Optimizing Facial Expression Recognition Tasks for Children's Faces Utilizing CNN Architectures

Hovhannes Broyan
hbroyan@ucsd.edu

Anthony Li
all010@ucsd.edu

Arjun Malleswaran
amallesw@ucsd.edu

Tauhidur Rahman
trahman@ucsd.edu

Abstract

Detecting psychological disorders in children is a complex undertaking. Children's behavior represents a large and unknown spectrum which can make it difficult to understand the upbringing of certain behaviors. To achieve this pre-diagnosis, previous studies have looked at using facial expression models to assist in detecting emotional responses to specific tasks. However, a pitfall presented is that these models tend to be trained on adult faces and do not generalize well when predicting children's emotional responses. To better address this challenge, our study utilizes a diverse array of pre-trained models that are based on advanced architectures of convolutional neural networks (CNNs) and transformers. Specifically, we explore the capabilities of models such as DeepFace and ResNet (which are CNN-based), alongside Vision Transformers (a transformer-based model), and EfficientNet (another CNN-based model). These models will be evaluated on the FER-2013 dataset, encompassing a wide age range, and the Dartmouth Database of Children's Faces, focused exclusively on children. By comparing the performance of these models across the two datasets using common metrics, we aim to delineate the differences in predicting emotional responses between children and adults.

Code: https://github.com/amallesw/DSC180B-Q2-Codebase/tree/main

1 Introduction

1.1 Context

Diagnosing psychological disorders in children remains a difficult and nuanced problem. Such diagnoses often require discerning whether certain children's behaviors are indicative of early psychopathology or simply deregulated behavior not associated with any disorder. To provide support for parental guardians in the pre-diagnosis process and provide them with a clearer understanding of their child's psychological health, accurate and reliable diagnostic tools are a necessity. However, the efficacy of current models for such tasks is compromised due to improper training, particularly with regard to the age of data image subjects.

1.2 Previous Study Limitations

Understanding and diagnosing early psychological disorders are important for addressing children's educational and developmental challenges. Previous studies Kalanadhabhatta et al. (2023, 2022); Zhang et al. (2023) have looked into children's emotional responses to various reactions to psychological tasks and games. While facial expression recognition models have been used to quickly assess these emotional responses, their effectiveness is limited due to the models being predominantly trained on adult faces. Children display a wider range of emotions than adults, making children's emotion detection a unique challenge. Along with factoring in children's facial structure compared to adults can present biases in facial expression models that cater much more to the way adults put forward emotions than children.

1.3 Approach and Model Choices

To address these challenges of facial expression recognition across children and adult faces, our study will use two datasets: the FER-2013 dataset, which includes faces from a broad age range, and the Dartmouth Database of Children's Faces, focused exclusively on images of children. We utilized a multitude of pre-trained architectures to analyze and predict emotions:

- **DeepFace and ResNet**: Chosen for their deep learning capabilities, offering robust feature extraction suitable for complex facial expression recognition.
- **EfficientNet**: Known for its efficiency in scaling model size, making it adept at handling varied datasets.
- **Vision Transformer (ViT)**: Incorporates self-attention mechanisms, providing a novel approach to understanding spatial hierarchies in images.
- **VGGNet**: Its deep, yet simple architecture is useful for analyzing image details, contributing to the nuanced detection of expressions.

- **Custom CNN**: Serves as a baseline to evaluate the advancements offered by more complex, pre-trained models.

To address these challenges p

Our project aims to tackle fine-tuning existing emotional recognition architectures and create a model that generalizes well to children's facial expressions. By leveraging techniques such as data augmentation and transfer learning from various pre-trained models, we aim to develop a model that thrives under real-world conditions.

we aim to develop a model that thrives under real-world conditions. Employing transfer learning from various pre-trained

2 Methods

2.1 Data Preprocessing

To ensure consistency and optimize runtime across both datasets, we implemented preprocessing steps to standardize the data format, making our code reusable for any dataset. This involved scaling all images to 48x48 pixels to reduce computational load while retaining sufficient detail for expression analysis. Each image's pixels were then converted to RGB values (ranging from 0 to 255) and concatenated into a single string for efficient storage and access. The datasets used, predominantly the Dartmouth Database of Children's Faces, contain images with perfect conditions of lighting and angles.

