Script

Introduction

* Basic Introduction

Motivation (1)

* Facial Emotion Recognition analyses facial expressions to detect emotions.
* FER is widely used in psychology, human-computer interaction, and various applications that enhance user experiences. (Mention some seen on screen)

Motivation (2)

* Despite its various use cases there are some glaring issues such as
  + Privacy Risks (Tailored Adverts)
  + Accuracy and Bias (Low -> For underrepresented ethnicities)
  + Misinterpretation (Public Safety -> Misidentified as a threat)

Research Topic (1)

* As outlined prior FER, is the process of using machine learning models to detect human emotions from facial images.
* The majority of FER models are typically trained on two types of datasets:
  + **In-lab FER**, where datasets like CK+ and JAFFE are collected under controlled conditions, ensuring clear and structured facial expressions.
    - Smaller in size
    - Uniform
  + **In-the-wild FER**, which uses real-world, unstructured datasets like AffectNet, FER2013, and RAF-DB, making models more adaptable to natural variations in facial expressions.
    - Larger
    - Less control over what images is present
  + Capture the following set of emotions
    - Anger, Disgust, Fear, Happy, Sad, Surprise,
      * Neutral, Contempt – Present based on the dataset
  + Some Datasets also include compound emotions (RAF-DB)

Research Topic (2)

To better understand how FER works, let’s break it down into five key steps:

1. **Preprocessing** – Images are converted to grayscale, normalized, resized, and aligned to ensure consistency before being analysed.
2. **Face Detection** – The system first identifies faces in an image using traditional methods like Haar Cascades or deep-learning techniques such as MTCNN.
3. **Feature Extraction** – Instead of manually defining features, modern FER models use deep learning to automatically learn spatial and temporal patterns in facial expressions.
4. **Classification** – The extracted features are then passed through deep learning architectures like CNNs, Transformers, RNNs, or LSTMs, which classify the expression into categories such as happiness, sadness, anger, or surprise.
5. **Emotion Prediction** – Finally, the model outputs an emotional label based on its trained dataset, completing the recognition process.

Research Topic (3)

One of the biggest research challenges in FER is **generalisability** – ensuring that models perform well across different datasets, environments, and demographics.

**Why is this an issue?**

* Many FER models achieve high accuracy on their **training datasets** but fail when tested on **new, unseen data**.
* The replicability of **state-of-the-art (SOTA)** models is often questionable, with results that are difficult to reproduce.
* Factors like **lighting conditions, facial occlusions, and diverse ethnic representations** can significantly affect model performance.

**Why does this matter?**

By studying generalisability, researchers can:  
✔ Compare different models to assess their real-world effectiveness.  
✔ Identify **biases** in emotion recognition, ensuring fairness across different demographics.  
✔ Improve **replicability** in FER research, challenging overhyped claims of SOTA performance.  
✔ Direct future research toward **more robust and reliable** emotion recognition models.

Addressing these issues is crucial for making FER applicable in **critical fields like healthcare, security, and education**.

Aims And Objectives

* As outlined prior Facial Expression Recognition has potential in fields such as healthcare, education, and public safety. But for these applications to work in real-world scenarios, we need models that generalise well beyond their training datasets.
* Through my research, I want to explore two key areas:
  + **Model Performance Replicability** – Can existing models consistently achieve their reported results?
  + **Cross-Dataset Performance** – If a model is trained on one dataset, how well does it perform on completely different datasets?
* By the end of this research, I aim to have a deeper understanding of FER models, identifying the best techniques in terms of accuracy and practical usability. Ultimately, I hope to **promote transparency and replicability** in FER research—ensuring that advancements in this field are built on reliable and reproducible results.

Existing Research (P1-1)

* Provides a deep dive into the state of FER research by comparing various approaches.
* Doesn’t implement the models but compares research only.
* Covers three types of FER models
  + **Traditional Machine Learning Models** such as SVM, KNN, Random Forest, CART, and Logistic Regression,
  + **Deep Learning Models** like MobileNet, CNN, DCNN, VGG16, and ResNet-50, and
  + **Hybrid Models** that combine approaches (for example, CNN + SVM or DBN + SVM), integrating feature extraction with classification.

Existing Research (P1-2)

* Common datasets used in the research covered include:
  + FER2013, FERPlus, JAFFE, CK, CK+, RAF-DB, and KDEF.
* Evaluation metrics:
  + Primary: Accuracy
  + Secondary: F1-score, precision, recall,

Existing Research (P1-3)

* Deep learning models generally outperform traditional machine learning techniques in terms of accuracy. However, this comes with some significant drawbacks:
  + Deep learning approaches demand extensive datasets,
  + They require substantial memory and processing power, and
  + They have longer training and testing durations.
* Data augmentation emerged as an overall beneficial strategy, as it enhances model flexibility, helps prevent overfitting, and boosts model accuracy.

The limitations of this research paper are as follows:

* Models were not compared under uniform conditions, meaning different datasets or hyperparameters were used,
* The review focused only on widely recognised models instead of incorporating the very latest state-of-the-art approaches, and
* The comparisons were primarily based on accuracy, which might not fully capture model performance.

