Early Adaptation of Deep Priors in Age Prediction from Face Images

Mahdi Hajibabaei
Computer Vision Lab
D-ITET, ETH Zurich
hmahdi@student.ethz.ch

Anna Volokitin
Computer Vision Lab
D-ITET, ETH Zurich
voanna@vision.ee.ethz.ch

Radu Timofte
CVL, D-ITET, ETH Zurich
Merantix GmbH
timofter@vision.ee.ethz.ch

Abstract

Age prediction from face images is a challenging task. Direct application of pre-trained models on new data leads to poor performance due to data and distribution mismatch and lack of newly annotated material.

In this work, we analyze the transfer of knowledge from deep models pre-trained on massive datasets to new target datasets with (very) little information available. We investigate (i) pre-training on massive datasets with an imposed target age label distribution, (ii) pre-training on massive face datasets but without age annotations, and (iii) fine-tuning on the target train data.

The experimental benchmark uses the massive IMDB-Wiki, VGG-Face and ImageNet datasets as sources and ChaLearn LAP and MORPH 2 as target datasets. The deep architectures/priors are based on the VGG-16 and the recent state-of-the-art DEX and VGG-Face models.

Our main findings are as follows. (i) Using deep priors (pre-trained models on similar data and/or task) boosts the performance on the target dataset. (ii) Imposing the target age label distribution on pre-trained models helps. (iii) The access to and the use of labeled target samples is critical - with as few as 12 samples used for fine-tuning a large performance gain is achieved, surpassing the impact of imposing target distribution for pre-training.

Early adaptation of deep priors to new target datasets can yield sufficiently good performance at a reasonably low computational cost.

1. Introduction

Convolutional Neural Networks (CNNs) have shown unprecedented performance in a wide range of prediction and estimation tasks [6, 15, 10, 18, 16] including real age and apparent age estimation [2, 12, 8, 1]. However, training such models from scratch for each new task that we face requires massive labeled datasets that might not be available. Thus, the only way to perform well in these new domains is to transfer the knowledge from a large available labeled dataset of a related task. In this work we focus on training and adapting models for predicting apparent or real age in different environments.

The prediction of real age from face images has a long-standing history [3, 9, 5] with multiple datasets being proposed. The largest datasets are manually labeled or web mined, containing tens to hundreds of thousand face images along their age labels. However, the prediction of apparent age is a relatively new task without many large datasets available, and creating such datasets is costly due to the need for multiple annotators per image. Currently, one of the most prominent datasets of apparent age voted is provided by the ChaLearn Looking at People (LAP) Apparent Age Prediction Challenge [2], in which contestants are provided with less than 4700 images to train, validate and test their approaches and models.

In this paper we address the training sample scarcity problem for a target dataset and cast it as an early adaptation of deep models (pre)trained on immense datasets with available labels to the new target dataset application.

Our main contributions are as follows:

- **i.** we show clearly the advantage of using pre-trained models and the importance of pre-training on similar data and/or labels;
- **ii.** knowing as little as the age distribution in the target data is beneficial as it can be imposed onto the pre-trained model for performance gains;
- **iii.** the availability of target age labeled samples is critical as very few such samples combined with a pre-trained model on a similar task can lead to a large boost in performance;
- **iv.** imposing age distribution onto a pre-trained deep model requires a reduced number of back propagations;
- **v.** early adaptation of deep priors to new target datasets can be done efficiently for good performance.

The remainder of the paper is structured as follows. In Section 2 we briefly review related works. In Section 3 we describe the experimental setup including datasets, models and evaluation techniques. In Section 4, we describe our main experiments evaluating the effect of fine-tuning our model on the target dataset, pre-training the model with different distributions and finally pre-training under computational and storage constraints. Conclusions are presented in Section 5.

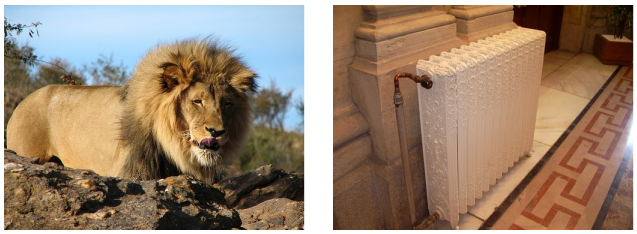
2. Related work

The specific problem of the small dataset provided for apparent age prediction is typically overcome by pre-training a CNN [7] on massive age labeled dataset of face images.

