Vision Transformer Equipped with Neural Resizer on Facial Expression Recognition Task

Paper ID: 1683

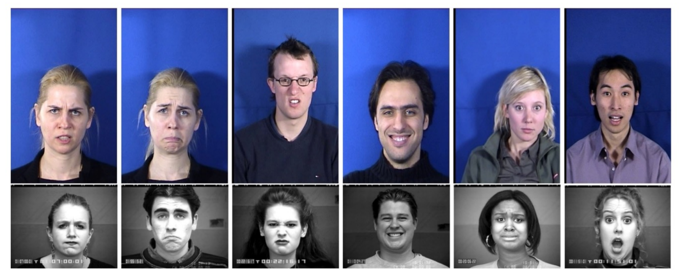
Hyeonbin Hwang[1], Soyeon Kim[1], Wei-Jin Park[2], Jiho Seo[2], Kyungtae Ko[2], Hyeon Yeo[1]

KAIST, Republic of Korea[1]
ACRYL, Republic of Korea[2]

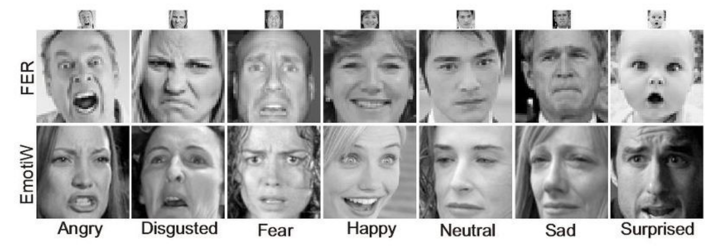
Facial Expression Recognition (FER) Task

Objective: classify expression on face images into several categories.

Dataset Type: "In the Lab" (ITL) vs "In the Wild" (ITW)



source: Deep Convolutional Neural Network for Expression Recognition



source: Deep learning for emotion recognition on small datasets using transfer learning (2015)

Challenges on FER In the wild (ITW) Dataset

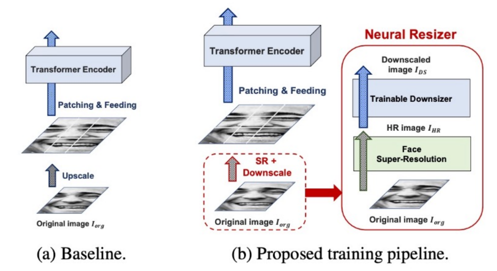
1. Large Image Variances
   - Occlusion, Pose, etc.
   - Generally Low and Non-uniform Image resolution
2. Annotator's subjectivity on emotion classification
   - Class ambiguity – no clear/gold standard.
   - Data imbalance

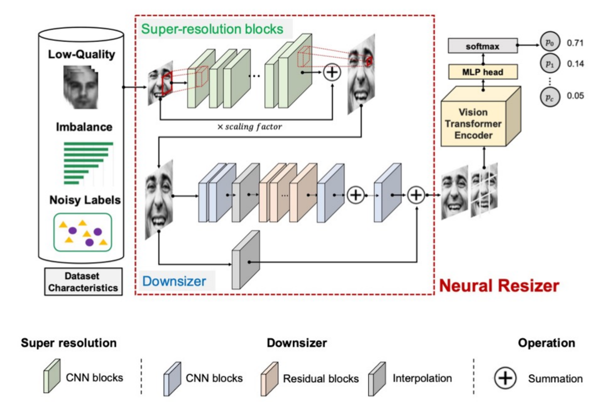
✓ Occlusion, quality(low, high)

Proposed Framework

a. Neural Resizer

- Inspired by [1], we propose a data-driven learnable resizer instead of conventional deterministic Interpolation methods.
- While [1] applies learnable resizer for CNN and downscaling only, our module super-resolutions the input image, and after downsizes the image according to the ViT input size.





b. F-PDLS (Focal Prior Distribution Label Smoothing)

- Assuming less data for a class implies harder classification difficulty, we adopt Focal Loss [2] perspective used in detection to alleviate class imbalance, for our loss function, extending the work in [3].

