



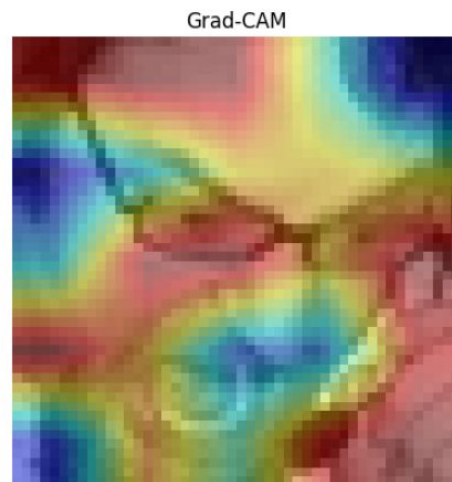
Interpretable Emotion Recognition: Visualizing Neural Attention in Facial Emotion Classification with Grad-CAM

Team M: Alex de la Haya, Gori Ribot and Marc Guiu

Introduction



FER 2013



State of Art

Human Accuracy (65%)

(2016) Pramerdorfer -> Deep CNNs (75%)

(2021) Chen -> VGG architecture (73%)

(2024) Lamichhane -> CNN-BiLSTM Hybrid model (79%)



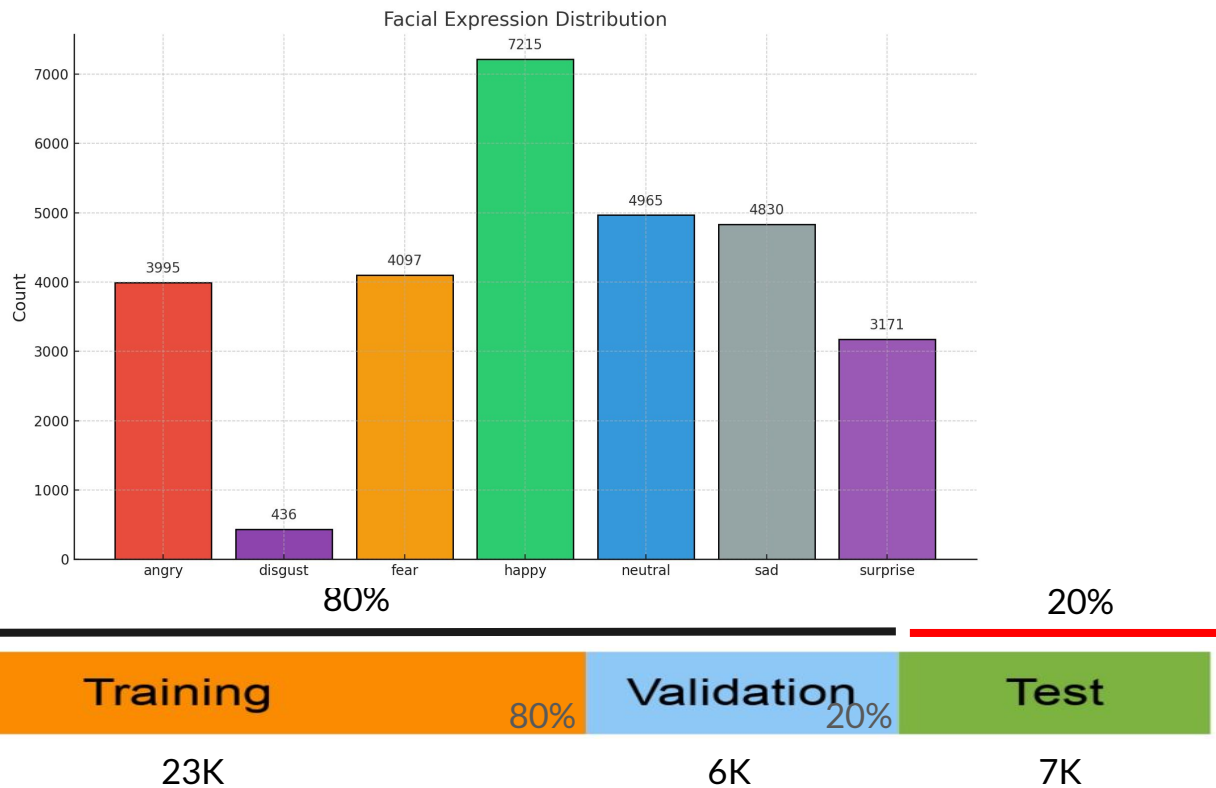
Methodology - Dataset

FER 2013 ->

36k images

CLASS

IMBALANCE



Methodology - Network

Preprocessing data:

Normalization

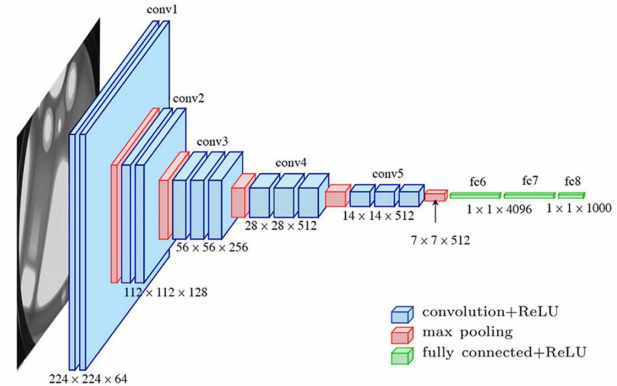
Transformation (rotate, flip, crop)

Convert to RGB

Adam optimizer, Cross Entropy Loss

VGG 19 - use pretrained weights (IMAGENET)

Transfer learning, Modify output



Configuration 1: Baseline Setup



Objective: Establish baseline performance

Model Setup:

- Pre-trained VGG19 as fixed feature extractor.
- All VGG19 convolutional layers frozen.
- Only custom classifier head trained (Linear, ReLU, Dropout)

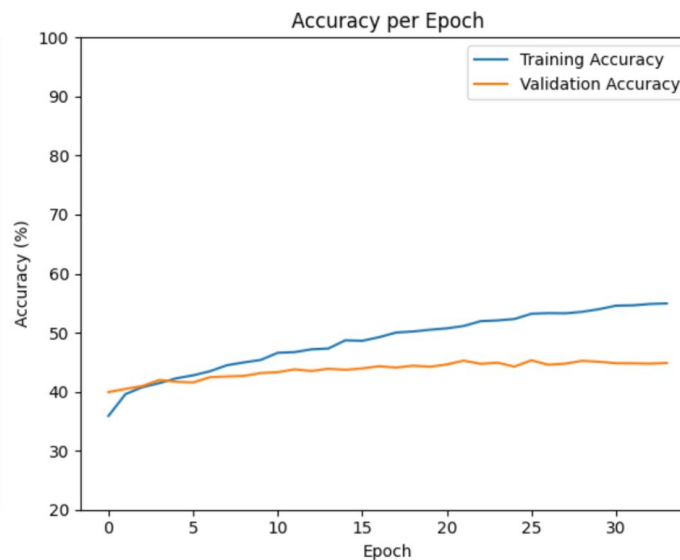
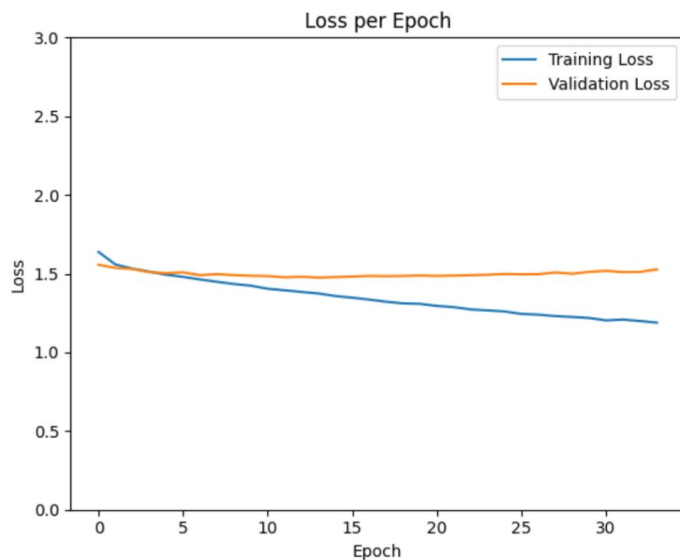
Data Prep & Augmentation:

- Images: 48x48, RGB,
- ImageNet normalization.