For example, in the ChaLearn LAP challenge [2], the winners Rothe *et al*. [12, 13] employed the VGG-16 [15] architecture trained on ImageNet dataset [14] for image classification and further adapted to their proposed DEX model (see Fig. 2) with pre-training on IMDB-Wiki dataset [13], collected by them from IMDb [1] and Wikipedia [2] websites, with face images and real age labels. Moreover, they fine-tuned this DEX network on the aligned and cropped face images from LAP training set augmented by 10 transformations (*i.e*. rotations) and inferred the predicted age by taking the Softmax expected value of 0-100 years age classes when the aligned faces of validation or test set were fed to the CNN. To address the scarcity of the training data the challenge runner up Lui *et al*. [8] trained two CNNs for face identity recognition on CASIA-WebFace [19] and further trained these models on multiple datasets of real ages, containing more than 1.2 million images to finally fine-tuned their models using the small training set provided in the challenge.

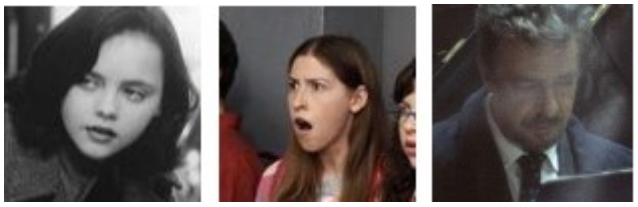
None of the aforementioned works analyzed the effect of target dataset size on adaptation performance because using the whole provided dataset always results in the best prediction accuracy. However, Geng *et al*. [4] analyzed the prediction performance of different models trained on MORPH dataset [11], including a 3 layer Neural Network (NN) and found out that not only the NN outperforms other models when trained on the whole dataset, but also when it was trained on only $\frac{1}{128}$ of MORPH dataset while other models were trained on whole the 55K samples within the MORPH. Though Geng *et al*. did not analyze deep CNNs for age prediction but they showed that a NN may not require training on the whole dataset to achieve a satisfactory result. Wagner *et al*. [17] showed that the transfer learning through coefficients of CNN trained on a similar dataset result in the best early adaptation performance compared to other simpler and less data hungry models trained from scratch.

In this work we analyze how transferring knowledge from (large) datasets of similar task improves the accuracy on a new target dataset when none or very few train samples from the target dataset are available.



(a) Non-face images within ImageNet dataset

(b) Images of seat belt and cowboy hat within ImageNet dataset that are usually accompanied with faces



(c) Faces of 17, 28 and 59 year old celebrities within IMDB dataset

(d) Faces of 17, 35 and 59 year old individuals within LAP dataset

(e) Faces of 18, 34 and 61 year old individuals within MORPH 2

Figure 1: Example images of all datasets used in this work.

[1] http://www.imdb.com/

[2] https://www.wikipedia.org/

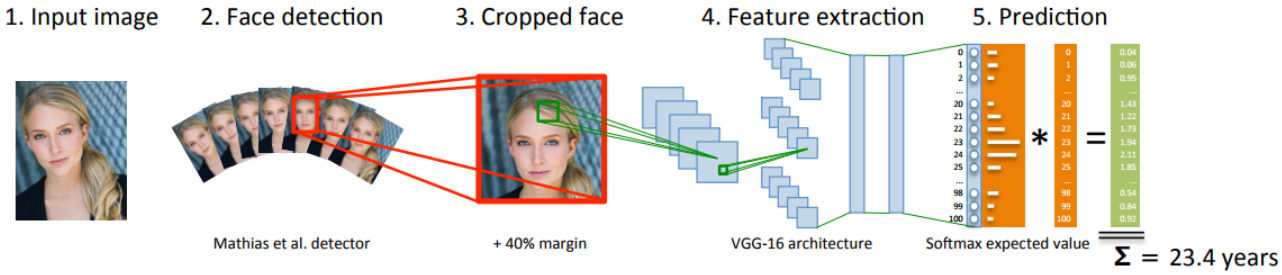


Figure 2: Pipeline of DEX age prediction method, courtesy of Rothe *et al*. [13].

3. Experimental setup

In our experiments we employ standard datasets as commonly used in the literature, a set of state-of-the-art recent CNN deep architectures, and a standard quantitative metric for the age prediction accuracy. These are introduced below.

3.1. Datasets

We work with four standard datasets: ImageNet [14], IMDb-Wiki [13], MORPH 2 [11] and LAP [2]. Examples from each dataset are shown in Fig 1.

ImageNet dataset of Russakovsky *et al*. [14] contains 1.2 million images labeled with 1000 object classes and it was meant for image classification. Although neither humans nor faces are a class in this dataset, many of the images contain humans, some examples are shown in Fig. 1b for the 'seat belt' and 'cowboy hat' classes.

IMDb-Wiki dataset was created by Rothe *et al*. [13] by crawling the profiles of more than 20,000 of the most popular person profiles of the Internet Movie Database (IMDb) and Wikipedia webpages for attributes, such as date of birth and gender, as well as correctly dated images of the individuals. The real age of each of 523K images was calculated by subtracting the date of birth from the date in which the photo was taken. The IMDb dataset is the subset of IMDb-Wiki comprising only the images taken from IMDb.

