Efficient Facial Feature Learning with Wide Ensemble-based Convolutional Neural Networks

Henrique Siqueira, Sven Magg and Stefan Wermter
Knowledge Technology
Department of Informatics, University of Hamburg
Vogt-Koelln-Str. 30, 22527 Hamburg, Germany
{siqueira, magg, wermter}@informatik.uni-hamburg.de

Abstract

Ensemble methods, traditionally built with independently trained de-correlated models, have proven to be efficient methods for reducing the remaining residual generalization error, which results in robust and accurate methods for real-world applications. In the context of deep learning, however, training an ensemble of deep networks is costly and generates high redundancy which is inefficient. In this paper, we present experiments on Ensembles with Shared Representations (ESRs) based on convolutional networks to demonstrate, quantitatively and qualitatively, their data processing efficiency and scalability to large-scale datasets of facial expressions. We show that redundancy and computational load can be dramatically reduced by varying the branching level of the ESR without loss of diversity and generalization power, which are both important for ensemble performance. Experiments on large-scale datasets suggest that ESRs reduce the remaining residual generalization error on the AffectNet and FER+ datasets, reach human-level performance, and outperform state-of-the-art methods on facial expression recognition in the wild using emotion and affect concepts.

Introduction

"We get resourcefulness from having many resources; not from having one very smart one" (Minsky 2014). In machine learning, ensemble methods refer to a set of models where an inference is made collectively based on individual predictions (Dietterich 2000). A well-trained ensemble can reduce the remaining residual generalization error, which results in predictions being more accurate than any single model in the ensemble. Traditional ensemble (TE) methods are built by independently training several models on the same or different data. They can be composed of a single type of machine learning method such as an ensemble of neural networks (Hansen and Salamon 1990), but the diversity is often higher when an ensemble is built from a library of different methods (Caruana et al. 2004).

At present, ensembling of deep networks is an important resource but requires high computational power. To make this training-intensive technology accessible to everyone, recent studies have explored ways to reduce redundancy in ensembling. Meshgi, Oba, and Ishii (2018) have exploited concepts from active learning to reduce training time and redundancy. Rather than using the whole dataset for training, their ensemble method is trained on the most informative samples that maximize learning based on the query by committee algorithm (Seung, Opper, and Sompolinsky 1992).

Another approach adopted a divide-and-conquer strategy (Li et al. 2019) where the input space is decomposed into multiple regions, and each region is used to train one convolutional neural network of the ensemble. Despite their progress on reducing redundancy, their approaches fall within the "explicit" ensemble methods, i.e., consist of independent models. Therefore, redundancy of low-level visual features is still high, and unnecessary computational resources have to be allocated for processing such features.

In the so-called "implicit" ensemble methods, a single network may generalize as well as an ensemble by distilling its knowledge (Hinton, Vinyals, and Dean 2015). By training a convolutional neural network (CNN) with the outputs of an ensemble of CNNs, Shen, He, and Xue (2019) have reduced inference time and redundancy while maintaining generalization power and similar intermediate representations under an adversarial training strategy. However, training time is greatly increased with their approach since a trained traditional ensemble is a fundamental pre-requisite.

Ensemble with Shared Representations (ESR), proposed in our previous work (Siqueira et al. 2018), offers the best of the two worlds. It is neither a fully implicit nor a fully explicit ensemble method. As depicted in Figure 1, the shared layers represent the implicit part. They are responsible for the reduction of redundancy, training, and inference time. The low-level features learned by them are shared with the ensemble of convolutional branches. The latter characterizes the explicit part and carries the diversity of the ensemble. The level to start the ensemble of branches plays an important role in the computational load and generalization power as well as for redundancy and diversity. However, the effect of the branching level is still an open question that needs careful analysis. In the context of facial expression recognition, for instance, starting branching too early (level 1) may result in high redundancy of low-level facial features where all branches have to learn skin textures and so forth. On the

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

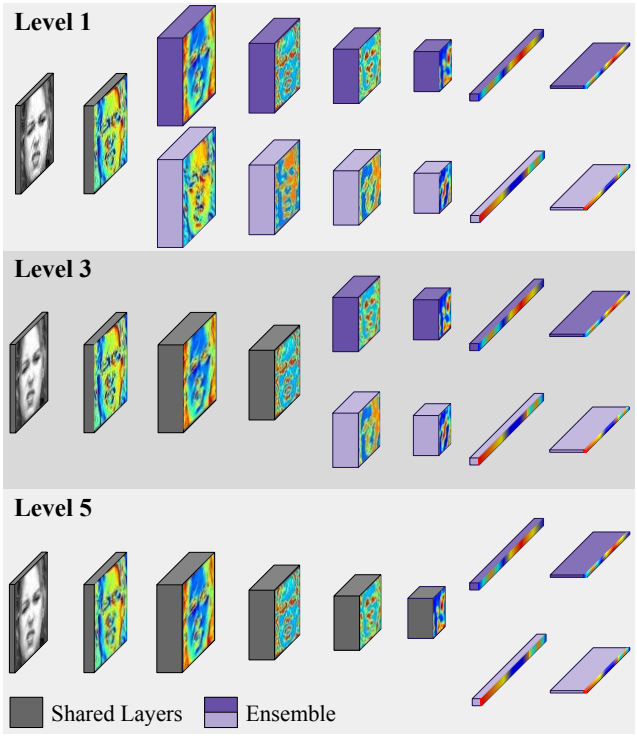

Figure 1: Ensemble with Shared Representations (ESR). Illustration of the experiments to investigate the effect of branching level on computational load and generalization.

other hand, branching too late may drastically decrease diversity in the ensemble where features from the shared layers no longer correspond to spatial facial features (level 5). We hypothesize that the optimal branching level may be located between the extremes, where the abstraction level of the facial features is high including smiling and frowning but have yet to be encoded into emotion concepts.

Another aspect that needs further understanding is the scalability of ESRs to large-scale datasets of facial expressions of emotion. Can ESRs reduce the remaining residual generalization error when training data is abundant? After reviewing prior work on facial expression recognition, we address these questions. In this paper, the effects of varying the branching level are extensively examined, quantitatively and qualitatively, first, on a small-scale but clean and well-structured dataset of facial expressions *in the lab*. Subsequently, experiments using a single GeForce GTX 1080 on large-scale benchmarks for facial expression recognition *in the wild* demonstrate the affordability and scalability of ESRs, followed by conclusions and future research. For reproducibility purposes, source code of our experiments, the ESR implementation in PyTorch, trained networks and supplementary material are available in our GitHub repository[1].