Optimization:

- Optimizer: Adam ($\text{lr}=1\text{e-}4$)
- Loss Function: CrossEntropyLoss.

Configuration 1: Results and Conclusion



Accuracy:

- Training: ~52%
- Validation: ~45%
- Testing: 44.22%

Loss:

- Training: 1.6 - 1.2
- Validation: 1.5 - 1.6

Conclusion: Model underfitting; limited capacity due to frozen backbone.

Configuration 2: Setup for Performance Boost



Objective: Improve performance by deeper fine-tuning and address class imbalance.

Model Setup:

- Unfroze last 9 VGG19 convolutional layers
- Custom classifier head still trained

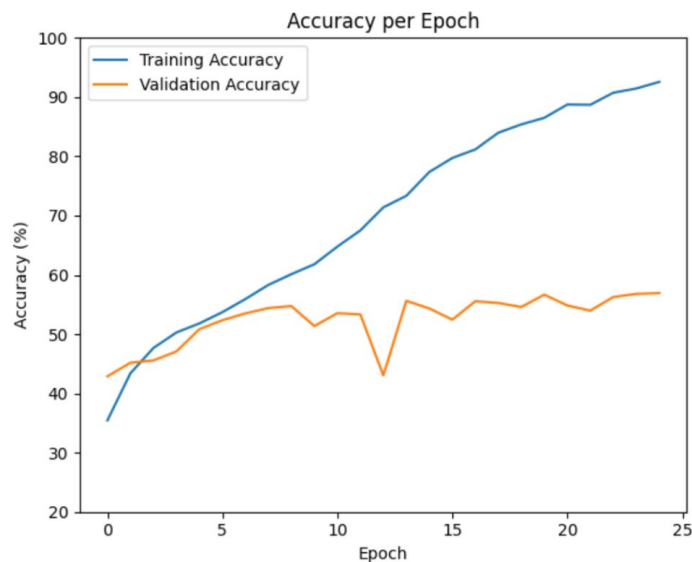
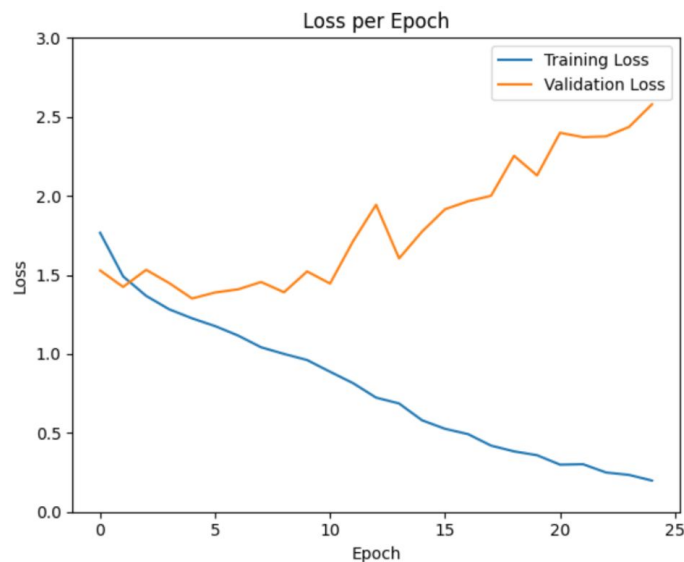
Data Prep & Augmentation:

- Same as configuration 1

Optimization:

- Optimizer: Adam ($\text{lr}=1\text{e-}4$) (same as configuration 1)
- Loss Function: CrossEntropyLoss with balanced class weights

Configuration 2: Results and Conclusion



Accuracy:


- Training: >90%
- Validation: ~57%
- Testing: 51.31%

Loss:

- Training: 1.7 - 0.2
- Validation: 1.5 - 2.5

Conclusion: Performance improved, but severe overfitting identified (large train-val gap).

Configuration 3: Reduced Fine-tuning, SGD and Learning Rate Scheduling



Objective: Drastically improve generalization and training stability.

