Age and Gender Classification

183.587 Computer Vision System Programming WS16

Christoph Körner (0726266), christoph.koerner@tuwien.ac.at
Patrick Wahrmann (1327120), patrick.wahrmann@student.tuwien.ac.at

Problem Definition

Our goal in this lecture is to develop a classification system for the task of age (in classes of around 10 years each) and gender estimation on human facial images. For this purpose we are going to use Deep Learning (Keras). The final input images are going to be cropped and registered face images from *"The Profiler"*, a project of the CVL for which we deliver an estimation of age and gender of the depicted person.

Dataset

For training and testing we use the IMDB-WIKI dataset of Rothe et al. [1]. This dataset consists of two parts (automatically crawled from IMDb and Wikipedia) and stands out because of its very big size (both parts together: 523,051 faces). The following Figure 1 shows 4 sample images per dataset. In addition, quadratic crops of the faces are also available (IMDb: 7GB, Wiki: 1GB).

Figure 1: Sample images from the IMDb/Wiki dataset [1]

The age labels can be computed from the meta attributes *date of birth* of the actor/person and *photo taken*, the year when the picture was taken; the gender label is available as enum.

A) Data Preprocessing

The dataset was generated from an automatic crawler and hence contains corrupted (1x1 pixel) and faulty images as well as images without faces (or a single face). The following Figures show some examples of these faulty images.

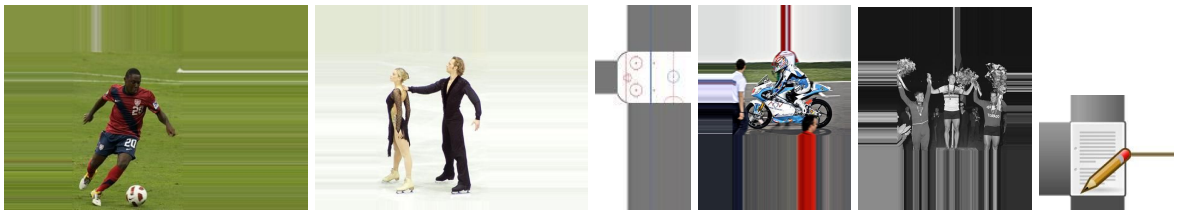

Figure 2: Faulty images of the Wiki dataset

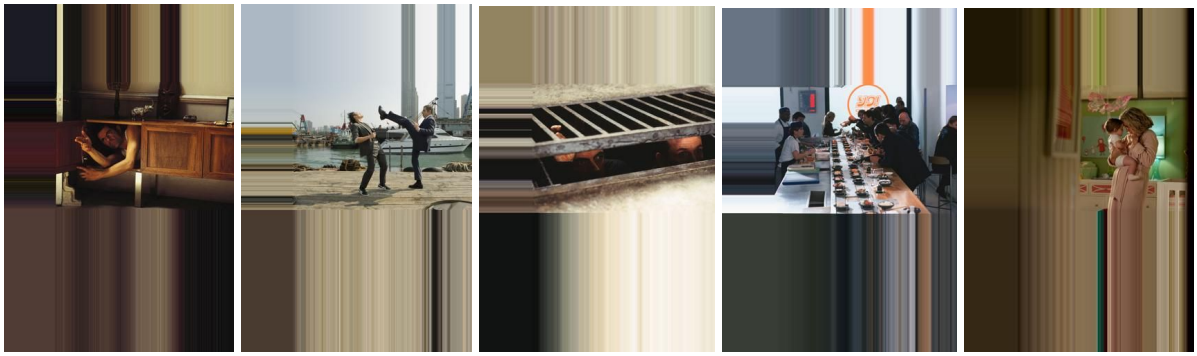

Figure 3: Faulty images of the IMDb dataset

We observe, that mainly non-square images are very likely to not contain a face. Hence, we filter out all non-square images. We as well remove corrupted images and images with faulty age labels (< 0 and > 100). The following Table 1 shows the number of samples removed during the cleaning process. Although the rigorous data cleaning process removes almost half of all training samples, we obtain a well-sized dataset for the classification task.

	WIKI	IMDb
total	62328	460723
image dimensions 1x1	8070	0
image non-square	24124	163452
age > 100	44	107
age < 1	57	244
total after cleaning	**30033**	**296920**

Table 1: Number of samples in the data cleaning process

The following Figure 4 shows the distributions of the age and gender values in the combined dataset after the data cleaning process.