LAP dataset was introduced by Escalera *et al*. [2] with the ChaLearn Looking at People Apparent Age Prediction Challenge. LAP has 2476 images for training, 1136 images for validation and 1079 images for testing. In this work up to 80% of the training dataset is used for training and the validation set is used for evaluation to make the results of our study comparable to results of Rothe *et al*. [12] that used 90% of the training set for training and reported the prediction error on validation set of the challenge.

MORPH 2 dataset is the second edition of the largest longitudinal database of adult age-progression database proposed by Ricanek and Tesafaye [11]. It contains more than 55K mugshots of 13,000 individuals aged from 18 to above 50. To make the results comparable to the results on LAP dataset and also to the results reported on MORPH 2 by Rothe *et al*. [13] and prior work, we adhere to the common setup from Rothe *et al*. [13] and use up to 4380 images for training and 1095 images for testing.

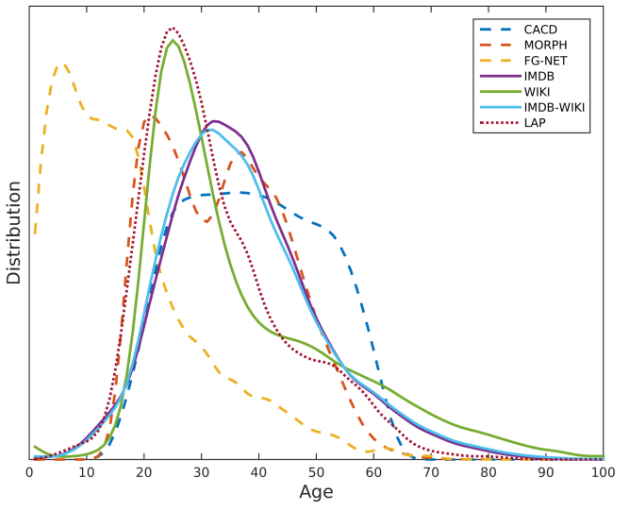


Figure 3: Label distributions of imposed datasets in pre-training DEX. Courtesy of Rothe *et al*. [13]

3.2. Deep models

For all our models we use the VGG-16 architecture [15]. As initializations of our models' weights, we choose weights trained for ImageNet image classification by Simonyan *et al*. [15], for face recognition (VGG-Face) by Parkhi *et al*. [10], and for real age prediction from face images (DEX) on IMDb-Wiki dataset by Rothe *et al*. [12, 13].

VGG-16 is a CNN architecture proposed by Simonyan *et al*. [15]. VGG-16 transforms an RGB image of the size 256×256 to a 4096-dimensional feature vector using a stack of 13 convolutional layers (with receptive field of 3×3 and Max Pooling layers placed every 2-3 layer in between) followed by two fully connected layers with output dimension of 4096. All layers use ReLU non-linear activation and the final output is fed to a fully connected layer with width

of 1000 corresponding to the 1000 classes from ImageNet dataset. In VGG-Face and DEX models the final output is fed to a fully connected layer with width of 2622 (identities) for face recognition (VGG-Face) and to 101 (age bins) for age estimation (DEX).

VGG-Face is a model based on VGG-16 and proposed by Parkhi *et al*. [10] for face recognition in unconstrained environment. The model was trained using the VGG-Face dataset. The best recognition performance was achieved by training on non aligned faces but testing on aligned face images.

DEX or Deep EXpectation model of Rothe *et al*. [12], shown in Fig. 2, is also based on VGG-16 and trained on the IMDb-Wiki dataset. The prediction is given by the softmax expected value over the 101 year bins output layer. Since the focus of our analysis is on the settings of pre-training and fine-tuning process, we skip the face detection and alignment steps and use the cropped and aligned face images as provided by Rothe *et al*. [12, 13] [3] and the 460K IMDb images within this dataset are further split into 80% for training, 10% for validation and 10% for testing. To make the results of our experiments comparable with the results of Rothe *et al*. [12] the same face detection and 10 fold augmentation is used for pre-processing the images for fine-tuning and inference.

3.3. Performance assessment - MAE

To make the evaluation on LAP and MORPH 2 datasets consistent, we use the standard Mean Absolute Error (MAE) measure for reporting quantitative results in all our experiments.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - \mu_i| \tag{1}$$

where n is the number of samples, x_i is the model's prediction and μ_i is the ground truth age label for i-th face image sample.

4. Experiments

In our experiments, we analyze how quickly a pre-trained model can be adapted to a new task when samples from this task are available, when only the target label distribution is known, and finally if only the target label distribution is known and we have limited computational resources available for pre-training.

