Universitat Pompeu Fabra Barcelona

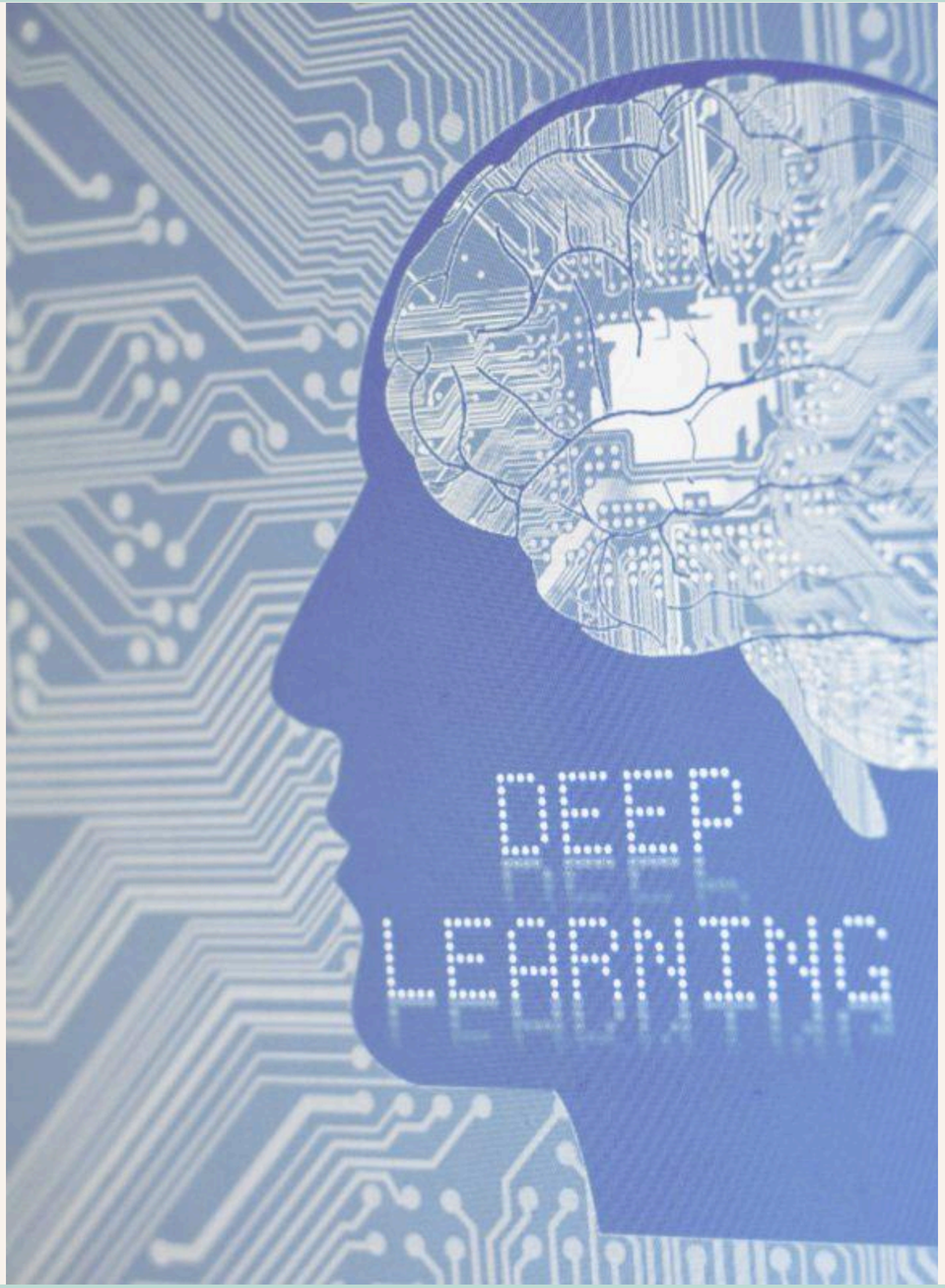
FACIAL EMOTION RECOGNITION

Tània Pazos & Yuyan Wang

Deep Learning

June 5, 2025

Index



1. Introduction

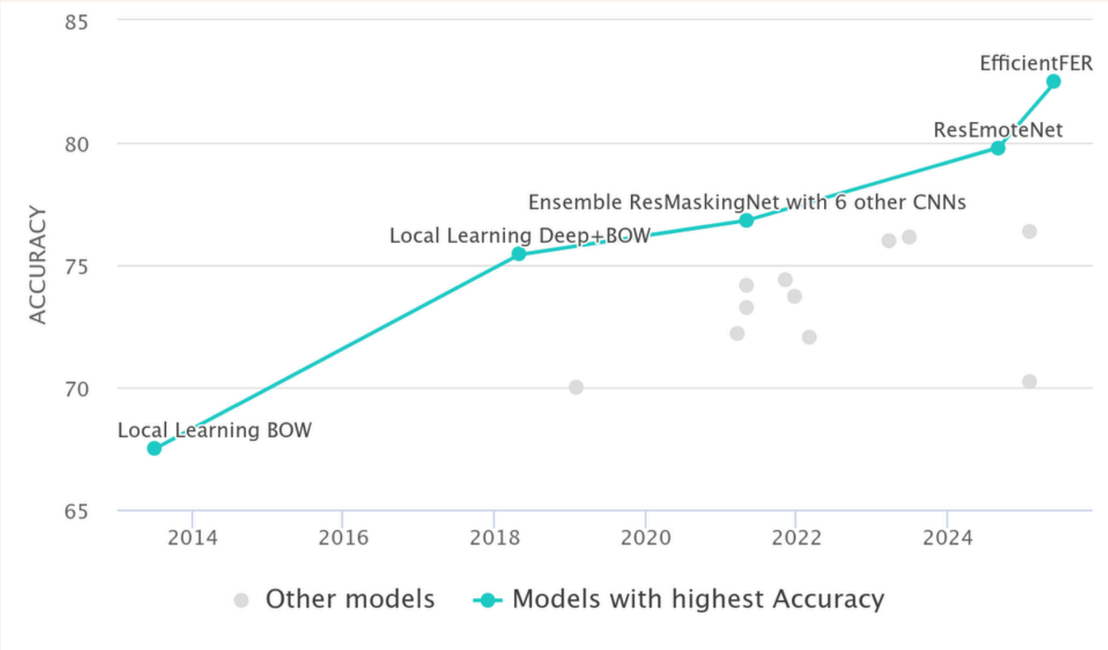
- “Facial Emotion Recognition (FER) is a technology used for analysing sentiments by different sources, such as pictures and videos.”
- Advances in biometric analysis, machine learning, and pattern recognition have contributed to the development of FER systems.
- Applications:
  - Healthcare (e.g., depression detection, patient monitoring during treatment)
  - Education (e.g., detect engagement in online learning)

Source: European Data Protection Supervisor TechDispatch #1/2021.

