Deep Affect:
Using objects, scenes and facial expressions in a deep neural network to predict arousal and valence values of images

George Parry & Quoc C. Vuong

Biosciences Institute & School of Psychology
Newcastle University, UK

Corresponding author: quoc.vuong@newcastle.ac.uk

Parry, G., & Vuong, Q. (2021, June 8). Deep Affect: Using objects, scenes and facial expressions in a deep neural network to predict arousal and valence values of images. https://doi.org/10.31234/osf.io/t9p3f

Abstract

Images are extremely effective at eliciting emotional responses in observers and have been frequently used to investigate the neural correlates of emotion. However, the image features producing this emotional response remain unclear. This study sought to use biologically inspired computational models of the brain to test the hypothesis that these emotional responses can be attributed to the estimation of arousal and valence of objects, scenes and facial expressions in the images. Convolutional neural networks were used to extract all, or various combinations, of high-level image features related to objects, scenes and facial expressions. Subsequent deep feedforward neural networks predicted the images' arousal and valence value. The model was provided with thousands of pre-annotated images to learn the relationship between the high-level features and the images arousal and valence values. The relationship between arousal and valence was assessed by comparing models that either learnt the constructs separately or together. The results confirmed the effectiveness of using the features to predict human emotion alongside their ability to augment each other. When utilising the object, scene and facial expression information together, the model classified arousal and valence to accuracies of 88% and 87% respectively. The effectiveness of our deep neural network of emotion perception strongly suggests that these same high-level features play a critical component in producing humans' emotional response. Moreover, performance increased across all models when arousal and valence were learnt together, suggesting a dependent relationship between these affective dimensions. These results open up numerous avenues for future work, whilst also bridging the gap between affective Neuroscience and Computer Vision.

1. Introduction

Images can portray a limitless array of scenes, objects and people, with each evoking a different emotional response[1]. The images in Figure 1, for example, can elicit anger, happiness or sadness. Emotion is a complicated, multifaceted state of feeling which induces both psychological and physiological changes within the observer[2]. The emotions elicited by images such as those in Figure 1 can be related to how observers' process the affective content of those images[3, 4]. The content can be characterized as continuous variations along an arousal and valence dimension[96]. Neuroimaging studies have provided insights into the cortical networks responsible for affective processing[13-16], but currently offer limited explanations of the underlying high-level image features that give rise to the perception of emotion. There is accumulating evidence that affective processing of images are predominantly driven by objects in the scene, the scene context and the presence of facial expressions[37-45]. The present study used state-of-the-art computational models of the brain to investigate the relative contributions of these high-level features to affective processing. Building on the work of Kim et al.[18], we focused on investigating whether such models can predict human ratings of arousal and valence when presented with natural images.

Figure 1. Images that elicit different affective responses.

1.1 Affective Processing

Studies of emotion typically adopt either Categorical Emotional States[5] (CES) or Dimensional Emotion Space (DES) models. CES models propose the existence of discrete and universal emotions, with clear boundaries separating each emotional state[6]. Typically, Ekman's Six[7] or Mikel's Eight[8] emotions are adopted as distinct emotion categories. Here we focused on DES models which propose

that emotions can be characterized by two fundamental dimensions of affective space wherein emotions are continuously represented as a linear combination of arousal and valence values[9]. Arousal represents the level of excitement, ranging from calm to excited, whereas valence ranges from unpleasant (negative) to pleasant (positive)[6].

Although researchers generally agree that independent brain regions process different discrete emotions[10-13], the literature fails to converge to similar neurological structures for arousal and valence. For example, although Kuniecki[15] and Viinikainen[16] used the same affective database, Viinikainen[16] reported that responses along the valence dimension varied with frontomedial and thalamic activity, whereas Kuniecki[15] found valence to be linked to activity in the anterior cingulate cortex and the right substantia innominata. Other studies suggest that there may be a non-linear relationship between arousal and valence[19-21, 96]. But these non-linear relationships may be related to the activation level of independent appetitive and defensive motivational systems[96].

1.2 Hierarchical Visual Pathways and Affective Processing

Although the manner in which affective images are processed remains unclear, there is a growing body of literature to suggest that the arousal and valence elicited by an image is related to the primary object, the context (i.e., scene background) and, if present, facial expressions[37-45]. By adulthood, observers are able quickly and accurately recognize objects, faces and bodies and scenes when presented with static images. Object recognition in the human cortex is mediated by the ventral visual pathway in which object features are processed in a hierarchical fashion[30]. The early stages of the ventral pathway (V1) process low-level features within visual stimuli such as edges, colour and motion[31]. These low-level features are detected by the amalgamation of the outputs of many 'simple cells' by the 'complex cells' within V1, acting as spatiotemporal filters[32]. The output from V1 is relayed into V2 and V4 which encode mid-level features such as contours, textures and individual parts of the stimuli[33]. Finally, neurons in inferior temporal cortex respond selectively to high-level features which are characterised by global and overall features such as objects[34, 35] and faces[36].

There are several reasons why the primary foreground object in a scene can contribute to the arousal and valence of an image. Indeed, many affective databases are separated into object categories (e.g., people, animal, etc.). First, viewers of emotional stimuli are drawn to the primary focal object of the image first, such as a boat or house, and spend considerable time examining these objects[37]. Additionally, when participants are asked to describe the contents of a scene, one of the first attributes of the image to be recalled is whether an object is present and if so, what that object is[38]. Interestingly, recent work has also shown a direct link between the primary object and the photo's emotional weighting. Kim et al.[18] investigated the link between the emotional value of an object in a photo and the photo's overall emotional score. The results showed a significant positive correlation between the arousal and valence scores of the primary object of an image and the overall image score. Taken

together, these studies not only indicate that the primary object is an important focal point of an image but also show the object can heavily influence the overall emotion value of the image.

The composition and semantics of the background may also contribute to the arousal and valence of images. For example, the valence score of emotional scenes can be successfully identified during very short display periods and even when outside the focus of foveal attention[39-40]. The heightened processing speed of emotionally charged scenes emphasises its importance when evaluating emotional images. Moreover, the importance of scene analysis in producing an emotional response is evident on both a neurological and a physiological level. Previous work reported that emotionally weighted images induce both cortical changes and physiological responses even as the picture quality is gradually degraded. Importantly, however, both these responses diminish once the viewer can no longer identify the background content of the image[41-42]. Together, these findings strongly imply that scene classification is a necessary prerequisite to affect processing. Additionally, it should be considered that objects in the scene interact with each other. For example in Figure 2, both scenes has a person and bouquet of flowers but would lead to different arousal and valence ratings.

Figure 2. An example of how background semantics can alter the meaning of an object. For example, the bouquet of flower, scene context and facial expression in each image interact to elicit different arousal and valence values.

A final high-level feature to consider is how the presence of a human figure, particularly the face, may influence the perceived emotion (see Figure 2). Eye-tracking studies have shown gaze to primarily follow the background of an image[43]. However, when the stimulus contained a human figure, viewers were instead drawn to examine the person, and most notably their face[43]. Not only are

viewers drawn to faces, but the existence of extensive cortical networks that preferentially respond to faces above all other stimuli further highlights their importance to humans[44-45].

1.3 Deep Neural Networks of Affective Processing

Deep neural networks (DNN) emulate the hierarchical layers in the ventral pathway and, importantly, can learn to accurately categorise object, scene and facial expressions in images that may be used during affective processing. Recently Kim et al.[18] proposed an emotion prediction model which could predict human ratings of arousal and valence to a high degree of accuracy. For their architecture, they used three separate convolutional neural networks (CNNs) trained for object categorization, scene segmentation and low-level feature extraction (e.g., colour and GIST descriptors). The outputs from these CNNs for a target image (i.e., the separate feature vectors) were concatenated and used as the input feature vector for a feed forward neural network which had three hidden layers and a single-value output layer (i.e., the predicted arousal or valence value). Separate deep networks were then trained to predict arousal or valence ratings after training on images rated by human observers. Similar architectures have been proposed specifically for faces and facial expressions[79, 98].

