In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp.1378-1384, November 4–8, 2019 in The Venetian Macao, Macau,China.

Exploring Low-level and High-level Transfer Learning for Multi-task Facial Recognition with a Semi-supervised Neural Network

Pablo Barros[1], Erik Fliesswasser[1], Matthias Kerzel[1], and Stefan Wermter[1]

Abstract— **Facial recognition tasks like identity, age, gender, and emotion recognition received substantial attention in recent years. Their deployment in robotic platforms became necessary for the characterization of most of the non-verbal Human-Robot Interaction (HRI) scenarios. In this regard, deep convolution neural networks have shown to be effective on processing different facial representations but with a high cost: to achieve maximum generalization, they require an enormous amount of task-specific labeled data. This paper proposes a unified semi-supervised deep neural model to address this problem. Our hybrid model is composed of an unsupervised deep generative adversarial network which learns fundamental characteristics of facial representations, and a set of convolution channels that fine-tunes the high-level facial concepts for the recognition of identity, age group, gender, and facial expressions. Our network employs progressive lateral connections between the convolution channels so that they share the high-abstraction particularities of each of these tasks in order to reduce the necessity of a large amount of strongly labeled training data. We propose a series of experiments to evaluate each individual mechanism of our hybrid model, in particular, the impact of the progressive connections on learning the specific facial recognition tasks and we observe that our model achieves a better performance when compared to task-specific models.**

I. Introduction

Different facial recognition tasks, such as identifying a person, the expressed emotion, or a possible age group are problems that have been tackled extensively in the past years. Most of the non-verbal communication Human-Robot Interaction (HRI) scenarios are based on the observation of facial characteristics [1], and thus, it is of much importance that a robot is able to process this information. Currently, the most common way to process this information is through the use of deep neural networks trained for processing the individual tasks [2], [3], [4]. These networks, although very effective in addressing the individual tasks, demand a large amount of strongly labeled data to be trained, and thus, require extensive training and fine-tuning procedures. Also, maximizing generalization is a problem, as each of these models usually needs specific pre-processing or image acquisition processes, which reduces their portability when used in robotic platforms.

Multi-task learning from facial expressions became a popular topic in very recent years, and it is one of the ways to address the problem of training individual classifiers. Most of the multi-task learning solutions [5], [6], [7], [8], however, only extend the problem even further, as they are based on supervised learning from strongly labeled data, but now with the necessity of multiple labels. This reduces the availability of data to train such models substantially.

One of the most common alternatives to strongly supervised training is the use of unsupervised learning of facial characteristics [9]. With the recent advent of deep generative adversarial networks, the learning of facial characteristics became easier, as unlabeled data is easily crawled from the internet. Recent solutions based on Generative Adversarial Networks (GANs) learn facial representations in order to provide an entangled representation which can be easily modified. The entangled representation can also be used as a general representation of facial characteristics. Applications of such techniques in editing facial attributes [10], [11], age [12], [13], [14], and pose [15], [16] became very popular and effective. Most of these models, however, learn very specific characteristics of the faces which are enforced by the task at hand. The discriminator of a GAN establishes certain particularities of the entangled representation to solve the desired task. Nevertheless, the learned representations of a network trained to edit facial attributes are not suited to recognize emotion expressions, for example.

As a way to provide a general emotion representation, different solutions to transfer learning were proposed. In a transfer learning method, knowledge is transferred from a previously learned task, such as learning general facial representations, to a related new task for which labeled data is lacking. Pre-training a network on a domain for which data is highly available and fine-tuning it for another domain where data is scarce is a paradigm that has become popular in recent years [17], [18], [19]. In particular with face representation, the use of the VGG-Face [20] became the basis for several transfer learning solutions [21]. The VGG-Face is a variation of the VGG network trained with an enormous person identity recognition dataset. The learned representations, however, need to be entirely fine-tuned with different specific datasets to be able to recognize emotions [22], [23], age [24] and gender [25].

Most of the learned representations, be it by unsupervised learning, or supervised learning, are usually not well suited for dealing with multi-task learning. When fine-tuning a model for a new task, usually the model is over-specified for that specific task. Re-using it in an open-task scenario would be impossible, as it would require foreknowledge about the subsequent tasks. For facial expressions, it is known that humans process faces in different steps [26]: first, the general

[1] Knowledge Technology, Department of Informatics, University of Hamburg, Germany. {barros, 5fliess, kerzel, wermter}@informatik.uni-hamburg.de

facial representation is obtained, and later on, it is used to extract higher-abstraction concepts such as age, identity, and emotions. One way to simulate this processing pathway in a robotic platform would be to use a modular hybrid network which applies transfer learning from known tasks to novel tasks. The progressive neural connections [27] were designed for this purpose. They are lateral connections between task-specific convolution layers which are trained to modulate the learning of novel tasks, using prior-knowledge about previously known tasks. Combining the learning of general facial representations with transfer learning with progressive connections seems to be a promising way to tackle the challenge of data sparsity for deep neural networks in the context of facial recognition tasks.

