Increasingly Packing Multiple Facial-Informatics Modules in A Unified Deep-Learning Model via Lifelong Learning

Steven C. Y. Hung
IIS, Academia Sinica
fevemania@iis.sinica.edu.tw

Jia-Hong Lee
IIS & CITI, Academia Sinica
honghenry.lee@iis.sinica.edu.tw

Timmy S. T. Wan
IIS & CITI, Academia Sinica
timmywan@iis.sinica.edu.tw

Chein-Hung Chen
IIS & CITI, Academia Sinica
redsword26@iis.sinica.edu.tw

Yi-Ming Chan
IIS, Academia Sinica; MOST Joint Research Center for AI Technology and All Vista Healthcare
yiming@iis.sinica.edu.tw

Chu-Song Chen
IIS & CITI, Academia Sinica; MOST Joint Research Center for AI Technology and All Vista Healthcare
song@iis.sinica.edu.tw

ABSTRACT

Simultaneously running multiple modules is a key requirement for a smart multimedia system for facial applications including face recognition, facial expression understanding, and gender identification. To effectively integrate them, a continual learning approach to learn new tasks without forgetting is introduced. Unlike previous methods growing monotonically in size, our approach maintains the compactness in continual learning. The proposed packing-and-expanding method is effective and easy to implement, which can iteratively shrink and enlarge the model to integrate new functions. Our integrated multitask model can achieve similar accuracy with only 39.9% of the original size.

CCS CONCEPTS

• **Computing methodologies → Lifelong machine learning**; **Neural networks**; **Computer vision tasks**.

KEYWORDS

Deep learning, neural networks, compact model, continual learning, lifelong learning, facial informatics

ACM Reference Format:
Steven C. Y. Hung, Jia-Hong Lee, Timmy S. T. Wan, Chein-Hung Chen, Yi-Ming Chan, and Chu-Song Chen. 2019. Increasingly Packing Multiple Facial-Informatics Modules in A Unified Deep-Learning Model via Lifelong Learning. In *2019 International Conference on Multimedia Retrieval (ICMR '19), June 10–13, 2019, Ottawa, ON, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3323873.3325053

1 INTRODUCTION

Facial informatics are fundamental to human-computer interaction in multimedia applications. Many face-related modules are required in a multimedia system, such as face recognition, facial expression, and gender classification. Each of these functionalities has its own state-of-the-art deep learning models, which have been well-trained on carefully collected benchmarks for specific purposes. For example, in face recognition, various solutions have been released and applied to modern life, such as FaceNet [21] and SphereFace [12]. In facial expression, the AffectNet [16] is a famous baseline in recent research and it also releases the largest facial expression dataset. This dataset becomes a fundamental evaluation in recent solutions, such as CAKE [7] and Zeng et al. [27]. In gender classification, Levi et al. [11] is the first research using CNN and it releases the Adience dataset [3] that becomes an evaluation baseline. Chalearn challenge [4] releases the dataset and several solutions as well.

When an effective face-recognition model is available, we would like to integrate other functionalities into this model, and hope that the additional functionalities (such as expression and gender prediction) remain accurate without performance sacrifice. A continual (lifelong) learning approach is introduced to integrate multi-functionality modules into a unified deep-learning model. Unlike previous lifelong learning approaches that may monotonically grow the deep-learning model architecture [20, 25] and result in a redundant model, our approach can perform continual learning while maintaining the compactness of the model. Besides, the functionality modules previously built are ensured to have exactly the same performance when new modules are incrementally added in our approach. To achieve the purpose, We introduce a packing-and-expanding (PAE) approach that shrinks and enlarges the model in an iterative manner, which can integrate the required functionalities in a compact model without affecting their accuracy in inference.

2 RELATED WORK

Lifelong Learning. The goal of lifelong learning [24] is to build a model capable of learning unknown sequential tasks while keeping the performance of previous tasks. As training a new task directly will force the model parameters to fit new data, which causes *catastrophic forgetting* [14, 15] on previously learned tasks, techniques leveraging on lifelong learning tend to avoids catastrophic forgetting in deep neural networks. In Kirkpatrick et al. [8] and Zenke et al. [28], the authors regularize the network weights and hope to search the common convergence for the current and previous tasks. In Yoon et al. [26], additional capacity is dynamically increased for the neural-network model when training new tasks, so that the old

