CONNECT.
SOLVE.
CREATE. + AI

kakaoenterprise

BigFace: A New Paradigm Toward High-Performance (Masked) Face Recognition

Technical Details of our submissions at
MFR-ICCV2021, WebFace260M track

Taewan Ethan Kim (ethan.y@kakaoenterprise.com),
Face Group (David), Vision team,

AI Lab@kakaoenterprise

2021 ICCV OCTOBER 11-17 VIRTUAL

FACIAL RECOGNITION WITH MASKS

Acknowledgement

Acknowledgement and Our Members

- We appreciate the organizers of this challenge for a chance to deep dive into the Masked Face Recognition Research.

- Our Members:
  - ***Jong-Ju Shin*** for the preparation of the (masked) dataset, and insightful discussion.
  - ***Pyeong-Gang Lim*** and ***Deepflow team*** for their devoted efforts to developing and maintaining our internal Distributed Machine Learning Infrastructure.
  - ***Yong-Hyeon Kim*** for the initial research of the Masked Face Recognition as our formal colleague.

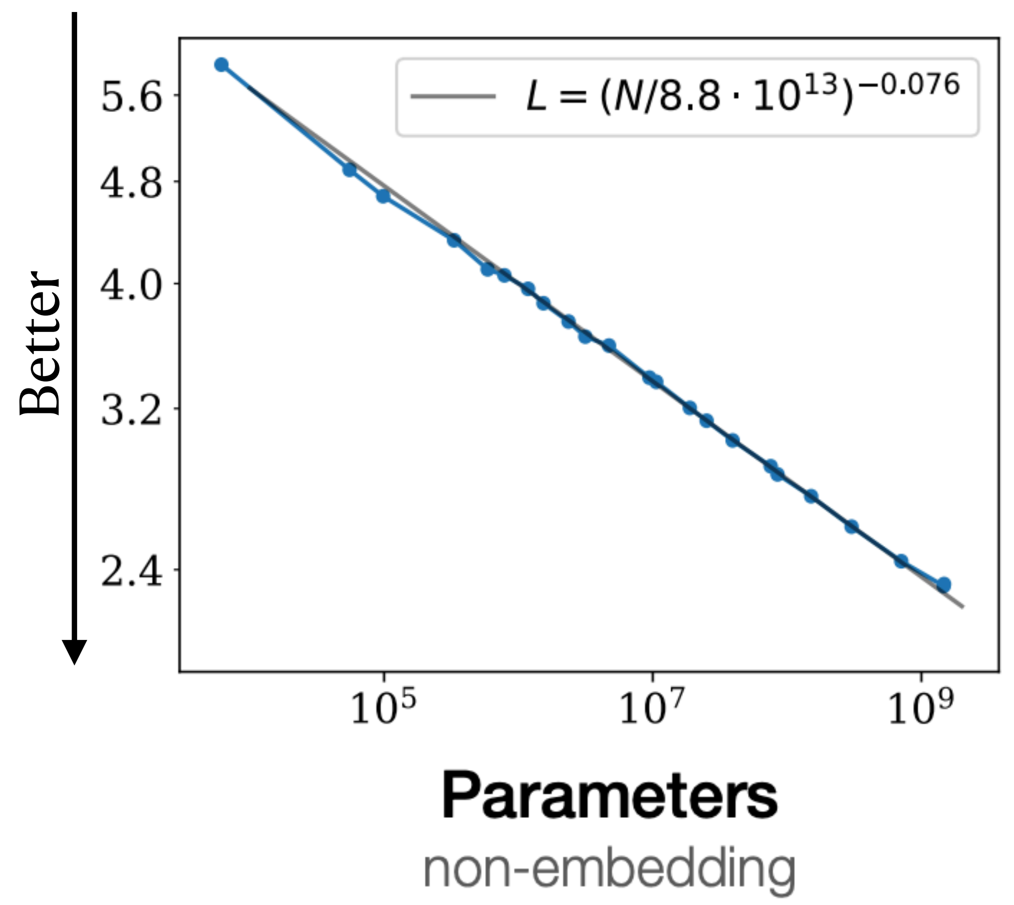
“*BigFace*” paradigm

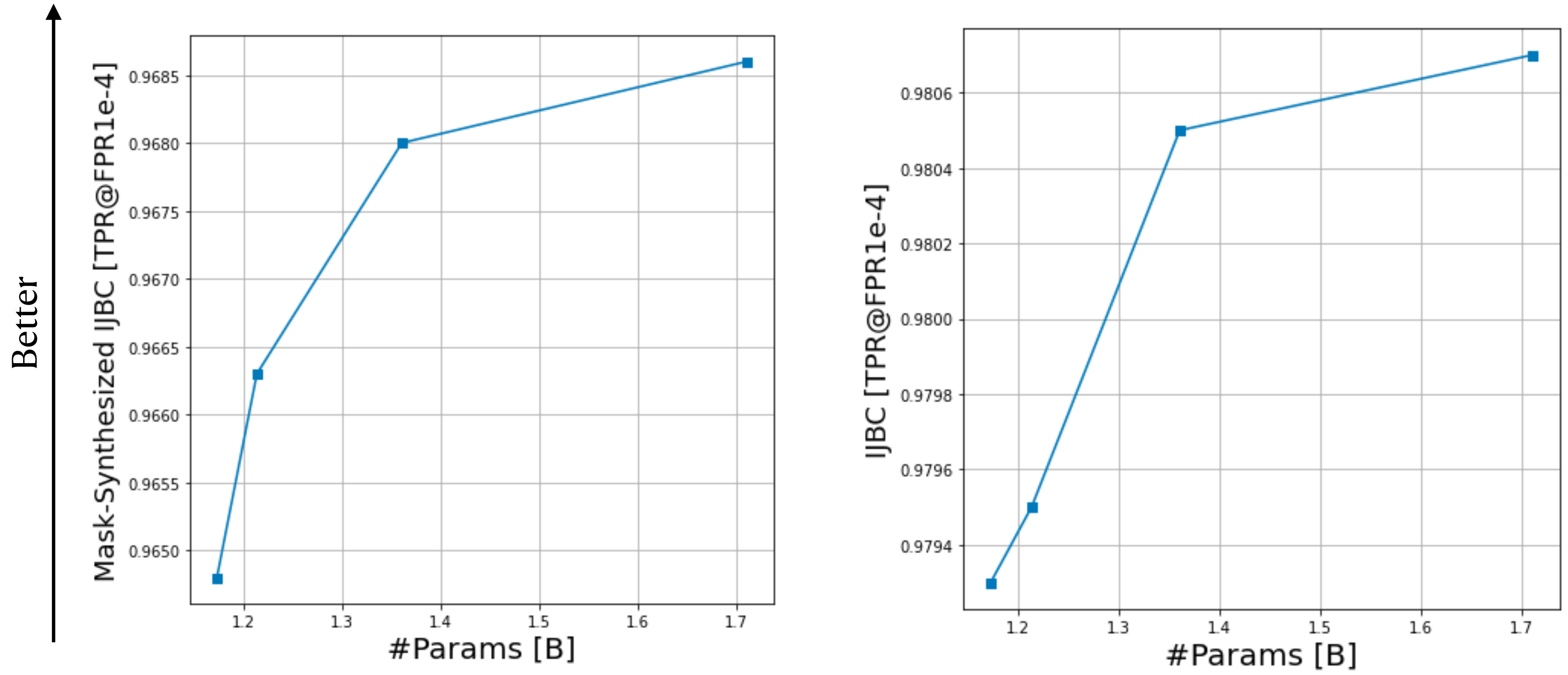
This challenge was a great chance to verify the effectiveness (usefulness) of our ***“BigFace”*** paradigm.

"BigFace" Paradigm

Linear Scaling Raw in the Face Recognition

- Inspired by a lighting surge of Big Models in the Language Modeling Field [1], we believe ***Big Face*** Recognition (FR) Models also give a great impact on performance, if there is no dataset bottleneck.
- We verified FR with mask also is governed by the Linear Scaling Raw (Power Raw) on the WebFace42M [2].