Overview of the Facial Emotion Recognition (FER) process.
Source: EDPS TechDispatch #1/2021.

2. State of the Art

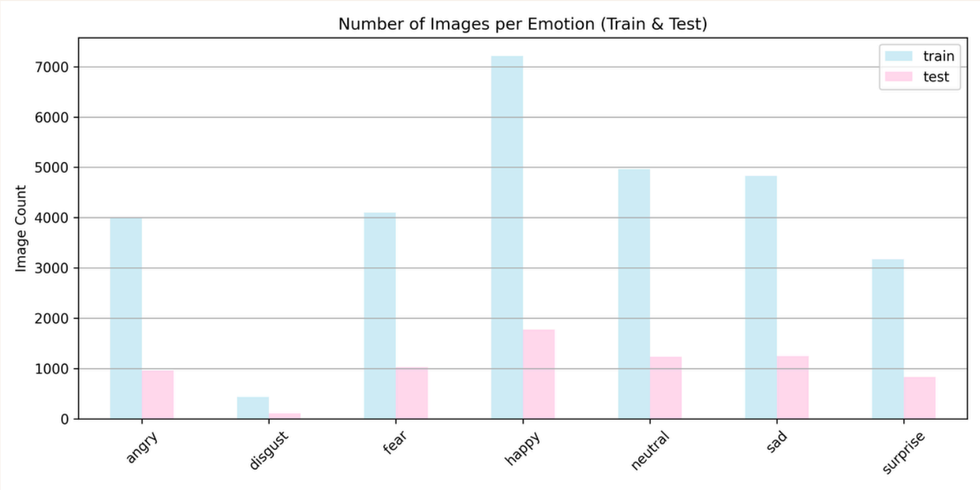
- Early models with >75% accuracy (2018): Fusion of handcrafted + deep features (BoVW + CNN)
- Recent top models:
  - ResMaskingNet (2021): ResNet + attention + segmentation
  - ResEmoteNet (2024): Residual + SE blocks
  - EfficientNetv2 (2025): Transfer + attention
- Limitation: 20M-120M+ parameters



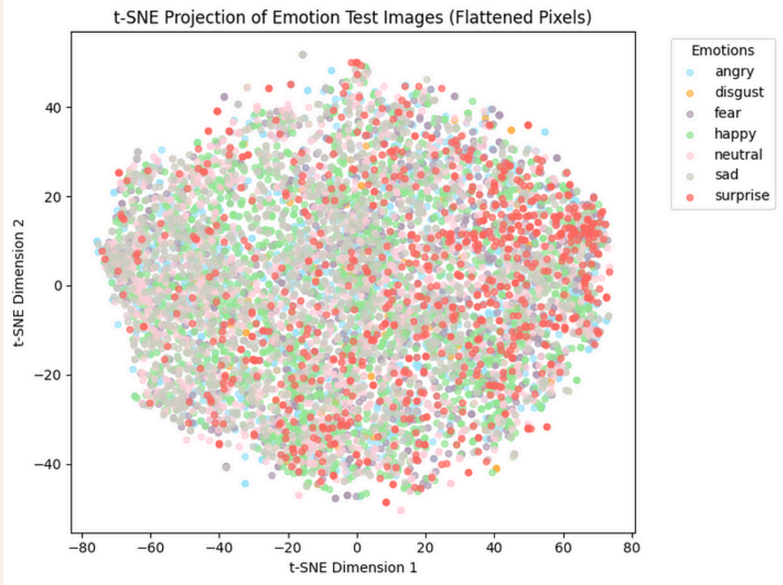
Progression of model accuracy on FER2013 over time.
Source: PapersWithCode FER2013.

3. Dataset & Exploratory Analysis

- Dataset: FER2013
  - 48x48 grayscale facial images
  - 7 emotion labels: 0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral
  - 28,709 training images, 3,589 test images
  - Class imbalance: Disgust underrepresented (436 samples), Happy overrepresented (7,215)



Number of images per emotion in the train and test sets of the original FER2013 dataset.



t-SNE projection of FER2013 test images based on raw pixel intensities (after PCA).

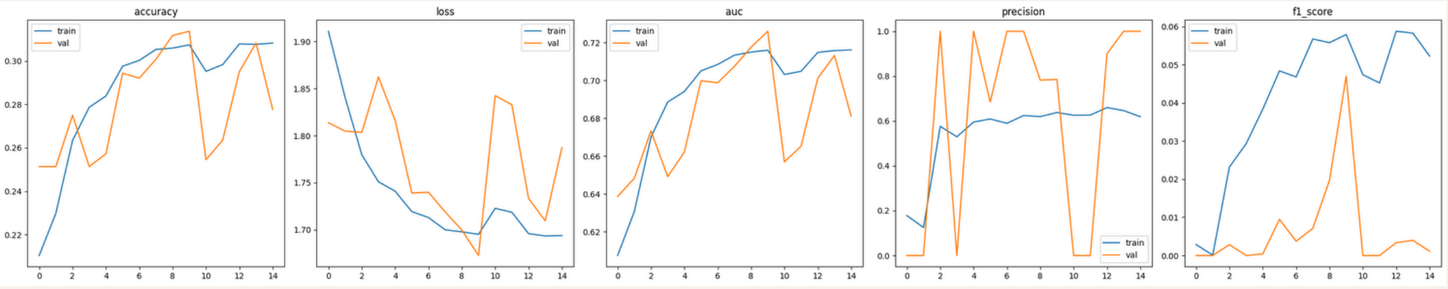
Sample images of the FER2013 dataset.

