Sequence-based Facial Emotion Recognition using EfficientNet and LSTM

Celestine Akpanoko, Alex Esser, Srikanth Narayanan, Chang-Yong Song, Hunter Mast
Vanderbilt University

Motivation

- This study performed analysis on sequence-based facial emotion recognition (FER) specifically for valence-arousal classification task using a combination of an efficient CNN and LSTM networks.
- Goal: Enhance an emotion-prediction model to assess emotions over a second interval.
- Hypothesis: Leveraging the aggregated spatial features from a pre-trained EfficientNet model, combined with the temporal modeling capabilities of a LSTM network, will enable a more contextual analysis of emotional expressions.
- Differences in our approach is using sequential image processing techniques like RNN to capture temporal dynamics.

Dataset

- We used the AFEW-VA dataset, found at https://ibug.doc.ic.ac.uk/resources/afew-va-database/ as a basis for this project.
- Armed with plenty of well labeled clips from professionally shot movies, we were able to feed the sequences of frames into the network.
- Preprocessing of data included cropping images to leave only the face, reducing the workload for the model by focusing on the important features in valence and arousal classification.
- Using these images gave a wide array of valence and arousal, with lots of variation between the faces and backgrounds.

Fig 1: Representative sample images from the AFEW-VA Dataset.

Original Model Architecture

{ }

- Found at https://github.com/av-savchenko/face-emotion-recognition/blob/main/models/affectnet_emotions/enet_b0_8_va_mtl.pt, we used a pre trained model designed to recognize facial emotion on still images.
- This model was trained on the AffectNet dataset.
- With over 3,000,000 parameters, the original model is very robust and able to both detect faces as well as classify the valence and arousal of those images with over 60% accuracy in ideal use cases on the original AFEW.
- This models is popular for being fast and somewhat lightweight for their capability, coming in at only 30MB

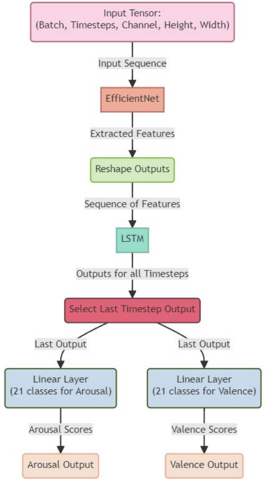
{ }

Model Modifications

- We employed the use of an LSTM layer.
  - 2 Internal layers & hidden dimension of 256 and 1280 input neurons.
- For classification, we take the last time step of the LSTM output.
  - Run through 2 parallel linear layers with 21 classes.
  - Each to give 2 separate outputs, one for arousal and a second for valence.

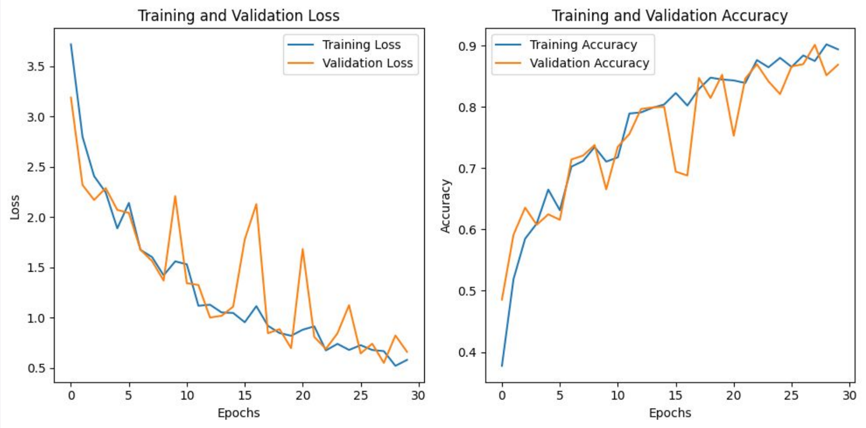
Algorithm 1 Training and Validation with Early Stopping

```
Initialize: best_val_loss ← ∞, trigger_times ← 0
for epoch = 1 to num_epochs do
    Train for one epoch and calculate train_loss, train_accuracy
    Validate and calculate val_loss, val_accuracy
    Print training and validation results
    Update train_losses, val_losses, train_accs, val_accs
    if val_loss < best_val_loss then
        best_val_loss ← val_loss
        trigger_times ← 0
    else
        trigger_times ← trigger_times + 1
        if trigger_times ≥ patience then
            Early stop
            break
        end if
    end if
end for
```



