Towards a General Model of Knowledge for Facial Analysis by Multi-Source Transfer Learning
Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, Frédéric Jurie

To cite this version:
Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, Frédéric Jurie. Towards a General Model of Knowledge for Facial Analysis by Multi-Source Transfer Learning. 2019. hal-02342091v2

HAL Id: hal-02342091
https://hal.archives-ouvertes.fr/hal-02342091v2
Submitted on 4 Nov 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.
L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Towards a General Model of Knowledge for Facial Analysis by Multi-Source Transfer Learning

Valentin Vielzeuf\textsuperscript{1,2}  Alexis Lechervy\textsuperscript{2}  Stéphane Pateux\textsuperscript{1}  Frederic Jurie\textsuperscript{2}

\textsuperscript{1}Orange Labs, Cesson-Sévigné, France  \textsuperscript{2}Université Caen Normandie, France

Abstract

This paper proposes a step toward obtaining general models of knowledge for facial analysis, by addressing the question of multi-source transfer learning. More precisely, the proposed approach consists in two successive training steps: the first one consists in applying a combination operator to define a common embedding for the multiple sources materialized by different existing trained models. The proposed operator relies on an auto-encoder, trained on a large dataset, efficient both in terms of compression ratio and transfer learning performance. In a second step we exploit a distillation approach to obtain a lightweight student model mimicking the collection of the fused existing models. This model outperforms its teacher on novel tasks, achieving results on par with state-of-the-art methods on 15 facial analysis tasks (and domains), at an affordable training cost. Moreover, this student has 75 times less parameters than the original teacher and can be applied to a variety of novel face-related tasks.

1. Introduction

An increasing number of deep neural networks has been implemented and trained during the last few years. These existing models can be seen as incredibly rich and compressed sources of knowledge about diverse domains, which can be reused to tackle novel tasks by transferring this knowledge. In this context, the standard way for knowledge transfer consists in selecting a single source, generally under the form of the parameters of a statistical model (e.g., a pre-trained network), and to re-use it on a new task by fine-tuning the parameters. The knowledge source is often empirically chosen, typically by selecting the closest task according to human judgment or a complex rich task such as the ImageNet classification task.

To automate this selection process, recent works \cite{40, 2, 67} have shown that a relational space between diverse basic tasks/models can be exploited, allowing to combine few potential candidate models and make the transfer more informative than using a single model.

However, the models discarded by this selection process may still contain useful knowledge. By analogy with multi-modal problems, one may consider each independent model (specifically the last hidden representation of the neural networks) as a modality for the new task. Some modalities taken in isolation can yield bad results while they carry useful information complementary to other modalities. On the contrary, two well-performing modalities can be redundant and combining them brings less improvement.

Extending this reasoning to \( M \) modalities can be done by learning a common representation embracing all the modalities. However, this is not a trivial problem. For example, naively concatenating the local embeddings provided by each model does indeed produce a common embedding, but does not work well when dealing with many tasks/modalities, as observed by Zamir \textit{et al.} \cite{67}.

We can formulate this multi-source transfer learning as follows. We define knowledge items as (domain, task) pairs, which can be associated with a model trained on this domain to fulfill this task. Supposing having \( M \) of such source knowledge, how to group them into a unique and general model of knowledge, performing well both on source knowledge and on novel target knowledge?

By addressing such a goal, this paper makes three main contributions: (a) the definition of a carefully designed ensemble of source and target knowledge related to facial analysis. Facial analysis is a topic of broad interest having received a lot of attention from the community, and for which large sets of knowledge (\textit{i.e.} pre-trained models) are available. Moreover, a general knowledge in such a field would be of practical interest, as the number of new face-related tasks and domains grows exponentially. (b) a simple yet efficient methodology to project the source embeddings into a unique one, which is accurate both on the source and the target knowledge. (c) A distillation process
is introduced to transfer the learned general knowledge into one lightweight convolutional neural network. This simple model outperformed its teacher, being on par with state-of-the-art models specifically built for solving specific tasks on specific domains. Moreover, it fits real-world application requirements, with 2 million parameters.

2. Related Work

The proposed work is related to several fields including information compression, transfer learning or distillation.

**Information Fusion and Multi-Task Learning** Combining the knowledge from several existing models into a single representation can be seen as a multimodal fusion, where model features are the input modalities. Classical methods exist such as Principal Components Analysis [35] or Canonical Correlation Analysis [30]. The recent trend is often focusing on equivalent neuronal methods such as the one Ngiam et al. [49] proposed, using multimodal autoencoders to learn a representation able to reconstruct all the modalities. The literature on deep multimodal fusion can be divided in i) multimodal architectures [62, 48, 53] focusing on where to fuse the information in the network, ii) representations based on constraints [3, 14, 61] building on the relations between the modalities (e.g. correlations). More details about this field can be found in these two surveys [5, 8].