[1]Source code: https://github.com/knowledgetechnologyuhh/Efficient-Facial-Feature-Learning-with-Wide-Ensemble-based-Convolutional-Neural-Networks

Prior Work on Facial Expression Recognition

Early approaches for automatic facial expression recognition have followed the general pipeline to tackle computer vision problems, which consist of pre-processing the facial images, appearance and/or geometric hand-crafted feature extraction and, in the final stage, the classification of such features (Tian, Kanade, and Cohn 2005). These methods are usually fast and accurate in indoor environments but frequently drop in performance under real-world conditions (Mollahosseini, Hasani, and Mahoor 2019).

The rapid progress in deep learning motivated researchers to develop facial expression recognition systems using deep neural networks. Since these networks can automatically learn features from data, hand-feature engineering was left out in the pipeline. Besides that, feature learning allows deep networks to learn a broader range of facial features than earlier approaches, including variation to rotations, and illumination changes. Indeed, as investigated by Khorrami, Paine, and Huang (2015), it has turned out that the features learned by a CNN trained for facial expression recognition reflect the facial features of emotion suggested by the psychologist Paul Ekman during his study of universal facial expressions of emotion (Ekman 1989). Recent approaches rely on well-established networks for object recognition such as AlexNet, MobileNet, ResNet, and VGGNet (Barsoum et al. 2016; Hewitt and Gunes 2018; Mollahosseini, Hasani, and Mahoor 2019). In visual perceptual tasks, certain features previously learned can be transferred among related tasks and the use of pre-trained networks often speed up learning and culminate in better accuracy than training them from scratch. These approaches represent the state of the art in the datasets utilized in our experiments (for a review, see (Poria et al. 2017)).

Ensembles with Shared Representations

Ensembles with shared representations exploit the fundamental properties of convolutional networks. A convolutional layer learns *local patterns* from the previous layer by convolving small filters over its input feature space (Chollet 2018). Thus, the patterns learned by convolutional layers are translation-invariant. Another property is the capability to learn spatial hierarchies of patterns by stacking multiple convolutional layers. Consider the intermediate representations exhibited in Figure 1. Early layers learn simple and local visual patterns such as oriented lines, edges, and colors. These low-level abstractions of input space are the reason for early feature maps resembling a face with emphasis on certain regions. Subsequent layers hierarchically combine local patterns from previous layers into increasingly complex concepts such as nose, mouth, and eyes. The level of abstraction increases as you go deeper into the network until the point where feature maps are no longer visually interpretable. Finally, the last layer encodes these representations into semantic concepts, for instance, concepts of emotion.

These properties are the foundations of ESRs and play a crucial role in reducing redundancy of visual features in the ensemble. An ESR consists of two building blocks. (1) The base of the network (gray blocks in Figure 1) is an array of convolutional layers for low- and middle-level feature learn-

ing. (2) These informative features are then shared with independent convolutional branches (purple blocks) that constitute the ensemble. From this point, each branch can learn distinctive features while competing for a common resource - the shared layers. This competitive training emerges from the minimization of a *combined loss function* defined as the summation of the loss functions of each branch as follows:

$$L_{esr} = \sum_{b} \sum_{i} L[P(f(x_i) = y_i | x_i, \theta_{shared}, \theta_b), y_i], \quad (1)$$

where b denotes the branch index, (x_i, y_i) random samples from the training set, θ_{shared} the parameters of the shared layers from the base of the network that acts as a regularizer for ESRs, and θ_b the parameters of a convolutional branch that composes the ensemble.

Because novel convolutional branches are added in sequence while training, as outlined in Algorithm 1, the shared layers turn out to be an efficient transfer learning mechanism that guides and accelerates learning as the ensemble grows. Besides that, the shared representations are conditioned to learn features that are suitable to different branches in the ensemble due to the inductive transfer learning from the combination of multiple loss functions from each convolutional branch. During inference time, a given input is classified by the ensemble through a collective decision such as plurality and majority voting.

Algorithm 1: Training ESRs.

```
initialize the shared layers with θ_shared
for b to maximum ensemble size do
    initialize the convolutional branch B_b with θ_b
    add the branch B_b to the network ESR
    sample a subset D' from a training set D
    foreach mini-batch (x_i, y_i) ~ D' do
        perform the forward phase
        initialize the combined loss function L_esr to
         0.0
        foreach existing branch B_b' in ESR do
            compute the loss L_b' with respect to B_b'
            add L_b' to L_esr
        end
        perform the backward phase
        optimize ESR
    end
end
```

Experimental Datasets

Over the last two decades, a number of datasets of facial expressions have been collected for research on affective computing (Mollahosseini, Hasani, and Mahoor 2019). Among the attributes that characterize these datasets (e.g. the number of subjects and representations of emotion), the nature of the facial expressions is critical for developing and assessing automatic facial expression recognition systems.

Some of the datasets rely on posed or simulated facial expressions of emotion. They are supported by Ekman and Friesen's work (Ekman and Friesen 1976; Ekman 1989) on universals in facial expressions of emotion. The arguments about universality suggest that when we feel certain emotions, some facial movements manifest regardless of age, culture, race, or sex. For example, when you are angry in a traffic jam, you scowl; when you are happy after an acceptance notification, you smile. These visible facial movements have been mapped latter to the Facial Action Code (FAC) (Ekman and Friesen 1976), where every single appearance change (action unit, AU) was categorized. These datasets are occasionally called *in-the-lab* datasets. As the name states, facial images are collected in controlled indoor environments where experimental variables (e.g. scene lighting and camera-view points) are accurately adjusted. They usually provide clean and high-quality data. Although posed emotional expressions are considered more expressive than natural expressions in everyday life (Koolagudi and Rao 2012), the datasets are well structured and carefully annotated from emotions to FAC (Lucey et al. 2010).

On the other end of the spectrum, there are the *in-the-wild* datasets with spontaneous facial expressions. Over a century, since Charles Darwin published *The Expression of the Emotions in Man and Animals* (Darwin 1872), the universality of emotional expressions has been called into question by distinguished psychologists including William James (James 1884; James et al. 1890), James A. Russell (Russell 2003) and Lisa F. Barrett (Barrett and Russell 2015; Barrett 2017; Gendron, Crivelli, and Barrett 2018). Their theses converge to the same conclusion: diversity of emotional expressions is the norm, not the uniformity. According to James et al. (1890), any categorization of an emotional expression can be seen "as true and as 'natural' as any other". Nevertheless, Russell argues for the minimum universality in his core affect theory, where emotions are described in an orthogonal dimensional space of arousal and valence levels. Therefore, even though we cannot claim that *in-the-wild* datasets contain emotional facial expressions, they do provide large and rich data of facial configurations captured in a vast range of environmental conditions. These variations are crucial to develop robust facial expression recognition systems. In most cases, the data is gathered from films or the Internet and annotated based on affect concepts and/or emotion concepts (Mollahosseini, Hasani, and Mahoor 2019).

