Title: Emergence of Emotion Selectivity in A Deep Neural Network Trained to Recognize Visual Objects

Authors: Peng Liu[1], Ke Bo[1], Mingzhou Ding[2,3]*, Ruogu Fang[2,3]*

Affiliations:

[1]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA
[2]J. Crayton Pruitt Family Department of Biomedical Engineering, Herbert Wertheim College of Engineering, University of Florida, Gainesville, FL, USA
[3]Center for Cognitive Aging and Memory, McKnight Brain Institute, University of Florida, Gainesville, FL, USA

*Corresponding author: mding@bme.ufl.edu; ruogu.fang@bme.ufl.edu

Abstract:

Visual cortex plays an important role in representing the affective significance of visual input. The origin of these affect-specific visual representations is debated: they are innate to the visual system versus they arise through reentry from frontal emotion processing structures such as the amygdala. We examined this problem by combining a convolutional neural network (CNN) model of the human ventral visual cortex pre-trained on ImageNet with two datasets of affective images. Our results show that (1) in all layers of the CNN model, there were artificial neurons that responded consistently and selectively to neutral, pleasant, or unpleasant images and (2) lesioning these neurons by setting their output to 0 or enhancing these neurons by increasing their gain lead to decreased or increased emotion recognition performance respectively. These results support the idea that the visual system may have the innate ability to represent the affective significance of visual input and suggest that CNNs offer a fruitful platform for testing neuroscientific theories.

Short title: Test Emotion Selectivity in A Deep Neural Network

Teaser: A convolutional neural network (CNN) model of the human ventral visual cortex shows that artificial neurons respond selectively to emotional images, supporting the idea of an innate ability to represent affective significance of visual input.

MAIN TEXT

Introduction

Human emotions are complex and multifaceted, and under the influence of many factors, including individual differences, cultural influences, and the context in which the emotion is experienced (*1–5*). Still, a large number of people, across different cultures, levels of knowledge, and backgrounds, experience similar feelings when viewing images of varying affective content (*6–9*). What fundamental principles in the functions of the human visual system underlie such universality requires elucidation.

Previous studies of emotion perception have primarily relied on empirical cognitive experiments (*10–12*). Some of them have focused on capturing human behavioral valence or arousal judgment on affective images (*13–16*), while others have recorded brain activities to look for neural correlates

of affective stimuli processing (*17–21*). Despites decades of effort, how the brain transforms visual stimuli into subjective emotion judgments (e.g., happy, neutral, or unhappy) remains not well understood. The advent of machine learning especially artificial neural networks (ANNs) opens the possibility of attacking this problem using a modeling approach.

In machine learning, artificial neural networks can project the raw visual information to a feature space and encode the feature as the activation patterns of hidden layers for object recognition. One type of artificial neural networks, convolutional neural networks (CNNs), owing to their hierarchical organization resembling that of the visual system, are increasingly used as models of visual processing in the primate brain (*22–26*). CNNs trained to recognize visual objects can achieve performance levels rivaling or even exceeding that of humans. Interestingly, CNNs trained on images from such databases as the ImageNet (*27*) are found to demonstrate neural selectivity to stimuli that are not in the training data. For instance, (*28*) showed that neurons in a CNN trained on ImageNet became selective for numbers without having been trained on any "number" datasets.(*29*) demonstrated that face-selective neurons could arise in a CNN model trained on recognizing nonface objects. These studies raise the natural question of whether it is possible that the brain uses similar neural mechanisms to extract relevant features from the stimuli and selectively use them to guide emotion judgments.

The role of the visual cortex in visual emotion processing is debated (*30*, *31*). (*32*) argued that emotion representation is an intrinsic property of the visual cortex. They used a CNN pre-trained on ImageNet to show that the model can accurately predict the emotion categories of affective images. (*20*), on the other hand, showed that the affective representations found in the visual cortex during affective scene processing might arise as the result of reentry from anterior emotion modulating structures such as the amygdala. The goal of this study is to further examine this question using CNN models.

CNN models are well suited for addressing questions related to the human visual system. Among the many well-established CNN models, VGG-16 (*33*) has an intermediate level of complexity and is shown to have superior object recognition performance (*34*). Using VGG-16, recent cognitive neuroscience studies have explored how encoding and decoding of sensory information are hierarchically processed in the brain (*23*, *35*, *36*). (*23*) used VGG-16 to quantitatively demonstrate an explicit gradient of feature complexity encoded in the ventral visual pathway. (*35*) used VGG-16 to model the visual cortical activity of human participants reviewing images of objects and demonstrated that different layers of the model highly correlate with brain activity in different visual areas. (*36*) investigated qualitative similarities and differences between VGG-16 and other feed-forward CNNs in the representation of the visual object and showed these CNNs exhibit multiple perceptual and neural phenomena such as the Thatcher effect (*37*), Weber's law (*38*), relative size, image object selectivity, and so on.

In this study, we adopted VGG-16 pre-trained on ImageNet as the mode of the human visual system. Using two well-established affective image datasets: International Affective Picture System (IAPS) (*15*) and Nencki Affective Picture System (NAPS) (*16*), we tested whether emotion selectivity can spontaneously emerge from such systems. Emotional selectivity was established by computing neural responses to three broad classes of images: pleasant, neutral, and unpleasant (tuning curves). A neuron is considered selective for a particular emotion if it exhibits the strongest responses to images of that category from both datasets. To test whether these neurons have a functional role, we replaced the last layer of the VGG-16 with two units and trained the connections to this layer to decode pleasant versus non-pleasant, neutral vs. non-neutral, and unpleasant vs. non-unpleasant images. Two neural manipulations were carried out: lesion and feature attention enhancements.

Lesioning the selective neurons for a specific emotion is expected to degrade the network's performance in recognizing that emotion, whereas applying attention enhancement to the selective neurons is expected to increase the network's performance in recognizing that emotion.

Results

We tested whether emotion selectivity can naturally arise in a CNN model trained to recognize non-emotional natural images. VGG-16 pre-trained on ImageNet data (*27*) was used for this purpose (see Figure 1). Selectivity for pleasant, neutral, and unpleasant emotions was defined for each neuron based on its response profiles to images from two affective picture sets (IAPS and NAPS). The functional significance of these neurons was then assessed using lesion and attention enhancement methods.

Emotion selectivity of neurons in different convolutional layers of VGG-16

The tuning curve for a neuron is defined as the normalized mean response to pleasant, neutral, and unpleasant images in a given dataset plotted as a function of the emotion category. The maximum of the tuning curve indicates the neuron's preferred emotion category for that picture set. Given that the two picture sets (IAPS and NAPS) each contain its own inevitable idiosyncrasies, to guard against noise and spurious effects, a neuron is defined to be selective for a given emotion if it exhibits maximum response to the same emotion category for both datasets.

Figure 2A (top) shows the tuning curves of three neurons from the Convolutional Layer 3 (an early layer). According to the definition above, these neurons are selective for pleasant, neutral, and unpleasant categories, respectively. For the top 100 images from IAPS and NAPS that elicited the strongest response in these neurons, Figure 2A (bottom) shows the valence distribution of these images. As can be seen, for these early layer neurons, while the pleasant neuron is more activated by images with high valence ratings (pleasant), for the neutral and unpleasant neurons, the patterns are less clear. For the neurons in Convolutional Layer 6 (a middle layer), however, as shown in Figure 2B, their emotion selectivity and the category of images they prefer show greater agreement. Namely, the pleasant neuron prefers predominately images with high valence (pleasant), the neural neuron prefers predominately images with intermediate valence (neutral), and the unpleasant neuron prefers predominately images with low valence (unpleasant). The results for the three neurons from Convolutional Layer 13 (a deep layer) are similar to that from Layer 6; see Figure 2C.

Do different emotion categories become better differentiated as we go from earlier to deeper layers? We used tuning quality, which is the maximum absolute tuning value of a neuron (*39*), to address this question. As shown in Figure 2D, emotional tuning became stronger as one ascended the layers from early to deep, an effect that is especially noticeable for the IAPS datasets, supporting the notion that emotion differentiability increases as we go from earlier to deeper layers. One of the computational principles embodied in CNN models of the brain is that earlier layer neurons encode lower-level stimulus properties (e.g., colors and edges), whereas deeper layer neurons encode higher-level properties such as semantic meaning (e.g., object identities) (*40–42*). The results in Figure 2, suggesting that from earlier to deeper layers, (1) the emotion selectivity of a neuron become more consistent with the valence of its preferred images and (2) different emotion categories become better differentiated, are seen to be in general agreement with this principle and add emotion as another higher-level concept encoded in deeper layers of a CNN model.

Generalizability of emotion-selective neurons

In the foregoing, a neuron was considered selective for a given emotion if it did so consistently for both IAPS and NAPS datasets. A natural question is whether such neurons arise as the result of random chance or as an emergent property of the network that is generalizable across datasets. To address this question, we quantified the overlap between two groups of randomly selected neurons using the Jaccard Index (JI) (*43–46*) and compared it to the overlap of the two groups of neurons defined to be selective for a given emotion by IAPS and by NAPS. The value of the Jaccard index ranges from 0.0 to 1.0, with 0 indicating nonoverlap and 1 total overlap. If emotion selectivity arises randomly, then the JI of two random sets of neurons will not be different from that derived from the overlap between IAPS identified and NAPS identified neurons selecting for the same emotion. As shown in Figure 3 (left), for the VGG16 pre-trained on the ImageNet, the overlap between IAPS identified and NAPS identified emotion-selective neurons is significantly higher than that between two randomly selected neurons ($p <= 1.00e-04$ in most layers), and the effect is more prominent in deeper than earlier layers. These results suggest that emotion selectivity is unlikely the result of random chance but an emergent property of the trained network. For the randomly initialized VGG-16, shown in Figure 3 (right), the overlap between emotion-selective neurons defined by IAPS and by NAPS, respectively, and that between randomly selectively two groups of neurons is largely the same, except for a small difference for the pleasant neurons.