However, the architecture used in Kim et al.[18] was not designed to fully consider the behavioural and neuroimaging data. First, Kim et al. produced two independent networks optimised to predict either arousal or valence *separately*. The researchers based this decision on the presumed independence of these two dimensions. However, as reviewed above, there is still some debate whether this is the case in human observers. Moreover, the same human observer typically rate both the arousal and valence of a given image, which can mutually influence the rating of both dimensions. Second, Kim and colleague's model did not consider facial expressions which, for the aforementioned reasons, are likely to have an impact on the human perception of the affective content of images. Finally, it is well documented that humans struggle to use low-level features, such as colour and texture, to accurately identify the emotion of an image[5, 71, 72]. Thus in the present study, we modified the architecture proposed in Kim et al.[18] in two ways to better mimic human emotion perception. First, we substituted the CNN for low-level feature extraction with a CNN for facial expression recognition. Second, we further trained the model to *simultaneously* predict both the arousal and valence of a given image. We hypothesise that the modified emotion prediction model would perform best when feature vectors associated with the primary object, scene and facial expression are used as the input to the model, and that performance would be better when the model is trained to predict both dimensions simultaneously rather than separately.

2. Method

2.1 Image Dataset

The study was conducted on five publicly available affective databases: The International Affective Picture System[73] (IAPS), The Nencki Affective Picture System[74] (NAPS), The Open Affective

Standardised Image Set[4] (OASIS), The Socio-Moral Image Database[75] (SMID) and The Image Emotion Dataset[18] (IED). Each dataset had pre-assigned arousal and valence values which were collected from a series of independent viewers, shown in Table 1. To allow the databases to be used in parallel, all values were normalised to be within the range of 1 to 9. After removing incompatible images, such as those with ".gif" or ".tiff" extensions, the final database consisted of 16,915 images.

Table 1. General information on the datasets used in this study.

	#Photos	Likert scale	Arousal descriptors used	Valence descriptors used	Country of origin	Ranked by	Number of rankers	Male / Female split
IAPS	956	1-9	Calm – Excited	Unpleasant – Pleasant	USA	Psychology Undergraduates	100	50/50
NAPS	1356	1-9	Relaxed – Aroused	Negative – Positive	Poland	College Students	204	41/59
OASIS	900	1-7	Low - High	Negative – Positive	USA	AMT*	822	49/51
SMID	2941	1-5	Calming – Exciting	Unpleasant – Pleasant	Australia	Psychology Undergraduates & AMT*	2716	35/65
IED	10,766	1-9	Calm - Exciting	Negative – Positive	South Korea	AMT*	1339	N/A

*Amazon Mechanical Turk[76]

2.2 Feature extraction using CNNs

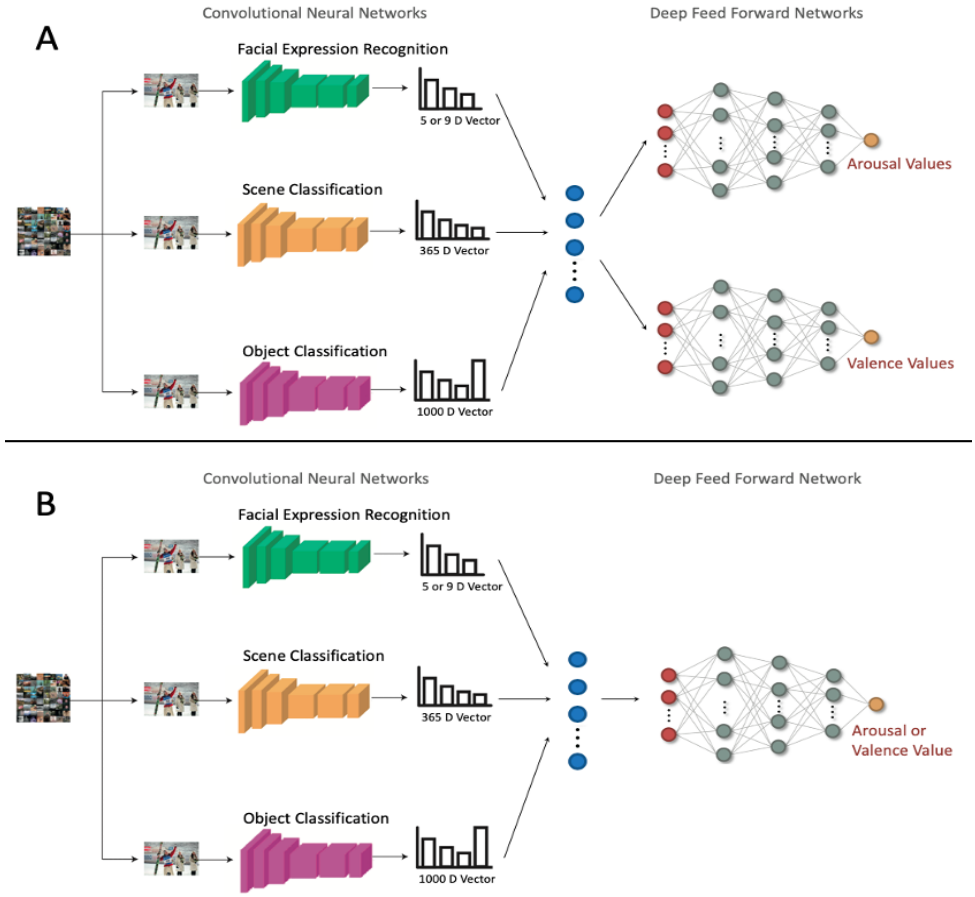
We used CNNs to extract high-level features as these architectures can perform comparably to humans on classification tasks[55-64], and using an analogous hierarchy[55, 65-70]. Specifically, we used a pre-trained VGG16 model to extract high-level features related to objects in training and test images[58, 78], and a pre-trained Places-365 model to extract features related to the scene context[62]. The output of these networks was either a 1000- or 365-D feature vector, respectively. Lastly, we trained two separate CNNs to extract high-level features for facial expression recognition[58, 79] (FER). One CNN classified static face images into five expressions (happiness, sadness, fear, anger and face absent; accuracy = 72%) and the other CNN classified face images into nine expressions (happiness, sadness, fear, anger, surprise, disgust, contempt, neutral and face absent; accuracy = 54%). We refer to these CNNs as the object, scene, 5-FER and 9-FER CNNs, respectively. The details of these four feature extractors are provided in the supplementary materials.

2.3 Affect Prediction Model

The general architecture for the proposed Affect Prediction Model consists of CNNs for feature extraction and Deep Feed Forward Networks (DFFNs) to predict arousal and valence values. Figure

3 illustrates the two types of architecture implemented, which we refer to as the split and shared architectures. In the shared architecture (Figure 3A), a single prediction model was implemented to predict both arousal and valence so that model training were not necessarily independent for these dimensions (i.e., one integrated affect prediction model). By comparison in the split architecture (Figure 3B), a prediction model was implemented separately for arousal and valence so that model training were independent for these dimensions (i.e. two independent affect prediction models). This architecture followed the approach taken by Kim et al.[18]

Figure 3. Schematics of the **(A)** shared and **(B)** split architecture. The different image features are extracted by the Convolutional Neural Networks. The output of these CNNs was then concatenated (blue circles) and used as the input (red layer) of the Deep Feed Forward Network/s. The input then flows through the three hidden layers (grey) and into the output layer (yellow) to predict arousal and valence values. Figure adapted from Kim et al.[18]



For both the shared and split architectures, the concatenated output of the CNN feature extractors (object, scene and FER) was used as the DFFN's input to predict arousal and valence values. The DFFN contained three hidden layers with 3000, 1000 and 500 neurons, respectively. The output layer consisted of a single neuron allowing it to predict either arousal or valence value. The

layers of the DFFN were fully connected. All layers in the feedforward network use Rectified Linear Units which have an activation function described by: $f(x) = \max(0, x)$, where x is the neuron's input. The edges that connected these neurons between each adjacent layer contained weights. These weights, along with the value of each neuron, are updated via backpropagation during the training phase.

To avoid overfitting the data, all the hidden layers in the DFFN are equipped with a regularisation technique known as *dropout*. Dropout refers to how the network randomly disconnects a set number of neurons, including its incoming and outgoing connections, from the network[80]. The probability that each neuron is retained is set at a fixed probability of p independent of other neurons. Each of the hidden layers was equipped with dropout with $p = 0.5$. Note that dropout is only applied during the training phase; during test phases the network retained all neurons.

During training, the loss function was set as Mean Squared Error (MSE) and the batch size set at 25. The weights of the network were updated via backpropagating the gradients through all the layers. By comparison to Kim et al.[18], the weights in the CNN layers were also updated during training. The network can then optimise the weights of the network by minimising the total cost of the loss function. The initial learning rate was set to 0.0004 and trained using the Stochastic Gradient Descent (SGD) optimisation method with a momentum of 0.9. Training under these conditions resumed until the MSE no longer diminished over four epochs at which point the learning rate was reduced by a factor of 0.1 to a minimum learning rate of 0.000006. If the error did not decrease over six consecutive epochs, training was terminated, and the model with the lowest total loss function was accepted. All experiments were implemented in the open-source deep learning framework Keras[81] using the Tensorflow[82] backend.

