Deep Face Recognition in Incremental Learning Scenario

Sayedmoslem Shokrolahi
Student #: 20264065
Department of Electrical and Computer Engineering
ELEC 872 AI and Interactive Systems
December 2021

***Abstract*—This research investigates the deep Face Recognition (FR) problem under Incremental Learning (IL) umbrella. Firstly, a deep face recognition model is proposed using the CNN structure. We explore different data augmentation methods and four CNN architectures considering three different classifiers. Then, we choose the best model to study the FR-IL scenario. In the IL-FR, we consider ten batches (tasks) with 100 identities in each batch. In this case, a method named ScaIL [32] is used to handle catastrophic forgetting. We use VGGFace2 as our benchmark dataset in this report. In both FR and FR-IL scenarios, our simulations show that the ResNet structure is the best choice compared to VGG-16 and Inception-V3. Moreover, kl-div loss with label smoothing outperforms SoftMax (cross-entropy) and SVM (hinge loss).**

Keywords—deep face recognition, incremental learning, CNN, VGGFace2

I. INTRODUCTION

Human-computer interactive systems based on artificial agents have been improved dramatically with significant advances in deep neural networks (DNNs) and modern learnable representation fields. Face Recognition (FR) is considered a key component in countless smart spaces where managing users and individuals matters. Loads of works have been done on designing practical face recognition models using deep learning concepts, and DNNs show superior performance in learning biometric features utilizing large-scale datasets [1] [2].

After face detection and alignment, a deep face recognition module consists of three main parts: 1-Face processing 2 – Deep feature extraction 3- Face matching by deep features [3].

The face processing stage consists of two main categories, one-to-many augmentation and many-to-one normalization [3]. Many images are generated with pose and illumination variability in the one-to-many augmentation method to make the network more robust against these variations. The focus of this research is also on one-to-many augmentation.

Considering the second part, the prime process of the deep feature extraction module is mapping the face image into a feature vector (embedding). With the knowledge that the ideal embedding has small intra-class and large inter-class variation (in the FR scenario, each class equals unique identity). To address this challenge, many solutions are proposed to train deep neural networks by either directly learning the embedding (e.g., Triplet loss [2]) or by learning and identity classification problem (e.g., Softmax loss [4]).

In face matching by the deep features part, the deep model uses a gallery of known subjects to be trained in a supervised manner. After training, a new subject feeds to the model, and its deep feature representation is obtained in order to identify (or verify) the subject [3].

However, distinguishing between the training and inference phases has always been an arguable basic paradigm in machine learning models. After the training phase, model parameters are set to be fixed in a static scenario, and inference is performed using unseen test data. The static case is not practical in many real-world problems, including FR, and considering dynamic strategies with the capability of online updating based on new input sequences is essential [5].

Continual learning (CL) or Incremental Learning (IL) is the capability of model learning from sequential input data without forgetting during the time (i.e., arriving tasks). In the rest of the report, the concept 'task' refers to an isolated training phase with a new batch of data belonging to a new group of classes, a new domain, or a different output space [6]. However, we also use *batch* instead of the *task* in this report with the same meaning. Based on the importance of IL in face recognition application, we will represent some results of deep face recognition under the paradigm of incremental learning.

Turning to face recognition under the umbrella of Incremental Learning, the face recognition problem could be formulated as a class incremental classification problem [7]. The training process is based on a sequence of different incoming tasks that in each new task (new state), images of new classes are presented. The goal is to obtain a model with the ability to learn new faces (in each new task) without previously learned faces. The tendency of the model to underfit past data with increasing tasks is the most challenging problem in incremental learning, known as catastrophic forgetting [8]. Many solutions are proposed to solve catastrophic forgetting. Many researchers assume that the model can grow by increasing the number of tasks with some limitations on the number of model parameters [9] [10] [11]. In some other approaches known as replay methods, using a limited number of previous tasks samples is essential in the current training stage to handle the issue of catastrophic forgetting [12] [13]. Some approaches known as regularization-based methods remove the need to use previous samples (or pseudo samples). In these approaches, an additional regularization term is added to the loss function to keep previously learned knowledge safe while learning new data [14] [15].

