Investigating Emotion-Color Association in Deep Neural Networks

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent research has shown that Deep Neural Networks (DNNs) correlate very well to neural responses, and are widely used by cognitive scientists as a proxy for human representation to model human behavior. But previously it has not been explored whether DNNs capture any aspects of stimuli association. In this study, we experimentally investigate if DNNs can learn implicit associations in stimuli, particularly, an emotion-color association between image stimuli. Our study was conducted in two parts. First, we collected human responses on a forced-choice decision task in which subjects were asked to select a color for a specified emotion-inducing image. Next, we modeled this decision task on neural networks using the similarity between deep representation (extracted using DNNs trained on object classification tasks) of the stimuli images and images of colors used in the task. We found that our model showed a fuzzy linear relationship between the two decision probabilities. This results in two interesting findings, 1. The representations learned by deep neural networks can indeed show an emotion-color association 2. The emotion-color association is not just random but involves some cognitive phenomena. Finally, we also show that this method can help us in the emotion classification task.

1 Introduction

Deep Neural Networks are widely being used in cognitive modeling to model human behavior because of their capability to capture meaningful and human like representations [8, 10, 11]. While deep neural networks show similarities with human representations, one fascinating question remains, can they learn implicit stimuli associations? And, can these deep neural networks show some emotional capabilities, like in humans? In this study, we try to answer this by analyzing the emotion-color association. Emotion is one of the most exciting aspects in human and is very extensively researched in emotion psychology. Different stimuli happen to elicit different kinds of emotions in humans. Psychologists have also extensively studied color perception for their special relationship with emotions, and findings suggest that different colors also elicit different emotions [1, 4]. Some studies have suggested that emotion-arousal is related to the visual cortex [7, 6]. Therefore, we decided to study this emotion-color association using deep neural networks trained on a visual task.

First, we conducted a behavioral experiment of a forced-choice decision task in which subjects were asked to select a specific color for a given emotion-inducing image stimuli. We estimated decision probabilities using the responses that we got from this experiment. Next, we developed a computational model for this decision task using similarities between deep representation (extracted using DNNs trained on object classification tasks) of the stimuli images and images of colors. We then examined the relationship between the two decision probabilities using Pearson's correlation coefficient (R). We found that the representation learned by deep neural networks indeed captures some emotion-color association. The representation extracted from the 'fc2' layer of VGG16 showed

Submitted to 2nd Workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM).

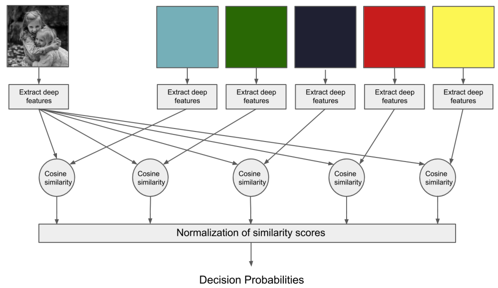

Figure 1: Modelling Decisions from Deep Representations

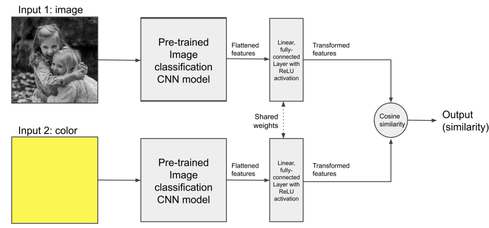

Figure 2: Feature Transformation

a fuzzy linear relationship. Similar to a study done by Peterson et al. [10], we tested our model after linearly transforming the raw representations to a smaller feature space. We found that the correlation score significantly improved, and the model showed a significant improvement in an emotion classification task compared to standard cross-entropy based classification model.

2 Behavioral Experiment

There are six basic emotions for which color is considered as a perceptual feature: Anger (red), disgust (green), fear (black), happiness (yellow), sadness (blue), and surprise (bright), which are the colors used in other studies on emotion-color association [1, 4, 12]. We only used the first five emotions as it was ambiguous to use any specific color for "bright".

Stimuli: Our stimulus set consisted of 50 grey-scaled images. These images were taken from an emotion data-set used by Machajdik and Hanbury [9] for affective image classification. The images were selected to include 10 images for each emotion. We converted the images to gray-scale so as to remove any bias because of dominant colors in the images themselves.

Participants: We distributed the experiment among students of the institute where this study was conducted. Their participation was completely voluntary, and none of them were forced to take part. A total of 56 different individuals completed the experiment.

Data analysis and procedure: The experiment was designed using jsPsych JavaScript library [2]. For an individual trial, a gray-scale image was shown along with the five colors. At the beginning of the experiment, participants were instructed to select a color that would best fit with the underlying emotion of the shown picture. To make sure that there was no bias, we added an additional instruction to each trial, "What color will you associate to this picture? Try to relate it to how the image makes you feel". After collecting the responses, we calculated histograms of chosen colors for each image stimuli. The normalised histogram was taken as the decision probabilities of choosing colors for an image.

