Face processing in humans and deep neural networks share similar computational principles

Devanand T[1], Yahav Atas[1], Ariel Goldstein[2,3], Dan Vilenchik [4], and Carmel Sofer[1,*]

[1]Department of Cognitive & Brain Sciences, Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Israel
[2]Department of Cognitive Science, The Hebrew University of Jerusalem, Jerusalem, Israel
[3]The Hebrew University Business School, Hebrew University of Jerusalem, Jerusalem, Israel
[4]School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Israel
[*]carmelso@bgu.ac.il

ABSTRACT

Facial expressions play a vital role in communication and interpreting emotional states. Previous studies have highlighted the parallels between hierarchical processing of vision, speech, and text in the brain and deep neural networks. Building on this perspective, our study investigates the use of neural network models, specifically deep convolutional neural networks (DCNNs), to understand facial expressions. Our computational experiments reveal that DCNNs process facial images in three stages: extracting identity information, followed by valence and arousal information, and finally recognizing basic categorical expressions. This three-stage processing which naturally emerged among multiple facial expression detection neural networks leads us to view discrete categorical facial expressions as components of underlying valence and arousal signals. Surprisingly enough, we were able to reproduce the IAPS-like valence/arousal representation from the penultimate layers of neural networks by simply feeding facial images to these DCNNs. Our analysis further revealed that identity information is the dominant signal in facial images, regardless of whether they were processed through unsupervised algorithms or supervised neural networks. Additionally, we observed that even untrained neural networks tend to detect identity information in the early stages/layers of processing.

[Dev says:TO DO: Maybe add Rebecca Saxe work 2019 psy-arxiv which got published recently.]

1 Introduction

Facial identity and emotional expression are important sources of information for social interaction. Our ability to recognize an individual's identity allows us to collect and retrieve information about them (Bruce & Young, 1986; Mende-Siedlecki, Cai, & Todorov, 2013). Recognizing an individual's expressions can infer their emotional states and predict their future behavior (Wagner, MacDonald, & Manstead, 1986). Due to their social importance, there has been much debate about the link between identity and emotional expression aspects of face processing, as well as about whether emotions are structured along dimensions of valence and arousal or as discrete entities.[Ariel says:The claim about debate should be backed with citations.]

1.1 Facial identity and emotional expressions

The majority opinion in face-perception research has been that facial identity and facial expression are identified using separate visual pathways at the neural and functional levels (Bruce & Young, 1986).[Ariel says:Are there more recent papers support this claim?] Neuropsychological evidence indicates that emotion and identity processing is divided into parallel routes; In some patients, facial identity recognition but not emotion is impaired (Bruyer et al., 1983; Todorov & Duchaine, 2008) , whereas in others, facial expression (or specific emotions) discrimination, but not identity, has been impaired (Adolphs, Tranel, Damasio, & Damasio, 1994). However, the recent opinion is that no solid evidence supports the "full separation" view (Calder & Young, 2005) ; Prosopagnosics often struggle to recognize expressions (Calder & Young, 2005) ; It is possible to separate expression recognition from identity recognition deficits (Etcoff, 1984; Young, Newcombe, Haan, Small, & Hay, 1993) , but such dissociations are possible later in processing, and they do not preclude substantial integration of expression and identity recognition before that point (Calder & Young, 2005) . Several other views have been advocated, the most dominant one being Haxby et al. (Haxby, Hoffman, & Gobbini, 2000) . The authors supported the assumption that two neural pathways exist. However, in their model, one pathway encodes invariant facial properties (like identity) while the other encodes changeable facial properties (like expression, lip speech, and eye gaze). Two additional views suggest that identity recognition has primacy to expression recognition. The first account argued that emotion processing depends on facial identity coding but not vice

versa (asymmetric dependency (Schweinberger & Soukup, 1998) ; Participants were able to respond to face identity, but ignored emotional expressions. However, when responding to emotional expressions, participants could not ignore identity. The second account proposed that representations of identity contribute to expression recognition, and vice versa (symmetric dependency; e.g., (Ganel & Goshen-Gottstein, 2002, 2004; Wang, Fu, Johnston, & Yan, 2013). (Ganel & Goshen-Gottstein, 2002, 2004) found symmetric interference between facial identity and emotions in familiar faces. They proposed that there is an interconnection between recognition of familiar identity and expressions, such that facial identity serves as a reference, easing the recognition of different expressions.

[Ariel says:In general I will say hat the citations are a bit old (the latest paper cited was published 10 years ago, and most of the 20). It will be more convincing if we could connect our work to recent work/questions.]

1.2 The structure of emotions

Psychological studies regarding emotion structure are also a matter of much debate. Researchers have argued over whether emotions are best explained using dimensions of valence and arousal (dimensional model) or categorical accounts (category-based accounts) (e.g., Barrett, 1998; Ekman, 2004; Ekman & Cordaro, 2011;Keltner & Cordaro, 2015).

The category-based account posits that a limited number of emotions have a 'basic' status, activating discrete category representations for each emotion. Ekman (Ekman, 2004; Ekman & Cordaro, 2011) proposed a list of "basic" discrete emotions that evoke a specific response tendency that addresses a specific evolutionarily significant need (e.g., avoiding potential aggression by fear, rejection of harmful substances by disgust). One of the main reasons these emotions have this status is that their corresponding facial expressions are recognized by several cultures worldwide (Ekman, 1994; Ekman, Sorenson, & Friesen, 1969).

In contrast to the category model, the dimensional model posits that facial expressions are recognized by their positions in a continuous two-dimensional space, which is often regarded as synonymous with a tendency to approach or avoid a person. Bradley and Lang (Bradley & Lang, 2007) suggested that the combined dimensions of arousal and valence form an urging or aversive emotive orientation (i.e., arousal determines how intense and vibrant the emotion is, and valence determines the emotional direction). Exponents of the model posit that it plays a significant role in face perception (Jack, Garrod, & Schyns, 2014; Mehu & Scherer, 2015) ,overcoming possible confusion between discrete entities such as disgust and anger or fear and surprise (Calder, Burton, Miller, Young, & Akamatsu, 2001; Woodworth & Schlosberg, 1954). Other researchers view the model as related only to the post-perceptual stage at which properties of the categorical expression are interpreted (e.g., (Calder et al., 2001)) .

Recently (Liu et al., 2022) used communication theory to interpret the system of facial expression communication.[Ariel says:Lets move to the intro and elaborate] They found that both category and dimensional information shape face perception. Some facial movements (Action units – AU) convey emotion category (e.g., anger); others convey dimensional information (e.g., negatively aroused); and some convey joint, multiplexed information of both dimensions and category. The authors suggested that AUs which convey dimensional information evolve before AUs that convey expression category (asymmetric dependency) because dimensional information can predict specific emotion categories, but not vice versa. Another possibility is that AUs are symmetrically dependent, producing a (quick) and clear face perception by emphasizing relevant AUs (categorial or dimensional) required for specific social communication.

Critically, all these previous studies were constrained by focusing, as separate subjects, on either identity-expression aspects or representation forms of expressions (dimensional or categorial). For example, most discussions about integrated /separated routes of identity/expression recognition assume, explicitly or implicitly, that expressions are categorial. Therefore, they might have overlooked a general system of rules governing face perception (Calder et al., 2001) . For example, suppose facial expressions (in part or as a whole) indeed are dimensionally perceived (Liu et al., 2022) ; Other relationships between identity and this form of expression information should be elaborated (e.g., different brain areas may be involved in face processing (Viinikainen et al., 2010)).[Ariel says:I read this paragraph several times, it seems very important because it hints as to what we are going to add to this literature. However, I don't understand the specifics. The second "for example" is an example of the first "for example"?. If the idea is that CNN can encompass two different perceptual processing (identity and emotion), which in the past were researched seperatly, I would phrase it directly.]

Taking it all together, we propose that a holistic model of face perception should include three perceptional aspects: recognition of identity, coarse expression information (dimensional factors), and exact discrete expressions (categorial aspects).[Ariel says:I see now what I just written.. but this make the previous paragraph more cryptic.] There is an agreement among researchers that identity coding is a prime aspect that affects emotion processing, asymmetrically or symmetrically (e.g., (Ganel & Goshen-Gottstein, 2002, 2004); (Schweinberger & Soukup, 1998); (Wang et al., 2013)) . Therefore, we expect this aspect to be readily available for perception – more distinctive and first to evolve before other aspects are noticed. The second postulated aspect is dimensional coding, enabling coarse expression recognition. Dependent on facial identity coding, the dimensional aspect is broad and robust. It can, for example, quickly derive "avoid" behavior before even establishing the cause of alarm

(Liu et al., 2022) . Dimensional coding can also reduce the chance of confusion between discrete emotions (later to come), such as disgust and anger (Calder et al., 2001; Woodworth & Schlosberg, 1954) . The third aspect is the expression category. Category aspects can refine the perception and subsequent behavioral responses. For example, the "avoid" response driven by the dimensional aspects can now be refined to potential danger (anger) or a sense of disgust.