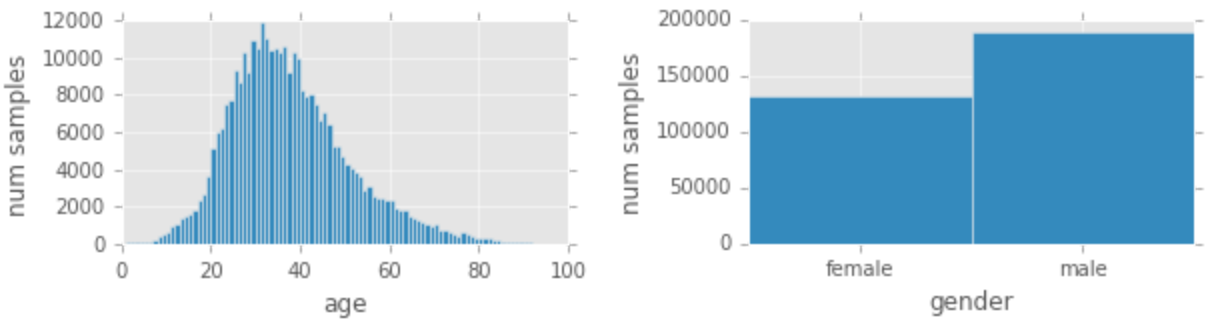

Figure 4: Age and gender distributions from the combined and cleaned dataset

We combine the age values into ordinal similar-sized age classes such that all classes contain approximately the same number of samples. The following figure shows an aggregation of the age label to the classes (0,15), (16,20), (21,25), (26,30), (31,35), (36,40), (41,45), (46,50), (51,55) and (56,100) for the combined and cleaned IMDb/Wiki dataset. However, we observe that the classes for teenagers will always contain less samples than the other classes.

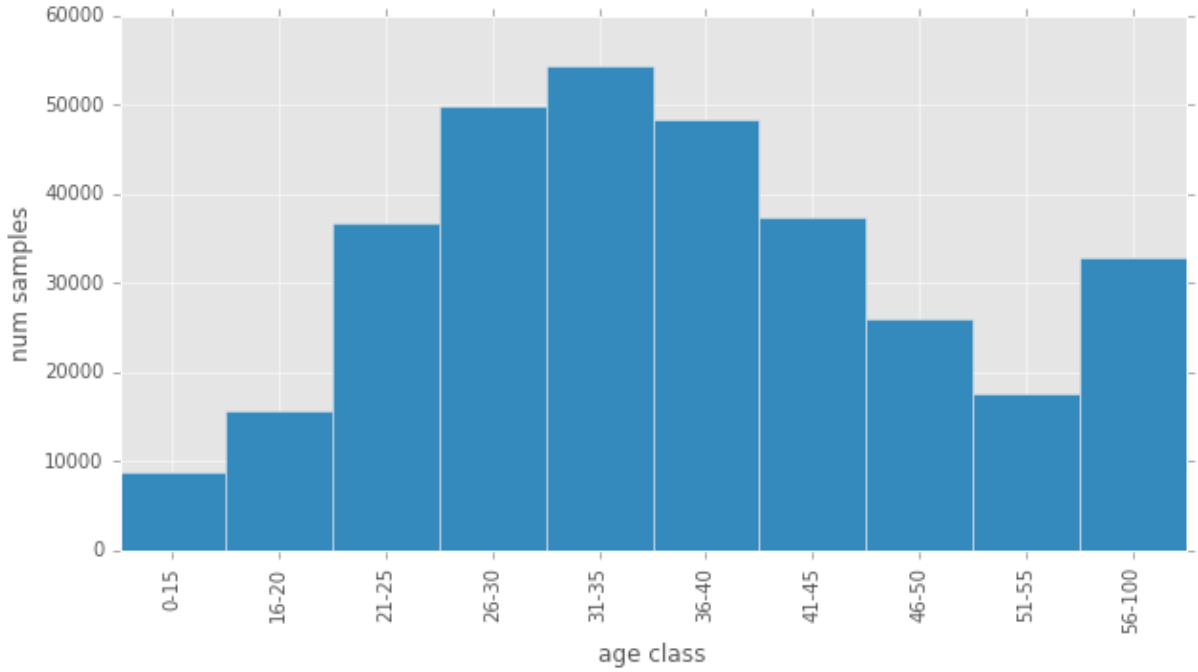

Figure 5: Age class distribution for the combined and cleaned dataset

B) Dataset Splits

Both datasets are cleaned and combined to a single face dataset. Then, the dataset is shuffled and randomly split it into train (80% of the dataset), validation (10% of the training set) and test data (20% of the dataset).

The random seed for the split has been fine-tuned, such that the distributions of the different sets look similar. The following Figures 6-8 shows the age class distributions of each split.