Information fusion and multitask learning approaches are often related, as in the multitask autoencoder proposed by Ghifary et al. [26] allowing to improve domain generalization. The recent work by Ruder et al. [59] on multitask architecture search allows to learn latent architectures for multitasking problems. Multitask approaches for face analysis have been proven to be as efficient or even better than single-task learning [15]. In the context of this paper, we can also see the fusion of diverse existing models both as a multimodal problem (each model is a modality) and as a multitask problem (the final student model has to be efficient on all tasks).

**Transfer Learning** As the goal of the paper is to transfer knowledge to novel tasks, it can be related to transfer learning. We share the same motivations as Taskonomy proposed by Zamir et al. [67], helping to select which combination of existing models to use when tackling a new task. Nevertheless in our case, because the existing models are already partially related to the target task, there is a benefit to select them all and keep what is useful in each model. Geyer et al. [25] propose to merge two pre-trained models before transferring them, based on incremental moment matching [39]. Chen et al. [17] designed a coupled end-to-end transfer learning, distilling the knowledge of one source model into the target model, while selective adversarial networks are proposed in [12] to select positive transfers and discard negative ones in the particular case where the target label space is a subspace of the source label space. Finally, regarding domain generalization, Mancini et al. [46] propose to fuse the outputs of domain-specific neural networks, after predicting the domain of target samples.

**Self-supervised Learning** Our work uses a large dataset to learn an autoencoder in an unsupervised fashion, and, by this means, exhibits the relations between the given models. Therefore, training the Teacher can be seen as a self-supervised learning process, which is a widely explored topic [69, 70, 51]. The goal of self-supervised learning is to design an efficient and cost-less proxy task helping to solve the target task for which we don't have enough annotations. Doersch et al. [23] showed the benefit of using several proxy tasks in self-supervised learning, which is basically what we are doing with the six different embeddings to reconstruct. The difference lies in the very definition of the proxy task. Ours are coming from previously learned knowledge. In traditional self-supervised, the tasks are low-level objectives such as solving a jigsaw puzzle [50] or evaluating the rotation of an image [27]. Radenovic et al. [54] used a proxy task closer to ours, using state-of-the-art models to extract edges of images as labels for their visual model.

Other methods falling in this unsupervised learning category can also be linked to ours, such as the deep clustering approach [13] where a convolutional neural network if trained using the output of a k-means clustering algorithm.

**Model Compression and Knowledge Distillation** A last important aspect of our method is to distill [32] the learned knowledge into a single lightweight model. In this spirit, Romero et al. proposed a deep and thin student named FitNet [56]. It learns from both last layer and hidden layers of the teacher and outperforms it. Aiming to ensure privacy of the learning dataset, Papernot et al. [52] proposed to use several teachers in a semi-supervised fashion.

With a goal close to ours, Chebotar et al. [16] used as a teacher a weighted average prediction of an ensemble of neural networks. More recently the idea of data distillation was developed by [55], applying the same teacher model to diverse transformations of the input image and taking the average prediction as a label for the student. Li et al. [43] use feature map attention to regularize the learning of the student. Finally, multimodal distillation bears similarities with our approach, using some well-known modalities of a given input to master other views with fewer annotations. For instance, the SoundNet model [6] learns an audio model from the labels yield by a visual model, while Xu et al. [65] proposed to fuse the predictions coming from diverse modalities to improve the quality of the final main task of the student.
3. Methodology

To the best of our knowledge, transferring multi-source knowledge across both tasks and domains has not been addressed yet in the literature. Therefore, we will first study how to formally and practically formalize such a problem. We will then detail the two main steps composing our approach: dimensionality reduction and distillation.

3.1. Multi-source Multi-domain Transfer Learning

General formulation Let’s first define the concept of knowledge as the abstract ability to perfectly solve a given task \( t \) on a given domain \( d \). We limit ourselves to the family of tasks including classification / regression with machine learning techniques. The knowledge extracted when solving this problem (e.g. through the training of a deep neural network) can be denoted as \( K_{(t,d)} = (E_{(t,d)}, C_{(t,d)}) \), where

- \( E_{(t,d)} \) is a function able to map each element \( x \) of domain \( d \) into a common embedding \( h \)
- \( C_{(t,d)} \) is a function able to map each embedding \( h \) to the expected output \( y \)

When dealing with a new target problem defined by a task \( t' \) on a domain \( d' \), \( K_{(t',d')} \) or \( K_{(t,d)} \) can be used to initialize a proposal for \( K_{(t',d')} \), that can be further adapted to fit the task/domain. Transfer learning consists in reusing (or fine-tuning) \( E_{(t,d)} \) and learning a new specific \( C_{(t',d')} \). While domain adaptation leads to learn \( E_{(t,d+d')} \) and \( C_{(t,d+d')} \).