The functionality of emotion-selective neurons

To test whether emotion-selective neurons have a functional role, we followed (*39*) and replaced the last layer of the VGG-16, which originally contained 1,000 units for recognizing 1000 different object categories, with a fully connected layer containing two units for recognizing two types of emotions. Three models were trained and tested for each of the two datasets: Model 1: pleasant versus non-pleasant, Model 2: neutral versus non-neutral, and Model 3: unpleasant versus non-unpleasant. Once these models were shown to have adequate emotion recognition performance, two neural manipulations were considered: feature attention enhancement and lesion. For feature attention enhancement (*47–49*), the gain of the neurons selective for a given emotion is increased by increasing the slope of the ReLU activation function (see Methods) (*50–53*), whereas for lesion, the output of the neurons selective for a given emotion was set to 0, which effectively removes the contribution of these neurons, i.e., they are lesioned. We hypothesized that (1) with attention enhancement, the network's ability to recognize emotion is increased and (2) with lesion, the network's ability to recognize emotion is decreased.

Feature attention enhancement: For IAPS images, Figure 4A compares performance changes after enhancing the emotion-selective neurons as well as enhancing the same number of randomly sampled neurons. The optimal tuning strength in which we achieved the best performance enhancement was chosen for each layer in the plot. As one can see, for pleasant versus non-pleasant, neutral versus non-neutral, and unpleasant versus non-unpleasant emotions, enhancing the gain of the neurons selective for a specific emotion can significantly improve the emotion recognition performance of the CNN model for that emotion. Moreover, deeper layer gain enhancement tends to yield greater performance improvements than earlier layer gain enhancement. Increasing the gain in randomly selected neurons, however, showed marginal changes in performance, which exhibited no systematic trends across layers. The feature-attention performance of emotion-selective neurons over random neurons is highly statistically significant in the middle and deeper layers ($p< 5e-2$ for the middle layers and $p<8e-05$ for the deeper layers). Figure 4A (right) shows the performance changes across layers as the tuning strength α varied from 0 to 5. Again, it is apparent that deeper layer tuning yields more performance increase than earlier layers. Interestingly, for a given layer,

the performance change is not necessarily a monotonic function of α. When α got too large, the performance started to decline. This could be because linear tuning may not fully reflect the non-linear property of the neuron. It might be instructed to recall the Yerkes-Dodson law, which posits that performance can improve as arousal (gain) increases up to a certain point, after which further increases in arousal can lead to a decline in performance (*54*, *55*).

We carried out the same analysis for the NAPS dataset in Figure 4B. The results largely replicated that in Figure 4A for the IAPS dataset. One thing worth noting is that enhancing unpleasant neurons did not result in significant performance improvement over enhancing randomly selected neurons in some of the layers. One possibility is that the emotion recognition performance of the original model is close to the ceiling for NAPS, which leaves little room for performance improvement. When we tested the NAPS trained model on IAPS, as shown in Figure 4C, a clear performance improvement was seen in deeper layers.

Lesion analysis: The functional importance of the emotion-selective neurons can be further assessed through lesion analysis (*56–59*). As shown in Figure 5, we compared the emotion recognition performance changes by setting the output from emotion-selective neurons to 0 as well as by setting the output of an equal number of randomly chosen neurons to 0. As can be seen, lesioning the emotion-selective neurons led to significant performance declines, especially for the deeper layers; the performance decline can be as high as 85%. In contrast, lesioning randomly selected neurons produces almost no performance changes. This result supports the hypothesis that the emotive-selective neurons are important for emotion recognition, and this is more so in deeper layers than in earlier layers.

Discussion

.

It has been argued that the human visual system has the innate ability to recognize the motivational significance of environmental inputs (*60*). We examined this problem using convolutional neural networks (CNNs) as models of the human visual system (*61–66*). Selecting the VGG16 pre-trained on non-emotional images from the ImageNet as our model (*67–69*) and using two sets of affective images (IAPS and NAPS) as test stimuli, we found the existence of emotion-selective neurons in all layers of the model even though the model has never been explicitly exposed to emotional content. Additionally, emotion selectivity becomes stronger and more consistent in the deeper layers, in agreement with prior literature demonstrating that the deeper layers of CNNs encode higher-level semantic information. Applying two manipulations: feature attention enhancement and lesion, we are able to show further that the emotion-selective neurons are functionally significant, specifically: (1) after increasing the gain of emotion-selective neurons (e.g., feature attention enhancement), the network's performance in emotion recognition is enhanced relative to increasing the gain of randomly selected neurons and (2) in contrast, after lesioning the emotion-selective neurons, the network's performance in emotion recognition is degraded relative to lesioning randomly selected neurons. These performance differences are stronger and more noticeable in deeper layers than in earlier layers. Together, these findings indicate that emotion selectivity can spontaneously emerge in CNN models trained to recognize visual objects, and these emotion-selective neurons play a significant role in recognizing emotion in natural images, lending credence to the notion that the visual system's ability to represent affective information may be innate.

Affective processing in the visual cortex

The perception of opportunities and threats in complex visual scenes represents one of the main functions of the human visual system. The underlying neurophysiology is often studied by having observers view pictures varying in affective content. (*70*) reported greater functional activity in the

visual cortex when subjects viewed pleasant and unpleasant pictures than neutral images. (*71*) showed the visual cortex has differential sensitivities in response to emotional stimuli compared to the amygdala. (*72*)demonstrated that emotional significance (e.g., valence or arousal) could modulate the perceptual encoding in the visual cortex. Two competing but not mutually exclusive groups of hypotheses have been advanced to account for emotion-specific modulations of activity in the visual cortex. The so-called reentry hypothesis states that the increased visual activation evoked by affective pictures results from reentrant feedback, meaning that signals arising in subcortical emotion processing structures such as the amygdala propagate to the visual cortex to facilitate the processing of motivationally salient stimuli (*73–75*). Recent work (*20*) provides support for this view. Using multivariate pattern analysis and functional connectivity, these authors showed that (1) different emotion categories (e.g., pleasant versus neutral and unpleasant versus neutral) are decodable based on the multivoxel patterns in the visual cortex and (2) the decoding accuracy is positively associated with reentry signals from anterior emotion-modulating regions. A second group of hypotheses states that the visual cortex may itself code for emotional qualities of a stimulus, without the necessity for recurrent processing (see (*76*) for a review). Evidence supporting this hypothesis comes from empirical studies in experimental animals (*77*, *78*) as well as in human observers (*79*), in which the extensive pairing of simple sensory cues such as tilted lines or sinusoidal gratings with emotionally relevant outcomes shapes early sensory responses (*80*). Beyond simple visual cues, recent computational work using deep neural networks has also suggested that the visual cortex may intrinsically represent emotional value as contained in complex visual media such as video clips of varying affective content (*32*). Our results, by showing that emotion-selective neurons exist in all layers of the visual cortex and that these neurons play an important role in emotion recognition, appear to support the view that the visual cortex has the innate ability to code the emotional qualities of visual stimuli.

Neural selectivity in ANNs and the brain

That CNNs, or more generally ANNs, can be trained to recognize a large variety of visual objects has long been recognized. Remarkably, recent studies note that ANNs trained on recognizing visual objects can spontaneously develop selectivity for other types of input, including visual numbers and faces (*81*). The number sense is considered an inherent ability of the brain to estimate the quantity of certain items in a visual set (*82*, *83*). There is significant evidence demonstrating that the number sense exists in both humans (e.g., adults and infants) (*84–86*) and non-human primates (e.g., numerically naïve monkeys) (*87–89*). (*90*) found that number-selective units spontaneously emerged in a deep artificial neural network trained on ImageNet for object recognition. (*91*) demonstrated that number selectivity can even arise spontaneously in randomly initialized deep neural networks without any training. Both studies focused on the last convolutional layers, in which the number-selective units were found, and they also demonstrated that the emergence of number-selective units could result from the weighted sum of both decreasing and increasing the activity of some units. In addition, it is well known that face-selective neurons exist in humans (*92*) and non-human primates. (*81*) showed that neurons in a randomly initialized deep neural network without training could selectively respond to faces, and the neurons in the deeper layers are more selective. (*93*) demonstrated that brain-like functional segregation can emerge spontaneously in deep neural networks trained on object recognition and face perception and proposed that the development of functional segregation of face recognition in the brain is a result of computational optimization in the cortex. Augmenting this rapidly growing literature, our study is the first to demonstrate that emotion selectivity can spontaneously emerge in a deep artificial neural network model of the human visual system trained to recognize non-emotional objects.

Layer dependence

Like the brain, the CNN model has a layered structure which allows the processing of information in a hierarchical fashion. Our layer-wise analysis showed that the extent and strength of emotion-neural selectivity are a function of the model layers. Compared to the early layers, the deeper layers have larger portions of neurons that show emotion selectivity, and the selectivity is stronger, consistent with the previous observations that deeper layers of CNN models encode more abstract concepts. For example, (*40*, *94*) examined the internal representations of different layers in a CNN and found that deeper layers of the network tend to encode more abstract concepts, such as object parts and textures. The layered processing of emotional information may have several functional benefits. First, by processing visual information in hierarchical stages, the brain can quickly and efficiently respond to stimuli without the need for a complete and detailed analysis of the entire stimulus at once (*95–97*). This is especially important for the processing of emotionally salient stimuli, as quick and accurate emotional responses can be crucial for survival. Second, it would offer more flexibility for the processing of emotion at different levels of detail, which may depend on the perception task and the environmental context. For example, If the stimulus is perceived as significant or crucial, it elicits a stronger and more widespread neural response, engaging multiple regions and processing stages. On the other hand, if the stimulus is not significant, it elicits a weaker and more limited neural response involving fewer regions or layers and processing stages (*98–100*). Third, the integration of information from different levels allows for a more complete and nuanced representation of the visual stimulus and emotional response. This allows for the creation of a final representation that takes into account not just the visual properties of the stimulus but also its emotional significance and its impact on the individual (*101–103*). Lastly, by processing information in a layer-dependent manner, the brain can adapt and change the processing of information based on experience and learning (*104*). This allows the brain to refine its processing strategies and improve its performance over time (*105*).