We trained and tested the ensemble with shared representations on in-the-lab and in-the-wild datasets for a couple of reasons. The former allows us to evaluate ESRs' inference performance when training data is scarce and to conduct a descriptive analysis of their predictions based on the FAC system. On the other hand, the latter permits us to asses the scalability of ESRs to large-scale datasets and to test their inference performance in a more challenging scenario which includes, among other aspects, a vast intraclass variation, rotations, occlusions, and a heavily imbalanced label distribution. Together, they provide evidence on how flexible and robust ensembles with shared representations are in dealing with different shortcomings on facial expression recognition in the lab and in the wild. A few samples from the datasets used in our experiments are shown in Figure 2 and the technical details are described as follows.

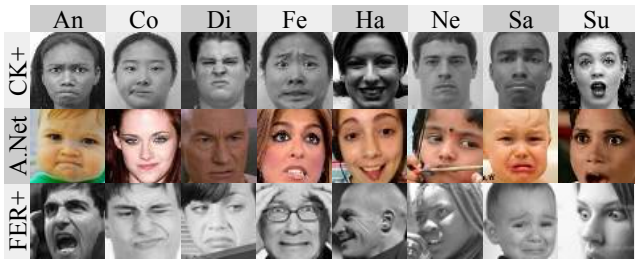

Figure 2: Experimental datasets. Extended Cohn-Kanade (CK+), AffectNet and FER+, from the top to the bottom.

In-the-Lab Dataset

The Extended Cohn-Kanade (CK+) dataset (Lucey et al. 2010) has been vastly used to develop action unit detection and facial expression recognition systems. 123 subjects between 18 and 50 years old from different races, sex, and ethnic groups were told to portray a series of facial configurations based on FAC. The onset facial expressions were recorded from frontal and 30-degree camera-view points, and their peaks were carefully annotated and validated in terms of 30 action units and 8 discrete emotion concepts.

In-the-Wild Datasets

AffectNet (Mollahosseini, Hasani, and Mahoor 2019) is the largest dataset of facial expressions in the wild publicly available. It contains more than one million images retrieved from the Internet using emotion keywords from different languages, where half of them were manually annotated by human experts using 8 discrete emotions, arousal and valence levels. In addition to its heterogeneity, the heavily imbalanced label distribution (e.g., contempt constitutes only 1% of the annotated images) and the strong baselines pose a real challenge for the affective computing community.

FER+ (Barsoum et al. 2016) derives from the re-annotation of the Facial Expression Recognition 2013 (FER-2013) dataset (Goodfellow et al. 2015) due to the originally high degree of noise presented in the annotations. FER-2013 was created by querying facial images from Google's image search engine using 184 emotion keywords. Each of the 35,887 facial images was then re-labeled by 10 annotators using crowd-sourcing, and the contempt category was added to the dataset as one of the possible 8 emotion labels.

Redundancy and Diversity Analysis

We start this section describing the methodology adopted to explore the impact of the branching level on redundancy, and diversity of ESRs. After discussing training strategies and architectural design, we present quantitative results on computational load, redundancy and recognition performance. We conclude this section by presenting evidence that ESRs converge faster than a TE while preserving diversity, by analyzing convergence graphs and saliency maps via Grad-CAM (Selvaraju et al. 2017) at different training milestones.

Methodology

We followed the subject-independent 10-fold cross-validation for comparison purposes based on our previous work (Siqueira et al. 2018). First, we extracted the first and last three frames from each sequence on CK+, converted them to gray-scale, cropped the faces using the Viola and Jones's algorithm (2004), and resized the facial images to 96 x 96 pixels. The first frame was labeled as neutral, whereas the last three frames received one of the seven basic emotion labels. Subsequently, the images were separated into 10 folds according to the subject's id available in the metadata. Each fold was populated with facial images from a subject by iterating the subject id and the fold id, which resulted in 12 subjects and 130.8 facial images on average for each fold. With the folds populated, we run the experiment 10 times. In each trial t, we selected fold-(t) for testing, fold-$(t+1)$ for validating, and only the first four folds from the remaining eight folds for training, i.e., 523.2 images on average on the training set.

Training ESRs at Different Branching Levels

How does the branching level affect computational load, redundancy and recognition performance on ESRs? This research question was addressed by training several ESRs at different branching levels and analyzing the impact on those aspects. Two baselines were defined according to our previous research (Siqueira et al. 2018). After an exhaustive search among different convolutional architectures and training strategies, the network with the best mean test accuracy on CK+ was selected as the first baseline. The network comprises five convolutional layers, each followed by a batch normalization layer. A max-pooling layer was also added after the second, third, and fourth batch normalization layers. On top, a global average pooling layer transforms the last feature maps into a vector and forwards it to the dense output layer for facial expression recognition. The ReLU activation function was applied after batch normalization layers as suggested by Ioffe and Szegedy (2015). A detailed architectural diagram of the network used in our experiments is presented at the top of Figure 3. The second baseline is a traditional ensemble with four of such networks.

The *single network* was trained on four folds using stochastic gradient descent (SGD) to minimize the cross-entropy loss, whereas different training strategies using SGD were tested to build ensembles with complementary representations of the data. Given the small number of training samples, the *traditional ensemble* was trained using bagging (Breiman 1996) due to its efficiency in dealing with the variance problem (Dietterich 2000). Since we have four folds for training, we built an ensemble of four networks where each network was trained on three folds following a leave-one-fold-out scheme. The shared layers of ESRs allow us to test some variations of bagging. After adding a new convolutional branch to the ESR, the shared layers (lr_{sl}) and already trained branches (lr_{tb}) continue learning on additional data using (1) the same initial learning rate (*fixed lr.*; $lr_{sl} = lr_{tb} = 0.1$), (2) a smaller learning rate (*varied lr.*; $lr_{sl} = 0.1$ and $lr_{tb} = 0.02$), or (3) not training at all (*frozen*