2.4 Model Evaluation

Various models were produced to evaluate our hypotheses. The first four models extracted the object, scene and 5- or 9-FER feature vectors giving rise to 1370 or 1374 features as input into the DFFN. Two of these models were trained under the shared architecture to learn both arousal and valence together (one for each FER CNN). Conversely, the second two models utilised the split architecture (again, one for each FER CNN), learning the two variables independently from one another. Subsequent to this, the highest performing model was further evaluated against models that extracted individual features (objects, scenes or facial expressions) and against combinations of features (objects and scenes).

Each model was subjected to 5-fold cross-validation to evaluate performance[83]. There were 15,560 images, divided into five groups, with 3,112 images in each group. The model was then trained five times using four of the groups for training data (n=12,448) and one for testing (n=3112). Each group was used as the test group once and all tests were run using the same model architecture and

data structure described above. Each model used the same five folds for training and testing. The NAPS dataset (n=1356) were used as an additional test set.

3. Results

Throughout this analysis, the values presented in each table represent the Mean Squared Error (MSE) between the ground truth (human average rating) and predicted rating. The MSE for training, testing (including the NAPS dataset) was averaged across the five folds. Finally, due to the exploratory nature of this study, uncorrected *p*-values are reported throughout.

3.1 Model Performance with Objects, Scenes and Facial Expressions

This section presents the affect prediction models' performance when utilising the object, scene and FER feature extractors. The primary aim of this section is to compare the shared and split models' performance and compare the outcomes of using a 5- or 9-FER feature extractor. The results are shown in Tables 2 and 3. Model performance for each fold is presented in supplementary materials.

Table 2. Means (M) and Standard Deviations (SD) of the 5-fold validation experiment on the shared and split models when using the 5-FER feature extractor.

	Ob, Sce & 5-FER Shared Model				Ob, Sce & 5-FER Split Model			
	Arousal		Valence		Arousal		Valence	
	M	SD	M	SD	M	SD	M	SD
Train	1.72	.17	1.12	.61	2.05	.12	1.74	.73
Test	1.89	.02	2.16	.01	2.13	.03	2.46	.13
NAPS	1.27	.07	2.23	.08	1.52	.06	2.36	.18

Table 2. Means (M) and Standard Deviations (SD) of the 5-fold validation experiment on the shared and split models when using the 9-FER feature extractor.

	Ob, Sce & 9-FER Shared Model				Ob, Sce & 9-FER Split Model			
	Arousal		Valence		Arousal		Valence	
	M	SD	M	SD	M	SD	M	SD
Train	1.75	.23	0.72	.5	2.08	.08	2.15	.16
Test	1.95	.003	2.26	.05	2.16	.03	2.45	.07
NAPS	1.39	.09	2.38	.06	1.54	.06	2.40	.09

To evaluate the models' performance under the shared or split architecture, a series of paired sample *t*-test was performed. The results revealed that both the shared 5- and 9-FER models significantly outperformed their split model counterparts (Table 4, rows 1-4). Next, it was assessed

how the overall model performance was affected by utilising a FER feature extractor that predicted fewer emotions at a higher accuracy (5-FER) or predicted more emotions at a lower accuracy (9-FER). The results, shown in Table 4, rows 5 and 6, suggest that the affect prediction model that utilised a 5-FER feature extractor led to a significantly increased model performance on both arousal and valence values compared to a 9-FER feature extractor. However, this effect was only present when using the shared architecture (Table 4, rows 7 and 8).

Table 4. *t*-test results comparing different affect prediction models where (A/V) represents whether arousal or valence is being compared and (M1) and (M2) represent the means of the first and second model respectively.

	Models Compared	A/V	M1	M2	*t*-value	*p*-value
1	Ob, Sce & 9-FER Shared Model vs Ob, Sce & 9-FER Split Model	A	1.95	2.15	13.41	.0001
2	Ob, Sce & 9-FER Shared Model vs Ob, Sce & 9-FER Split Model	V	2.25	2.45	8.47	.001
3	Ob, Sce & 5-FER Shared Model vs Ob, Sce & 5-FER Split Model	A	1.89	2.13	14.89	.0001
4	Ob, Sce & 5-FER Shared Model vs Ob, Sce & 5-FER Split Model	V	2.16	2.46	4.52	.01
5	Ob, Sce & 9-FER Shared Model vs Ob, Sce & 5-FER Shared Model	A	1.95	1.89	5.80	.004
6	Ob, Sce & 9-FER Shared Model vs Ob, Sce & 5-FER Shared Model	V	2.26	2.16	3.74	.02
7	Ob, Sce & 9-FER Split Model vs Ob, Sce & 5-FER Split Model	A	2.16	2.13	1.51	.2
8	Ob, Sce & 9-FER Split Model vs Ob, Sce & 5-FER Split Model	V	2.45	2.46	.14	.9

Overall the object, scene and 5-FER shared model achieved the highest performance at predicting arousal and valence ratings. Figures 4-6 present this model's prediction of human arousal and valence ratings for example images. Arousal and valence values are arranged so that the emotion values match their location in valence arousal space. For these figures, the mean absolute percent correct is also shown for each image using the equation: $accuracy = 100 \times \left(1 - \frac{abs(predicted\ value - actual\ value)}{9}\right)$. The average accuracy for the entire dataset was 88% for arousal and 87% for valence.

Figure 4. Outputs from the object, scene and 5-FER affect prediction model. The images vary from low (top row) to mid (bottom row) arousal, and low (left column) to high (right column) valence. The values below each image show the prediction made by the model and the ground truth value rating by humans, respectively. The accuracy for A/V is also shown.

Figure 5. Outputs from the object, scene and 5-FER affect prediction model. The images have mid arousal, and vary from low (left column) to high (right column) valence. The values below each image show the prediction made by the model and the ground truth value rating by humans, respectively. The accuracy for A/V is also shown.

Figure 6. Outputs from the object, scene and 5-FER affect prediction model. The images vary from mid (top row) to high (bottom row) arousal, and low (left column) to high (right column) valence. The values below each image show the prediction made by the model and the ground truth value rating by humans, respectively. The accuracy for A/V is also shown.

3.2 Further Analysis of the Best Performing Affect Prediction Model

It was not only hypothesised that the object, the scene and the facial expressions are important features to emotion, as shown above, but it was also suggested that they would augment and interact with one another. To evaluate this, it was assessed how the object, scene and 5-FER shared model performed when it only had access to each of the features individually or a combination of object and scene information (Table 5). A series of *t*-tests were conducted to evaluate further how model performance was affected as it accessed more features. The results, presented in Table 6, strongly confirm the

notion that a more extensive knowledge of high-level features leads to an increase in model performance.

Table 5. Means (M) and Standard Deviations (SD) of models with access to differing levels of feature information.

	Individual Features						Combined Features			
	Object		Scene		5-FER		Object & Scene		Object, Scene & 5-FER	
	M	SD	M	SD	M	SD	M	SD	M	SD
Arousal	2.27	.15	2.13	.11	2.23	.13	1.92	.03	1.89	.02
Valence	2.25	.06	2.34	.09	2.46	.09	2.20	.04	2.16	.01

Table 6. *t*-test results comparing models with different feature information where (A/V) represents whether arousal or valence is being compared, and (M1) and (M2) represent the means of the first vs second model respectively.