Limitations

Several limitations of our study should be noted. First, we tested our hypothesis on a VGG-16 architecture. Although VGG-16 is a common architecture and widely used in studies to model the human visual system, it remains possible that the neuronal selectivity reported is model specific. Future work can address this problem by examining a broader range of CNN models as well as other deep neural network architectures to test the generality of the results. Second, the present study considered whether neuronal selectivity for emotion could arise in neurons of a visual system model that has not been exposed to affective content. It remains to be determined whether the selectivity could arise inherently from the hierarchical network structures, even in the absence of any training. Third, in our study, emotion was divided into three broad categories: pleasant, unpleasant and neutral. While this is in line with many neurophysiological studies in humans, future work should examine finer differentiations of emotion, e.g., joy, sadness, horror, disgust, and so on, and their neural representations in the brain. Lastly, despite the repeated demonstration that CNNs provide good models of human visual processing, caution must be exercised when extending the findings made on CNN models to the human brain.

To conclude, the present study shows that emotion selectivity can spontaneously emerge in a deep neural network trained to recognize non-emotional visual objects, which becomes stronger as we move deeper into the network. Two manipulations, attention enhancement, and lesion, further demonstrate the functional significance of the emotion-selective neurons. In addition to offering support to the idea that the visual system may have an innate ability to represent motivational values of sensory input, our findings also suggest that CNNs offer a valuable platform for testing neuroscience ideas in a way that is not practical in empirical studies.

Materials and Methods

Affective picture sets

Two sets of widely used affective images were used in this study. The IAPS library includes 1,182 images covering approximately 20 subclasses of emotions such as joy, surprise, entrancement, sadness, romance, disgust, and fear. The NAPS library has 1,356 images that can be divided into similar subclasses. For both libraries, each image has a normative valence rating, ranging from 1 to 9, indicating whether the image expresses unpleasant, neutral or pleasant emotions; the distributions of the valence rating from the two datasets were given in Fig.S1-C (right). In this study, for simplicity and in accordance with the common practice in human imaging studies of emotion (*20*, *32*, *106–108*), we classified images into three main categories based on their valence scores: "pleasant," "neutral," and "unpleasant." For images that fell near the boundary between categories, we used soft thresholds of 4.3±0.5 and 6.0±0.5 to determine their classification as either "unpleasant" or "neutral," or "neutral" or "pleasant." We also visually examined each image to confirm its category. Finally, any images that we could not confidently classify were marked as "unknown" and removed from the analysis. This process resulted in some differences in the number of images in each category from the original datasets. After this categorization, IAPS images were divided into 296 pleasant, 390 neutral, and 341 unpleasant images, and NAPS into 352 pleasant, 477 neutral, and 281 unpleasant images (see Fig.S1-B).

The convolutional neural network model

VGG-16, a well-tested deep convolutional neural network for natural image recognition, was used in this study to evaluate emotion selectivity. It has 13 convolutional layers followed by three fully connected layers, with the last fully connected layer containing 1000 units for recognizing 1000 different types of visual objects. Each layer of VGG-16 contains a large number of artificial neurons (also referred to variously as filters or channels or feature maps). Each neuron is characterized by a ReLU activation function (see Fig.S1-A). Through this function, neurons within a given layer, upon receiving and processing the input from the previous layer, yield activation maps called feature maps which become the input for the next layer. Previous studies have compared the activation patterns of the VGG-16 model have been compared with experimental recordings from both humans and non-human primates and found that early layers of the model behave similarly to early visual areas such as V1, whereas deeper layers of the model are more analogous to higher-order visual areas such as the object-selective lateral occipital areas (*22*, *109–111*).

In this study, VGG-16 was used in two ways. First, to examine whether emotional selectivity emerges in neurons trained to recognize non-emotional objects, we took the VGG-16 model pre-trained on 1.2 million natural images from the ImageNet, presented affective pictures from the two aforementioned affective picture datasets to the model, and analyzed the activation profiles of neurons from each layer. Emotional selectivity of each neuron was determined from these activation profiles (see below). Second, to test the functionality of the emotion-selective neurons, we replaced the last layer of the VGG-16 with a two-unit fully connected layer and trained the connections to this two-unit layer to recognize two categories of emotion: pleasant versus non-pleasant, neutral versus non-neutral, or unpleasant versus non-unpleasant. The training of the two-unit last layer used cross-entropy as the objective function. The other weights in the network remained the same as that trained on the ImageNet data (i.e., they are frozen).

Emotional selectivity calculation

The output from each unit in a feature map (see Fig.S1-A) can be written as:

$$X^{lk} = (1+\alpha)max[0, w^{lk} * X^{l-1}] \tag{1}$$

where w^{lk} indicates the kernel weight of the kth filter in the lth convolutional layer, and $*$ indicates mathematical convolution which applies matrix multiplication between w and the outputs X from the $(l-1)^{th}$ layer. Of note in Eq. (1) is that the ReLU activation function typically has a slope of 1 ($\alpha = 0$). Here in this work the slope is a tunable parameter. By tuning the slope of the ReLU function, we change the gain of the neuron, simulating the effect of feature based attention control (*39, 58*).

Let $X_{i,j}^{lk}(n)$ represents the response of the unit located at coordinates (i,j) in the kth filter in layer l to image n. Then

$$\overline{p}^{lk}(n) = \frac{1}{WH}\sum_{i=1,j=1}^{W,H} X_{i,j}^{lk}(n) \tag{2}$$

is the response to the image averaged across the entire filter. Here W and H represent the width and height of the feature map. Thus, the mean activity of the filter k in layer l in response to all images in a dataset can be formulated as:

$$\hat{p}^{lk} = \frac{1}{N}\sum_{n=1}^{N} \overline{p}^{lk}(n) \tag{3}$$

where N represents the total number of images in a given set. Emotional selectivity of the filter is calculated according to

$$S_e^{lk} = \frac{\frac{1}{N_e}\sum_{n=1}^{N_e} \overline{p}^{lk}(n) - \hat{p}^{lk}}{\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\overline{p}^{lk}(n) - \hat{p}^{lk}\right)^2}} \tag{4}$$

where S_e^{lk} represents the normalized activation of filter k in layer l in response to all images of emotion category e, where $e \in \{pleasant, neutral, unpleasant\}$. A neuron is considered selective for a specific emotion if the normalized activation for the images within that emotion category is highest among the three possible values. For example, if $S_{e=pleasant}^{lk}$=-0.1, $S_{e=neutral}^{lk} =$ 0.2, and $S_{e=unpleasant}^{lk}$=0.3, the artificial neuron k is considered selective for "unpleasant images".

To mitigate the possible influence of the idiosyncrasies within each picture dataset set on emotion selectivity definition, we further specified that a neuron is said to be selective for a given emotion if it is selective for the same emotion defined by both the IAPS dataset and the NAPS dataset.

Testing the functionality of the emotion-selective neurons

Do the emotion-selective neurons defined above have a functional role? We applied two different approaches to examine this question: lesion and attention enhancement.

Lesion. If the emotion-selective neurons are functionally important, then lesioning these neurons should lead to degraded performance in recognizing the emotion of a given image. Here the lesion

of a specific neuron is achieved by setting its output to 0 (namely, setting $\alpha = -1$ in Eq. (1)). In our experiments, we lesioned the neurons selective for a given emotion as well as randomly selected neurons in a particular layer and observed the changes in the emotion recognition performance of the model.

Attention enhancement. We further tested whether enhancing the activity of an emotion-selective neuron can lead to performance improvement in emotion recognition. Following (*39*), the strength of α was increased from 0 to 5 with interval step size 0.1, where $\alpha = 0$ is the conventional choice and $\alpha > 0$ represents increased neuronal gain (i.e., enhanced feature attention). According to the feature similarity gain theory, increasing the gain of a neuron leads to enhanced performance of the neuron in perceiving stimuli with the relevant features. In our experiments, we enhanced the neurons selective for a given emotion as well as randomly selected neurons in a particular layer and observed the changes in the emotion recognition performance of the model (*39*) (see Fig.S2-A,B).

The generalization and overlap of the emotion-selective neurons

To test whether a neuron's selectivity for a particular emotion is innate and independent of the idiosyncrasies of different picture sets, we compared neurons determined to be selective for a certain type of emotion by the IAPS dataset and by the NAPS dataset. The overlap between the two groups of neurons was quantified by the Jaccard index as follows:

$$U_e^l = \frac{O_e^l}{K_{e,IAPS}^l + K_{e,NAPS}^l - O_e^l} \tag{5}$$

where U_e^l represents the Jaccard index of the overlapped neurons in layer l in response to emotion category e. O_e^l represents the number of overlapping neurons in layer l in response to emotion category e across two datasets: IAPS and NAPS. K is the total number of selective neurons in layer l in response to emotion category e defined on dataset IAPS or NAPS. To validate the results are not by chance, we randomly selected the same number of neurons as O_e^l defined on the dataset IAPS and NAPS separately to get the overlap between two groups of randomly selected neurons of U_e^l.

References

1. S. Kitayama, *Emotion and culture: Empirical studies of mutual influence* (American Psychological Association, Washington, DC, US, 1994).

2. E. D. McCarthy, The Social Construction of Emotions: New Directions from Culture Theory. *Sociology Faculty Publications* (1994) (available at https://research.library.fordham.edu/soc_facultypubs/4).