4.1. Fine-tuning on target train data

In the first experiment, we analyze how many labeled training examples are needed to adapt a model from one domain to another. We sample varying numbers of examples from the target distribution, use them to fine-tune DEX [12], and analyze how quickly adaptation occurs.

[3] https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/

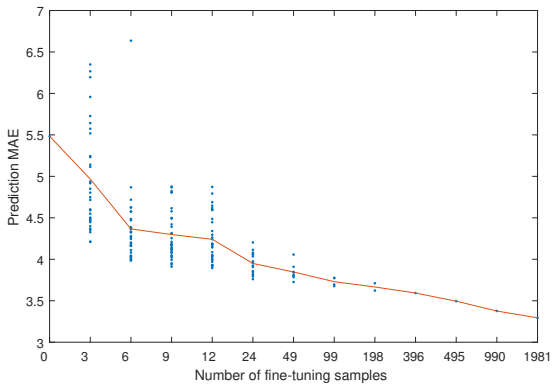


Figure 4: DEX performance on LAP after fine-tuning with different numbers of LAP examples. Points show different trials.

When fine-tuning with a very small number of samples, the fine-tuned model's performance depends strongly on the examples sampled from the target domain. For this reason, we repeated these experiments multiple times. We repeated fine-tunings using less than 0.5% of the training set (*i.e*. 12 LAP samples or less) 32 times, using less than 1% of the training set (*i.e*. 24 LAP samples) 16 times and so on. When fine-tuning with more than 8% of the training set, we only do one trial.

We scale the training steps and number of training iterations linearly with the number of samples available. The fine-tuning with the maximal dataset was done for 5 steps of 2000 iterations. The optimizer is Stochastic Gradient Descent (SGD) with an initial learning rate of 0.0001 in all settings.

Results We report the prediction error for each model as we increase the number of training samples from target dataset in the first columns of Tables 1 and 2 for LAP and MORPH 2, respectively. The adaptation performance on the LAP dataset is also visualized in in Fig. 4).

The plot shows that in some cases, when fine-tuning with only three samples, and in one case when fine-tuning with six, we end up with models that perform worse than the original model we chose as our starting point. On average, however, we see that fine-tuning with only six samples decreases the MAE from 5.5 to 4.4, more than one year of improvement. The total improvement possible by fine-tuning is 2.2 years MAE, but most of the gains come in the first few samples. In MORPH 2, fine-tuning with six samples gives a gain of 0.7 years, out of the total 1.9 years of improvement available through fine-tuning. Thus we see that *if the*

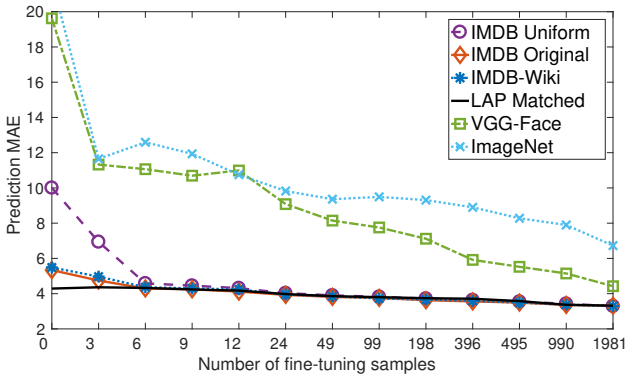


Figure 5: Adaptation performance on LAP of different CNN models.

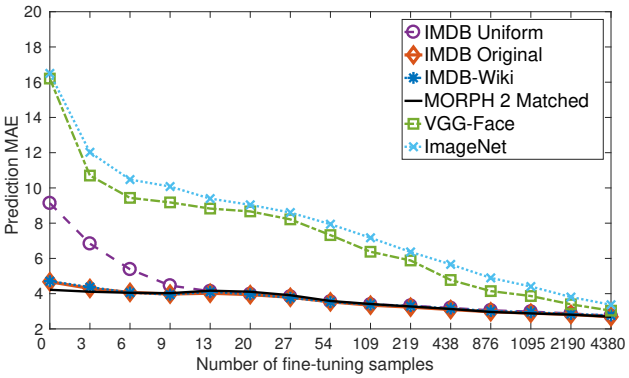


Figure 6: Adaptation performance on MORPH 2 of different CNN models.

source task is sufficiently close to the target task, adaptation occurs very quickly given just a few labeled examples from the target domain.

4.2. Pre-training distribution

We now evaluate how well a model can perform if it is not trained with any samples from the target domain. For this analysis, we evaluate the DEX model [12] with different label distributions imposed, as well as two models not trained for age prediction, *e.g*. VGG-Face [10] and the original VGG-16 trained on ImageNet [15]. We also fine-tune each of these models with data from the new task to quantify the improvement possible by fine-tuning the models directly.