4. Baseline Model: VGG16

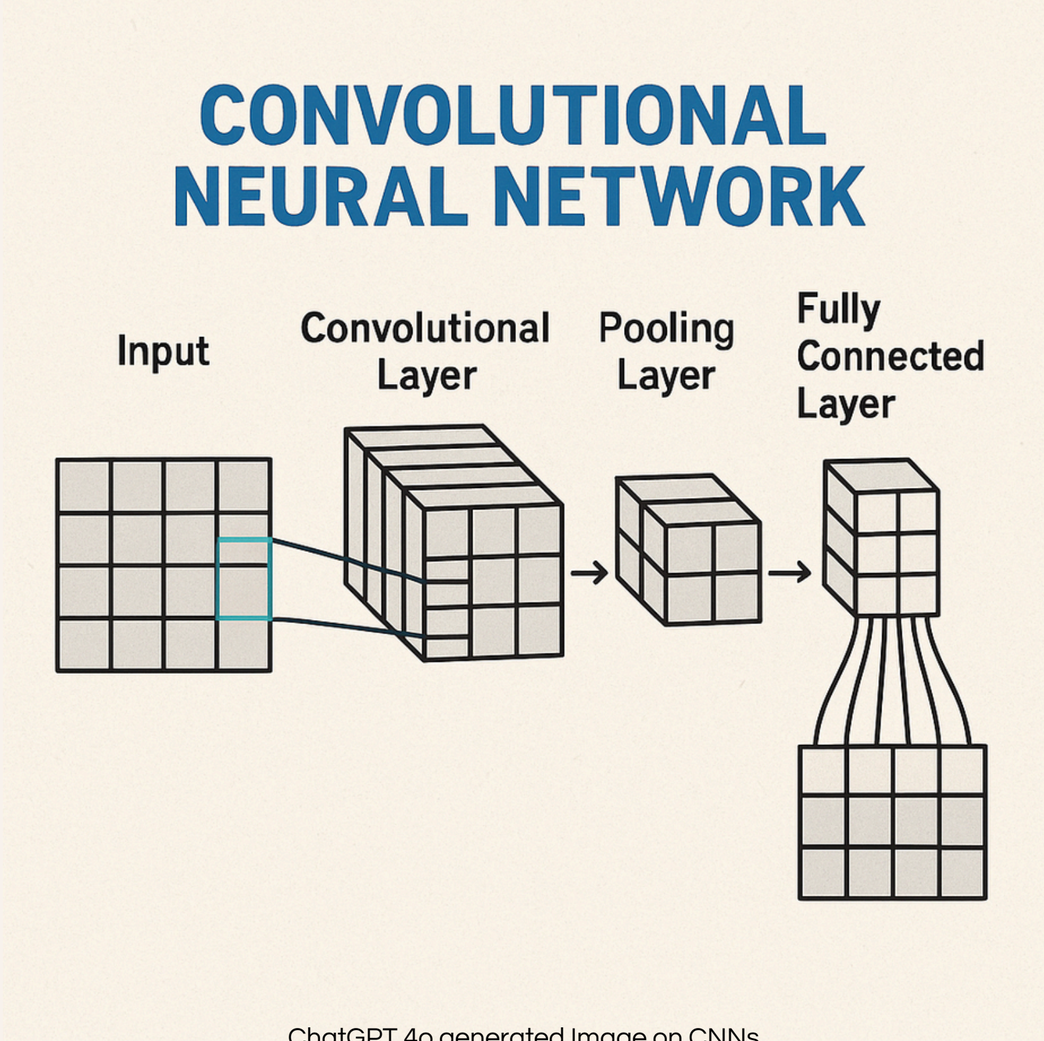
- Pretrained on ImageNet, fine-tuned on FER2013
- Architecture based on a public Kaggle implementation (Y. H. Shakir)
- Original metrics corrected (used binary accuracy instead of categorical)
- Kept training setup and architecture: 3 FC layers + dropout + batch norm
- 14,7 M total parameters
- After 15 epochs:

Metrics	Train	Validation
Accuracy	0,2994	0,2777
Loss	1,7023	1,7874
F1-score	0,0488	0,0011

Performance of the VGG16 baseline model after 15 training epochs.



Training and validation curves for the VGG16 baseline model over 15 epochs.



ChatGPT 4o generated Image on CNNs

5. Project Goals

01 Develop lightweight CNNs for classifying facial expression images into 7 basic emotions. No more than 500k parameters.

02 Compare performance across different training configurations, including optimizers (Adam, SGD), learning rates, dropout rates...
Benchmark against the predefined baseline.

03 Use macro F1-score and class-level precision/recall to fairly assess performance across potentially imbalanced emotion classes.

6. Methodology

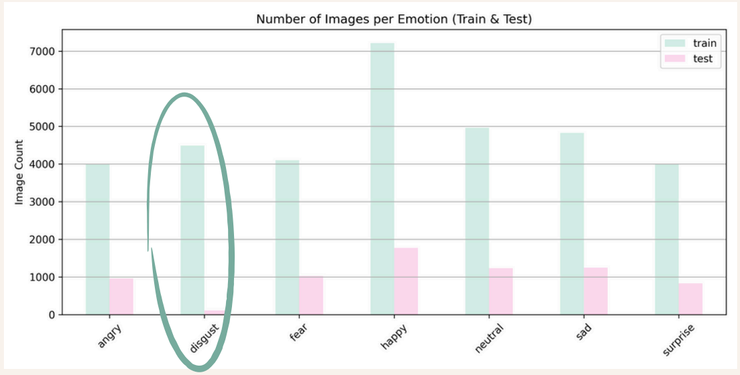
01

Dataset & Preprocessing

- Collected images from the AffectNet dataset with the goal of balancing class distribution to at least 4k images per emotion
- Converted all images to grayscale; used 3-channel inputs; Normalized pixel values to the range [-1, 1]
- Underrepresented classes like disgust, oversampling was performed to achieve comparable sample sizes

02

Data Augmentation techniques



Number of images per emotion in the train and test sets of the dataset after offline data augmentation.

