www.youtube.com>

Asians, 14% Africans, and 24% are multi-ethnic.

A private dataset was collected by the Rhode Island State Police (RISP) containing 1,109 videos and 84,619 images of child pornography collected over a multi-year period from approximately 20 real world cases. At no point did the authors of this paper have direct access to this dataset. Testing was done through coordination with a liaison at the state police. Each video is a known instance of child pornography. The true label of each image in the dataset is somewhat ambiguous as an image may not necessarily contain nudity and a minor for it to be included in the set. Each image comes from a larger collection of identified child pornography involving the given victims for each case, but is not clear whether actual nudity or the face of a child is present in each image. Therefore, as of now, we can only approximate the generalization of the NSFW detection and age estimation model on the images in this set. Keep in mind that the labels for the videos *are* reliable and metrics such as accuracy and F2 score may be calculated.

4.5 Training Details

Models were evaluated locally on an Ubuntu 14.04 server configured with 4x Nvidia Titan X Pascals, 2x Intel Xeon CPU E5-2620v4, and 64 gb of memory. All CNNs were trained and inferenced using the GPUs. OpenNSFW and the S^3FD face detector used the Caffe deep learning framework while the age estimation model runs on PyTorch. Yahoo!’s OpenNSFW nudity detection model [10] was initialized with ImageNet weights and then trained on a private dataset of both NSFW and SFW images. The S^3FD face detector [20] uses a VGG-16 architecture and is fine-tuned from ImageNet onto the WIDER FACE face detection dataset [60]. We used these two models as provided by the authors on their respective github pages with their provided weights.

Our age estimation model uses a DenseNet-161 architecture pre-trained on

ImageNet, fine-tuned onto a collection of 130,000 images of actors and actresses crawled from the publicly available IMDB-Wiki dataset [61] for the task of label distribution learning. Since apparent age guesses are not available for this dataset, the distribution was parameterized by a fixed standard deviation of 3 and a mean equivalent to the age of the subject as a proxy. Finally, our model is fine-tuned onto APPA-REAL [19] for the task of apparent age estimation using normal label distributions parameterized by the mean and standard deviation of the per-image human guesses.

Analysis of the runtime of the models deployed at the local law enforcement agency is also provided. This workstation is less powerful than the one the models were trained on, and only has 1 Nvidia Quadro P4000. In total, 1,956 videos were split into 1,044,577 frames. 123,200 images were detected. It took the file crawler 11 minutes to identify all files (single thread), 3 hours and 4 minutes for the frame extractor to extract the frames at a rate of 1 per second (using a single thread), 25 hours to detect the faces in the resulting 1,167,777 files (batch size of 1, single GPU), 56 minutes to crop the detected faces into 599,232 files (single thread), 1 hour and 13 minutes to run the age estimation model on each detected face (batch size of 64, single GPU), and 8 hours and 56 minutes to run the nudity detection model on the 1,167,777 total images/frames (batch size of 1, single GPU) for a total runtime of 39 hours.

Significant improvements in runtime could be made by altering the face detector and OpenNSFW models to support batch sizes larger than 1, and additional improvements could be made by parallelizing the file crawler, frame extractor, and face cropper. In the future, we plan to implement these improvements in addition to exploring options in reducing the complexity of the CNNs used in order to speed up inference time.

Keep in mind that the CNN used for the face detector, nudity detector, and age estimator were not fine-tuned for the task of SEIC content detection or trained on video frames at all. In the next section, we present empirical results that demonstrate the generalization abilities of these models in challenging and diverse real world environments. The age estimation model is shown to generalize to challenging images of adults and children, the never before seen space of NSFW images and the variance of the model’s predictions on video frames is analyzed. The OpenNSFW model is also validated on the RedLight dataset. To the best of our knowledge, all reported results come from proper “test” sets, that is data that the model has never been exposed to during training. Our uncertainty stems from the possibility that OpenNSFW’s private training dataset may have included some of our test data, however this is unlikely.

As mentioned in Section 4.3, a meta-classifier was trained on the age and nudity score predictions of each frame of every video. When evaluating results for video classification, a different protocol is used since some amount of model training is performed. In all cases, the data is split into two groups. 75% of the data is used to perform a 5-fold cross validation before the model is retrained on that whole subset of data to be evaluated on a held out 25% test set. Confusion matrices, plots, and the reported true positive rates come from the test set.

		Predicted		
		+	-	
Actual	+	228	38	266
	-	19	1693	1712

(a) Child ≤ 14

		Predicted		
		+	-	
Actual	+	245	66	311
	-	13	1654	1667

(b) Child ≤ 18

Figure 8: Confusion matrices for minor classification on the test fold of the APPA-Real dataset using cutoff ages 14 and 18. The model rarely mistakes an adult (-) for a child (+), but is prone to calling children adults.

4.6 Robustness of NSFW and Age Models

In this section we establish the generalization capabilities of our age estimation and nudity detection models on NSFW and SFW content, different ethnic groups, challenging images, and on videos.

4.6.1 Evaluation of Age Estimation and Nudity Detection Models on SFW and NSFW Images

Results for minor classification on the APPA-Real dataset’s test fold are presented in Fig. 8. When the age estimator is given a properly extracted image of a face, it is quite good at correctly identifying adults as adults yielding false positive rates of only 1.11% and 0.78% for cutoff ages of 14 and 18 respectively.

Next, the performance of the age estimator on all the NSFW material included in the RedLight dataset is analyzed. Since all the data was collected from legal pornography websites, it is assumed that there are no minors present in any of the images, i.e. a perfect model would label every face as being an adult. The true positive rate for adult classification is given in Table 4. In total, 8% of the faces detected in the dataset were falsely classified as containing minors. Note that the Count column represents the number of faces detected, and is different than the

	Minors	Adults	Count	Acc.
Partial Nudity	6	198	204	0.97
Partial Porn	29	450	479	0.94
Single Nude Male	6	216	222	0.97
Single Nude Female	28	469	497	0.94
3+	15	372	387	0.96
Close Pen	1	7	8	0.88
Close Up	4	18	22	0.82
Oral	26	311	337	0.92
Generic Porn	919	10326	11245	0.92

Table 4: Binary classification of faces extracted from NSFW images into adults and minors using a cutoff age of 14. The accuracy of the model degrades significantly here in comparison to APPA-Real, but still generalizes well to this far more uncontrolled environment.

number of images in the dataset. The face detector may not have detected a face in every image, or in some instances multiple faces were detected and analyzed.