This paper proposes a novel deep neural model for multi-task learning of identity, age, gender, and facial expressions. The network is divided into two modules: a face representation channel, and a series of convolution columns, with lateral progressive connections, for learning the specific characteristics of each of the recognition tasks. The face representation channel is pre-trained using an unsupervised adversarial learning and unlabeled training samples, so its filters are tuned to represent general facial characteristics. Each individual column is trained with specific datasets for each of the learning tasks. Our semi-supervised model presents state-of-the-art performance on the recognition of each of these tasks, and we provide a holistic insight on how progressive connections affect learning and general performance. Finally, we also present the visualization of learned filters which show that the transfer of high-level features of similar tasks is beneficial when the amount of labeled data is very small for the target task.

II. Transfer Learning of Facial Representations

Our semi-supervised network architecture is illustrated in Figure 1. The first module of the network, the face representation channel, is implemented as a series of convolution layers, and acts as a facial encoder. Connected to the facial encoder, a series of individual convolution columns is used to learn high-level features used in the recognition of each specific task.

To obtain a true general facial representation, we trained the facial encoder as part of a boundary equilibrium Generative Adversarial Network (BEGAN) [28]. The BEGAN achieved promising results on the encoding and decoding of facial representations, keeping the structural characteristics of the face while being able to reconstruct it with high-fidelity [29]. The facial encoder acts as a low-level general representation transfer learning to the specific recognition tasks. To achieve high-level concept transfer learning, each specific convolution column implements lateral progressive connections.

A. Facial encoder

The first step of the proposed semi-supervised strategy is the unsupervised training of the facial encoder. Therefore, we train a generative adversarial network (GAN) due to its ability to learn representations that are able to generate images with a higher fidelity when compared to more traditional unsupervised learning models, such as the auto-encoders [30]. However, the original GAN solutions usually tend to have unstable training progress and run the risk of mode collapse. These drawbacks are addressed by the boundary equilibrium GAN (BEGAN) [28] that introduces an enhanced objective function (Eq. 1) in which an equilibrium (Eq. 2) between the generator and the discriminator loss is explicitly forced. In this way, the training procedure is stabilized and the risk of mode collapse is reduced. Furthermore, the BEGAN implements an encoder/decoder architecture which is optimized for a minimal pixel-wise reconstruction error. hence, the goal is to match the distributions of the reconstruction error of generated and real images instead of matching the generated distribution with the input distribution. Formally, the new objectives for the discriminator loss $\mathcal{L}_D$ and the generator loss $\mathcal{L}_G$ are defined as follows:

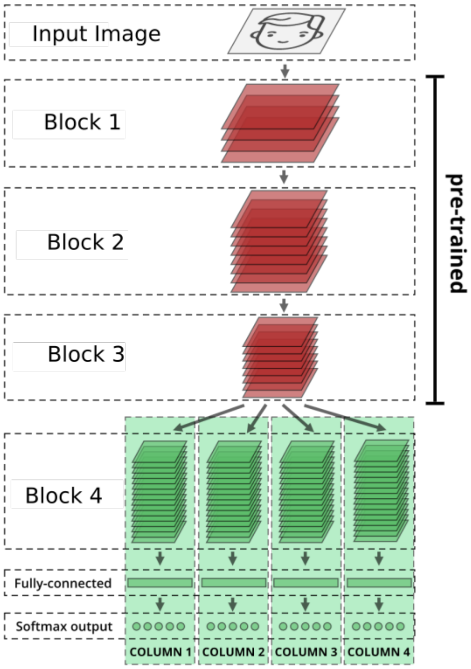


Fig. 1. Illustration of the proposed network that is composed of three pre-trained convolutional blocks and task-specific columns, each consisting of another convolutional block and fully-connected layers with a softmax output.