03

2 Model Architectures

- ResNet (Custom Lightweight Variant)
  - residual connections, which help mitigate the vanishing gradient problem
- MobileNetV2 (Custom Lightweight Variant)
  - computationally efficient and suitable for low-resource environments

04

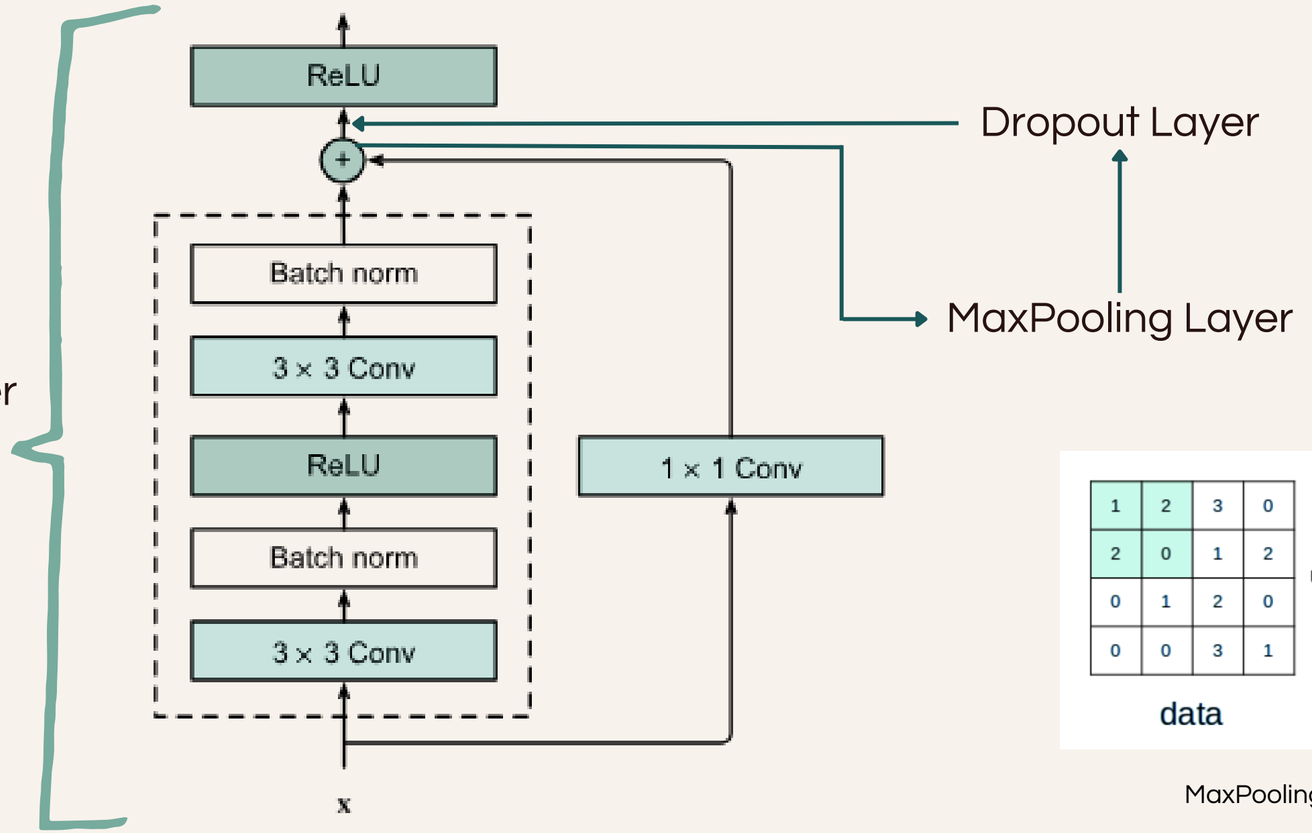
Evaluation & Metrics

- Macro F1-score (main metric), Precision, Recall per emotion
- Accuracy, training/validation loss
- Visual embedding analysis using t-SNE to project CNN features

