Don't Take Things at Face Value: Facial Age Classification

Sriya Bulusu
University of Washington
bulusu@cs.washington.edu

Lam Pham
University of Washington
lampham@cs.washington.edu

Deeksha Vatwani
University of Washington
vatwani2@cs.washington.edu

Abstract

Facial age verification is increasingly important for protecting safety and enforcing age-restricted services, yet current methods are unreliable or invasive. We investigate deep learning–based facial age classification specifically within the challenging 9–25 age range, where legal thresholds (e.g., 13, 18, 21) are critical. After discovering severe skew and imbalance in the UTK-Face dataset, we constructed a custom dataset by filtering IMDB-Wiki and supplementing it with UTK-Face to achieve bin-balanced representation. We trained three VGG-inspired CNN architectures of increasing depth and evaluated them across four legal-age bins. While models achieved high training accuracy, test performance remained poor and Grad-CAM analysis revealed reliance on background artifacts rather than facial features. We conclude that dataset limitations and architectural pitfalls hinder performance, and propose future work including dataset refinement, adversarial robustness testing, and architecture redesign for improved interpretability and generalization.

1. Introduction

As digital platforms and online services increasingly permeate everyday life, the need for reliable age verification has become more critical than ever. In an era of online anonymity, accurate age estimation plays a vital role in protecting user safety and ensuring compliance with legal regulations. From social media platforms that require users to be at least 13, to e-commerce and delivery services that sell age-restricted products like alcohol and tobacco to only those 21 and over, to entertainment platforms offering age-restricted content, demand for more granular age classifications is only expanding.

Humans also struggle with the facial estimation task due to subjective biases and natural variance in aging appearance. For instance, lifestyle, genetics, and cultural differences can all impact how young or old someone appears, making visual estimation difficult even for trained professionals. CNNs could provide more objective feature extraction and discrimination for this purpose.

Thus, we introduce a deep learning facial age estimation model as a robust approach to age verification across our digital concerns. Our goal is to enable accurate, privacy-conscious age verification for account creation, purchases of restricted goods, and access to age-gated content—while highlighting interpretability challenges inherent to this domain.

2. Related Works

This work relies on an understanding of use-cases for age-verification and current deep learning methods, both of which are described below.

2.1. Age Verification in Practice

Age verification is crucial for a range of online platforms in the U.S., from social media to e-commerce of age-restricted goods. Social media services like Facebook, Instagram, and Twitter officially bar children under 13 (to comply with COPPA regulations) by requiring a date-of-birth attestation during sign-up [2]. In practice, this self-reported age gate is often ineffective—for example, a survey found 45% of 12-year-olds in the U.S. were active on social networks despite being underage, often with parental awareness or assistance [2]. Simply asking users to input their birthdate fails as minors can easily lie about their age to gain access.

E-commerce and delivery services for alcohol, tobacco, and other adult goods face similar challenges. Many online retailers rely on rudimentary age checks (a checkbox or birthdate entry), ID upload at purchase, or manual verification at delivery [8]. Studies show these measures are often circumvented. In a controlled experiment of 100 underage purchase attempts from Internet alcohol vendors, 45% of orders succeeded outright; most sellers used only weak age verification, and half of the "successful" orders had no age check at all at the point of order [8]. Even when IDs are checked upon delivery, compliance is inconsistent and fails about half the time [8]. Current methods like simple birth-date gating, perfunctory ID checks, or even scanning a driver's license or uploading a government ID do not adequately prevent minors' access to purchasing these products [4].

It is clear that current age verification practices fall short; date-of-birth self-attestation is easily bypassed, and strict ID checks are intrusive and often circumvented—creating demand for more robust and user-friendly solutions.

2.2. Current Deep Learning Methods

There have been significant advances in estimating a person's age from a facial image using deep learning. Levi and Hassner's CNN on the Adience dataset (unconstrained images labeled into age brackets) was an early success, attaining around 50–60% accuracy on exact age-group classification and 84% within one off [3]. Our specific classification thresholds of interest—under 13, 13–17, 18–20, and 21+ —align with U.S. legal boundaries (children vs. teens, minors vs. adults, and 21 for alcohol/tobacco). Some research has directly targeted such categories, achieving over 85% accuracy in differentiating adolescent faces from mature adult faces on the AAF dataset, however, current solutions target more generalized age estimations (for example, adolescent vs. mature adult) rather than classifying ages at crucial boundaries [5].