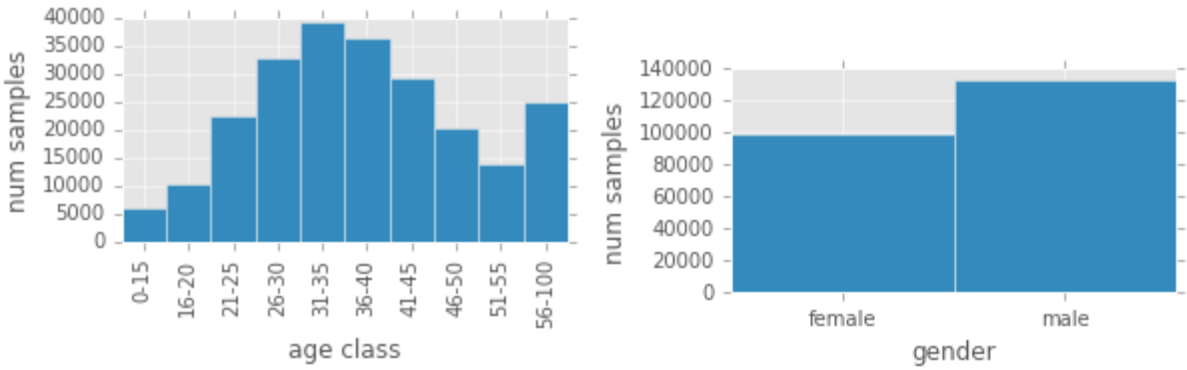

Figure 6: Age class and gender distribution of the training split

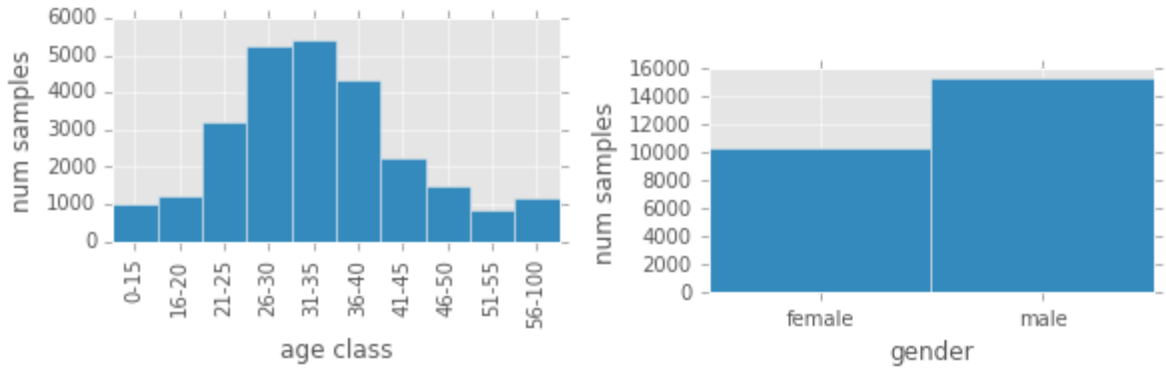

Figure 7: Age class and gender distribution of the validation split

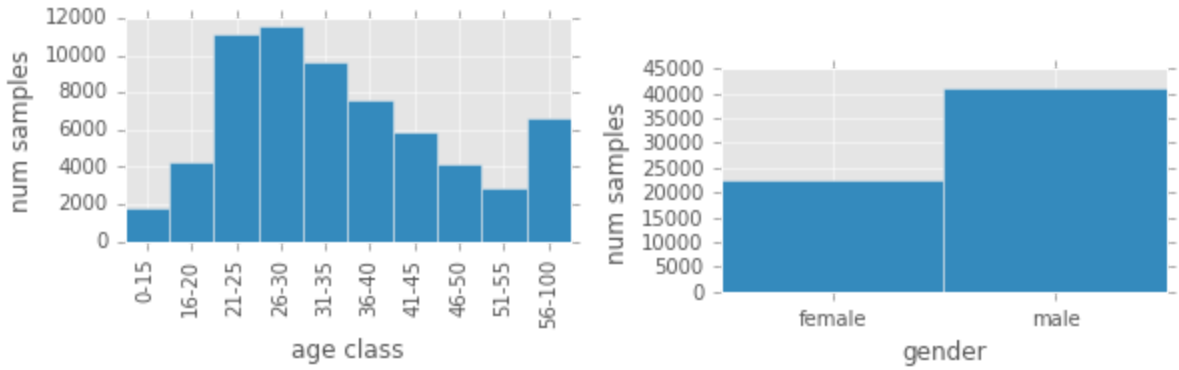

Figure 8: Age class and gender distribution of the testing split

C) Loss Functions

For computing the loss of the gender estimate we will use the standard cross-entropy loss function which is commonly used for binary classification. For computing the loss of the age estimate we will at first also use cross-entropy loss.
If there is still time at the end of the project, we will try a different loss function that also takes the ordinal property of the classes into account: hence, we will use Mean Squared Error (MSE) or Cohen's Weighted Kappa [2].