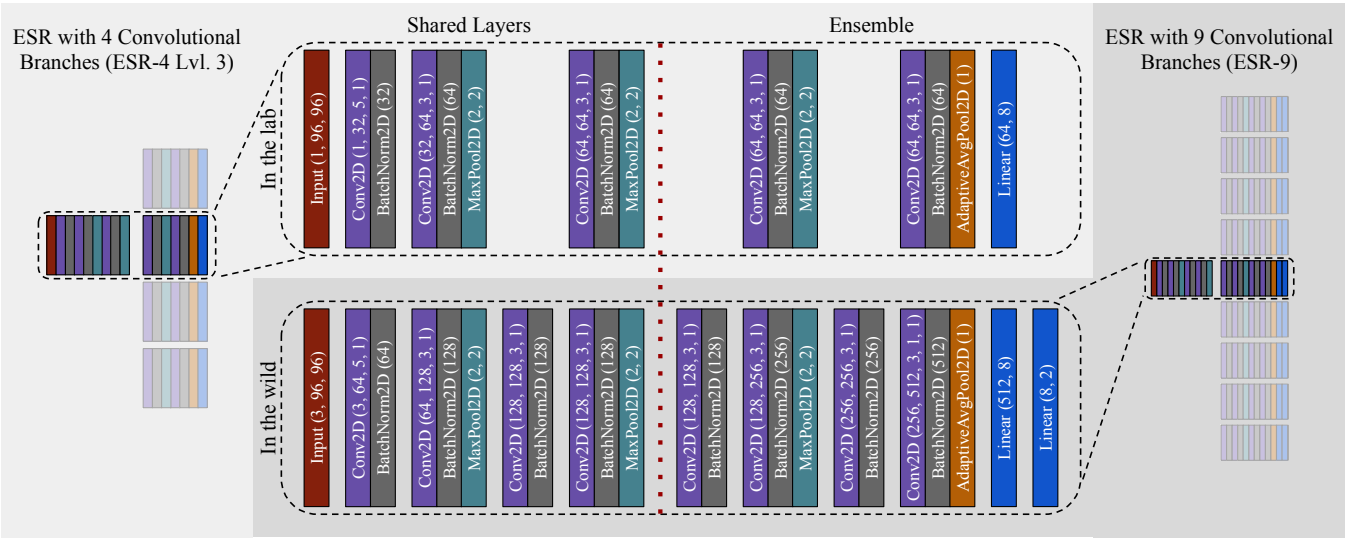

Figure 3: On the top, the architecture used in the in-the-lab experiments and an illustration of ESR-4 Lvl. 3 on the left. On the bottom, the architecture used in the in-the-wild experiments and an illustration of ESR-9 on the right. The latter architecture was designed to be equivalent to the former with respect to the spatial information of the features. The ReLU activation function is applied after batch normalization layers. The last linear layer in the bottom architecture was added for the affect perception experiment only. Each color represents a different type of layer and the PyTorch nomenclature was followed for reproducibility.

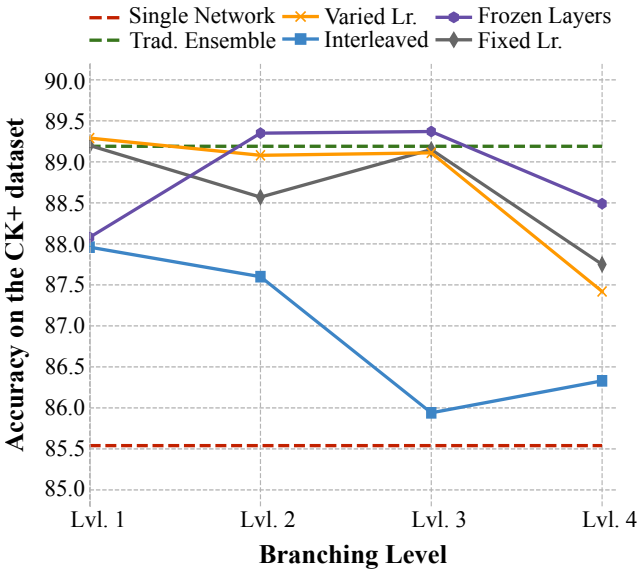

Figure 4: Accuracy on the Extended Cohn-Kanade dataset increasing branching level for different training strategies.

layers; $lr_{sl} = lr_{tb} = 0.0$). We adopted a momentum factor of 0.9 on SGD and a learning rate decay with a multiplicative factor of 0.5 applied after every 250 epochs. Finally, we also included the *interleaved* training strategy adopted in our previous work (Siqueira et al. 2018) in this experiment where all branches were trained iteratively on random mini-batches from the four folds. Data augmentation was randomly applied in all of the cases including brightness and contrast changes, horizontal flips, rotations up to 30 degrees, translations, and rescaling.

Table 1: Test accuracy (%) of the most accurate networks and baselines on CK+ and their number of parameters.

Approach	#	Accuracy
Single Network	**131.208**	$85.5 \pm 3.5\%$
Traditional Ensemble	524.832	$89.2 \pm 1.2\%$
ESR-4 Lvl. 3	355.104	$\mathbf{89.4 \pm 2.2\%}$
ESR-4 Lvl. 4	243.936	$88.5 \pm 3.8\%$

Table 2: Paired *t*-test (*p*-values) to compare Single Network, Trad. Ensemble, ESR-4 Lvl. 3, and ESR-4 Lvl. 4 on CK+.

	TE	**Lvl. 3**	Lvl. 4
Single Network	0.004 ✓	**0.005** ✓	0.043 ✓
Trad. Ensemble (TE)	–	**0.956** ✗	0.614 ✗
Lvl. 3	–	–	0.514 ✗

Figure 4 displays the mean accuracy on the CK+ test set with increasing branching level for every approach as well as the baselines (dashed lines). Consistent with ensemble literature, the ensemble methods achieved higher accuracies than the single network. The interleaved approach, however, demonstrated inferior performance among the ensembles. We believe the poorer performance might have been caused by low diversity in the ensemble. In interleaved training, diversity derives only from different starting points and different data augmentation executions on shuffled mini-batches. The accuracies obtained by the ESRs especially at level 3 were as high as the traditional ensemble method but the advantage is evident in the number of trainable parameters used by each approach, as shown in Table 1. ESRs need far