In general, CNN models are quite adept at separating broad age groups (e.g., identifying a child vs. an adult) because the facial differences at opposite ends of the age spectrum are usually pronounced. The hardest cases are those near the decision boundaries—for example, distinguishing a 17-year-old from an 18-year-old purely by face can be very challenging for humans and deep learning models alike. VGG-16 [7], a CNN architecture with deep layers and small convolutional filters, has demonstrated exceptional feature extraction capabilities that indicate CNNs can learn hierarchical facial representations transferable across multiple face-related tasks. Based on these promising foundations from CNNs, we aim to follow a similar architecture to create a robust feature learning approach in our model, specifically exploring how CNN depth affects feature extraction.

3. Methodology

3.1. Dataset Collection

Facial age estimation initially appears to be a simple classification task. Around six major online datasets (e.g., AAF, Adience, AFAD, IMDB-Wiki, MORPH, UTK-Face) exist for this purpose, each containing images of faces with age labels (and often other metadata like gender and race). However, many of these datasets are incompatible with our specific application due to severe under-representation or omission of key age groups, poor distribution of data (e.g., only Asian faces), or bin labels that are misaligned with our classification bins.

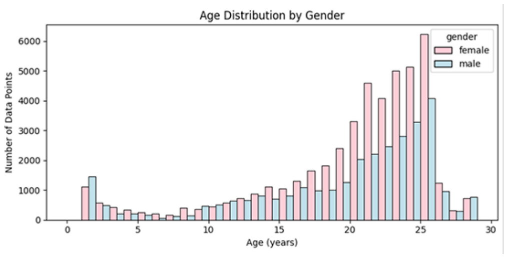

Figure 1. Distribution of gender across ages in UTK-Face and IMDB-Wiki datasets. Note that we are not using the entirety of both datasets, and are only focusing on images in the 0-30 age range.

We have chosen the UTK-Face and IMDB-Wiki datasets [9] [6], filtered on ages 1-25, to test/train our model because they contain a moderate (20k+ images) amount of data that encompasses age and gender distributions, within our ages of interest. A distribution of gender across our data is shown in Figure 1.

To use this data, we manually downloaded both datasets to a shared Google Drive location, randomly distributed our data into train/test/validation folders with a standard 70/20/10 split, and instantiated PyTorch dataloaders to process our data, which consisted of face detection/cropping (a provided feature of IMDB-Wiki dataset), resizing to 64x64, and normalizing [1].

3.2. Experiments

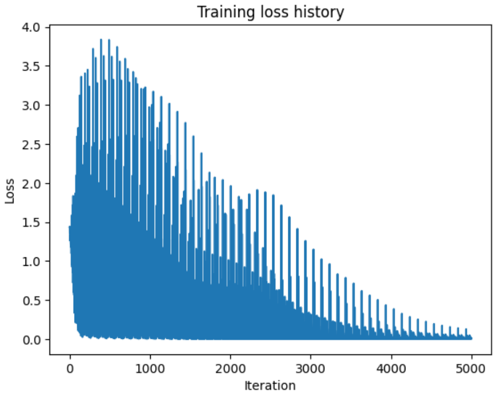

Figure 2. Training loss for simple CNN (Model 0.1) over a subset of 100 images with 50 epochs.

We initially solely relied on UTK-Face for this task. Our preliminary experimentation involved evaluating two simple hand-designed architectures inspired by our previous work in Assignment 4, with one model (Model 0.1) following a CNN structure consisting of [conv-batchnorm-relu-maxpool] x 2 → [affine-relu-dropout] → [affine] → [softmax] and the other model (Model 0.2) following a deeper

CNN structure of [conv-relu-maxpool] x 4 → [affine-relu-dropout] x 2 → [affine] → [softmax]. Before running any in-depth training experimentation, we confirmed the validity of our data-processing and architectures by overfitting Model 0.1 to a subset of 100 images over 50 epochs. Figure 2 demonstrates these results.