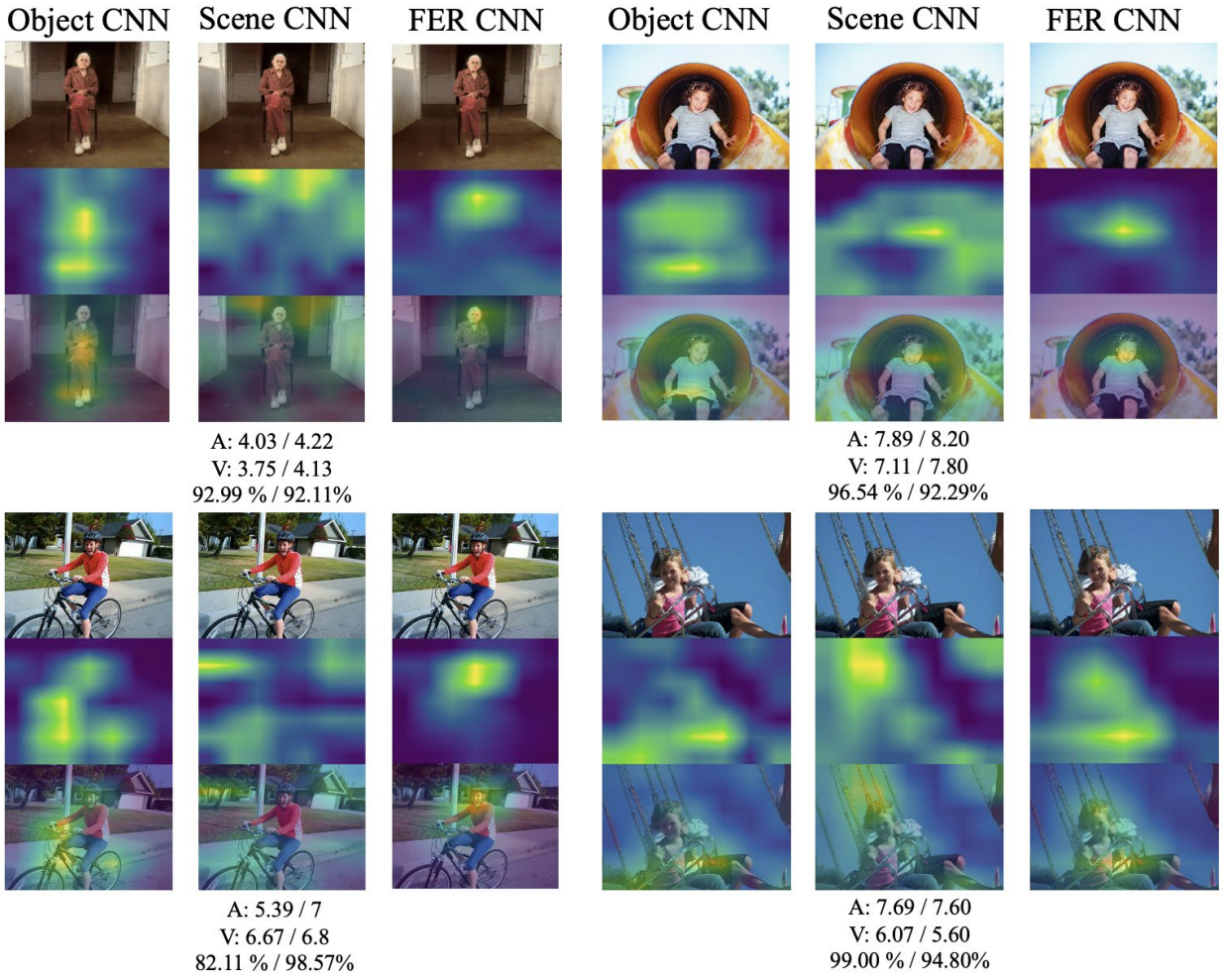
	Models Compared	A/V	M1	M2	*t*-value	*p*-value
1	Object, Scene & 5-FER vs Object Only	A	1.89	2.27	6.68	.003
2	Object, Scene & 5-FER vs Object Only	V	2.16	2.25	2.79	.05
3	Object, Scene & 5-FER vs Scene Only	A	1.89	2.13	5.13	.007
4	Object, Scene & 5-FER vs Scene Only	V	2.16	2.34	4.54	.01
5	Object, Scene & 5-FER vs 5-FER Only	A	1.89	2.23	5.72	.005
6	Object, Scene & 5-FER vs 5-FER Only	V	2.16	2.46	5.93	.004
7	Object, Scene & 5-FER vs Object & Scene	A	1.89	1.92	2.91	.04
8	Object, Scene & 5-FER vs Object & Scene	V	2.16	2.20	2.95	.04

3.3 Visualising CNN Feature Extractors Post-training

As noted above, the weights were updated via backpropagating the gradients through all the layers, including the CNN layers. Consequently, the trained CNNs may no longer extract the expected high-level features (i.e. objects, scenes or facial expressions). We therefore used Gradient-weighted Class Activation Mapping (Grad-CAM) to visualise what aspects of the image were deemed relevant by the trained CNNs. This was achieved by compartmentalising the overall model and segmenting the three individual CNNs (feature extractors) into independent models while keeping their final neuronal weights and gradients. Grad-CAM then was programmed to independently target the class-specific gradient information flowing into the final convolutional layer of each of the three feature extractors. After gathering the gradient data, a coarse localisation map of the regions considered important by

the given CNN was outputted and overlaid onto the original image[84]. The Grad-CAM activation maps can be seen in Figure 7.

Figure 3. Grad-CAM activations from each of the CNN feature extractors. A lighter green/yellow region represents an area of increased interest, whereas a darker blue region represents low interest. The values below each image represent the predicted and the ground truth emotion values, respectively. The prediction accuracies (A/V) are also shown.



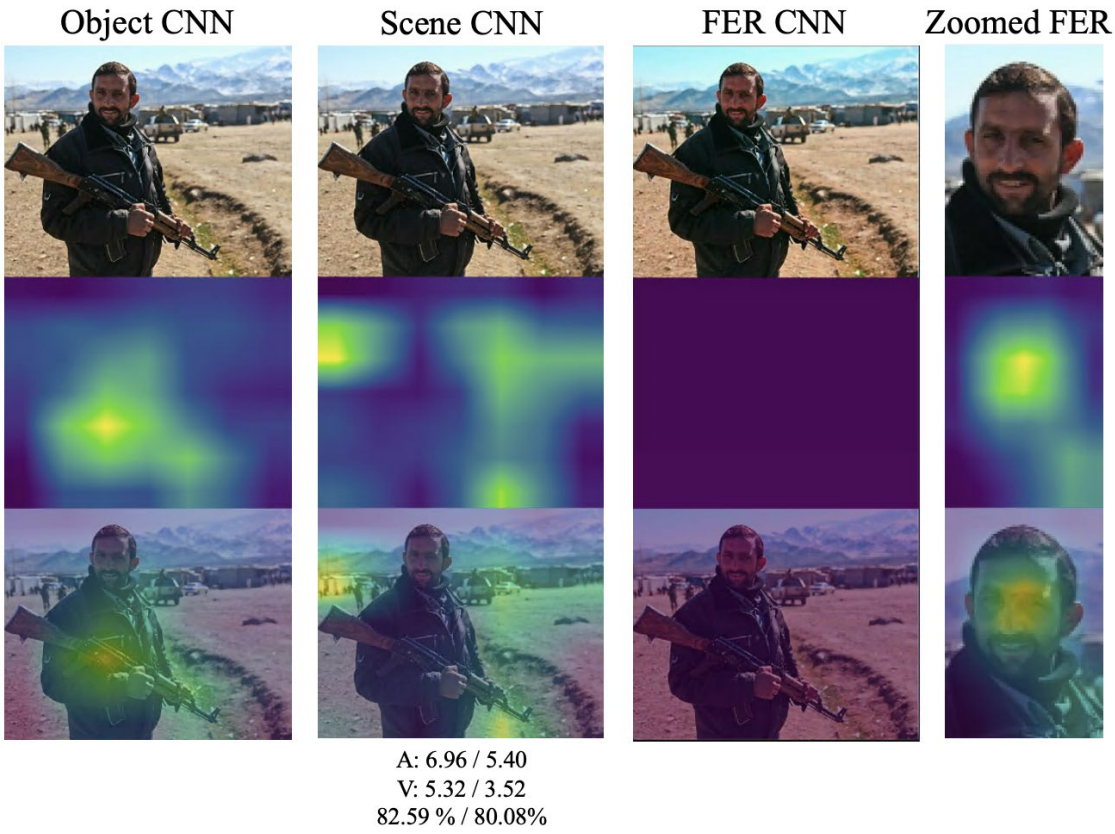
4. Discussion

The present study investigated the relative contribution of the object, the scene context and the facial expression in an image to the perception of emotions. Building a computational model that paralleled the hierarchical feature extraction in the cortex allowed the evaluation of the hypothesis that affective processing stems from the consolidation of high-level representations of objects, scenes and facial expressions. This hypothesis was confirmed. Each of the models presented here was able to use these high-level features to accurately predict the emotional response from humans to the same images. For example, when utilising the object, scene and 5-FER information, the shared model was able to, on average, predict humans' affective response to images to accuracies of 88% and 87% for arousal and valence values, respectively.

It was further hypothesised that not only would the individual features be good predictors, but it was suggested that the features would interact with each other to augment the models' prediction of human observers' affective responses. This hypothesis was also confirmed. When the object, scene and 5-FER shared model was further evaluated, it was not only evident that each of the features alone were good predictors of affect, but also displayed the features' ability to compound and interact with one another to increase model performance. These findings suggest that the estimation of these same high-level features plays a critical role in producing an affective response in the human cortex[10-16].