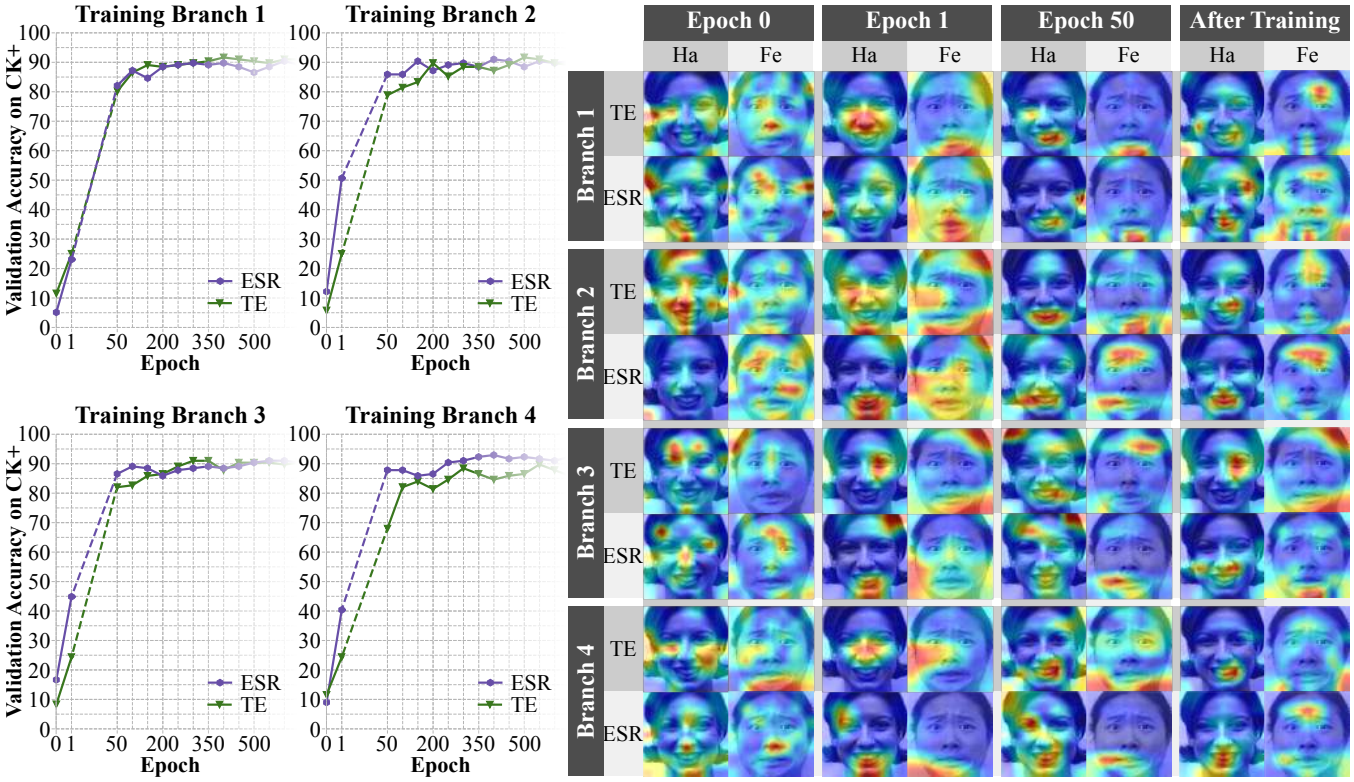

Figure 5: Comparison of the accuracy (%) on the validation set of CK+ over epoch between the ensemble with shared representations and the traditional ensemble. On the right, Grad-CAM visualization at different training milestones including before any weight update (epoch 0). The saliency maps were generated using the jet colormap, where red regions indicate facial features that contributed the most to the high activation of the output neurons, in this example, happy or fear. Best viewed in color.

less trainable parameters than traditional ensembles with a substantial decrease of 32% at level 3 and 54% at level 4, while achieving the same generalization power confirmed by the paired *t*-test in Table 2. Positive markers indicate statistically significant differences ($p < 0.05$). The improvement in recognition performance is clear when compared with a single network. The high *p*-value between the ESR with four branches at level 3 and TE indicates that the generalization abilities are equivalent while the redundancy and computational load are significantly reduced by ESR-4 Lvl. 3.

Transfer Learning and Diversity Analysis

Training time is an important factor when training deep neural networks, especially, ensembles of them. Figure 5 shows both how transfer learning in ESRs accelerates and guides the learning of new branches, as well as the diversity analysis of learned facial features. The graphs, on the left, compare the convergence of the ESR and TE over epoch for every branch, or network, added to the ensemble. The convergence curve follows the same pattern as the ensemble size increases in TEs since any new model is trained from scratch. On the other hand, the convergence speed of additional branches in ESRs increases due to the prior knowledge stored in the shared layers. Even after the first update, the ESR was already twice as accurate as TEs, and this gap was only closed around epoch 50.

These quantitative results suggest that the shared layers learned informative facial features of emotion concepts. To support our claim with visual evidence, we generated saliency maps with respect to the ESR and TE predictions at different training milestones using Grad-CAM. Note, on the right, that the learning progress of facial features advanced at the same pace while the ensemble size was one (both methods are identical). When training the second branch, however, the ESR already learned after the first update that the region around the mouth is relevant for recognizing happy facial expressions, whereas the TE took around 50 epochs to discover the same pattern.

In fact, AU-12 from the FAC must be visible on happy facial expressions in CK+ (Lucey et al. 2010). When the facial muscle underlying AU-12 (i.e., *Zygomatic Major*) is fired, it pulls the corner of the lips up, producing a smile. The smile is one of the most discriminative and repetitive facial features presented in CK+ that distinguishes happy facial expressions among other categories. Fear, for instance, is categorized from the combination of more complex facial features, which would require more time for the ESR and TE to learn such features. Nevertheless, the ESR learned around epoch 50 that frowning is one of the features necessary to recognize fear. This appearance change is produced by the *frontalis* muscle that covers the forehead and, when activated, can raise the inner brow (AU-1). The other feature is

the lip stretcher coded by AU-20 and presented in fear facial expressions in CK+. In general, the TE needs more training epochs than ESR to learn informative facial features. Finally, note that the diversity of features of the ESR is as high as the TE. In the happy facial expression example, while branches 2 and 4 captured the smile after training, branch 3 focused on the nasolabial furrows and branch 1 captured the wrinkles in the outer eyes caused by raising the cheeks (AU-6).

Training ESRs on Large-Scale Datasets

Nowadays, training ensembles on large-scale datasets became impracticable for those who have limited computational resources because even a single deep neural network may take over a month of training using several GPUs in large data centers (Chollet 2017). This section supports that ESRs are affordable for ensembling on large-scale datasets. Besides short training time and low computational cost, ESRs can reduce the remaining residual generalization error which led to higher accuracies than state-of-the-art methods in facial expression recognition benchmarks.

Methodology

Along with data, benchmark datasets usually provide standard experimental protocols and baseline results. AffectNet and FER+ have divided the dataset into training, validation, and test sets, and published them for the scientific community, except for the test set of the former. Meanwhile, researchers have utilized the validation set for evaluation and comparisons, as suggested by the AffectNet authors. We followed the same methodologies as the state-of-the-art methods for fair comparisons on both datasets. In experiments on AffectNet, the best inference performance on the validation set is reported, whereas the mean and standard deviation of the test accuracies after five trials are used as an evaluation metric for FER+ (Barsoum et al. 2016).