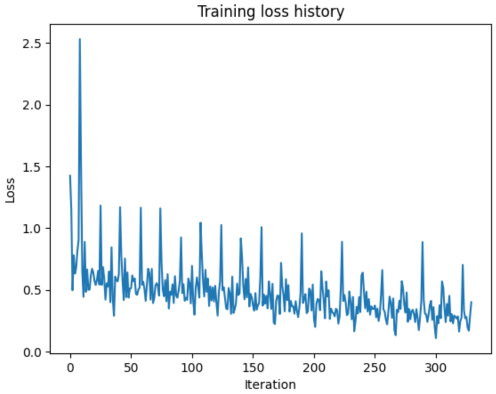

Figure 3. Training loss for Model 0.1 over about 1000 images with 10 epochs.

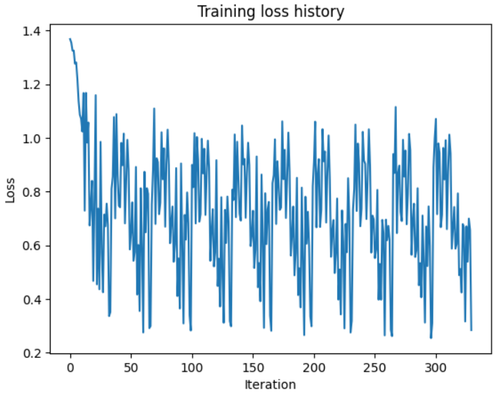

Figure 4. Training loss for Model 0.2 over about 1000 images with 10 epochs.

For both models, we then performed short training runs on a subset of approximately 1000 images using a batch size of 32 over 10 epochs. Training loss curves for each model are shown in Figures 3 and 4. After tuning learning rate and weight decay, we evaluated performance on our validation set of 3615 images from UTK-Face:

- Model 0.1 achieved 88.14% training accuracy and 81.24% validation accuracy.
- Model 0.2 achieved 78.27% training accuracy and 79.42% validation accuracy.

However, upon further inspection of these results, we found that the promising validation accuracies were not due to actual success in facial estimation, but rather highlight a significant limitation in the UTK-Face dataset for this task, as there is a heavy representation skew/bias towards the youngest and oldest age bins, which allowed the model to rely on bins 1 and 4 and still achieve high accuracy, even while refraining from learning any granular facial features for the smaller, more subtle age bins (2 and 3) This imbalance in data is shown in Figure 5.

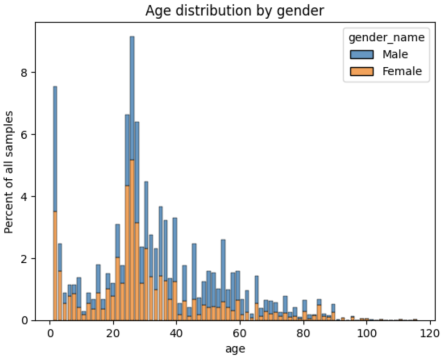

Figure 5. Distribution of gender across ages in UTK-Face dataset.

After this, we revised our dataset to rely more heavily on IMDB-Wiki, as it had a higher representation in our categories of interest (mostly bins 2 and 3), and decided to narrow our usage of the combined dataset to ages 1-25 only to avoid the abundant 0-1 and 25-60 age classification data that prevents models from focusing on the finer feature discrimination task at hand for our specific bins. Similarly, we decided to pivot to using per-bin accuracies as our primary evaluator, since overall accuracy can be artificially inflated by bin imbalances. Per-bin accuracy better reflects a model's ability to distinguish between adjacent age groups—particularly important in our context where legal thresholds (e.g., 17 vs. 18) demand precise classification.

