MicroExpNet: An Extremely Small and Fast Model For Expression Recognition From Face Images

İlke Çuğu, Eren Şener, Emre Akbaş

Department of Computer Engineering
Middle East Technical University

IPTA 2019

Outline

1. **Introduction**
2. **Search for a Compact FER Model**
   - Architecture: Max pooling vs. No pooling
   - Dataset: Random split vs. Subject-independent split
   - Performance: Model size and speed
3. **Knowledge Distillation for FER**
   - Regularization: Model size vs. Teacher's Supervision
   - Hyperparameters: Temperature Analysis
4. **Future Research Directions**

Facial Expression Recognition

- Automatic recognition of basic emotions
- Anger, contempt, disgust, fear, happy, sadness, surprise
- Datasets:
  - CK+ [1]
  - Oulu-CASIA [2]

[1] Lucey, Patrick, et al. "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression." 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. IEEE, 2010.

[2] Zhao, Guoying, et al. "Facial expression recognition from near-infrared videos." Image and Vision Computing 29.9 (2011): 607-619.

Model Family

Model	# of neurons in *fc1*
M	256
S	64
XS	32
XXS	16

- 2 conv layers (conv1, conv2)
- 2 fully-connected layers (fc1, fc2)
- Rectified linear units (ReLU)[3] as activation functions.
- Most of the parameters are at fully-connected layers
- Grid search to test size/performance trade-off

[3]Nair, Vinod, and Geoffrey E. Hinton. "Rectified linear units improve restricted boltzmann machines." Proceedings of the 27th international conference on machine learning (ICML-10). 2010.

Max pooling vs. No pooling

- Facial expressions are located mostly on eyes and mouth [4]
- Hypothesis: max-pooling layers hurt the performance of a FER model as the expressions are sensitive to small, pixel-wise changes around the eye and the mouth.
  - FAILED
  - However, testing environment shapes the problem definition (memorization vs. learning)
  - Specifically, for train/val/test set separation:
    - If random split, **YES**
    - If subject-independent split, **NO**

[4] Ekman, Rosenberg. What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press, USA, 1997.

Notation

- v: no pooling layer
- p_1: only one max pooling layer after conv1
- p_2: only one max pooling layer after conv2
- p_{12}: each conv layer is followed by a max pooling layer

Train/Val/Test Set Separation

	Model	CK+	Oulu-CASIA	Model	CK+	Oulu-CASIA
Random	v_M	97.93%	97.68%	v_{XS}	**93.41%**	**88.73%**
	$p1_M$	**97.99%**	**97.79%**	$p1_{XS}$	91.85%	80.16%
	$p2_M$	97.41%	96.64%	$p2_{XS}$	86.84%	77.88%
	$p12_M$	97.39%	97.47%	$p12_{XS}$	88.07%	77.04%
	v_S	96.65%	92.95%	v_{XXS}	**81.91%**	**73.64%**
	$p1_S$	**96.73%**	**93.22%**	$p1_{XXS}$	69.05%	52.99%
	$p2_S$	94.09%	88.61%	$p2_{XXS}$	77.74%	66.84%
	$p12_S$	94.39%	88.72%	$p12_{XXS}$	78.52%	61.71%

Random split

- Model may see images of the same subject both in training & testing
- Images are numerically different, but visually very similar
- No pooling gives the best result for XS and XXS models
- **Q: Does preserving the pixel information ease memorization?**