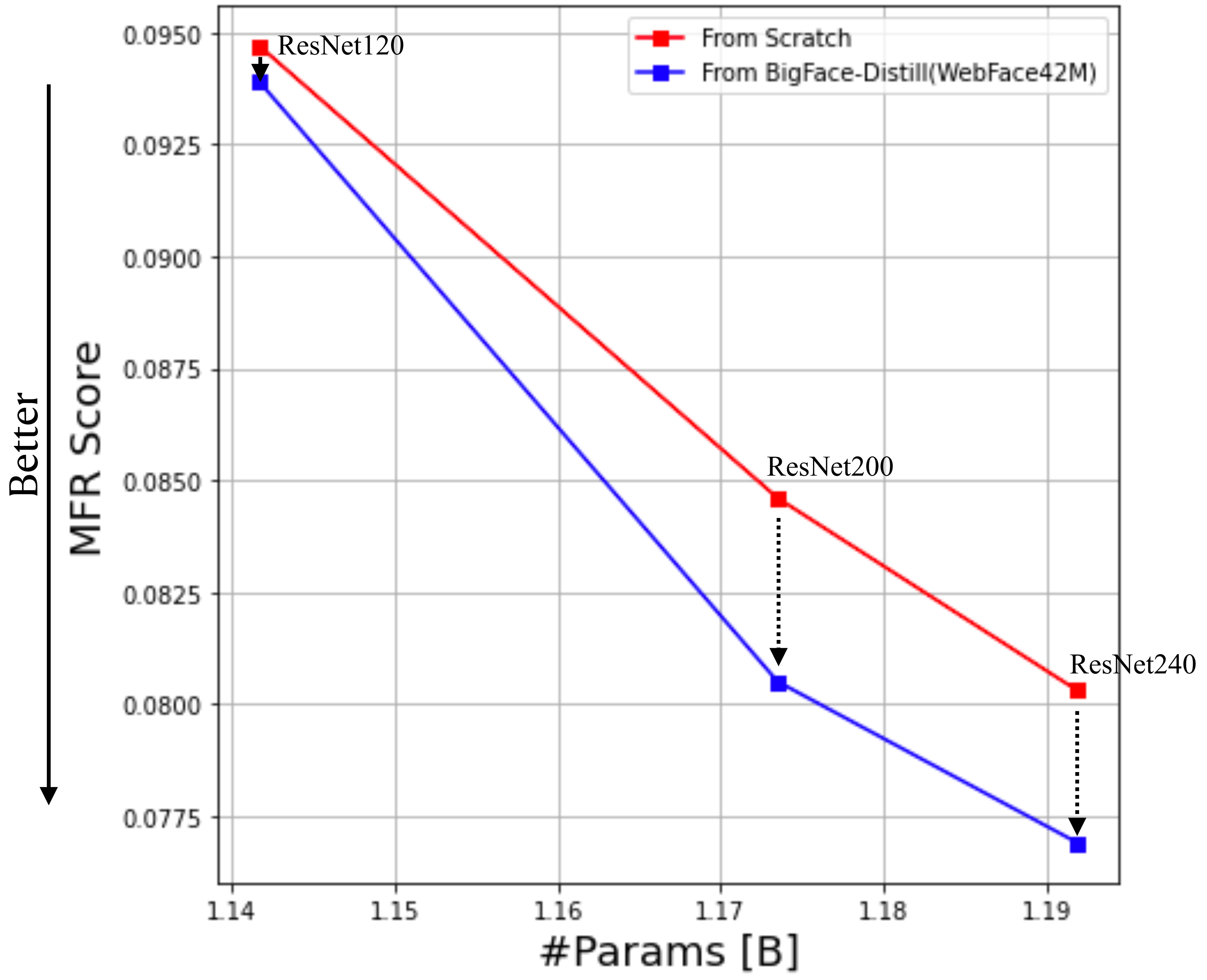
BigFace: Performance as a function of #params. Mask-Synthesized IJBC is generated via two off-the-shelf masked face generation tools. Classification weight (on the WebFace42M) is included to measure the number of parameters.

[1] J Kaplan et. al, Scaling Laws for Neural Language Models, arXiv:2001.08361v1, 2020
[2] Z. Zhu et. al, WebFace260M: A Benchmark Unveiling the Power of Million-scale Deep Face Recognition, CVPR, 2020

“*BigFace*” Paradigm

“BigFace-Distill”, a promising algorithm exploiting such *BigFace* Models

- Many FR applications do not allow the use of such ***BigFace*** Models due to a huge amount of their memory consumption and latency.
- To tackle the aforementioned limitations, we can employ the ***Knowledge Distillation (KD)*** algorithms for extracting informative knowledge from the ***BigFace*** Models into smaller models application-specific.
- This challenge was a great chance to verify the effectiveness (usefulness) of our ***“BigFace”*** paradigm.
- After finishing a training of ***BigFace*** model once, it can alleviate a cost of labeled train-dataset building via the ***Semi-Supervised KD***.