In this research, subject-dependent face identification setting is considered in terms of training protocol and evaluation tasks. Our proposed solution is based on deep CNN networks in order to create proper embedding (Fig. 1). Our simulations will involve different data augmentation, network

architecture, and loss functions (classifiers) using the VGGFace2 dataset [4]. Then the best model will be selected to investigate mentioned IL condition in deep face recognition.

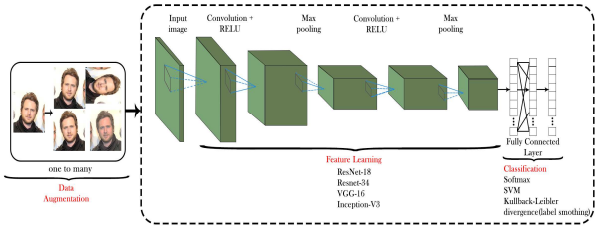

Fig. 1. Proposed structure for deep face recognition problem

The rest of the report is organized as follows: In Sect. II, some related works are mentioned. In Sect. III, deep face recognition formulization under the concept of continual learning is explained. In Sect. IV, the simulation results will be discussed. A discussion on the whole results will be presented in Sect. V, and finally, the conclusion is drawn in section VI.

II. Related Works

A. Deep Face Recognition

After strong image recognition and visual tasks performance using CNNs, countless CNN architectures are proposed as the backbone network for the deep face recognition problem. The DeepFace method is considered as a turning point in deep face recognition [16]. This work uses a CNN with SoftMax loss to tackle the FR problem as a multi-class classification problem. DeepFace is enhanced in several other papers with more challenging conditions [17] [18]. Deep face recognition approaches have been improved dramatically by introducing different loss functions based on specific FR tasks. For example, these works adjust SoftMax loss to enhance the performance of FR systems [19] [20].

Jointly using face verification and face identification results in better feature extraction using CNN named DeepID2 [21]. In DeepID2, features are learned under face identification and face verification signals. This work was improved in [22] and [23].

Using RGB-D images is also investigated by many researchers in order to achieve a more robust FR model. Two modalities approach using a two-level attention-based network is proposed for taking advantage of the image's depth [24]. In this approach, a Siamese CNN is used for RGB and a VGG net for depth modalities, resulting in improved deep FR performance. In the same way, a teacher-student generative adversarial network (TS-GAN) is proposed for obtaining depth from a single RGB image [25]. In this method, a teacher (including a generator and a discriminator) generates synthesized depth, and a student network utilizes those synthesized depths to perform deep face recognition.

B. Continual Learning

In recent works, many researchers have provided approaches for continual learning using longer tasks and more examples. Represented methods so far could be categorized into three main groups based on how task-specific information is stored and used throughout the sequential learning process [26]:

- Replay methods
- Regularization-based methods
- Parameter isolation methods

A generative model is used in replay methods to generate pseudo-samples of previous tasks [12] [13]. Adding regularization terms to loss function and using the concept of distillation loss is another group of works that remove the needs of previous samples (or pseudo samples) [14] [27]. In the parameter isolation method, each task changes its parameters to prevent interference with other tasks and solve the problem of forgetting. Hence, the size of the network structure is considered dynamic[28].

C. Deep Face Recognition with Considering Incremental Batches

Despite the tremendous interest in deep face recognition, a few pieces of research deal with online learning in FR systems. A task-free continual learning method with limited results in online deep face recognition is proposed in [5]. In this work, they attempt to estimate the arrival of a new task by detecting sudden changes in the loss function. An IL structure using distillation loss and neighborhood selection for continual face recognition is introduced in [7]. They presented results using particular parts of LFW [29] and MegaFace [30] datasets.