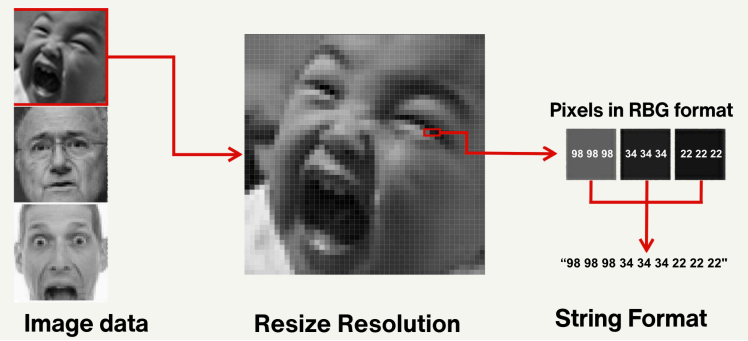

Figure 1: Data preprocessing; creating flattened, concatenated RGB values in string format.

2.2 Data Augmentation

Since real-world environmental conditions vary, we incorporated data augmentation techniques to enhance model exposure to diverse lighting conditions and partial face occlusions. This involved using OpenCV for eye detection and occluding them with black squares, simulating scenarios where facial expressions might be partially obscured. Additionally, we

adjusted the brightness of images randomly within a range of 0.3 to 0.9, effectively doubling our dataset and ensuring our models are trained on data that closely mimics natural variations in real-world images.

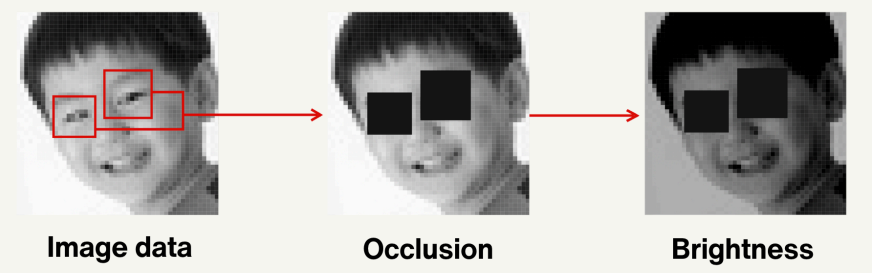

Figure 2: Data augmentation; occlusion of eyes & simulating lighting conditions.

2.3 Models and Metrics

Our evaluation framework targets the effectiveness of six architectures, along with a custom CNN, against the intricacies of children's facial expressions within the FER 2013 and Dartmouth datasets. This section details their application and performance insights:

- DeepFace and ResNet were used in benchmarking deep learning efficacy, with ResNet's residual connections highlighting the potential for deep networks without the gradient vanishing issue.
- EfficientNet's complexity and scalability was tested against the unique demands of facial expression data.
- Vision Transformer offered a contrasting approach to conventional CNNs, its performance shedding light on the role of self-attention mechanisms in facial expression recognition.
- VGGNet allowed for an examination of how deeper convolutional layers affect the model's ability to discern subtle expressions.
- The Custom CNN provided a foundational comparison point, illustrating the value added by complex model architectures and pre-training.

Models were assessed using metrics like accuracy, F1 Score, and a normalized confusion matrix, with a data split of 70/15/15 for training, validation, and testing, to offer a comprehensive view of each model's capability in addressing the challenges of facial expression recognition in children.

2.4 Transfer Learning Methods

Recognizing the specific challenges of facial expression detection in children, we applied transfer learning to tailor model performance. This involved two strategies:

1. **Adapting Output Layers**: The first transfer learning method we used was changing the last layer of the model to output seven features, to match the number of emotions

we are classifying. This proved to not be beneficial in boosting performance and thus, was left out in the final renditions of our models.