The general problem we tackle is the one of adapting transfer learning and domain adaptation to the case where we have not only one already solved problem, but a set \( S \) of solved problems.

Following transfer learning and domain transfer, our approach is based on defining an operator \( \mathcal{E}^G \) = \( \mathcal{G}(E_{(t_i,d_i)}, i \in S) \) where \( \mathcal{G} \) is a combination function using all the \( h_i \) embedding (coming from the \( K_{(t_i,d_i)} \)). It aims at gathering all knowledge information regarding the \( t_i \) and the \( d_i \) into one embedding \( h_G \).

One possible solution is to concatenate all the outputs of \( E_{(t_i,d_i)}, i \in S \) to generate \( h_G \). But as observed in Taskonomy [67], it generally suffers from lack of generalization.

Recent approaches such as Taskonomy [67, 2] have addressed this question by selecting the \( K_{(t_i,d_i)} \) where \( t_i \) and \( t_j \) are most correlated.

In our approach (see Section 3.2), \( \mathcal{G} \) is a neuronal encoder allowing to estimate a representation \( h_G \) of reduced dimensionality, still approximating well the \( h_i \) and potentially leading to better generalizations by removing biases due to over-complexity.

Nevertheless, \( \mathcal{E}^G \) is then composed of all \( E_{(t_i,d_i)} \) and of \( \mathcal{G} \) and thus mapping a given \( x \) to \( h_G \) has a high computational cost. In Section 3.3 we propose the use of distillation to transform \( \mathcal{E}^G \) into a unique and lightweight model \( \mathcal{E}^G_{\text{unique}} \), directly projecting \( x \) to \( h_G \) and therefore allowing easier transfer when using it to get a new knowledge.

Source and target knowledge for facial analysis As building a general knowledge on all tasks and domains existing would not be feasible, we propose to validate our approach experimentally by carefully designing a set \( S \) of Source-Knowledge and a set \( T \) of Target-Knowledge to cover a facial analysis general knowledge, including the 15 different knowledge described in Table 1.

\( S \) gathers M=6 source knowledge, which leads to a domain \( \bigcup d^S_i \) of around 22 million faces with M=6 different tasks contains in around 106 million parameters. To train \( G \) (in an unsupervised fashion), instead of using \( \bigcup d^S_i \), we will use a domain \( D_{\text{unsup}} \) of 4.12 millions of faces (discarding annotations), extracted from VGGFace2 [10] (3.14 million), EmotioNet [24] (0.72 million) and IMDb-WIKI [57] (0.26 million).

\( T \) contains 9 various knowledge on facial analysis. Note that two tasks with the same name (for instance Expression Classification) may still be different (for instance facial expression is subjective and depends on its annotators [68]).

<table>
<thead>
<tr>
<th>Source-Knowledge</th>
<th>Target-Knowledge</th>
<th>Task Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expr-AffectNet</td>
<td>Emotion Classif.</td>
<td>0.3M [47]</td>
</tr>
<tr>
<td>Expr-RAF</td>
<td>Emotion Classif.</td>
<td>15,339 [42]</td>
</tr>
<tr>
<td>Expr-SFEW</td>
<td>Emotion Classif.</td>
<td>1,766 [22]</td>
</tr>
<tr>
<td>Identity-MS</td>
<td>Identity Matching</td>
<td>6.5M [28]</td>
</tr>
<tr>
<td>Identity-LFW</td>
<td>Identity Matching</td>
<td>13,000 [33]</td>
</tr>
<tr>
<td>Gender-IMDb</td>
<td>Gender Prediction</td>
<td>0.5M [57]</td>
</tr>
<tr>
<td>Gender-UTK</td>
<td>Gender Prediction</td>
<td>20,000 [44]</td>
</tr>
<tr>
<td>Attrib-CelebA</td>
<td>Attrib. Detection</td>
<td>0.2M [44]</td>
</tr>
<tr>
<td>AgeR-IMDb</td>
<td>Age Regression</td>
<td>0.5M [57]</td>
</tr>
<tr>
<td>AgeR-FG</td>
<td>Age Regression</td>
<td>1,000 [38]</td>
</tr>
<tr>
<td>AgeR-UTK</td>
<td>Age Regression</td>
<td>20,000 [71]</td>
</tr>
<tr>
<td>AgeC-UTK</td>
<td>Age Classif.</td>
<td>20,000 [71]</td>
</tr>
<tr>
<td>Ethnic-UTK</td>
<td>Ethnicity Classif.</td>
<td>20,000 [71]</td>
</tr>
<tr>
<td>Pain-UNBC</td>
<td>Pain Estimation</td>
<td>48,000 [45]</td>
</tr>
<tr>
<td>Object-ImageNet</td>
<td>Object Classif.</td>
<td>14M [21]</td>
</tr>
</tbody>
</table>

Table 1. The M=6 source and 9 target knowledge (source is in bold). The task, as well as the used pre-trained model for source knowledge, are described in the associated papers. Information about the used domain can be found in the second citation on each row. Moreover, more details about all the selected knowledge will be provided in supplementary materials.