ICMR '19, June 10–13, 2019, Ottawa, ON, Canada

ACM ISBN 978-1-4503-6765-3/19/06…$15.00
https://doi.org/10.1145/3323873.3325053

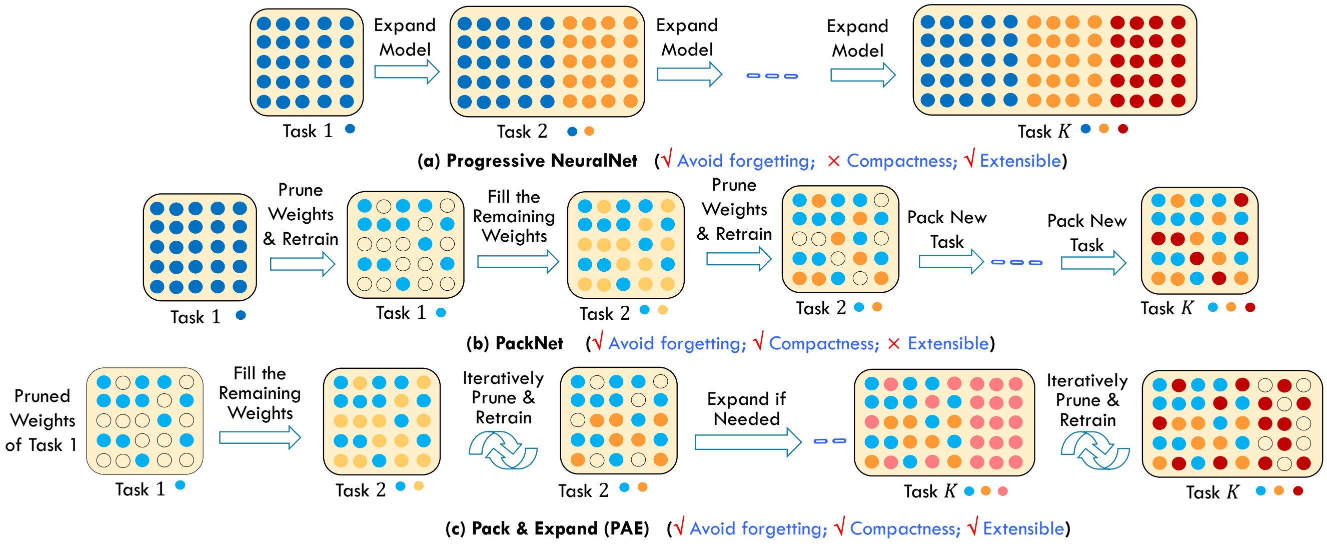
Expand Model
Expand Model
Expand Model
Task 1 ●
Task 2 ● ●
Task K ● ● ●
(a) Progressive NeuralNet (√ Avoid forgetting; × Compactness; √ Extensible)
Prune Weights & Retrain
Fill the Remaining Weights
Prune Weights & Retrain
Pack New Task
Pack New Task
Task 1 ●
Task 1 ●
Task 2 ● ●
Task 2 ● ●
Task K ● ● ●
(b) PackNet (√ Avoid forgetting; √ Compactness; × Extensible)
Pruned Weights of Task 1
Fill the Remaining Weights
Iteratively Prune & Retrain
Expand if Needed
Iteratively Prune & Retrain
Task 1 ●
Task 2 ● ●
Task 2 ● ●
Task K ● ● ●
Task K ● ● ●
(c) Pack & Expand (PAE) (√ Avoid forgetting; √ Compactness; √ Extensible)

Figure 1: Illustration of related lifelong learning methods. (a) ProgressiveNet [20] handles lifelong learning via expanding the architecture and re-using the fixed old-task weights. (b) PackNet [13] adds new tasks via compressing the model for old tasks at first, and the previous-task weights are re-used and fixed as well. (c) Our PAE compresses and (selectively) expands the model in an alternating manner, which maintains the model compactness, avoids forgetting, and allows model expansion.

tasks are less influenced. In Rebuffi et al. [18], Shin et al. [22] and Rannen et al. [17], representation or distribution of the training data of previous tasks are recorded to prevent the learned models from forgetting the previous tasks when training new tasks.