$$\begin{cases} \mathcal{L}_D = \mathcal{L}(x) - k_t \mathcal{L}(G(z)) & \text{for } \theta_D \\ \mathcal{L}_G = -\mathcal{L}_D & \text{for } \theta_G \\ k_{t+1} = k_t + \lambda_k(\gamma \mathcal{L}(x) - \mathcal{L}(G(z))) & \text{for each } t \end{cases} \tag{1}$$

where $\mathcal{L}(\cdot)$ is the reconstruction error, k_t the parameter ensuring the equilibrium in training step t, and λ_k its learning rate. $\gamma \in [0, 1]$ is a predefined ratio of the generator and discriminator loss defined as

$$\gamma = \frac{\mathbb{E}[\mathcal{L}(G(z))]}{\mathbb{E}[\mathcal{L}(x)]}. \tag{2}$$

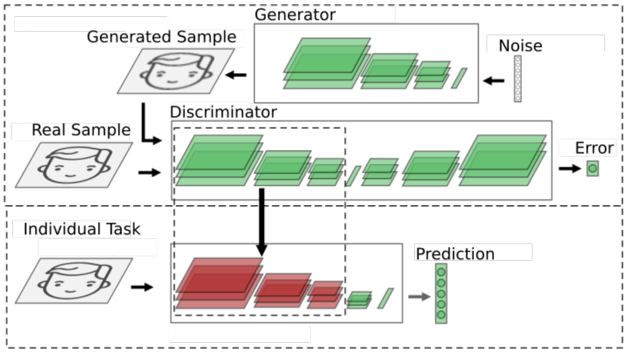


Fig. 2. Illustration of the proposed model showing the unsupervised face representation learning (top) and the subsequent supervised task learning (bottom).

After training the BEGAN, we use the network's discriminator as our facial encoder, as shown in Figure 2.

B. Progressive neural network

Progressive connections [27] are an effective way of reusing acquired knowledge while learning a sequence of tasks. The transfer is realized by lateral connections between the networks which are trained to solve certain tasks, and a new network which is trained to solve a similar task. Each task-specific network is considered as a column such that connections are established between new and old columns. As a result, each layer in a newly learned column is connected to the previous layer of its own column and to an adaption layer that entails input of the previous layers of all previous columns. The first specific column of our network is connected directly with the output of the facial encoder, and it is trained on a task via back-propagation, as usually done in convolution networks. For each of the other tasks, a new convolution column with lateral connections with the previous ones is added, as shown in Figure 3.

In order to prevent the network to unlearn the previous task, the weights of the first column are frozen and will not undergo any changes during the training of the second column. Formally, the hidden activation $h_i^{(k)} \in \mathbb{R}^{n_i}$ of layer i in column k with n_i the number of units at layer $i \leq L$ for a neural network having L layers is written as follows:

$$h_i^{(k)} = f\left(W_i^{(k)} h_{i-1}^{(k)} + \sum_{j<k} U_i^{(k:j)} h_{i-1}^{(j)} \right)$$

where $W_i^{(k)} \in \mathbb{R}^{n_i \times n_{i-1}}$ are the weights of layer i in column k, $U_i^{(k:j)} \in \mathbb{R}^{n_i \times n_j}$ are the lateral connections from layer $i-1$ of column j to layer i of column k, while h_0 is the input to the network and f the nonlinear activation function.

The problem of scalability for a growing number of subsequent tasks is solved by adaptation layers [27] denoted by *a* in Figure 3. Before feeding the activation of the previous column into a new column, lateral connections are combined and processed by 1×1 convolution in order to reduce their dimensionality [31].

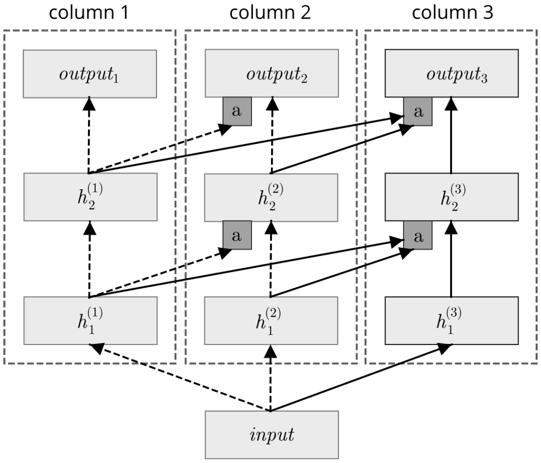


Fig. 3. Architecture of a 3-column progressive neural network. Lateral connections are established between the new column 3 and the previous columns 1 and 2 in order to have access to previously learned features. Adapted from [27].