The interplay between these three aspects may affect face perception differently than known today. Adding the identity aspect to the equation can lead to new predictions regarding the extent to which the identity aspect is linked with valence, arousal, and specific emotion categories.

1.3 Parallels in hierarchy learned by brain and DNN models

[Ariel says:As this paper relates to face perception I would also cite: Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks , Face dissimilarity judgments are predicted by representational distance in morphable and image-computable models.] There have been studies which indicated that the brain and Deep Neural Network (DNN) models share hierarchical similarities in processing various kinds of stimuli, including speech, language, and vision (Goldstein et al., 2022; Khaligh-Razavi & Kriegeskorte, 2014; Millet et al., 2022; Schyns, Snoek, & Daube, 2022; Yamins & DiCarlo, 2016). In the language processing context, the human brain and autoregressive deep learning models perform continuous context-dependent next-word prediction, match pre-onset predictions to incoming words in order to generate surprise or prediction-error signals, and represent words using contextual embeddings (Goldstein et al., 2022). In the context of speech processing, Juliette Millet et al. (Millet et al., 2022) showed that self-supervised algorithms trained on raw speech waveform input might be a promising candidate for generating brain-like responses. Through experiments involving functional magnetic resonance imaging (fMRI) of English, French, and Mandarin speakers listening to audiobooks, the authors found that the self-supervised algorithm Wav2Vec 2.0 (Baevski, Zhou, Mohamed, & Auli, 2020) can learn brain-like representations with as little as 600 hours of unlabeled speech. Its functional hierarchy aligns with the brain's cortical hierarchy of speech processing. Similarly, in the case of vision, we have studies from Nikolaus Kriegeskorte group (Khaligh-Razavi & Kriegeskorte, 2014; Kriegeskorte, Mur, & Bandettini, 2008) which showed that the deep convolutional neural network (DCNN) model explained Inferior temporal (IT) cortex in human and nonhuman primates very well. They compared the representational dissimilarity matrices (RDMs) of the DCNN with RDMs derived from human and monkey IT for the same set of stimuli and found similarities. Ans more specifically, in the case of face recognition, remarkable similarity is observed between the tuning properties of V1 neurons and the artificial neurons located in the initial (most upstream) layers when examining their characteristics across various levels of the deep convolutional neural network hierarchy. (Grossman, Malach, et al., 2023; Vogelsang et al., 2018) Also evidence(Grossman et al., 2019) suggests that the involvement of face-selective areas in mid-high visual cortex in face perception and recognition based on fMRI, clinical, and stimulation data. However, understanding the precise role of neuronal groups in the processing cascade and overall human perceptual capabilities remains challenging. Their study highlights the pictorial function of human face selective neuronal groups, showing consistent matches between DCNN and neural representations.

In short, many recent studies indicate shared computational principles, including hierarchies of processing when it comes to the functional units of deep learning models and the brain. In this work, we are interested in how the human brain perceives human faces with respect to identity information versus emotion information, discrete versus continuous perception of emotions, and whether deep neural networks capture those features and hierarchies at their functional computational units.

2 Methodology

We intend to tackle the central problem of facial perception with a popular computational tool called Deep Convolutional Neural Network (DCNN). Deep Convolutional Neural Networks are a category of biologically inspired supervised learning algorithms that have revolutionized computer vision problems from image classification and detection to localization tasks (Krizhevsky, Sutskever, & Hinton, 2017; Sermanet et al., 2013; Zeiler & Fergus, 2014). The DCNNs inspired by the visual cortex capture the stages of human visual processing in the brain from early visual areas to the dorsal and ventral streams in their hidden layers (Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016). The convolutional layer ("convnet") is the core unit of DCNNs, which has a set of filters that learn the basic building blocks of images. Another component called pooling layers is responsible for down-sampling the data. As the convnet layers progress from early to later stages, the network captures local features initially and later global ones. A typical DCNN will have numerous "convnet" layers of different dimensions and later fully connected flat layers before the final output layer. While the initial convolution filters are usually smaller relative to the input image dimension, the extracted features are local. However, the higher-order convolutions capture global features which depend on all pixels roughly in the images. The local features mentioned here are edges at different angles, curves, and other round features. And the higher-order layers will capture more complicated structures. Although DCNNs do not model the temporal structure of information flow through its layers, it is surprising to note that the sequential breakdown of individual intermediate layer outputs seems to correlate with data representation in visual processing units of the brain (Cichy et al., 2016).

One specific attractive feature of Convolutional neural networks is the fact that there are tools like Grad-CAM (Selvaraju et al., 2017) which make it possible to dissect what each convnet block does. The Grad-CAM analysis makes interpreting DCNN functioning possible in a very elegant way, which will come in handy in our context of understanding facial features learned by the network.

In the context of facial expression detection, DCNNs were used in past too and has added valuable insights to the field. One key study was by Khorrami et al (Khorrami, Paine, & Huang, 2015) who unraveled the fact that CNNs trained to detect facial expressions also discovered facial action units (FAUs) in the intermediate layers. Such studies present motivation for our work which intend to inspect crucial information of identity and graded valence/arousal dimensional signals, etc in the process of detecting categorical expressions.

From a mechanistic perspective, our problem is primarily a computer vision task that will shed insights to cognitive understanding of emotions from facial perception. Therefore, DCNN seems to be a suitable tool in this context. The neural network that we chose was designed in an archived Kaggle Notebook by Gaurav Sharma (Sharma, 2020) aimed at detecting expressions from facial images. (Please see Table 1 and Figure 1 for architecture, specifications, and different variants of this Neural Network.)

As the first step, we used a lightweight DCNN (Sharma, 2020) to detect emotions. We trained and tested this network on FER2013 (Goodfellow et al., 2013) and DISFA (Mavadati, Mahoor, Bartlett, Trinh, & Cohn, 2013) datasets respectively. We inspected data representation in different intermediate layers of this DCNN. Because understanding what the intermediate DCNN layers do to the facial expression images is crucial, we resort to Grad-CAM (Selvaraju et al., 2017) analysis, and dimensionality reduction algorithms like PCA (Calder et al., 2001) and UMAP (McInnes, Healy, & Melville, 2018). We use these algorithms as the computational lens on the outputs generated at intermediate layers.

UMAP (McInnes et al., 2018) is an algorithm that can be used to visualize high-dimensional data in a low-dimensional space, like a 2D or 3D plot. When we have a large dataset with many features (like pixel values in an image or gene expression levels), it can be difficult to understand the relationships between different data points. UMAP helps us overcome this challenge by reducing the dimensionality of the data, while preserving important relationships between points. The UMAP algorithm works by measuring the distances between each data point in the high-dimensional space, and then finding a way to represent these distances in a lower-dimensional space. This new space is chosen to minimize a cost function that balances the preservation of local and global relationships between data points. Essentially, UMAP tries to preserve the structure of the data in the new, lower-dimensional space, while also ensuring that the points are spread out enough to be easily distinguishable. Once the algorithm has computed the new, lower-dimensional space, we can plot the data points in this space to visualize their relationships. This can help us understand the structure of the data and identify patterns that might be difficult to see in the original high-dimensional space. Overall, UMAP is a powerful tool for exploring and visualizing complex datasets, and it has many applications in fields like machine learning, biology, and social science.

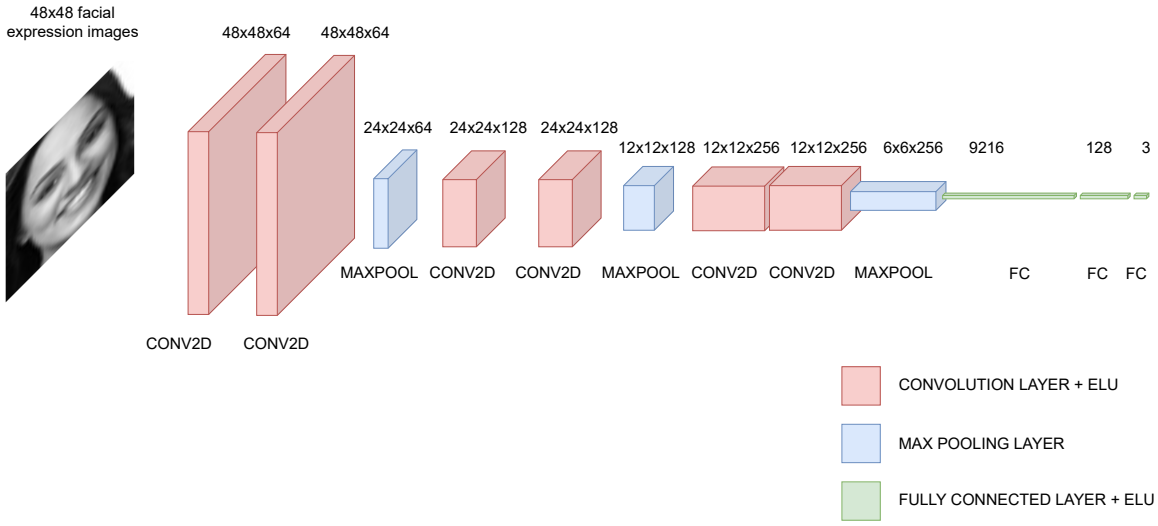
ID	Training Data	Labels	Accuracy
NN2	FER2013(35887 samples)	AN(13.8%),DI(1.5%),FE(14.2%),HA(25%),NE(17.3%),SA(16.9%),SU(11.1%)	65.90%
NN1	FER2013	HA,SA,NE	82.32%
NN3	FER2013	Untrained Network	31.41%
RMN	FER2013 and,	(see first row)	74.14%,
	VEMO(36470 samples)	AN(16.2%),DI(5.2%),FE(6.8%),HA(19.3%),NE(17.9%),SA(15.1%),SU(19.4%)	65.94%