The trade-off between the prediction accuracy of the FER feature extractor and the number of expressions it classified on arousal and valence scores was also assessed. The results clearly showed that, under a shared architecture, the FER framework that predicted fewer facial expressions at a higher accuracy led to a significant increase in model performance. Considering the profound impact of facial expressions on emotional response[46-48], this result is somewhat unsurprising. For example, the cost of misidentifying an angry face as a fearful face would likely drastically impair the accuracy of the overall emotion prediction. Consequently, in the context of affect, the quality of prediction is clearly more valuable than a wider variety of expressions. Notably, considering the extensive literature backing the impact facial expressions should have on the affective response[46-48], a more significant increase in emotion prediction accuracy might have been expected when the FER information was added. Upon evaluating the visualisations, this effect was explained by the difficulties the network displayed when localising faces in the image (e.g., if a person is photographed from a distance). One example of this is shown in Figure 8. However, succeeding a deeper analysis of Grad-CAM, it became apparent that a large proportion of faces were missed by the FER CNN for the reason described above. To combat this, future work should first locate and enlarge the face before being passed through the CNN (see Figure 8).

Figure 8. An example of where Grad-CAM revealed an error in the functioning of a feature extractor. The Object CNN (left) and Scene CNN (middle left) activated on the correct region of the image. The FER CNN (middle right), however, failed to activate in response to facial information. Only once the face in the image was zoomed in on did the FER CNN (right) correctly activate in response to the stimulus.



This study additionally aimed to evaluate the relationship between arousal and valence. This was achieved by training the models under a shared or split architecture in which the former but not the latter had access to both affective dimensions simultaneously. Within the field of machine learning, the split architecture would be considered to be solving a simple regression problem where the model is attempting to predict a single numeric output (A or V) given an input (image). Conversely, the shared architecture is attempting to solve a multi-output regression problem where the model is predicting two or more numeric outputs (A and V) given an input (image). Typically, multi-output regression is implemented when the outputs are not independent of one another given the same input. If the two variables are at least partially dependent, then the network can learn this relationship to better predict each output. Conversely, if the outputs are completely independent of one another, learning both within one network causes competition for neuronal resources at the expense of performance. Throughout this study, various models were compared under the split and shared architectures, and in each case, the shared model significantly outperformed the split model. In line with our results, one previous study found that training neural models to estimate the emotion

category and either arousal or valence value of *faces* simultaneously improved model performance[98]. Our results, in conjunction with the theory behind multi-output models and work with faces, suggest that the emotional constructs of arousal and valence are not independent and that training on both dimensions simultaneously can improve model performance.

This study, to our knowledge, has been the first to empirically evaluate the relationship between arousal and valence within computational models of the brain and will undoubtedly play a crucial part in the ongoing debate surrounding the DES model of emotion. Numerous studies have argued arousal and valence to be independent of one another, often citing neuroimaging studies that report each construct to be processed by independent cortical regions[14, 15]. In opposition to this, the results presented here suggest there is a more convoluted relationship than previously suggested. Despite some studies supporting the notion of a non-linear relationship between the two variables[19-21], these studies are less supported by the neuroimaging literature. Given the evident relationship reported here, future work should aim to identify potential brain regions responsible for the estimation of arousal and valence values. For example, despite Kuniecki[15] mostly reporting independence, they observed valence-dependency of arousal in small cortical regions offering some evidence of where this valuation may occur.

One potential limitation of this study resides in the choice to allow gradient descent to propagate through all of the convolutional layers. This decision is in line with feedback signals reported in the visual hierarchy[97]. By allowing all of the neuronal weights in the CNN and FFN to be adjusted during training, the criticism arises that the network may no longer be extracting the suggested high-level features but has instead learnt to find new patterns within the data that better predict arousal and valence values. For example, although colour alone struggles to elicit an affective response[5, 71, 72], it has still been documented to influence emotion perception[92, 93]. This, in conjunction with a CNN's inherent ability to extract colour and other low-level features, leave a possibility that during training the network learnt to use these image features instead to predict the emotional value of the images. Nevertheless, the final model presented here was passed through Grad-CAM to allow the visualisation of activity from the final convolutional layer. The outputs from these visualisations, seen in Figure 7, provided reason to reject these criticisms by showing the CNNs to individually assess and analyse their respective high-level features. That said, future work can investigate two related lines of inquiry. First, it would be interesting to map the FNN patterns of activity to neural patterns to identify which cortical regions process the high-level features to produce affective responses. This would be similar to how neuronal activity of different layers within CNNs have been regressed onto V1[67, 91], V2[67], V4[68] and IT[70] to show similar neuronal patterns. Second, it would interesting to investigate how neural signals from these identified regions can feedback and potentially change neural activities in earlier cortical areas like V1, V2 and IT.

There can also be limitations due to the dataset used to train the network. Affective responses are well known to vary across many different psychological, individual and cultural factors[94]. By combining several data sets ranked by individuals from a wide cultural background, these discrepancies in affective opinion may lead to similar images having inconsistent arousal and valence scores. However, this variability appeared not to influence, or noticeably influence, performance. For example, had this been a significant confound, it is likely the model would have struggled to learn the emotional values at all due to each high-level feature being associated with a differing emotional score across datasets. Moreover, the final training dataset used was ranked by individuals from the USA (IAPS, OASIS), Australia (SMID) and Asia (IED). Yet the models all performed well when tested on the NAPS dataset which was ranked by Europeans. Although this result is likely to be partially due to the tighter distribution of the NAPS database, it suggests that the high-level features extracted can be robust against cultural variations in affect. An interesting avenue for future work would be to investigate further how culture impacts high-level features influence on affect.

The current study extends a deep learning emotion prediction framework to better relate the computational models to neurobiological processes in the cortex. The results provide evidence of the partially dependent relationship between the emotional construct's arousal and valence. Moreover, they reveal not only that the object, scene and facial expressions within an image are good predictors of affective response, but also revealed an interaction and compounding effect between the features. Overall the study highlights numerous avenues for further research in both the neuroscience and computing science communities.

References

1. Kurdi B, Lozano S, Banaji MR. Introducing the open affective standardised image set (OASIS). Behavior research methods. 2017 Apr 1;49(2):457-70.

2. Niedenthal PM, Ric F. Psychology of emotion. Psychology Press; 2017 Apr 20.

3. Harmon-Jones E, Gable PA, Price TF. Does negative affect always narrow and positive affect always broaden the mind? Considering the influence of motivational intensity on cognitive scope. Current Directions in Psychological Science. 2013 Aug;22(4):301-7.

4. Gerger G, Leder H, Kremer A. Context effects on emotional and aesthetic evaluations of artworks and IAPS pictures. Acta Psychologica. 2014 Sep 1;151:174-83.

5. Zhao S, Ding G, Huang Q, Chua TS, Schuller BW, Keutzer K. Affective Image Content Analysis: A Comprehensive Survey. In IJCAI 2018 Jul 13 (pp. 5534-5541).

6. Hanjalic A. Extracting moods from pictures and sounds: Towards truly personalised TV. IEEE Signal Processing Magazine. 2006 Apr 24;23(2):90-100.

7. Ekman P. An argument for basic emotions. Cognition & emotion. 1992 May 1;6(3-4):169-200.

8. Mikels JA, Fredrickson BL, Larkin GR, Lindberg CM, Maglio SJ, Reuter-Lorenz PA. Emotional category data on images from the International Affective Picture System. Behavior research methods. 2005 Nov 1;37(4):626-30.

9. Matsuda YT, Fujimura T, Katahira K, Okada M, Ueno K, Cheng K, Okanoya K. The implicit processing of categorical and dimensional strategies: an fMRI study of facial emotion perception. Frontiers in human neuroscience. 2013 Sep 26;7:551.

10. Damasio AR, Grabowski TJ, Bechara A, Damasio H, Ponto LL, Parvizi J, Hichwa RD. Subcortical and cortical brain activity during the feeling of self-generated emotions. Nature neuroscience. 2000 Oct;3(10):1049-56.

11. Paradiso S, Robinson RG, Andreasen NC, Downhill JE, Davidson RJ, Kirchner PT, Watkins GL, Boles Ponto LL, Hichwa RD. Emotional activation of limbic circuitry in elderly normal subjects in a PET study. American Journal of Psychiatry. 1997 Mar 1;154(3):384-9.