C. Architectural design and parameters

Our model was designed in two steps. To obtain a general facial encoder, the topological structure of the BEGAN was implemented based on the work of Berthelot et al. [28]. The facial encoder receives as input a gray-scale image, and it is implemented as a series of four blocks, each of them containing a convolutional layer and a subsequent strided convolution for subsampling. A flattened fully-connected layer is used as the latent representation. The decoder network is also implemented as four blocks, each of them composed of convolution layers and upsampling layers. The generator architecture is built with the same topological configuration of the decoder.

The BEGAN parameters were optimized using a tree-structured Parzen estimator (TPE) [32], and the final configuration of the encoder/decoder and generator are listed in Tables I and II.

TABLE I
DETAILED PARAMETERS OF THE ENCODER LAYER.

# layers × type	filter size	# filters	output dim.
1 ×input	-	1 (channel)	128×128
2 × conv	3×3	64	128×128
1 × strided conv	3×3	64	64×64
2 × conv	3×3	128	64×64
1 × strided conv	3×3	128	32×32
3 × conv	3×3	128	32×32
1 × strided conv	3×3	128	16×16
3 × conv	3×3	256	16×16
1 × fully-conn.	-	-	1 ×128

Each of the specific convolution columns is implemented using the same topology which facilitates the training of the model, and produced a robust performance on all the evaluated tasks. It is composed of four blocks, the first two composed of convolution layers, and the last two composed of convolution layers followed by a max-pooling operator. Similar as the BEGAN, the parameters, exhibited in Table

TABLE II

DETAILED PARAMETERS OF THE DECODER AND GENERATOR LAYERS.

# layers × type	**filter size**	**# filters**	**output dim.**
1 × input	-	-	128
1 × fully-conn.	-	-	1024
2 × conv	3 × 3	16	8 × 8
1 × up-sampling	-	-	8 × 8
2 × conv	3 × 3	16	16 × 16
1 × up-sampling	-	-	32 × 32
3 × conv	3 × 3	16	32 × 32
1 × up-sampling	-	-	64 × 64
3 × conv	3 × 3	16	64 × 64
1 × up-sampling	-	1 (channel)	128 × 128

III of the specific channels were optimized using a TPE algorithm. The optimization maximized the recognition of all the tasks.

TABLE III

DETAILED PARAMETERS OF THE INDIVIDUAL TASK CHANNELS.

# layers × type	**filter size**	**# filters**	**output dim.**
1 × input	-	128	256 × 16 × 16
3 × conv	3 × 3	128	16 × 16
3 × conv	3 × 3	256	16 × 16
3 × conv	3 × 3	128	16 × 16
3 × conv	3 × 3	256	16 × 16
3 × conv	3 × 3	128	16 × 16
3 × conv	3 × 3	256	16 × 16
1 × max-pooling	3 × 3	2 ×2	8 × 8
3 × conv	3 × 3	128	16 × 16
3 × conv	3 × 3	256	16 × 16
1 × max-pooling	3 × 3	2 ×2	4 × 4
1 × fully-conn.	-	-	1 ×200
1 × dropout	-	-	-
1 × fully-conn.	-	-	2 / 7 / 8 / 50

III. EXPERIMENTAL SETUP

Our model is trained in a two-step process: first, we need to pre-train the facial encoder, and then the specific classifiers. To provide a holistic evaluation, and to be able to compare it with existing solutions, we use five different datasets:

- CelebFaces Attributes dataset [33] (**CelebA**): 202,599 face images of 10,177 different celebrities. Although this dataset contains annotations for facial attributes, it is treated as an unlabeled dataset and used for training the **facial encoder**.
- Facial Expression Recognition 2013 + dataset [34] (**FER+**): 35,887 images of faces expressing seven different basic **emotions** that are labeled accordingly.
- Labeled Faces in the Wild - Identity [35] (**LFW-identity**): A subset of the original Labeled Faces in the Wild dataset [36] containing 1,000 images of 50 different subjects, each providing 20 images and the corresponding **identity** labels.
- Labeled Faces in the Wild - Gender [37] (**LFW-gender**): Another subset of the Labeled Faces in the Wild dataset that has a perfectly balanced **gender** distribution and a total amount of 5,810 images.
- **Adience** [38]: 26,560 face images showing 2,284 different subjects with varying **age**. Each image is labeled with one of eight age groups.

Our model tackles two problems of general transfer learning, and provides two sets of experiments: one to evaluate the transfer of low-level feature representation through the facial encoders, and another to evaluate the transfer of high-level representations through the progressive connections.