(a) Details of Neural Networks used in *Study1* to *Study3*. All 7 emotion categories are AN(Anger),DI(Disgust),FE(Fear),HA(Happy),NE(Neutral),SA(Sad), and SU(Surprise). NN1 and NN2 are the lightweight CNNs(Sharma, 2020) trained to detect 3(HA,SA,NE) and all 7 expressions respectively. NN3 is an untrained neural network with the same architecture as NN1. RMN stands for ResMaskNet(Pham, Vu, & Tran, 2021) neural network and VEMO is a dataset which was created by Pham et al.

Dataset	Sample Size	Label Info	Resolution
DISFA	27 videos with 15 male and 12 female participants responding to 4 minute stimuli (∼5700 frames selected)	Unlabeled (Expression labeling was done via ResMaskNet)	1024×768
FER2013	∼ 36K samples	7 expression labels	48×48
KDEF	980 images of 35 males, 35 females spanning 7 expressions each comprising of 140 samples	Yes (identity and expression)	562×762

(b) Datasets used in this study.

Table 1. (a)Neural Network specifications, and (b)Datasets used in this study. Also see Figure 1



(a) The convolutional neural network with a three cascades of two sequential convnets and one maxpool layer connected finally with fully connected flatten layer and intermediate fully connected layer before output layer. Exponential Linear Unit (ELU) activation functions were used for convnet units, and 128 dimensional fully connected layer, and finally softmax activation for the final layer.

CNN Model
INPUT: (48 x 48) pixel gray-scale images
CONV2D FILTERS (5x5,64) [``**conv2d_1**''], BATCH NORM [``**batchnorm_1**'']
CONV2D FILTERS (5x5,64) [``**conv2d_2**''], BATCH NORM [``**batchnorm_2**'']
MAX-POOL2D (2x2,64) [``**maxpool2d_1**''], DROPOUT (0.4) [``**dropout_1**'']
CONV2D FILTERS (3x3,128) [``**conv2d_3**''], BATCH NORM [``**batchnorm_3**'']
CONV2D FILTERS (3x3,128) [``**conv2d_4**''], BATCH NORM [``**batchnorm_4**'']
MAX-POOL2D (2x2,128) [``**maxpool2d_2**''], DROPOUT (0.4) [``**dropout_2**'']
CONV2D FILTERS (3x3,256) [``**conv2d_5**''], BATCH NORM [``**batchnorm_5**'']
CONV2D FILTERS (3x3,256) [``**conv2d_6**''], BATCH NORM [``**batchnorm_6**'']
MAX-POOL2D (2x2,256) [``**maxpool2d_3**''], DROPOUT (0.5) [``**dropout_3**'']
FLATTEN FULLY CONNECTED (9216) [``**flatten**'']
FULLY CONNECTED (128) [``**dense_1**''], BATCH NORM [``**batchnorm_7**'']
DROPOUT (0.6) [``**dropout_4**'']
OUTPUT: (3) [``**out_layer**'']

(b) The basic architecture of the DCNN in table format with relevant layer names in square brackets

Figure 1. The basic architecture of trained convolutional neural networks (Sharma, 2020). We have three neural networks with same architecture: "NN1", "NN2", and "NN3". The NN1 is trained to detect 3 expressions (Happy,Sad, and Neutral) and so it has an outputsize of 3. NN2 is designed to detect all 7 expressions (So it will have an outputsize of 7). NN3 has identical architecture of NN1, but is untrained.[Ariel says:We cannot present this. Either we put this in the appendix, or we present visually (you can see my nature neuroscience paper for an example ()]

3 Data and Networks

The data used for the five studies were FER2013 (Goodfellow et al., 2013), KDEF dataset (Lundqvist, Flykt, & Öhman, 1998), and DISFA (Mavadati et al., 2013) datasets. (Also see Table 1) FER2013 is a collection of around 36K (48 x 48 pixels) grayscale facial images with specific expressions, which are labeled into seven categories. The seven categories are Happy, Sad, Surprise, Disgust, Fear, Anger, and Neutral expressions. We trained "NN1" with this FER2013 data from three categories Happy, Sad, and Neutral. And "NN2" was trained with all 7 labeled data. "NN1" gave superior accuracy as you might notice in Table 1 (a). One reason for this is balanced data in the cases of 3 labeled training. But we will need "NN2" in upcoming experiments as it turns out that it captures dimensional representation of expressions better in the penultimate layer. DISFA dataset comprises of 4-minute facial expression videos from 27 individuals as a result of them watching a specific stimuli video that is capable of eliciting positive, neutral, and negative expressions at different intensities. DISFA dataset is suitable for this study because of its high resolution and limited number of identities, which helps in clustering the data in a very clear way. On the other hand FER2013, which is only used for training of Neural Network has around 36K samples with that many identities. For the testing purposes of this study, unlabeled DISFA dataset is sufficient. But for *Study 3*, we will assign expression labels for DISFA dataset via ResMaskNet neural network.

For the training part we used only the FER2013 data. We chose FER2013 for training because the number of samples is very high which is necessary for NN training (around 36000). Regardless of its low resolution, Neural Networks (Goodfellow et al., 2013; Li & Deng, 2020; Zhang, Zhang, & Tang, 2023) were able to detect crucial features and get a result of around 65 to 70 % accuracy.

4 Results

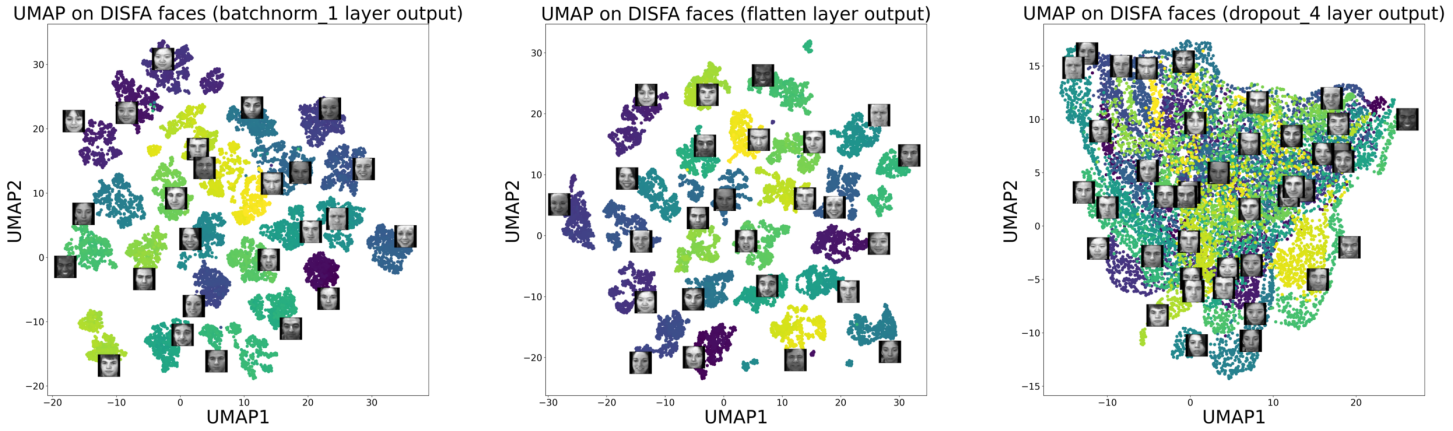
We conducted five studies via data-driven computational analyses. The goal of *Study 1* is to illustrate graphically the fact that the Neural Networks trained to detect emotions extract identity signals first, dimensional representation of emotions next, and finally, categorical expressions. The *Study 2* is aimed to illustrate graphically that the identity signal is the primary dominant semantic signal in the facial images; this point is illustrated using a host of unsupervised learning algorithms like PCA and UMAP, untrained DCNNs, and also trained neural networks but for other tasks than identity detection: the examples in this study (trained for emotions), VGG16 (Simonyan & Zisserman, 2014 (trained for general objects), and ResMaskNet (Pham, Vu, & Tran, 2021) etc. *Study 3* will show that supervised learning algorithms trained to detect expressions, like DCNNs lifts the "expression" signal component from facial images as we progress to final layers. This study will quantify the neural network's capability to disentangle the expression signals from identity signal at different stages. And finally, *Study 4* will show that the penultimate layers in different neural networks trained for emotions (categorical/dimensional) reflect the dimensional representation of the expressions.