Though the above approaches lessen the effect of forgetting old skills, they do not guarantee to preserve the performance for previous tasks. To exactly avoid forgetting, a promising way is to keep the old-task weights already learned, and enlarge the network by adding nodes or weights for training new tasks [20][25]. In ProgressiveNet [20], to ease the training of new tasks, the old-task weights are shared with the new ones but remain fixed, where only the new weights are adapted for the new task. As the old-task weights are kept, it ensures to preserve the performance of learned tasks. However, as the model structure is getting increased with the tasks, it would result in a redundant structure for keeping multiple models. Recently, PackNet [13] packs multiple tasks sequentially in a single model. According to deep-net compression [6], there is much redundancy in a neural network, and removing the redundant weights does not affect the network performance. PackNet [13] exploits this property, which compresses an old task by deleting neglectable weights at first and the deleted weights are then saved for packing the next task. The old-task weights are also re-used for the new tasks but remain fixed, and thus the old tasks are ensured unforgettable. However, PackNet only compresses a single network without expansion, and thus the tasks to be packed is limited.

In this work, we introduce PAE that avoids the above limitations. Our approach ensures keeping the old-task performance, yields a compact model, and the model is extensible for lifelong learning.

Facial Informatics. Various facial informatics have been studied. We briefly review the works on face recognition, age prediction, gender and expression classifications in recent years. In face recognition, recent researches attempt to obtain a better deep learning model by either modifying the objective functions or redesigning the network structures. For example, FaceNet [21] proposes triplet loss and redesigns the network structure; SphereFace [12] proposes additive angular margin loss and also redesigns the network structure. In age and gender prediction, Levi et al. [11] introduce a five-layer CNN with the Adience dataset [3] released to public for age and gender classification. Lee et al. [10] follow the study of [11] to build a lightweight CNN that can be deployed on mobile devices. Besides, the Chalearn challenge [4] hold in 2016 also releases the FotW dataset for smile and gender classification and ChaLearn-AgeGuess dataset for age estimate, respectively. The technique of winners leverages the ensemble method to vote the result of the multiple CNN models. In expression classification, AffectNet [16] introduces the largest in-the-wild facial expression dataset and evaluates itself via a baseline model. To classify the expression based on latent emotion representation, CAKE [7] proposes a pipeline to predict discrete emotions by discriminating among the compact emotion embedding learned in multi-domain fashion.

In this work, we technically demonstrate that PAE can sequentially integrate facial understanding tasks in a compact model, and yields comparable performance to the state-of-the-art approaches using individual models.

3 METHOD

An illustration of our approach is given in Figure 1(c). As shown in the figure, ProgressiveNet [20] (Figure 1(a)) keeps the function mappings learned for previous tasks because the old-task weights are re-used and remain unchanged for the new tasks. It avoids catastrophic forgetting via architecture expansion, but the model is monotonically increased and a redundant structure is yielded. PackNet [13] (Figure 1(b)) compresses the deep model by pruning small weights and maintains the prediction accuracy via re-training the remaining weights. The pruned weights are saved for the new task, and the old weights are re-used and remain unchanged during re-training the new task as well. The newly trained weights could still be pruned so that the saved weights can be used for the next task. PackNet maintains the performance of all learned tasks and avoids forgetting. However, it does not allow extending the model architecture, and thus the tasks to be packed in a model is restricted.

Our PAE compresses the deep model and (selectively) expands the architecture alternatively. First, a compressed model is built via pruning. Unlike PackNet that directly prunes the model, we employ the iterative pruning procedure [29], which prunes a small portion of weights and retrains the remaining weights to restore the performance iteratively, and thus it tends to find a more compact model with the classification accuracy preserved. Next, we re-use the old-task weights that remain fixed and add extra weights from the previously saved ones by repeating iterative pruning. If the accuracy goal is not attained yet, the architecture can be expanded by adding filters and resuming the procedure. The algorithm of PAE is listed below.

Algorithm 1: Packing and Expanding (PAE)

```
Input: given task 1 and an original model trained on task 1.
Set an accuracy goal for task 1;
Alternatingly remove small weights and re-train the remaining
  weights for task 1 via iterative pruning [29], until meeting the
  accuracy goal;
Let the model weights reserved for task 1 be $W_1$ (referred to as task-1
  weights), and those that are removed by the iterative pruning be $W_1^r$
  (referred to as the saved weights);
for task $i = 2 \cdots K$ (let the saved weights of task $i$ be $W_{i-1}^r$) do
    Set an accuracy goal for task $i$;
    Use the weights $W_1$ and $W_{i-1}^r$ to train task $i$, with $W_1$ fixed;
    If the accuracy goal is not achieved by the trained model, expand
      the number of filters (wights) in the model, and reset
      $W_{i-1}^r \leftarrow W_{i-1}^r \cup W_E$, where $W_E$ denotes the expanded weights;
    Alternatingly remove small weights from $W_{i-1}^r$ and re-train the
      remaining weights (with $W_1$ fixed) for task $i$ via iterative
      pruning, until meeting the accuracy goal;
end
```