In the first set of experiments, we train the facial encoder with the CelebA dataset. We then feed its output to one convolution column, freeze its layers, and train only the specific column for one specific task. We repeat the same experiment training the entire network with the FER+ dataset for emotion expression, the LFW-identity dataset for identity, the LFW-gender dataset for gender, and the Adience dataset for age recognition. We explore the impact of the transfer-learning by also providing experiments where we do not freeze the facial encoder layers, and re-train the entire network.

Our second set of experiments uses the same setup as the first set, but with the inclusion of the progressive connections. For a given task, the network column is trained with either one, two, or three pre-trained columns that are wired by lateral connections. We measure the performance of each of these combinations and discuss how they contribute to the general performance of the model.

In all of our experiments, each network is trained for 35 epochs with a batch size of 32 using the Adam optimizer with an initial learning rate of 0.0001 which is reduced by a factor of 0.5 once the loss stagnates for two consecutive training steps. As a performance measure, we compute the mean accuracy and standard deviation on the test set when running each experiment 10 times.

Finally, we provide a visualization of the learned features in order to explain how the two transfer learning strategies implemented by our model affect the multi-task recognition. We apply the guided backpropagation [39], the Grad-CAM, and the guided Grad-CAM [40] techniques to illustrate the learned convolution filters of the last convolution layer of the facial encoder and each individual column.

IV. RESULTS

The results of our first set of experiments, evaluating the low-level transfer learning, are exhibited in Table IV. We observe that the performance when re-training the facial encoder are slightly better than when freezing its weights, except for the gender task. These results indicate that the facial encoder learns general facial representations, which are, at some extent, important to each of the individual tasks. By re-training the facial encoder we are re-adapting the general characteristics to the ones found in each of the specific datasets, and thus, achieving better results.

Table V exhibits the results of our second set of experiments. We observe that in general the progressive connections improve the accuracy of the specific tasks. Most notable are the improvements on the emotion, age, and

TABLE IV

ACCURACY OF THE FACIAL RECOGNITION TASK LEARNING, REPORTED AS MEAN ACCURACY ± STANDARD ERROR.

Facial encoder	Emotion	Age	Gender	Identity
Freezing	83.66 ±0.27	84.01 ±0.46	**90.78** ±0.79	91.20 ±6.00
Re-training	**84.3** ±0.36	**85.12** ±0.45	90.21 ±0.31	**93.96** ±4.52

gender tasks, when compared to our first set of experiments. The identity task, however, shows worst results when using the progressive connections. In fact, every time the identity column is connected to any of the others, the results seem to decrease, reaching results which are lower than our first set of experiments.

By analysing the combination of the lateral connections which did not involve identity, we also observe that the best results, in all the experiments, are obtained when the gender column is present.

TABLE V

RESULTS OF PROGRESSIVE TRANSFER LEARNING EXPERIMENTS, REPORTED AS MEAN ACCURACY.

source	emotion	age	gender	identity
emotion	-	84.20	92.27	87.32
age	85.7	-	89.42	82.50
gender	86.8	86.36	-	89.43
identity	80.32	71.80	83.21	-
age, gender	**90.41**	-	-	92.54
age, identity	82.51	-	87.91	-
age, emotion	-	-	**93.02**	90.45
gender, identity	76.30	45.63	-	-
gender, emotion	-	**90.31**	-	**93.10**
emotion, identity	-	79.36	86.19	-
age, gender, identity	86.32	-	-	
emotion, gender, identity	-	81.65	-	
identity, age, emotion	-	-	90.45	-
emotion age, gender	-	-	-	89.27

When compared to state-of-the-art results in each specific task, our network achieves the best performance in all tasks except identity, as exhibited in Table VI. Georgescu et al. [41] present an ensemble of handcrafted features and convolution neural networks to achieve 87.76% of accuracy on the FER+ dataset. Our network presents an improvement of almost 3% accuracy when implementing the progressive connections coming from the age and gender columns. A similar result appears when comparing our model with the works of Levi et al [3], which provides a supervised deep neural network for age classification, and with the work of Jalal et al. [37], who implement a deep convolution network for gender recognition. An ensemble of deep neural networks which learn embedded representations [35] outperform our best results on identity recognition by at least 2% of accuracy, but demands a much more extensive training procedure than ours.

TABLE VI

ACCURACY OF THE FACIAL RECOGNITION TASK LEARNING, REPORTED AS MEAN ACCURACY ± STANDARD ERROR.