3 Methods

Modelling Decisions from Deep Representations: We extracted features using the intermediate layers of the state of the art deep learning models trained on the imagenet dataset [3]. We refer to these extracted features as deep representations. We extracted these deep representations of stimuli images and color images by passing them through VGG16, DenseNet, ResNet, and MobileNet architectures for our study. Then we calculated cosine similarities between extracted representations of stimuli images and color images. Finally, to get overall decision probabilities from our model, we normalised the similarity scores for a given stimuli image among the five color images. So, our decision model outputs a probability of choosing a particular color for a given image (See Figure 1). To evaluate the correspondence between the model decision and human decisions, we calculated the Pearson's correlation coefficient (R) between the two decision probabilities.

Evaluating Transformed Representations: On similar lines as the methods used by Peterson et al. and Jha et al. of transforming deep representations to capture psychological representations of

Model	R_r	R_t
VGG16	0.33	0.63 ± 0.01
DenseNet169	0.29	0.57 ± 0.03
ResNet50	0.28	0.60 ± 0.02
MobileNet	0.26	0.53 ± 0.03

Table 1: R scores found using various pre-trained deep learning models. The second column (i.e. R_r) corresponds to the R score calculated using raw representations and the third column (i.e. R_t) corresponds to the R score calculated using transformed representations. For transformed representation, mean scores and standard deviations are reported over 50 independent runs.

Color Sequence	R_r	R_t
[0, 1, 2, 3, 4] (original seq.)	0.33	0.63 ±0.01
[4, 3, 0, 2, 1]	0.01	0.04
[2, 3, 1, 4, 0]	-0.09	-0.34
[2, 4, 3, 1, 0]	0.11	0.05
[1, 0, 4, 2, 3]	0.06	-0.03

Table 2: R score against wrong color labels. The second column (i.e. R_r) corresponds to the R score calculated using raw representations and the third column (i.e. R_t) corresponds to the R score calculated using transformed representations. The first row corresponds to original labels of the colors. Except the original sequence, other sequence were evaluated for a single iteration.

similarity judgement, we introduced a linear transform of the deep representations extracted using pre-trained models to a smaller number of features. And then, we performed the previous analysis as we did for raw representations on the transformed representations, as shown in Figure 2. The results are shown in Table 4, fourth column (R_t). These R scores are calculated using the similarity obtained on the validation set of five-fold cross validation method. Means and standard deviations are reported for 50 different independent runs.

Evaluating on Classification Task: We also evaluated this method for its ability to classify images into the five emotions. We considered two possibilities for true class labels, 1. As predicted by humans in the experiment (color chosen the most), 2. Class labels in the original dataset. We compared the following four methods: Raw similarity (color with the maximum similarity based on the 'fc2' layer of VGG16), Transformed similarity (color with the maximum similarity based on transformed representation), Standard classification model trained on cross entropy loss between model predictions and human predictions, and Standard classification model trained on cross entropy loss between model prediction and actual class labels. Results are shown in Table 3.

4 Results and Discussions

Results are shown in Table 4 (R_r is the correlation score evaluated using raw representation. R_t is the correlation score evaluated using transformed representations). The R scores on transformed representation reported here are calculated using the similarity obtained on the validation set of five fold cross-validation method. So, for each fold, we get 50 similarity scores corresponding to the validation set of that fold, comprising a total of 250 similarities for the overall run. Also, note that reported results are averaged over 50 independent runs (we have reported mean along with standard deviation). For raw representation VGG16 showed the best results with a correlation score of $R = 0.33$ with $pvalue < 0.0001$ (null hypothesis being zero correlation). While $R = 0.33$ indicates a moderate linear relationship between the model decisions and human decisions, the score is still small. So, before making any claims, we checked the R scores against wrong colored images, i.e., we changed labels of the colored images, so as to result in wrong similarity scores for image-color pairs. We found that this decreases the R score significantly. This supports the hypothesis that images associate with specific colored images. And the low R score could be attributed to the following reasons: 1. There's no straight association between color and emotion-inducing images or 2. The features extracted using VGG16 don't directly correspond to representations of emotions and need to be transformed to some other dimension, which could better associate with color and emotions.