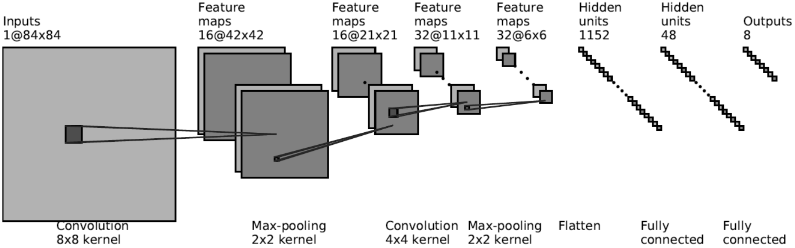
Train/Val/Test Set Separation

	Model	CK+	Oulu-CASIA	Model	CK+	Oulu-CASIA
Subject-independent	v_M	81.23%	60.87%	v_{XS}	77.14%	53.73%
	$p1_M$	**81.57%**	**62.46%**	$p1_{XS}$	77.14%	53.41%
	$p2_M$	78.77%	60.21%	$p2_{XS}$	78.42%	57.51%
	$p12_M$	79.95%	60.53%	$p12_{XS}$	**79.78%**	**57.54%**
	v_S	79.73%	58.18%	v_{XXS}	71.36%	44.33%
	$p1_S$	**81.25%**	**59.49%**	$p1_{XXS}$	67.04%	34.04%
	$p2_S$	78.75%	57.37%	$p2_{XXS}$	76.91%	54.62%
	$p12_S$	79.71%	57.25%	$p12_{XXS}$	**78.44%**	**55.03%**

Subject-independent split

- Model trains with a set of subjects s_1
- It is tested with another set of subjects s_2
- where $s_1 \cap s_2 = \emptyset$
- Having 2 max-pooling gives the best result for XS and XXS models
- **Q: Does information loss improve generalization?**

MicroExpNet Architecture



The final architecture with max-pooling layers

Model size and speed

Model	# of params	Size (MB)	i7-7700HQ	GTX1050	Tesla K40
TeacherExpNet	21.8M	88.13	124.22 ms	83.25 ms	-
FN2EN [2]	11M	42.42	96.08 ms	23.81 ms	13.09 ms
PPDN [1]	6M	23.93	57.18 ms	9.12 ms	13.11 ms
StudentExpNet$_M$	900K	10.88	0.89 ms	1.13 ms	1.74 ms
StudentExpNet$_S$	232K	2.91	0.78 ms	1.08 ms	1.69 ms
StudentExpNet$_{XS}$	121K	1.52	0.63 ms	0.97 ms	1.63 ms
MicroExpNet	**65K**	**0.88**	**0.53 ms**	**0.97 ms**	**1.52 ms**

Memory requirements and average per-image running times

Recap: Knowledge Distillation [5]

Let p_t and p'_s be the softened softmax of the student and teacher respectively whereas p_s is the vanilla softmax of the student:

$$p_t = \frac{e^{z_i/T}}{\sum_j e^{z_j/T}}, \quad p'_s = \frac{e^{v_i/T}}{\sum_j e^{v_j/T}}, \quad p_s = \frac{e^{v_i}}{\sum_j e^{v_j}} \tag{1}$$

Then the cost function becomes:

$$\mathcal{L} = \lambda(\frac{1}{N}\sum_{n=1}^{N}\mathcal{H}(p_t, p'_s)) + (1-\lambda)(\frac{1}{N}\sum_{n=1}^{N}\mathcal{H}(y, p_s)). \tag{2}$$

[5] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 (2015).

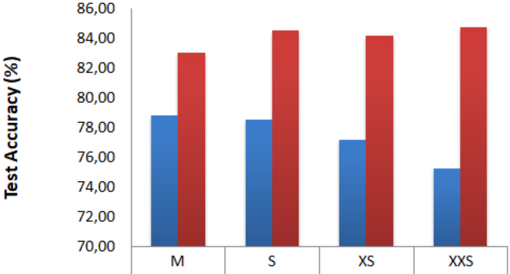
Teacher and Student Models

- TeacherExpNet: Inception_v3 [6] network trained on ImageNet[7]
- StudentExpNet:
  - p_{12}: each conv layer is followed by a max pooling layer
  - M, S, XS, XXS

[6] Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[7] Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." International journal of computer vision 115.3 (2015): 211-252.

