EmoCAM: Toward Understanding What Drives CNN-based Emotion Recognition

Youssef Doulfoukar*[1] **Laurent Mertens***[1,2] **Joost Vennekens**[1,2,3]

[1]KU Leuven, De Nayer Campus, Dept. of Computer Science
J.-P. De Nayerlaan 5, 2860 Sint-Katelijne-Waver, Belgium
[2] Leuven.AI - KU Leuven Institute for AI, 3000 Leuven, Belgium
[3]Flanders Make@KU Leuven, 3000 Leuven, Belgium
laurent.mertens@kuleuven.be
*Authors contributed equally

Abstract

Convolutional Neural Networks are particularly suited for image analysis tasks, such as Image Classification, Object Recognition or Image Segmentation. Like all Artificial Neural Networks, however, they are "black box" models, and suffer from poor explainability. This work is concerned with the specific downstream task of Emotion Recognition from images, and proposes a framework that combines CAM-based techniques with Object Detection on a corpus level to better understand on which image cues a particular model relies to assign a specific emotion to an image. We demonstrate our framework using the EmoNet model and show that it mostly focuses on human characteristics, but also explore the pronounced effect of specific image modifications.

1 Introduction

Thanks to recent progress, image analysis problems such as Object Detection using Artifical Neural Networks (ANN) can be more or less considered to be solved [20, 8]. However, higher-order tasks, such as identifying the emotion content of an entire image, remain more challenging. Convolutional Neural Network (CNN) models such as EmoNet [10] present a promising approach in this area, but its results are not yet completely convincing. This raises the question to which extent this network is actually picking up meaningful cues in the images, and to what extent it is learning spurious correlations that may be present in the private dataset on which it was trained.

ANNs are still considered "black box" models, and the domain that attempts to untangle how they make the predictions they make, i.e., to improve their *explainability*, is a very active one [17, 19, 1]. One of the techniques for this is Class Activation Mapping [21], or CAM, which allows to highlight those parts of the image that contributed most to a model's (say, a CNN image classifier) output. This technique allows to visually inspect individual images or videos, but does not immediately allow for an automated global analysis on a corpus level. To answer the earlier question of what image cues a CNN-based Emotion Recognition network, in casu EmoNet, most relies on, we propose EmoCAM. Our framework combines two information streams, namely CAM and Object Detection, to build a pipeline that allows to determine those object classes that most contributed to the model's decision making on a corpus level. Besides better understanding what object classes the model relies most upon, we also want to explore the potential of applying minor changes to the input images that steer the model towards a specific emotion by leveraging the obtained information from our EmoCAM analysis. Our source code can be found at `https://gitlab.com/EAVISE/lme/emocam`.

Workshop on Scientific Methods for Understanding Deep Learning, NeurIPS 2024.

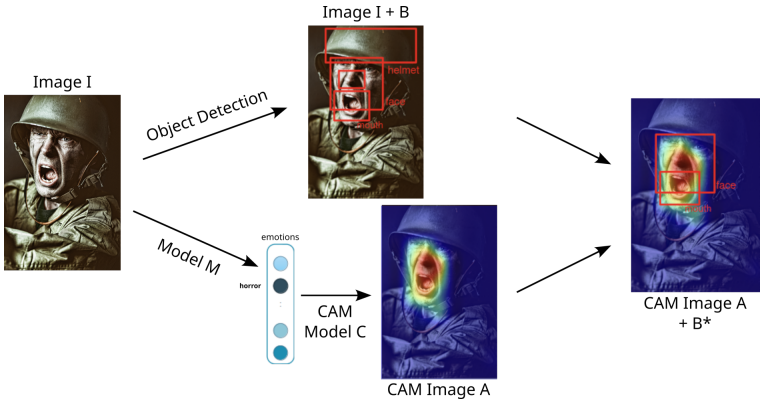

Figure 1: Schematic illustrating pipeline combining CAM with Object Detection.

The remainder of this paper is organized as follows: in Section 2 we describe our proposed framework in detail, followed by Section 3 where we look at a concrete case using the EmoNet network and FindingEmo [12] image dataset. Limitations and roads for future work are explored in Section 4 and we present concluding remarks in Section 5. Photo credits have been gathered in Appendix A.4.