A detailed description of each of the categories shown as rows in Table 4 are as follows. Partial nudity includes suggestive imagery such as lingerie models and people in bathing suits, some of which are provocatively posed but without exposed genitalia. Partial porn contains instances of pornographic images, with clear sexual context, in situations where the participants are still partially clothed. The next three categories refer to the gender or number of subjects in the images. Close Pen refers to nearly full frame images of penetration, while close up more generally refers to closely cropped images of pornographic content. Images residing in the Oral category depict acts of Fellatio. Finally, the last category contains general instances of NSFW or SFW images without any fine-grained labels. Most of these images are of very high quality, and provide a good upper bound for both models’ performance on NSFW images.

Since RedLight was not collected with the intention of performing facial analysis some detected faces may be of significantly worse quality or smaller sized than those in APPA-REAL, and some images may not contain faces at all. Addition-

ally, there are some artifacts unique to pornographic images, such as strong facial expressions, or seminal fluid obstructing a face that the age estimation model was not exposed to during its training. In contrast to APPA-REAL, no human verification was performed to ensure proper detection of faces. This may help explain the higher false positive rate on this particular dataset. The imperfection of the face detector is showcased by the diminutive number of faces detected in the Close Pen and Close Up categories. Very few faces should have been detected in images of these categories.

It is suspected that the feature space of pornographic images is too different from “natural” images, and the quality of the age estimation model is suffering. The number of false positives (minors) detected in RedLight may be significantly reduced by fine-tuning the model on such content, but that would create a serious issue with class imbalance as no labeled faces of children in that context are available. A more feasible approach would be to fine-tune the nudity detection model directly on SEIC material to discriminate SEIC vs. non-SEIC content, as was done in other recent works, so there would be no concerns about gathering the detailed facial labels required to fine-tune the age estimator.

The adult actresses found in this type of material are also typically biased towards the younger side as many popular categories of pornography predominantly feature young looking adults. Making the distinction between these adult actresses and actual minors is extremely difficult, even for trained professionals such as the law enforcement agents interviewed in [54]. One participant of this study is reported as saying

“I was starting to struggle around teenage years, because some I thought, oh, they could be 12 or they could be 18.”. Another participant said “The ones that are difficult are when there’s sort of . . . well, its the age, isn’t it, whether you are looking at them thinking, well, are you 15 or are you a young-looking 18-year-old, or are you an old-looking 15-year-old, and it’s that area that’s difficult” [54].

	NSFW	SFW	Count	Acc.
Partial Nudity	146	44	190	0.77
Partial Porn	282	79	361	0.78
Solo Nude Male	222	1	223	1
Solo Nude Female	498	2	500	1
3+	174	1	175	0.99
Close Pen	174	0	174	1
Close Up	183	0	183	1
Oral	294	10	304	0.97
Generic Porn	9679	522	10201	0.95
Generic Nonporn	960	14416	15376	0.94

Table 5: Nudity detection on RedLight using the Yahoo Open NSFW model with a sensitivity level of 0.402. The model has almost perfect accuracy on the easiest examples in the dataset (solo nude male, female, and 3+) but makes some mistakes in the partial porn category.

Table 5 provides insight into the OpenNSFW model’s performance on the RedLight dataset’s images using $P_{nsfw} = 0.42$. The model is remarkably accurate on these images, except in the case of the much more ambiguous categories of partial nudity and partial porn where membership to the NSFW class is more difficult to identify due to the smaller amount of visible skin.

4.6.2 Racial Bias

	NSFW	SFW	Count	Acc.
African	194	10	204	0.95
Asian	428	15	443	0.97
Caucasian	400	0	400	1
Hispanic	267	7	274	0.97
Indian	145	5	150	0.97

Table 6: OpenNSFW accurately identifies Caucasian pornographic images but makes a few mistakes for other ethnicities.

Analyzing the racial bias of machine learning models is a relevant issue, particularly in the area of nudity detection. Traditional skin-detection based models often fail to generalize to multiple ethnicities. OpenNSFW’s ability to detect

pornography across several ethnicities is shown in Table 6. It is evident that performance does not degrade significantly from the overall average accuracy among each category. Although it is worth mentioning that the Caucasian images were classified perfectly, one cannot be certain whether the perfect performance is an artifact of the moderately low sample size, or some implicit bias; however an observer would be unlikely to say the model “breaks down” or performs poorly for non-Caucasian ethnic groups.

	Minors	Adults	Count	Acc.
African	15	175	190	0.92
Asian	46	321	367	0.87
Caucasian	27	352	379	0.93
Hispanic	18	273	291	0.94
Indian	29	101	130	0.78

Table 7: Binary classification of ethnic faces extracted from NSFW images into adults and minors using a cutoff age of 14.

In Table 7, some bias is observed in detecting the age of two different ethnicities - Asian and Indian. Overall, the model generalizes well to the difficult task of age estimation across ethnicities in NSFW images despite some loss of performance when dealing with images of people from Asia. The deficit in accuracy on Asian and Indian images is likely due to smaller presence of training examples from people of these races in the APPA-REAL dataset. A rigorous analysis of human and machine learning bias in apparent age estimation on the APPA-REAL dataset due to subject gender, ethnicity, emotional expression, and presence of makeup is given in [62].

4.6.3 Challenging Images

A dataset of challenging images of children and adults was collected for further analysis. Each image contains either children or adults wearing a costume, hat, or sunglasses. No images in the dataset contain pornography. All images are run

	+Hat	+Glasses	+Costume	-Hat	-Glasses	-Costume
Child	86	47	95	0	2	3
Adult	41	66	57	111	112	109
Acc.	0.68	0.42	0.63	1.00	0.98	0.97

Table 8: Binary classification of extremely challenging images of children (+) and adults (-) using a cutoff age of 14.

through our framework; the results for age estimation are presented in Table 8. The key observation from this experiment is that the age estimation model relies heavily on features around the eyes of the face, as is evident by the significant drop in accuracy for the group of children wearing glasses. The reliance of the model on eye features makes sense in the context of the data the age estimator was trained on. Not only are most facial images centered around the eyes, but images where the subject has blinked or the eyes are not visible due to red-eye or other artifacts are usually discarded and not uploaded to the internet.

The OpenNSFW model was also run on this dataset. Out of the 693 images in the dataset, only 12 were incorrectly classified as pornographic (1.7% false positive rate).