BigFace-Distill: MFR Score as a function of #params. We verified KD from our BigFace model (i.e. ResNet600) improved the Masked Face Recognition (MFR) score over our student models.

Our Approach

Mask-Augmentation, Models, and BigFace-Distill

Mask Augmentation

Combination of two off-the-shelf masked face synthesis tools

- We use two off-the-shelf masked face generation tools [1,2]

Figure 1. Examples of synthesized masked-face images from the WebFace42M . Samples in the first and second row are generated via [1], and [2] respectively. From [1] we can augment synthesized images with various shapes and textures of face masks, but resulting images are slightly unnatural. In contrast, [2] generates relatively natural masked-face images, however, limited to synthesis with various commodity face masks.

- We sampled mask-augmented face images from the WebFace42M ratio of 0.35 rather than 0.5.
- By the sampling ratio, we could roughly steer our interest between MFR and SFR (Standard Face Recognition).

[1] J. A. Anwar., Mask the face, https://github.com/aqeelanwar/MaskTheFace, 2020. Online; accessed 05-Oct-2021
[2] J. Guo et. Al, Insightface, https://github.com/deepinsight/insightface/tree/master/recognition/_tools_, 2020. Online; accessed 05-Oct-2021.

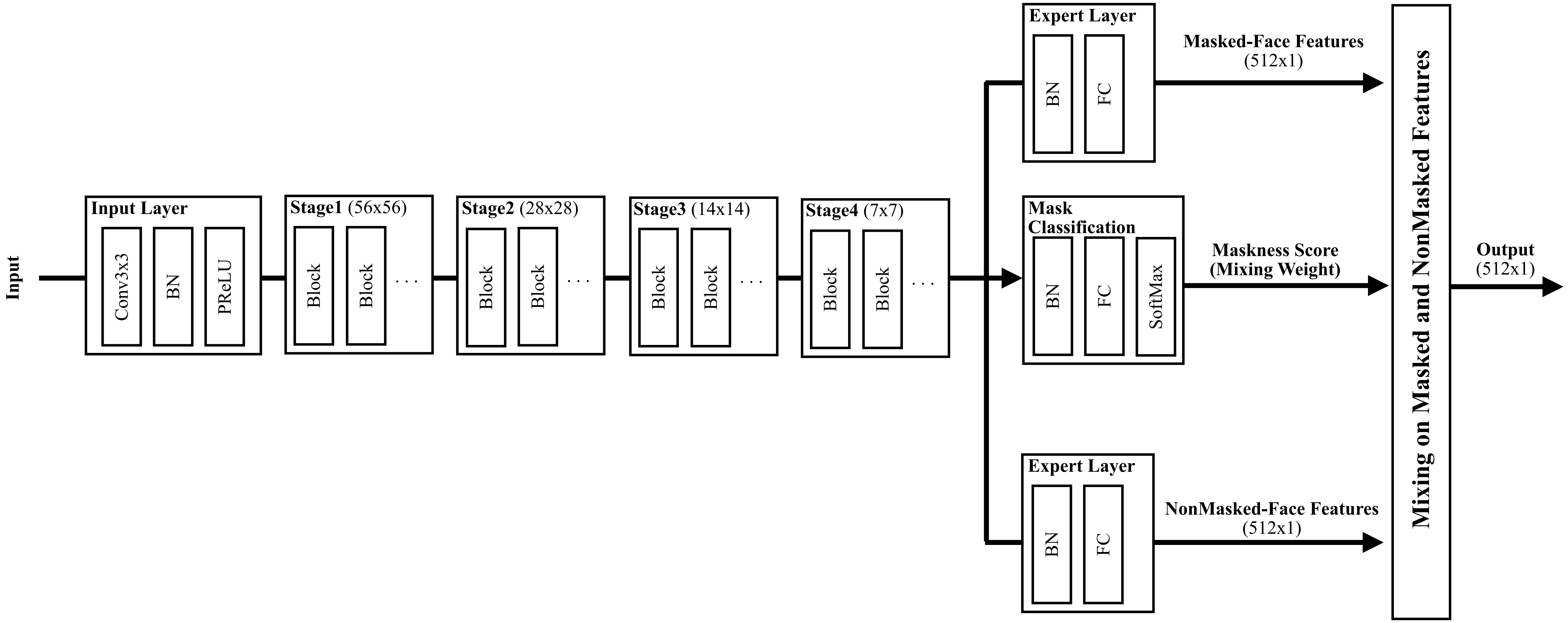
CONNECT.
SOLVE.
CREATE. + AI

kakaoenterprise

Model

Plain ResNet is Enough

- In the first round, we adopted a simple Mixture-Of-Expert (MoE) [1] structure



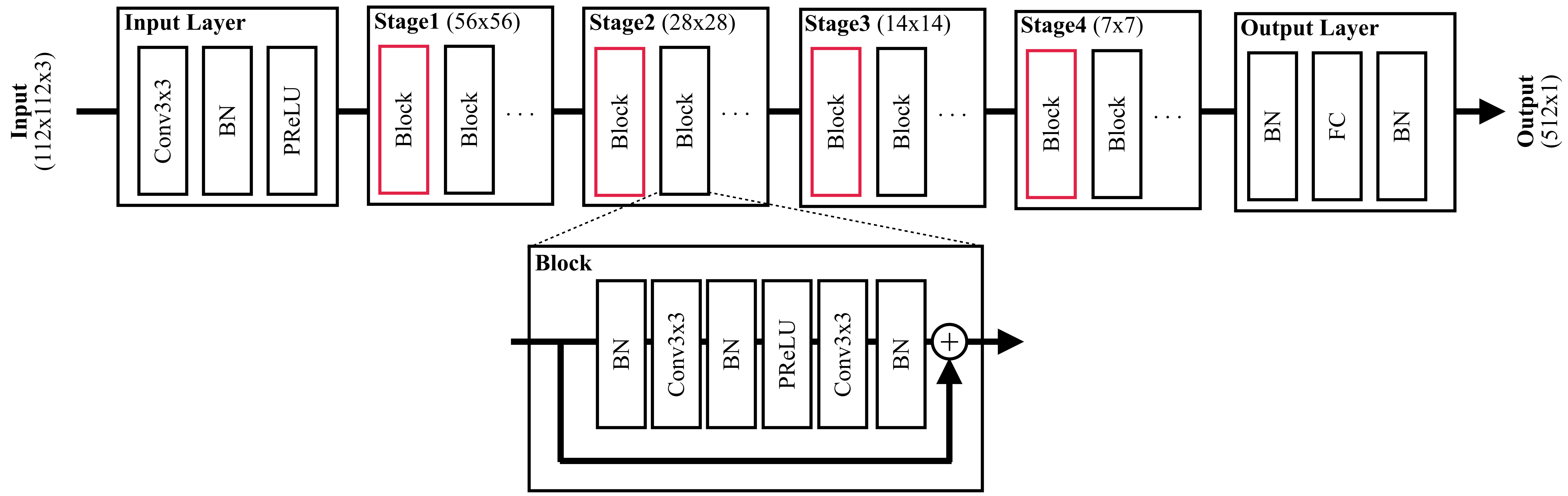
- However, we observed the plain ResNet shows a better performance than that of our MoE variant.

[1] N. Shazeer et. al, OUTRAGEOUSLY LARGE NEURAL NETWORKS: THE SPARSELY-GATED MIXTURE-OF-EXPERTS LAYER, ICLR, 2017

Model

Plain ResNet is Enough

- Block Diagram of the model we used in this challenge [1]:



- Plain ResNet shows much faster inference than that of our MoE.
- Since we did not know the largest model acceptable under the FRUITS-1000 protocol [2], we gradually increased the trainable parameters of our model.
- Finally, ResNet240 was employed in our feature-extraction network.
- In this model, most BN weights and running statistics are absorbed into the previous convolution weights manually after finishing a training process.

