Abstract

In this project, competition-winning deep neural network architectures with pretrained weights are used for image-based gender recognition and age estimation. The effects of various design schemes and training techniques are observed in order to improve prediction accuracy.

What are Deep CNNs?

A convolutional neural network (CNN) is a type of AI used for analyzing images. Convolving images allows patterns to be extracted from them that are then interpreted by the neural network in order to make classification decisions. Deep CNNs use many layers of convolutional filters, pooling, and fully-connected neurons.

Transfer Learning

Because of the vastness and complexity of deep neural network (DNN) architecture, designing and testing models is expensive and time-consuming. In transfer learning, the weights and convolutional filters that are proficient at one task, can be re-used for a different task requiring only a small amount of retraining. This involves using a network architecture with pretrained weights, modifying it slightly, and then retraining part or all of the model to output predictions for the new task. For this study, the VGG19 network was used along with its ImageNet weights [3]. The network was modified to work with the MORPH-II dataset as shown in the image below.

The Dataset

The dataset used during this project was MORPH-II. It consists of 55,134 mugshot images with subject ages ranging from 16 to 77 years old. 84.6% of the dataset is male, and 77.22% of the dataset is black. The dataset contains all of the remaining images. Set 1 and 2 both contain 10,280 images, and set 3 contains 34,344 images.

Training Parameters

To compare the effects of changes in transfer learning techniques, all training parameters are kept consistent unless otherwise specified. All input is standardized before being fed into the network. Gender models use binary cross entropy loss, and age models use mean absolute error (MAE) loss. During training, models are supplied with a validation set of 500 images randomly selected from S2. Results for age estimation are reported as an MAE. MAE = $\sum_{i} |a_{i} - \hat{a}_{i}|$. For gender, the results are reported as an accuracy, the number of correct predictions over the size of the test set.

- **Training Set:** MORPH-II S1
- **Test Set:** MORPH-II S1 S2 S3
- **Input Size:** 200 x 240 x 3
- **Batch Size:** 50
- **Epochs:** 60
- **Dropout Ratio:** 0.5
- **Activation:** ReLU
- **Optimizer:** Adam

Input Standardization

To perform input standardization on an image dataset, the formula:

\[
\text{input} = \text{input} - \mu_n \times \sigma_n
\]

is used, where \(\mu_n\) is the mean of all pixel values (red, green, and blue) in \(\text{input}\) which is the set of all pixel values (red, green, and blue) in \(\text{input}\). \(\sigma_n\) is the standard deviation of the age probabilities. The network can then be trained using LDAE labels which produces consistently better results than one-hot vector encoding.

Epochs

An epoch is one pass through all of the data in the training set. Depending on the dataset, depth of the network, regularization techniques, and a variety of other factors, an optimal number of epochs might be low or high. Too few epochs and the network will be underlearned. Too many epochs and the model becomes overfit. In both of these instances, validation loss will be higher than normal, and it is unlikely that a near-optimal model will be produced. The table below shows how validation accuracy is affected during training using varying numbers of epochs.

<table>
<thead>
<tr>
<th>Epochs</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
</tr>
</thead>
<tbody>
<tr>
<td>AvgMax</td>
<td>5.175</td>
<td>5.286</td>
<td>4.908</td>
<td>4.006</td>
<td>4.913</td>
</tr>
<tr>
<td>Expected Value</td>
<td>5.234</td>
<td>5.238</td>
<td>4.915</td>
<td>4.843</td>
<td>4.919</td>
</tr>
</tbody>
</table>

Data Augmentation

Data augmentation is a general term for increasing the size of a dataset. 12-crop resampling was used to augment data. If more labeled data can be obtained, neural network accuracy can increase by quite a bit. By training on \(S_1\) and \(S_2\), the size of the training set was doubled and an MAE of 4.690 was obtained on \(S_3\).

Dropout Regularization

One method of combatting overfitting is to add dropout to weight layers. In the 2014 paper that introduces dropout, Srivastava et al. state that it “provides a way of approximately combining exponentially many different neural network architectures efficiently” [4]. When dropout is added to a weight layer, neurons are randomly selected to be removed from the network at each iteration.

Label Distribution Age Encoding (LDAE)

LDAE uses the formula:

\[
\frac{1}{2\pi\sigma}\exp\left(-\frac{1}{2\sigma^2}(i - \mu)^2\right)
\]

\[\mu = \text{age}
\]

\[\sigma = \text{standard deviation}
\]

\(i\) is the age at which a probability should be produced, and \(\sigma\) is a hyperparameter that affects the spread of the age probabilities. The network can then be trained using LDAE labels which produces consistently better results than one-hot vector encoding.

Acknowledgements

This research was conducted under the careful guidance of Dr. Cuixian Chen. The project was produced with funding from the National Science Foundation under grant DMR1365528. The contents of this paper do not necessarily reflect or coincide with the views of the NSF.

References