**EmoNeXt Paper Summary**

1. **EmoNeXt** - FER framework based on an adapted **ConvNeXt** architecture network.
   1. **Spatial Transformer Network** (STN) to focus on feature-rich regions of the face
   2. **Squeeze-and-Excitation** **block** to capture channel-wise dependencies.
   3. **Self-attention regularisation term** & **Cross-Entropy Loss,** encouraging the model to generate compact feature vectors
2. Mentions two approaches for FER:
   1. **Geometric** methods - capture the shape, location and interconnections of facial components during expressions.
   2. **Appearance-based** methods - focus on the variations in facial appearance, such as wrinkles and furrows, and can be extracted from the whole face or specific regions.
3. **FER** approaches follow a two-step process: 1) analyse and define facial features, 2) utilise these features for inference. However, as these **two steps are performed separately**, sub-optimal performance is obtained, particularly when dealing with complex datasets containing numerous sources of variability. Consequently, **it is more advantageous to perform these two steps together for better recognition performance**.
4. **Spatial Transformer Network** (STN):
   1. Spatial Transformer Networks (STN) are a special type of neural network component that helps a model adjust images before analysing them. They allow a network to automatically correct distortions like changes in size, rotation, or position, which is especially useful in Facial Expression Recognition (FER). STNs work in three main steps:
      1. **Localisation Network** – This part of the network examines the input image and predicts how it should be transformed.
      2. **Grid Generator** – Based on the predictions, this step creates a grid that determines how to shift or warp the image.
      3. **Sampler** – The final step applies the transformation, adjusting the image before passing it to the next layers of the neural network.
   2. A key benefit of STNs is that they learn these transformations automatically during training, making models more adaptable and accurate.

A diagram of a grid generator

AI-generated content may be incorrect.

1. **ConvNeXt**
   1. ConvNeXt is a pure convolutional model inspired by Vision Transformers but maintains the efficiency of CNNs.
   2. It improves upon the standard **ResNet** architecture by using larger kernels, depth wise convolutions, and an inverted bottleneck to boost performance while reducing computational cost. Key architectural changes:
      1. Activation Function: Replaces **ReLU** with **GELU**
      2. Normalization: Uses **Layer Normalization** (LN) instead of **BatchNorm** (BN) for better stability.
      3. Multiple versions exist, varying in the number of channels and blocks, with configurations from **Tiny** to **XLarge**.

A diagram of a algorithm

AI-generated content may be incorrect.

1. **Squeeze-and-Excitation** **block**
   1. Enhances CNNs by allowing them to **focus on important features in an image** by assigning importance weights to each individual channel. Still works on grayscale images as it works on the extracted feature maps of which their can be plenty and determines which is most important.
   2. How It Works:
      1. **Squeeze** – Uses global average pooling to compress each feature channel into a single value, summarising the entire feature map.
      2. **Excitation** – Passes these values through small fully connected layers to learn the importance of each channel.
      3. **Scale** – The learned weights are multiplied with the original feature map, highlighting important channels and suppressing less useful ones.

A diagram of a diagram of a diagram

AI-generated content may be incorrect.

1. **Self-Attention Mechanism**
   1. Enables models to focus on important parts of an input sequence by capturing dependencies and relationships effectively.
   2. Read that section of the paper as its hard to summarise mathematical equations.
2. **EmoNeXt Final Architecture**
   1. **STN** integrated at the beginning to handle variations in scale, rotation, and translation in facial images.
   2. **Downsamples** the input using a non-overlapping convolution with a kernel size of 4, reducing dimensionality and efficiently capturing features.
   3. **ConvNeXt Stages with SE Blocks:**  
      Processes the downscaled inputs through multiple stages, each followed by an SE block to recalibrate feature maps and enhance discriminative facial feature extraction.
   4. **Overall Performance:**  
      Combines these techniques to achieve robust and accurate facial emotion detection, effectively handling variations in facial expressions.

A diagram of a block diagram

AI-generated content may be incorrect.

1. **Results**
   1. Model was trained on the **FER2013** dataset which has a lot of images but **imbalanced** **class distribution**.
   2. Data Augmentation – Random Cropping & Random Rotation
   3. Prevent Overfitting – Stochastic Depth & Label Smoothing
   4. Alleviate overfitting – Exponential Moving Average
   5. Memory constraints – Mixed Precision
   6. Pre-Trained Image weights trained on the 1k or 22k ImageNet dataset
   7. Images resized to 2242 – standard industry practice

A screenshot of a document

AI-generated content may be incorrect.