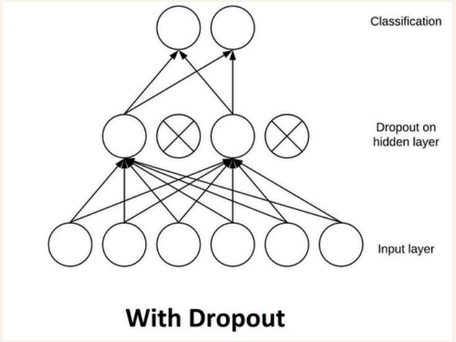
7. Residual Neural Network

2 blocks + FC Layer

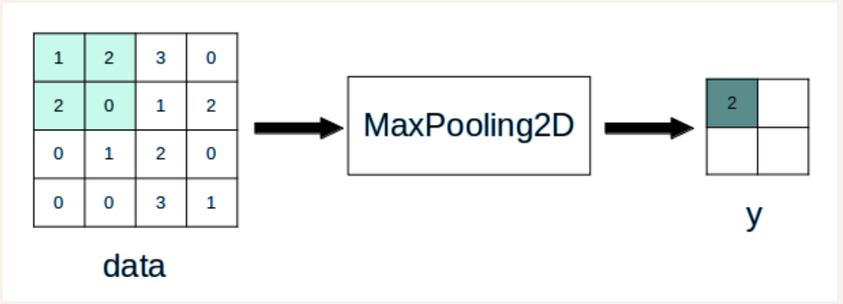
418k parameters



Architecture overview ResNet. Source: Dive Into Deep Learning



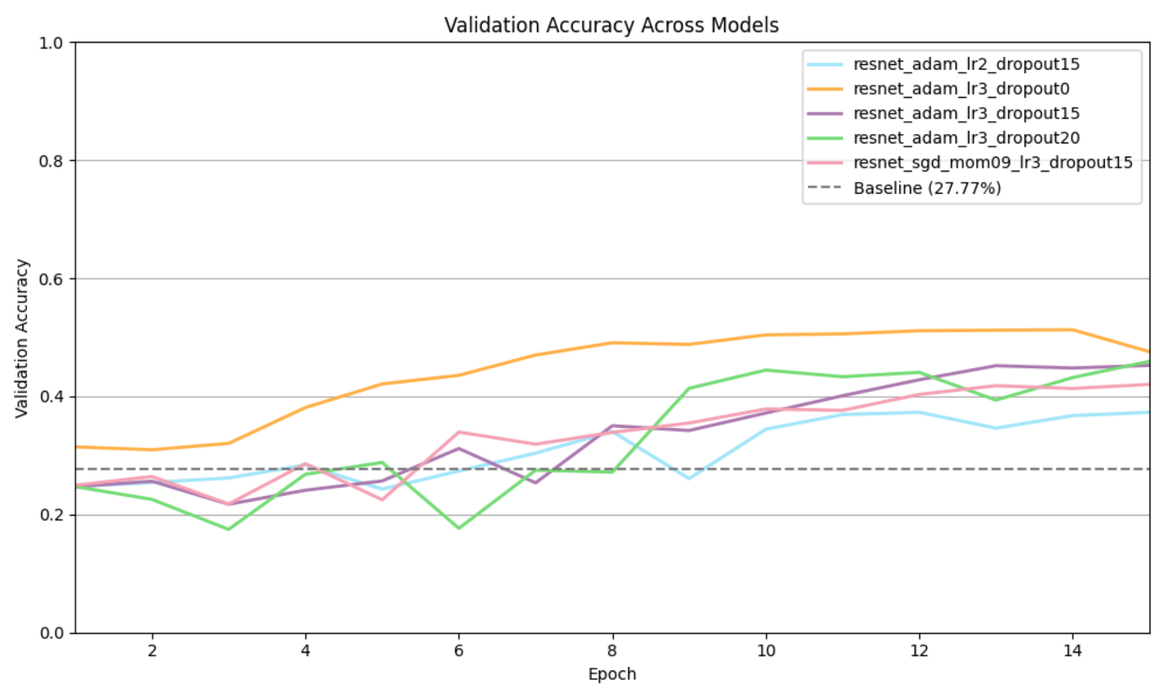
Dropout Illustration. Source: Medium



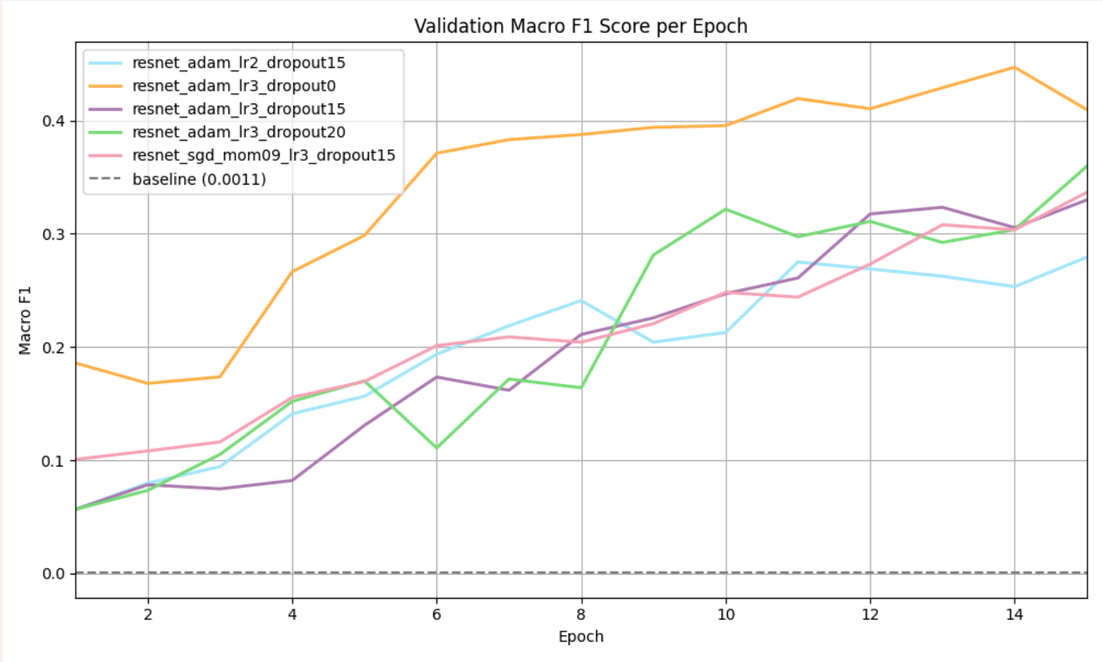
MaxPooling Illustration. Source: tmytokai.github.io