3. S. J. Banks, K. T. Eddy, M. Angstadt, P. J. Nathan, K. L. Phan, Amygdala–frontal connectivity during emotion regulation. *Social Cognitive and Affective Neuroscience*. **2**, 303–312 (2007).

4. J. J. Gross, "Emotion regulation: Conceptual and empirical foundations" in *Handbook of emotion regulation, 2nd ed* (The Guilford Press, New York, NY, US, 2014), pp. 3–20.

5. L. F. Barrett, M. Lewis, J. M. Haviland-Jones, *Handbook of Emotions* (Guilford Publications, 2016; https://books.google.com/books?id=cbKhDAAAQBAJ).

6. H. A. Elfenbein, N. Ambady, On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychol Bull*. **128**, 203–235 (2002).

7. S. Hareli, K. Kafetsios, U. Hess, A cross-cultural study on emotion expression and the learning of social norms. *Frontiers in Psychology*. **6** (2015) (available at https://www.frontiersin.org/articles/10.3389/fpsyg.2015.01501).

8. B. Q. Ford, I. B. Mauss, Culture and emotion regulation. *Curr Opin Psychol*. **3**, 1–5 (2015).

9. S. Olderbak, O. Wilhelm, Emotion perception and empathy: An individual differences test of relations. *Emotion*. **17**, 1092–1106 (2017).

10. R. S. Lazarus, *Emotion and Adaptation* (Oxford University Press, 1991).

11. J. A. Coan, *Handbook of emotion elicitation and assessment* (Oxford University Press, New York, NY, US, 2007), *Handbook of emotion elicitation and assessment*.

12. V. LoBue, Behavioral evidence for a continuous approach to the perception of emotionally valenced stimuli. *Behavioral and Brain Sciences*. **38**, e79 (2015).

13. M. K. Greenwald, E. W. Cook, P. J. Lang, Affective judgment and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli. *Journal of Psychophysiology*. **3**, 51–64 (1989).

14. M. M. Bradley, P. J. Lang, Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*. **25**, 49–59 (1994).

15. P. Lang, International affective picture system (IAPS) : affective ratings of pictures and instruction manual. *undefined* (2005) (available at https://www.semanticscholar.org/paper/International-affective-picture-system-(IAPS)-%3A-of-Lang/788e2f5a24784ce952eec8a57902a6f03cd9318c).

16. A. Marchewka, Ł. Żurawski, K. Jednoróg, A. Grabowska, The Nencki Affective Picture System (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database. *Behav Res*. **46**, 596–610 (2014).

17. T. Canli, Z. Zhao, J. E. Desmond, E. Kang, J. Gross, J. D. E. Gabrieli, An fMRI study of personality influences on brain reactivity to emotional stimuli. *Behavioral Neuroscience*. **115**, 33–42 (2001).

18. P. Vrticka, S. Simioni, E. Fornari, M. Schluep, P. Vuilleumier, D. Sander, Neural Substrates of Social Emotion Regulation: A fMRI Study on Imitation and Expressive Suppression to Dynamic Facial Signals. *Frontiers in Psychology*. **4** (2013) (available at https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00095).

19. M. Résibois, P. Verduyn, P. Delaveau, J.-Y. Rotgé, P. Kuppens, I. Van Mechelen, P. Fossati, The neural basis of emotions varies over time: different regions go with onset- and offset-bound processes underlying emotion intensity. *Social Cognitive and Affective Neuroscience*. **12**, 1261–1271 (2017).

20. K. Bo, S. Yin, Y. Liu, Z. Hu, S. Meyyappan, S. Kim, A. Keil, M. Ding, Decoding Neural Representations of Affective Scenes in Retinotopic Visual Cortex. *Cerebral Cortex*. **31**, 3047–3063 (2021).

21. H. Saarimäki, Naturalistic Stimuli in Affective Neuroimaging: A Review. *Frontiers in Human Neuroscience*. **15** (2021) (available at https://www.frontiersin.org/articles/10.3389/fnhum.2021.675068).

22. D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, J. J. DiCarlo, Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*. **111**, 8619–8624 (2014).

23. U. Güçlü, M. A. J. van Gerven, Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *J. Neurosci.* **35**, 10005–10014 (2015).

24. D. L. K. Yamins, J. J. DiCarlo, Using goal-driven deep learning models to understand sensory cortex. *Nat Neurosci*. **19**, 356–365 (2016).

25. A. H. Marblestone, G. Wayne, K. P. Kording, Toward an Integration of Deep Learning and Neuroscience. *Front. Comput. Neurosci.* **10** (2016), doi:10.3389/fncom.2016.00094.

26. B. A. Richards, T. P. Lillicrap, P. Beaudoin, Y. Bengio, R. Bogacz, A. Christensen, C. Clopath, R. P. Costa, A. de Berker, S. Ganguli, C. J. Gillon, D. Hafner, A. Kepecs, N. Kriegeskorte, P. Latham, G. W. Lindsay, K. D. Miller, R. Naud, C. C. Pack, P. Poirazi, P. Roelfsema, J. Sacramento, A. Saxe, B. Scellier, A. C. Schapiro, W. Senn, G. Wayne, D. Yamins, F. Zenke, J. Zylberberg, D. Therien, K. P. Kording, A deep learning framework for neuroscience. *Nat Neurosci*. **22**, 1761–1770 (2019).

27. J. Deng, W. Dong, R. Socher, L. Li, Kai Li, Li Fei-Fei, "ImageNet: A large-scale hierarchical image database" in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), pp. 248–255.

28. K. Nasr, P. Viswanathan, A. Nieder, Number detectors spontaneously emerge in a deep neural network designed for visual object recognition. *Science Advances*. **5**, eaav7903 (2019).

29. K. Dobs, A. Kell, J. Martinez, M. Cohen, N. Kanwisher, N. Kanwisher, Why Are Face and Object Processing Segregated in the Human Brain? Testing Computational Hypotheses with Deep Convolutional Neural Networks (2020).

30. P. Vuilleumier, M. P. Richardson, J. L. Armony, J. Driver, R. J. Dolan, Distant influences of amygdala lesion on visual cortical activation during emotional face processing. *Nat Neurosci*. **7**, 1271–1278 (2004).

31. M. G. Shuler, M. F. Bear, Reward Timing in the Primary Visual Cortex. *Science*. **311**, 1606–1609 (2006).

32. P. A. Kragel, M. C. Reddan, K. S. LaBar, T. D. Wager, Emotion schemas are embedded in the human visual system. *Science Advances*. **5**, eaaw4358 (2019).

33. K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]* (2015) (available at http://arxiv.org/abs/1409.1556).

34. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vis*. **115**, 211–252 (2015).

35. K. Seeliger, M. Fritsche, U. Güçlü, S. Schoenmakers, J.-M. Schoffelen, S. E. Bosch, M. A. J. van Gerven, Convolutional neural network-based encoding and decoding of visual object recognition in space and time. *NeuroImage*. **180**, 253–266 (2018).

36. G. Jacob, R. T. Pramod, H. Katti, S. P. Arun, Qualitative similarities and differences in visual object representations between brains and deep networks. *Nat Commun*. **12**, 1872 (2021).

37. P. Thompson, Margaret Thatcher: A New Illusion. *Perception*. **9**, 483–484 (1980).

38. P. T. Sowden, "Psychophysics" in *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics* (American Psychological Association, Washington, DC, US, 2012), *APA handbooks in psychology®*, pp. 445–458.

39. G. W. Lindsay, K. D. Miller, How biological attention mechanisms improve task performance in a large-scale visual system model. *eLife*. **7**, e38105 (2018).

40. M. D. Zeiler, R. Fergus, "Visualizing and Understanding Convolutional Networks" in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars, Eds. (Springer International Publishing, Cham, 2014), pp. 818–833.

41. G. Lee, Y.-W. Tai, J. Kim, Deep Saliency with Encoded Low level Distance Map and High Level Features. *arXiv:1604.05495 [cs]* (2016) (available at http://arxiv.org/abs/1604.05495).

42. G. W. Lindsay, Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of Cognitive Neuroscience*, 1–15 (2020).

43. A. Pimashkin, I. Kastalskiy, A. Simonov, E. Koryagina, I. Mukhina, V. Kazantsev, Spiking Signatures of Spontaneous Activity Bursts in Hippocampal Cultures. *Frontiers in Computational Neuroscience*. **5** (2011) (available at https://www.frontiersin.org/article/10.3389/fncom.2011.00046).

44. J. Winnubst, E. Bas, T. A. Ferreira, Z. Wu, M. N. Economo, P. Edson, B. J. Arthur, C. Bruns, K. Rokicki, D. Schauder, D. J. Olbris, S. D. Murphy, D. G. Ackerman, C. Arshadi, P. Baldwin, R. Blake, A. Elsayed, M. Hasan, D. Ramirez, B. Dos Santos, M. Weldon, A. Zafar, J. T. Dudman, C. R. Gerfen, A. W. Hantman, W. Korff, S. M. Sternson, N. Spruston, K. Svoboda, J. Chandrashekar, Reconstruction of 1,000 Projection Neurons Reveals New Cell Types and Organization of Long-Range Connectivity in the Mouse Brain. *Cell*. **179**, 268-281.e13 (2019).

45. C. M. Palmateer, S. C. Moseley, S. Ray, S. G. Brovero, M. N. Arbeitman, Analysis of cell-type-specific chromatin modifications and gene expression in Drosophila neurons that direct reproductive behavior. *PLOS Genetics*. **17**, e1009240 (2021).

46. M. T. Schmitz, K. Sandoval, C. P. Chen, M. A. Mostajo-Radji, W. W. Seeley, T. J. Nowakowski, C. J. Ye, M. F. Paredes, A. A. Pollen, The development and evolution of inhibitory neurons in primate cerebrum. *Nature*. **603**, 871–877 (2022).