To record the weights reserved for tasks, we use an integer mask as PackNet [13] did. In the following, we apply PAE to lifelong learning of face-related tasks for multi-purpose facial understanding.

3.1 Initial Facial Task

To build a multi-purpose deep-learning system for various facial information understanding tasks, we would start from a base module that can extract fundamental facial representations. We choose face recognition (i.e., person identification from faces) as the first task. Among the existing well-trained face-recognition models, we use directly the FaceNet model[1] available publicly as our first module, which is an Inception-ResNet-v1 [23] network trained on the VGGFace2 [2] dataset for face recognition. We then retrain the model with iterative pruning [29] to prune away a dynamic fraction of redundant weights per a certain number of iterations. It can find an appropriate pruning ratio and maintain the accuracy after pruning.

After condensing and retraining the model with iterative pruning, we obtain a model with sparse weights, where the accuracy remains approximately the same as that of the original model. The surviving weights in each layer of the first module are kept via a mask, which records the protected weights that cannot be modified when training the new module. The space associated with the deleted weights is then saved for the other tasks.

3.2 Progressive Packing and Expanding

The saved space from the pruned weights of the previous module is allocated for training and pruning the weight of the newly coming task. Then, as depicted above, when the space is insufficient to maintain the accuracy of a new task, PAE will expand the network's connections by adjusting the number of convolution filters like [5].

Our PAE approach does not restrict the order of the incoming tasks. Without loss of generality, we test two combinations: identity+gender+age and identity+gender+expression. The results are shown in the next section.

4 EXPERIMENTAL RESULTS

We implement PAE using Tensorflow [1]. In the experiments, we verify the proposed PAE by comparing the accuracy and model size with other approaches of face-related tasks.

4.1 Face Verification, Gender and Age Modules

In the first experiment, we utilize PAE to increasingly integrate three modules: Face verification (identity), gender classification, and age estimation into one Inception-ResNet-v1 [23] model, called as *PAENet*. The datasets of these modules include:

VGGFace2 dataset [2] contains approximately 3.3 millions facial images with pose and emotion variations and different lighting and occlusion conditions, which has more than 9000 identities spanning a wide range of different ethnicities, accents, professions and ages.

LFW dataset [9] contains more than 13,000 images of faces collected from the Web and approximately has 1.680 identities.

Adience dataset [3] composed of pictures taken by camera from smartphones or tablets and its images are filled with extreme variations, including extreme blur (low-resolution), occlusions, out-of-plane pose variations, expressions. It includes 26,580 unconstrained images of 2,284 subjects. Its age labels contain eight groups, including (0−2), (4−6), (8−13), (15−20), (25−32), (38−43), (48−53), (60+).

We compare the accuracy and model size of our PAENet with individual models, such as FaceNet [21], AgeNet, GenderNet, and Levi_Hassner CNN [11]. The AgeNet and GenderNet are trained by fine-tuning the Inception-ResNet-v1 model (pretrained on VGGFace2 dataset [2]) with Adience dataset. We also compare PAENet with lightweight multi-task CNN model, LMTCNN-2-1 [10]. To calculate the accuracy of each module, we leverage the evaluation protocol of the FaceNet [21] that trains the model using VGGFace2 dataset [2] and tests on LFW dataset [9] for face verification, and we

[1] https://github.com/davidsandberg/facenet

Table 1: The accuracy of each method in the first experiment. The initial network model of FaceNet [21], AgeNet, GenderNet and PAENet are built via Inception-ResNet-v1 [23].

Methods	Face	Age	Gender
	LFW Acc.(%)	Top-1 Avg. Acc.(%)	Avg. Acc.(%)
Levi_Hassner CNN [11]		44.14	82.52
LMTCNN-2-1 [10]		44.26	85.16
FaceNet [21]	99.55±0.34		
AgeNet		56.37	
GenderNet			89.50
PAENet	**99.67±0.39**	**57.30**	89.08

Table 2: Model size of each method in the first experiment.