7. ResNet

Validation Accuracy Performance



Validation F1-Score Performance

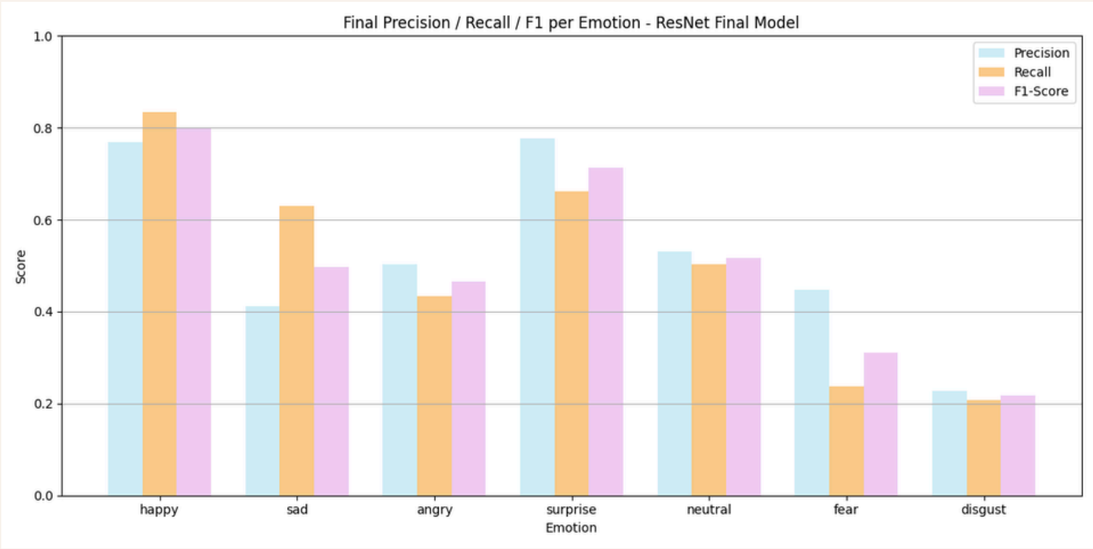


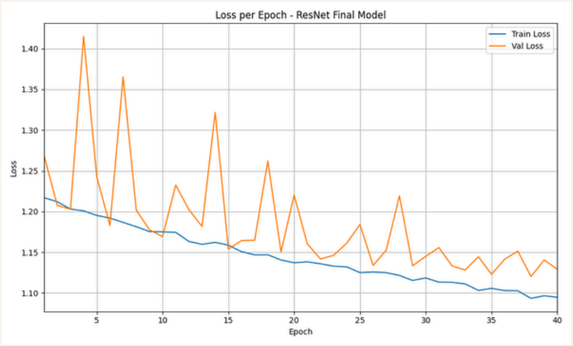
Model	Val Accuracy	Val F1	Happy Val F1	Disgust Val F1
Baseline (VGG16)	27,77%	0,11%	42,00%	0,00%
adam_lr3_dropout0_bn	47,60%	40,95%	76,30%	14,77%
adam_lr3_dropout15_bn	45,20%	32,99%	72,10%	17,82%
adam_lr2_dropout15_bn	37,30%	27,93%	61,60%	4,11%
adam_lr3_dropout20_bn	45,90%	35,96%	74,40%	10,80%
sgd_mom09_lr3_dropout15_bn	42,00%	33,65%	63,40%	3,54%

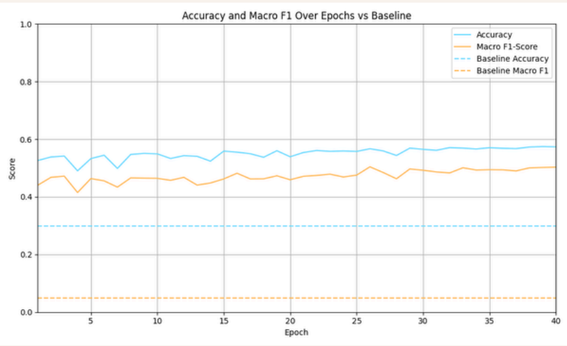
Table summary of performance of the baseline model and ResNet models after 15 training epochs.

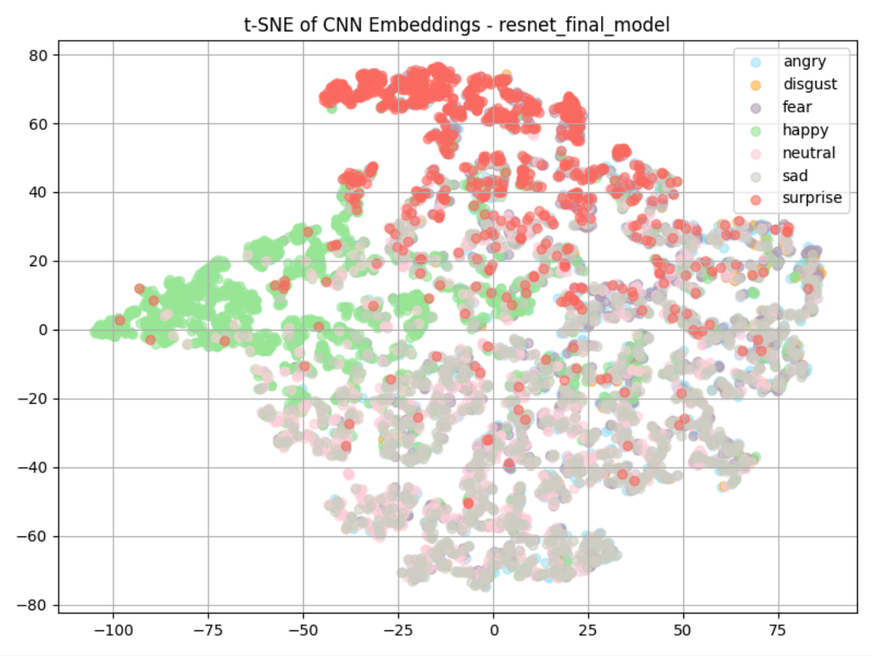
7. ResNet – Final Model

Performance of the ResNet adam_lr3_dropout0_bn after 40 epochs

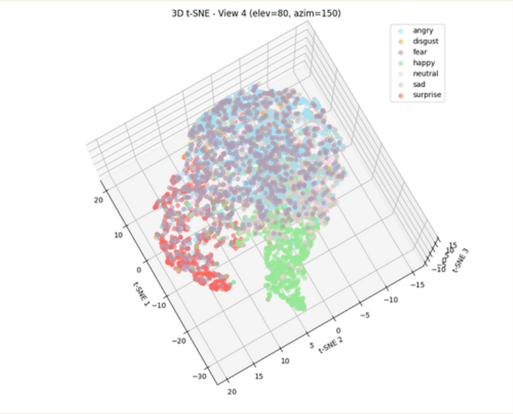




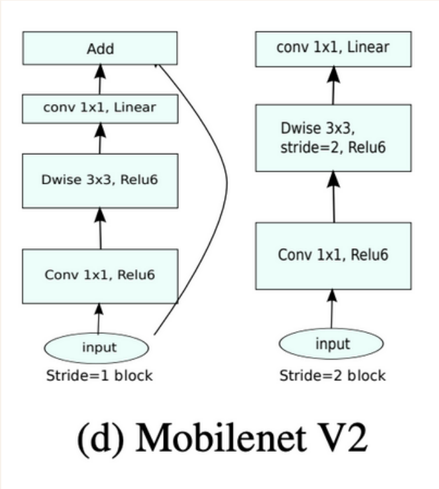




t-SNE projection of FER2013 test images after training.



8. MobileNetV2



Overview of the original MobileNetV2 architecture.