12. Schwartz GE, Davidson RJ. Neuroanatomical correlates of happiness, sadness, and disgust. The American journal of psychiatry. 1997 Jul;154(7):926-33.

13. Phan KL, Wager T, Taylor SF, Liberzon I. Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. Neuroimage. 2002 Jun 1;16(2):331-48.

14. Anders S, Lotze M, Erb M, Grodd W, Birbaumer N. Brain activity underlying emotional valence and arousal: A response-related fMRI study. Human brain mapping. 2004 Dec;23(4):200-9.

15. Kuniecki M, Wołoszyn KB, Domagalik A, Pilarczyk J. Effects of scene properties and emotional valence on brain activations: a fixation-related fMRI study. Frontiers in human neuroscience. 2017 Aug 31;11:429.

16. Viinikainen M, Jääskeläinen IP, Alexandrov Y, Balk MH, Autti T, Sams M. Nonlinear relationship between emotional valence and brain activity: evidence of separate negative and positive valence dimensions. Human brain mapping. 2010 Jul;31(7):1030-40.

17. Osgood CE. The nature and measurement of meaning. Psychological bulletin. 1952 May;49(3):197.

18. Kim HR, Kim YS, Kim SJ, Lee IK. Building emotional machines: Recognising image emotions through deep neural networks. IEEE Transactions on Multimedia. 2018 Apr 20;20(11):2980-92.

19. Kuppens P, Tuerlinckx F, Russell JA, Barrett LF. The relation between valence and arousal in subjective experience. Psychological bulletin. 2013 Jul;139(4):917.

20. Kuppens P, Tuerlinckx F, Yik M, Koval P, Coosemans J, Zeng KJ, Russell JA. The relation between valence and arousal in subjective experience varies with personality and culture. Journal of personality. 2017 Aug;85(4):530-42.

21. Feldman LA. Valence focus and arousal focus: Individual differences in the structure of affective experience. Journal of personality and social psychology. 1995 Jul;69(1):153.

22. Qian S, Zhang T, Xu C. Multi-modal multi-view topic-opinion mining for social event analysis. InProceedings of the 24th ACM international conference on Multimedia 2016 Oct 1 (pp. 2-11).

23. Balouchian P, Foroosh H. Context-Sensitive Single-Modality Image Emotion Analysis: A Unified Architecture from Dataset Construction to CNN Classification. In2018 25th IEEE International Conference on Image Processing (ICIP) 2018 Oct 7 (pp. 1932-1936). IEEE.

24. Zhao S, Yao H, Gao Y, Ji R, Xie W, Jiang X, Chua TS. Predicting personalised emotion perceptions of social images. InProceedings of the 24th ACM international conference on Multimedia 2016 Oct 1 (pp. 1385-1394).

25. Wilson TD, Gilbert DT. Affective forecasting: Knowing what to want. Current directions in psychological science. 2005 Jun;14(3):131-4.

26. Chen T, Xu R, He Y, Wang X. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. Expert Systems with Applications. 2017 Apr 15;72:221-30.

27. Lu X, Lin Z, Jin H, Yang J, Wang JZ. Rapid: Rating pictorial aesthetics using deep learning. InProceedings of the 22nd ACM international conference on Multimedia 2014 Nov 3 (pp. 457-466).

28. Wang W, He Q. A survey on emotional semantic image retrieval. In2008 15th IEEE International Conference on Image Processing 2008 Oct 12 (pp. 117-120). IEEE.

29. Zhu X, Li L, Zhang W, Rao T, Xu M, Huang Q, Xu D. Dependency exploitation: A unified CNN-RNN approach for visual emotion recognition. InProceedings of the 26th International Joint Conference on Artificial Intelligence 2017 Aug 19 (pp. 3595-3601).

30. Serre T, Oliva A, Poggio T. A feefeedforwardchitecture accounts for rapid categorisation. Proceedings of the national academy of sciences. 2007 Apr 10;104(15):6424-9.

31. Ungerleider LG, Haxby JV. 'What' and 'where' in the human brain. Current opinion in neurobiology. 1994 Jan 1;4(2):157-65.

32. Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of physiology. 1962 Jan;160(1):106.

33. Kravitz DJ, Saleem KS, Baker CI, Ungerleider LG, Mishkin M. The ventral visual pathway: an expanded neural framework for the processing of object quality. Trends in cognitive sciences. 2013 Jan 1;17(1):26-49.

34. Schwartz EL, Desimone R, Albright TD, Gross CG. Shape recognition and inferior temporal neurons. Proceedings of the National Academy of Sciences. 1983 Sep 1;80(18):5776-8.

35. Desimone R, Albright TD, Gross CG, Bruce C. Stimulus-selective properties of inferior temporal neurons in the macaque. Journal of Neuroscience. 1984 Aug 1;4(8):2051-62.

36. Baylis GC, Rolls ET, Leonard CM. Selectivity between faces in the responses of a population of neurons in the cortex in the superior temporal sulcus of the monkey. Brain research. 1985 Sep 2;342(1):91-102.

37. Kirtley C. How Images Draw the Eye: An Eye-Tracking Study of Composition. Empirical Studies of the Arts. 2018 Jan;36(1):41-70.

38. Patterson G, Hays J. Sun attribute database: Discovering, annotating, and recognising scene attributes. In2012 IEEE Conference on Computer Vision and Pattern Recognition 2012 Jun 16 (pp. 2751-2758). IEEE.

39. Nummenmaa L, Hyönä J, Calvo MG. Semantic categorisation precedes affective evaluation of visual scenes. Journal of Experimental Psychology: General. 2010 May;139(2):222.

40. Calvo MG, Nummenmaa L. Processing of unattended emotional visual scenes. Journal of Experimental Psychology: General. 2007 Aug;136(3):347.

41. Codispoti M, Mazzetti M, Bradley MM. Unmasking emotion: Exposure duration and emotional engagement. Psychophysiology. 2009 Jul;46(4):731-8.

42. De Cesarei A, Codispoti M. Scene identification and emotional response: which spatial frequencies are critical?. Journal of Neuroscience. 2011 Nov 23;31(47):17052-7.

43. Massaro D, Savazzi F, Di Dio C, Freedberg D, Gallese V, Gilli G, Marchetti A. When art moves the eyes: a behavioral and eye-tracking study. PloS one. 2012 May 18;7(5):e37285.

44. Dricu M, Frühholz S. Perceiving emotional expressions in others: activation likelihood estimation meta-analyses of explicit evaluation, passive perception and incidental perception of emotions. Neuroscience & Biobehavioral Reviews. 2016 Dec 1;71:810-28.

45. Dols JM, Russell JA, editors. The science of facial expression. Oxford University Press; 2017.

46. Bruder M, Dosmukhambetova D, Nerb J, Manstead AS. Emotional signals in nonverbal interaction: Dyadic facilitation and convergence in expressions, appraisals, and feelings. Cognition & emotion. 2012 Apr 1;26(3):480-502.

47. Peters K, Kashima Y. A multimodal theory of affect diffusion. Psychological Bulletin. 2015 Sep;141(5):966.

48. Chartrand TL, Bargh JA. The chameleon effect: the perception–behavior link and social interaction. Journal of personality and social psychology. 1999 Jun;76(6):893.

49. Prochazkova E, Kret ME. Connecting minds and sharing emotions through mimicry: A neurocognitive model of emotional contagion. Neuroscience & Biobehavioral Reviews. 2017 Sep 1;80:99-114.

50. Sato W, Fujimura T, Kochiyama T, Suzuki N. Relationships among facial mimicry, emotional experience, and emotion recognition. PloS one. 2013 Mar 25;8(3):e57889.

51. Harada T, Hayashi A, Sadato N, Iidaka T. Neural correlates of emotional contagion induced by happy and sad expressions. Journal of psychophysiology. 2016 May 3.

52. Kegel LC, Brugger P, Frühholz S, Grunwald T, Hilfiker P, Kohnen O, Loertscher ML, Mersch D, Rey A, Sollfrank T, Steiger BK. Dynamic human and avatar facial expressions elicit differential brain responses. Social cognitive and affective neuroscience. 2020 Mar;15(3):303-17.

53. Scherer KR. Unconscious Processes in Emotion: The Bulk of the Iceberg.

54. Winkielman P, Berridge KC. Unconscious emotion. Current directions in psychological science. 2004 Jun;13(3):120-3.

55. Lindsay G. Convolutional neural networks as a model of the visual system: past, present, and future. Journal of Cognitive Neuroscience. 2020 Feb 6:1-5.

56. Yamins DL, DiCarlo JJ. Using goal-driven deep learning models to understand sensory cortex. Nature neuroscience. 2016 Mar;19(3):356-65.

57. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. InAdvances in neural information processing systems 2012 (pp. 1097-1105).

58. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014 Sep 4.

59. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. InProceedings of the IEEE conference on computer vision and pattern recognition 2015 (pp. 1-9).

60. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. InProceedings of the IEEE conference on computer vision and pattern recognition 2015 (pp. 3431-3440).

61. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. InProceedings of the IEEE conference on computer vision and pattern recognition 2014 (pp. 580-587).

62. Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A. Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence. 2017 Jul 4;40(6):1452-64.

63. Qassim H, Verma A, Feinzimer D. Compressed residual-VGG16 CNN model for big data places image recognition. In2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC) 2018 Jan 8 (pp. 169-175). IEEE.

64. Verma A, Qassim H, Feinzimer D. Residual squeeze CNDS deep learning CNN model for very large scale places image recognition. In2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON) 2017 Oct 19 (pp. 463-469). IEEE.

65. Seeliger K, Fritsche M, Güçlü U, Schoenmakers S, Schoffelen JM, Bosch SE, Van Gerven MA. Cnn-based encoding and decoding of visual object recognition in space and time. BioRxiv. 2017 Jan 1:118091.

66. Bálya D, Roska B, Roska T, Werblin FS. A CNN framework for modeling parallel processing in a mammalian retina. International Journal of Circuit Theory and Applications. 2002 Mar;30(2-3):363-93.

67. Laskar MN, Giraldo LG, Schwartz O. Correspondence of deep neural networks and the brain for visual textures. arXiv preprint arXiv:1806.02888. 2018 Jun 7.

68. Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. Performance-optimised hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences. 2014 Jun 10;111(23):8619-24.

69. Seeliger K, Fritsche M, Güçlü U, Schoenmakers S, Schoffelen JM, Bosch SE, Van Gerven MA. Convolutional neural network-based encoding and decoding of visual object recognition in space and time. NeuroImage. 2018 Oct 15;180:253-66.

70. Eickenberg M, Gramfort A, Varoquaux G, Thirion B. Seeing it all: Convolutional network layers map the function of the human visual system. NeuroImage. 2017 May 15;152:184-94.

71. Rhodes LJ, Ríos M, Williams J, Quiñones G, Rao PK, Miskovic V. The role of low-level image features in the affective categorisation of rapidly presented scenes. PloS one. 2019 May 1;14(5):e0215975.

72. Zhao S, Gao Y, Jiang X, Yao H, Chua TS, Sun X. Exploring principles-of-art features for image emotion recognition. InProceedings of the 22nd ACM international conference on Multimedia 2014 Nov 3 (pp. 47-56).

73. Lang PJ, Bradley MM, Cuthbert BN. International affective picture system (IAPS): Technical manual and affective ratings. NIMH Center for the Study of Emotion and Attention. 1997;1:39-58.

74. Marchewka A, Żurawski Ł, Jednoróg K, Grabowska A. The Nencki Affective Picture System (NAPS): Introduction to a novel, standardised, wide-range, high-quality, realistic picture database. Behavior research methods. 2014 Jun 1;46(2):596-610.

75. Crone DL, Bode S, Murawski C, Laham SM. The Socio-Moral Image Database (SMID): A novel stimulus set for the study of social, moral and affective processes. PloS one. 2018 Jan 24;13(1):e0190954.

76. Buhrmester M, Kwang T, Gosling SD. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data?

77. Chen M, Zhang L, Allebach JP. Learning deep features for image emotion classification. In2015 IEEE International Conference on Image Processing (ICIP) 2015 Sep 27 (pp. 4491-4495). IEEE.

78. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In2009 IEEE conference on computer vision and pattern recognition 2009 Jun 20 (pp. 248-255). Ieee.

79. Mollahosseini A, Hasani B, Mahoor MH. Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing. 2017 Aug 21;10(1):18-31.

80. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research. 2014 Jan 1;15(1):1929-58.

81. F Cholletetal., "Keras," https://github.com/fchollet/keras,2015.

82. M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," arXiv preprint arXiv:1603.04467, 2016.

83. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localisation. InProceedings of the IEEE international conference on computer vision 2017 (pp. 618-626).

84. Anders S, Lotze M, Erb M, Grodd W, Birbaumer N. Brain activity underlying emotional valence and arousal: A response-related fMRI study. Human brain mapping. 2004 Dec;23(4):200-9.

85. Campos V, Salvador A, Giro-i-Nieto X, Jou B. Diving deep into sentiment: Understanding fine-tuned CNNs for visual sentiment prediction. In Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia 2015 Oct 30 (pp. 57-62).

86. Peng KC, Chen T, Sadovnik A, Gallagher AC. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In Proceedings of the IEEE conference on computer vision and pattern recognition 2015 (pp. 860-868).

87. You Q, Luo J, Jin H, Yang J. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In Twenty-ninth AAAI conference on artificial intelligence 2015 Feb 9.

88. You Q, Luo J, Jin H, Yang J. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In Thirtieth AAAI conference on artificial intelligence 2016 Feb 21.

89. Xu C, Cetintas S, Lee KC, Li LJ. Visual Sentiment Prediction with Deep Convolutional Neural Networks (2014). arXiv preprint arXiv:1411.5731.

90. Hanjalic A. Extracting moods from pictures and sounds: Towards truly personalised TV. IEEE Signal Processing Magazine. 2006 Apr 24;23(2):90-100.

91. Hu Y, Qiao K, Tong L, Zhang C, Gao H, Yan B. A CNN-based computational encoding model for human V1 cortex. In 2018 Tenth International Conference on Advanced Computational Intelligence (ICACI) 2018 Mar 29 (pp. 408-413). IEEE.

92. Csurka G, Skaff S, Marchesotti L, Saunders C. Learning moods and emotions from color combinations. In Proceedings of the Seventh Indian Conference on Computer Vision, Graphics and Image Processing 2010 Dec 12 (pp. 298-305).

93. He L, Qi H, Zaretzki R. Image color transfer to evoke different emotions based on color combinations. Signal, Image and Video Processing. 2015 Nov 1;9(8):1965-73.

94. Mesquita B, Frijda NH. Cultural variations in emotions: a review. Psychological bulletin. 1992 Sep;112(2):179.

95. Lewandowska-Tomaszczyk B, Wilson PA. Self-conscious emotions in collectivistic and individualistic cultures: A contrastive linguistic perspective. In Yearbook of Corpus Linguistics and Pragmatics 2014 (pp. 123-148). Springer, Cham.

96. Bradley MM, Lang PJ. The International Affective Picture System (IAPS) in the study of emotion and attention. In J. A. Coan & J. J. B. Allen (Eds.), Series in affective science. Handbook of emotion elicitation and assessment 2007 (p. 29–46). Oxford University Press.

97. Ahissar M, Hochstein S. The reverse hierarchy theory of visual perceptual learning. Trends in Cognitive Sciences. 2004 Oct;8(10):457-64.

98. Handrich S, Dinges L, Al-Hamadi A, Werner P, Al Aghbari Z. Simultaneous prediction of valence/arousal and emotions on AffectNet, Aff-Wild and AFEW-VA. Procedia Computer Science 2020 170:634-41.

Supplementary Materials

1. Details of the CNN feature extractors

To build the proposed model of human emotion perception, it was first necessary to extract the same features from the images as the cortex, including object, scene and facial expression. All features were normalised to [0,1].

1.1 Object Features

In recent years, various CNN architectures have performed comparably to humans in object classification tasks[58, 59, 77]. Here, the VGG16 CNN was used to extract the object feature vector from the image[58]. VGG16 was pre-trained on the ImageNet[78] database which contained millions of annotated images and achieved a classification accuracy of 92.7% across the 1000 object categories. VGG16 accepted a 224x224 RGB image as the input to the model. The image was then passed through thirteen convolutional layers which utilised 3x3 receptive fields and a convolutional stride of one pixel. As the image moves deeper into the VGG16 architecture, the number of feature maps produced becomes progressively larger from 64 to 512. Spatial pooling was conducted by five max-pooling layers after convolution layers two, four, seven, ten and thirteen. Each max-pooling layer used a stride of two and a 2x2 pixel window.

Following the stack of convolutional layers, the image was passed through three dense layers with 4096, 4096 and 1000 neurons respectively. All hidden layers (layers that are not the input or output of the model) use the Rectified Linear Unit (ReLU) as the non-linear activation function. The final output layer consists of a SoftMax activation layer, which produces an output of the probability that a given image contains an object from 1000 different categories.