Model	Task	Accuracy
Ensemble of features[41]	Emotion	87.76
Deep CNN [3]	Age	84.70
Deep CNN[37]	Gender	87.95
OpenFace [35]	Identity	95.00
age, gender	Emotion	90.41
gender, emotion	Age	90.31
age, emotion	Gender	93.02
gender, emotion	Identity	93.20

V. DISCUSSIONS

Our proposed model takes advantage of two transfer learning processes: low-level features with the facial encoder, and high-level features with the progressive connections. Our experiments demonstrate that the facial encoder architecture is able to learn robust facial representations which are shared among the individual tasks. The strong evidence of this assumption is the small increase on the recognition rate when the facial encoder is re-trained. The freezing of the facial encoder, thus, guarantees that each convolution column will receive the same facial representation. It is the role of each column to learn the high-level features for each individual task.

Regarding the high-level feature transfer, we observe that in most of the cases the progressive connections actually improve the recognition of the other tasks. This was clearer when we performed the experiments with connections coming from more than one column. We reache the best performances of our model following this scheme. Interestingly, the identity column seems to not have contributed at all to the other tasks. Every time the identity column is used as a source column, the performance of the target column drops considerably. Also, the identity column is the only one which does not surpass the benchmark results.

To help to understand better the impact of the identity column, we performed a visualization experiment in each of the last convolution layers of the individual columns, and illustrate it in Figure 4. While the identity recognition network only highlights the mouth and its surrounding regions (left-most column), the columns trained with the other tasks highlight more general regions, such as eyes, cheeks, and nose areas.

Based on the visualizations, we assume that the high-level features involved in identity recognition are not shared by the other networks. Thus, the identity column does not contribute to the other tasks. Likewise, the features learned by the other columns also do not contribute to the identity recognition task.

One of the most important advantages or our model, when compared with other state-of-the-art solutions, is its inherited modularity and scalability. By using the low-level and high-level transfer learning, we were able to embed the prior knowledge of facial representation into the specific

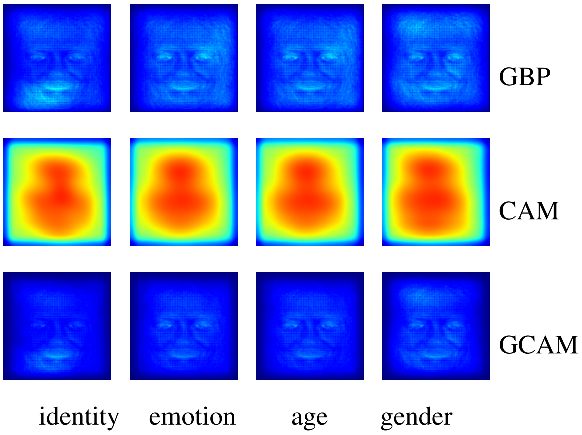
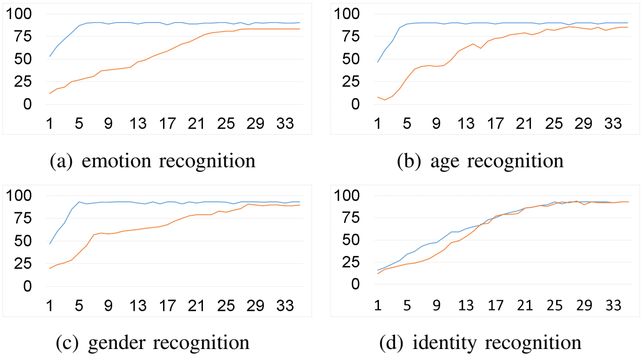


Fig. 4. Visualization of learned features of the regular identity recognition (left column) and the three 1-column ProgNet configurations shown in the columns two to four (from left). The first row shows the output of the guided backpropagation (GBP), the second row shows the output of the Grad-CAM (CAM), and the last row shows the fusion of both techniques called guided Grad-CAM (GCAM).



(a) emotion recognition

(b) age recognition

(c) gender recognition

(d) identity recognition

Fig. 5. Learning curves of each individual recognition task with (blue line) and without (organge line) the progressive connections. The plots depict the accuracy (%) and the number of epochs.

knowledge learned by the individual channels for a novel task, without the overhead of re-training the entire model. This alone is a much important feature for models aiming to be deployed in robotic platforms, where fast adaptability to novel scenarios is mandatory [42].