Architecture	Number of Blocks			
	Stage1	Stage2	Stage3	Stage4
ResNet120	3	13	40	3
ResNet140	3	15	48	3
ResNet200	6	26	60	6
ResNet240	3	25	88	3
ResNet600	3	70	220	3
ResNet1.2K	3	80	512	3

Table 1. ResNet variants we used in this challenge. The ResNet240 is adapted for the final submission. We utilized the ResNet600 and ResNet1.2K as a teacher model for the knowledge distillation

[1] J. Deng et. al, ArcFace: Additive Angular Margin Loss for Deep Face Recognition, arXiv:1801.07698v3, 2018
[2] Z. Zhu et. al, WebFace260M: A Benchmark Unveiling the Power of Million-scale Deep Face Recognition, CVPR, 2020

BigFace-Distill

Loss Function and the Final Results

- We adopted a convex combination of the CosFace [1] and Angular-Distillation [2,3] loss.

$$\mathcal{L} := \mathcal{L}_{\text{CosFace}} + \lambda \mathcal{L}_{\text{distill}}$$

,where

$$\mathcal{L}_{\text{CosFace}} := -\frac{1}{N}\sum_i \log \frac{e^{s(\cos(\theta_{y_i})-m)}}{e^{s(\cos(\theta_{y_i})-m)} + \sum_{j=1, j\neq y_i} e^{s\cos(\theta_j)}},$$

$$\theta_{y_i} := \frac{\mathbf{W}_{y_i}}{\|\mathbf{W}_{y_i}\|} \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|},$$

$$\mathcal{L}_{\text{distill}} := \frac{1}{N}\sum_i^N \left\| \frac{\mathbf{x}_i^t}{\|\mathbf{x}_i^t\|_2} - \frac{\mathbf{x}_i^s}{\|\mathbf{x}_i^s\|_2} \right\|_2^2 \quad , m = 0.4, s = 64, \text{ and } \lambda = 9.$$

- Our last two submission results:

Submission	Score	
	MFR	SFR
ResNet240	0.0803	0.0235
ResNet240-Distill	**0.0769**	0.0241

Table 2. The last two submission results. We used the ResNet600 as our teacher network in the *ResNet240-Distill* submission.

[1] H. Wang et. al, CosFace: Large Margin Cosine Loss for Deep Face Recognition, arXiv:1801.09414v2, 2018
[2] M. Yan et. al, VarGFaceNet: An Efficient Variable Group Convolutional Neural Network for Lightweight Face Recognition, arXiv:1910.04985v4, 2019
[3] C. Nhan et. al, ShrinkTeaNet: Million-scale Lightweight Face Recognition via Shrinking Teacher-Student Networks, arXiv:1905.10620v1, 2019

BigFace-Distill

Training Details

- The data-parallelism for the feature-extraction network, and model-parallelism for the face-classification weights.
- In the model-parallelism, we adopted an idea from the PartialFC [1] (excluding sampling of classes).
- Micro-batching [2], Gradient-Checkpoint [3], ZeRO-Redundancy Optimizer [4], Mixed-Precision [5] on activation and gradient (including communication of the grad.) etc. were employed as need.
- Batch-size of 256 per each rank, LARS [6] optimizer, initial Learning Rate (LR) of 0.01 (Linear LR Scaling [7].), warm-up LR [7], weight-decay of 0.0005, multi-step LR scheduling, and maximum epochs of 30 were set for all training sessions.
- All models were trained on the PyTorch1.8 framework.
- Training sessions were conducted on 32~64 NVIDIA's Tesla V100 GPU cards.

[1] X. An et. al, Partial FC: Training 10 Million Identities on a Single Machine, arXiv:2010.05222v2, 2021
[2] E. Hoffer et. al, Train longer, generalize better: closing the generalization gap in large batch training of neural networks, arXiv:1705.08741v2, 2018
[3] T. Chen et. al, Training Deep Nets with Sublinear Memory Cost, arXiv:1604.06174v2, 2016
[4] S. Rajbhandari et. al, ZeRO: Memory Optimizations Toward Training Trillion Parameter Models, arXiv:1910.02054v3, 2020
[5] S. Narang et. al, Mixed Precision Training, ICLR, 2018
[6] Y. You et. al, Large Batch Training of Convolutional Networks, arXiv:1708.03888v3, 2017
[7] P. Goyal et. al, Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour, arXiv:1706.02677v2, 2018

Thank you

The "BigFace" paradigm is still ongoing research and we plan to be published shortly.

Taewan Ethan Kim (ethan.y@kakaoenterprise.com),
Face Group (David), Vision team,

AI Lab@kakaoenterprise