3.2. Dimensionality Reduction for a More General Knowledge

We explain in the previous subsection that we dispose of M source knowledge encoders \( E_{(t_i,d_i)} \), allowing to extract M \( h_i \) embedding from a given face \( x \). We study here
Step 1: Dimensionality Reduction

Step 2: Distillation

Step 3: Estimating a new knowledge

Figure 1. Overview of the method. Step 1 is the training of $\mathcal{G}$ on $D_{unsup}$, to obtain an encoder $E^G$ of $x$ in a compact embedding $h_G$, gathering all sources of knowledge (represented by the extracted $h_i$). Step 2 aims at reducing the number of parameters used to get $h_G$ by distilling the knowledge of 130 millions parameters into a much simpler encoder $E^G_{unique}$, leading to $\tilde{h}_G$. In step 3 this lightweight encoder $E^G_{unique}$ can then be plugged with a MLP $C(t_n,d_n)$ and all parameters adapted to learn new knowledge $K_{(t_n,d_n)}$.

the operator $\mathcal{G}$ allowing to combine these embedding into a compact and general embedding $h_G$.

Motivations  Defining $\mathcal{G}$ as a basic concatenation function implies a very large and redundant $h_G$. Therefore there is a meaning in reducing the dimension of $h_G$, leading to discard some redundancies and exploit the complementarity between the $h_i$. Dimensionality reduction is a well-explored topic, from linear projections such as PCA [35] to more complex non-linear approaches such as manifold learning [34]. The major drawback of a linear approach is that the representation is projected into a plane and therefore can miss the real shape of the data. Thus, learning the manifold represented by $h_G$ with a (non-linear) neural network [41] makes sense.

Adopted Approach  As we want to reduce the dimensionality without supervision, we propose to train $\mathcal{G}$ with an auto-encoding objective $L_{Rec}$, as shown in Step 1 of Figure 1. In other words, we are building $h_G$ as an answer to all source tasks (represented by the $h_j$) but on the $D_{unsup}$ domain. For that, we optimize the parameters of both $\mathcal{G}$ (as an encoder) and of $\tilde{\mathcal{G}}$ (as a decoder reconstructing the $h_i$):

$$h_G = \mathcal{G}((h_i)_{i=1..M})$$  

$$h_i = 1..M = \tilde{\mathcal{G}}(h_G)$$  

$$L_{Rec} = \sum_{m=1}^{M} ||\tilde{h}_i - h_i||^2$$  

Note that we choose $\tilde{\mathcal{G}}$ to have a symmetric architecture to $\mathcal{G}$. We will further discuss the architecture choice of an autoencoder in the experiments section, by also experimenting with PCA, regular autoencoders [7], variational autoencoders [37] and denoising autoencoders [64].

Another important choice is the choice of the dimensionality of $h_G$, as it drives the knowledge compression process. As we are aiming at creating a unique general knowledge, we consider as a rule of thumb to fix it as the average dimensionality of the $h_i$. We will conduct an empirical study in subsection 4.4 on the impact of this dimension on the quality of $h_G$ and check that there is no optimal dimension, only extreme values (very low-dimensional or high-dimensional) clearly degrading our approach.

3.3. Real-world Transfer Learning by Distillation

As mentioned in the problem definition, obtaining $h_G$ by step 1 implies using $E^G$, composed of $M=6$ different pre-trained models combined to $\mathcal{G}$. It leads to a huge number of parameters (130M) and greatly limits the possibility of domain adaptation when dealing with the target knowledge, as we can’t adapt such a large number of parameters on a new domain. A natural way to solve this problem would then to compress the so-obtained model. Model compression can be addressed with several approaches as shown in the recent literature [18], mainly gathered into four categories: parameters pruning, low-rank factorization, compact convolutional filters and knowledge distillation.

Our current model is composed of $M=6$ different specific
branches and we do not only want to reduce the number of parameters: we want to achieve a unique encoder. It is the promise of the distillation approach, allowing to train a new Student model, supervised by the previous big Teacher model [32]. Thus, we will not only make $\mathcal{E}^G$ lighter but also transform it to a conventional estimator $\mathcal{E}_G^{unique}$, such as a classic convolutional neural network architecture, on which well-known methods such as data augmentation can be easily applied. Moreover previous works [19] have shown that fine-tuning all the parameters of a model may lead to better domain adaptation and improve the quality of the transfer.