An end-to-end model in an incremental learning scenario is proposed in [31] (named IL2M) using VGGFace2 for testing face recognition results. In the IL2M method, statistics of old tasks are stored in a second memory for handling model bias problems. In the same concept, another state of art considers deep FR and IL together named ScaIL [32]. In the ScaIL method, catastrophic forgetting is solved by storing classifier weights of past learned tasks and using them to reduce bias toward new tasks.

III. Deep Face Recognition Problem Formulation Under the Incremental Learning Condition

Suppose a model $\mathcal{M}_0$ is trained from scratch using a training dataset $\mathcal{X}_0 = \{(X_0^j, Y_0^j),\ j = 1,2,\dots,P_0\}$ containing P_0 identities. This state is called the initial state (non-incremental state) denoting by $\mathcal{S}^0$. Then in each incremental state $\mathcal{S}^k,\ k > 0$, a new task (batch) with P_k new identities enter, and the previous model $\mathcal{M}_{k-1}$ needs to be updated into the current model state $\mathcal{M}_k$ with the ability to recognize $N_k = \{P_0 + P_1 + \cdots, P_k\}$ identities (classes) are the state of $\mathcal{S}^k$. The training dataset at incremental state $\mathcal{S}^k$ is $\mathcal{X}_k = \{(X_k^j, Y_k^j),\ j = 1,2,\dots,P_k\} \cup \mathcal{B}$ which means all data of P_k identities are available at the state $\mathcal{S}^k$ but only a limited number of exemplars from past tasks $N_{k-1} = \{P_0 + P_1 + \cdots, P_{k-1}\}$ are allowed indicating by $\mathcal{B}$. Finally, using a deep network, the input dataset $\mathcal{X}_{N_k}$ are transformed into a D dimensional feature space: $\mathcal{F}_k: \mathcal{X}_{N_k} \to \mathbb{R}^D$ and using the classifier part (e.g. $C_{N_k} = \boldsymbol{f}_{N_k}^x . \boldsymbol{w}_{N_k}^T + \boldsymbol{b}_{N_k}$) transformation of $\mathcal{X}_{N_k}$ dataset into a set of raw classifiers (logits) is performed. The whole transformation is such as $M_k: \mathcal{X}_{N_k} \to C_{N_k}$. The classifier weights learned in state $\mathcal{S}^k$ and j[th] class denotes as $C_k^j = \{w^1(C_k^j), w^2(C_k^j), \dots, w^D(C_k^j)\}$.

The end-to-end structure for solving the FR-IL problem in this report is illustrated in Fig. 2. In each new task (batch), $\mathcal{X}_{new}$ images of the current task ($\mathcal{S}^k$) along with $\mathcal{B}$ images

belonging to previous batches ($\mathcal{S}^0$ to $\mathcal{S}^{k-1}$) are available, and $\mathcal{X}_{new} \gg \mathcal{B}$. The whole training dataset will be created using both of them. After the preparation of the training dataset, unbalanced fine-tuning will be applied to the whole model. We call it unbalanced fine-tuning because the training dataset is highly biased toward current identities. In each task, model weights are also stored in a second memory and will be used for weight scaling in the next step. Then, model weights are scaled using stored weights in previous tasks. The scaling method (denoting as ScaILin Fig. 2) is described in [32], and the reason is preventing catastrophic forgetting phenomena. Finally, the gallery of previous tasks images will be updated.

Scaling classifier weights performs as follow:

$$w_{sc}^{h}(C_{sc}^{j}) = \frac{\mu_k^{r(h)}}{\mu_i^{r(h)}} \times w^{h}(C_i^{j}) \qquad (1)$$

Where $w_{sc}^{h}(C_{sc}^{j})$ is the scaled version of $w^{h}(C_i^{j})$, $\mu_k^{r(h)}$ is mean activation of new classes, and $\mu_i^{r(h)}$ mean activation of past classes in their initial state. Hence, each weight w^h is scaled using the ratio between mean activations of current and initial states $\mathcal{S}^k$ and $\mathcal{S}^i$ respectively (more detail in [32]).

