**DDAMFN++**

1. Proposes a new backbone called **Mixed Feature Network (MFN)** which is built upon the **MobileFaceNets** foundation a lightweight network tailored for face verification tasks.
2. Attention mechanisms offer a potential solution to two significant FER challenges: small inter-class difference and large intra-class difference by establishing connections among various facial regions.
3. The **Dual-Direction Attention (DDA)** generates attention maps from both vertical and horizontal directions. Subsequently, multiplying the attention map obtained from the **Dual-Direction Attention Network (DDAN)** with the input feature map results in a new feature map. This feature map undergoes a linear **Global Depthwise Convolution (GDConv)** layer, followed by a reshaping operation. A fully connected layer is employed to generate the conclusive results. By integrating the proposed DDA head and subsequent processing steps, the model’s ability can be enhanced.
4. This work integrates the **MFN** and the **DDAN**, presenting a novel model named the Dual-Direction Attention Mixed Feature Network (DDAMFN)

A screenshot of a document

AI-generated content may be incorrect.

1. Architecture consists of 2 main components, the **MFN** which produces basic features maps from the image input. The resultant feature maps are then fed into the **DDAs** which generate vertical and horizontal direction feature maps. Eventually, attention maps are reshaped to specific dimensions, and the expression category of the images is predicted by a fully connected lay

A diagram of a block diagram

AI-generated content may be incorrect.

1. The MFN architecture is built using two key types of blocks:
   1. **Residual Bottleneck Block:** This block uses shortcut connections to capture complex features and improve gradient flow, helping the network avoid degradation during training.
   2. **Non-Residual Block:** This block enhances the network’s ability to represent diverse and discriminative facial features, which is important for facial expression recognition (FER).
   3. Additionally, the MFN employs several strategies to improve performance:
      1. **MixConv Operation:** Integrated into the bottleneck, this operation uses multiple kernel sizes simultaneously, allowing the network to capture a wider range of features.
      2. **Activation and Depth Adjustments:** PReLU is chosen over ReLU for better feature extraction, and the network depth is carefully tuned to prevent overfitting, especially on smaller datasets.
      3. **Coordinate Attention:** This mechanism is added to each bottleneck to further refine the feature extraction process.
2. The DDAN architecture uses multiple independent attention heads that work in two directions—horizontal and vertical—to capture long-range dependencies in the network. Here’s a simplified breakdown:
   1. **Dual-Direction Attention Heads:** Each head creates two sets of direction-aware feature maps: one for horizontal features and one for vertical features.
   2. **Modified Pooling:** Instead of average pooling, a linear GDConv is used. This allows the model to assign different importance to various spatial positions.
   3. **Attention Map Creation:** The two feature maps (one horizontal and one vertical) are multiplied element-wise to produce a combined attention map that matches the size of the input feature map.
   4. **Selection of the Best Map:** Among the attention maps generated by all the heads, the one with the highest saliency (i.e., the most informative) is selected as the final attention map.
   5. **Applying Attention:** The final attention map is multiplied element-wise with the input feature map to highlight the most important features.
   6. **Attention Loss:** To ensure that each head focuses on different facial areas, an attention loss is introduced. This loss is calculated by taking the Mean Squared Error (MSE) between every pair of attention maps from different heads, and then taking the reciprocal of their sum.
3. Finally, the feature map (7×7×512) from the DDAN is first processed by a linear GDConv layer and then a linear layer, after which it is reshaped into a 512-dimensional vector. A fully connected layer then produces the class confidence scores. The model is trained using a loss function that combines the standard cross-entropy loss (L₍cls₎) with an attention loss (Latt). The overall loss is defined as: L = Lcls + αLatt Here, α is a hyperparameter set to 0.1 by default.
4. 2 attention heads had the best result
5. Performance:
   1. AffectNet-7 – 67.03%
   2. AffectNet-8 – 64.25%
   3. RAF-DB – 91.35%
   4. FERPlus 90.74%

**DDAMFN:** The Dual-Direction Attention Mixed Feature Network (DDAMFN) as denoted in is specifically designed for facial expression recognition (FER), boasting both robustness and lightweight characteristics. The network architecture comprises two primary components: the Mixed Feature Network (MFN) serving as the backbone, and the Dual-Direction Attention Network (DDAN) functioning as the head . The architecture design for the DDAMFN model consists of a Mixed Feature Network (MFN) backbone, enhanced with mixed-size kernels to extract resilient features, alongside a Dual-Direction Attention Network (DDAN) head that captures long-range dependencies by generating attention maps in both horizontal and vertical directions .

The Mixed Feature Network (MFN) is built upon MobileFaceNets and enhanced by introducing mixed depthwise convolutional kernels, which exploit advantages from different size kernels . Coordinate attention is introduced into the MFN architecture to facilitate the capture of long-range dependencies, extracting meaningful features for FER . The MFN leverages convolution kernels of varying sizes to capture diverse spatial information from the facial images . The MixConv operation, consisting of multiple-size kernels arranged as depicted in Figure 4, is integrated into the network’s bottleneck .

The Dual-Direction Attention Network (DDAN) generates attention maps in both the vertical and horizontal directions . The DDAN consists of multiple independent DDA heads, each contributing to capturing long-range dependencies within the network . Attention heads generate direction-aware feature maps from both the horizontal and vertical directions . The average pooling operation is replaced by linear Global Depthwise Convolution (GDConv), which helps to learn very different importances at different spatial positions . To obtain the final attention map, xh and xw are multiplied element-wise, resulting in an attention map of the same size as the input feature map .

A novel attention loss mechanism is applied to ensure the attention heads of DDAN are focusing on distinct areas, which leads to a notable enhancement in overall performance and discriminative power of the model . The Mean Squared Error (MSE) loss is calculated between each pair of attention maps generated from different dual-direction heads . The attention loss is then defined as the reciprocal of the sum of these MSE losses .