### Distillation

To achieve distillation, we still are training our new $\mathcal{E}^G_{unique}$ on $\mathcal{D}_{unsup}$, as illustrated in Figure 1. We consider $\mathcal{E}^G_{unique}$ as a neural network of arbitrary architecture (we choose a ResNet-18 [31] for all experiments), taking the face image $x$ and directly projecting it into a representation $h_G$.

The training is done by minimizing $\mathcal{L}_{Distillation}$

$$\mathcal{L}_{Distillation} = D_c(h_G, \hat{h}_G)$$

where $D_c$ is the cosine metric (i.e. $D_c(a, b) = \frac{a^\top b}{\|a\|\|b\|}$).

### Using $\mathcal{E}^G_{unique}$ to estimate knowledge from $\mathcal{T}$

Finally when $\mathcal{E}^G_{unique}$ has been trained, the last step is straight-forward and only consists in adding a Multi-Layer-Perceptron on top of it and train all parameters on to estimate the new target knowledge.

### 3.4. About Using 2 Step Training

The two previous steps may be seen as training two parts of the same model and may be done at the same time, by training $\mathcal{E}^G_{MT} = g \circ \mathcal{E}^G_{unique}$ to directly fit the $h_t$. This approach may be considered as a weakly supervised multi-task (MT) learning. In practice we observe that it does not converge as well as our two-step approach, which is disentangling the processing relative to the tasks (first step) and to the domains (second step). It also is in line with several progressive approaches observed in the literature [36], achieving better model convergence by progressively training different parts of the model.

### 4. Experiments

This section validates our approach, first by detailing results of $\mathcal{E}^G_{unique}$ on the different source and target knowledge, compared to state-of-the-art dedicated approaches, then by running an ablation study on the different steps of our method and on the benefit to decompose the learning process into these steps.

#### Table 2. Number of detected/undetected faces for each domain.

<table>
<thead>
<tr>
<th>Domain</th>
<th># Detected Faces</th>
<th># Undetected Faces</th>
</tr>
</thead>
<tbody>
<tr>
<td>CelebA</td>
<td>202442</td>
<td>177</td>
</tr>
<tr>
<td>UTKFace</td>
<td>24018</td>
<td>98</td>
</tr>
<tr>
<td>FG-NET</td>
<td>1002</td>
<td>0</td>
</tr>
<tr>
<td>SFEW</td>
<td>1732</td>
<td>37</td>
</tr>
<tr>
<td>RAF</td>
<td>15330</td>
<td>9</td>
</tr>
<tr>
<td>ShoulderPain</td>
<td>48391</td>
<td>0</td>
</tr>
<tr>
<td>LFW</td>
<td>13233</td>
<td>0</td>
</tr>
<tr>
<td>$\mathcal{D}_{unsup}$</td>
<td>4.12 M</td>
<td>0.02M</td>
</tr>
</tbody>
</table>

### 4.1. Implementation details

We pre-process all faces from all domains with same operations. We use a private face detector to first detect and loosely crop the detected faces. If more than one face is detected, we select the closest to the image center. A landmark-based aligner is then applied on the detected faces, which are finally resized to $300 \times 300 \times 3$ pixels. Table 2 reports the number of images where a face was detected for each domain (first column). When a face is not detected (second column), we will consider during evaluation of our models (a) that the prediction is not correct when addressing classification problems (b) and that the prediction is the average between minimum and maximum values when dealing with regression problems. Note that the rest of the paper is following this rule.

As we use different tasks and domains, the evaluation protocols are changing for each knowledge. We follow the exact evaluation protocol used by the state-of-the-art dedicated approaches we are comparing to.

### 4.2. Experimental Validation

We first propose to evaluate our final approach $\hat{h}^{(t,d)}_G$ on target knowledge. Looking at the two last columns of Table 3, we observe that our approach adapt well to all target knowledge, even outperforming dedicated state-of-the-art approaches on AgeC-UTK, Expr-SFEW, Expr-RAF, Ethnic-UTK, AgeR-UTK. Note that the model for $\hat{h}^{(t,d)}_G$ counts only 2.2 millions parameters, which is almost always far less than the parameters used by other approaches, helping to better generalize and better fits real-world applications constraints.

#### Target Knowledge

Moreover, this level of performance with a lightweight model is difficult to achieve, as shown by the results of the CNN baseline, which has the same architecture than $\mathcal{K}^G_{unique}$ but is trained from scratch on the domain of the target knowledge. Indeed, this CNN is not always converging (e.g. on AgeR-UG with only 1002 images) and there is a huge gap between its scores and our approach, highlighting the clear benefit of transferring knowledge. A second stronger baseline called P-CNN is basically the clas-
Ablation study

4.3 Ablation study

As described in Section 3, \( G \) takes as input the concatenation of the 6 embeddings \( \hat{h}_i \), one per task (of size 5488 in our case). Then, the goal is to generate a compact representation \( \hat{h}_G \) (which size is fixed to 1024 in all the experiments) of the M=6 embeddings \( h_i \). As this dimension reduction step is a crucial operation of our approach, we propose to study the impact of different variations on this step.