After validating that a standard CNN architecture seems to learn somewhat well for this task even on a poor dataset distribution, we designed three models to train and test on our revised dataset: VGG-16 out of the box to serve as our baseline, a simpler CNN consisting of 3 of VGG-16's CNN layers (Model 1) and a deeper CNN consisting of 9 of VGG-16's CNN layers. The intention here was to see how CNN depth affected both performance and the granularity of the extracted features. A visual of these architectures is shared in Figure 6.

Our goal was to evaluate how varying depths and complexities of a model would impact performance on fine vision tasks. Our models were trained over a period of 10 epochs[*], using Adam optimizer with a learning rate

[*] Note that only Model 1 was trained on 30 epochs, but after observing

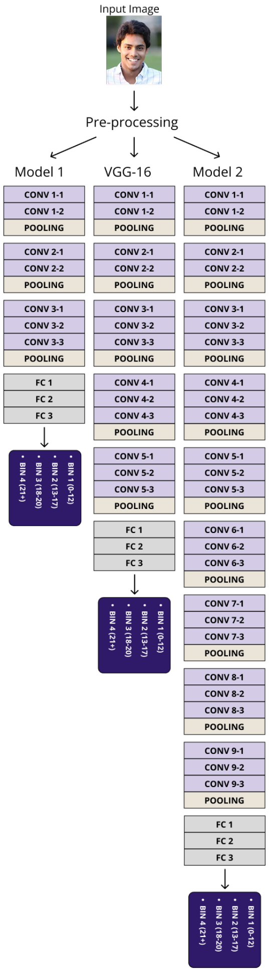

Figure 6. Model Architectures

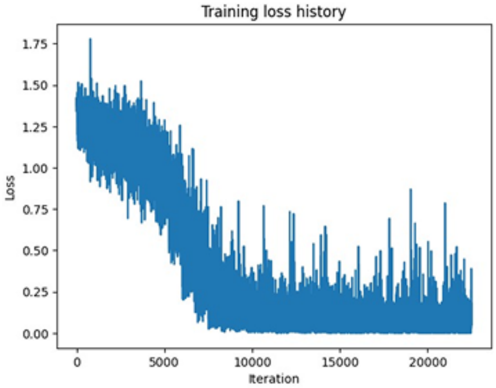

Figure 7. Training loss for Model 1

of 1e-4 (after tuning), a standard batch size of 32, and softmax/cross-entropy loss. Architecture specific decisions

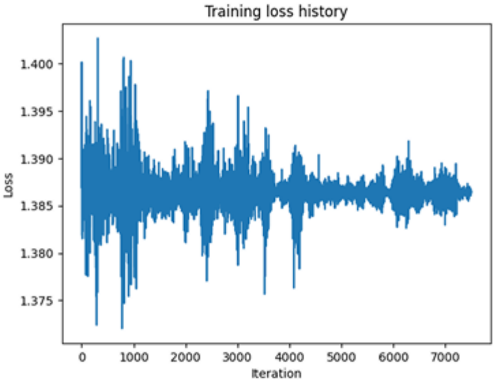

Figure 8. Training loss for Model 2

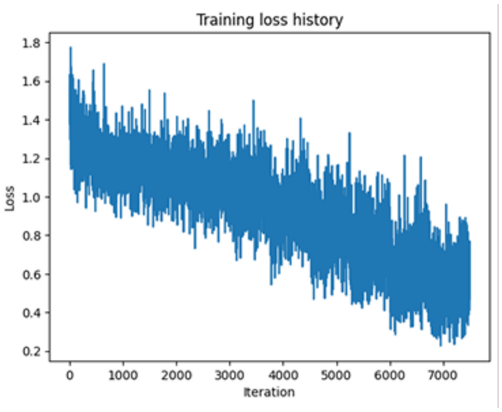

Figure 9. Training loss for VGG-16

were to include a ReLU nonlinearity after each convolution layer, use Maxpool for pooling, and consistently utilize 3x3 kernel sizes for convolutions and 2x2 kernel sizes for pooling. We also refrained from using dropout or adding additional regularization techniques as the existing mislabel noise from the IMDB-Wiki dataset caused worry that the models would not be able to sufficiently learn.