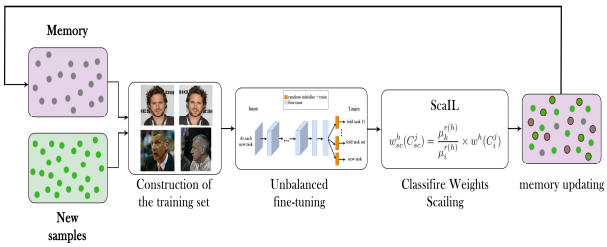

Fig. 2. The proposed method for deep face recognition in the incremental learning scenario

IV. Experiments

A. Deep Face Recognition Results

Deep FR Settings. VGGFace2 dataset is used in this report. We consider the training and testing data as in [32]. In this way, it is possible to compare the results. The training data consist of 50000 images, and the test data are 5000 with considering100 identities.

SGD + momentum(= 0.9) is used as the optimizer with *learning rate* = 0.001 which is divided by 10 when the error plateaus for 15 consecutive epochs and *weight decay* = 0.0001. The total number of *epoch* is 100 with *batch size* = 256. Due to the totally balanced identity distribution, *accuracy* will be used as the performance metric, and there is no need to check other metrics (e.g., F1 score, recall). we consider four different CNN architectures (i.e., ResNet-18, ResNet-34, Inception-V3 and VGG-16) with different loss function and data augmentation.

1) Different data augmentation results: six different data augmentation methods are tested using ResNet-18. We use softmax as the classifier(cross-entropy loss) in this stage.

- **RandomResizedCrop(224)** is the first type of augmentation that crops a random portion of the image and resizes it to a 224*224 image size. The accuracy for this scenario is presented in the first row of Table I.
- **RandomHorizontalFlip()** is the second data augmentation that horizontally flips the given image randomly with a probability of 0.5. This case's performance is shown in the second row of Table I. It could be seen that using this data augmentation improves the performance.
- **RandomRotation(degrees=(-45, 45))** is the third method that rotates the image by angle range [-45 45]. the third row of Table I shows that this method degrades the performance so that it will be removed for the rest simulations.
- **RandomAdjustSharpness(2, p=0.5)** is the fourth data augmentation method that increases the sharpness by a factor of 2 of the image randomly selected with the probability of 0.5. Based on Table I (fourth row), this method also slightly degrades the performance so that it will be ignored in the rest simulations.
- **GaussianBlur((5, 9), sigma=(0.1, 2.0))** is the fifth data augmentation method that blurs images with randomly chosen Gaussian blur using (5, 9) Gaussian kernel. Again this method slightly degrades the performance(fifth row of Table I), so that it will be ignored in the rest simulations.
- **RandomAffine(degrees=0, translate=(0.1, 0.1), scale=(0.2, 0.5))** is the last data augmentation that randomly apply affine transformation of the image keeping center invariant. This data augmentation highly degrades the performance shown in the last row of Table I.

By looking at Table I, it can be seen that the best result obtains (97.7%) when using *ResizedCrop + HorizentalFlip* data augmentation techniques. Therefore, only these two data augmentation methods will be used in the rest of the simulations.

TABLE I. Results of different data augmentation using ResNet-18

ROW#	DATA AUGMENTATION TECHNIQUES	ACC.(%)
1	RandomResizedCrop	97
2	RandomResizedCrop + RandomHorizontalFlip	**97.7**
3	RandomResizedCrop + RandomHorizontalFlip + RandomRotation	96.1
4	RandomResizedCrop + RandomHorizontalFlip + RandomAdjustSharpness	96.8
5	RandomResizedCrop + RandomHorizontalFlip + GaussianBlur	96.7
6	RandomResizedCrop + RandomHorizontalFlip + RandomAffine	90.1

2) Using four different CNN architectures (i.e., ResNet-18, ResNet-34, Inception-V3, and VGG-16) with three different Classifiers (i.e., SoftMax, SVM, Kullback-Leibler divergence): now data augmentation is fixed, and the CNN model along with loss function are changed.

