DEX: Deep EXpectation of apparent age from a single image

September 15, 2015

1 Team details

- Team name: CVL_ETHZ
- Team leader name: Rasmus Rothe
- Team leader address, phone number and email:
  Office ETF D115, Sternwartstrasse 7, 8092 Zurich, Switzerland
  +41 (0) 78 633 02 22
  rrothe@vision.ee.ethz.ch
- Rest of the team members: Radu Timofte, Luc Van Gool
- Team website URL (if any): `www.vision.ee.ethz.ch`
- Affiliation: Computer Vision Laboratory, D-ITET, ETH Zurich

2 Contribution details

The propose method Deep EXpectation (DEX) of apparent age from a single image first detects the face in the test image (face detector by Mathias *et al.* [1]) and then extracts the CNN predictions from an ensemble of 20 networks on the cropped face. Each network (VGG-16 architecture [2]) was pre-trained on Imagenet and then fine-tuned on face images from IMDB and Wikipedia.

The resulting network was then again fine-tuned on the actual dataset from the challenge. The networks were trained for classification with 101 output neurons, each neuron corresponding to an integer age (0-100). The final prediction is the expected value of the softmax-normalized output of the last layer, averaged over the 20 networks.

3 Face Detection Stage

For both training and testing images, we run the off-the-shelf face detector by Mathias *et al.* [1] to detect the location of the face.

In order to align the faces we run the face detector not only on the original image but also on all rotated versions between $-60°$ and $60°$ in $5°$ steps. As a few of the training images were upside down or rotated by $90°$, we also run the detector at $-90°$, $90°$, and $180°$. We take the face with the strongest detection score and rotate it accordingly to a up-frontal position.

For very few images ($< 0.2\%$) the face detector is not able to find a face. In those cases we just take the entire image. On the final test set this applies only to 1 image.

We then extend the face size and take 40% of its width to the left and right and 40% of its height above and below. Adding this context helps the prediction accuracy. If the face already covers most of the image, we just pad with the last pixel at the border. This ensures that the face is always at the same location of the image.

The resulting image is then squeezed to 256×256 pixels and used as an input to a deep convolutional network.

4 Face Apparent Age Estimation Stage

4.1 Prediction

The apparent age prediction is obtained by applying a deep convolutional neural network to the detected face from the previous processing stage. Our method uses the VGG-16 architecture [2] which has shown impressive results on the Imagenet challenge.

The output layer was adapted to 101 output neurons followed by a softmax normalization. Each output neuron corresponds to one integer age (0 to 100). The prediction is obtained by taking the expected value over the softmax output probabilities. During experiments we noticed that this works much better than a) training a regression, b) learning a regression (i.e. SVR) on top of the CNN features of the previous layer, or c) just taking the age of the neuron with the highest probability.

The final prediction is the average of an ensemble of 20 networks trained on slightly different splits of the data.

4.2 Training

Instead of training the CNN from scratch, we start with the pre-trained weights from Imagenet. As the provided training set is relatively small (3612 images in the training and validation set combined) we first fine-tune the network on images from Wikipedia and IMDB. For Wikipedia we crawled all profile images from pages of people and only used the ones for training where we have the date of birth and the year when the photo was taken in the caption. This resulted in

62,359 images. For IMDB we crawled all images for the 100,000 most popular celebrities. For training we used the images where we have both the date of birth in the IMDB profile as well as the year when it was taken in the caption of the image. This results in 461871 images.

As some of the images (especially from IMDB) contain several people we only use the photos where the second strongest face detection is below a threshold. For the network to be equally discriminative for all ages, we equalize the age distribution, i.e. we randomly ignore some of the images of the most frequent ages. This leaves us with 260282 training images.

After this fine-tuning stage, we then fine-tune the resulting network on 20 different splits of the ChaLearn dataset. In each split we use 90% of the images for training and 10% for validation. The splits are chosen randomly for each age separately, i.e. the age distribution in the training is always the same. We then train the 20 networks on an augmented version of the ChaLearn dataset, as we add 10 augmented versions of each image. Each augmentation randomly rotates the image by -10° to 10°, translates it by -10% to 10% of the size and scales it by 0.9 to 1.1 of the original size. Note that we do the augmentation after splitting the data into the training and validation set to ensure that there is no overlap between the two sets. Each network is then trained and we pick the weights with the best performance on the validation set.

5 Global Method Description

We use the pre-trained network on Imagenet from Oxford [1].

On top of the data that has been provided by ChaLearn for training and validation data, we crawled images from IMDB and Wikipedia. See Sec. 4.2 for details. The images are stored in the provided Google Drive folder (imdb.tar and wiki.tar).

Qualitative results showed that our proposed solution is able to predict apparent age of faces in the wild – to an extend that for a human it his impossible to judge if the prediction is correct. This is partly enabled by learning from a large dataset crawled from Wikipedia and IMDB depicting faces in the wild. Taking the expectation of the softmax-normalized output neurons robustifies the prediction, minimizing the number of outliers.

The proposed solution has not been published anywhere else. Training the network for classification even though it is a regression problem and then taking the expected value over the softmax-normalized output can be considered as novel. We are also the first to crawl more than 500,000 images with known age and show that it improves the age prediction.

We consider to submit a paper to the ChaLearn workshop describing our approach [3] in depth.

[1] `http://www.robots.ox.ac.uk/~vgg/software/very_deep/caffe/VGG_ILSVRC_16_layers.caffemodel`

6 Other details

The pipeline is written in Matlab. The CNNs are trained on Nvidia Tesla K40C GPUs using the Caffe framework. The face detection was parallized over a Sun Grid Engine which was essential for the IMDB and Wikipedia images.

The proposed method was developed over about 8 weeks with an involvement of 40% of the main author's working time. Developing and implementing took roughly 6 weeks and the last 2 weeks were used for validation and fine-tuning.

Training the network on the IMDB and Wikipedia image took around 5 days. Fine-tuning a single network on the ChaLearn dataset takes about 3h. For testing the face detection at each rotation takes around 1s. The feature extraction per image and network takes 200ms.

We enjoyed taking part in the challenge and consider to participate again. It would be great if the size of the dataset is increased next time to reduce the advantage of training on external data. Thanks for organizing!

References

[1] Markus Mathias, Rodrigo Benenson, Marco Pedersoli, and Luc Van Gool, "Face detection without bells and whistles," in *ECCV*. 2014.

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

[3] Rasmus Rothe, Radu Timofte, and Luc Van Gool, "Dex: Deep expectation of apparent age from a single image," in *ICCV Workshops*. 2015.