A dataset of 24k images within a 1-25 age range was used, balanced such that each bin contained 6k images (countering any representation skew issues from previous attempts). Figures 7, 8, and 9 show the loss history throughout training.

4. Results

Despite balancing the dataset, none of the models generalized well. This suggests that dataset balance alone does not solve the underlying challenges of fine-grained age estimation.

We found that all three models achieved poor test accuracy (<40% overall), with performance varying dramatically across age bins. This contrasts with our very high training accuracy (90%+), indicating severe overfitting.

a training loss plateau at about 10 epochs, we chose to train remaining two models for 10 epochs only.

Model / Bin	0-12	13-17	18-20	21+	Overall
Model 1	23.51	50.15	26.71	36.62	35.54
VGG-16	29.73	43.05	24.46	41.31	37.80
Model 2	0.00	100.00	0.00	0.00	15.66

Table 1. Accuracy (as a percent) for each model, segmented by age bins.

We can also observe inconsistent performance across age groups; the per-bin accuracy shows uneven performance across the different age ranges. Notably, the 13-17 age group shows the best performance across all models, while the other age bins perform much worse. Most concerning is Model 2, showing 100% accuracy for the 13-17 age bin and 0% for all of the others.

5. Discussion

5.1. Models

As shown in 7 and 9, our models seemed to perform quite well during training and eventually plateaued at low loss values (around 0.2). However, the low test accuracy is indicative of model overfitting and memorization of training data, implying that our models are not learning facial age classification features. While these models' accuracies remained lower than expected, they yielded their highest accuracies for predictions within the 13-17 and 21+ age bins. This makes sense; intuitively, these age ranges have the most distinct feature set and are likely easiest, even for humans, to predict. Model 2 seems to have collapsed entirely with vanishing gradients and simply predicts the 13-17 age bin on every input. We believe Model 2 is exhibiting this behavior due to a complex architecture that does not have residual connections between layers to allow for gradients to flow back.

5.2. Feature Extraction

Returning to feature extraction and interpretation, our Grad-CAM analysis exposes that all three models are not learning facial age features. Model 1 reveals random, scattered activation patterns suggesting spurious correlation learning. VGG-16 had focused on the clothing and background of the image, rather than facial characteristics. Model 2 shows that the activation maps were uniform across the entire input image, rather than highlighting specific features. This suggests that this model suffers from gradient saturation and learning failure.