4.1 Study 1

In this study, we trained a DCNN to detect emotions and used UMAP as the dimensionality reduction technique to inspect what aspect of the facial expression is represented by each intermediate layer. Our main finding is that the Neural Network detects identity signals in the first layers, then dimensional representation of emotions next, and finally, categorical expressions.

We applied UMAP algorithm on different intermediate and final layer output vectors from neural networks used in this work. So if you feed 'N' samples of facial images to the DCNN, and the respective neural network layer (say for example "dropout_4" layer) is 128 dimensional one, you end up with Nx128 sized array. Alongside UMAP we also used Grad-CAM analysis. Grad-CAM shows the most contributing regions necessary for the detection of expressions from facial images with respect to different convnet filter blocks. For implementing various algorithms in this work, like UMAP and the neural network etc we used *umap-learn*, *scikit-learn*, and *tensorflow* python programming language packages. Several other packages like *numpy*, *pandas* and *matplotlib* used for data analysis and visualization purposes.

We fed the DISFA dataset into "NN1" (refer to Table 1). The main results are shown in Figure 2. As the first step, we analyzed individual layer outputs in the "NN1" network. The analysis comprised applying the UMAP algorithm to get the first two modes on these high-dimensional data and then plotting the scatter plots versus those modes (UMAP1 vs. UMAP2). The convnet blocks ("batchnorm_1" here is used as the example) and "flatten" layer were primarily sensitive towards the identities, which can be inferred from the fact that all these layers clustered into identity-based clusters. Each dot in the scatter plots of Figure 2 corresponds to individual facial snapshot input from the DISFA dataset we feed to the "NN1" network. Figure 2 (a) and (b) scatter plots represent identity clusters, but (c) is a continuous manifold, where UMAP1 and UMAP2 represent the dimensional representation of emotions in terms of valence and arousal. We also computed silhouette scores with respect to identity labels (from DISFA dataset) for clustered data in Figure 2 (a)–(c). For "batchnorm_1" and "flatten" layers we got 0.417 and 0.498 as the silhouette scores, which are high values quantitatively proving the fact that these layers are sensitive towards identity signal in the facial images. For the dropout-4 later (c), we got a negative sillohouete, showing that the cluster stutterer

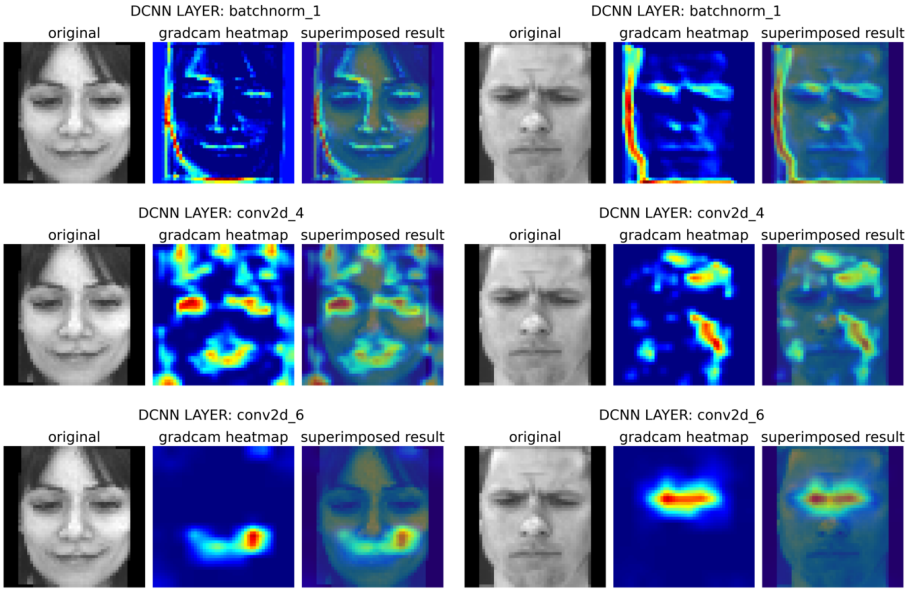


(a) "batchnorm_1" layer (the final layer of first convolutional block) output. Silhouette score with respect to identity labels is 0.417.

(b) 'flatten" layer output. Silhouette score with respect to identity labels is 0.498.

(c) "dropout_4" layer (the 128 dimensional penultimate layer in NN1) output (Silhouette score with respect to identity labels is -0.138)

(d) The first row represents uniformly projected DISFA faces onto Y-axis (UMAP2) from "dropout_4" scatterplot (c). Similarly the second row represents projection on X-axis (UMAP1). From qualitative observation the first and second row represents arousal (passive to active faces) and valence (negative to positive faces) respectively.



(e) Person with with smiling face.

(f) Person with eyebrow frown expression.

Figure 2. Scatter plot results between first 2 modes of UMAP algorithm applied to the outputs from (a)"batchnorm_1", (b)"flatten", and (c)"dropout_4" layers of Neural Network "NN1" (See Table 1 (a)). In (a) and (b), each cluster loosely identifies individual identity (at least one facial image snapshot is projected on most clusters). Visually, the UMAP is better at capturing identity clusters in Convnet blocks and the "flatten" layer (also supported quantitatively with silhouette scores). The disentanglement of identities occurs at the "dropout_4" layer, which can be visualized from the scatter plot in (c) and quantitatively understood via negative silhouette score with respect to identities. Instead of identity clusters, the output represents the dimensional emotional representation, where the projections on UMAP1 and UMAP2 represent valence and arousal, respectively. (d)21 equally spaced DISFA faces projected onto the X(UMAP1) and Y(UMAP2) axis of UMAP scatterplot applied on "dropout_4" outputs. (e) and (f) shows sample Grad-CAM plots from "batchnorm_1", "conv2d_4", and "conv2d_6" layers. The dark blue represents zeros and dark red represents ones in the heatmap figures.

vanished. The semantic transition of facial representations from "flatten" layer to "dropout_4" layer indicates the process of *disentanglement* from identities to emotions.

In Figure 2 (c) we observe that the identity structure is gone. We now explore which structure supplants it. Figure 2 (d) shows the projection of dimensional scatterplot (c) on the UMAP1 and UMAP2 axis. The top row in Figure 2 (d) is the projection on UMAP2, which shows a qualitative increase in the arousal signal as we go from left to right. Similarly, moving from left to right in the bottom row, we observe faces transitioning from negative to positive expressions(valence). This implies that the primary components(latent variables) of information in the penultimate layer "dropout_4" captures dimensional representation of expressions in the form of arousal and valence. The "NN1" network basically achieved 82.32% accuracy in detecting categorical expressions. So from this experiment it is very clear that this network acts like a three block unit, where the first unit(convnet and "flatten" layer) is tuned to detect identities, the penultimate layer detect dimensional expressions and final layer outputs categorical expressions.

Figure 2 (e) and (f) give another visualization of the disentanglement procedure using the Grad-CAM analysis. Grad-CAM shows relevant regions specific to the input while classification process in "NN1" CNN. Applying Grad-CAM analysis to the convnet blocks layer revealed what information each convnet block captured. We see that the first convnet block and the corresponding batch normalization layer "saw" edges of the faces. This block acted as an edge detector. The corresponding Grad-CAM visualization examples are given via two examples. See Figure 2 (e) and (f). As we progress from early to final convnet blocks, the features being captured by this analysis become more specific with respect to the features relevant to expression detection. The "conv2d_4" layer captures intermediate features like eyes, lips, etc. The final convnet block "convd2d_6" is more focused towards particular regions like lips for the smiling face and eyebrow region for the person doing eyebrow frowning.

4.2 Study 2

This study is aimed to prove that identity is the primary and dominant semantic feature extracted from facial expression images. We show this result via unsupervised machine learning algorithms like PCA, UMAP, etc. We also showed this in the "NN1" network previously. Also, as a surprising result supported by Seungdae Baek et al. study(2021), we see that even in the untrained DCNNs (NN3), identity detection happens "innately" purely out of statistical variation in the feedforward projections in the neural network. We also analyzed how VGG16 (Simonyan & Zisserman, 2014) neural network (which is trained to classify object categories) treats facial images. For this purpose, we fed facial images as input to VGG16 network and extracted corresponding intermediate outputs from penultimate "fc2" layer.