$$L_{PDLS} = -\sum_{c\in C}\left(t_c * \alpha + d_{kc} * (1-\alpha)\right) * \log(\sigma(z_c))$$

$$L_{F-PDLS} = -\sum_{c\in C}\left((1-\sigma(z_c))^{\gamma} * L_{PDLS}\right)$$

[1] Learning to Resize Images for Computer Vision Tasks, Hossein Talebi, Peyman Milanfar, 2021
[2] Focal Loss for Dense Object Detection, Tsung-Yi Lin et al. 2017
[3] Pyramid With Super Resolution for In-the-Wild Facial Expression Recognition, TH Vo et al. 2020

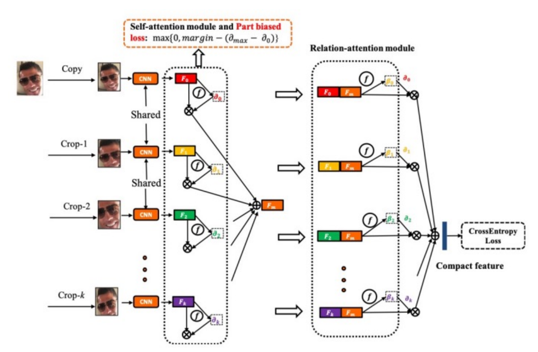
Key Questions

1. How can we effectively feed **low-resolution images** into **Pretrained Vision Transformer?**
2. How can we consider **label ambiguity** and **imbalance simultaneously** in the training process?

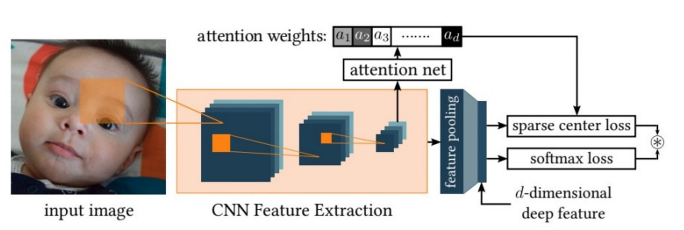
Related Works / Background

Q1. How to capture **fine-grained features** when performing FER task with low-resolution images?

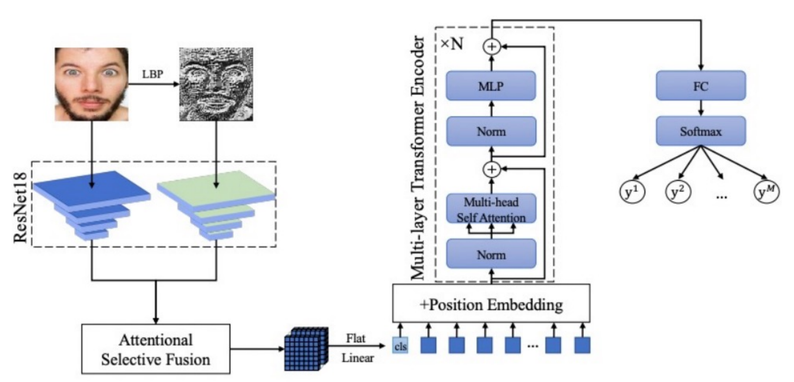
Related Work ①: Use deeper model or (and) employ **attention** mechanism.



Kai Wang et al. Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition, 2019



Related Work ②: Use Vision Transformer, but with input pre-processing network.



Fuyan Ma et al. "Robust facial expression recognition with convolutional visual transformers", 2021

- ✓ In order to use pretrained Vision Transformer, we need to rescale input image to pre-defined resolution.
- ✓ ViT paper (Dosovitskiy, 2020) recommends using images of higher-resolution when fine-tuning.

⇒ **Is traditional interpolation-based upscaling approaches the best in this scenario?**

Q2. How to effectively alleviate **class ambiguity and imbalance** at the same time?

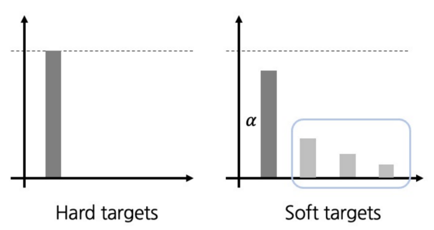
Related Work ①: Use prior distribution to solve ambiguity.

