Emotion Recognition from Images

Michel Chamoun | Camille Duchesne | Gurpreet Singh

Problem

Emotion detection through facial expression recognition using various CNN and transformer architectures and testing the impact of transfer learning to solve this task.

Methodology

VGG16 & EfficientNetB0 as a benchmark, we build these models with 7 output classes and we trained the weights from scratch. We then used transfer learning, and used their pre-trained ImageNet weights with the first layers frozen and only trained the top layers. Finally, we used fine-tuning.

Five Layer & Deep CNN we first started building a neural network with only five convolution layers. Then we trained deeper neural networks by varying the convolution layers hyperparameters and adding additional fully connected layers.

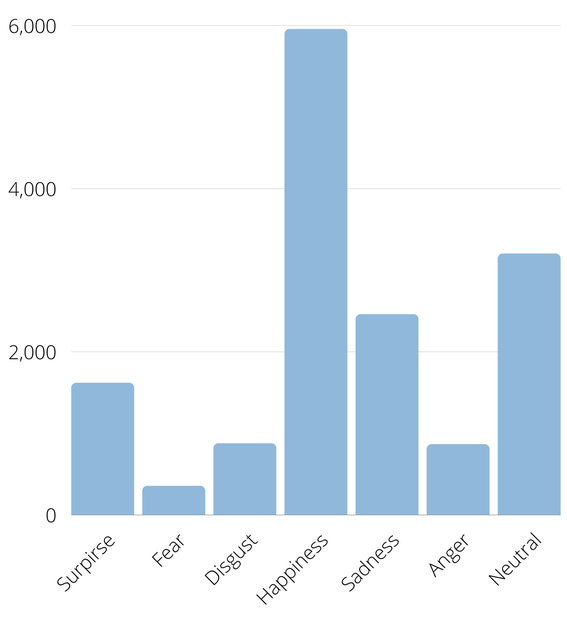
ViT (Visual Transformer) is a SOTA technique in image classification. We started by using the ViT model with attention architecture from the PyTorch library and trained on the RAF-BD dataset. In the second model, we used a pre-trained ViT (imagenet21k) model and fine-tuned it on the RAF-BD dataset.

RAF-BD

15339 colored images of size 100x100

10738 train 4601 validation

7 labels:
Surprise, Fear, Disgust, Happiness, Sadness, Anger, Neutral



CNN

Hyperparameters	
Optimizer	Adam
Loss	Categorical cross entropy
Metric	Accuracy
Learning rate	0.001
Padding	To conserve the size
Max pool	Kernel size = 2x2 Strides = 2
Dropout	Rate = 0.3

Five-layer CNN	
3 back-to-back CNN	64 neurons per layer Kernel size = 3x3 Batch normalization ReLu
Max pool and Dropout	
2 back-to-back CNN	128 neurons per layer Kernel size = 3x3 Batch normalization ReLu
Max pool and Dropout	
Flatten followed by a Dense layer	512 neurons Batch normalization ReLu
Dropout	
Dense layer	7 Neurons Soft-max

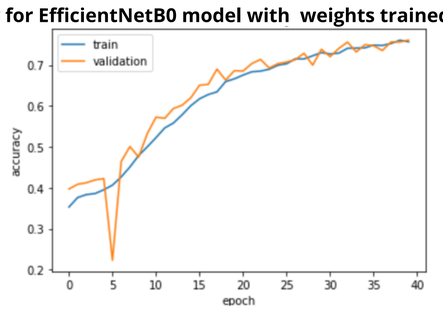
Accuracy: *72.6% after 10 epochs*

Deep CNN	
Stage 1	Convolution layer: 64 neurons, kernel size = 5 Batch normalization ReLu
Stage 2	3 back-to-back convolution layers as in stage 1 Max pooling Dropout
Stage 3	2 back-to-back convolution layers: 128 neurons, kernel size = 3 Max pooling Dropout
Stage 4	Flatten Dense 512 Batch normalization ReLu Dropout Dense 256 Batch normalization ReLu Dropout Dense 128 Batch normalization Dropout
Stage 5	Dense 7 Soft-max

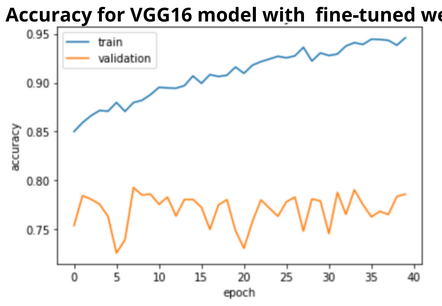
Accuracy: *74.24% after 10 epochs*

VGG16 & EfficientB0

Accuracy for EfficientNetB0 model with weights trained from scratch



Accuracy for VGG16 model with fine-tuned weights



	Efficient NetB0	VGG16
Weights trained from scratch	76.1%	39.7 %
Transfer learning with pre-trained weights	53.8 %	52.9 %
Fine-Tuning (Top 20 layers)	70.1 %	79.3 %

Both models are trained on ImageNet, VGG16 has 138 million parameters while EfficientNetB0 only has 5.3 million parameters.

Visual Transformers

ViT

batch size	64
lr	3.00E-05
gamma	0.7
epochs	30
efficient tranformer	linformer
dim	128
image_size	224
patch_size	32
num_classes	7
channels	3
solver	Adam

Pre-trained ViT

batch size	16
lr	2.00E-04
num_attention_heads	12
epochs	4
transformer	hugging face feature extractor
num_hidden_layers	12
image_size	224
patch_size	16
num_labels	7
channels	3
hidden_size	768

Results & Discussion

We tested our models on 4601 images:

Five Layer CNN	Deep CNN	EfficientNetB0	VGG16	ViT	Pretrained ViT
72%	74%	76%	79%	45%	84%

https://github.com/gurpreetsingh14/emotion_recognition

HEC MONTRÉAL