Lightweight Deep Convolutional Neural Networks for Facial Expression Recognition

Yanan Wang, Jianming Wu, Keiichiro Hoashi
KDDI Research, Inc., Japan

KDDI Research

Objectives

- *Facial Expression Recognition* (FER) is an intuitive way to analyze human emotions
- *Deep Convolutional Neural Networks* (DNN) can capture fine-grained facial expression features to achieve high FER accuracy

To maintain the high performance and reduce computational resources, we propose a method to build a lightweight DNN based on a VGG pre-trained network for FER task

Motivation

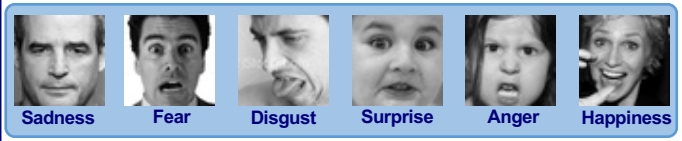

Fig 1. Examples of facial expressions

- *Facial expressions differ among individuals*
- Some facial expressions are very hard to discriminate (ex. “fear” and “disgust”)

The fine-grained facial expression features can not be obtained by current lightweight DNN without increasing the number of parameters *(Such as MobileNets, DenseNet, Xception)*

Proposed method

1. Pretrain a high performance VGG model for FER

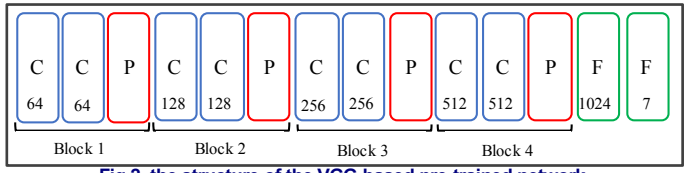

Fig 2. the structure of the VGG-based pre-trained network.
“C”, “P”, and “F” denote the convolution, pooling and fully connected layers

2. Conduct two re-training networks through connecting two different part of the pre-trained networks with **lightweight architectures**

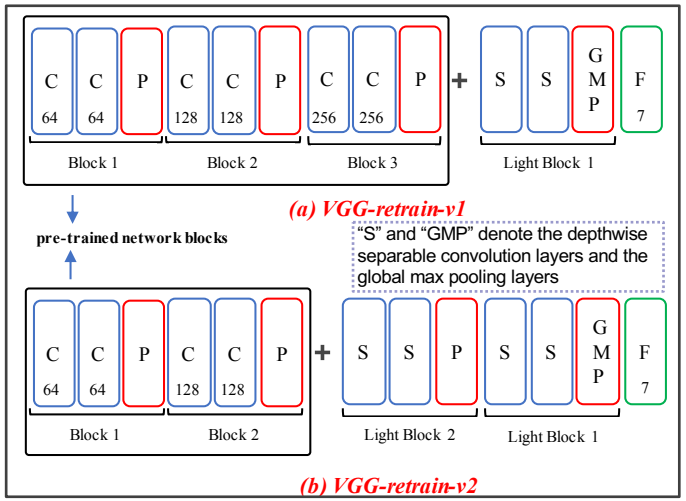

Fig 3. Two types of re-training network structures

Lightweight architectures

Depthwise separable convolutions [Fig 4]

(a) Depthwise: apply a $D_k \times D_k$ *kernel* to each channel
(b) Pointwise: apply a 1×1 *kernel* to the output of depthwise convolution

Number of standard convolution parameters
$D_F * D_F * M * D_k^2 * N$

Approximately D_k^2 times reduction

Number of depthwise separable convolution parameters
$D_F * D_F * M * D_k^2 + D_F * D_F * M * N$

Depthwise Pointwise

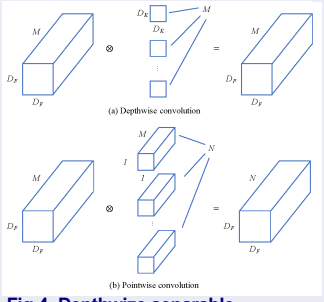

Fig 4. Depthwise separable convolutions structure

Global max pooling [Fig 5]

- Avoid overfitting problem
- Remove the fully connected layers and heavily reduce parameters

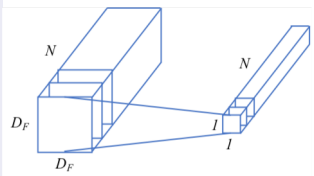

Fig 5. Global max pooling structure

Experiment

Dataset

- ***FER2013:*** Training, validation, and test sets with 28.7K, 3.6K and 3.6K samples
- ***AffectNet:*** Training, validation sets with 58.1K and 3.5K samples
- 7 facial expressions classification task

Table 1. Results of comparison on FER2013

Model	Depth	Accuracy	Parameters
Xception	40	67.4%	20.9M
MobileNets	28	61.8%	3.2M
MobileNetV2	54	62.1%	2.3M
DenseNet-40	40	66.6%	1.0M
VGG-pretrain	10	70.1%	9.4M
VGG-retrain-v1	9	69.5%	2.4M
VGG-retrain-v2	9	**69.8%**	**1.0M**

Table 2. Results of comparison on AffectNet

Model	Depth	Accuracy	Parameters
Xception	40	53.3%	20.9M
MobileNets	28	55.4%	3.2M
MobileNetV2	54	57.9%	2.3M
DenseNet-40	40	59.4%	1.0M
VGG-pretrain	16	61.3%	20.5M
VGG-retrain-v1	9	**61.3%**	2.4M
VGG-retrain-v2	9	61.1%	**1.0M**

Discussion [Table1/2]

- ***Accuracy:***
  VGG-retrain-v1/v2: High accuracy comparable to VGG pretrained model (Retrain models retained the fine-grained facial expression features of *VGG-pretrain* model)
- ***Lightweight strategy:***
  VGG-retrain-v1/v2: Fewer parameters than other lightweight *DNNs (Due to lightweight architectures)*
- ***Network architecture:***
  VGG-retrain-v2 can reduce more parameters than *VGG-retrain-v1* while attaining comparable accuracy

Conclusions

Our proposed lightweight FER method reduced parameters approximately ***1/20 times*** that of the *VGG-pretrain model*, and achieved the highest FER accuracy over other lightweight methods (FER2013: 69.8% and AffectNet: 61.1%)