Report: gender classification

Outline

1. The task
2. Exploratory data analysis
3. Choice of model
4. Performance of model

The task

The task is to develop a classifier model from scratch which takes images as input and returns as output a gender label for each image. I have addressed this by (1) exploring the data to inform my choices of how to build the classifier, (2) training a model, and (3) evaluating its performance, which are described below.

Exploratory data analysis

The training and validation data consist of images of the same size (height, width and channels), all cropped to show the face, mostly facing the front, and roughly upright. This homogeneity helps the training and means that minimal pre-processing of the images is required. They are all labelled as either male or female, making it a binary classification task, with some class imbalance of 58% female images in the training set and 57% female images in the validation set.

Choice of model

Convolutional neural networks (CNNs) have shown great success for many image classification tasks, including with complex natural stimuli like faces. Given time and computational constraints, I chose a simple end-to-end CNN of two convolutional layers and two fully connected layers. For scaling up, it is likely that a deeper network with 10 or more layers would give much better performance. Other complex network architectures like transformers or separate models for feature extraction and classification may also perform well, but are infeasible given the constraints.

I judged that the model choices had to satisfy these requirements:

- Historically good performance for binary image classification tasks
- Complex enough to achieve good performance on the training set
- Simple enough to run with low computational cost
- Trainable in a few minutes or hours
- Likely to generalise well to the validation and test sets

Network architecture

Several features of the network architecture make it well suited to this task:

- The convolutional layers learn translation-invariant features in the images, because weight parameters are shared over space. This is necessary because useful features of the images (for example, curvature of the jaw) may not always appear in the same pixels. The relatively small kernel sizes of 3 are suited to these small images with small features, and also reduce the computational cost compared to larger kernels.
- The subsequent fully connected layers learn to combine these features in a way that enables classification.
- The model also contains two dropout layers which randomly remove half and one quarter (respectively) of the weights before a fully connected layer, following each minibatch of training; this is a form of regularisation which prevents overfitting and avoids convergence to local minima during training.
- ReLU activation functions after the first three layers, which not only introduce non-linearity to the network, but also reduce the number of units activated at each pass to make the model more computationally efficient than alternative activation functions. It is also often observed to produce faster and more reliable convergence during learning.
- Finally, a max-pool layer following the convolutional layers which gives additional translational invariance and helps to extract the sharpest features such as edges, with less computational cost compared to an additional convolutional layer.

Other model considerations

Aside from the network architecture, some other choices were made to suit the task:

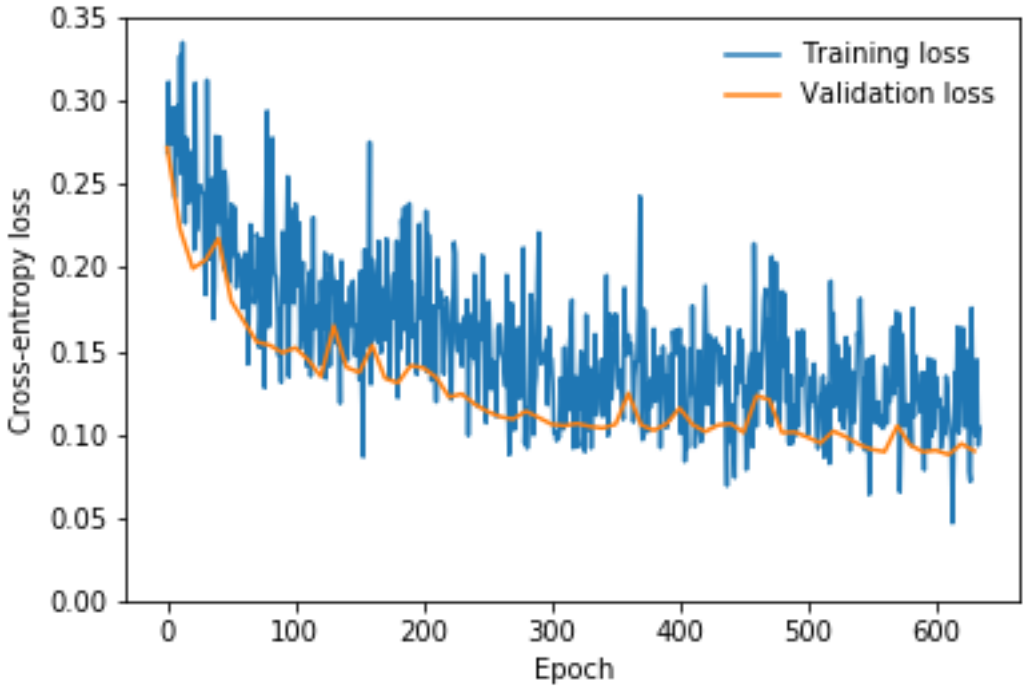
- Weighted cross-entropy as the loss function. Binary cross-entropy is a common choice for binary classifiers because it minimises the dissimilarity between training data and model distributions. Because of the slight class imbalance in the training set, the cross-entropy is weighted slightly higher for male images to ensure good performance on both classes.
- Adam as the optimiser, which adapts the learning rates of each parameter according to its gradient. This ensures large updates at the beginning of learning, for faster convergence, and smaller updates as it nears the optimum, for better convergence to the optimum.
- Data augmentation. The training set offers tens of thousands of examples per class, but to improve learning and prevent overfitting the data can be augmented further by applying transformations to the existing training set. For each batch of training, several transformations are randomly applied to each image in the batch: horizontal flips, rotations and perspective changes, and random erasing of small patches of the image. These transformations are small because the training and validation sets are fairly homogeneous: for example, a rotation of 90º would be outside the expected test distribution and therefore unhelpful, so rotation is limited to 5º.