Model Setup:

- Reduced fine-tuning to last 5 VGG19 feature modules

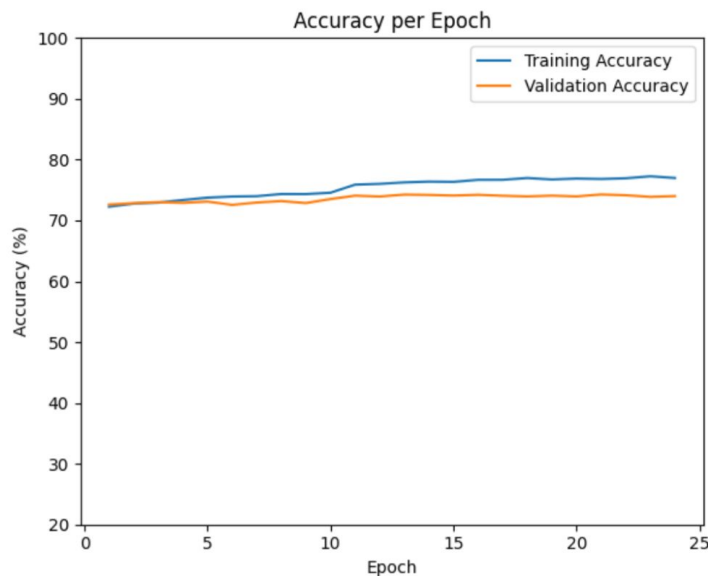
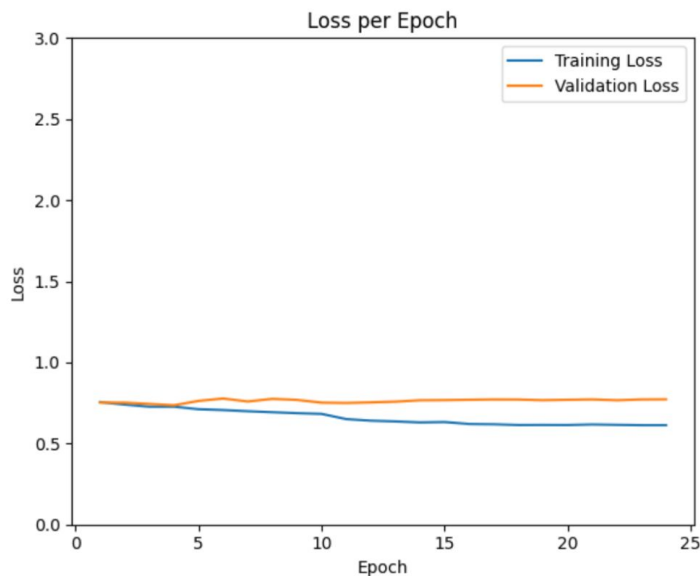
Data Prep & Augmentation:

- Same as configuration 1 and 2

Optimization:

- Changed optimizer to SGD (momentum=0.9, lr=0.01)
- Added L2 regularization (weight_decay=1e-4)
- Introduced Learning Rate Scheduler (ReduceLROnPlateau): Reduces the learning rate when validation loss stops improving.

Configuration 3: Results and Conclusion



Accuracy:

- Training: ~77%
- Validation: ~74%
- Testing: 57.91%

Loss:

- Training: 0.7 - 0.6
- Validation: 0.7 - 0.8

Conclusion: Transformative improvement in generalization (minimal train-val gap) but still room for absolute accuracy improvement

Configuration 4: Aggressive Augmentation, Label Smoothing, and Enhanced Regularization

Objective: Maximize robustness and absolute performance

Model Setup:

- Same as Configuration 3 (last 5 layers unfrozen)