4.6.4 Evaluation of Age Estimation Models on Videos

Our framework operates under the assumption that if our models work well on images then they will also perform well at inference on video frames even though neither of our models were trained on video frames at all. This is a problem because the quality of video frames is generally much worse than that of images. Additionally, when a face is detected in a video the quality or resolution of the cropped image may be too degraded to get an accurate estimation of age, especially when the face is blown up to the required input size for our model (224x224). In the future, performance may be increased substantially by fine-tuning both the NSFW and age estimation modules on individual frames from videos.

Figure 9 shows our age estimation model’s predictions in a controlled setting. The video corresponding to this analysis annotated with our face detections and age predictions can be found here ⁴. The harsh variations in the model’s predictions when small rotations or changes in facial expression are induced is interesting to note, although relating these variations to more fundamental properties of convolutional neural networks is outside the scope of this thesis. A general trend to notice is that the model consistently overestimates the age of children in video.

Figure 9: Frame-by-frame analysis of age estimation model on controlled aging video. The video begins with a child of age 5 facing the camera with a monochromatic backdrop in an interview setting. Every few seconds, a new subject is introduced exactly 1 year older than the previous, up to age 75. For each frame of the video, age estimation was performed and plotted versus the ground truth age.

⁴<https://jaredrondeau.com/2018/08/30/age-estimation/>

4.7 Classification of SEIC Videos and Images

Even though deep learning models perform well individually on RedLight and other legal media, our analysis must include the target domain of child pornography detection. In this section results on SEIC image classification, NSFW video detection, and SEIC video detection are presented in addition to a discussion on the advantages of our framework from an interpretability standpoint and practical considerations.

4.7.1 SEIC Image Detection

	# Images	# NSFW	# Faces	# Minors	# CP
RISP	84619	65664	69564	27875	17248
Redlight	14200	13427	15346	1188	1116

Table 9: Speculative results for SEIC image detection by combining nudity and minor detection. A threshold of 0.402 is used to consider an image NSFW and an age cutoff of 14 is used. The number of false positives from the RedLight dataset is also provided for “loose” comparison.

A large number of illicit images were collected by a Rhode Island law enforcement agency for evaluation. All 84,619 images came from real cases and have been identified by law enforcement agents as being known instances of child pornography. However, not every image may necessarily contain nudity, a child, or a face due to the method of data collection. Each image comes from a set of known SEIC images pertaining to a particular victim, some of which may contain clothed children in non-sexual contexts. The agents method of bulk gathering of images is different than the collection criteria for RedLight, where each image contains an obvious instance of pornography. The authors of this paper cannot release the SEIC dataset that they themselves did not have direct access to for obvious legal and moral reasons.

The following conclusions may be drawn from the SEIC image dataset:

OpenNSFW is successful at detecting nudity in SEIC (although it cannot distinguish SEIC images from adult pornography) with 77.6% of such images being flagged. Less faces are present in SEIC content than in traditional pornography (0.82 faces per image in the SEIC content versus 1.08 in RedLight), and a larger proportion of those faces will contain minors (40.1% versus 7.7%). A false positive rate for SEIC versus pornographic image detection of 7.9% may be calculated based on the RedLight data.

4.7.2 NSFW Video Detection

In Table 10 results for NSFW video classification are presented from the SVM trained on histogram feature sets of age, nudity, or age and nudity model predictions. The nudity detection abilities of OpenNSFW are competitive with state-of-the-art results for pornographic video classification on NPDI even though the base model has never been trained on *any* videos itself. There is no significant improvement in NSFW video detection by adding age predictions to the classifier. It is important to note that we did not utilize the official 5 fold cross validation sets proposed by the authors of the dataset, and deviated from the other authors by using the frames we extracted instead of the key frames already provided.

	Acc
Age Features	76.0 ± 3.7
NSFW Features	94.5 ± 3.7
NSFW and Age Features	94.5 ± 3.9
Moustafa [63] 2015	94.1 ± 2.0
Perez [21] 2017	97.9 ± 0.7

Table 10: Performance on the task of pornographic video detection using the meta-classifier compared to other deep learning methods on the same dataset.

Figure 10: Sample frames drawn from the NPDI dataset. The top row displays frames from the non-porn hard subset of data. The bottom row are samples from some non-porn easy videos. Even though the content of the non-porn easy videos may be easy to distinguish from pornography, it may still be subject to challenges such as blur and poor quality.

4.7.3 SEIC vs. NPDI Videos

Finally, the main contribution of this work is presented; results for classification of child pornography videos by training an SVM on age and nudity features, referred to as our meta-classifier. The NPDI video pornography classification dataset is used as a baseline to compare our model against. The NPDI dataset contains both pornographic and non-pornographic videos. The benign videos are separated into two categories: non-porn hard and non-porn easy. Examples from the publicly available NPDI dataset are given in Fig. 10. All results in this section use a sensitivity threshold (probability) of 0.50 for classification.

In Table 11 the performance of the meta-classifier is explored as we vary the subset of NPDI the classifier is trained and evaluated on. For each experiment the subset used was split into 75% training and 25% for testing. Five-fold cross validation was used to select hyper parameters to train the final SVM on. The

NSFW	SFW Hard	SFW Easy	TPR	TNR	Acc	F2
		X	0.98	0.93	0.97	0.99
	X		0.91	0.79	0.90	0.96
	X	X	0.92	0.82	0.89	0.93
X			0.92	0.81	0.89	0.92
X		X	0.91	0.83	0.88	0.90
X	X	X	0.87	0.80	0.84	0.86

Table 11: Breakdown of model performance varying the subsets of NPDI. For each experiment the SEIC dataset was used as the positive class label while the subset composing the benign label varied as indicated by the ‘X’.

true positive rate (TPR), true negative rate (TNR), accuracy (Acc), and F2 score is presented on the held out testing set after the model was retrained on the full training set with the best performing hyper parameters.

It is important to reiterate that neither the age estimation model or nudity detection model have ever been exposed to frames/videos from NPDI or the SEIC dataset. By posing the final classification in this way the base models can be trained and tested locally on easy to obtain pornography and aging datasets. Careful evaluation and analysis of ethnic bias, challenging examples, and common mistakes may be performed that is not possible on the target dataset due to its grotesque and restricted nature.