47. J. H. R. Maunsell, S. Treue, Feature-based attention in visual cortex. *Trends in Neurosciences*. **29**, 317–322 (2006).

48. G. W. Lindsay, Feature-based Attention in Convolutional Neural Networks. *arXiv:1511.06408 [cs]* (2015) (available at http://arxiv.org/abs/1511.06408).

49. C.-H. Yeh, M.-H. Lin, P.-C. Chang, L.-W. Kang, Enhanced Visual Attention-Guided Deep Neural Networks for Image Classification. *IEEE Access*. **8**, 163447–163457 (2020).

50. J. A. Cardin, L. A. Palmer, D. Contreras, Cellular mechanisms underlying stimulus-dependent gain modulation in primary visual cortex neurons in vivo. *Neuron*. **59**, 150–160 (2008).

51. E. Eldar, J. D. Cohen, Y. Niv, The effects of neural gain on attention and learning. *Nat Neurosci*. **16**, 1146–1153 (2013).

52. S. Jarvis, K. Nikolic, S. R. Schultz, Neuronal gain modulability is determined by dendritic morphology: A computational optogenetic study. *PLOS Computational Biology*. **14**, e1006027 (2018).

53. H. Bos, A.-M. Oswald, B. Doiron, Untangling stability and gain modulation in cortical circuits with multiple interneuron classes (2020), p. 2020.06.15.148114, , doi:10.1101/2020.06.15.148114.

54. P. L. Broadhurst, Emotionality and the Yerkes-Dodson Law. *Journal of Experimental Psychology*. **54**, 345–352 (1957).

55. L. K. A. Sörensen, S. M. Bohté, H. A. Slagter, H. S. Scholte, Arousal state affects perceptual decision-making by modulating hierarchical sensory processing in a large-scale visual system model. *PLOS Computational Biology*. **18**, e1009976 (2022).

56. R. Aharonov, L. Segev, I. Meilijson, E. Ruppin, Localization of Function via Lesion Analysis. *Neural Computation*. **15**, 885–913 (2003).

57. L. J. Chareyron, D. G. Amaral, P. Lavenex, Selective lesion of the hippocampus increases the differentiation of immature neurons in the monkey amygdala. *Proceedings of the National Academy of Sciences*. **113**, 14420–14425 (2016).

58. G. R. Yang, M. R. Joglekar, H. F. Song, W. T. Newsome, X.-J. Wang, Task representations in neural networks trained to perform many cognitive tasks. *Nat Neurosci*. **22**, 297–306 (2019).

59. S. Cohen-Zimerman, H. Khilwani, G. N. L. Smith, F. Krueger, B. Gordon, J. Grafman, The neural basis for mental state attribution: A voxel-based lesion mapping study. *Human Brain Mapping*. **42**, 65–79 (2021).

60. P. J. Lang, M. M. Bradley, B. N. Cuthbert, "Motivated attention: Affect, activation, and action" in *Attention and orienting: Sensory and motivational processes* (Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, 1997), pp. 97–135.

61. N. Kriegeskorte, Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*. **1**, 417–446 (2015).

62. A. Brachmann, E. Barth, C. Redies, Using CNN Features to Better Understand What Makes Visual Artworks Special. *Frontiers in Psychology*. **8** (2017) (available at https://www.frontiersin.org/articles/10.3389/fpsyg.2017.00830).

63. K. Iigaya, S. Yi, I. A. Wahle, K. Tanwisuth, J. P. O'Doherty, Aesthetic preference for art can be predicted from a mixture of low- and high-level visual features. *Nat Hum Behav*. **5**, 743–755 (2021).

64. L. E. van Dyck, R. Kwitt, S. J. Denzler, W. R. Gruber, Comparing Object Recognition in Humans and Deep Convolutional Neural Networks—An Eye Tracking Study. *Frontiers in Neuroscience*. **15** (2021) (available at https://www.frontiersin.org/articles/10.3389/fnins.2021.750639).

65. J. J. D. Singer, K. Seeliger, T. C. Kietzmann, M. N. Hebart, From photos to sketches - how humans and deep neural networks process objects across different levels of visual abstraction. *Journal of Vision*. **22**, 4 (2022).

66. J. Lee, M. Jung, N. Lustig, J.-H. Lee, Neural representations of the perception of handwritten digits and visual objects from a convolutional neural network compared to humans. *Human Brain Mapping*. **n/a** (2023), doi:10.1002/hbm.26189.

67. J. Kauramäki, I. P. Jääskeläinen, M. Sams, Selective Attention Increases Both Gain and Feature Selectivity of the Human Auditory Cortex. *PLOS ONE*. **2**, e909 (2007).

68. S. Moldakarimov, M. Bazhenov, T. J. Sejnowski, Top-Down Inputs Enhance Orientation Selectivity in Neurons of the Primary Visual Cortex during Perceptual Learning. *PLOS Computational Biology*. **10**, e1003770 (2014).

69. T. Pasternak, D. Tadin, Linking Neuronal Direction Selectivity to Perceptual Decisions About Visual Motion. *Annu Rev Vis Sci*. **6**, 335–362 (2020).

70. P. J. Lang, M. M. Bradley, J. R. Fitzsimmons, B. N. Cuthbert, J. D. Scott, B. Moulder, V. Nangia, Emotional arousal and activation of the visual cortex: An fMRI analysis. *Psychophysiology*. **35**, 199–210 (1998).

71. P. Rotshtein, R. Malach, U. Hadar, M. Graif, T. Hendler, Feeling or Features: Different Sensitivity to Emotion in High-Order Visual Cortex and Amygdala. *Neuron*. **32**, 747–757 (2001).

72. H. T. Schupp, J. Markus, A. I. Weike, A. O. Hamm, Emotional Facilitation of Sensory Processing in the Visual Cortex. *Psychol Sci*. **14**, 7–13 (2003).

73. D. Sabatinelli, M. M. Bradley, J. R. Fitzsimmons, P. J. Lang, Parallel amygdala and inferotemporal activation reflect emotional intensity and fear relevance. *Neuroimage*. **24**, 1265–1270 (2005).

74. P. J. Lang, M. M. Bradley, Emotion and the motivational brain. *Biol Psychol*. **84**, 437–450 (2010).

75. L. Pessoa, Emotion and Cognition and the Amygdala: From "what is it?" to "what's to be done?" *Neuropsychologia*. **48**, 3416–3429 (2010).

76. V. Miskovic, A. K. Anderson, Modality general and modality specific coding of hedonic valence. *Curr Opin Behav Sci*. **19**, 91–97 (2018).

77. N. M. Weinberger, Specific long-term memory traces in primary auditory cortex. *Nat Rev Neurosci*. **5**, 279–290 (2004).

78. Z. Li, A. Yan, K. Guo, W. Li, Fear-Related Signals in the Primary Visual Cortex. *Curr Biol*. **29**, 4078-4083.e2 (2019).

79. N. N. Thigpen, F. Bartsch, A. Keil, The malleability of emotional perception: Short-term plasticity in retinotopic neurons accompanies the formation of perceptual biases to threat. *Journal of Experimental Psychology: General*. **146**, 464–471 (2017).

80. V. Miskovic, A. Keil, Acquired fears reflected in cortical sensory processing: A review of electrophysiological studies of human classical conditioning. *Psychophysiology*. **49**, 1230–1241 (2012).

81. S. Baek, M. Song, J. Jang, G. Kim, S.-B. Paik, Face detection in untrained deep neural networks. *Nat Commun*. **12**, 7328 (2021).

82. D. Burr, J. Ross, A Visual Sense of Number. *Current Biology*. **18**, 425–428 (2008).

83. A. Nieder, The neuronal code for number. *Nat Rev Neurosci*. **17**, 366–382 (2016).

84. F. Xu, E. S. Spelke, Large number discrimination in 6-month-old infants. *Cognition*. **74**, B1–B11 (2000).

85. F. Xu, E. S. Spelke, S. Goddard, Number sense in human infants. *Dev Sci*. **8**, 88–101 (2005).

86. S. Santens, C. Roggeman, W. Fias, T. Verguts, Number Processing Pathways in Human Parietal Cortex. *Cerebral Cortex*. **20**, 77–88 (2010).

87. M. D. Hauser, S. Carey, L. B. Hauser, Spontaneous number representation in semi-free-ranging rhesus monkeys. *Proc Biol Sci*. **267**, 829–833 (2000).

88. H. Sawamura, K. Shima, J. Tanji, Numerical representation for action in the parietal cortex of the monkey. *Nature*. **415**, 918–922 (2002).

89. M. D. Hauser, F. Tsao, P. Garcia, E. S. Spelke, Evolutionary foundations of number: spontaneous representation of numerical magnitudes by cotton–top tamarins. *Proceedings of the Royal Society of London. Series B: Biological Sciences*. **270**, 1441–1446 (2003).

90. K. Nasr, P. Viswanathan, A. Nieder, Number detectors spontaneously emerge in a deep neural network designed for visual object recognition. *Science Advances*. **5**, eaav7903.

91. G. Kim, J. Jang, S. Baek, M. Song, S.-B. Paik, Visual number sense in untrained deep neural networks. *Science Advances*. **7**, eabd6127 (2021).

92. N. Kanwisher, J. McDermott, M. M. Chun, The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *J. Neurosci.* **17**, 4302–4311 (1997).

93. K. Dobs, J. Martinez, A. J. E. Kell, N. Kanwisher, Brain-like functional specialization emerges spontaneously in deep neural networks. *Science Advances*. **8**, eabl8913 (2022).

94. C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning requires rethinking generalization. *arXiv:1611.03530 [cs]* (2017) (available at http://arxiv.org/abs/1611.03530).

95. R. VanRullen, S. J. Thorpe, The time course of visual processing: from early perception to decision-making. *J Cogn Neurosci*. **13**, 454–461 (2001).