1.2 Scene Features

We used the Places-365 CNN model which achieved an accuracy of 84.91% on the Places dataset, which contained 1.8 million images from 365 scene categories[62]. Places-365 accepted a 224x224 RGB input image. The image was then passed through thirteen convolutional layers with very small (3x3) receptive fields and convolutional strides of one or two. All convolutional layers were regularised using L2 regularisation, which applied penalties across the layer parameters. During backpropagation, these penalties were summed into the loss function to help the model converge and prevent overfitting. Spatial pooling was carried out by five max-pooling layers after convolution layers two, four, seven, ten and thirteen. Each max-pooling layer used a stride of two with a 2x2 pixel window and valid padding. Two dense layers containing 4096 neurons followed the stack of convolutional layers. ReLU activation was applied to all hidden layers. The final layer contained 365 neurons, each representing a different scene category. A SoftMax activation function was applied to the final layer to produce the probability that a photo contains a scene from one of the 365 categories.

1.3 Facial Expression Recognition

Unlike object and scene classification, Facial Expression Recognition (FER) currently does not have a publicly available pre-trained model. It was, therefore, necessary to build a FER prediction model from scratch before it could be used as a feature extractor for arousal and valence. The AffectNet Database[79], which contained 420,000 images manually annotated for the presence of eleven discrete facial emotion categories consisting of Neutral, Happiness, Sadness, Surprise, Fear, Disgust, Anger, Contempt, None, Uncertain and No-Face, was used to train the FER model. Due to computational memory constraints, the dataset was reduced to roughly 1,500 images from each category, excluding the "None" and "Uncertain" categories. The final dataset used in this study consisted of 13,968 photos ranging across nine emotions.

The FER model utilised the pre-trained convolutional base of the VGG19 model to extract the features from the images[58]. The features were then passed into a custom-built linear classifier which predicted the images into one of nine emotion categories. To allow the convolutional base to adapt which features were extracted to the new task of FER, backpropagation was allowed to continue through the convolutional layers allowing their weights to be adjusted in response to the new input data. Images from the AffectNet database were first converted into 224x224 RGB images before being passed into the model. It was essential to train the FER model to use these input dimensions to allow the final network to be employed alongside the object and the scene classifiers. The images were then passed through sixteen convolutional and five max-pooling layers. Spatial pooling was conducted after convolutional layers two, four, six, ten and fourteen. The output from the fifth max-pooling layer was then passed into the custom-built feedforward network containing four dense layers with 2048, 1024, 512 and 9 neurons respectively. All hidden layers utilised the ReLU activation function. The final output layer was a SoftMax layer to normalise the outputs to values between 0 and 1, allowing the model to predict the emotion category.

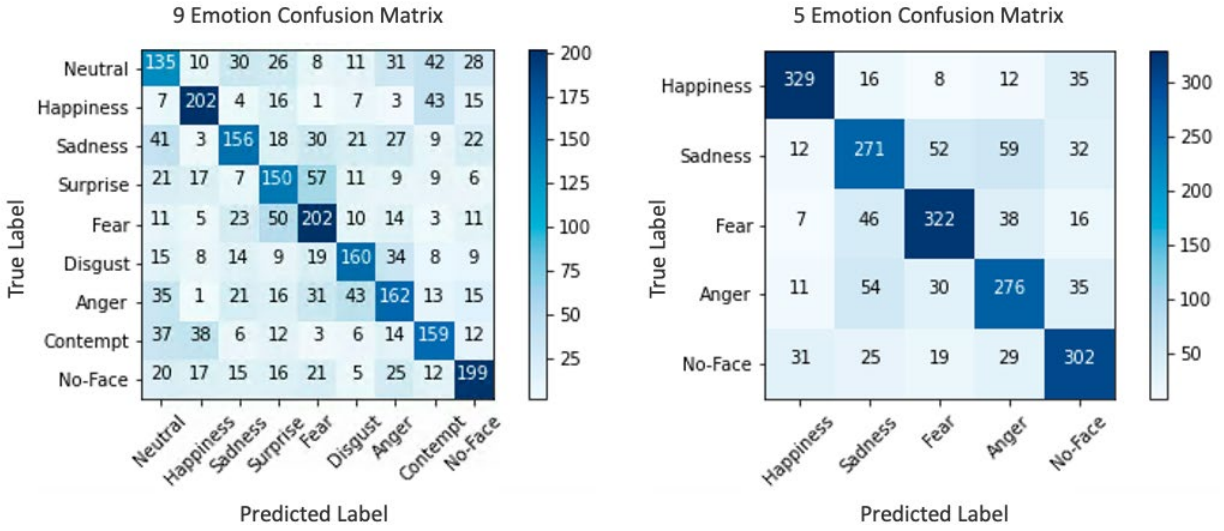
Using the architecture described above, two FER models were trained with the intention of examining the trade-off between higher prediction accuracy and the total number of expressions predicted. The first model was trained on eight emotions, as described above. The second FER model was trained on only five emotion categories consisting of Happiness, Sadness, Fear, Anger, and No-Face. Note that in the five-emotion model, the final dense layer's neurons were changed from nine to five.

Table S1 gives the average training and test accuracies when classifying the data into 9- and 5-emotions. The best network from both the 9- and 5-emotion models was taken forward to be used as the FER feature extractor used in the overall emotion prediction model. The confusion matrices for the best models are also shown in Figure S1.

Table S1. Training and test accuracies (Mean) and Standard Deviations (SD) for the 5-Fold validation experiment on the 9- and 5-FER models.

	9-FER Model		5-FER Model	
	M	SD	M	SD
Train	73.41%	5.20	88.51%	3.16
Test	53.75%	1.22	71.71%	1.35

Figure S1. Confusion matrices for the best performing 9-emotion and the 5-emotion models. The value in each box represents the number of times the model predicted each true label. For example, the 5-emotion model correctly predicted a fearful face 322 times but mistook a fearful face for an angry face 38 times.



2. Additional results

2.1 Object, Scene and FER Shared Models

Table S2. Model performance for each fold (K1-K5) for the training, test and NAPS datasets. The 5-FER CNN was used.

		K1	**K2**	**K3**	**K4**	**K5**	**Mean**
Arousal	Train	1.84	1.79	1.43	1.85	1.70	1.72
	Test	1.90	1.86	1.90	1.88	1.90	**1.89**
	NAPS	1.30	1.22	1.34	1.19	1.33	1.27
Valence	Train	0.72	2.01	0.81	1.49	0.57	1.12
	Test	2.15	2.14	2.19	2.13	2.17	**2.16**
	NAPS	2.27	2.33	2.25	2.17	2.13	2.23

Table S3. Model performance for each fold (K1-K5) for the training, test and NAPS datasets. The 9-FER CNN was used.

		K1	**K2**	**K3**	**K4**	**K5**	**Mean**
Arousal	Train	1.66	1.47	1.70	2.07	1.86	1.75
	Test	1.95	1.96	1.95	1.95	1.94	**1.95**
	NAPS	1.29	1.54	1.42	1.34	1.38	1.39
Valence	Train	0.50	0.21	0.85	1.51	0.54	0.72
	Test	2.21	2.31	2.21	2.27	2.29	**2.26**
	NAPS	2.39	2.48	2.37	2.33	2.34	2.38

2.2 Object, Scene and FER Split Models

Table S4. Model performance for each fold (K1-K5) for the training, test and NAPS datasets. The 5-FER CNN was used.

		K1	**K2**	**K3**	**K4**	**K5**	**Mean**
Arousal	Train	2.17	2.07	1.89	1.96	2.14	2.05
	Test	2.18	2.13	2.09	2.11	2.13	**2.13**
	NAPS	1.49	1.49	1.62	1.47	1.54	1.52
Valence	Train	2.41	1.06	2.06	0.85	2.31	1.74
	Test	2.36	2.68	2.39	2.49	2.37	**2.46**
	NAPS	2.28	2.24	2.18	2.55	2.55	2.36

Table S4. Model performance for each fold (K1-K5) for the training, test and NAPS datasets. The 9-FER CNN was used.

		K1	**K2**	**K3**	**K4**	**K5**	**Mean**
Arousal	Train	2.12	2.07	2.10	2.15	1.94	2.08
	Test	2.17	2.14	2.18	2.11	2.18	**2.16**
	NAPS	1.46	1.50	1.58	1.57	1.59	1.54
Valence	Train	2.20	2.11	2.11	2.36	1.91	2.15
	Test	2.39	2.46	2.38	2.55	2.47	**2.45**
	NAPS	2.43	2.37	2.30	2.35	2.54	2.40