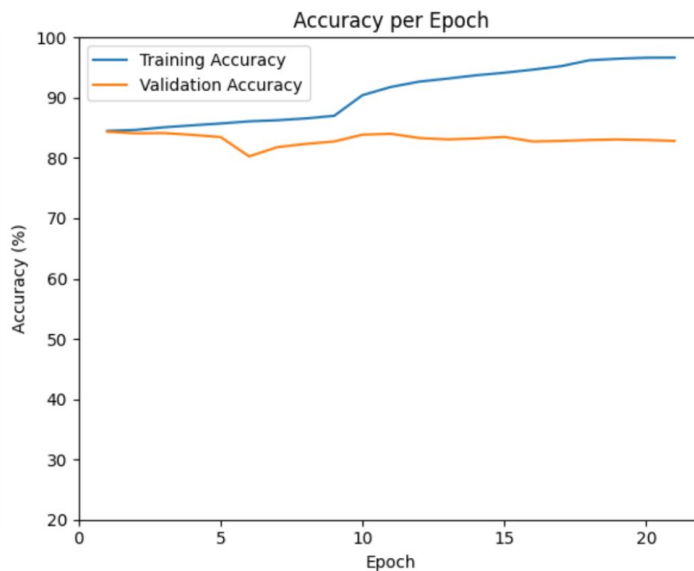
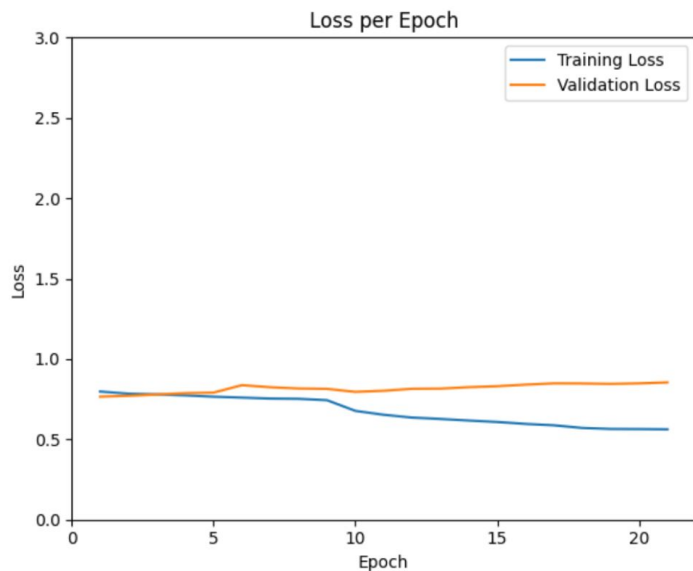
Data Prep & Augmentation:

- Expanded and aggressive pipeline: Stronger rotations (15°), broader crop scale (0.7-1.0)
- Added: ColorJitter, ElasticTransform, RandomPerspective, RandomAffine

Optimization:

- Increased L2 weight_decay to 1e-3
- Loss Function: LabelSmoothingCrossEntropy (replaces CrossEntropyLoss, reduces overconfidence)

Configuration 4: Results and Conclusion



Accuracy:

- Training: >90%
- Validation: ~83%
- Testing: 59.11%

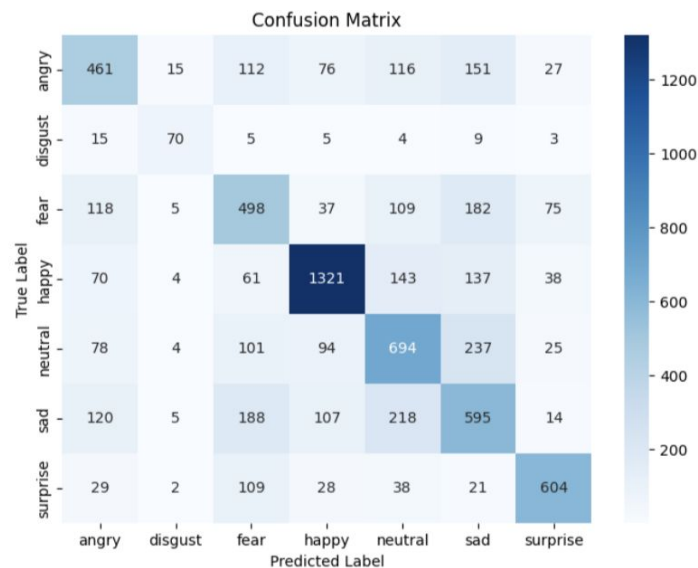
Loss:

- Training: 0.7 - 0.5
- Validation: 0.7 - 0.8

Conclusion: Significant boost in robustness and absolute performance. However, performance ceiling around 60% suggests dataset inherent challenges.