96. N. Srinivasan, R. Gupta, Rapid communication: Global-local processing affects recognition of distractor emotional faces. *Q J Exp Psychol (Hove)*. **64**, 425–433 (2011).

97. L. Cabral, B. Stojanoski, R. Cusack, Rapid and coarse face detection: With a lack of evidence for a nasal-temporal asymmetry. *Atten Percept Psychophys*. **82**, 1883–1895 (2020).

98. K. Zipser, V. A. F. Lamme, P. H. Schiller, Contextual Modulation in Primary Visual Cortex. *J. Neurosci.* **16**, 7376–7389 (1996).

99. S. Tschechne, H. Neumann, Hierarchical representation of shapes in visual cortex—from localized features to figural shape segregation. *Front Comput Neurosci*. **8**, 93 (2014).

100. R. M. Willems, M. V. Peelen, How context changes the neural basis of perception and language. *iScience*. **24**, 102392 (2021).

101. M. M. Bradley, P. J. Lang, Affective reactions to acoustic stimuli. *Psychophysiology*. **37**, 204–215 (2000).

102. E. Harmon-Jones, P. A. Gable, C. K. Peterson, The role of asymmetric frontal cortical activity in emotion-related phenomena: A review and update. *Biological Psychology*. **84**, 451–462 (2010).

103. P. M. Niedenthal, A. Wood, Does emotion influence visual perception? Depends on how you look at it. *Cognition and Emotion*. **33**, 77–84 (2019).

104. G. Li, M. G. Forero, J. S. Wentzell, I. Durmus, R. Wolf, N. C. Anthoney, M. Parker, R. Jiang, J. Hasenauer, N. J. Strausfeld, M. Heisenberg, A. Hidalgo, A Toll-receptor map underlies structural brain plasticity. *eLife*. **9**, e52743 (2020).

105. A. L. Tierney, C. A. Nelson, Brain Development and the Role of Experience in the Early Years. *Zero Three*. **30**, 9–13 (2009).

106. W. Sato, T. Kochiyama, S. Yoshikawa, E. Naito, M. Matsumura, Enhanced neural activity in response to dynamic facial expressions of emotion: an fMRI study. *Brain Res Cogn Brain Res*. **20**, 81–91 (2004).

107. R. M. Cichy, D. Pantazis, A. Oliva, Resolving human object recognition in space and time. *Nat Neurosci*. **17**, 455–462 (2014).

108. P. T. Putnam, K. M. Gothard, Multidimensional Neural Selectivity in the Primate Amygdala. *eNeuro*. **6** (2019), doi:10.1523/ENEURO.0153-19.2019.

109. C. F. Cadieu, H. Hong, D. L. K. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, J. J. DiCarlo, Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLOS Computational Biology*. **10**, e1003963 (2014).

110. R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, A. Oliva, Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci Rep*. **6**, 27755 (2016).

111. M. Eickenberg, A. Gramfort, G. Varoquaux, B. Thirion, Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*. **152**, 184–194 (2017).

Acknowledgments

Funding: This work was supported by the National Institutes of the National Institutes of Health/National Institute of Mental Health grants MH112558 and MH125615, the National Science Foundation grant 1908299, the University of Florida Artificial Intelligence Research Catalyst Fund, the University of Florida Informatics Institute Graduate Student Fellowship, the University of Florida McKnight Brain Institute, the University of Florida Center for Cognitive Aging and Memory, and the McKnight Brain Research Foundation.

Author contributions:
Conceptualization: P.L., R.F., and M.D.; investigation: P.L., K.B., R.F., and M.D.; methodology: P.L., K.B., R.F., and M.D.; software: P.L.; writing—original draft: P.L.; writing—review and editing: P.L., K.B., R.F., and M.D.; visualization: P.L.; supervision: R.L. and M.D.

Competing interests: The authors declare that they have no competing interests.

Data and materials availability: All data are publicly available in the main text or the supplementary materials. The analysis code will available and may be requested from the authors.

Figures and Tables

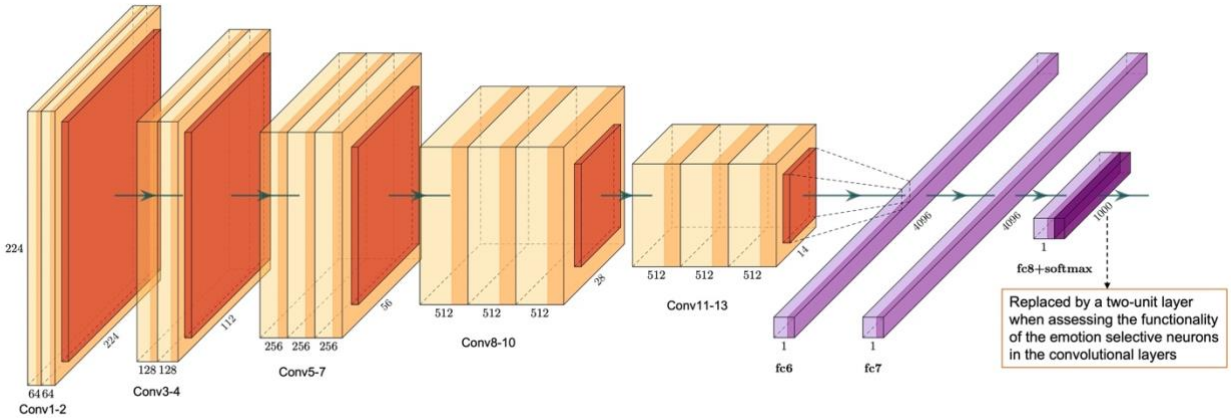


Fig. 1. The architecture of VGG-16. We used the VGG-16 pre-trained on ImageNet to model the visual system. VGG-16 has 13 convolutional layers and three fully connected (FC) layers. Each convolutional layer (light yellow color) is followed by a ReLU activation layer (yellow color) and a max-pooling layer (red color). Each FC layer (light purple color) is followed by a ReLU layer (purple color). The last FC layer is followed by a ReLU and a SoftMax layer (dark purple color). Affective images from two datasets (IAPS and NAPS) were presented to the model to define emotion-selectivity of neurons in the convolutional layers (see details in the Methods section). For attention enhancement and lesion experiments, we replaced the last FC layer with 2 units to represent different emotion categories: (1) pleasant vs. non-pleasant; (2) neutral vs. non-neutral; (3) unpleasant vs. non-unpleasant.

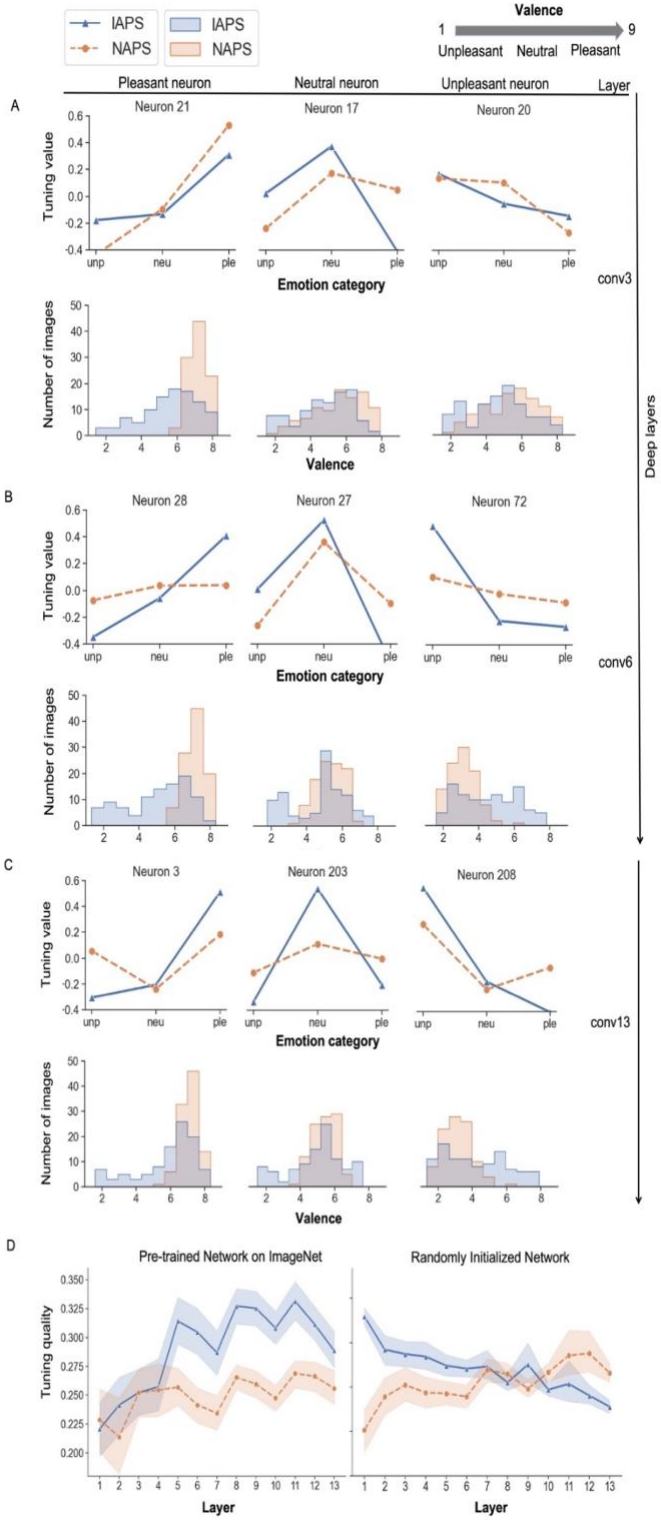


Fig. 2. Tuning curves and emotion selectivity. (**A**-**C**) Tuning curves of example neurons from different convolutional layers (top panel) along with the valence distribution of the top 100 images that elicited the strongest responses from a given neuron. (**D**) Tuning quality as a function of layers, calculated on the VGG-16 pre-trained on ImageNet (left) and on a randomly initialized network (right), respectively. The neurons in deeper layers of the pre-trained network show stronger tuning quality, especially for the IAPS dataset, suggesting that they are better able to differentiate different emotional categories, while no such effects are found for the randomly initialized network.