In order to illustrate the fast adaptability of our model, we plot, in Figure V, the convergence curves for each task trained with and without the progressive connections. We observe that every time the lateral connections are present, except for the identity task, the network presents a much faster convergence, in terms of less training epochs, to achieve the maximum performance. We assume the nature of the features which are shared from the other tasks also affect the learning curve for identity recognition. In the end, this specific task learning behavior is the same with and without the high-level features transfer learning strategy.

Similar behavior on fast convergence was observed in the application of the progressive connections on the transfer learning of auditory information [43]. We also observe that almost in all the cases, the combination of low-level and high-level feature sharing allow that each task, besides the identity one, starts with higher accuracy than when no features are shared.

To be able to adapt to new tasks as quickly as our model does from the beginning of the training for each task, and on the faster learning curve, allow our proposed solution to have the potential to be successful in real-world robotic applications.

VI. Conclusion and Future Work

In this work, we propose a novel semi-supervised deep learning architecture for multi-task recognition of emotion expressions, age, gender, and identity. Our network implements two strategies for transfer learning: the learning of low-level features based on unsupervised learning of a boundary equilibrium generative adversarial network (BEGAN), and the supervised fine-tuning of high-level features through the use of lateral progressive connections.

We evaluate our model's performance in two sets of experiments, one to evaluate each of the transfer learning strategies. Our results demonstrate the proposed model does learn general facial representations which are common for all the recognition tasks. Our second set of experiments demonstrates that the lateral progressive connections improve the recognition of individual tasks, except for the identity recognition task, making our model achieve state-of-the-art results.

To understand better why the identity recognition column does not contribute to the other columns, neither is affected by them, we also performed a visualization experiment where we analyzed the learned representations. We conclude that the network learns different high-level features for the identity task, which are not compatible with the other tasks.

In order to underline the observations made in this work and to draw stronger conclusions regarding multi-task facial representation learning, experiments with other datasets for the same facial recognition tasks should be conducted.

The potential of the proposed network and training strategy could be explored regarding their suitability for robotic applications. Due to the highly reduced amount of trainable parameters, i.e. a reduced training time, this approach seems to be very appropriate for task-specific fine-tuning on Human-Robot Interaction (HRI) scenarios.

Acknowledgment

The authors gratefully acknowledge partial support from the German Research Foundation DFG under project CML (TRR 169).

References

[1] E. Cha, Y. Kim, T. Fong, M. J. Mataric, *et al.*, "A survey of nonverbal signaling methods for non-humanoid robots," *Foundations and Trends® in Robotics*, vol. 6, no. 4, pp. 211–323, 2018.

[2] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1701–1708.

[3] G. Levi and T. Hassner, “Age and gender classification using convolutional neural networks,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2015, pp. 34–42.

[4] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, “Training deep networks for facial expression recognition with crowd-sourced label distribution,” in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ser. ICMI 2016. New York, NY, USA: ACM, 2016, pp. 279–283.

[5] H. Han, A. K. Jain, F. Wang, S. Shan, and X. Chen, “Heterogeneous face attribute estimation: A deep multi-task learning approach,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 11, pp. 2597–2609, 2018.

[6] X. Yin and X. Liu, “Multi-task convolutional neural network for pose-invariant face recognition,” *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 964–975, 2018.

[7] R. Ranjan, V. M. Patel, and R. Chellappa, “Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 121–135, 2019.

[8] J. Cao, Y. Li, and Z. Zhang, “Partially shared multi-task convolutional neural network with local constraint for face attribute learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4290–4299.

[9] S. Datta, G. Sharma, and C. Jawahar, “Unsupervised learning of face representations,” in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 2018, pp. 135–142.

[10] B. Huang, W. Chen, X. Wu, C.-L. Lin, and P. N. Suganthan, “High-quality face image generated with conditional boundary equilibrium generative adversarial networks,” *Pattern Recognition Letters*, vol. 111, pp. 72–79, 2018.

[11] G. Zhang, M. Kan, S. Shan, and X. Chen, “Generative adversarial network with spatial attention for face attribute editing,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 417–432.

[12] Z. Wang, X. Tang, W. Luo, and S. Gao, “Face aging with identity-preserved conditional generative adversarial networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7939–7947.

[13] H. Yang, D. Huang, Y. Wang, and A. K. Jain, “Learning face age progression: A pyramid architecture of gans,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 31–39.

[14] S. Palsson, E. Agustsson, R. Timofte, L. Van Gool, and K. ESAT, “Generative adversarial style transfer networks for face aging,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2084–2092.

[15] L. Q. Tran, X. Yin, and X. Liu, “Representation learning by rotating your faces,” *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[16] K. Cao, Y. Rong, C. Li, X. Tang, and C. Change Loy, “Pose-robust face recognition via deep residual equivariant mapping,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5187–5196.