The results (except for the trained neural network case) are summarized in Figure 3. For the trained DCNN case, see Figure 2 (a) and (b). We also see that object classifiers like VGG16 network treats each identity as a separate entity (or cluster here). This can be seen from the UMAP scatter plot in Figure 3 (d). All these scatter plots visualize identity clusters. Each cluster in the scatter plot represents an individual identity, and also we have quantified it using average silhouette scores with respect to identity labels. Silhouette scores confirm our premise quantitatively.

4.3 Study 3

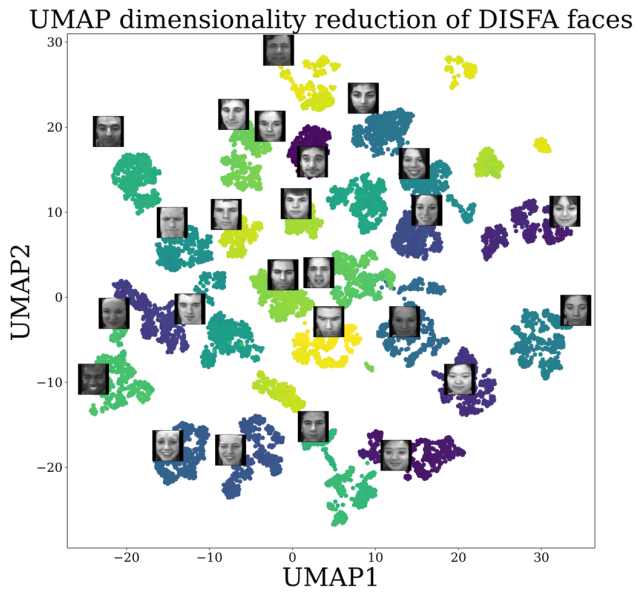
This study aims at a quantitative proof of the disentanglement phenomenon. For this experiment, we used all our networks from Table 1: NN1, NN2, RMN and NN3 as a control group.

For testing, we selected around 970 snapshots from 12 female identities in the DISFA dataset, which were easy to label due to being separate identity-wise videos. Additionally, we used a subset of the KDEF dataset (Lundqvist et al., 1998), consisting of front-pose facial images from 70 identities and all seven expressions, totaling 980 images.

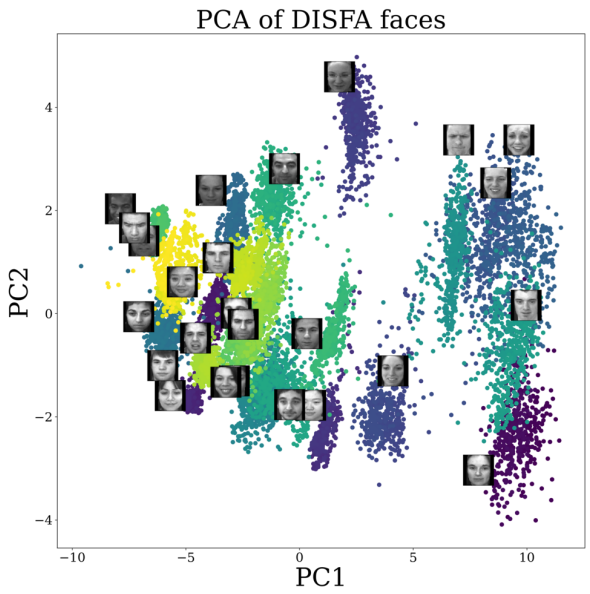
To analyze the expression and identity signal detection strength of the intermediate layers we computed silhouette scores for the embedding of the data at each relevant layer. The clustering was induced by the expression label and by the identity labels (two separate silhouette scores). The distance between two points (needed to compute the silhouette score) was the ℓ_2 Euclidian distance between vectors. We noticed that reducing the dimension of the embedding to 10 (by taking the projection on the first 10 principal components) gave sharper results. Indeed the 10 PCs explain more than 60% of the variance in each case, and we treat the remaining PCs as noise.[Ariel says:Sharper in what way?This feels a bit arbitrary criteria]

Silhouette scores range between +1 and -1, with higher scores indicating clearer data clustering. We plotted these scores with respect to different layers to examine the expression or identity signal detection strength in each layer. This analysis enabled us to identify the layers with the strongest signal detection for expression and identity.

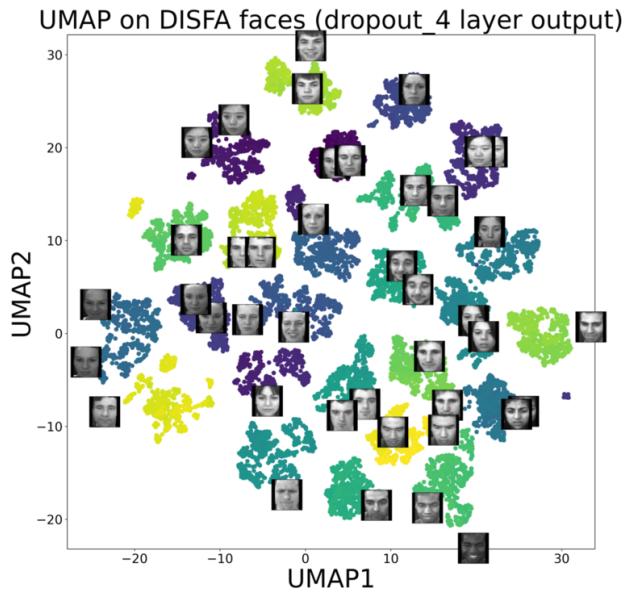
While observing the results in Figure 4, the most crucial aspect which is evident is the *lifting* up of the emotion/expression signals (Fig 4 (a)) from identity signal which happens in all neural networks except "NN3" (which is an untrained neural network used as a control scenario). The results quantify the dominant sensitivity towards identity signals in the early stages (Fig 4 (b)) of neural networks. Another important aspect that is revealed by these analyses is the importance of supervised learning in DCNN, which disentangles expressions from identities as the layers progress towards output (See Figure 4 (a)). The important point to note from this result is that the overall trend for expression detection goes up as the neural network



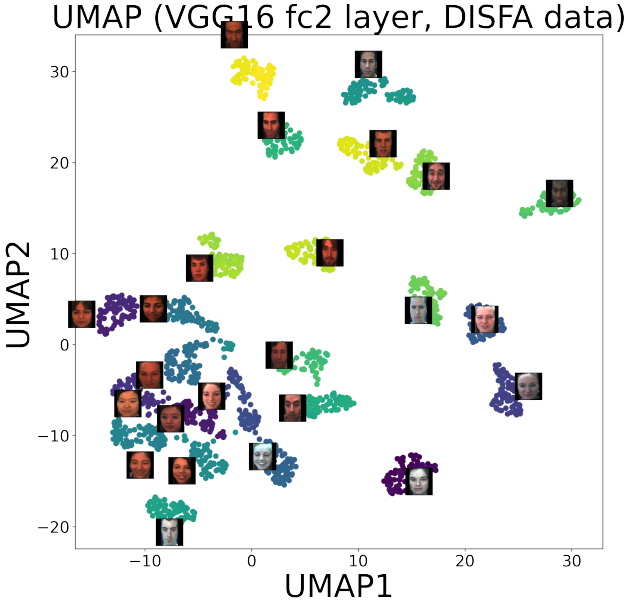
(a) Scatter plot between first 2 modes from UMAP dimensionality reduction applied on DISFA faces (Silhouette score=0.458 with respect to identity labels)



(b) Scatter plot between PC1 (61.6%) and PC2 (6.1%) from PCA applied on DISFA faces (Silhouette score=0.301 with respect to identity labels and considering first 2 PCs. 10 PCs will increase the score to 0.501)