To impose different label distributions onto DEX [12], we resample a dataset that is available to us (*i.e*. IMDB), and use this to further train the pre-trained model. The distributions imposed are: (i) the uniform distribution, (ii) the IMDb dataset, and (iii) the LAP/MORPH2 label distribution. We also evaluate the original DEX model (effectively the IMDb-Wiki distribution). These are shown in Fig 3.

To resample the IMDb dataset to a uniform label distribution, we take the following number of samples for each age:

$$n_{\text{uniform}}(\text{age}) = \frac{4 \times N_{\text{IMDb}}}{101} \tag{2}$$

where N_{dataset} is the total number of samples in *dataset*.

For the matched distribution, we have

$$n_{\text{matched}}(\text{age}) = n_{\text{target}}(\text{age}) \times \frac{4 \times N_{\text{IMDb}}}{N_{\text{target}}} \tag{3}$$

Due to overfitting, training with the uniformly distributed dataset was only done for one epoch with the learning rate of 0.00001. All other models were trained to convergence.

Results The model performances before fine-tuning are shown in the top row of Tab. 1 for the LAP and in Tab 2 for the MORPH 2. We see that *imposing the target label distribution improves model performance* - relative to DEX, the MAE decreases from 5.484 to 4.289 for LAP, and from 4.708 to 4.220 for MORPH 2. However, although the label distribution of the target has been matched, the pixel distribution of the target is not, which is why considerable improvements are still possible by fine-tuning. Imposing a distribution which does not match the true target distribution is very harmful - imposing the uniform distribution almost doubles the MAE in both LAP and MORPH 2.

To adapt VGG-Face [10] and VGG-16 [15] to age prediction, we replace their final layers with a randomly initialised 101-dimensional fully connected layer. Thus the performances reported for these models before fine-tuning are equivalent to random guessing.

We now examine how quickly each of these models converges to the new task when fine-tuning, as shown in Figs 5 and 6. On average, imposing the LAP distribution only is better than than the fine-tuned DEX model when less than nine samples are used for fine-tuning. In the case of MORPH 2 , this crossover point happens after fine-tuning with six samples. This could be because the uniform background of the MORPH 2 mugshots makes the images less variable and easier to learn. We also see that the effect of imposing a distribution which is far away from the target, *e.g*. the uniform distribution, is erased when we fine-tune with six samples in the case of LAP and with nine samples for MORPH 2. Indeed, all models based on DEX converge to performances within a spread of 0.1 MAE after being fine-tuned with just 24 samples for LAP, and 13 samples for MORPH 2. Thus, *fine-tuning with labeled examples is very important and allows the model to adapt very quickly,* i.e. *after being exposed to only a few target examples.*

This point is made even more strongly by the experi-

	DEX Model								VGG-Face		VGG-16	
dataset:	IMDB-Wiki		IMDB		IMDB		IMDB		VGG-Face		ImageNet	
distribution:	original		uniform		original		LAP matched		original		original	
# LAP	MAE	SD	MAE	SD	MAE	SD	MAE	SD	MAE	SD	MAE	SD
0	5.484		10.007		5.337		4.289		19.626		21.769	
3	4.967	±0.623	6.940	±1.114	4.743	±0.556	4.362	±0.384	11.325	±2.160	11.652	±2.960
6	4.366	±0.468	4.584	±0.571	4.295	±0.543	4.325	±0.386	11.063	±2.172	12.590	±4.727
9	4.298	±0.291	4.454	±0.351	4.246	±0.281	4.240	±0.217	10.694	±1.793	11.934	±2.306
12	4.241	±0.278	4.315	±0.317	4.117	±0.258	4.173	±0.240	10.99	±1.585	10.751	±1.135
24	3.950	±0.127	4.031	±0.149	3.931	±0.134	3.977	±0.081	9.081	±0.675	9.823	±0.686
49	3.847	±0.094	3.902	±0.069	3.817	±0.082	3.858	±0.088	8.146	±0.354	9.361	±0.281
99	3.730	±0.045	3.822	±0.082	3.758	±0.049	3.802	±0.027	7.756	±0.280	9.493	±0.063
198	3.666	±0.044	3.722	±0.038	3.628	±0.050	3.742	±0.054	7.117	±0.176	9.310	±0.123
396	3.593		3.636		3.560		3.712		5.917		8.899	
495	3.495		3.546		3.494		3.578		5.523		8.278	
990	3.376		3.428		3.361		3.368		5.146		7.897	
1981	3.294		3.294		3.314		3.318		4.423		6.726	

Table 1: Adaptation performance on LAP of different models with new imposed distributions after fine-tuning with LAP.