Methods	Modules	Model Size (MB)
FaceNet [21]	Face	89.6
AgeNet	Age	89.7
GenderNet	Gender	89.6
Levi_Hassner CNN [11]	Age	35.4
	Gender	35.4
LMTCNN-2-1 [10]	Age+Gender	30
PAENet (masks included)	Face	**68.9**
	Face+Age	**84.2**
	Face+Gender	**82.6**
	Face+Age+Gender	**97.8**

utilize the evaluation protocol of the LMTCNN-2-1 [10] that trains and tests on Adience dataset [3] using five-fold cross-validation and single-crop per image for Age and Gender classification. In Tables 1 and 2, we demonstrate that the accuracy of PAENet is higher than that of the other approaches [10, 11, 21] in all three tasks. PAENet also outperforms FaceNet and AgeNet, though has a slight accuracy drop with the GenderNet. Besides, the model size (masks included) of PAENet, which can process these three tasks simultaneously, is smaller than the sum of model sizes of individual-module models.

4.2 Face Verification, Gender and Expression

To verify whether expression and gender can be packed with face verification, we replace the task of age prediction with facial expression recognition in the second experiment. Details of the additional datasets are described in the following:

IMDb-Wiki dataset [19]: The refined dataset contains 260,282 face images from 20,284 celebrities crawled from IMDb and Wikipedia websites. Its labels have real age, gender and celebrity name.

FotW dataset [4] is used for the third-track competition of smile and gender classification in Chalearn big challenge [4]. It contains 9,258 facial images; 6,171 for training and 3,087 for validation.

AffectNet dataset [16] contains 287,401 facial images; 283,901 for training and 3,500 for validation. In the training set, the number of images per expression is imbalance. Its labels have seven primary expressions, Neutral, Happy, Sad, Surprise, Fear, Disgust and Anger.

Table 3: The accuracy of methods in the second experiment. The initial network model of FaceNet [21], EmotionNet, GenderNet and PAENet are with Inception-ResNet-v1 [23].

Methods	Face	Expression	Gender
	LFW Acc.(%)	Top-1 Acc.(%)	Acc.(%)
AffectNet [16]		58	
CAKE [7]		61.7	
SIAT MMLAB [4]			92.69
FaceNet [21]	99.55±0.34		
EmotionNet		64.74	
GenderNet			94.45
PAENet	**99.67±0.39**	**65.29**	92.93

Table 4: Model sizes in the second experiment.

Methods	Modules	Model Size (MB)
FaceNet [21]	Face	89.6
EmotionNet	Expression	89.6
GenderNet	Gender	89.6
PAENet	Face	**68.9**
	Face+Expression	**95.1**
	Face+Gender	**81.2**
	Face+Expression+Gender	**107.3**

We compare the accuracy and model size of PAENet with that of individual-task models, including FaceNet [21], EmotionNet and GenderNet. We train the EmotionNet by fine-tuning the Inception-ResNet-v1 model (pre-trained on VGGFace2 [2]) with AffectNet dataset [16], and the GenderNet by fine-tuning that model (pre-trained on VGGFace2 [2] and IMDb-Wiki [19] datasets) with FotW dataset [4], respectively. We also compare the accuracy of our PAENet with the state of the art and baseline approaches in expression and gender classification, such as AffectNet [16], CAKE [7] and SIAT MMLAB [4]. To calculate the accuracy of each module, we utilize the evaluation protocol of AffectNet [16] in the expression classification and the protocol of third-track challenge in Chalearn big challenge 2016 [4] for gender classification. In Tables 3 and 4, we demonstrate that the model size of our PAENet, which can handle the three tasks simultaneously, is only 39.9% of the sum of individual models. Besides, the accuracy of our PAENet is higher than the other approaches [16, 21] on gender and expression classifications.

In sum, the above results reveal that our PAE method can incrementally add various facial-informatics modules in a single neural network with little to no loss in the final model accuracy while maintaining the model compactness.

5 CONCLUSION AND FUTURE WORK

In this paper, we introduce a simple but effective method for lifelong deep learning, which employs an iterative compression and expansion principle to keep the previous-tasks performance, model compactness, and model extensiblility. We technically demonstrate fusing several face-related modules in a condensed model, and show the effectiveness of our method. Though the tasks of facial informatics are tested in this work, our method can be applied to other lifelong learning tasks as well, which will be our future work.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv* (2016).