We found that the R score significantly improved for the transformed representation for all the four models. Most importantly, VGG16 performed best (with $R = 0.63$ and $pvalue < 0.0001$) which is consistent with the evaluation done on psychological representation [10]. We also performed the analysis on wrong classes for transformed representation on VGG16. Interestingly, we found that the R-scores for wrong color labels were significantly low than the correct color labels. We also evaluated features extracted across different pooling layers of VGG16 to check if they produce similar trends as

Method	wrt Human Prediction	wrt Actual Class
Chance (averaged over 1000 runs)	7.9 ± 2.88	20.01 ± 4.44
Raw similarity model	40	24
Transformed similarity model	$\mathbf{56.12 \pm 3.21}$	$\mathbf{40.20 \pm 2.89}$
Standard model (trained on human prediction)	43.84 ± 4.70	32.60 ± 3.49
Standard model (trained on actual class)	31.80 ± 5.49	30.44 ± 4.74

Table 3: Accuracy with respect to human prediction and actual class labels.

with psychological representation [10] i.e., deeper layers better capture human judgement. We found that the results are indeed valid with the results of Peterson et al. on psychological representations.

Table 3 shows results for the classification tasks. Accuracy reported is average accuracy over 50 training trials. We were amazed to find that emotion classification using raw representations yields 40% classification accuracy on classes predicted by humans, which is way above chance (8% accuracy). It's also important to note that we did not explicitly train our model to classify to those specific emotions that humans predicted. This further validates our point that Deep Neural Networks are capable of capturing emotion-color association. The model's performance further increased after we linearly transformed the features, achieving 40.20% on actual classes and 56.12% on human predictions. In both of the cases, the similarity-based model (using transformed representation) performed better than the standard classification model. We also compared the maximum accuracy achieved by different models among the 50 trials. For the similarity based model (transformed representation), max accuracy achieved was 46% on actual class and 64% on human prediction. While the Standard classification model (trained on human prediction) achieved 40% accuracy on actual class and 52% accuracy on human prediction.

5 Conclusions

In this analysis, we show that representations learned by Deep Neural Networks are capable of capturing emotion-color association. Though comparing raw representations yielded a low correlation score, the representations show a greater generality and correlation to human decisions when linearly transformed. We also showed how we could use this overall method to train deep learning models for an emotion classification task. Our analysis answers an interesting question in Cognitive Sciences. The human emotion-color association is not random but could possibly be learned while performing other cognitive tasks. If not, wrong labeled colors should have shown comparable correlation scores for the transformed representation, as the network was exclusively trained to do that. But we see a big difference between the correlation scores of correctly labeled colors and wrong labeled ones. The method could also be very beneficial to the Machine Learning community on finding alternative ways to train deep learning models for classification problems, which could probably improve the performance when we have smaller dataset. However, a potential limitation of this would be that you will need to identify which alternate associative feature to use for a specific task; for example, we used colors for emotion classification. The study also needs more and more replication work on different datasets to validate the point for the generality of this method to study stimuli association in deep neural networks. We also see a great potential for this result and method for advancement in affective computing in developing artificial emotional intelligence and in emotion psychology.

References

[1] Roy D'Andrade and M. EGAN. The colors of emotion. *American Ethnologist*, 1:49 – 63, 10 2009. doi: 10.1525/ae.1974.1.1.02a00030.

[2] Joshua R. de Leeuw and Benjamin A. Motz. Psychophysics in a web browser? comparing response times collected with JavaScript and psychophysics toolbox in a visual search task. *Behavior Research Methods*, 48(1):1–12, March 2015. doi: 10.3758/s13428-015-0567-2. URL `https://doi.org/10.3758/s13428-015-0567-2`.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[4] Avery N. Gilbert, Alan J. Fridlund, and Laurie A. Lucchina. The color of emotion: A metric for implicit color associations. *Food Qual. Preference*, 52:203–210, Sep 2016. ISSN 0950-3293. doi: 10.1016/j.foodqual.2016.04.007.

[5] Aditi Jha, Joshua Peterson, and Thomas L. Griffiths. Extracting low-dimensional psychological representations from convolutional neural networks, 2020.

[6] Philip A. Kragel, Marianne C. Reddan, Kevin S. LaBar, and Tor D. Wager. Emotion schemas are embedded in the human visual system. *Science Advances*, 5(7), 2019. doi: 10.1126/sciadv.aaw4358. URL `https://advances.sciencemag.org/content/5/7/eaaw4358`.

[7] Peter J. Lang, Margaret M. Bradley, Jeffrey R. Fitzsimmons, Bruce N. Cuthbert, James D. Scott, Bradley Moulder, and Vijay Nangia. Emotional arousal and activation of the visual cortex: An fMRI analysis. *Psychophysiology*, 35(2):199–210, March 1998. doi: 10.1111/1469-8986.3520199. URL `https://doi.org/10.1111/1469-8986.3520199`.

[8] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. doi: 10.1038/nature14539. URL `https://doi.org/10.1038/nature14539`.

[9] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 83–92, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605589336. doi: 10.1145/1873951.1873965. URL `https://doi.org/10.1145/1873951.1873965`.

[10] Joshua C. Peterson, Joshua T. Abbott, and Thomas L. Griffiths. Adapting deep network features to capture psychological representations: An abridged report. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, Aug 2017. doi: 10.24963/ijcai.2017/697. URL `http://dx.doi.org/10.24963/ijcai.2017/697`.