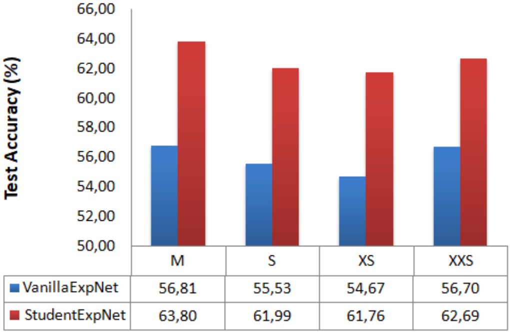
Regularization on CK+ Performance



	M	S	XS	XXS
VanillaExpNet	78,79	78,55	77,22	75,25
StudentExpNet	83,02	84,53	84,18	84,75

The effect of supervision on CK+ for 3000 epochs of training

Regularization on Oulu-CASIA Performance

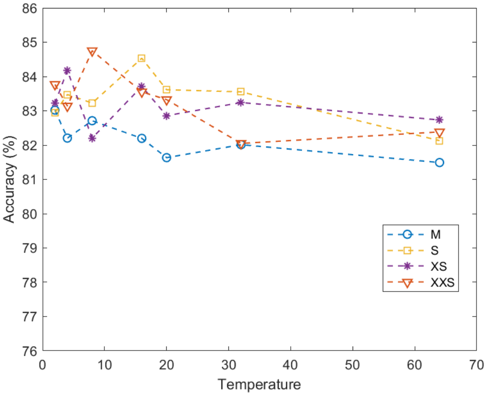


The effect of supervision on Oulu-CASIA for 3000 epochs of training

Temperature Analysis

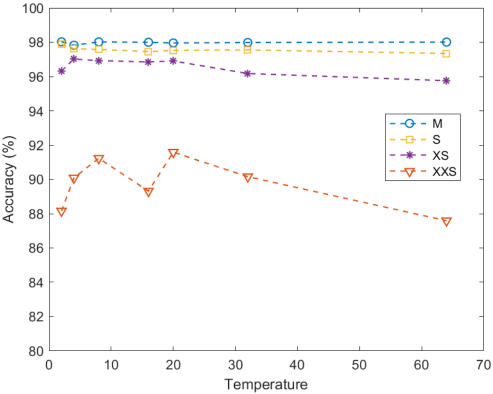
- Grid search for temperatures: $T \in [2, 4, 8, 16, 20, 32, 64]$
- Random split vs. subject-independent split

Temperature Analysis using CK+



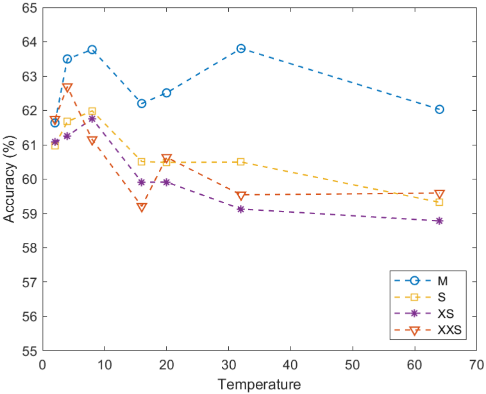
Classification performances of the student networks across different temperatures on the CK+ dataset using **subject-independent splits**

Temperature Analysis using CK+



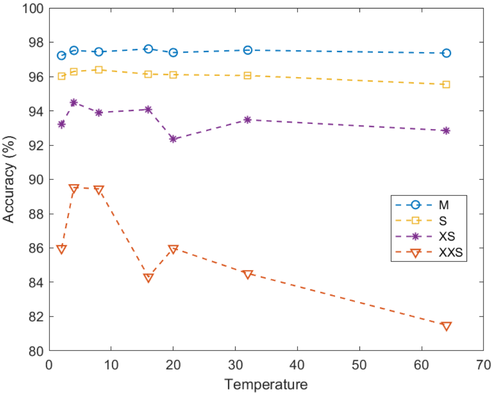
Classification performances of the student networks across different temperatures on the CK+ dataset using **random splits**

Temperature Analysis using Oulu-CASIA



Classification performances of the student networks across different temperatures on the Oulu-CASIA dataset using **subject-independent splits**

Temperature Analysis using Oulu-CASIA



Classification performances of the student networks across different temperatures on the Oulu-CASIA dataset using **random splits**

Future Research Directions

- Is information loss essential for generalization?
- Is a smaller model more open to teacher's supervision?
- Why does the classification accuracy fluctuate as the temperature T is changed?