Therefore, we first evaluate the ability of \( G \) to reconstruct the original \( h_i \) features from \( \hat{h}_G \). We report in Table 5 the normalized Root Mean Squared Error between each \( h_i \) and \( \hat{h}_i \) with diverse variations on \( G \). PCA stands for Principal Components Analysis, AE for standard Autoencoder, VAE for Variational Autoencoder and DAE for Denoising Autoencoder. Note that AE, VAE and DAE have the same number of parameters (the encoder \( G \) has 3 fully connected layers: 5488×3136, 3136×1792 and 1792×1024 and the decoder \( \tilde{G} \) is the symmetrical). The linear PCA baseline is easily outperformed by other approaches. The two best performing are the AE and the DAE. Surprisingly, the DAE does not generalize better than the AE, and is performing better only for the reconstruction of the ImageNet embedding, which is the less relevant in a context of face inputs. Still note that the gap of performance is not very important and all non-linear methods allow a fairly decent reconstruction of the embeddings.

### Table 3. Performance of diverse end-to-end approach evaluated on the target knowledge. All approaches are described in the methods or detailed in the text. Note that for the specific case of LFW, two types of results are reported. For result between parenthesis, before evaluation on LFW, the network is first trained (or fine-tuned) on a subset of 100,000 faces from VGGFace2 to predict Identity. The other LFW results are obtained directly from embedding \( \hat{h} \). Acc. is Accuracy and ER is the Average Error Rate.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Metric</th>
<th>CNN</th>
<th>P-CNN</th>
<th>( \mathcal{E}^G )</th>
<th>( \mathcal{E}^G_{MT} )</th>
<th>( \mathcal{E}_{\text{unique}}^G )</th>
<th>State-of-the-art (# parameters)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AgeC-UTK</td>
<td>Acc.</td>
<td>57.80 64.30</td>
<td>68.80 67.90 68.40</td>
<td>68.80 70.40</td>
<td>70.10 [20](5 M)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender-UTK</td>
<td>Acc.</td>
<td>90.00 96.50</td>
<td>97.20 93.15 94.4</td>
<td>97.58 97.90</td>
<td>98.23 [20] (5 M)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expr-SFEW</td>
<td>Acc.</td>
<td>22.00 53.60</td>
<td>53.10 50.20 52.20</td>
<td>54.00 57.2</td>
<td>55.40-58.14 [63] [1] (1.7 M/5 M)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expr-RAF</td>
<td>Acc.</td>
<td>69.00 85.70</td>
<td>86.51 82.40 87.20</td>
<td>83.70 89.3</td>
<td>86.77 [68] (35 M)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethnic-UTK</td>
<td>Acc.</td>
<td>62.20 84.90</td>
<td>89.20 81.20 82.20</td>
<td>88.2 91.20</td>
<td>90.10 [20] (5 M)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Identity-LFW</td>
<td>Acc.</td>
<td>98.90 99.65</td>
<td>99.10 94.27 (99.1)</td>
<td>98.92 (99.42)</td>
<td>99.65-99.87 [60] (25 M)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AgeR-UTK</td>
<td>MAE</td>
<td>6.38 4.70</td>
<td>4.39 4.70 4.42</td>
<td>4.24 4.05</td>
<td>5.39 [1] [21](8 M)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain-UNBC</td>
<td>MAE</td>
<td>0.89 0.56</td>
<td>0.54 0.64 0.72</td>
<td>0.56 0.52</td>
<td>0.51 [72] (0.0001 M)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AgeR-FG</td>
<td>MAE</td>
<td>11.10 3.10</td>
<td>2.85 3.95 3.12</td>
<td>2.95 3.05</td>
<td>2.81-3.00 [58] [4] (25 M)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Atrib-CelebA</td>
<td>ER</td>
<td>8.60 8.04</td>
<td>7.70 8.12 8.04</td>
<td>7.81 7.67</td>
<td>7.02 [9] (16 M)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 4. Performance of the original models \( \mathcal{E}^G_{(S_j, d_j)} \), of \( \mathcal{E}^G \) and of \( \mathcal{E}_{\text{unique}}^G \) on the original tasks.