2. **Training on Embeddings**: The second method we used was training on embeddings. When letting the pre-trained models process our images, we extracted the output of the last layers and saved it as an embedding. This allows us to save trained features from pre-trained models, which can be used on classification models such as a neural network, SVM, or random forest. This results in faster learning and boosting performance while allowing the reduction of computational resources. When extracting embeddings from pre-trained models, two different routes were provided: full layer training or partial layer training.
   (a) Full layer training involves fine-tuning all layers of a pre-trained model and extracting those embeddings used for a classification model.
   (b) Partial layer training involves freezing layers while keeping a select few layers for fine-tuning throughout the training process. Layers that showed more important core features of an image were kept unfrozen.

These methods allowed us to leverage pre-trained model strengths while focusing the learning process on features most relevant to children's facial expressions, optimizing performance and efficiency.

3 Results

The evaluation focused on three models which demonstrated superior baseline performance: Resnet18, Vision Transformer, and EfficientNetB0. These models underwent further fine-tuning through transfer learning methods. Performance was assessed on the FER2013 and the Dartmouth Database of Children's Faces under three conditions: baseline, partial layer training, and after data augmentation on the training data. The models were evaluated based on accuracy and F1 Score.

3.1 Resnet18

Resnet18 Performance (FER2013)		
Model Kind	**Accuracy**	**F1 Score**
Baseline	0.548	0.542
Partial Layers	0.611	0.608
Data Augmentation	0.649	0.641

Resnet18 Performance (Dartmouth)		
Model Kind	**Accuracy**	**F1 Score**
Baseline	0.604	0.597
Partial Layers	0.653	0.655
Data Augmentation	0.757	0.754

Figure 3: Results of Resnet18's Performance on FER2013 and Dartmouth.

For the ResNet18 architecture, the baseline model achieved an accuracy of 0.548 and an F1 Score of 0.542 on the FER2013 dataset. With partial layer training, a slight improvement was observed, bringing the accuracy to 0.611 and the F1 Score to 0.608. The application of

data augmentation techniques yielded the most substantial improvement, with the accuracy increasing to 0.619 and the F1 Score to 0.641 on the same dataset.

On the Dartmouth Database, baseline figures were 0.604 (accuracy) and 0.597 (F1 Score). Partial layer training led to slight improvements, achieving an accuracy of 0.653 and F1 Score of 0.655. Data augmentation had a significant impact, increasing accuracy to 0.757 and F1 Score to 0.754.

3.2 Vision Transformer

Vision Transformer Performance (FER2013)		
Model Kind	**Accuracy**	**F1 Score**
Baseline	0.637	0.593
Partial Layers	0.657	0.652
Data Augmentation	0.697	0.694

Vision Transformer Performance (Dartmouth)		
Model Kind	**Accuracy**	**F1 Score**
Baseline	0.642	0.631
Partial Layers	0.668	0.661
Data Augmentation	0.729	0.728

Figure 4: Results of Vision Transformer's Performance on FER2013 and Dartmouth.

The Vision Transformer model began with a baseline accuracy of 0.637 and F1 Score of 0.593 on the FER2013 dataset. Partial layer training marginally increased these to 0.657 (accuracy) and 0.652 (F1 Score). Data augmentation significantly boosted performance, yielding an accuracy of 0.697 and F1 Score of 0.694.

On the Dartmouth dataset, the baseline yielded an accuracy of 0.642 and an F1 Score of 0.631. The application of partial layer training improved the accuracy to 0.668 and the F1 Score to 0.661. Notably, data augmentation provided the greatest boost, resulting in an accuracy of 0.729 and an F1 Score of 0.728.

3.3 EfficientNetB0

EfficientNetB0 Performance (FER2013)		
Model Kind	**Accuracy**	**F1 Score**
Baseline	0.515	0.511
Partial Layers	0.530	0.496
Data Augmentation	0.571	0.534

EfficientNetB0 Performance (FER2013)		
Model Kind	**Accuracy**	**F1 Score**
Baseline	0.520	0.511
Partial Layers	0.558	0.542
Data Augmentation	0.585	0.549

Figure 5: Results of EfficientNetB0's Performance on FER2013 and Dartmouth.