Performance of model

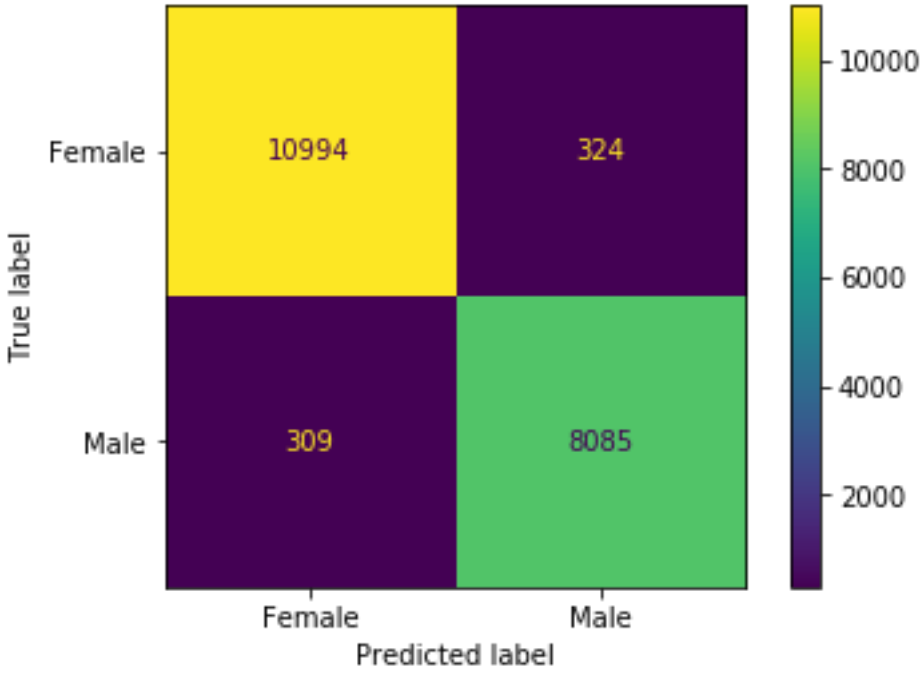
Model performance

The cross-entropy loss over 635 training batches showed a gradual reduction, for both the training loss (calculated on one batch) and the validation loss (calculated on the full

validation set). This indicates that the model converged well. The decrease in the training loss indicates no substantial underfitting, and the decrease in the validation loss indicates no substantial overfitting.



The overall accuracy of the model predictions on the validation set can be assessed using the F1 score, which combines precision and accuracy, weighted according to the class sizes because of the mild class imbalance in the validation set. The weighted F1 score is 0.9679, indicating that the model correctly predicts gender from an image in 97% of cases.



Further improvements

These metrics indicate that the model is good, although there is scope to increase the accuracy still further towards 100%. Tuning hyperparameters such as learning rate, batch size and optimiser could improve performance; changing the data augmentation procedure could also improve performance; and developing a more complex model with additional layers or larger kernels could allow more features to be learned.

Improving the model could also benefit from insight into how the model works and where its areas of improvement are. For example, the validation set contains a number of attributes such as whether the image is of someone wearing eyeglasses or a hat, whether they have a moustache or a beard; calculating the F1 score for each of these classes would indicate which features the model learns and relies on and which features cause lower performance. This approach would have a further benefit of flagging potential biases which might have detrimental ethical or practical consequences in production, such as poorer performance on darker skin. A second source of interpretability could be saliency maps, which can be built to indicate which pixels contribute the most to the classification of the image, such as pixels corresponding to the mouth.