Nonetheless, as shown in the confusion matrix presented as Fig. 11a the final classifier for SEIC vs. easy non-porn achieves an accuracy of 97% on the hold-out set. Of the 49 instances of non-porn, only 6 were falsely flagged as SEIC. Likewise, of the 278 instances of SEIC only 3 were mistaken for being benign.

This is the most ideal case for the classifier, as no nudity is present in the easy non-porn videos. Distinguishing pornography from child pornography poses a much more difficult task, as seen in Fig. 11b, as both types of videos contain a large amount of nudity. This leaves the output of the age estimation model, a task much more sensitive to the content and quality of the extracted frames and faces,

		Predicted										
		+	-			+	-			+	-	
Actual	+	275	3	278	+	248	20	268	+	239	40	279
	-	6	43	49	-	22	87	109	-	35	163	198

(a) Non-porn Easy

(b) Pornography

(c) All NPDI

Figure 11: Confusion matrices for SEIC (+) video classification, and the subset used for comparison (-) varies.

as the deciding feature set for the SVM.

A visualization of the feature space in terms of the proportion of faces containing minors and frames nudity is given in Fig. 12. It is evident that the number of false-positives given by the age estimator is not negligible, as many pornographic videos are detected to have many minors in them. In general, the pornographic videos are clustered to the top left of the plot. The non-porn easy videos are nearly all in the bottom left. The non-porn hard videos are scattered, as are the SEIC videos.

Another surprising observation may be made from this plot. Many SEIC videos are found to contain no faces of children, but a large amount of nudity. Pornographic videos do not have this property; almost all of these videos have between 5% and 20% of the faces mistakenly classified as under 18. It seems that the meta-classifier learned this strange relationship in our datasets and was still able to distinguish between the two groups successfully.

4.7.4 Interpretability (Sample Report/Qualitative Analysis)

One advantage of this hybrid classification approach to SEIC content detection is the large amount of interpretability available. For each image the number of detected faces as well as the suspected ages and nudity score can be quantified and

Classification Performance using NPD1

Figure 12: Plot of errors for RISP vs. all of NPD1 classification. The proportion of faces where a minor (age ≤ 18) is detected is plotted against the proportion of frames considered NSFW by OpenNSFW using a threshold of 0.402. Mistakes are marked by an ‘X’.

presented to law enforcement agents. Similar numbers may be displayed for each video as well. The selectivity of the model may also be tuned by the agent. The nudity score and age used for the SEIC image detection results was chosen as a baseline for comparison to NSFW image detection, and if more false positives are tolerable to the agent, these values could be tuned via a “slider” option by the law enforcement agent. If the minor-filtering is greatly relaxed (perhaps to age 25), the agent will be left with a world-class pornography scanner to help speed up the search of seized media.

On the other hand, if a tool was to be deployed to monitor images and videos passing through a network or another high traffic application where even a 1% false positive rate is not acceptable, these options could be made very restrictive and only flag the most suspicious content for human review i.e, images where nudity and a child’s face is detected with high probability or videos where most of the frames are identified similarly. This level of interpretability may be used to

generate a report with many options for filtering and sorting for the user. Perhaps an agent wants to sort by the probability of SEIC, or only review content with faces and nudity (irrespective of the presence of a detected child).

Another avenue, unexplored so far in this work, is file system analysis. Due to the structured nature of the file system on seized computer, it seems natural to assume that directories adjacent to those containing SEIC content are also likely to have this type of content as well. Instead of ranking images and videos by probability of containing SEIC, entire directories for the agent to look through could be displayed with sorting features. A plot similar to Fig. 12 could be presented for the agent, and a pop up could appear with additional details such as the filename, key frames, and other useful information when the mouse hovers over a point on the plot. Points in the top right corner of the plot have a large proportion of children and NSFW content and almost certainly contain SEIC.

Finally, we present an additional more qualitative analysis of the mistakes made by the meta-classifier in the case of SEIC video versus benign classification on NPDI videos. In this particular experiment, 489 videos were used in the held out test set. Of these videos 290 and 199 were SEIC and benign respectively. For the benign videos, 13/91 consisted of pornographic content and were misclassified, 3/57 of the non-porn easy subset and 10/47 from non-porn hard.

The benign pornographic videos mistaken as SEIC have a sizable number of detected faces and a good proportion of the faces were considered to be children using a cutoff of 14 years (10 to 30 percent), although two of the videos had 0 children detected in total. Of the 3 misclassified non-porn easy videos, two of them had a large number of frames misclassified as containing NSFW content. The other video had no frames considered to contain NSFW content. Manual inspection of the non-porn hard videos was performed to help understand the errors made. Of

the 10 mistakes, 8 may be considered understandable as there is a lot of nuance to capture in these videos. These 8 videos contain either infants breastfeeding or children and adults on the beach in swim wear. The other two mistaken hard videos were a girl twerking and MMA fighters wrestling while shirtless.

On the other hand, of the 39 SEIC videos falsely classified as being benign 26 had less than 5 children’s faces detected. All but one of these videos are longer than 1 minute which implies that a child’s face was visible on camera for about 5 seconds, although it is quite likely the model mistakenly called some children adults in other frames. It is not clear why the remaining 13 videos could have been misclassified.

4.7.5 Practical Limitations and Considerations

While the results can be seen as impressive, this framework should only be used as a filter to flag suspected instances of digital media for further human review. The law enforcement agent in the middle can not be removed from the process of tagging SEIC content, and the agent should be aware that while the age estimation model has been rigorously validated in many different scenarios, it may still fail for not obvious reasons. Some SEIC content will go undetected on a suspected hard drive, and some borderline material will be falsely flagged as being SEIC. However it is often the case the suspects of child pornography cases have many gigabytes of SEIC. The age estimation model will only work best on images where faces are present and in reasonable quality. In the end, it must be up to the human agent to decide whether or not a particular file should be used for prosecution or identified as SEIC content.

As is evident by the discussion between law enforcement agents regarding the classification of this type of material, there is also a large amount of nuance to be captured that we can not explicitly model:

“But in terms of any sort of sexual nature, actually in the image, there isn’t any. Because it is actually a family photograph (Agent 1). Yeah, and in fact, despite the fact that they’re naked, it’s completely unisexual . . . It’s not . . . there’s nothing about it that’s indecent (Agent 3).” [54]

In this particular scenario our model would almost certainly flag this particular suspected image as SEIC. The human agents must be kept in the loop to rectify these types of mistakes. The true ability of an automated tool to cut down on the subjectivity and personal bias exhibited by law enforcement agents, as called for in [54], would have to be further analyzed before replacing humans is even considered. If this is validated, however, the potential for further cutting down on the emotional damage and time sink incurred upon the law enforcement personnel manually identifying this material is great.