- **SoftMax** loss, a Softmax Activation plus a Cross-Entropy Loss, is the most common classifier used after the CNN feature extractor. The result of face

recognition using mentioned four CNNs is shown in Table II. It could be seen that ResNet-18 and ResNet-34 show almost the same performance. However, ResNet-18 is slightly better, with an accuracy of 97.7%. Moreover, the performance curve of all models is illustrated in Fig. 3.

TABLE II. RESULT OF SOFTMAX CLASSIFIER WITH DIFFERENT NETWORK ARCHITECTURES

Network Architecture	ACC.(%)
VGG-16	96.2
ResNet-18	**97.7**
ResNet-34	97
Inception-V3 (without using AUX. output)	94.2

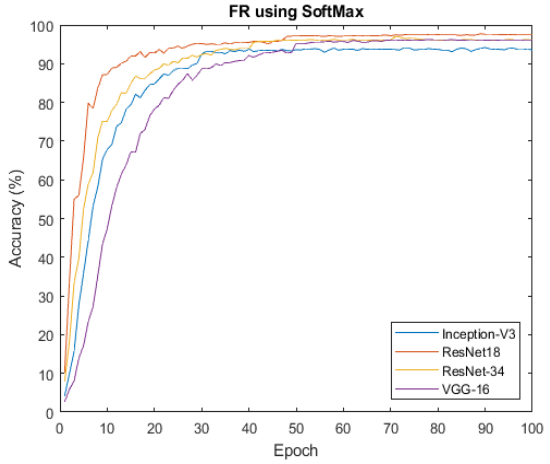

Fig. 3. Performance curve for different CNNs using SoftMax classifier

- **Hinge** loss (margin-based loss) with margin = 1, is another option for the classifier part, based on maximizing the margins of support vectors. This case's performance is shown in Table III, and the performance curve of all models is illustrated in Fig. 4, as well. Again, the best accuracy belongs to ResNet-18 with an accuracy of 89%. However, comparing Table II and TableIII, the superior of Softmax is totally visible. Another difference, compared to Table II, is the result of VGG-16, which shows the lowest performance.

TABLE III. RESULTS OF SVM CLASSIFIER WITH DIFFERENT NETWORK ARCHITECTURES

Network Architecture	ACC.(%)
VGG-16	72.3
ResNet-18	**89**
ResNet-34	88.9
Inception-V3 (without using AUX. output)	85

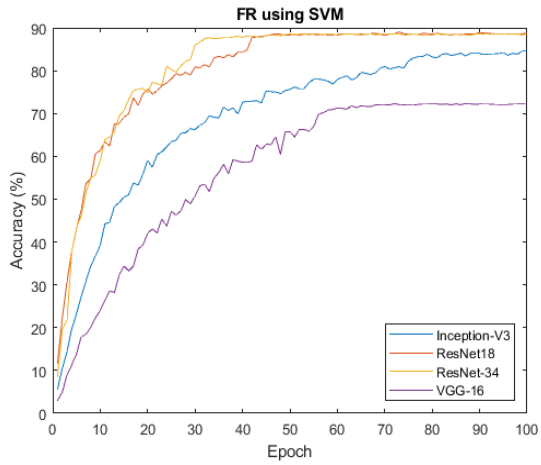

Fig. 4. Performance curve for different CNNs using SVM classifier

- **KL-divergence with label smoothing (kl-div)**, in this loss function, labels will be smoothed as follow:

$$y_{smooth} = (1-\alpha) \times y_{one-hot} + {}^{\alpha}/_{k} \qquad (2)$$

Where k is the number of identities and $0 \leq \alpha \leq 1$ is a hyperparameter that controls the degree of smoothing. In our simulation $\alpha = 0.15$. Then the KL-divergence between smoothed ground truth probability (Q) and probability computed by the model (P) is minimized as (3).

$$D_{KL}(P||Q) = \sum_{x \in \boldsymbol{X}} P(x) \log {}^{P(x)}/_{Q(x)} \qquad (3)$$