2 Methodology

We start our analysis by applying, for a given CNN model M and corpus D, the following steps to each image $I \in D$, schematically illustrated in Fig. 1.

First, we process I with the object detection network of our choice, in casu, YOLOv3 [15] trained on the Open Images dataset.[1] We opted for this particular pretrained network as other popular Object Detection dataset choices such as PASCAL VOC (20 classes) and MS-COCO (80 classes) are too restricted in the classes they propose. By contrast, Open Images, which contains 601 classes, presents a nice balance between human-related classes (e.g., "human face", "mouth", etc.), and more general classes representing contextual elements (e.g., "car", "tree", etc.). The result of this operation is a list of detected objects and their corresponding bounding boxes B. We filter the YOLOv3 output by keeping only bounding boxes with an IoU score > 0.005.

Second, we process I with M, and apply a CAM-based technique C to the last convolutional layer of M.[2] This gives us an activation map that we overlay on top of I to obtain a new image A.

Finally, we lay the bounding boxes B on top of A, and look for those boxes $b \in B$ for which the average CAM activation, or *importance*, $C_{Act} > 0.3$, with C_{Act} defined as the sum of the CAM activations within the box divided by the area of the box.[3] The threshold was heuristically determined by visually inspecting a limited set of images. We refer to these boxes as the set B^*, and interpret these as those objects that most contributed to the model's decision.

Once we have found B^* for every image in our corpus, we then analyse these data to find associations between object classes and output labels by constructing an association matrix M_A, where $a_{ij} \in M_A$ represents the number of images labeled with the j^{th} EmoNet emotion label in which the i^{th} object class has been detected at least once. By dividing each column j through by the total number of images labeled with the j^{th} emotion such as to obtain percentages (after doing $\times 100$), we obtain M'_A which allows to ignore imbalances in the prediction rates of the different emotions.

[1] We use the PyTorch LightNet [13] implementation, and OpenImages weights available from `https://pjreddie.com/darknet/yolo/`.

[2] We tried (combinations of) other layers, but the best results were obtained using only the last layer.

[3] We take the activations from the original grayscale CAM output, not from the colorized version.

3 Results

We tested our proposed approach using the EmoNet model and the FindingEmo dataset. EmoNet is a model obtained through replacing the last layer of an AlexNet model pretrained on the ImageNet [3] corpus. This last layer was then trained on a private dataset of 137,482 images annotated for the emotion they evoke in the observer with one of 20 custom emotion labels. We use the Python port by L. Mertens [11] of the original Matlab release. FindingEmo is an image dataset consisting of 25,869 images annotated for, a.o., the dominant emotion in the picture, using one of the 24 emotion labels in Plutchik's Wheel of Emotions [14]. All images represent multiple people in various natural settings and with varying degrees of interaction among them. We present detailed results for Grad-CAM [18] in Section 3.1. An exploration of the effect of using other CAM-based methods can be found in Appendix A.2.[4] Finally, we briefly explore the effect on the predicted label of artificially adding certain objects to images, attempting to answer the question whether the presence of certain objects can *cause* a specific label to be predicted.

3.1 Results for Grad-CAM

A heatmap depicting M'_A as obtained using Grad-CAM together with EmoNet applied on the FindingEmo corpus can be found in Fig. 2. We limit ourselves to the 25 most prominent Open Images classes (as determined by the average of the corresponding row in M'_A). A clear conclusion to be drawn from this graph is that human features do indeed contribute the most to the decision making, most particularly the human face which, except for "Clothing", represents the most important class for each EmoNet label.