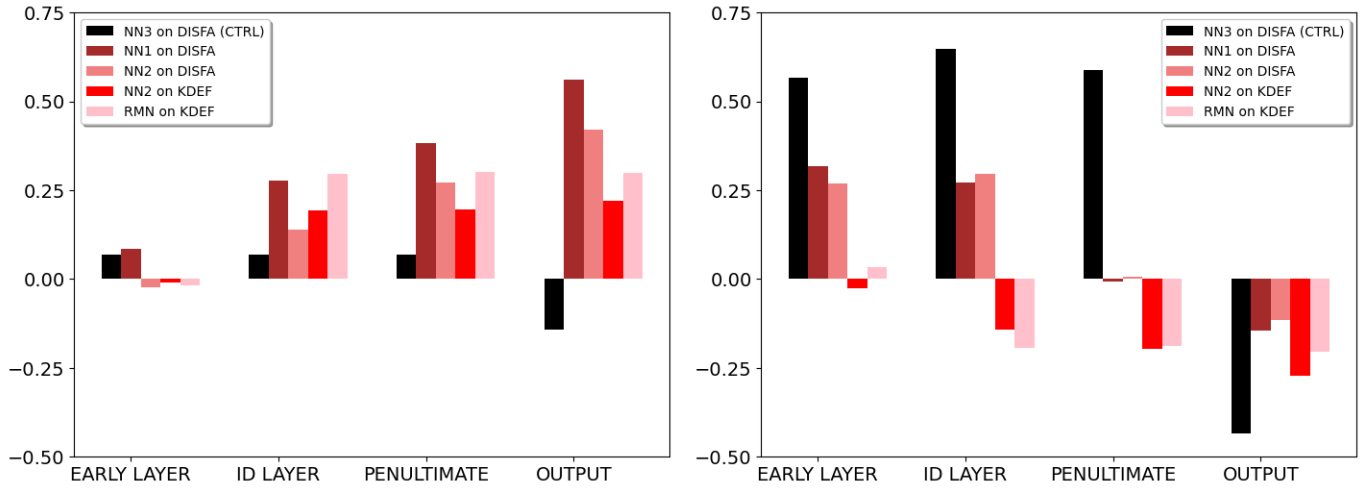


(c) Scatter plot between first 2 modes from UMAP applied on penultimate "dropout_4" layer of untrained "NN3" network (Silhouette score=0.561)



(d) Scatter plot between first 2 modes from UMAP applied on penultimate "fc2" layer of VGG16 network (Silhouette score=0.63)

Figure 3. Scatter plot results between first 2 modes from (a)UMAP algorithm, (b)PCA on DISFA facial images, (c)UMAP on the penultimate "dropout_4" layer outputs of untrained "NN3" network, and (d)UMAP applied on "fc2" layer of VGG16 neural network. Each cluster loosely identifies individual identities (snapshots projected on most clusters). Silhouette scores are used to quantitatively support the argument.



(a) Facial expression signal measured using silhouette scores (with respect to expression labels) based on PCs of different layer outputs

(b) Identity signal detection with silhouette scores of PCs (with respect to identity labels) at different stages of Neural networks

Figure 4. Neural Networks discriminating between facial expression signals and identity information with respect to different layers. We have considered four layers in each. The first layer (EARLY LAYER) considered is "batchnorm_1" for NNs and "maxpool" layer for ResMaskNet (RMN), second laye (ID LAYER) is "flatten" for NNs and "mask4.conv2.downsample" for RMN, the third layer (penultimate layer) is "dropout_4" for NNs and "fc.0" for RMN respectively. The final layer is "out_layer" for NNs, and "fc.1" for RMN. See the details of how scores are computed in *Method* of *Study 3* subsection.

layers progresses towards final layer, and a downward trend is seen in identity signals with respect of layers. The weak identity scores for KDEF dataset cases is due to lack of sufficient samples per identity. KDEF has 980 frontal face samples with 70 identities each with only 14 samples. The higher signal for CTRL case (untrained "NN3") in identity scores until final layer can be attributed to aforementioned study by Seungdae Baek et al. study (Baek, Song, Jang, Kim, & Paik, 2021).

4.4 Study 4

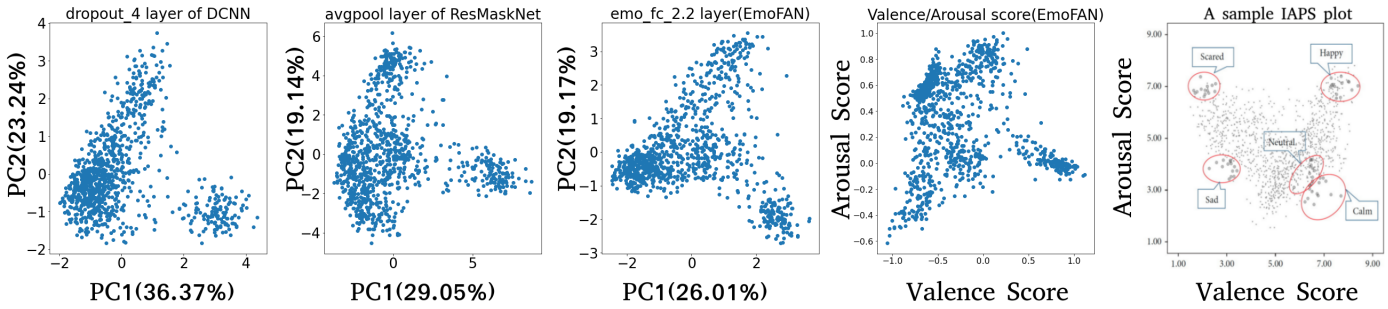
This study aims to prove that several neural networks trained to detect categorical emotions from facial images discover valence/arousal dimensional representation in their penultimate layers. We confirmed this claim for our neural network "NN2", and aforementioned ResMaskNet, and EmoFAN. We chose "NN2" instead of "NN1" since the former was trained on all 7 expression labels. EmoFAN also detects Valence/Arousal scores along with categorical emotions. However, we show that the Valence/Arousal dimensional information emerges in the neural networks before the output layer.

We fed all the selected 980 KDEF facial images to pre-trained networks ResMaskNet, EmoFAN, and our "NN2" networks. Then extracted the penultimate layer outputs and applied PCA on those. We used python packages *tensorflow* and *PyTorch* for analyzing pre-trained networks.

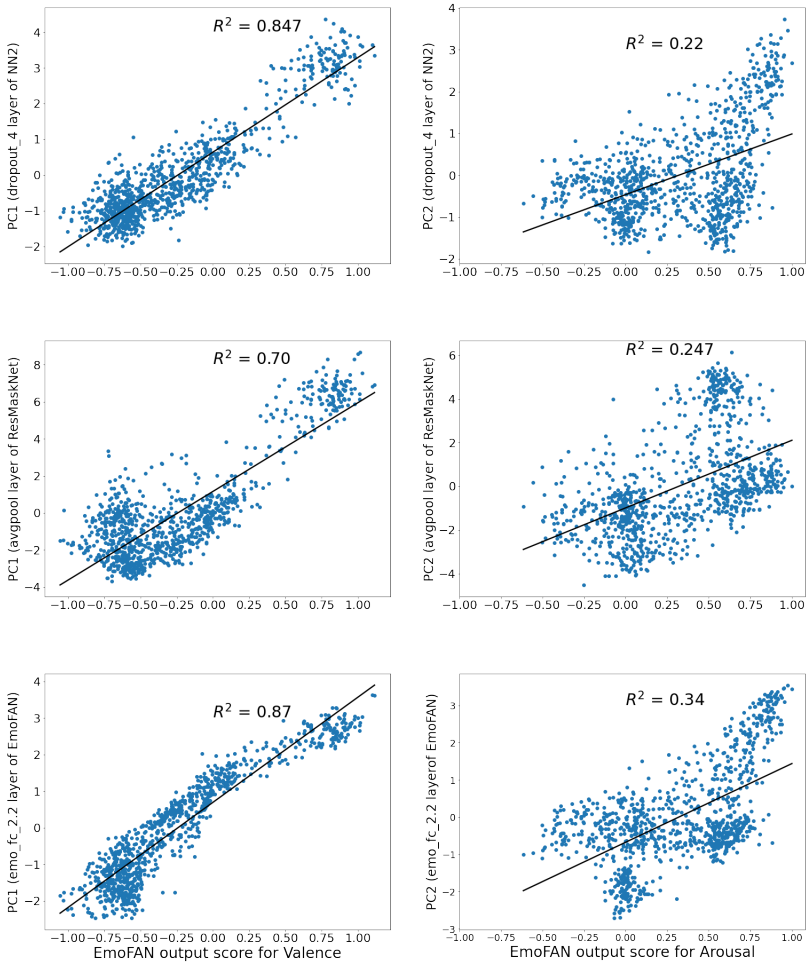
The results are summarized in Figure 5, which shows the scatter plots between PC1 and PC2 of the PCA analysis. For reference, we also plotted the valence and arousal outputs from EmoFAN, and a sample IAPS plot. The IAPS scatter is a result of human participants rating valence and arousal scores from images which can elicit different kinds of expressions. On other hand, the neural networks here "guesses" dimensional representation of expressions. And they both have similar signature feature of a U-curve. The fact that neural networks are able to learn this in the process of discovering categorical expressions is something very interesting and of utmost importance to the cognitive aspect of the research. All the PC1 vectors and valence scores from the EmoFAN network are strongly correlated. The correlation, although weak, is also evident in PC2 vectors and Arousal scores from EmoFAN. This is visible in our linear regression analysis (Figure 5 (b)). Our analysis reveals that the penultimate layers in all these networks discover Valence and Arousal scores in the form of PC1 and PC2, respectively. To validate these results, we presented face projections (from PC1 and PC2 of different networks and the Valence/Arousal outputs of EmoFAN) as done in Figure 2 (d). See Figure 6 for face projections corresponding to what is shown in Figure 5.