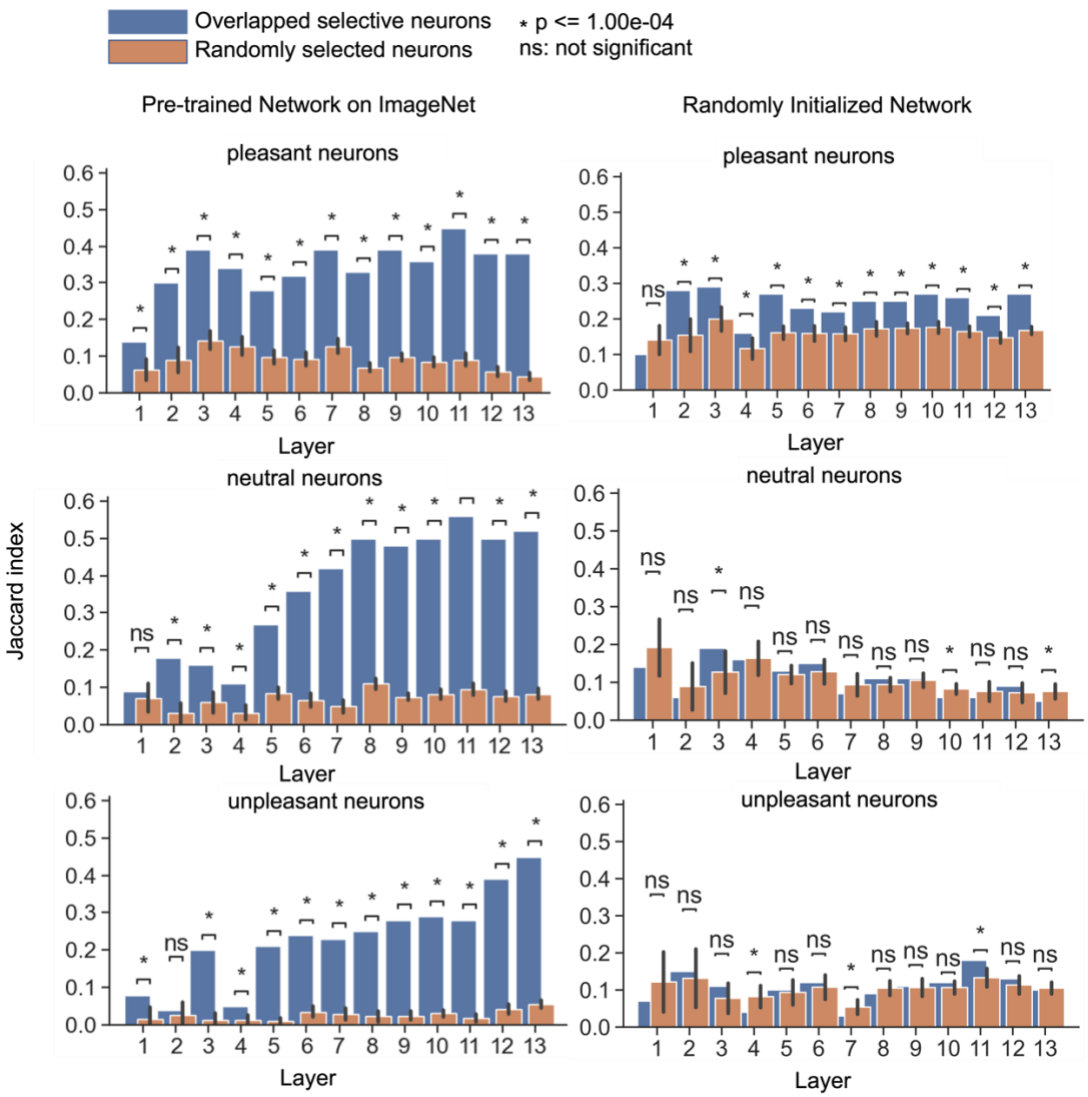


Fig. 3. Generalizability of emotion selectivity across the two datasets. Blue: the overlap between the group of neurons defined to be selective for a given emotion category by IAPS and that by NAPS was assessed using Jaccard Index in each layer of the pre-trained VGG-16 (left column) and a randomly initialized network (right column). Orange: the overlap between two groups of randomly selected neurons.

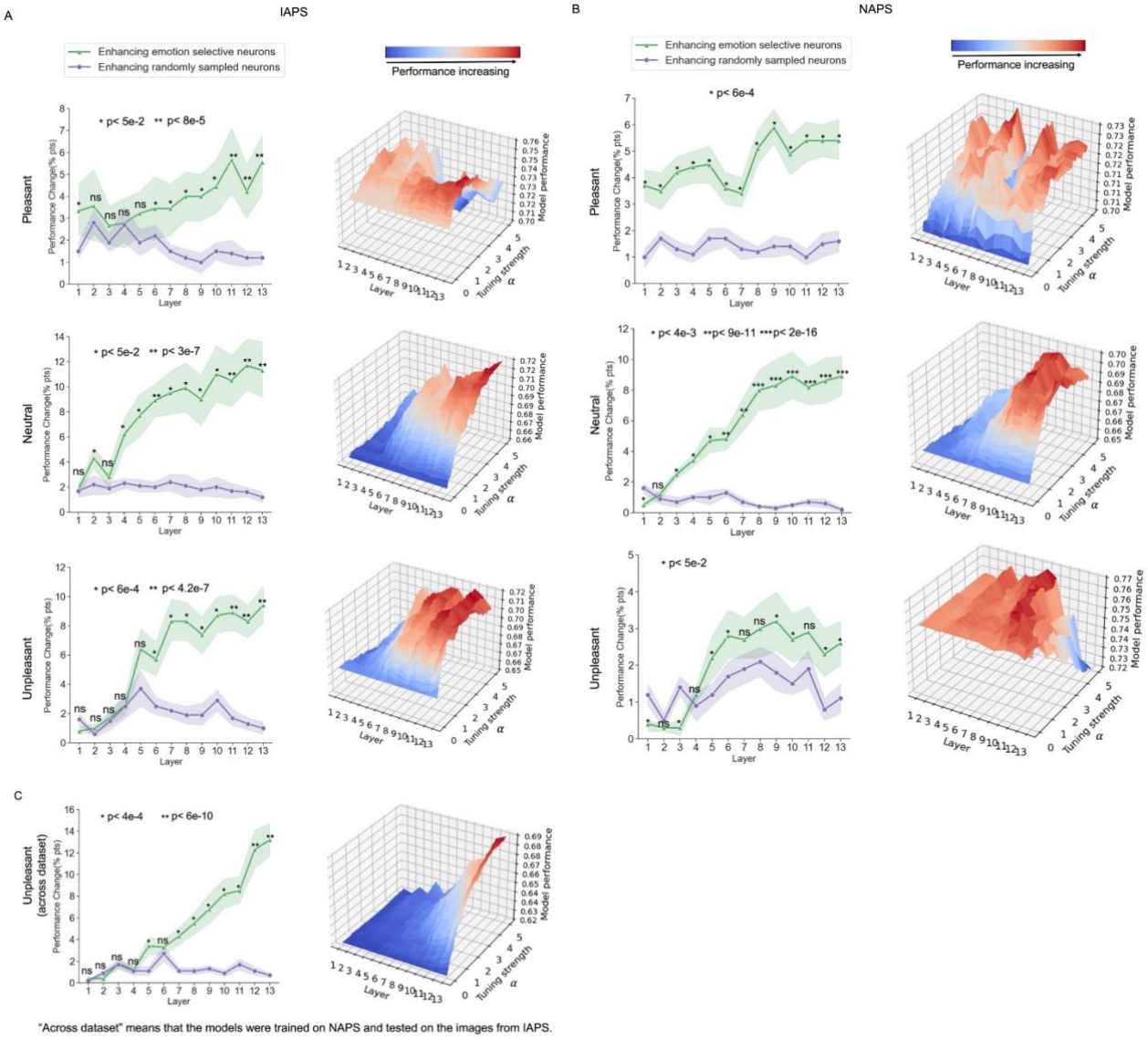


Fig. 4. Effects of enhancing emotion-selective neurons and randomly selected neurons. (**A**) IAPS dataset. (**B**) NAPS dataset. (**C**) Model trained on NAPS dataset tested on IAPS data.