	DEX Model								VGG-Face		VGG-16	
dataset:	IMDB-Wiki		IMDB		IMDB		IMDB		VGG-Face		ImageNet	
distribution:	original		uniform		original		MORPH 2 matched		original		original	
# MORPH 2	MAE	SD	MAE	SD	MAE	SD	MAE	SD	MAE	SD	MAE	SD
0	4.708		9.136		4.673		4.220		16.206		16.499	
3	4.373	±0.428	6.838	±0.897	4.296	±0.386	4.110	±0.301	10.702	±2.773	12.023	±3.777
6	4.060	±0.382	5.384	±0.729	4.072	±0.357	4.050	±0.307	9.436	±1.105	10.475	±2.074
9	3.947	±0.255	4.460	±0.491	3.977	±0.180	4.028	±0.233	9.179	±1.130	10.079	±2.232
13	4.107	±0.532	4.139	±0.487	4.024	±0.420	4.168	±0.467	8.828	±0.895	9.394	±1.679
20	3.968	±0.364	3.977	±0.387	3.944	±0.355	4.109	±0.486	8.667	±0.822	9.040	±0.820
27	3.771	±0.316	3.833	±0.301	3.794	±0.320	3.911	±0.404	8.215	±0.953	8.603	±0.685
54	3.526	±0.122	3.552	±0.148	3.534	±0.170	3.584	±0.111	7.322	±0.399	7.941	±0.354
109	3.383	±0.082	3.388	±0.068	3.346	±0.076	3.414	±0.090	6.378	±0.363	7.171	±0.495
219	3.294	±0.050	3.316	±0.047	3.233	±0.085	3.273	±0.049	5.883	±0.259	6.367	±0.043
438	3.156	±0.005	3.188	±0.026	3.098	±0.012	3.131	±0.043	4.776	±0.036	5.661	±0.224
876	3.014		3.054		2.972		2.962		4.145		4.895	
1095	2.945		2.980		2.893		2.880		3.866		4.404	
2190	2.843		2.876		2.820		2.808		3.394		3.806	
4380	2.768		2.743		2.689		2.691		3.035		3.370	

Table 2: Adaptation performance on MORPH 2 of different models with new imposed distributions after fine-tuning with MORPH 2.

ments on VGG-Face and VGG-16. Although the discrepancy between VGG-Face and DEX after fine-tuning with 3 samples is 6.3 MAE for LAP (6.6 for MORPH 2), after fine-tuning with all available samples, the difference in performance is only 1.1 MAE for LAP and 0.3 MAE for MORPH 2, despite the fact that the models started from very different initializations. Even more impressively, VGG-16 trained on ImageNet fine-tuned with 2190 samples manages to outperform MORPH 2 imposed on DEX. Thus, *having access to labeled examples from the target domain enables the model to overcome initializations that are only weakly related to the target task.*

4.3. Pre-training dataset size

In the previous section, we showed that imposing the target distribution improves performance in the absence of samples from the new domain. However, pre-training for multiple epochs on a large re-sampled dataset requires massive amounts of computation. For example, each 10 epoch pre-training done in Section 4.2 took about a day on 3 Maxwell Titan X GPUs.

In this experiment, we investigate whether we can im-

	Number of samples from IMDB used to impose LAP distribution									
	0		100		1,000		10,000		368,553	
# LAP	MAE	S.D.	MAE	S.D.	MAE	S.D.	MAE	S.D.	MAE	S.D.
0	5.484		4.416		4.280		4.148		4.289	
3	4.967	±0.623	4.527	±0.440	4.339	±0.415	4.351	±0.331	4.362	±0.384
6	4.366	±0.468	4.449	±0.544	4.330	±0.382	4.310	±0.481	4.325	±0.386
9	4.298	±0.291	4.524	±0.433	4.304	±0.329	4.266	±0.376	4.240	±0.217
12	4.241	±0.278	4.339	±0.292	4.250	±0.324	4.241	±0.244	4.173	±0.240
24	3.950	±0.127	4.008	±0.091	4.007	±0.148	4.036	±0.098	3.977	±0.081
49	3.847	±0.094	3.882	±0.102	3.892	±0.097	3.928	±0.038	3.858	±0.088
99	3.730	±0.045	3.776	±0.040	3.785	±0.040	3.840	±0.048	3.802	±0.027
198	3.666	±0.044	3.641	±0.066	3.645	±0.031	3.756	±0.069	3.742	±0.054
396	3.593		3.596		3.620		3.589		3.712	
495	3.495		3.450		3.594		3.529		3.578	
990	3.376		3.404		3.419		3.407		3.368	
1981	3.294		3.283		3.271		3.331		3.318	

Table 3: Adaptation of models with LAP distribution imposed in pre-training using 0, 100, 1000, 10000 or 368,553 samples after fine-tuning with different number of LAP samples.