[Dev says:TO DO: CATEGORICAL DESCRIPTION + PLOTS]

Furthermore, we projected the ground truth labels to the "dimensional" representation (PC1 vs PC2 plots as in Figure 6 (a)) resulted from penultimate layers of "NN2" and ResMaskNet to see visualize the accuracy of the categorical classification of expressions. Figure 7 shows the results. One important point from these plots is that there is less ambiguity for the network to correctly classify when the data belongs to regimes of extremities whether in Y-axis or X-axis of these scatter plots. The prime



(a) PCA on penultimate layer outputs of "NN2", ResMaskNet, and EmoFAN networks (first three), Valence/Arousal scores from EmoFAN (fourth), and a sample IAPS (International Affective Picture System) plot (final one). The IAPS plot taken and adapted from (Mehmood & Lee, 2015) (CC BY 3.0 licensed).



(b) Linear Regression plots of PC1 and PC2 of penultimate layers from different neural networks ("NN2", ResMaskNet, and EmoFAN)) onto EmoFAN valence/arousal scores. It is obvious from the plots that PC1s are strongly correlated with EmoFAN valence output. The arousal scores are weakly correlated, as expected. All *P-values* < 0.001.

Figure 5. (a)Comparison between EmoFAN valence/arousal scores with PCs from penultimate layers of neural networks via scatter plots, and (b)Linear Regression analysis between EmoFAN results with corresponding PCs.

examples for this phenomenon is "HAPPY", and "SURPRISE" case for "NN2", and "DISGUST", and "HAPPY" case for ResMaskNet. All these cases are maximum value regions of either PC1 or PC2. This makes sense from a cognitive perspective because faces which exhibit high valence/arousal is easier to recognize. [Dan says:relevant citation necessary]

(a) Projections from Neural Network "NN2".

(b) Projections from ResMaskNet

(c) Projections from EmoFAN

(d) Valence-Arousal projections from output layer of EmoFAN

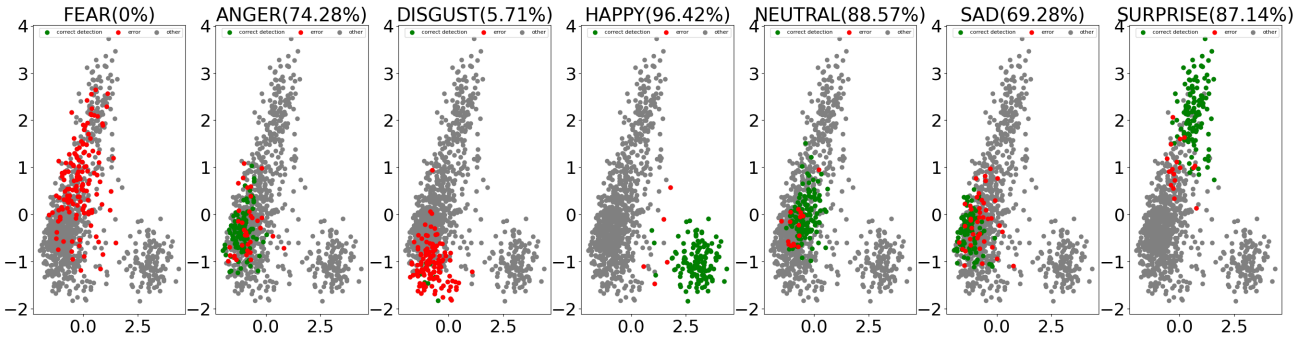
Figure 6. 21 equally spaced face Projections on PC1 and PC2 axes from (a)"NN2", (b)ResMaskNet, (c)EmoFAN, and (d)Valence/Arousal scores from EmoFAN network. The first and second rows corresponds to PC2 and PC1 respectively in (a), (b) and (c). For (d), the first row represents face projections on valence and second row on arousal axis.

5 Discussion

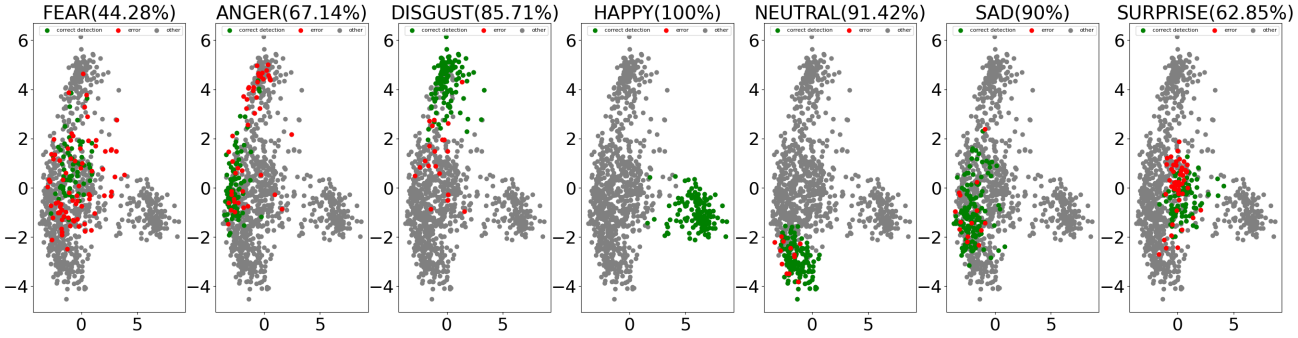
We support the claim that the DCNNs trained to detect emotions category-wise do it in a 3-step process.[Ariel says:I think that this point should be explained. We show that information can be decoded from the neural net and we interpret it as a process the layer/NN is implementing. I agree with this interpretation, but I think it sould be justified. Reviewers not necessarly know that each layer can be thought of a process.] The first process is the detection of identities. As previously mentioned, CNNs capture local features (edges at different angles, curved features) in the initial convnet layers. However, these local features may not be semantically meaningful, like facial expression-specific ones (smiling lips, wide open eyes, or action units related, etc.). In the process of capturing local to semantic features, identity information gets captured in convnet layers. A surprising result about identity detection in convnet layers is that it is an "innate" capability of untrained deep neural networks, which is verified by our results and supported by Seungdae Baek et al. study (Baek et al., 2021). This innate capability of detecting facial identities emerges solely via statistical variation in the feedforward projections in DCNNs. We also showed that identities emerge as the dominant components even in unsupervised algorithms like PCA or UMAP when applied to cropped and normalized facial images. The semantically meaningful local features emerge at a later stage (expression-specific local features), and those "signals" get boosted due to the training process with appropriate emotion labels.

[Ariel says:We need to have a paragraph that start with "the third process"]

The second process in a DCNN trained for category-wise expressions detects dimensional information. We see that by applying PCA on intermediate layer outputs and visually projecting the primary and secondary principal components(PC1 and PC2) to individual faces. This observation is consistent with other neural networks, which we inspected (Pham et al., 2021; Toisoul, Kossaifi, Bulat, Tzimiropoulos, & Pantic, 2021). It is also worth mentioning that EmoFAN (Toisoul et al., 2021) network is trained to explicitly output Valence and Arousal information along with categorical emotion scores. While inspecting the fully connected layer of EmoFAN ("the intermediate layer") prior to the output layer, we saw that PC1 and PC2 from that



(a) "NN2". The overall accuracy is 60.20%. Expression-wise accuracy shown in bracket.