Configuration 4: Detailed Performance and Interpretability

Label	Precision	Recall	F1-score	Support
angry	0.52	0.48	0.50	958
disgust	0.67	0.63	0.65	111
fear	0.46	0.49	0.47	1024
happy	0.79	0.74	0.77	1774
neutral	0.52	0.56	0.54	1233
sad	0.45	0.48	0.46	1247
surprise	0.77	0.73	0.75	831
accuracy			0.59	7178
macro avg	0.60	0.59	0.59	7178
weighted avg	0.60	0.59	0.59	7178



Grad-CAM

Happy: faces often show attention around the mouth and cheeks and around the eyes, which is consistent with human perception of smiles.



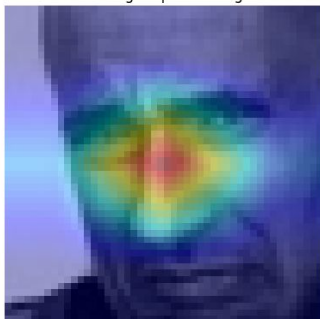
Fear: activations typically concentrate around the eyes and eyebrows, capturing widened eyes or raised brows, key features in fearful expressions.



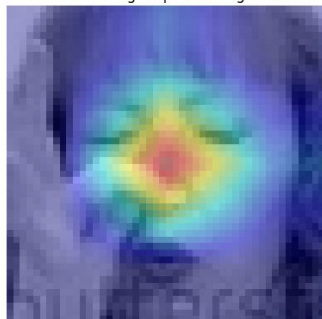
Grad-CAM

Disgust: the model frequently focuses on the nose and upper lip area, which aligns with crinkling and muscle movements common in disgusted faces.

True: Disgust | Pred: Disgust



True: Disgust | Pred: Disgust



True: Angry | Pred: Angry



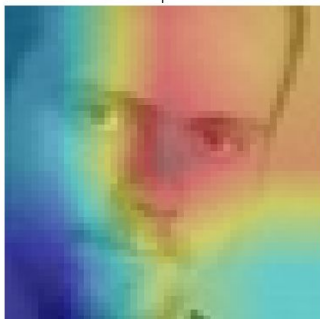
True: Angry | Pred: Angry



Grad-CAM

Sadness: the model focuses on the eyes and eyebrows. In some cases, it also highlights the mouth, especially when a frown is present.

True: Sad | Pred: Sad



True: Sad | Pred: Sad



True: Neutral | Pred: Neutral



True: Neutral | Pred: Neutral



We also found **labeling errors** between Surprise and Neutral, which introduced bias during training and made the model learn from inconsistent examples.

Conclusions



- CustomVGG19 with fine-tuning significantly outperforms the base CNN in accuracy and generalization.
- Data augmentation and class weighting are essential for handling the class imbalance in FER2013, leading to better model robustness.
- Grad-CAM visualizations effectively reveal which facial regions influence each emotion prediction, increasing model transparency and trust.
- The model still struggles to distinguish between visually or semantically similar emotions.
- Transfer learning, combined with domain-specific adaptation, proves to be a powerful approach when working with limited or noisy data.
- Results align with known limitations of FER2013: low resolution, imbalanced classes, and ambiguous labels.

References



- Pramerdorfer, C., & Kampel, M. (2016).** Facial Expression Recognition using Convolutional Neural Networks: State of the Art. *arXiv preprint arXiv:1612.02903*. <https://arxiv.org/abs/1612.02903>
- Khanzada, S. I., Mustafa, M. W., & Jamal, T. (2020).** Deep Learning-Based Facial Emotion Recognition Using FER Dataset. *arXiv preprint arXiv:2004.11823*. <https://arxiv.org/abs/2004.11823>
- Khairuddin, M. A., & Chen, J. (2021).** Performance Enhancement of Facial Expression Recognition System Using Deep Learning with Data Augmentation. *arXiv preprint arXiv:2105.03588*. <https://arxiv.org/abs/2105.03588>
- Oguine, G. C., Eya, I. N., & Akintola, A. G. (2022).** Real-time Facial Emotion Recognition Using Deep Learning Algorithms. *arXiv preprint arXiv:2206.09509*. <https://arxiv.org/abs/2206.09509>
- Lamichhane, S., & Karn, R. K. (2024).** Facial Emotion Recognition Using Hybrid CNN-BiLSTM Architecture. *International Journal of Engineering and Technology*. <https://www.nepjol.info/index.php/injet/article/view/72579>

Questions