Input	Operator	t	c	n	s
$224^2 \times 3$	conv2d	-	32	1	2
$112^2 \times 32$	bottleneck	1	16	1	1
$112^2 \times 16$	bottleneck	6	24	2	2
$56^2 \times 24$	bottleneck	6	32	3	2
$28^2 \times 32$	bottleneck	6	64	4	2
$14^2 \times 64$	bottleneck	6	96	3	1
$14^2 \times 96$	bottleneck	6	160	3	2
$7^2 \times 160$	bottleneck	6	320	1	1
$7^2 \times 320$	conv2d 1x1	-	1280	1	1
$7^2 \times 1280$	avgpool 7x7	-	-	1	-
$1 \times 1 \times 1280$	conv2d 1x1	-	k	-	

Layer configuration of MobileNetV2.

Original MobileNetV2 (2018)

- Designed for 224x224 RGB images
- Multiple stride=2 layers
- 3.4M parameters

Our adaptation (48x48, ~130k parameters)

- Input resolution is low, so we avoid early stride=2 to preserve spatial features
- Fewer bottleneck repeats and removed heavy blocks (160, 320)
- Final projection layer reduced from 1280 to 128 channels
- Dropout applied after final conv block, following Kim et al. (2023), to improve regularization before classification.

Source: Sandler et al., 2019, MobileNetV2: Inverted Residuals and Linear Bottlenecks.

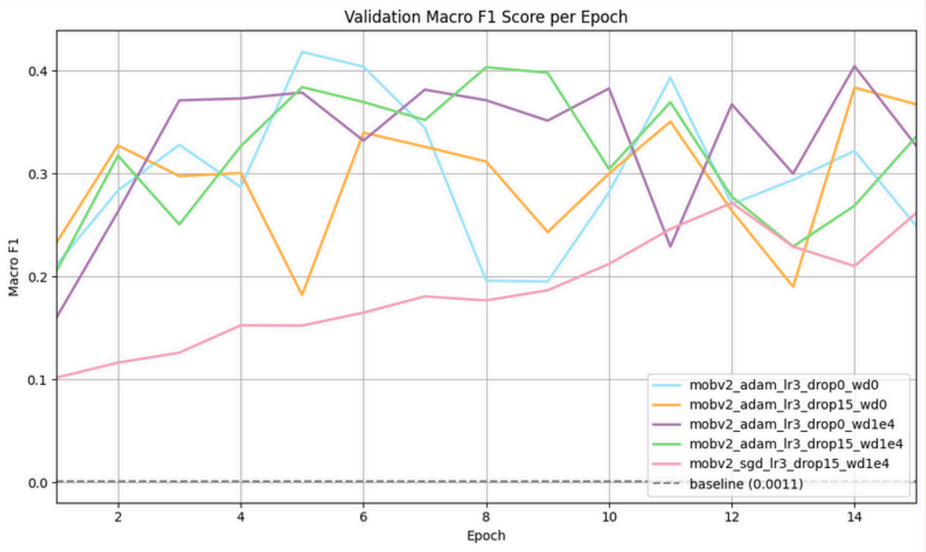
8. MobileNetV2

- Our MobileNetV2 showed signs of overfitting, so we systematically tested dropout, weight decay, and their combination
- Compared Adam vs. SGD with momentum to evaluate optimizer impact under the same regularization settings
- Goal: Identify the best trade-off between generalization and training stability within our <500k param budget

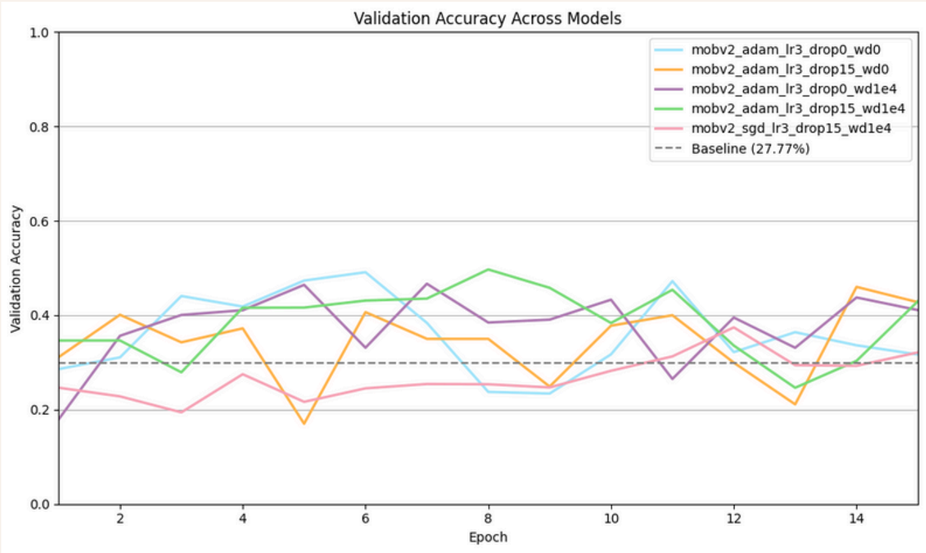
Model	Val Accuracy	Val F1	Happy Val F1	Disgust Val F1
Baseline (VGG16)	27,77%	0,11%	42,00%	0,00%
adam_lr3_drop0_wd0_bn	31,70%	24,92%	66,35%	7,22%
adam_lr3_drop0_wd1e4_bn	41,08%	32,71%	17,34%	11,61%
adam_lr3_drop15_wd0_bn	42,76%	36,73%	75,81%	9,07%
adam_lr3_drop15_wd1e4_bn	42,96%	33,56%	74,32%	9,44%
sgd_lr3_drop15_wd1e4_bn	32,13%	26,13%	56,73%	2,66%

Baseline (VGG16) vs. MobileNetV2 validation performance across regularization and optimizer settings after 15 epochs.