TABLE IV. RESULT OF THE KL-DIV CLASSIFIER WITH DIFFERENT NETWORK ARCHITECTURES

Network Architecture	ACC.(%)
VGG-16	96.7
ResNet-18	**98.5**
ResNet-34	98.2
Inception-V3 (without using AUX. output)	95.5

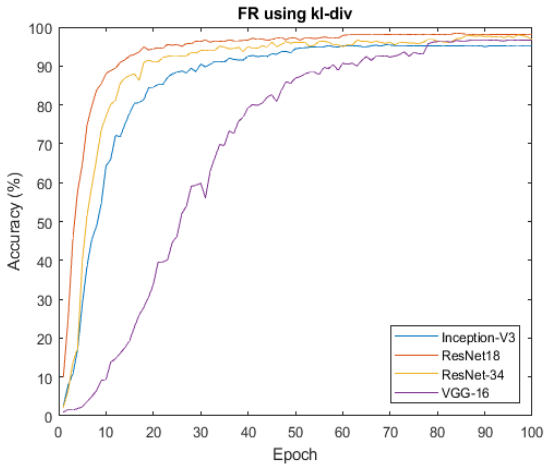

Fig. 5. Performance curve for different CNNs using kl-div loss

By comparing Table II, TableIII, and Table IV, we see that ResNet architecture shows the best results in all cases, and the best result among all cases is using ResNet-18 + kl-div loss.

B. FR-IR Results

In this section, deep face recognition will be investigated under the paradigm of incremental learning (IL). The FR solution in the previous section will consider as the null (non-incremental) state ($\mathcal{S}^0$). Du to high computation resources in FR-IL, we only provide FR-IL results for the best solution, i.e., ResNet-18 with a kl-div loss function.

FR-IL setup. We consider the FR-IL setup as in [32]. The number of whole batches (tasks) is ten ($\boldsymbol{N_k}$**=10**). This includes one non-incremental and nine incremental states. More specifically, after a non-incremental state, 100 new identities will be presented to the model in each new task (i.e., $\boldsymbol{P_k} = \mathbf{100}, k = 1, \dots, 10$), and the model must be trained with these new data. A fixed memory for holding images from previous batches is considered with the limitation of 1% of the full training sets. Therefore, the total training images in each new batch (e.g., kth batch) is 50000 containing 100 new identities, and the number of images in memory for all previous batches (0 to k-1) is only 5000 ($\boldsymbol{\mathcal{B}} = \mathbf{5000}$). It is clear that the number of identities in the null state is 100, and after the occurrence of all ten tasks, there are 1000 identities. In the fine-tuning stage (see Fig. 2), the number of the epoch is 35, $learning\ rate = \frac{0.01}{k+1}, 1 \leq k \leq 9$, and other settings are the same as the Deep FR settings.

The accuracy of each incremental state after fine-tuning without applying the ScaIL method and after using the ScaIL method is illustrated in Fig. 6. As expected, accuracy decreases after the arrival of each new task. However, using the ScaIL technique will improve the performance notably(green). Due to the reporting of top-5 accuracy in the [32] VGGFace2 dataset, we also provide the same metric in our report so that we can compare the results. Fig. 7 shows the top-5 accuracies before using the ScaIL method (yellow and red) and after using the ScaIL method (green and magenta) for both architectures. Looking at Fig. 7, we can see that after using the ScaIL method, the problem of catastrophic forgetting is handled significantly. As before, the kl-div loss results in better performance both in fine-tuning without classifier wight scaling and after classifier weight scaling.