(b) ResMaskNet. This network has a superior accuracy of 77.34%

Figure 7. PC1 versus PC2 scatter plots from penultimate layer outputs of (a)"NN2", and (b)ResMaskNet. These plots are the same as in the first 2 plots of Figure 6 (a). We have highlighted expression-wise classification with respect to ground truth labels of 980 frontal-face KDEF images. The comparison is between the respective network output labels and the ground truth. Green dots represent correct classification and red dots represents errors. The Gray dots represents data corresponding to expressions other than what is highlighted. It is clear that Resmasknet with more green dots is far superior than the lightweight network "NN2". The Gray dots represents data corresponding to expressions other than what is highlighted.

layer respectively captured Valence and Arousal already, which strengthens our argument that DCNNs capture dimensional information first. The fully connected layer in the EmoFAN network has already captured dimensional information before the final processing towards the output layer. We see that the Valence and Arousal information in the output is a redundant reflection that requires no further processing other than scaling. The final process is the clustering of dimensional information into categorical expressions. The emergence of the Valence-Arousal dimensional representation (in the form of the first two principal components) from the penultimate layer outputs prior to categorical class emotion outputs indicates the possibility of it acting as the "*basis components*" components of categorical representation. We also showed that the sequential order of dimensional representation and categorical representation in intermediate layers is preserved in different neural network architectures like EmoFAN (Toisoul et al., 2021), and ResMaskNet (Pham et al., 2021). Furthermore, our inspection of final layer results exhibit a fuzzy continuum expression landscape than very discretely segregated clusters. This is in line with the findings of some of the early studies (Dailey, Cottrell, Padgett, & Adolphs, 2002) using Neural Networks to detect expressions from facial images. Dailey et al (Dailey et al., 2002) also argued that the psychological phenomena behind facial expression perception can be explained as a result of how the brain processes the specific task. Their neural network model supports both categorical perception and graded dimensional perception arguments. With evidence from our computational experiments we argue that this is indeed the case. We can see discrete categorical and graded valence/arousal representation of expressions as "duals" in the sense that valence and arousal signals act as basis components for categorical perception data. The computational results from our study agree with Rachael E Jack et al.'s (Jack et al., 2014; Liu et al., 2022) suggestion that the facial signals with dimensional information can describe specific emotion categories, but not vice versa. Although DCNNs cannot temporally decompose expressions into dimensional or categorical components, a recent report (Cichy et al., 2016) compared DCNNs with human visual processing in the brain and found that the intermediate layers in DCNNs and stages in visual processing decompose information in a semantically correlated fashion. The Cichy et al. study (Cichy et al., 2016) lets us suggest that the dimensional components may be the representations that emerge first in the visual processing pathway of the brain before

categorical ones as we observe sequentially in DCNNs.

References

Adolphs, R., Tranel, D., Damasio, H., & Damasio, A. (1994). Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature*, *372*(6507), 669–672.

Baek, S., Song, M., Jang, J., Kim, G., & Paik, S.-B. (2021). Face detection in untrained deep neural networks. *Nature communications*, *12*(1), 1–15.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449–12460.

Barrett, L. F. (1998). Discrete emotions or dimensions? the role of valence focus and arousal focus. *Cognition & Emotion*, *12*(4), 579–599.

Bradley, M. M., & Lang, P. J. (2007). Emotion and motivation.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British journal of psychology*, *77*(3), 305–327.

Bruyer, R., Laterre, C., Seron, X., Feyereisen, P., Strypstein, E., Pierrard, E., & Rectem, D. (1983). A case of prosopagnosia with some preserved covert remembrance of familiar faces. *Brain and cognition*, *2*(3), 257–284.

Calder, A. J., Burton, A. M., Miller, P., Young, A. W., & Akamatsu, S. (2001). A principal component analysis of facial expressions. *Vision research*, *41*(9), 1179–1208.

Calder, A. J., & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience*, *6*(8), 641–651.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, *6*(1), 1–13.

Dailey, M. N., Cottrell, G. W., Padgett, C., & Adolphs, R. (2002). Empath: A neural network that categorizes facial expressions. *Journal of cognitive neuroscience*, *14*(8), 1158–1173.

Ekman, P. (1994). Strong evidence for universals in facial expressions: A reply to russell's mistaken critique.

Ekman, P. (2004). Emotions revealed. *Bmj*, *328*(Suppl S5).

Ekman, P., & Cordaro, D. (2011). What is meant by calling emotions basic. *Emotion review*, *3*(4), 364–370.

Ekman, P., Sorenson, E. R., & Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. *Science*, *164*(3875), 86–88.

Etcoff, N. L. (1984). Selective attention to facial identity and facial emotion. *Neuropsychologia*, *22*(3), 281–295.

Ganel, T., & Goshen-Gottstein, Y. (2002). Perceptual integrality of sex and identity of faces: Further evidence for the single-route hypothesis. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(4), 854.

Ganel, T., & Goshen-Gottstein, Y. (2004). Effects of familiarity on the perceptual integrality of the identity and expression of faces: The parallel-route hypothesis revisited. *Journal of Experimental psychology: Human perception and Performance*, *30*(3), 583.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., … Cohen, A., et al. (2022). Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, *25*(3), 369–380.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., … Lee, D.-H., et al. (2013). Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing* (pp. 117–124). Springer.

Grossman, S., Gaziv, G., Yeagle, E. M., Harel, M., Mégevand, P., Groppe, D. M., … Mehta, A. D., et al. (2019). Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nature communications*, *10*(1), 4934.

Grossman, S., Malach, R. et al. (2023). Brain-machine convergent evolution: A window into the functional role of neuronal selectivity.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in cognitive sciences*, *4*(6), 223–233.

Jack, R. E., Garrod, O. G., & Schyns, P. G. (2014). Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current biology*, *24*(2), 187–192.

Keltner, D., & Cordaro, D. T. (2015). Understanding multimodal emotional expressions. *The science of facial expression*.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, *10*(11), e1003915.

Khorrami, P., Paine, T., & Huang, T. (2015). Do deep neural networks learn facial action units when doing expression recognition? In *Proceedings of the ieee international conference on computer vision workshops* (pp. 19–27).

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 4.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90.

Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, *13*(3), 1195–1215.

Liu, M., Duan, Y., Ince, R. A., Chen, C., Garrod, O. G., Schyns, P. G., & Jack, R. E. (2022). Facial expressions elicit multiplexed perceptions of emotion categories and dimensions. *Current Biology*, *32*(1), 200–209.

Lundqvist, D., Flykt, A., & Öhman, A. (1998). Karolinska directed emotional faces. *Cognition and Emotion*.

Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P., & Cohn, J. F. (2013). Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, *4*(2), 151–160.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Mehmood, R. M., & Lee, H. J. (2015). Exploration of prominent frequency wave in eeg signals from brain sensors network. *International Journal of Distributed Sensor Networks*, *11*(11), 386057.

Mehu, M., & Scherer, K. R. (2015). Emotion categories and dimensions in the facial communication of affect: An integrated approach. *Emotion*, *15*(6), 798.

Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2013). The neural dynamics of updating person impressions. *Social cognitive and affective neuroscience*, *8*(6), 623–631.

Millet, J., Caucheteux, C., Orhan, P., Boubenec, Y., Gramfort, A., Dunbar, E., … King, J.-R. (2022). Toward a realistic model of speech processing in the brain with self-supervised learning. *arXiv preprint arXiv:2206.01685*.

Pham, L., Vu, T. H., & Tran, T. A. (2021). Facial expression recognition using residual masking network. In *2020 25th international conference on pattern recognition (icpr)* (pp. 4513–4519). IEEE.

Schweinberger, S. R., & Soukup, G. R. (1998). Asymmetric relationships among perceptions of facial identity, emotion, and facial speech. *Journal of Experimental Psychology: Human perception and performance*, *24*(6), 1748.

Schyns, P. G., Snoek, L., & Daube, C. (2022). Degrees of algorithmic equivalence between the brain and its dnn models. *Trends in Cognitive Sciences*.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the ieee international conference on computer vision* (pp. 618–626).

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2013). Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*.

Sharma, G. (2020). Facial emotion recognition. Retrieved September 17, 2022, from https://www.kaggle.com/code/gauravsharma99/facial-emotion-recognition/data

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Todorov, A., & Duchaine, B. (2008). Reading trustworthiness in faces without recognizing faces. *Cognitive Neuropsychology*, *25*(3), 395–410.

Toisoul, A., Kossaifi, J., Bulat, A., Tzimiropoulos, G., & Pantic, M. (2021). Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, *3*(1), 42–50.

Viinikainen, M., Jääskeläinen, I. P., Alexandrov, Y., Balk, M. H., Autti, T., & Sams, M. (2010). Nonlinear relationship between emotional valence and brain activity: Evidence of separate negative and positive valence dimensions. *Human brain mapping*, *31*(7), 1030–1040.

Vogelsang, L., Gilad-Gutnick, S., Ehrenberg, E., Yonas, A., Diamond, S., Held, R., & Sinha, P. (2018). Potential downside of high initial visual acuity. *Proceedings of the National Academy of Sciences*, *115*(44), 11333–11338.

Wagner, H. L., MacDonald, C. J., & Manstead, A. (1986). Communication of individual emotions by spontaneous facial expressions. *Journal of personality and social psychology*, *50*(4), 737.

Wang, Y., Fu, X., Johnston, R. A., & Yan, Z. (2013). Discriminability effect on garner interference: Evidence from recognition of facial identity and expression. *Frontiers in Psychology*, *4*, 943.

Woodworth, R. S., & Schlosberg, H. (1954). Experimental psychology, rev.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, *19*(3), 356–365.

Young, A. W., Newcombe, F., Haan, E. H. d., Small, M., & Hay, D. C. (1993). Face perception after brain injury: Selective impairments affecting identity and expression. *Brain*, *116*(4), 941–959.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818–833). Springer.

Zhang, W., Zhang, X., & Tang, Y. (2023). Facial expression recognition based on improved residual network. *IET Image Processing*.