D) Deep Neural Network Model

Initially, we wanted to use the GoogLeNet [3] architecture to train both the age and gender classification models. According to [5] this model achieves the highest image classification accuracy (top 1 accuracy on ImageNet) with the lowest number of parameters. However, due to the image dimensions we had to choose a different architecture.

Most of the images in the dataset are slightly bigger or slightly smaller than 100x100 pixels and hence we cannot use the "standard ImageNet" dimensions 3x224x224 which is also used in GoogLeNet. In addition, the data set would result in a size of 140GB (using a float 32 data type) when scaling all images to 224x224 pixel. Therefore, we choose the input dimensions 3x112x112 for our deep learning model.
Changing the spatial dimensions of the input data usually requires to change the network architecture. In GoogLeNet for example, we would have to reduce the spatial dimensions of the filter pool5/7x7_s1 to 3x3. However, we don't know how this would affect the model

training and convergence and we decide to choose the VGG [4] architecture. Concerning model complexity (see Figure in the Appendix), VGG seems to be easier to understand, and hence for us to train.
In some cases (especially when the model flattens the layers before the first dense layer) this also results in a huge reduction of parameters. Reducing the RGB channels to a single grayscale channel however does not reduce the number of parameters noticeably. The following Table 2 summarizes the numbers of parameters for each model.

Hence, for training both age and gender classification models we will use a VGG-16 architecture with input size 3x112x112.

Model	Input Dimensions	No. Parameters	Needs Adaption
VGG-16	3x224x224	138,357,544	-
VGG-16	3x112x112	69,151,528	no
VGG-16	1x112x112	69,150,376	no
GoogLeNet	3x224x224	6,998,552	-
GoogLeNet	3x112x112	6,998,552	pool5/7x7_s1 → pool5/3x3_s1
GoogLeNet	1x112x112	6,992,280	pool5/7x7_s1 → pool5/3x3_s1

Table 2: Image Dimensions and number of parameters in deep learning models

Evaluation
The age and gender estimation algorithms will both be evaluated on the test set with reported accuracy, precision, recall and F1-score values. These scores will be compared to state-of-the-art age and gender estimation algorithms and to human classification accuracy (if available).
We will as well evaluate the performance of our algorithm to the pretrained model provided in [1], a state-of-the art deep neural network trained on the same IMDb/Wiki dataset for both age and gender classification.

References
[1] https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/
[2] B. Arie, “Comparison of classification accuracy using Cohen’s Weighted Kappa”, *Expert Systems with Applications*, vol. 34/2, pp. 825-832, 2008.
[3] C. Szegedy et al., “Going Deeper with Convolutions”, CoRR, 2014.
[4] K. Simonyan and A.Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition”, CoRR, 2014.
[5] A. Canziani, A. Paszke and E. Culurciello, “An Analysis of Deep Neural Network Models for Practical Applications”, CoRR, 2016.

Appendix

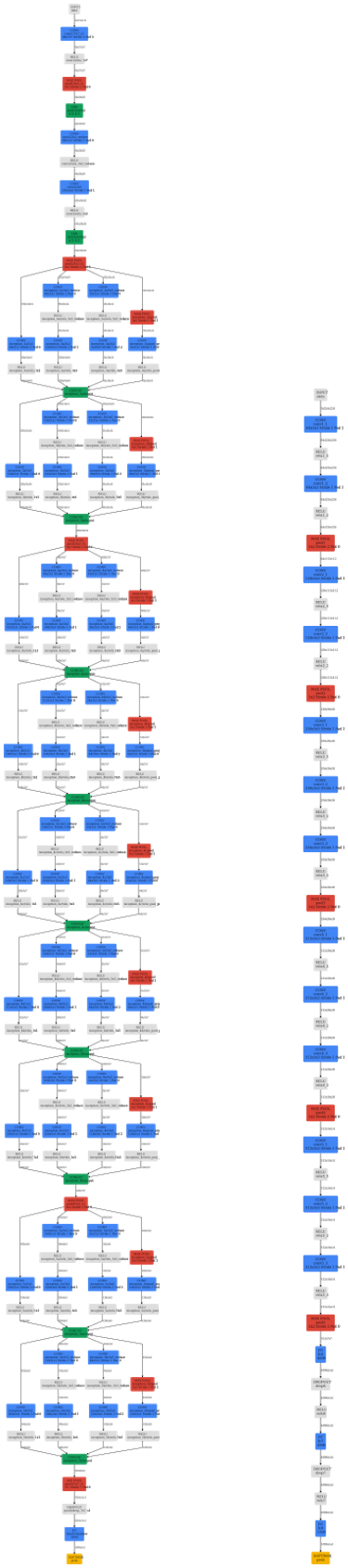

Figure 9: GoogLeNet (left) vs. VGG-16 (right)