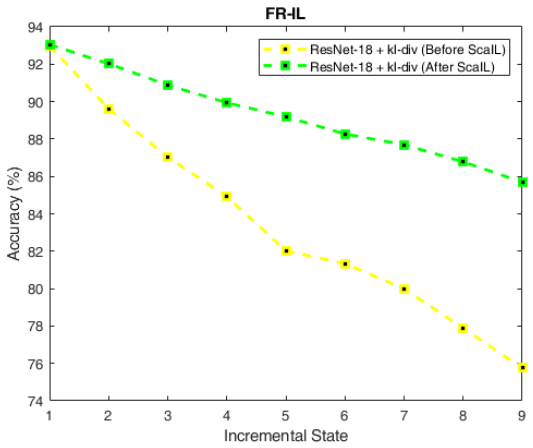

Fig. 6. Accuracy of nine incremental states for ResNet-18 + kl-div, before applying ScaIL (yellow) and after using ScaIL(green)

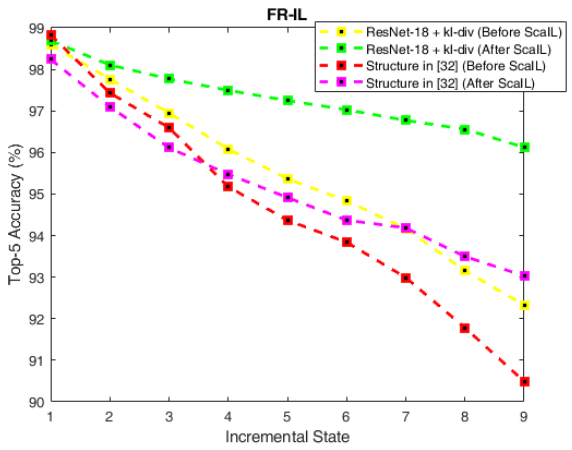

Fig. 7. Top-5 accuracy of nine incremental states using kl-div loss and SoftMax loss

For the final comparison between "ResNet-18 + kl-div" and [32], the mean of accuracy and top-5 accuracy in incremental states (excluding initial state) before using the ScaIL method and after using the ScaIL technique is presented in Table V.

TABLE V. AVERAGE ACCURACY AND TOP-5 AVERAGE ACCURACY WITH TWO DIFFERENT LOSS FUNCTIONS

Model Combination	**Top-5 ACC.(%) (average)**	**ACC.(%) (average)**
ResNet-18 + kl-div	95.47	83.5
ResNet-18 + kl-div **with ScaIL**	**97.3**	89.3
Model in [32]	94.1	Not provided
Model in [32] **with ScaIL**	95.6	Not provided

V. DISCUSSION

Deep face recognition has a wide range of applications in the real world. Hence, we tried to work on some parts of it. **Data augmentation:** The simulations show that model performance will be degraded by adding some augmentation methods (e.g., rotation, gaussian blurring, affine transform) in the data augmentation part. These kinds of augmentations are for testing how robust the model is. We saw that severe changes in scale and translation (i.e., affine transformation)result in high degradation, but minor noise such as blurring and pose changes (rotation) will handle using CNNs.

CNN architectures: in our simulation with all combinations of preprocessing and classifiers, the ResNet structure shows the best performance. It means that using the residual layer is helpful in deep face recognition problems. Looking at [3], we can see that most of the recent state of arts are also used ResNet as their main backbone architecture. The worse performance belongs to VGG-16, which has pretty simple layers. Furthermore, Inception-V3 seems to overfit in our FR simulations.

Loss function: Based on large intra-variations in face images, loss function plays a crucial role in deep face recognition. In our simulation, hinge loss (margin-based loss) showed worse performance. The reason could be focusing on maximizing the absolute difference between margin and instance of input. The softmax loss (using cross-entropy)

showed better performance than hinge loss. The reason could be using entropy instead of the absolute difference margin. Finally, the best result is obtained by using the Kullback-Leibler divergence loss (kl-div) that is a measure to indicate how a probability distribution is different from another probability distribution. In our case, one distribution is the model probability prediction, and the second one is smoothed labels distribution. It seems that (not surprisingly) the distribution of face images contains various vital statistics that cross-entropy with hard labels can not tackle.

FR-IL: we assumed that the best result in a non-incremental state would also result in the best accuracy in incremental states. Hence, we used ResNet-18 + kl-div, and after nine incremental batches the accuracy outperformed [32] (which used Resnet-1 8 + cross-entropy).