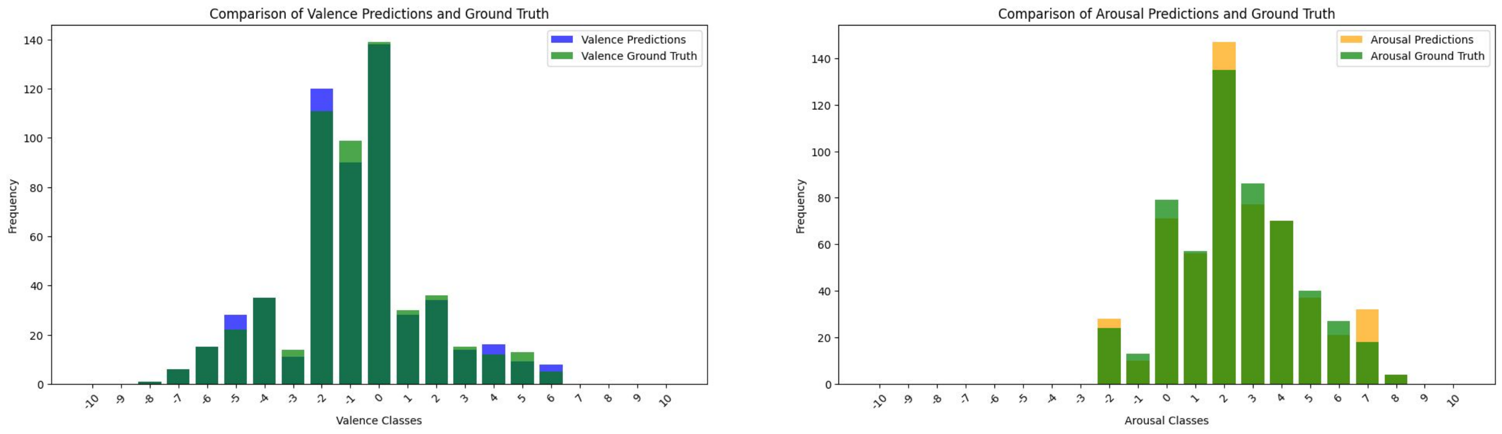
Results

- Our sequence-based FER model integrating CNNs and LSTMs show results in accurately identifying facial expressions over time.
- Training and validation phases depicted shows the effectiveness in learning and generalizing across multiple datasets.
- Consistent decrease in loss and increase in accuracy demonstrates proficiency in capturing and understanding dynamic emotional cues.



Results

- Performance was measured using loss values and accuracy measures.
  - Lower loss values shows agreement between predicted and actual arousal and valence scores.
  - Accuracy measures computed based on predictions of arousal and valence categories.



Conclusion

- We proposed an approach for FER utilizing a hybrid model combining LSTM networks with a pre-trained EfficientNet architecture.
- Our methods leveraged temporal dependencies in video frame sequences to enhance expression analysis.
    - Focus on capturing dynamic changes in valence and arousal over time.
- Utilizing AFEW-VA dataset allowed for a rich resource for training and evaluating model by continuous annotations of emotional states in diverse contexts.
- Demonstrated effectiveness by capturing emotional expressions and predicting categorical valence-arousal values
- Overall, we contributed to the advancement of FER techniques by introducing a novel hybrid model architecture and demonstrating the capability of capturing temporal dynamics of emotions.