EfficientNetB0 started with a baseline accuracy of 0.515 and F1 Score of 0.511 on the FER2013 dataset. Partial layer training slightly increased accuracy to 0.530 but reduced the F1 Score to 0.496. Data augmentation improved both metrics to an accuracy of 0.571 and F1 Score of 0.534.

On the Dartmouth dataset, the baseline showed an accuracy of 0.520 and F1 Score of 0.511. With partial layer training, these metrics improved to 0.558 (accuracy) and 0.542 (F1 Score). Data augmentation positively impacted the results, enhancing accuracy to 0.585 and F1 Score to 0.549.

4 Discussion

4.1 The Impact of Data Augmentation

Data augmentation was shown to be a beneficial technique in enhancing the performance of facial recognition models. By introducing variations such as altered lighting conditions and simulated occlusions, data augmentation significantly improved both accuracy and F1 Scores across all models. This reinforces the concept that training models on a more varied dataset can lead to better generalization, and required qualities for real-world applications where lighting and facial obstructions can vary widely.

4.2 Partial Layer Training's Subtler Role

Although partial layer training resulted in more modest improvements compared to data augmentation, its contribution towards refining the models' capabilities proved beneficial. By fine-tuning the models closer to their output layers, and taking advantage of pre-trained models' ability to learn deep features only, we observed an improvement in their ability to extract and interpret relevant features from children's facial expressions.

4.3 Architectural Sensitivities in Model Performance

The differential improvements observed among the ResNet18, Vision Transformer, and EfficientNetB0 models show the importance of architecture choice in facial recognition tasks. ResNet18 and Vision Transformer's notable responsiveness to data augmentation and partial layer training suggests these architectures may be better suited for capturing children's expressions. In contrast, the relatively poor performance of EfficientNetB0 highlights the necessity of matching model architecture to the specific challenges of the task at hand.

5 Conclusion

5.1 Summary of Findings

Our focus on facial recognition problems geared toward children's emotional expressions has displayed the performance of many different neural network architectures and trans-

ferring learning methods. The usage of the FER2013 and the Dartmouth Database of Children's Faces dataset has proven a difficult task of accurate emotion classification for children's faces in comparison to adult faces.

The performance outputs for each pre-trained model indicate that data augmentation and transfer learning, specifically partial layer training, helped in improving accuracy and F1 scores. Data augmentation has proven to significantly improve performance, demonstrating that models trained on more diverse datasets can help better capture the subtle emotional expressions seen in children's faces. Both ResNet18 and Vision Transformer, along with the previously mentioned techniques used, excelled in performance in comparison to other architectures.

5.2 Importance of Age-Specific Data

The difference in performance metrics between the Dartmouth and FER2013 datasets demonstrates the importance of training the models on age-specific data. As our goal was to refine the recognition of children's facial expressions, using the Dartmouth dataset with its exclusive focus on children's images yielded better model performance than the broader age-ranged data in FER2013.

5.3 Implications and Future Directions

This research offers implications for the further development of tools to assist in the early diagnosis of psychological disorders among children. Future work will focus on creating datasets that offer a wider range of emotions and environmental conditions, such as lighting, angles, and blur. With this, more opportunities for deeper testing and fine-tuning can be done to understand the applications of real-world diagnostic tools.

References

Kalanadhabhatta, Manasa, Shaily Roy, Trevor Grant, Asif Salekin, Tauhidur Rahman, and Dessa Bergen-Cico. 2023. "Detecting PTSD Using Neural and Physiological Signals: Recommendations from a Pilot Study." In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE

Kalanadhabhatta, Manasa, Adrelys Mateo Santana, Deepak Ganesan, Tauhidur Rahman, and Adam Grabell. 2022. "Extracting Multimodal Embeddings via Supervised Contrastive Learning for Psychological Screening." In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE

Zhang, Saining, Yuhang Zhang, Ye Zhang, Yufei Wang, and Zhigang Song. 2023. "A Dual-Direction Attention Mixed Feature Network for Facial Expression Recognition." *Electronics* 12(17), p. 3595