[17] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, “Cnn features off-the-shelf: An astounding baseline for recognition,” in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, ser. CVPRW ’14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 512–519.

[18] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, “Decaf: A deep convolutional activation feature for generic visual recognition,” in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ser. ICML’14. JMLR.org, 2014, pp. I–647–I–655.

[19] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, “Deep learning for emotion recognition on small datasets using transfer learning,” in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ser. ICMI ’15. New York, NY, USA: ACM, 2015, pp. 443–449.

[20] O. M. Parkhi, A. Vedaldi, A. Zisserman, *et al.*, “Deep face recognition.” in *BMVC*, vol. 1, no. 3, 2015, p. 6.

[21] K. Sundararajan and D. L. Woodard, “Deep learning for biometrics: A survey,” *ACM Computing Surveys (CSUR)*, vol. 51, no. 3, p. 65, 2018.

[22] X. Xia, J. Liu, T. Yang, D. Jiang, W. Han, and H. Sahli, “Video emotion recognition using hand-crafted and deep learning features,” in *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*. IEEE, 2018, pp. 1–6.

[23] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang, and Y. Zong, “Multi-cue fusion for emotion recognition in the wild,” *Neurocomputing*, 2018.

[24] L. Chen, C. Fan, H. Yang, S. Hu, L. Zou, and D. Deng, “Face age classification based on a deep hybrid model,” *Signal, Image and Video Processing*, pp. 1–9, 2018.

[25] A. Manyala, H. Cholakkal, V. Anand, V. Kanhangad, and D. Rajan, “Cnn-based gender classification in near-infrared periocular images,” *Pattern Analysis and Applications*, pp. 1–12, 2018.

[26] P. C. Quinn, K. Lee, and O. Pascalis, “Face processing in infancy and beyond: The case of social categories,” *Annual review of psychology*, vol. 70, pp. 165–189, 2019.

[27] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, “Progressive neural networks,” *CoRR*, vol. abs/1606.04671, 2016.

[28] D. Berthelot, T. Schumm, and L. Metz, “BEGAN: boundary equilibrium generative adversarial networks,” *CoRR*, vol. abs/1703.10717, 2017.

[29] J. Hah, W. Lee, J. Lee, and S. Park, “Information-based boundary equilibrium generative adversarial networks with interpretable representation learning,” *Computational Intelligence and Neuroscience*, vol. 2018, 2018.

[30] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” *CoRR*, vol. abs/1511.06434, 2015.

[31] M. Lin, Q. Chen, and S. Yan, “Network in network,” *CoRR*, vol. abs/1312.4400, 2013.

[32] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, “Algorithms for hyper-parameter optimization,” in *Advances in neural information processing systems*, 2011, pp. 2546–2554.

[33] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[34] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, “Training deep networks for facial expression recognition with crowd-sourced label distribution,” in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 279–283.

[35] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, “Openface: A general-purpose face recognition library with mobile applications,” CMU-CS-16-118, CMU School of Computer Science, Tech. Rep., 2016.

[36] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[37] A. Jalal and U. Tariq, “The lfw-gender dataset,” in *Computer Vision – ACCV 2016 Workshops*, C.-S. Chen, J. Lu, and K.-K. Ma, Eds. Cham: Springer International Publishing, 2017, pp. 531–540.

[38] E. Eidinger, R. Enbar, and T. Hassner, “Age and gender estimation of unfiltered faces,” *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2170–2179, Dec 2014.

[39] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, “Striving for simplicity: The all convolutional net,” in *ICLR (workshop track)*, 2015.

[40] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, “Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization,” *CoRR*, vol. abs/1610.02391, 2016.

[41] M.-I. Georgescu, R. T. Ionescu, and M. Popescu, “Local learning with deep and handcrafted features for facial expression recognition,” *arXiv preprint arXiv:1804.10892*, 2018.

[42] T. M. Moerland, J. Broekens, and C. M. Jonker, “Emotion in reinforcement learning agents and robots: a survey,” *Machine Learning*, vol. 107, no. 2, pp. 443–480, 2018.

[43] T. Moriya, R. Masumura, T. Asami, Y. Shinohara, M. Delcroix, Y. Yamaguchi, and Y. Aono, “Progressive neural network-based knowledge transfer in acoustic models,” in *Proceedings, APSIPA Annual Summit and Conference*, vol. 2018, 2018, pp. 12–15.