<table>
<thead>
<tr>
<th>Knowledge</th>
<th>Metric</th>
<th>Original</th>
<th>( \mathcal{E}^G )</th>
<th>( \mathcal{E}_{\text{unique}}^G )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expr-AffectNet</td>
<td>Accuracy</td>
<td>63.5</td>
<td>64.0</td>
<td>64.4</td>
</tr>
<tr>
<td>Identity-LFW</td>
<td>Accuracy</td>
<td>99.65</td>
<td>99.1</td>
<td>99.42</td>
</tr>
<tr>
<td>Gender-UTK</td>
<td>Accuracy</td>
<td>96.5</td>
<td>97.2</td>
<td>97.9</td>
</tr>
<tr>
<td>Object-TinyImageNet</td>
<td>Accuracy</td>
<td>76.2</td>
<td>71.8</td>
<td>56.5</td>
</tr>
<tr>
<td>AgeR-FG</td>
<td>MAE</td>
<td>2.85</td>
<td>2.85</td>
<td>3.05</td>
</tr>
<tr>
<td>Atrib-CelebA</td>
<td>ER</td>
<td>8.05</td>
<td>7.7</td>
<td>7.67</td>
</tr>
</tbody>
</table>
Table 5. Reconstruction normalized RMSE obtained on $D_{unsup}$ test set by considered methods for each source knowledge embedding $h_i$ and on average.

<table>
<thead>
<tr>
<th>$h_i$</th>
<th>PCA</th>
<th>AE</th>
<th>VAE</th>
<th>DAE</th>
<th>$E_{MT}^{G}$</th>
<th>$E_{unique}^{G}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expr</td>
<td>0.28</td>
<td><strong>0.23</strong></td>
<td>0.26</td>
<td>0.26</td>
<td>0.55</td>
<td>0.39</td>
</tr>
<tr>
<td>Identity</td>
<td>0.25</td>
<td><strong>0.22</strong></td>
<td>0.23</td>
<td>0.26</td>
<td>1.12</td>
<td>0.35</td>
</tr>
<tr>
<td>Object</td>
<td>0.48</td>
<td>0.26</td>
<td>0.30</td>
<td><strong>0.25</strong></td>
<td>0.67</td>
<td>0.44</td>
</tr>
<tr>
<td>Age</td>
<td>0.23</td>
<td><strong>0.19</strong></td>
<td>0.22</td>
<td><strong>0.19</strong></td>
<td>0.77</td>
<td>0.38</td>
</tr>
<tr>
<td>Attrib</td>
<td>0.33</td>
<td><strong>0.27</strong></td>
<td>0.31</td>
<td>0.29</td>
<td>0.97</td>
<td>0.35</td>
</tr>
<tr>
<td>Gender</td>
<td>0.25</td>
<td><strong>0.25</strong></td>
<td>0.27</td>
<td>0.27</td>
<td>1.22</td>
<td>0.30</td>
</tr>
<tr>
<td>Average</td>
<td>0.31</td>
<td><strong>0.24</strong></td>
<td>0.27</td>
<td>0.25</td>
<td>0.92</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Table 5. Reconstruction normalized RMSE obtained on $D_{unsup}$ test set by considered methods for each source knowledge embedding $h_i$ and on average.

**Table 5.**

### Contribution of the distillation: $E_{MT}^{G}$ versus $E_{unique}^{G}$

The last two columns of the Table 5 allow to evaluate $E_{MT}^{G}$ and $E_{unique}^{G}$ by their ability to reconstruct the $h_i$ from the embedding they produced. To achieve such a reconstruction, we extracted the $h_{MT}$ and $h_{G}$ from all element of $D_{unsup}$ and then trained a decoder (with a similar architecture to $G$) to reconstruct the $h_i$.

We can observe that the reconstruction error of $E_{MT}^{G}$ is sometimes very high (e.g. for the identity embedding, explaining the low results of $h_{MT}$ when use for Identity-LFW in Table 3). Note that simply generating random embeddings (according to a uniform distribution in the range of the target embeddings) allows to achieve a normalized RMSE of 1.4. Therefore a score of 1.22 is almost random.

Without surprise, the reconstruction error of $E_{unique}^{G}$ is higher than the one obtained by $G$, which can be explained by both the distillation approximation between $h_G$ and $h_{G}'$ and the lower capacity of our small model. Nevertheless, a low reconstruction error is not a guarantee of better performance when trying to apply the $h_G$ representation in other tasks and domains. Thus, we propose to evaluate the very transfer learning operation using $E_{G}$ instead of $E_{unique}^{G}$. The first column of Table 3 reports the results obtained by $E_{G}$ on the target knowledge, while Table 4 provides the results on the source knowledge. On almost all knowledge, $E_{G}$ is outperformed by $E_{unique}^{G}$ by a significant margin.

Yet, if we try to understand from where the distillation improvement is coming, we can observe that only using a MLP on top of $h_{G}$ (without fine-tuning all parameters) does not bring improvements compared with using a similar MLP on top of $h_{G}$, some small changes in term of performance being observed. Thus, the distillation error has a limited impact when dealing with transfer. The bigger improvement is observed when using $h_{G'}^{(t,d)}$, meaning that we fine-tune all parameters of $E_{unique}^{G}$ on the new knowledge. It illustrates that the main contribution of the distillation lies in the reduced size of the encoder, giving it the ability to more easily adapt to new tasks and domains.