* PDLS = Prior distribution label smoothing

$$PDLS = -\sum_{c\in C}\left(t_c * \alpha + d_{kc} * (1-\alpha)\right) * \log(\sigma(z_c))$$

$$= -\sum_{c\in C} t_c * \alpha * \log(\sigma(z_c)) - \sum_{c\in C} \boxed{d_{kc} * (1-\alpha)(\log(\sigma(z_c))}$$

where
t_c : grount truth target
$\sigma(z_c)$: softmax function for model output z_c
α : hyperparameter related to label smoothing
d_{kc} : prior distribution for correct label k and confusion label c



⇒ How can we design an effective loss function that can also consider **class imbalance**?

Dataset

a. FERPlus: collected from google search engine and consists of about 30K images. All images are in grayscale and resized to 48 x 48. It is also comprised of 8 different emotions - neutral, happiness, surprise, sadness, anger, disgust, fear, contempt -

b. RAF-DB: We use aligned RAF-DB, which consists of about 12K training images, which are all resized to 100 x 100. RAF-DB has images from 7 emotions, excluding contempt from the emotions in FERPlus.

Results

Quantitative

- Neural Resizer with F-PDLS loss generally improves performance with vision transformer variants in general, shown in Table 1.
- We conduct an ablation study on our Neural Resizer and witness the importance of data quality before resizing, shown in Table 2.
- We also conduct an ablation study on our loss function and observe our loss function benefits exclusively with our proposed framework. We hypothesize that our Neural Resizer plays a role as a magnifier to F-PDLS which puts importance on minor and ambiguous samples.
- Finally, we show a performance comparison with some of the state-of-the-art works with respect to the date when the paper was written in table 4.

Table 1: Comparison with various state-of-the-art **small-sized** Transformers on FERPlus, tested with sole backbone architecture and ours

Models	CE + Vanilla	F-PDLS + Proposed
ViT [12]	88.84	88.87
DeiT [41]	88.00	88.09
ConViT [9]	88.12	88.53
XCiT [13]	88.22	88.81
Swin-S	88.69	**89.28**

Table 2: Ablation study on the effect of each module, when **downscaling** and **upscaling** images, tested with **Swin-S**, on FERPlus, using F-PDLS

Setting	model	STN	Up.	Down.	Acc.
a	Swin-S	-	Bi.	-	88.69
b	Swin-S	-	Bi.	LTR	88.53
c	Swin-S	-	SR	Bi.	89.03
d	Swin-S	-	SR	LTR	89.28
e	**Swin-B**	✓	**SR**	**LTR**	**89.50**

Table 3: Comparison across the effect of the loss function, tested on both Vanilla Swin-T and Proposed architecture, **with Swin-B**

Loss	Vanilla	Proposed
Cross-Entropy	88.72	88.87
PDLS [43]	88.69	88.91
F-PDLS (ours)	**88.78**	**89.50**

Table 4: Comparison with other state-of-the-art methods for In-the-wild FER task. * denotes accuracy trained with Swin-Large

Type	Method	FERPlus	RAF-DB
CNN	RAN [45]	89.16	86.90
	SCN [44]	89.35	88.14
	PSR [43]	89.75	88.98
Transformer	LBP + CVT [32]	88.81	88.14
	MVT [24]	88.88	87.03
	VIT + SE [2]	-	86.18
	ours	**89.50**	**88.57***

Qualitative

- To visually deliver the efficacy of our model, we also show some output examples of the Neural Resizer.
- The images show that our framework successfully captures fine-grained features like the line of the wrinkles compared to the deterministic interpolation approaches.
- That is, image shape is not notably changed, but the edges of the discriminant features are more conspicuously accented which facilitates the classification process.

Figure 3: Example of the trainable resizer output. **First row :** the result of deterministic resizer(e.g, bilinear interpolation), the proposed trained resizer and the absolute difference between (a) and (b) from the left. **Second row:** More examples of the difference.

Conclusion

1. We propose a novel training framework to leverage Transformer under the realistic FER with Neural Resizer and F-PDLS.
2. We experimentally show our framework with loss function to improve the performance of Transformer variants in general.
3. We further show that Swin-Transformer achieves competitive results compared to the strong baseline.

Code:
https://github.com/hbin0701/VT_with_NR_for_FER

2022 IEEE International Conference on Acoustics, Speech and Signal Processing