Existing Research (P2-1)

* Compared 3 SOTA models (Present results of ResNet50 only, other are in appendix)
  + ResNet50 (PreTrained on VGGFace2)
  + Inception-ResNet (PreTrained on CASIA-WebFace)
  + ResNet50 + Entropy Regularisation (PreTrained on VGGFace2)
* Also Compared 3 API models:
  + Face++
  + Amazon Rekognition
  + Microsoft Azure
* Tested on 12 datasets
  + 6 In-Lab: JAFFE, CK+, Oulu-CASIA, KDEF, IASLab, GEMEP
  + 6 In-The-Wild: EmotioNet, SFEW, RAF-DB, Aff-Wild2, FER2013, AffectNet

Existing Research (P2-2)

* Preprocessed Images
  + MTCNN (Facial Detection) + Facial Alignment (Eye Location)
  + Grayscale
  + Resized (ResNet50 – 2242, Inception-ResNet – 1602)
  + 50% Horizontal Flip (Reduce Overfitting)
* Same Hyperparameters used across all models & experiments:
  + **Learning Rate**: **0.001**
  + **Learning Rate Scheduler**: Step size of **10** with a decay factor (γ) of **0.5**.
  + **Optimizer**: Stochastic Gradient Descent (SGD) with a momentum value of **0.9**.
  + **Epochs**: **30 epochs**.
  + **Batch Size**:
    - **64** for single-source experiments.
    - **128** for multi-source experiments.
* 3 experiments

Experiment 1: Single-Source Generalisation

Assess how well models trained on a single dataset can generalise to other datasets.

**Method:**

* Models trained on 1 source dataset from the 12.
  + Tested on remaining 11 other datasets (cross-corpus testing).
* This setup allowed for **132** permutations (1 dataset as a source and others as target datasets).

**Results:**

* Evaluation Metric: Accuracy
* The best generalising models **(BOLD RESULTS)** were trained mostly on the **AffectNet** (one of the largest) datasets
* Diagonal Entries: Within-Corpus -> Avg Accuracy 76.4%
* Horizontal Entries: Cross-Corpus -> Avg Accuracy 42.0%

Experiment 2: Multi-Source Generalisation

Assess whether using multiple source datasets improves model generalisation.

**Method**:

* The experiment had three conditions for training on multiple source datasets:
  + **Within-setting**: Trained on five in-lab datasets and tested on the remaining in-lab dataset.
  + **Cross-setting**: Trained on five in-lab datasets and evaluated on all the in-the-wild datasets and vice-versa.
  + **Leave-one-out**: Trained on eleven source datasets and tested on the remaining dataset.

**Results:**

* **Best** **Models** irrelevant of evaluation type were derived from **In-the-wild datasets** most likely due to these datasets generally **being of a larger size**.
* **Within-Setting** - Average Accuracy (61.7%)
* **Cross-Setting** - Average Accuracy (42.76%)
* **Leave-one-out** - Average Accuracy (65.6%)
  + Improved when compared to Within & Cross-Setting but still worse than the single-source results from experiment 1
  + Further lends itself to the idea that larger datasets improve performance overall

Experiment 3: Commercial APIs

Assess the performance of custom-built models against commercial facial expression recognition APIs.

**Method**:

* Same method as used prior altered for testing on these models
  + Test sets only were used since no training was required

**Results:**

* API average performance underperforms by 25% across all datasets compared to within-corpus results
* Microsoft Azure (Discontinued -> Missing Results) (not reliable as other models)
* Constantly Changing hence these models’ performance easily fluctuates

My Plan (1)

* Datasets + Models from papers with code
  + Unofficial datasets as some **RAF-DB**, **AffectNet** only accessible to lecturers or researchers
* Transform datasets to the correct format required by the models
  + Augment -> Face Detection, Alignment, Resizing, Grayscale
  + In-built Model specific transformations
  + Implement Models via GitHub code
  + Train each model 3 times on 3 In-the-wild datasets (AffectNet, Fer2013, RAF-DB)
    - Research in paper 2 outlined that In-the-wild performed better overall
    - Same Hyperparameters were possible
  + Evaluate Models:
    - Their Within-Corpus Testing Set
    - Derive Results for the rest of the datasets
      * (Within & Cross Setting (In-Lab / In-The-Wild))
  + Conclude on best model, taking into consideration:
    - Training Time
    - Difficulty of Setup
    - Performance Replicability
      * Not sure on this due to not having exact datasets
    - General Performance
    - Generalisability

My Plan (2)

* Models Used:
  + ResEmoteNet, ResNet50, EmoNeXt, DDAMFN++, PAtt-Lite
* Datasets Used:
  + AffectNet, FER2013, RAF-DB, CK+, JAFFE (Last 2 only for evaluation)
* Performance Metric:
  + Accuracy as mostly used in research
  + Precision, Recall, F1-score - if they produce interesting insights