CHAPTER 5

Conclusion

In this thesis we presented a study of convolutional neural networks trained on label distributions for the task of predicting the apparent age of face images. We provided analysis on different ways to pose the objective function for this task, and showed empirically that label distributions are the most intuitive solution. Models pre-trained using label distributions learn better underlying feature representations even when the true distribution is not available. We also achieved a new state-of-the-art mean absolute error in apparent age estimation on the APPA-Real dataset. Our best model implements a KL-divergence loss function and defines the ground-truth label of each image as a nonparametric density function estimated from its corresponding human guesses.

Additionally, we presented a framework for the automated detection of Sexually Exploitative Imagery of Children images and video by utilizing a fusion of age estimation and nudity detection model predictions. A rigorous analysis of the models' performance at age estimation and nudity detection was conducted, including experiments on ethnically diverse and challenging data. We examined the variability of modern age estimation models by utilizing the unique nature of the GLAMOUR YouTube videos. Finally, our framework was validated on a real world dataset consisting of 1,109 videos and 84,619 images of SEIC content from 20 cases collected by local law enforcement agents. We achieve 97% accuracy in SEIC video classification (SEIC versus easy non-porn), 94.5% accuracy in NSFW video classification (NSFW versus non-porn), 89% accuracy in SEIC versus NSFW video classification, and finally, 84% accuracy in SEIC versus all of NPDI video classification.

In the future, we hope to expand upon our work by analyzing the interpretability of our predicted label distributions and by applying unsupervised learning to exploit the massive amount of unlabeled face-containing data. It has been recently shown that fine-tuning the nudity detection module on both NPDI video frames and the target illicit material will result in a significant boost in performance [64] (9 percentage point boost in accuracy over OpenNSFW model in their experiments). Additionally, models taking advantage of spatio-temporal features have been shown to improve results in pornographic video detection as well [21]. The SEIC video detection procedure could be extended to have the frame extractor make several passes over a video, extracting additional frames to better inform the final classifier if the nature of the video is ambiguous.

Perhaps most exciting, however, is the prospect of investigating minor detection and general apparent age estimation in videos. By utilizing the GLAMOUR YouTube videos, we hope to soon release a framework for evaluating age estimation models in videos. Preliminary results on this proposed work resulted in analysis of the variability and sensitivity of our age estimation models' predictions in videos. This served as inspiration for a new semi-supervised learning algorithm to take advantage of the massive amount of unlabeled facial data available in videos.

The proposed algorithm will train a model with age-labeled facial images jointly with unlabeled frames of videos. Suppose we have a video of the popular American actor Nicholas Cage. This video will be split up into a series of frames, and a face detector will be run to extract all faces. Next, a facial verification model will be used to identify all facial images of Nicholas Cage. All other faces will be thrown out. We are then left with a large set of frames all containing the same person throughout the video. Such a dataset already exists with millions of unlabeled faces organized by human subject.

This algorithm will exploit the assumption that a person will not age throughout a video. If a person's age does not change, then our model should predict the same age consistently for every extracted facial image. We can therefore enforce some consistency loss penalizing the model for predicting different ages for the same person. The model will then be jointly optimized with the labeled image data to train a model which should be far more robust to spatial rotation of the face, closed eyes, and other phenomenon found in videos. The predictions of our model should also be much more consistent for the same person.

The overarching goal of this future work is to improve the age estimation models and further tackle the problem of SEIC detection. It is the belief of the author that based on the results shown thus far, this is an extremely important real world problem that is ready to be solved.

LIST OF REFERENCES

- [1] E. Agustsson, R. Timofte, S. Escalera, X. Baró, I. Guyon, and R. Rothe, “Apparent and real age estimation in still images with deep residual regressors on appa-real database,” in *IEEE International Conference on Automatic Face Gesture Recognition (FG)*, 2017, pp. 87–94.
- [2] “NCMEC quick facts page,” <http://www.missingkids.org/KeyFacts>, Last accessed 23 Apr 2018.
- [3] H. Han, C. Otto, and A. Jain, “Age estimation from face images: human vs. machine performance,” in *International Conference on Biometrics (ICB)*, 2013, pp. 1–8.
- [4] A. Lanitis, “A survey of the effects of aging on biometric identity verification,” *International Journal of Biometrics*, vol. 2, no. 1, pp. 34–52, 2009.
- [5] Y. Zhu, Y. Li, G. Mu, and G. Guo, “A study on apparent age estimation,” in *IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2015, pp. 267–273.
- [6] R. Rothe, R. Timofte, and L. Van Gool, “Deep expectation of real and apparent age from a single image without facial landmarks,” *International Journal of Computer Vision*, 2016.
- [7] S. Escalera, X. Baró, H. Escalante, and I. Guyon, “Chalearn looking at people: A review of events and resources,” in *International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 1594–1601.
- [8] X. Geng, “Label distribution learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1734–1748, 2016.
- [9] J. Rondeau and M. Alvarez, “Deep modeling of human age guesses for apparent age estimation,” in *International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro, Brazil, 2018.
- [10] J. Mahadeokar and G. Pesavento, “Open sourcing a deep learning solution for detecting nsfw images,” 2016.
- [11] C. Fernández, I. Huerta, and A. Prati, “A comparative evaluation of regression learning algorithms for facial age estimation,” in *Face and Facial Expression Recognition from Real World Videos*. Springer, 2015, pp. 133–144.
- [12] G. Guo, G. Mu, Y. Fu, and T. S. Huang, “Human age estimation using bio-inspired features,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 112–119.