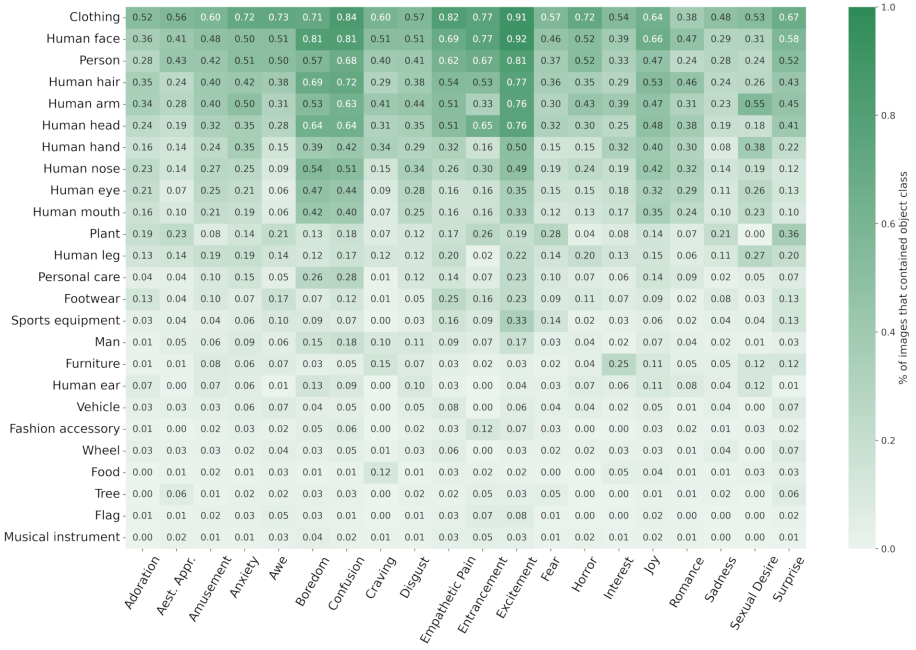

Figure 2: Association between Open Images classes and predicted EmoNet label. Heatmap entries represent the percentage of images labeled with a certain EmoNet label for which at least one object of the corresponding Open Images class was detected with high enough importance. "Aest. Appr." = Aesthetic Appreciation.

Additionally, some more specific associations do manifest themselves. Clear examples are the association between "Sports equipment" and "Excitement", and "Food" and "Craving", both of which seem logical. Less clear is, e.g., the association between "Furniture" and "Interest", or "Plant" and "Surprise", which hint of spurious associations resulting from biases in either or both the EmoNet training dataset and FindingEmo.

[4]For the CAM analysis, We use the Python packages `grad-cam` [6] and `captum` [9].

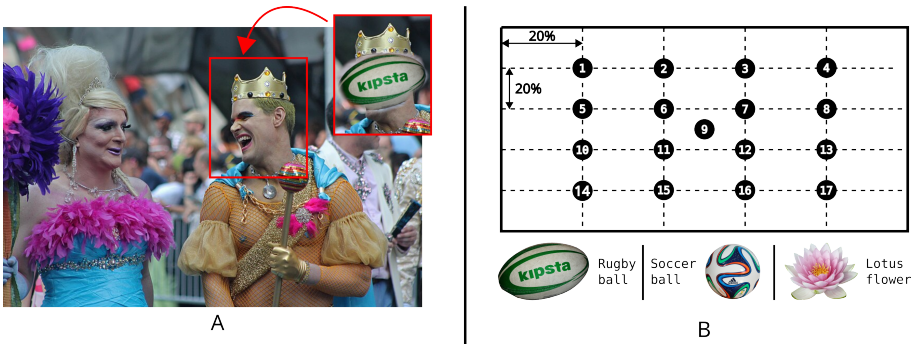

Figure 3: Fig. A: Adversarial example. Original image labeled by EmoNet as 92.9% "Joy". Modified part in upper right box; modified image labeled as 66.1% "Excitement". Fig. B: Schematic illustration of "paste object in image" experiment. The grid illustrates the relative positions within the image the objects are pasted and centered at, with the considered objects shown at the bottom.