	Number of samples from IMDB used to impose MORPH 2 distribution									
	0		100		1,000		10,000		368,553	
# MORPH 2	MAE	S.D.	MAE	S.D.	MAE	S.D.	MAE	S.D.	MAE	S.D.
0	4.708		3.994		3.945		3.844		4.220	
3	4.373	±0.428	3.998	±0.332	3.923	±0.256	3.824	±0.303	4.110	±0.301
6	4.060	±0.382	3.879	±0.276	3.866	±0.301	3.750	±0.285	4.050	±0.307
9	3.947	±0.255	3.912	±0.267	3.853	±0.241	3.752	±0.174	4.028	±0.233
13	4.107	±0.532	4.103	±0.543	4.165	±0.652	4.041	±0.540	4.168	±0.467
20	3.968	±0.364	4.013	±0.422	4.112	±0.483	4.019	±0.492	4.109	±0.486
27	3.771	±0.316	3.839	±0.340	3.927	±0.378	3.804	±0.407	3.911	±0.404
54	3.526	±0.122	3.547	±0.117	3.594	±0.100	3.543	±0.123	3.584	±0.111
109	3.383	±0.082	3.427	±0.096	3.478	±0.117	3.423	±0.073	3.414	±0.090
219	3.294	±0.050	3.322	±0.055	3.347	±0.041	3.295	±0.067	3.273	±0.049
438	3.156	±0.005	3.167	±0.010	3.202	±0.020	3.158	±0.009	3.131	±0.043
876	3.014		3.025		3.044		3.035		2.962	
1095	2.945		2.966		2.965		2.964		2.880	
2190	2.843		2.867		2.837		2.854		2.808	
4380	2.768		2.756		2.744		2.727		2.691	

Table 4: Adaptation of models with MORPH 2 distribution imposed in pre-training using 0, 100, 1000, 10000 or 368,553 samples after fine-tuning with different number of MORPH 2 samples.

pose the target distribution using less computation and using less memory to store the entire source dataset. To do this, we impose the target distribution using only 100, 1000 and 10000 samples from the IMDB dataset. Such pre-trainings took less than a minute, 5 minute and 45 minutes on the same 3 GPU system respectively.

Again, after imposing these distributions, we fine-tune the obtained models to see the relative improvement of imposing a distribution versus fine-tuning.

Results The results are shown in Tables 3 and 4, as well as visualized in Fig 7 and 8. Here we see that imposing the distribution with just 100 samples already give more than one year improvement in MAE, in LAP and 0.7 MAE in MORPH 2. Increasing our training set tenfold only gives an additional gain of 0.13 and and 0.05 MAE respectively.

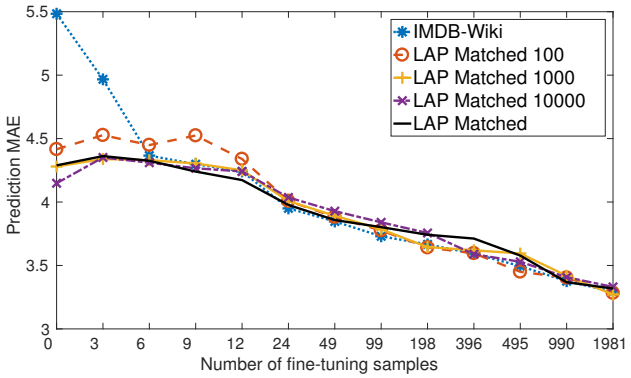


Figure 7: Adaptation performances on LAP test data of LAP matched pre-trainings with different sizes.

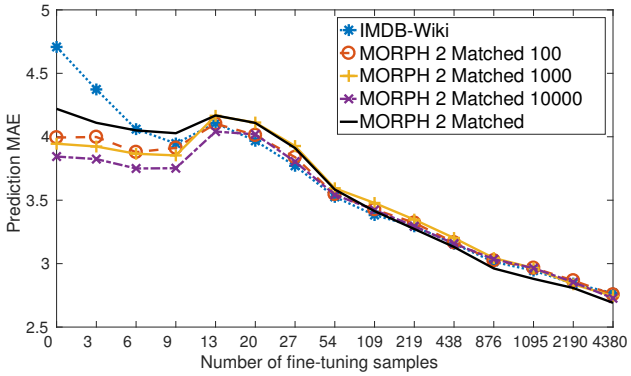


Figure 8: Adaptation performances on MORPH 2 test data of MORPH 2 matched pre-trainings of different sizes.

Thus, most of the benefit of imposing the label distribution can already be obtained by using only a 100 samples to train the network.

Imposing MORPH 2 onto DEX using only 100 samples outperforms DEX fine-tuned with less than 9 samples. Using 1000 samples to impose the LAP distribution outperforms DEX fine-tuned with 9 samples. Thus, if we have no access to the samples from the target domain, and would like to impose a label distribution, we can obtain satisfactory results using only 100 or 1000 samples, cutting computational time by a factor of around 370 in our case.