[11] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? September 2018. doi: 10.1101/407007. URL `https://doi.org/10.1101/407007`.

[12] Tina M. Sutton and Jeanette Altarriba. Color associations to emotion and emotion-laden words: A collection of norms for stimulus construction and selection. *Behav. Res.*, 48(2):686–728, Jun 2016. ISSN 1554-3528. doi: 10.3758/s13428-015-0598-8.

Appendices

A. Model Details

A.1 Deep Feature Extraction

VGG16, DenseNet169, ResNet, and MobileNet are the four DNN models for which we have reported the results. We used the pre-trained weights provided by TensorFlow deep learning library. All the models were trained on the imagenet dataset to classify 1000 object categories. For our analysis, we mostly used the last layer of each model before the final classification layer to extract image features. The corresponding number of features and layer name available in TensorFlow model is shown in Table below:

Model	Layer Name	Number of Features
VGG16	fc2	4096
DenseNet169	avg_pool	1664
ResNet50	avg_pool	2048
MobileNet	reshape_2	1000

Table 4: Model and layer name as in tensorflow for the corresponding layers used to find the results shown in the main paper

A.2 Training Details for Transformed Representation

We first tested for various number of output features starting from 0 to 175 with a step size of 25. We found that the correlation score maximized around output features = 75. So, we used 75 numbers of output features for further analysis. We trained the weights for this linear layer using the similarity scores obtained from the behavioral experiment. We used L2 loss function between the human similarity and similarity predicted by the model. The model was evaluated using five fold cross-validation for its generalisation performance. Note that we have a total of 250 different similarity scores corresponding to 50 different stimuli images and 5 color images. For each cross-validation set, only 200 similarity pairs were used for training, and rest 50 were used for model evaluation. During training, we shuffled the 200 input data.

Training Parameters: adam optimiser with learning rate = 0.001, batch size = 10, and number of epochs = 30

A.3 Details for Classification Model

Raw similarity: No training involved; we predict the classes based on the color which gives maximum similarity to the input images based on the features extracted using the 'fc2' layer of VGG16.

Transformed similarity: We predict the classes based on the color which gives maximum similarity to the input images based on the transformed representation. We trained the model using five fold cross-validation, and the accuracy reported here is based on the label predicted using test cases 'only'. The training parameters were the same as shown in Appendice A.2

Standard classification (on human prediction): We replaced the last layer of VGG16 with a fully connected layer with 75 output units and 'relu' activation and then added one another layer to give five outputs and 'softmax' function to predict among 5 class labels. We trained the model on human prediction using categorical cross-entropy loss. The reported accuracy is for test cases only in a five fold cross-validation. Our training parameters were: adam optimiser with learning rate = 0.001, batch size = 10, and number of epochs = 15. (We trained it using fewer epochs compared to the similarity model, because this model converges faster than the similarity model. Even if we take epochs = 30, the results were not significantly different).

Standard classification (on actual classes): Similar to the previous one, but the model was trained using actual class labels.

B. Example from the Experiment Trial

Figure 3: An illustration of the trial from the behavioral experiment. Subjects were asked to select a single color from the five available options. Note that the stimuli image shown here is for illustration purpose which is free to use.

C. Correlation Score Vs VGG16 Layers

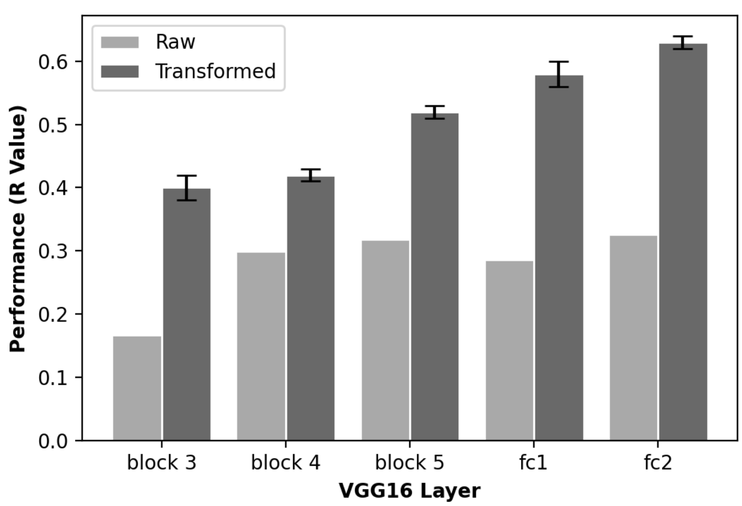

Figure 4: VGG 16 performance across different pooling layers. Bars shows the average accuracy over 10 trials and the error bars show the standard deviation. For ’fc2’ accuracy is averaged over 50 trials.