3.2 Prediction Stability

To illustrate how the obtained knowledge can be applied to create an adversarial attack, consider the image shown in Fig. 3.A. We know from Section 3.1 that there is a high association between the object category "Sports equipment" and EmoNet label "Excitement". This inspired us to take an image labeled with high probabilty as "Joy" (92.9%; Excitement: 1.8%). After altering this image by pasting a rugby ball on top of the head of one of the two main subjects, the prediction changes to 66.1% "Excitement" ("Joy": 30.2%), demonstrating the dramatic effect the presence of a particular object can have on the model's output. Note that the position of the pasted object greatly influences the effect it has. Moving the rugby ball to the immediate right of the subject's face alters the predictions to 43.9% "Joy" and 42.1% "Excitement", while moving it to the immediate left only alters the predictions by 4% in the same directions ("Joy": 34.2%; "Excitement": 62.1%). Covering the other subject's head instead, we obtain 52.0% "Joy" and 30.7% Excitement.

We further investigate this effect by performing the following experiment. For each I in D, we paste a given object $O \in \{\text{Rugby ball, Soccer ball, Lotus flower}\}$, resized such that its height equals $0.2 \times \text{height}(I)$, in I centered at each one of a set of predefined relative positions P within the image, resulting in $size(P)$ alterations to I. The positions and objects considered are illustred in Fig. 3.B. We then send the altered images through EmoNet, and observe how the prediction was affected. Finally, we determine for what percentage of images the predicted label changed, and determine the label most often switched to.

The results are shown in Appendix A.3, and summarized here. The experiment confirms that the model shows high sensitivity to certain objects. Although differences of up to more than 10% can be observed between positions, specifically for the rugby ball, no real tendencies reveal themselves. In combination with the example in Fig. 3.A, we hypothesize the differences are not so much due to the *absolute* position of the object, but to what it *occludes*. For the rugby ball, 4 out of 17 positions resulted most often in a label switch to "Excitement". For the soccer ball, the number increases to 7. The Lotus flower clearly results in much less label shifts overall, with not a single position favoring "Excitement", confirming the importance of the object class in effecting a label switch.

4 Limitations and Future Work

Although the currently described approach already provides valuable insights, some limitations are to be noted.

First, the approach is, by definition, heavily dependent on the choice of Object Detection network and its corresponding classes and performance. The upside is that, as a plug-and-play component, different Object Detection networks can be chosen for different tasks, allowing to pick object classes tailored to the task at hand.

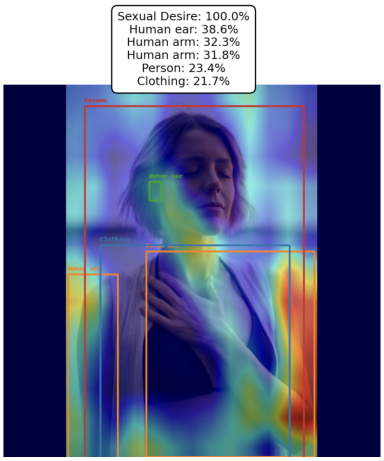

Figure 4: An instance were the current EmoCAM approach fails to detect the most important class.

Second, our current implementation does not take into account the size of the bounding boxes, which can result in suboptimal results. Consider, e.g., the example shown in Fig. 4. Although the subject's ear is clearly not the most important contributing element in the picture, because of the small size of the "Human ear" bounding box the average CAM activation is nonetheless the highest, spuriously pushing this object class to the top. Two main paths could be explored to counter this issue. The most straightforward would be to develop a scoring function that does take into account the bounding box size, or the activation distribution within it. Alternatively, segmentation models could be used instead of bounding box detection models, so as to obtain clearly delineated zones representing the different objects. Then of course, our first limitation still applies, i.e., the segmentation classes need to be relevant to the task at hand.

Third, with regard to our experiment described in Section 3.2, a more interesting approach might be to, instead of, or along with, considering only a fixed set of positions, paste the object at the center of bounding boxes relating to specific features such as human heads, thus investigating the effect of masking specific objects. We also intend to apply EmoCAM to the modified images to explore if the shift in label is reflected in a shift in focus in the modified image.