### Concatenation, reduction with $G$ or selection?

If we come back to the operator $G$, we discussed in section 3 the good reasons to reduce dimensionality of the $h_i$. We propose to empirically compare this choice to other methods and validate the intuition that selection may discard useful information. Thus, for each target knowledge, we propose to evaluate $G$ and several alternatives summarized in Table 6. We first evaluate a Multi Layer Perceptron taking as input the concatenation (Concat of all $h_i$) and having $G$ architecture. The MLP parameters are trained from scratch on the target knowledge. Our $G$ is then evaluated by training a small MLP on top of $h_{G}$ (extracted from all images of the target domain). We also compare the benefits brought by using a non-linear $G$ by also reporting the Principal Components Analysis (PCA) results.

Finally, another concurrent approach consists in selecting the embedding (resp. the combination of embedding) yielding the Best Transfer (BT) (resp. the Best Combination Transfer (BCT)). Several works [67, 40] proposed a method to automatically select such a best combination of embeddings. We choose here to reproduce this approach in a naive way, by brute force testing all the possible single transfers (BT) or combinations (by concatenation) of transfer (BCT) and reporting the best found results on each target knowledge in the two last columns of Table 6.

Table 6 shows that the Concat baseline is outperformed by all other approaches on almost all target knowledge. It is in line with what Zamir et al. [67] observed and it might be explained by the large dimensionality of the input of the Concat method and the limited size of some of the target domains. Then, it is interesting to see that the simple PCA method is nevertheless several times on par with the BT results, while $G$ performances are better than BT results and often better than the BCT. Thus, it validates the choice of a reduction of dimensionality applied on all embedding $h_i$, moreover illustrating the intuition developed in introduc-
Figure 2. Impact of the size of $h_G$ on the reconstruction error of $\tilde{G}$ and on the transfer for some of the target knowledge. Black cross is the optimal representation size for a given knowledge.

4.4. Impact of $h_G$ dimension

During the presentation of the method, we propose to choose the size of $h_G$ as the average size of $h_i$. In Figure 2 we discuss this choice, by showing that there is no optimal representation size for all knowledge. Nevertheless, a too small representation conduct to low results and a large representation implies a high-computational cost and lower performance.

4.5. Effect of Learning in 2 Stages

During the construction of our model, we argue that there was a benefit to adopt a 2 stage approach, disentangling the compression and the distillation steps. We already have observed the low ability of $h_{MT}$ to reconstruct the source embedding $h_i$. Yet, what is the real impact in term of transfer learning of this insufficient convergence.

Let’s study in details the columns dedicated to $\mathcal{E}_{MT}^G$ and $\mathcal{E}_{unique}^G$ in Table 3. We observe that if we are not fine-tuning the whole parameters of the model and only using the $h_{MT}$ and $h_G$ as features, the gap is significant on most of the source and target knowledge. Moreover, these results are correlated task by task to the reconstruction error observed on each $h_i$. Nevertheless $h_{MT}$ allows to achieve far better performance than the CNN baseline, still underlining the benefit of a pre-training, even when it is a noisy one.

4.6. Training Time

Finally, we report from Figure 3 the training time spent on $D_{unsup}$ and on the target domains for different approaches. Even if the training on $D_{unsup}$ is relatively long (almost 70 hours), the transfer learning step is then really faster than training a model from scratch. Moreover compared to the brute-force baseline of selecting all possible combinations (BCT), the total training time of the $\mathcal{E}_{unique}^G$ (taking the training time of $G$ into account) is divided by 4. We can also project ourselves in the case where we had much more knowledge to master. For instance, multiplying the number of target knowledge by 2 will imply a factor of time of 8 between $\mathcal{E}_{unique}^G$ and BCT.

5. Conclusions

This paper has introduced a novel approach to the problem of multi-source transfer learning, validated in the context of facial analysis. A unique and general model $\mathcal{E}_{unique}^G$, obtained by merging six different source knowledge, can be transferred on 9 different target knowledge. Building this model is done through two successive training steps. First, an autoencoder $G$ if trained to combine the hidden representations of the existing models into one single unifying embedding $h_G$. Then, distilling this model to a light-weight student CNN allows to reduce the number of parameters and improve the adaptation ability of the model. The approach was experimentally validated by an exhaustive ablation study and performances on par with state-of-the-art methods on the 15 different knowledge, with a single simple model. On overall, the approach provides an efficient way to obtain universal models compressing the knowledge included in several existing models, without loss in performance, allowing an easy exploitation in real-world applications.
References