VI. CONCLUSION

In this research, we investigated deep face recognition using CNN. Firstly, we explored the effect of noise and pose variation on the performance by using various data augmentation methods. Then we presented the result of four CNN architectures with three different loss functions. Finally, deep face recognition is investigated under the incremental learning assumptions to show the problem of online face recognition in deep networks, which could be considered an open challenge in deep face recognition.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.

[2] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815-823.

[3] M. Wang and W. Deng, "Deep face recognition: A survey," Neurocomputing, vol. 429, pp. 215-244, 2021.

[4] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018), 2018: IEEE, pp. 67-74.

[5] R. Aljundi, K. Kelchtermans, and T. Tuytelaars, "Task-free continual learning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 11254-11263.

[6] E. Belouadah, A. Popescu, and I. Kanellos, "A comprehensive study of class incremental learning algorithms for visual tasks," Neural Networks, vol. 135, pp. 38-54, 2021.

[7] B. Zhao, S. Tang, D. Chen, H. Bilen, and R. Zhao, "Continual representation learning for biometric identification," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1198-1208.

[8] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends in cognitive sciences,* vol. 3, no. 4, pp. 128-135, 1999.

[9] R. Aljundi, P. Chakravarty, and T. Tuytelaars, "Expert gate: Lifelong learning with a network of experts," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3366-3375.

[10] A. Mallya and S. Lazebnik, "Packnet: Adding multiple tasks to a single network by iterative pruning," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 7765-7773.

[11] Y.-X. Wang, D. Ramanan, and M. Hebert, "Growing a brain: Fine-tuning by increasing model capacity," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2471-2480

[12] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "icarl: Incremental classifier and representation learning," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 2001-2010.

[13] A. Chaudhry *et al.*, "Continual learning with tiny episodic memories," 2019.

[14] A. Rannen, R. Aljundi, M. B. Blaschko, and T. Tuytelaars, "Encoder based lifelong learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1320-1328.

[15] . Li and D. Hoiem, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence,* vol. 40, no. 12, pp. 2935-2947, 2017.

[16] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701-1708.

[17] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212-220.

[18] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1891-1898.

[19] A. Hasnat, J. Bohné, J. Milgram, S. Gentric, and L. Chen, "Deepvisage: Making face recognition simple yet with powerful generalization skills," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 1682-1691.

[20] B. Chen, W. Deng, and J. Du, "Noisy softmax: Improving the generalization ability of dcnn via postponing the early softmax saturation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5372-5381.

[21] Y. Sun, *Deep learning face representation by joint identification-verification*. The Chinese University of Hong Kong (Hong Kong), 2015.

[22] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2892-2900.

[23] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," *arXiv preprint arXiv:1502.00873,* 2015.

[24] H. Uppal, A. Sepas-Moghaddam, M. Greenspan, and A. Etemad, "Two-level attention-based fusion learning for RGB-D face recognition," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021: IEEE, pp. 10120-10127.

[25] H. Uppal, A. Sepas-Moghaddam, M. Greenspan, and A. Etemad, "Teacher-Student Adversarial Depth Hallucination to Improve Face Recognition," *arXiv preprint arXiv:2104.02424,* 2021.

[26] M. Delange *et al.*, "A continual learning survey: Defying forgetting in classification tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2021.

[27] J. Kirkpatrick *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences,* vol. 114, no. 13, pp. 3521-3526, 2017.

[28] J. Serra, D. Suris, M. Miron, and A. Karatzoglou, "Overcoming catastrophic forgetting with hard attention to the task," in *International Conference on Machine Learning*, 2018: PMLR, pp. 4548-4557.

[29] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[30] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4873-4882.

[31] E. Belouadah and A. Popescu, "Il2m: Class incremental learning with dual memory," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 583-592.

[32] E. Belouadah and A. Popescu, "Scail: Classifier weights scaling for class incremental learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 1266-1275.