- [13] A. Lanitis, C. Draganova, and C. Christodoulou, “Comparing different classifiers for automatic age estimation,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 1, pp. 621–628, 2004.
- [14] X. Liu, S. Li, M. Kan, J. Zhang, S. Wu, W. Liu, H. Han, S. Shan, and X. Chen, “Agenet: Deeply learned regressor and classifier for robust apparent age estimation,” in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 16–24.
- [15] Y. Fu, G. Guo, and T. Huang, “Age synthesis and estimation via faces: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 1955–1976, 2010.
- [16] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, “Ordinal regression with multiple output cnn for age estimation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4920–4928.
- [17] S. Escalera, J. Fabian, P. Pardo, X. Baro, J. Gonzalez, H. J. Escalante, D. Mivsevic, U. Steiner, and I. Guyon, “Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results,” in *The IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2015.
- [18] S. Escalera, M. Torres Torres, B. Martinez, X. Baro, H. Jair Escalante, I. Guyon, G. Tzimiropoulos, C. Corneou, M. Oliu, M. Ali Bagheri, and M. Valstar, “Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2016.
- [19] E. Agustsson, R. Timofte, S. Escalera, X. Baro, I. Guyon, and R. Rothe., “Apparent and real age estimation in still images with deep residual regressors on appa-real database.” in *12th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2017*. IEEE, 2017.
- [20] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, “S3fd: Single shot scale-invariant face detector,” *CoRR*, vol. abs/1708.05237, 2017. [Online]. Available: <http://arxiv.org/abs/1708.05237>
- [21] M. Perez, S. Avila, D. Moreira, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, and A. Rocha, “Video pornography detection through deep learning techniques and motion information,” *Neurocomputing*, vol. 230, pp. 279–293, 2017.
- [22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

- [23] T. Weyand, I. Kostrikov, and J. Philbin, “Planet-photo geolocation with convolutional neural networks,” in *European Conference on Computer Vision*. Springer, 2016, pp. 37–55.
- [24] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *European Conference on Computer Vision*. Springer, 2016, pp. 694–711.
- [25] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?” in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 3320–3328.
- [26] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *nature*, vol. 521, no. 7553, p. 436, 2015.
- [27] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, <http://www.deeplearningbook.org>.
- [28] B. Xu, N. Wang, T. Chen, and M. Li, “Empirical evaluation of rectified activations in convolutional network,” *CoRR*, vol. abs/1505.00853, 2015. [Online]. Available: <http://arxiv.org/abs/1505.00853>
- [29] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *CoRR*, vol. abs/1502.03167, 2015. [Online]. Available: <http://arxiv.org/abs/1502.03167>
- [30] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [31] X. Geng, C. Yin, and Z. Zhou, “Facial age estimation by learning from label distributions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 10, pp. 2401–2412, 2013.
- [32] X. Yang, X. Geng, and D. Zhou, “Sparsity conditional energy label distribution learning for age estimation,” in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2016, pp. 2259–2265.
- [33] A. Berger, V. Pietra, and S. Pietra, “A maximum entropy approach to natural language processing,” *Computational Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.
- [34] X. Geng and P. Hou, “Pre-release prediction of crowd opinion on movies by label distribution learning,” in *International Conference on Artificial Intelligence (IJCAI)*, 2015, pp. 3511–3517.

- [35] X. C., G. X., and X. H., “Logistic boosting regression for label distribution learning,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4489–4497.
- [36] W. Shen, K. Zhao, Y. Guo, and A. Yuille, “Label distribution learning forests,” in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 834–843.
- [37] B. Gao, C. Xing, C. Xie, J. Wu, and X. Geng, “Deep label distribution learning with label ambiguity,” *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2825–2838, 2017.
- [38] S. T. Roberts, “Content moderation,” *Encyclopedia of Big Data*, pp. 1–4, 2017.
- [39] A. Chen, “The laborers who keep dick pics and beheadings out of your facebook feed,” Octobet 2014.
- [40] C. Ries and R. Lienhart, “A survey on visual adult image recognition,” *Multimedia tools and applications*, vol. 69, no. 3, pp. 661–688, 2014.
- [41] C. Caetano, S. Avila, S. Guimarães, and A. d. A. Araújo, “Representing local binary descriptors with bossanova for visual recognition,” in *Proceedings of the 29th Annual ACM Symposium on Applied Computing*. ACM, 2014, pp. 49–54.
- [42] E. Valle, S. de Avila, A. d. Luz Jr, F. de Souza, M. Coelho, and A. Araújo, “Content-based filtering for video sharing social networks,” *arXiv preprint arXiv:1101.2427*, 2011.
- [43] K. Tanner, “Modeling automated detection of children in images,” Master’s thesis, University of Rhode Island, 2011.
- [44] N. Sae-Bae, X. Sun, H. T. Sencar, and N. D. Memon, “Towards automatic detection of child pornography,” in *ICIP*, 2014, pp. 5332–5336.
- [45] S. Karavarsamis, N. Ntarmos, K. Blekas, and I. Pitas, “Detecting pornographic images by localizing skin rois,” *International Journal of Digital Crime and Forensics (IJDCF)*, vol. 5, no. 1, pp. 39–53, 2013.
- [46] A. Ulges and A. Stahl, “Automatic detection of child pornography using color visual words,” in *Multimedia and Expo (ICME), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1–6.
- [47] M. Moustafa, “Applying deep learning to classify pornographic images and videos,” *arXiv preprint arXiv:1511.08899*, 2015.

- [48] 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.
- [49] J. Wehrmann, G. S. Simões, R. C. Barros, and V. F. Cavalcante, “Adult content detection in videos with convolutional and recurrent neural networks,” *Neurocomputing*, vol. 272, pp. 432–438, 2018.
- [50] S. E. F. de Avila, N. Thome, M. Cord, E. Valle, and A. de Albuquerque Araújo, “Pooling in image representation: The visual codeword point of view,” *Computer Vision and Image Understanding*, vol. 117, pp. 453–465, 2013.
- [51] “Microsoft PhotoDNA,” <https://www.microsoft.com/en-us/PhotoDNA>, Last accessed 23 Apr 2018.
- [52] C. Schulze, D. Henter, D. Borth, and A. Dengel, “Automatic detection of csa media by multi-modal feature fusion for law enforcement support,” in *Proceedings of International Conference on Multimedia Retrieval*. ACM, 2014, p. 353.
- [53] P. Vitorino, S. Avila, M. Perez, and A. Rocha, “Leveraging deep neural networks to fight child pornography in the age of social media,” *Journal of Visual Communication and Image Representation*, vol. 50, pp. 303–313, 2018.
- [54] J. A. Kloess, J. Woodhams, H. Whittle, T. Grant, and C. E. Hamilton-Giachritsis, “The challenges of identifying and classifying child sexual abuse material,” *Sexual Abuse*, p. 1079063217724768, 2017.
- [55] B. Silverman, *Density estimation for statistics and data analysis*. CRC Press, 1986, vol. 26.
- [56] G. Ozbulak, Y. Aytar, and H. Ekenel, “How transferable are cnn-based features for age and gender classification?” in *International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2016, pp. 1–6.
- [57] H. Ng, V. Nguyen, V. Vonikakis, and S. Winkler, “Deep learning for emotion recognition on small datasets using transfer learning,” in *ACM International Conference on Multimodal Interaction (ICMI)*, 2015, pp. 443–449.
- [58] G. Huang, Z. Liu, L. Van Der Maaten, and K. Weinberger, “Densely connected convolutional networks,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [59] S. P. Alvarez, *RedLight an efficient illicit image detection application for law enforcement*. University of Rhode Island, 2012.