[2] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. 2018. VGGFace2: A dataset for recognising faces across pose and age. In *Proceedings of IEEE FG.*

[3] Eran Eidinger, Roee Enbar, and Tal Hassner. 2014. Age and gender estimation of unfiltered faces. *IEEE Trans. Inf. Forensics Security* (2014).

[4] Sergio Escalera, Mercedes Torres Torres, Brais Martinez, Xavier Baró, Hugo Jair Escalante, Isabelle Guyon, Georgios Tzimiropoulos, Ciprian Corneou, Marc Oliu, Mohammad Ali Bagheri, et al. 2016. Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In *Proceedings of IEEE CVPRW.*

[5] Ariel Gordon, Elad Eban, Ofir Nachum, Bo Chen, Hao Wu, Tien-Ju Yang, and Edward Choi. 2018. Morphnet: Fast & simple resource-constrained structure learning of deep networks. In *Proceedings of IEEE CVPR.*

[6] Song Han, Huizi Mao, and William J Dally. 2016. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *ICLR* (2016).

[7] Corentin Kervadec, Valentin Vielzeuf, Stéphane Pateux, Alexis Lechervy, Frédéric Jurie, and Cesson-Sévigné. 2018. CAKE: Compact and Accurate K-dimensional representation of Emotion. In *Proceedings of BMVC.*

[8] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences* (2017).

[9] Erik Learned-Miller, Gary B Huang, Aruni RoyChowdhury, Haoxiang Li, and Gang Hua. 2016. Labeled faces in the wild: A survey. In *Advances in face detection and facial image analysis.* Springer.

[10] Jia-Hong Lee, Yi-Ming Chan, Ting-Yen Chen, and Chu-Song Chen. 2018. Joint Estimation of Age and Gender from Unconstrained Face Images using Lightweight Multi-task CNN for Mobile Applications. In *Proceedings of IEEE MIPR.*

[11] Gil Levi and Tal Hassner. 2015. Age and gender classification using convolutional neural networks. In *Proceedings of IEEE CVPRW.*

[12] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2017. SphereFace: Deep Hypersphere Embedding for Face Recognition. In *Proceedings of IEEE CVPR.*

[13] Arun Mallya and Svetlana Lazebnik. 2018. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 7765–7773.

[14] James L McClelland, Bruce L McNaughton, and Randall C O'reilly. 1995. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review* (1995).

[15] Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation.* Elsevier.

[16] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. 2017. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Trans. Affective Comput.* (2017).

[17] Amal Rannen Ep Triki, Rahaf Aljundi, Matthew Blaschko, and Tinne Tuytelaars. 2017. Encoder based lifelong learning. In *Proceedings of ICCV.*

[18] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of IEEE CVPR.*

[19] Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2016. Deep expectation of real and apparent age from a single image without facial landmarks. *Int. J. Comput. Vis. (IJCV)* (2016).

[20] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. 2016. Progressive neural networks. *arXiv* (2016).

[21] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of IEEE CVPR.*

[22] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. 2017. Continual learning with deep generative replay. In *Proceedings of NeurIPS.*

[23] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning.. In *Proceedings of AAAI.*

[24] Sebastian Thrun. 1995. -A Lifelong Learning Perspective for Mobile Robot Control. In *Intelligent Robots and Systems.* Elsevier.

[25] Tianjun Xiao, Jiaxing Zhang, Kuiyuan Yang, Yuxin Peng, and Zheng Zhang. 2014. Error-driven incremental learning in deep convolutional neural network for large-scale image classification. In *Proceedings of ACM-MM.*

[26] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. 2018. Lifelong Learning with Dynamically Expandable Networks. In *Proceedings of ICLR.*

[27] Jiabei Zeng, Shiguang Shan, and Xilin Chen. 2018. Facial Expression Recognition with Inconsistently Annotated Datasets. In *Proceedings of ECCV.* Springer.

[28] Friedemann Zenke, Ben Poole, and Surya Ganguli. 2017. Continual Learning Through Synaptic Intelligence. In *Proceedings of ICML.*

[29] Michael Zhu and Suyog Gupta. 2017. To prune, or not to prune: exploring the efficacy of pruning for model compression. In *Proceedings of NeurIPS Workshop on Machine Learning of Phones and other Consumer Devices.*