Evaluation on the AffectNet Dataset

As the body of features increases, the memory capacity of the neural network shall also increase to account for the higher volume of patterns. Therefore, the architecture used from this point is based on the previous ESR but with a few more convolutional layers, and batch normalization layers, as well as more convolutional filters per layer. In order to preserve the spatial information of the features, we adopted the same spatial reduction rate of the feature maps from the previous experiment by adding a max-pooling layer after every two convolutional layers. The ensemble of convolution branches begins in an equivalent spatial level to ESR-4 Lvl. 3, where the shape of the feature maps are similar, as depicted in Figure 3.

Discrete emotion perception. One of the challenges of facial expression recognition on AffectNet is the imbalance problem. We coped with this problem by training the branches of the ESR on balanced subsets from the whole training set containing up to 5000 samples of each emotion. Through an empirical analysis, subsets with fewer samples of each category resulted in lower performance, while more samples provided no significant gain in accuracy. The stochastic gradient descent was used to minimize the cross-entropy loss function with an initial learning rate of 0.1, a momentum of 0.9, and a learning rate decay with a multiplicative factor of 0.5 applied after every 10 epochs. Convolutional branches were added to the ensemble until no significant gain in accuracy was achieved by the collective classification. Trained branches were continually updated on additional training data with a lower initial learning rate of 0.01 for their adaptation to the representational changes in the shared layers.

Table 3: Accuracy (%) on AffectNet for discrete emotions and number of emotion labels used in the experiments.

Approach	#	Acc ↓
ESR-9 (Our network)	**8**	**59.3%**
AlexNet-WL (Mollahosseini et al. 2019)	8	58.0%
VGGNet (Hewitt and Gunes 2018)	8	58.0%
MobileNet (Hewitt and Gunes 2018)	8	56.0%
AlexNet (Hewitt and Gunes 2018)	8	56.0%
AlexNet-US (Mollahosseini et al. 2019)	8	47.0%
AlexNet-DS (Mollahosseini et al. 2019)	8	40.0%
gACNN (Li et al. 2019)	7	58.8%
IPA2LT (Zeng, Shan, and Chen 2018)	7	57.3%
pACNN (Li et al. 2019)	7	55.3%

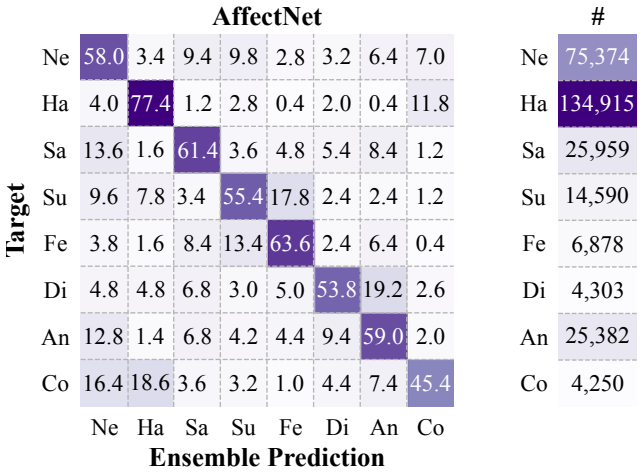

Figure 6: Normalized confusion matrix of the ensemble predictions on the AffectNet dataset and the emotion label distribution.

Our results are reported in Table 3 and Figure 6. The ESR with 9 convolutional branches (ESR-9) achieved the highest recognition performance on AffectNet in comparison to state-of-the-art methods. It is important to note that no single branch in the ensemble achieved an accuracy higher than 58.0%, only the collective classification made by the ensemble reached 59.3% of accuracy, which reveals that the remaining residual generalization error was reduced by the ESR. In the confusion matrix, we can see that ESR-9 was more accurate in the recognition of the happiness category but under-represented categories such as fear, disgust, and contempt were still well recognized given the disparity of

Table 4: Root-mean-square error (RMSE) for arousal (aro) and valence (val) prediction on the AffectNet dataset.

Approach	RMSE	
	Aro ↓	Val
ESR-9 (Our network)	**0.33**	**0.36**
VGGNet (Hewitt and Gunes 2018)	0.37	0.41
MobileNet (Hewitt and Gunes 2018)	0.38	0.42
AlexNet (Hewitt and Gunes 2018)	0.39	0.41
AlexNet (Mollahosseini et al. 2019)	0.41	0.37
VGG16-Based (Lindt et al. 2019)	0.41	0.45

the label distribution. Finally, even though Li et al. (2019)'s ensemble has obtained an accuracy of 58.8%, the contempt category was removed from their experiments, which greatly reduced the complexity of the task since the chance of the network for learning undesired features decreased.

Continuous affect perception. Predicting arousal and valence levels of facial images in a continuous space is a complex task where disagreement levels between human annotators are usually higher than in discrete emotion annotations. Thus, we trained the ESR in a curriculum learning fashion (Bengio et al. 2009), where ESR-9, trained for discrete facial expression recognition, was fine-tuned for arousal and valence prediction. We assumed that some facial features learned by ESR-9 from the previous task would lead the network to learn faster and become more accurate for inferring affect concepts than training it from scratch. For example, a smile detector usually learned after a few training epochs, as shown in our experiments on the in-the-lab dataset, can be associated with positive arousal and positive valance levels. Instead of replacing the output layer of ESR-9 to account for arousal and valence predictions, we added two neurons on top of each branch, as shown in Figure 3, and trained only the weights connected to those neurons. Since the relation of discrete emotions and continuous affect is non-linear, we applied the ReLU function to the second last layer that is related to discrete emotion concepts.

We followed the same training procedure as in our previous experiments where each branch is sequentially trained on a balanced subset with up to 5000 samples from each quadrant of the arousal and valence circumplex to reduce bias. However, since arousal and valence prediction in the continuous domain is a regression problem, we minimize the root-mean-square error using stochastic gradient descent with a momentum of 0.9 and a learning rate of 0.01. Trained branches were continually updated with a lower learning rate of 0.001. The results are reported in Table 4. ESR-9 outperformed state-of-the-art methods based on established pre-trained deep neural networks for visual classification tasks with a significant margin on both of the arousal and valence dimensions by achieving 0.33 and 0.36 RMSEs, respectively. Moreover, since only the output layer was trained, ESR-9 can still perform discrete emotion perception which resulted in a great drop in computational load and redundancy. In comparison to Mollahosseini, Hasani, and Mahoor (2019)'s approach which has approximately $180M$ parameters in total, ESR-9 has 9 times fewer parameters ($\approx 20M$). In their work, three AlexNets were trained, one for each of the three facial expression perception problems. Finally, ESR-9 reached the performance of human experts in facial expression annotations which have a disagreement level of 0.36 and 0.34 RMSEs for arousal and valence prediction.