- [60] S. Yang, P. Luo, C. C. Loy, and X. Tang, “Wider face: A face detection benchmark,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [61] R. Rothe, R. Timofte, and L. V. Gool, “Deep expectation of real and apparent age from a single image without facial landmarks,” *International Journal of Computer Vision (IJCV)*, July 2016.
- [62] A. Clapés, O. Bilici, D. Temirova, E. Avots, G. Anbarjafari, and S. Escalera, “From apparent to real age: gender, age, ethnic, makeup, and expression bias analysis in real age estimation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2373–2382.
- [63] M. Moustafa, “Applying deep learning to classify pornographic images and videos,” *arXiv preprint arXiv:1511.08899*, 2015.
- [64] P. Vitorino, S. Avila, M. Perez, and A. Rocha, “Leveraging deep neural networks to fight child pornography in the age of social media,” *Journal of Visual Communication and Image Representation*, vol. 50, pp. 303–313, 2018.

APPENDIX

Additional Figures

A.1 Learned Label Distributions

Additional examples of predicted label distributions using Kernel Density Estimation (KDE) and 101-way classification are given in this section. Note the smoother output distributions from KDE. The classification model is generally very confident in a single age class resulting in sharp peaks in the output distribution while the label distribution models are much smoother. These figures are identical, enlarged versions of the ones given in Figure 5.

BIBLIOGRAPHY

- “Microsoft PhotoDNA,” <https://www.microsoft.com/en-us/PhotoDNA>, Last accessed 23 Apr 2018.
- “NCMEC quick facts page,” <http://www.missingkids.org/KeyFacts>, Last accessed 23 Apr 2018.
- Agustsson, E., Timofte, R., Escalera, S., Baró, X., Guyon, I., and Rothe, R., “Apparent and real age estimation in still images with deep residual regressors on appa-real database,” in *IEEE International Conference on Automatic Face Gesture Recognition (FG)*, 2017, pp. 87–94.
- Agustsson, E., Timofte, R., Escalera, S., Baro, X., Guyon, I., and Rothe., R., “Apparent and real age estimation in still images with deep residual regressors on appa-real database.” in *12th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2017*. IEEE, 2017.
- Alvarez, S. P., *RedLight an efficient illicit image detection application for law enforcement*. University of Rhode Island, 2012.
- Berger, A., Pietra, V., and Pietra, S., “A maximum entropy approach to natural language processing,” *Computational Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.
- C., X., X., G., and H., X., “Logistic boosting regression for label distribution learning,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4489–4497.
- Caetano, C., Avila, S., Guimarães, S., and Araújo, A. d. A., “Representing local binary descriptors with bossanova for visual recognition,” in *Proceedings of the 29th Annual ACM Symposium on Applied Computing*. ACM, 2014, pp. 49–54.
- Chen, A., “The laborers who keep dick pics and beheadings out of your facebook feed,” Octobet 2014.
- Clapés, A., Bilici, O., Temirova, D., Avots, E., Anbarjafari, G., and Escalera, S., “From apparent to real age: gender, age, ethnic, makeup, and expression bias analysis in real age estimation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2373–2382.
- de Avila, S. E. F., Thome, N., Cord, M., Valle, E., and de Albuquerque Araújo, A., “Pooling in image representation: The visual codeword point of view,” *Computer Vision and Image Understanding*, vol. 117, pp. 453–465, 2013.

- Escalera, S., Baró, X., Escalante, H., and Guyon, I., “Chalearn looking at people: A review of events and resources,” in *International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 1594–1601.
- Escalera, S., Fabian, J., Pardo, P., Baro, X., Gonzalez, J., Escalante, H. J., Misevic, D., Steiner, U., and Guyon, I., “Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results,” in *The IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2015.
- Escalera, S., Torres Torres, M., Martinez, B., Baro, X., Escalante, H., Guyon, I., Tzimiropoulos, G., Corneou, C., Oliu, M., Ali Bagheri, M., and Valstar, M., “Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2016.
- Fernández, C., Huerta, I., and Prati, A., “A comparative evaluation of regression learning algorithms for facial age estimation,” in *Face and Facial Expression Recognition from Real World Videos*. Springer, 2015, pp. 133–144.
- Fu, Y., Guo, G., and Huang, T., “Age synthesis and estimation via faces: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 1955–1976, 2010.
- Gao, B., Xing, C., Xie, C., Wu, J., and Geng, X., “Deep label distribution learning with label ambiguity,” *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2825–2838, 2017.
- Geng, X., “Label distribution learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1734–1748, 2016.
- Geng, X. and Hou, P., “Pre-release prediction of crowd opinion on movies by label distribution learning,” in *International Conference on Artificial Intelligence (IJCAI)*, 2015, pp. 3511–3517.
- Geng, X., Yin, C., and Zhou, Z., “Facial age estimation by learning from label distributions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 10, pp. 2401–2412, 2013.
- Goodfellow, I., Bengio, Y., and Courville, A., *Deep Learning*. MIT Press, 2016, <http://www.deeplearningbook.org>.
- Guo, G., Mu, G., Fu, Y., and Huang, T. S., “Human age estimation using bio-inspired features,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 112–119.