Note about Fig. 8, on MORPH 2 - the 'bump' in error around 20 samples comes from the fact that our learning policy was selected for training with smaller datasets and kept fixed for all fine-tuning sizes. Despite the increase in error around 20 samples, with more training, the models adapt to larger fine-tuning sizes. This 'bump' is less noticeable in the case of LAP dataset as shown in Fig. 7.

In both the LAP and MORPH 2 experiments, imposing the target distribution using 10000 samples has lower MAE than using the entire dataset. This could be due to the inherited settings from DEX, a model optimized for the LAP dataset with a number of train samples in a comparable range.

A target age label distribution can be imposed onto a pre-trained model by fine-tuning with only 100 or 1000 samples from the pre-train dataset.

5. Conclusion

In our experiments, we found that having even a little bit of information from the target domain helps to adapt a pre-trained model to a new task.

When we have access to samples directly from the target domain, training even with a small number of these decreases the the MAE significantly on average. Already fine-tuning the DEX model with six samples decreases the MAE by more than one year in the LAP dataset and by 0.7 in the MAE. This is a significant part of the total 2.2 and 1.9 year MAE improvements that are possible when we fine-tune with the entire dataset. Having access to samples from the target domain also lets us use models trained on non-face tasks, such as VGG-16 trained on ImageNet, and adapt them. Fine-tuning allowed this model to go from 12 MAE to 3.4 MAE - within 0.7 MAE of the best-performing fine-tuned model on MORPH 2.

However, even if we do not have any samples from the target dataset available, but know its label distribution, we can also adapt a pre-trained model by imposing the label distribution of the target task. When less than 12 (LAP, or 13, MORPH 2) samples are available for fine-tuning, imposing the distribution has better performance than fine-tuning the DEX model.

In an even more constrained scenario, we show that even when we are limited in how much pre-training we can do, we can still have gains in imposing the target distribution using only 100 samples. This amount of pre-training is equivalent to fine-tuning with six to nine samples, in both datasets.

In summary, we have characterized how quickly a pre-trained model can be adapted to a new task, both when data from this new task is available, and when only the label distribution of this task is known. When there are less than around ten samples available from the target domain, imposing a distribution is helpful to the model performance. However, having access to labeled data is the best signal for learning.

Acknowledgments

This work was supported by the ETH Zurich General Fund (OK), the Fashwell project, and an NVIDIA GPU grant.

References

[1] E. Agustsson, R. Timofte, S. Escalera, X. Baro, I. Guyon, and R. Rothe. Apparent and real age estimation in still images with deep residual regressors on appa-real database. In *12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017)*, 2017.

[2] S. Escalera, J. Fabian, P. Pardo, X. Baró, J. Gonzalez, H. J. Escalante, D. Misevic, U. Steiner, and I. Guyon. Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1–9, 2015.

[3] Y. Fu, G. Guo, and T. S. Huang. Age synthesis and estimation via faces: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 32(11):1955–1976, 2010.

[4] X. Geng, C. Yin, and Z.-H. Zhou. Facial age estimation by learning from label distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(10):2401–2412, 2013.

[5] H. Han, C. Otto, and A. K. Jain. Age estimation from face images: Human vs. machine performance. In *Biometrics (ICB), 2013 International Conference on*, pages 1–8. IEEE, 2013.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[8] X. Liu, S. Li, M. Kan, J. Zhang, S. Wu, W. Liu, H. Han, S. Shan, and X. Chen. Agenet: Deeply learned regressor and classifier for robust apparent age estimation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 16–24, 2015.

[9] G. Panis and A. Lanitis. An overview of research activities in facial age estimation using the fg-net aging database. In *European Conference on Computer Vision*, pages 737–750. Springer, 2014.

[10] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In *BMVC*, volume 1, page 6, 2015.

[11] K. Ricanek and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 341–345. IEEE, 2006.

[12] R. Rothe, R. Timofte, and L. Van Gool. Dex: Deep expectation of apparent age from a single image. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, December 2015.

[13] R. Rothe, R. Timofte, and L. Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, pages 1–14, 2016.

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[15] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[16] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, L. Zhang, et al. NTIRE 2017 challenge on single image super-resolution: Methods and results. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July.

[17] R. Wagner, M. Thom, R. Schweiger, G. Palm, and A. Rothermel. Learning convolutional neural networks from few samples. In *Neural Networks (IJCNN), The 2013 International Joint Conference on*, pages 1–7. IEEE, 2013.

[18] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: towards good practices for deep action recognition. In *European Conference on Computer Vision (ECCV)*, volume 9912, pages 20–36, 2016.

[19] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.