Fine-tuning on the FER+ Dataset

In our experiments on FER+, we rescaled the images from 48 x 48 pixels to 96 x 96 pixels and fine-tuned ESR-9 trained on AffectNet. Before any training on FER+, ESR-9 achieved a test accuracy of 57.92%, similar to its performance on AffectNet. This cross-dataset evaluation indicates that ESR-9 generalizes well to different data distributions. It is important to note that facial images from FER+ are gray-scale images with low resolution and are not as centralized as AffectNet's images. These aspects may deteriorate certain facial features in images relevant for emotion perception until a point where they can no longer be detected, as argued by Tian, Kanade, and Cohn (2005). We fine-tuned ESR-9 using the stochastic gradient descent with a momentum of 0.9, an initial learning rate of 0.1, and a learning rate decay with a multiplicative factor of 0.75 applied after every 10 epochs. The learning rate decay was increased due to the faster convergence of ESR-9 on FER+. Trained branches were continually updated with a lower initial learning rate of 0.02.

After fine-tuning each branch sequentially on random subsets with up to 5000 training samples per emotion category on FER+, ESR-9 reached an average test accuracy of 87.153% with a very low standard deviation of 0.097%, outperforming the current state-of-the-art method (Barsoum et al. 2016). Our results are reported in Table 5 and Figure 7. Also, ESR-9 generalized reasonably well to under-represented categories. When compared to Barsoum et al. (2016)'s approach, for instance, ESR-9 correctly recognized 20.0% of the contempt test samples and 56.2% of the disgust test samples, while their approach recognized only 4.17%, and 26.32% respectively. The bias towards neutral classifications was also reduced in almost all categories. While our approach misclassified 40% of contempt samples as neutral and had no misclassification of disgust samples as neutral, their approach misclassified 54.17% of contempt and 10.53% of disgust samples. The bias problem in facial expression recognition is not solely related to the unbalance problem, but also to the inherent subjectivity of emotion perception on faces where humans may perceive different emotions in the same facial expression as illustrated in Figure 8 (Barrett 2017; Mollahosseini, Hasani, and Mahoor 2019).

It is relevant to mention that we extensively investigated the effects of varying the maximum number of training samples for each emotion category on FER+. When trained with lower upper bounds, ESRs increased recognition on under-represented categories but the overall accuracy decreased. If the upper bound is too high, the diversity of the ensemble decreased. In this experiment, an upper-bound of 5000 samples for each category was the "optimal" value to achieve high overall accuracy and a relatively high correct classification of under-represented categories. Finally, our findings suggest that our approach to ESRs is an important contribution to alleviate the bias problem in machine learning.

Table 5: Mean and standard deviation of the test accuracy on FER+. Some authors only reported the best accuracy.

Approach	Acc ↓
ESR-9 (Our network)	$\mathbf{87.15 \pm 0.1\%}$
SHCNN (Miao et al. 2019)	86.54%
VGG16-PLD (Barsoum et al. 2016)	$84.99 \pm 0.37\%$
VGG16-CEL (Barsoum et al. 2016)	$84.72 \pm 0.24\%$
TFE-JL (Li et al. 2018)	84.3%
VGG16-ML (Barsoum et al. 2016)	$83.97 \pm 0.36\%$
VGG16-MV (Barsoum et al. 2016)	$83.85 \pm 0.63\%$
ResNet18 + FC (Li et al. 2018)	83.4%
ResNet18 (Li et al. 2018)	83.1%

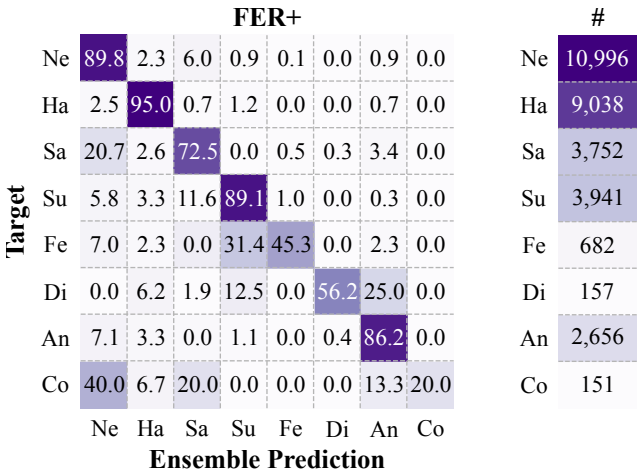

Figure 7: Normalized confusion matrix of the ensemble predictions on FER+ and the emotion label distribution.

Conclusions

Referring to Minsky at the beginning of this paper, one may think that single deep neural networks trained on large-scale datasets are enough to build rich, robust and highly accurate perceptual models. However, in certain domains where label distribution is unbalanced, for instance, those networks tend to become highly biased to the most representative categories. We demonstrated that ensembles with shared representations cope with this problem by training "many resources" (i. e., convolutional branches) on balanced subsets from the training data. Together, through the collective classification made by the ensemble, ESRs outperformed state-of-the-art deep neural networks on AffectNet and FER+ with low redundancy and an efficient transfer learning mechanism from the shared layers. Moreover, we showed that the branching level directly impacts ensemble diversity, generalization, and computational load.

Artificial neural networks, when trained under continual learning settings, typically suffer from a phenomenon called *catastrophic forgetting*. Correct classified samples become misclassified when the network is continually trained on a different data distribution due to its inability to keep learned information. The same phenomenon occurs when training additional branches in ESRs having a direct impact on the generalization performance. To address the effects of catastrophic forgetting on ESRs, learning rates of the trained branches and shared layers should be carefully defined. High differences in learning rates may cause trained branches to forget learned information, whereas similar learning rates may foster co-adaptation between branches and decrease ensemble diversity. In the future, we will investigate approaches to overcome catastrophic forgetting in ensembles with shared representations.

Figure 8: Subjective perception of facial expressions. Samples annotated as fear by one expert human annotator perceived differently by another expert. Adapted from (Mollahosseini, Hasani, and Mahoor 2019).