- Han, H., Otto, C., and Jain, A., “Age estimation from face images: human vs. machine performance,” in *International Conference on Biometrics (ICB)*, 2013, pp. 1–8.
- Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K., “Densely connected convolutional networks,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- Ioffe, S. and Szegedy, C., “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *CoRR*, vol. abs/1502.03167, 2015. [Online]. Available: <http://arxiv.org/abs/1502.03167>
- Johnson, J., Alahi, A., and Fei-Fei, L., “Perceptual losses for real-time style transfer and super-resolution,” in *European Conference on Computer Vision*. Springer, 2016, pp. 694–711.
- Karavarsamis, S., Ntarmos, N., Blekas, K., and Pitas, I., “Detecting pornographic images by localizing skin rois,” *International Journal of Digital Crime and Forensics (IJDCF)*, vol. 5, no. 1, pp. 39–53, 2013.
- Kloess, J. A., Woodhams, J., Whittle, H., Grant, T., and Hamilton-Giachritsis, C. E., “The challenges of identifying and classifying child sexual abuse material,” *Sexual Abuse*, p. 1079063217724768, 2017.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E., “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- Lanitis, A., “A survey of the effects of aging on biometric identity verification,” *International Journal of Biometrics*, vol. 2, no. 1, pp. 34–52, 2009.
- Lanitis, A., Draganova, C., and Christodoulou, C., “Comparing different classifiers for automatic age estimation,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 1, pp. 621–628, 2004.
- LeCun, Y., Bengio, Y., and Hinton, G., “Deep learning,” *nature*, vol. 521, no. 7553, p. 436, 2015.
- Liu, X., Li, S., Kan, M., Zhang, J., Wu, S., Liu, W., Han, H., Shan, S., and Chen, X., “Agenet: Deeply learned regressor and classifier for robust apparent age estimation,” in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 16–24.
- Mahadeokar, J. and Pesavento, G., “Open sourcing a deep learning solution for detecting nsfw images,” 2016.
- Moustafa, M., “Applying deep learning to classify pornographic images and videos,” *arXiv preprint arXiv:1511.08899*, 2015.

- Moustafa, M., “Applying deep learning to classify pornographic images and videos,” *arXiv preprint arXiv:1511.08899*, 2015.
- Ng, H., Nguyen, V., Vonikakis, V., and Winkler, S., “Deep learning for emotion recognition on small datasets using transfer learning,” in *ACM International Conference on Multimodal Interaction (ICMI)*, 2015, pp. 443–449.
- Nian, F., Li, T., Wang, Y., Xu, M., and Wu, J., “Pornographic image detection utilizing deep convolutional neural networks,” *Neurocomputing*, vol. 210, pp. 283–293, 2016.
- Niu, Z., Zhou, M., Wang, L., Gao, X., and Hua, G., “Ordinal regression with multiple output cnn for age estimation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4920–4928.
- Ozbulak, G., Aytar, Y., and Ekenel, H., “How transferable are cnn-based features for age and gender classification?” in *International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2016, pp. 1–6.
- Perez, M., Avila, S., Moreira, D., Moraes, D., Testoni, V., Valle, E., Goldenstein, S., and Rocha, A., “Video pornography detection through deep learning techniques and motion information,” *Neurocomputing*, vol. 230, pp. 279–293, 2017.
- Ries, C. and Lienhart, R., “A survey on visual adult image recognition,” *Multimedia tools and applications*, vol. 69, no. 3, pp. 661–688, 2014.
- Roberts, S. T., “Content moderation,” *Encyclopedia of Big Data*, pp. 1–4, 2017.
- Rondeau, J. and Alvarez, M., “Deep modeling of human age guesses for apparent age estimation,” in *International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro, Brazil, 2018.
- Rothe, R., Timofte, R., and Van Gool, L., “Deep expectation of real and apparent age from a single image without facial landmarks,” *International Journal of Computer Vision*, 2016.
- Rothe, R., Timofte, R., and Gool, L. V., “Deep expectation of real and apparent age from a single image without facial landmarks,” *International Journal of Computer Vision (IJCV)*, July 2016.
- Sae-Bae, N., Sun, X., Sencar, H. T., and Memon, N. D., “Towards automatic detection of child pornography,” in *ICIP*, 2014, pp. 5332–5336.
- Schulze, C., Henter, D., Borth, D., and Dengel, A., “Automatic detection of csa media by multi-modal feature fusion for law enforcement support,” in *Proceedings of International Conference on Multimedia Retrieval*. ACM, 2014, p. 353.

- Shen, W., Zhao, K., Guo, Y., and Yuille, A., “Label distribution learning forests,” in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 834–843.
- Silverman, B., *Density estimation for statistics and data analysis*. CRC Press, 1986, vol. 26.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R., “Dropout: a simple way to prevent neural networks from overfitting,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- Tanner, K., “Modeling automated detection of children in images,” Master’s thesis, University of Rhode Island, 2011.
- Ulges, A. and Stahl, A., “Automatic detection of child pornography using color visual words,” in *Multimedia and Expo (ICME), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1–6.
- Valle, E., de Avila, S., Luz Jr, A. d., de Souza, F., Coelho, M., and Araújo, A., “Content-based filtering for video sharing social networks,” *arXiv preprint arXiv:1101.2427*, 2011.
- Vitorino, P., Avila, S., Perez, M., and Rocha, A., “Leveraging deep neural networks to fight child pornography in the age of social media,” *Journal of Visual Communication and Image Representation*, vol. 50, pp. 303–313, 2018.
- Vitorino, P., Avila, S., Perez, M., and Rocha, A., “Leveraging deep neural networks to fight child pornography in the age of social media,” *Journal of Visual Communication and Image Representation*, vol. 50, pp. 303–313, 2018.
- Wehrmann, J., Simões, G. S., Barros, R. C., and Cavalcante, V. F., “Adult content detection in videos with convolutional and recurrent neural networks,” *Neurocomputing*, vol. 272, pp. 432–438, 2018.
- Weyand, T., Kostrikov, I., and Philbin, J., “Planet-photo geolocation with convolutional neural networks,” in *European Conference on Computer Vision*. Springer, 2016, pp. 37–55.
- Xu, B., Wang, N., Chen, T., and Li, M., “Empirical evaluation of rectified activations in convolutional network,” *CoRR*, vol. abs/1505.00853, 2015. [Online]. Available: <http://arxiv.org/abs/1505.00853>
- Yang, S., Luo, P., Loy, C. C., and Tang, X., “Wider face: A face detection benchmark,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

- Yang, X., Geng, X., and Zhou, D., “Sparsity conditional energy label distribution learning for age estimation,” in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2016, pp. 2259–2265.
- Yosinski, J., Clune, J., Bengio, Y., and Lipson, H., “How transferable are features in deep neural networks?” in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 3320–3328.
- Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., and Li, S. Z., “S3fd: Single shot scale-invariant face detector,” *CoRR*, vol. abs/1708.05237, 2017. [Online]. Available: <http://arxiv.org/abs/1708.05237>
- Zhu, Y., Li, Y., Mu, G., and Guo, G., “A study on apparent age estimation,” in *IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2015, pp. 267–273.