5 Conclusion

We propose the novel EmoCAM approach to explaining CNN decisions, specifically with the downstream task of Emotion Recognition from images in mind. Our objective is threefold: 1) better understanding what parts of the input image the model uses to make its decision, 2) allowing to check whether or not the information used by the model aligns with expectations from a human perspective, and 3) uncovering potential model biases. We have demonstrated our approach using the EmoNet model, FindingEmo dataset and multiple CAM techniques. Using our approach, we found that EmoNet indeed shows a strong focus on human elements, most notably (parts of) the human face, which is encouraging as it aligns with our understanding of human emotion recognition from Psychology. Nevertheless, we also found the model output to be quite unstable, in that adding specific objects (e.g., a rugby ball) to an image can dramatically alter its output and steer it towards a specific target emotion (e.g., "Excitement").

References

[1] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J.M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., Herrera, F.: Explainable Artificial Intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. Information Fusion **99**, 101805 (2023). https://doi.org/https://doi.org/10.1016/j.inffus.2023.101805, https://www.sciencedirect.com/science/article/pii/S1566253523001148

[2] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE **10** (2015), https://api.semanticscholar.org/CorpusID:9327892

[3] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)

[4] Desai, S., Ramaswamy, H.G.: Ablation-CAM: Visual explanations for deep convolutional network via gradient-free localization. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 972–980 (2020). https://doi.org/10.1109/WACV45572.2020.9093360

[5] Dimsdale-Zucker, H.R., Charan, R.: Chapter 27 - representational similarity analyses: A practical guide for functional mri applications. In: Manahan-Vaughan, D. (ed.) Handbook of in Vivo Neural Plasticity Techniques, Handbook of Behavioral Neuroscience, vol. 28, pp. 509–525. Elsevier (2018). https://doi.org/https://doi.org/10.1016/B978-0-12-812028-6.00027-6, https://www.sciencedirect.com/science/article/pii/B9780128120286000276

[6] Gildenblat J. and contributors: Pytorch library for CAM methods. https://github.com/jacobgil/pytorch-grad-cam (2021)

[7] Jung, H., Oh, Y.: Towards better explanations of class activation mapping. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1316–1324 (2021). https://doi.org/10.1109/ICCV48922.2021.00137

[8] Kaur, R., Singh, S.: A comprehensive review of object detection with deep learning. Digital Signal Processing **132**, 103812 (2023). https://doi.org/https://doi.org/10.1016/j.dsp.2022.103812, https://www.sciencedirect.com/science/article/pii/S1051200422004298

[9] Kokhlikyan N., Miglani V. et al.: Captum. https://pypi.org/project/captum/ (2019)

[10] Kragel, P.A., Reddan, M.C., LaBar, K.S., Wager, T.D.: Emotion schemas are embedded in the human visual system. Science Advances **5**(7), eaaw4358 (2019)

[11] Mertens, L.: EmoNet: A pytorch port. https://gitlab.com/EAVISE/lme/emonet (2022)

[12] Mertens, L., Yargholi, E., Op de Beeck, H., Van den Stock, J., Vennekens, J.: FindingEmo: An image dataset for emotion recognition in the wild (2024), accepted at the Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track

[13] Ophoff, T.: Lightnet: Building blocks to recreate darknet networks in pytorch. https://gitlab.com/EAVISE/lightnet (2018)

[14] Plutchik, R.: Chapter 1 - a general psychoevolutionary theory of emotion. In: Plutchik, R., Kellerman, H. (eds.) Theories of Emotion, pp. 3–33. Academic Press (1980). https://doi.org/https://doi.org/10.1016/B978-0-12-558701-3.50007-7, https://www.sciencedirect.com/science/article/pii/B9780125587013500077

[15] Redmon, J., Farhadi, A.: YOLOv3: An incremental improvement. CoRR **abs/1804.02767** (2018), http://arxiv.org/abs/1804.02767

[16] Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 1135–1144. KDD '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939778, https://doi.org/10.1145/2939672.2939778

[17] Saeed, W., Omlin, C.: Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. Knowledge-Based Systems **263**, 110273 (2023). https://doi.org/https://doi.org/10.1016/j.knosys.2023.110273, https://www.sciencedirect.com/science/article/pii/S0950705123000230

[18] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-Cam: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 618–626 (2017). https://doi.org/10.1109/ICCV.2017.74