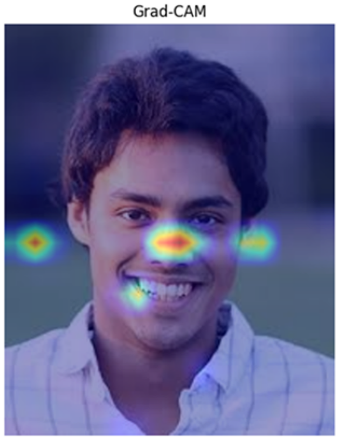

Figure 10. Grad-CAM analysis on Model 1

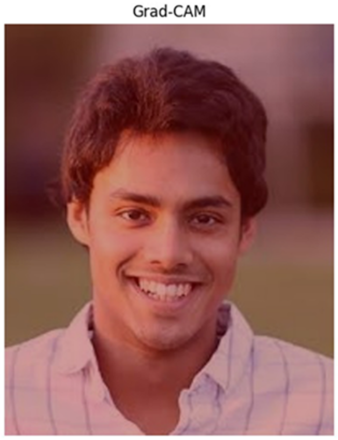

Figure 11. Grad-CAM analysis on Model 2

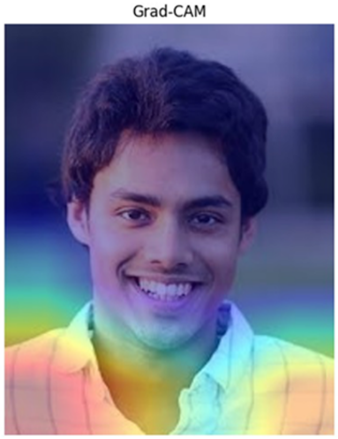

Figure 12. Grad-CAM analysis on VGG-16

5.3. Limitations

Based on our analysis, one prominent challenge in facial age estimation within the critical 9-25 age range stems from dataset quality. Current standard datasets reveal significant under-representation within the relevant 9-25 age range, noisy mislabeling, and distribution bias towards older adults. This results in larger biases for age groups that were not relevant to our project, making it difficult to provide our models with abundant high-quality data within our target demographic. We believe these limitations contributed towards the overfitting we saw in our models. Additionally, the fact that performance peaks on broader bins (13–17 and 21+) further highlights how challenging fine-grained distinctions are; after all, even humans struggle to tell a 17-year-old from an 18-year-old or a 20-year-old from a 21-year-old.

Adversarial robustness is also an emerging concern. Like other deep networks, face-age classifiers can be vulnerable to adversarial manipulation. Even a subtle, imperceptible perturbation to an image can dramatically change the model's output without changing the face's true appearance. A determined user could potentially evade automated age checks by digitally altering their selfie in a way that a human eye would not detect, but an algorithm would misread. Physical-world attacks are also plausible – for example, specialized makeup or accessories might confuse the age estimator.

5.4. Ethics Statement

Facial age classification entails important considerations around privacy, consent, fairness, and potential misuse. In this work, all images are drawn from publicly available datasets with explicit research-use permissions.

A key concern in deploying deep age classifiers is dataset representativeness and bias. Face datasets used for training may over- or under-represent certain populations or age ranges. If a model is trained mostly on, say, adult Caucasian faces, its accuracy may degrade on younger faces or people of other ethnicities. While we intended to ensure our training data adequately represents differing demographics in order to minimize bias and deliver consistent, equitable age-classification performance across all age, race, and gender groups, the sparse existence of reliable age-classification data within our categories did not guarantee such even distributions.

We stress that this system is intended only as an assistive tool—not a definitive arbiter of age—and should always be used in conjunction with human review. Finally, we explicitly discourage any deployment for unauthorized surveillance, profiling, or decision-making that could adversely affect individuals without their knowledge or consent.

6. Future Work

One of the main takeaways from this experiment is the lack of a high-quality, application-accurate dataset for facial age estimation, especially within the 9-25 age range. Given that our analysis revealed that current standard datasets suffer from this under-representation of this demographic, we believe that curating a larger and more representative dataset would be a good approach to solving this issue, but would have to navigate the legal and ethical concerns with collecting such data for our target age group of ages 9-25.

In addition, future applications of age-classification models could explore the impact of feature-altering procedures, such as lighting, clothing, Botox, or makeup. This analysis will help us to understand in more detail which features contribute to mis-classifications in age.

Lastly, exploring attention mechanisms within our model architecture may allow for better focus and identification of age-critical regions of the face.

References

[1] Link to our google drive can be found here. 2

[2] Danah Boyd, Eszter Hargittai, Jason Schultz, and John Palfrey. Why parents help their children lie to facebook about age: Unintended consequences of the children's online privacy protection act. *First Monday*, 16(11), Oct. 2011. 1

[3] Gil Levi and Tal Hassncer. Age and gender classification using convolutional neural networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 34–42, 2015. 2

[4] Julia Martinez, Patricia Rutledge, and Kenneth Sher. Fake id ownership and heavy drinking in underage college students: Prospective findings. *Psychology of addictive behaviors : journal of the Society of Psychologists in Addictive Behaviors*, 21:226–32, 06 2007. 2

[5] Muhammad Mustapha, Nur Mohamad, Ghazali Osman, and Siti Ab Hamid. Age group classification using convolutional neural network (cnn). *Journal of Physics: Conference Series*, 2084:012028, 11 2021. 2

[6] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2-4):144–157, 2018. 2

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[8] Rebecca S. Williams and Kurt M. Ribisl. Internet alcohol sales to minors. *Archives of Pediatrics Adolescent Medicine*, 166(9):808–813, 09 2012. 1

[9] Song Yang Zhang, Zhifei and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. 2