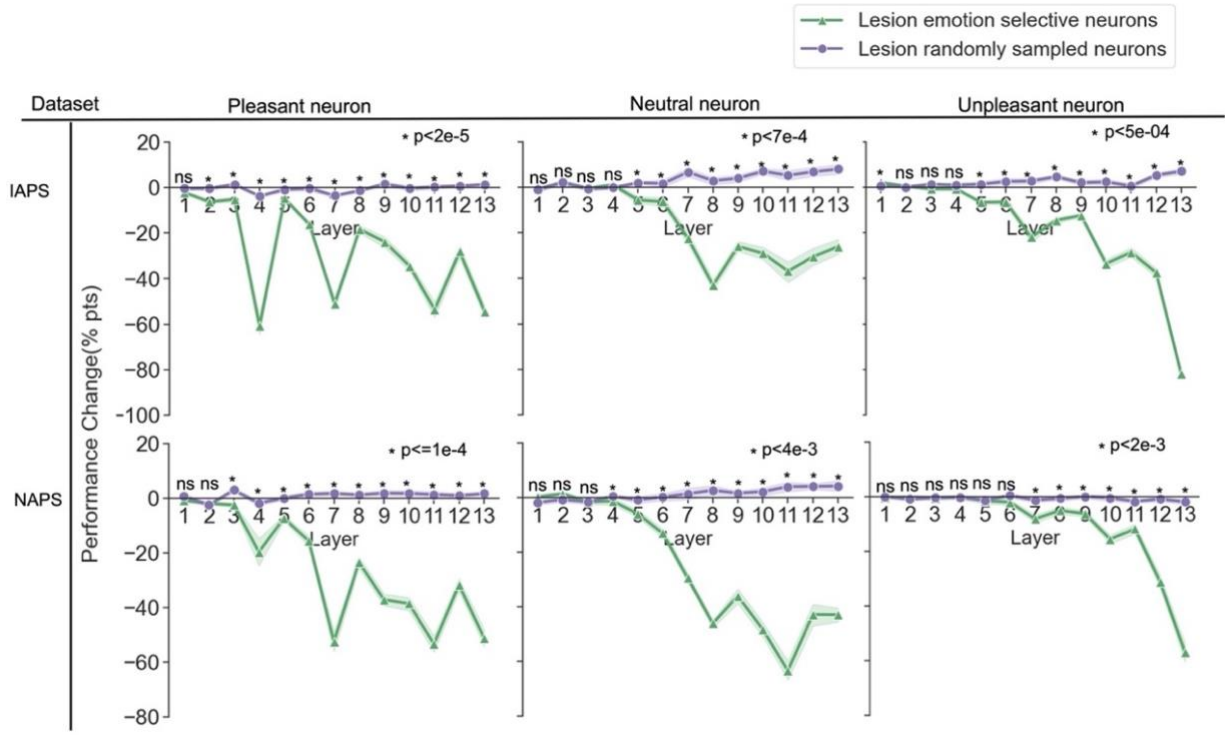


Fig. 5. Lesion Analysis. Performance changes were compared between lesioning emotion-selective neurons and randomly selected neurons.

Supplementary Materials for

Emergence of Emotion Selectivity in A Deep Neural Network Trained to Recognize Visual Objects

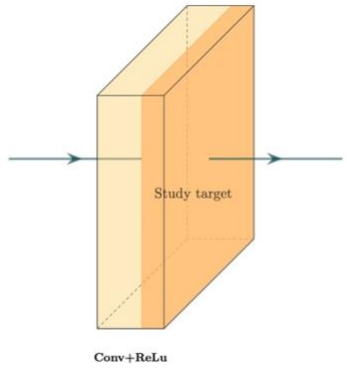
Peng Liu *et al.*

*Corresponding author. Email: mding@bme.ufl.edu; ruogu.fang@bme.ufl.edu

This PDF file includes:

Figs. S1 to S2

A



B

Dataset	Pleasant	Neutral	Unpleasant	Total
IAPS	296	390	341	1,027
NAPS	352	477	281	1,110

C

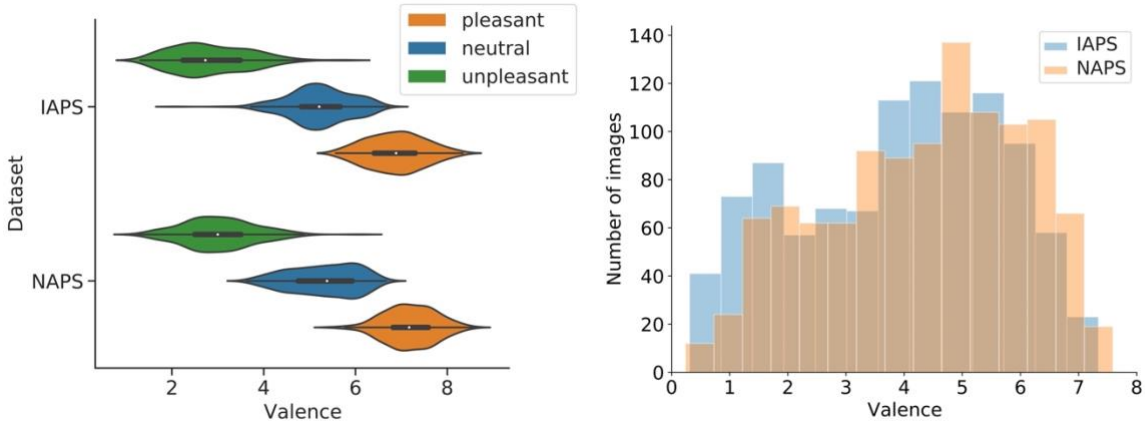
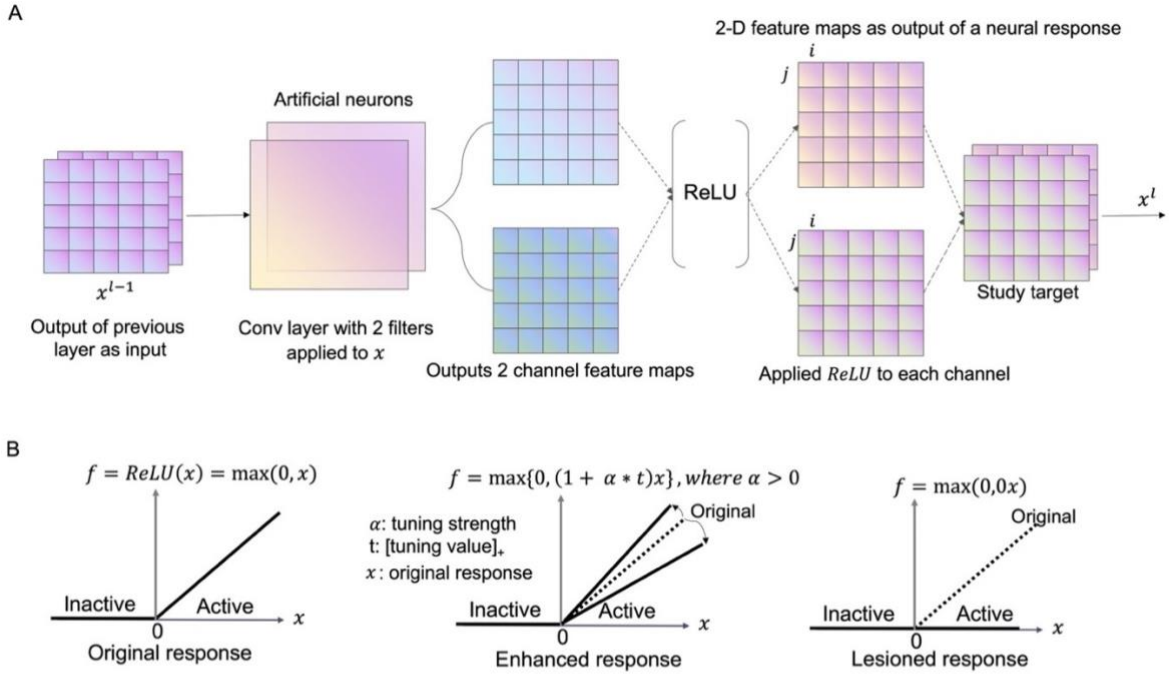


Fig. S1 Model development details. (**A**) Each convolutional layer is followed by one $ReLU$ layer, which reflects the emotional responses of the artificial neurons of the convolutional layer. Thus, in this study, the output of a $ReLU$ layer is our study target for understanding the activity of the artificial neurons. (**B**) It shows the number of images of each emotion category in the two datasets used in this study. Two datasets were treated equally for defining selective neurons, tuning study, lesion study, etc. (**C**) It shows how the divided categorial images match the valence score originally rated by human subjects in the two datasets. The C (left) shows the valence score distribution and the boundary score between the pleasant and neutral category: 4.3 ± 0.5 and between the neutral and unpleasant category: 6.0 ± 0.5. The C (right) shows the number of images per valence score across two datasets. Basically, this figure illustrates the details of the model development in terms of the model itself and datasets.



C

The datasets and models used in study

#emotion categories	Dataset	The usage	Balanced #image per category(e.g., pleasant versus non-pleasant)	Associated model
Binary (2 categories) in each dataset	IAPS-pleasant	Tuning study, random sampling study, and lesion study	296	IAPS-pl
	IAPS-neutral		390	IAPS-ne
	IAPS-unpleasant		341	IAPS-un
	NAPS-pleasant		352	NAPS-pl
	NAPS-neutral		477	NAPS-ne
	NAPS-unpleasant		281	NAPS-un
-	IAPS	defining selectivity and overlapped neurons for each emotion category	See S. Fig.1B	No need model training
	NAPS			

Fig.S2 The activation of calculation with convolution and $ReLU$. (**A**) It shows that the information flow of applying convolution and $ReLU$ operation to the input x, which represents all the feature maps from the previous layer $l-1$. In this demo, x is composed of two channels of feature maps. Two filters (artificial neurons) in the following convolutional layer are applied to x separately. Each filter convolution results in a channel of a new feature map, which then is passed through an activation function, such as $ReLU = \max(0, x)$. $ReLU$ function indicates which features after the convolutional operation is activated or inactivated. The active units are valued with positive numbers, and the inactive units are valued with zeros by the $ReLU$. Fundamentally, this figure indicates how an artificial neuron responds to a stimulus and how the response activations are calculated in a CNN. (**B**) The bold black line represents the $ReLU$ response (activation) values, and dash lines in the middle and right sub-figures represent the normal activation value of the $ReLU$. Three cases of activation behavior were investigated empirically. The normal excitation (left) is applying the original $ReLU$; the enhanced excitation (middle) is applying a positive weight α to the activation value x; the inhabited activation through lesion is setting the activation values to be zeros instead, which performs like a lesion study. (**C**) It summarizes the datasets and the models used in the study. The number of images in the binary model is balanced. The non-* category images were randomly selected from another two categories.