Despite reaching human-level performance in facial expression recognition on AffectNet, human-level affect inference under real-world conditions is far to be reached. To do so, computational models closer to recent findings that are changing and enhancing our understanding of emotions under the theory of psychological construction (Barrett and Russell 2015) should be developed. It is important to take into consideration not only cross-modal learning of emotional expressions but also temporal and contextual information during emotional episodes. As the next step, we will implement a model closer to the theory of constructed emotion (Barrett 2017) by adopting ESR-9's high-level representations as "proto concepts" of facial expressions to guide learning of emotion concepts in a hybrid neural system based on an intermediate view between empiricism and nativism of the cognition theory (Ullman 2019).

Acknowledgments

This work has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 721619 for the SOCRATES project. The authors thank Prof. Dr. Thomas Hellström for his insightful questions that motivated the development of this paper.

References

Barrett, L. F., and Russell, J. A. 2015. An introduction to psychological construction. *The psychological construction of emotion* 1–17.

Barrett, L. F. 2017. *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt.

Barsoum, E.; Zhang, C.; Ferrer, C. C.; and Zhang, Z. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ICMI*, 279–283.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proc. of ICML*, 41–48.

Breiman, L. 1996. Bagging predictors. *Machine Learning* 24(2):123–140.

Caruana, R.; Niculescu-Mizil, A.; Crew, G.; and Ksikes, A. 2004. Ensemble selection from libraries of models. In *Proc. of ICML*, 18–25.

Chollet, F. 2017. Xception: Deep learning with depthwise separable convolutions. In *IEEE Conference on CVPR*.

Chollet, F. 2018. *Deep Learning with Python and Keras: The Handbook by the Developer of the Keras Library*. MITP-Verlag GmbH & Co. KG.

Darwin, C. 1872. *The expression of the emotions in man and animals*. John Murray, UK.

Dietterich, T. G. 2000. Ensemble methods in machine learning. In *Multiple Classifier Systems*, 1–15. Berlin, Heidelberg: Springer Berlin Heidelberg.

Ekman, P., and Friesen, W. V. 1976. Measuring facial movement. *Envir. Psycho. and Nonverbal Behavior* 1(1):56–75.

Ekman, P. 1989. The argument and evidence about universals in facial expressions. *Handbook of Social Psychophysiology* 143–164.

Gendron, M.; Crivelli, C.; and Barrett, L. F. 2018. Universality reconsidered: Diversity in making meaning of facial expressions. *Current Directions in Psychological Science* 27(4):211–219.

Goodfellow, I. J.; Erhan, D.; Carrier, P. L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.-H.; Zhou, Y.; Ramaiah, C.; Feng, F.; Li, R.; Wang, X.; Athanasakis, D.; Shawe-Taylor, J.; Milakov, M.; Park, J.; Ionescu, R.; Popescu, M.; Grozea, C.; Bergstra, J.; Xie, J.; Romaszko, L.; Xu, B.; Chuang, Z.; and Bengio, Y. 2015. Challenges in representation learning: A report on three machine learning contests. *Neural Networks* 64:59 – 63. Special Issue on Deep Learning of Representations.

Hansen, L. K., and Salamon, P. 1990. Neural network ensembles. *IEEE Transactions on PAMI* 12(10):993–1001.

Hewitt, C., and Gunes, H. 2018. Cnn-based facial affect analysis on mobile devices. *arXiv preprint arXiv:1807.08775*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Bach, F., and Blei, D., eds., *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, 448–456. Lille, France: PMLR.

James, W.; Burkhardt, F.; Bowers, F.; and Skrupskelis, I. K. 1890. *The principles of psychology*, volume 1. Macmillan London.

James, W. 1884. What is an emotion? *Mind* 9(34):188–205.

Khorrami, P.; Paine, T. L.; and Huang, T. S. 2015. Do deep neural networks learn facial action units when doing expression recognition? In *IEEE on CVPR - Workshops*, 19–27.

Koolagudi, S. G., and Rao, K. S. 2012. Emotion recognition from speech: a review. *IJST* 15(2):99–117.

Li, M.; Xu, H.; Huang, X.; Song, Z.; Liu, X.; and Li, X. 2018. Facial expression recognition with identity and emotion joint learning. *IEEE Trans. on Affective Computing* 1–1.

Li, Y.; Zeng, J.; Shan, S.; and Chen, X. 2019. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Trans. on IP* 28(5):2439–2450.

Lindt, A.; Barros, P.; Siqueira, H.; and Wermter, S. 2019. Facial expression editing with continuous emotion labels. In *14th IEEE Int. Conf. on FG 2019*, 1–8.

Lucey, P.; Cohn, J. F.; Kanade, T.; Saragih, J.; Ambadar, Z.; and Matthews, I. 2010. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on CVPR - Workshops*, 94–101.

Meshgi, K.; Oba, S.; and Ishii, S. 2018. Efficient diverse ensemble for discriminative co-tracking. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Miao, S.; Xu, H.; Han, Z.; and Zhu, Y. 2019. Recognizing facial expressions using a shallow convolutional neural network. *IEEE Access* 7:78000–78011.

Minsky, M. 2014. Is the singularity near? Available at: https://www.youtube.com/watch?v=RZ3ahBm3dCk. [Accessed 23 Jul. 2019].

Mollahosseini, A.; Hasani, B.; and Mahoor, M. H. 2019. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* 10(1):18–31.

Poria, S.; Cambria, E.; Bajpai, R.; and Hussain, A. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37:98–125.

Russell, J. A. 2003. Core affect and the psychological construction of emotion. *Psychological review* 110(1):145.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE Int. Conf. on Computer Vision (ICCV)*, 618–626.

Seung, H. S.; Opper, M.; and Sompolinsky, H. 1992. Query by committee. In *Proc. of the 5th Workshop on Comp. Learning Theory*, 287–294.

Shen, Z.; He, Z.; and Xue, X. 2019. Meal: Multi-model ensemble via adversarial learning. In *Proceedings of AAAI Conference on Artificial Intelligence*.

Siqueira, H.; Barros, P.; Magg, S.; and Wermter, S. 2018. An ensemble with shared representations based on convolutional networks for continually learning facial expressions. In *IEEE/RSJ Int. Conf. on IROS*, 1563–1568.

Tian, Y.-L.; Kanade, T.; and Cohn, J. F. 2005. *Facial Expression Analysis*. New York, NY: Springer New York. 247–275.

Ullman, S. 2019. Using neuroscience to develop artificial intelligence. *Science* 363(6428):692–693.

Viola, P., and Jones, M. J. 2004. Robust real-time face detection. *Int. Journal of Computer Vision* 57(2):137–154.

Zeng, J.; Shan, S.; and Chen, X. 2018. Facial expression recognition with inconsistently annotated datasets. In *The European Conference on Computer Vision (ECCV)*.