Validation F1-Score Performance



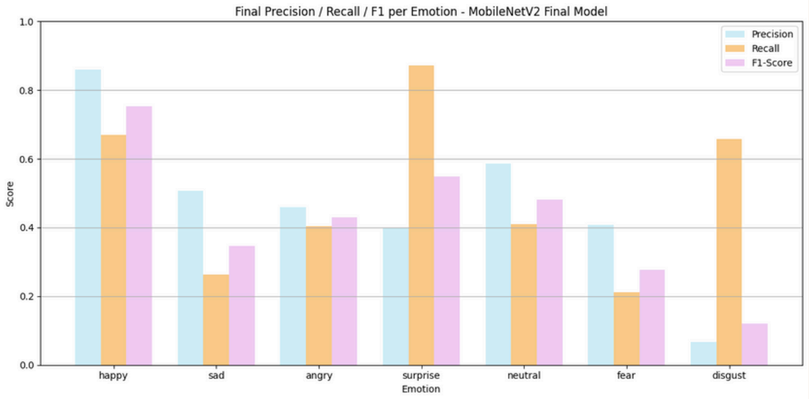
Validation Accuracy Performance

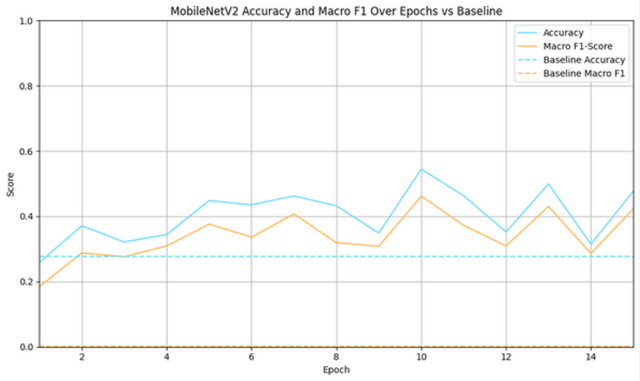


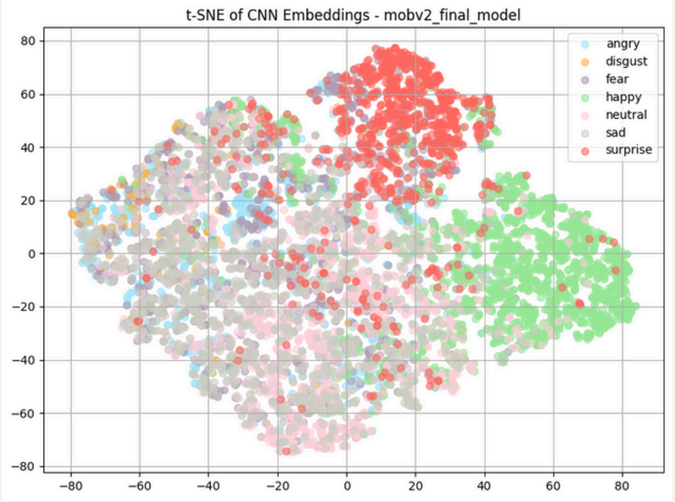
8. MobileNetV2 – Final Model

- Final training (40 epochs): Used early stopping, LR scheduler, and checkpointing to stabilize training and prevent overfitting observed in short runs

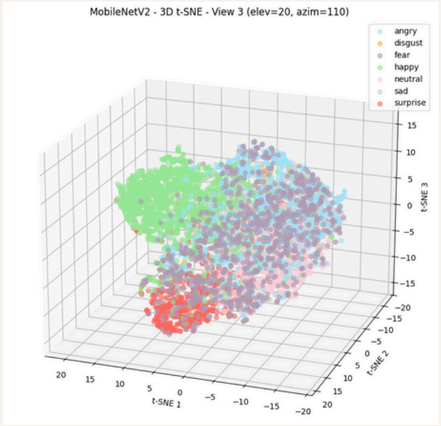
Performance of the MobileNetV2 adam_lr3_drop15_wd1e4_bn







t-SNE projection of FER2013 test images after training



9. Limitations & Future Work

Limitations

- Limited GPU and time budget restricted the number of training runs and hyperparameter tuning
- Low input resolution (48×48) may limit the ability to extract subtle facial features (especially important in classes like Disgust)
- Even with data augmentation, Disgust remains the most underperforming class in terms of F1-score

Future work

- Train on color images to add texture information and improve emotion discrimination
- Further fine-tune MobileNetV2, since validation accuracy and F1-score still show noticeable fluctuation across epochs

10. Conclusions

01 We developed and compared lightweight deep learning models to classify 7 human emotions using facial images. Both achieved strong results with fewer parameters.

02 Data augmentation and oversampling, especially for imbalanced classes like disgust, helped models perform better.

03 ResNet performed well with more training epochs, MobileNetv2 benefited from early stopping.

04 Happy and Surprise emotions are easier to differentiate from the other 5 emotions.

11. References

European Data Protection Supervisor. (2021, May). Facial emotion recognition (TechDispatch #1/2021). https://www.edps.europa.eu/system/files/2021-05/21-05-26_techdispatch-facial-emotion-recognition_ref_en.pdf

Kim, B. J., Choi, H., Jang, H., Lee, D., & Kim, S. W. (2023, July). How to use dropout correctly on residual networks with batch normalization. In Uncertainty in Artificial Intelligence (pp. 1058–1067). PMLR.

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. IEEE Transactions on Affective Computing, 10(1), 18–31. https://doi.org/10.1109/TAFFC.2017.2740923

Pham, L., Vu, T. H., & Tran, T. A. (2021, January). Facial expression recognition using residual masking network. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 4513–4519). IEEE. https://doi.org/10.1109/ICPR48806.2021.9412403

Sambare, M. (n.d.). FER2013 facial emotion recognition dataset. Kaggle. Retrieved from https://www.kaggle.com/datasets/msambare/fer2013

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4510–4520). https://doi.org/10.1109/CVPR.2018.00474

Shakir, Y. H. (2021). Emotion recognition with VGG16. Kaggle. https://www.kaggle.com/code/yasserhessein/emotion-recognition-with-vgg16

Papers with Code. (2025). Facial expression recognition on FER2013. Retrieved from https://paperswithcode.com/sota/facial-expression-recognition-on-fer2013

Universitat Pompeu Fabra Barcelona

Thank you for your attention

Tània Pazos & Yuyan Wang

Deep Learning

June 5, 2025