Architectures (Existing and Fresh ideas) - Ifreen

1. Introduction

The motivation for this report is to highlight the standard architectures that could be employed for masked facial identification and most importantly, The new architectures I believe would be particularly useful for our use case. 3 of them don't exist out in the wild and 1 of them are inspirations from other modalities.

2. Baseline Architectures

2.1. Standard Architecture 1

The starting point for this work is the well-established **Standard Architecture 1 (Siamese Network)**, which operates as follows

Input Processing:
The system accepts an image where key facial features are masked.
Siamese Network Usage:
Both the masked image and a corresponding unmasked image are processed using a siamese network
Weight Sharing
The network employs shared weights across both branches, ensuring that comparable features are extracted.
Loss Optimization:
A contrastive loss function is used to optimize the similarity between the feature representations of masked and unmasked images.

2.2. Standard Architecture 1 Variant

Building on the baseline, a variant is considered

Non-Weight-Shared Siamese Configuration:
Separate pathways for masked and unmasked images are used instead of weight sharing.
Swapped Image Pairs for Training:
Training involves creating swapped image pairs to ensure that both pathways learn to produce aligned feature representations.
Purpose:
This modification is designed to ensure both the architectures see both the images. (Indirect Weight Sharing)

New ARCHITECTURAL IDEAS

3.1. Idea #1: Feature-by-Feature Matching and Aggregation

Overview

I propose a fine-grained approach that extracts and matches individual facial features from the exposed regions. The key idea is to process each exposed facial component separately and then aggregate the resulting feature embeddings.

Methodology

Exposed Region Identification:

Detect and isolate the exposed parts of the face from the masked image.

Regional Feature Extraction:

Process each exposed region through a dedicated CNN to extract localized embeddings.

In parallel, extract features from the full-face (base) image using a similar CNN.

Weight Sharing Across CNNs:

Employ tied weights between the CNNs to ensure consistency in the feature extraction process.

Similarity Scoring:

Compute a weighted score by comparing corresponding feature embeddings from the masked and unmasked images.

Emphasize regions that are critical for identity verification.

What could be the benefits ?

- Allows for a feature-level examination of the face.
- Weighted aggregation helps us to understand which part of the face is more crucial in making a decision.

Potential Issues ?

- Will independent facial feature recognition work as robust as to looking at the entire face and making a decision ?
- Computational Feasibility. Multiple passes through the network for the same image.

3.2. Idea #2: Architecture Approach (Neuron Zeroing)

Overview

This proposal introduces neuron zeroing—a strategy that segments the CNN outputs to focus specifically on the neurons corresponding to the exposed areas of the face.

Methodology

Dual Input Processing:
Process both the masked and fully exposed images using identical CNN architectures with shared weights.
Flattening of CNN Outputs:
Flatten the outputs to create a one-dimensional feature vector for each image.
Neuron Selection:
For the masked image, selectively retain neurons that correspond to the exposed facial regions (a choice we make), zeroing out the rest of the neurons.
For the fully exposed image, retain the entire feature vector.
Integration via Fully Connected Layer:
Pass the modified masked image output (with most neurons zeroed) through a fully connected neural network.
Loss Function:
Use a contrastive loss to minimize the difference between the masked and unmasked representations.

What could be the benefits ?

- Ensures that only the relevant neurons contribute to the final representation.
- With limited dataset, This kind of acts like a regularizer ? The subset of neurons lowers the risk of overfitting ?

Potential Issues ?

- While the idea of fixing certain neurons for certain features isn't new, How well will this translate to during backpropagation ?
- Training time ? How slow would be the convergence if only a portion of neurons is being made to act at each forward pass.

3.3. Idea #3: Architecture Approach (Neuron Zeroing Variant)

Overview

A variant of the neuron zeroing concept.

Methodology

Single CNN Processing:
Process the masked image with a CNN and flatten the output.
Partition the flattened neurons into 10 corresponding segments.
Selective Neuron Activation:
Based on which segments of the face are exposed, retain the corresponding neurons while zeroing out the others.
Classification via Softmax:
Use a softmax classifier to compare the resulting feature representation against a database of known faces.

3.4. Idea #4: Cross-Encoder Based Approach (Sentence Similarity)

Overview

This proposal involves a cross-encoder strategy where masked and unmasked images are processed together to derive a similarity score.

Methodology

Image Stacking:
Stack the masked and unmasked images (either vertically or channel-wise) to form a composite input.
Joint CNN Processing:
Process the composite image through a CNN, enabling joint feature extraction from both images.
Embedding Extraction:
Joint embeddings of both images.
Similarity Estimation:
Train a simple similarity scorer function.

What could be the benefits ?

Joint processing allows for the capture of complex interdependencies between the two images.