[19] Yang, W., Wei, Y., Wei, H., Chen, Y., Huang, G., Li, X., Li, R., Yao, N., Wang, X., Gu, X., Amin, M.B., Kang, B.: Survey on Explainable AI: From approaches, limitations and applications aspects. Human-Centric Intelligent Systems **3**, 161 – 188 (2023), `https://api.semanticscholar.org/CorpusID:260817864`

[20] Zhao, Z.Q., Zheng, P., Xu, S.T., Wu, X.: Object detection with deep learning: A review. IEEE Transactions on Neural Networks and Learning Systems **30**(11), 3212–3232 (2019). https://doi.org/10.1109/TNNLS.2018.2876865

[21] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2921–2929 (2016). https://doi.org/10.1109/CVPR.2016.319

A Appendix

A.1 Preliminaries

In this section we provide brief introductions to the Image Classification and Object Detection tasks, and the CAM visualization technique, which form the basis for the current manuscript.

Figure 5: Illustration of Image Classification (A), Object Detection (B) and Class Activation Mapping (C). In (A), a (hypothetical) model has assigned the label "Flower" to the image. In B, a (hypothetical) Object Detection model has detected one instance of "Flower" and "Bee". In C, a (hypothetical) CAM output is superimposed on A, showing what parts of the image contributed to the model assigning the label "Flower". The redder the pixel, the more it contributed.

A.1.1 Image Classification

The task of Image Classification consists of training a network to assign a class—or label—c out of a set C of possible classes to an input image I. Consider, e.g., Fig. 5.A, showing an image for which a (hypothetical) image classifier determined the label "Flower" to be the most likely. In our case, the interest lies with emotion labels instead of object labels, but the principle remains the same.

A.1.2 Object Detection

The task of Object Detection is related to Image Classification, but instead of assigning a label to the entire image, the goal is to detect all instances of a fixed set of classes C that are represented in the image. A way of doing so is to determine the bounding box of each object, and assign the correct label to it. Fig. 5.B shows the same image as Fig. 5.A, but this time a (hypothetical) Object Detection network trained to recognize, a.o., objects of class "Flower" and "Bee" has detected one instance of each class.

A.1.3 Class Activation Mapping

A general problem with ANNs is that they are not explainable, i.e., it is not clear how they came to make a particular prediction. CAM, originally introduced by Zhou et al. [21], is one technique that attempts to answer this particular question for the task of Image Classification. The original technique works only for CNNs that contain a Global Average Pooling (GAP) layer at the tail of the sequence of convolutional layers. The GAP layer converts each filter in the preceding convolutional layer to the average activation of its features, essentially converting a collection of N filters to an N-dimensional vector V, which is then typically fed to a linear output layer O of size $|C|$ to obtain the output probabilities for each class. The idea behind CAM is to use the weights connecting V to O to compute a weighted sum of the filters corresponding to the entries in V to obtain a weighted average filter, called the "Class Activation Map", that encodes the importance of each filter to the model's output. This (1-channel) Class Activation Map is then upscaled to the size of I, and typically converted to an RGB image that is superimposed on I. An illustration is shown in Fig. 5.C. This technique was further generalized in Grad-CAM [18] to allow usage with models that initially do not use a GAP layer, and with any task, by using the gradients flowing back into the last convolutional layer as basis for the weights assigned to the filters. Subsequent techniques mentioned later in this paper essentially only differ in how they define these weights.

A.2 Comparison of CAM Methods

To answer the question to what extent different CAM methods yield different results, we performed a Representational Similarity Analysis [5] as follows. For each CAM method $C \in \{$Grad-CAM, Ablation-CAM[4], LIME[16], LRP[2], LIFT-CAM[7]$\}$, we determine the association matrix M_A with all Open Images classes as described in Section 2, keeping the same emotion and class ordering for each C[5]. We then flatten each matrix by concatenating all rows, turning it in to a 1D vector V_{M_C}. Finally, we construct a matrix R where each entry $R_{CC'}$ represents the Spearman Correlation rank between V_{M_C} and $V_{M_{C'}}$. The resulting matrix is shown in Fig. 6. All related p-values were $\ll 0.05$, indicating statistical significance.

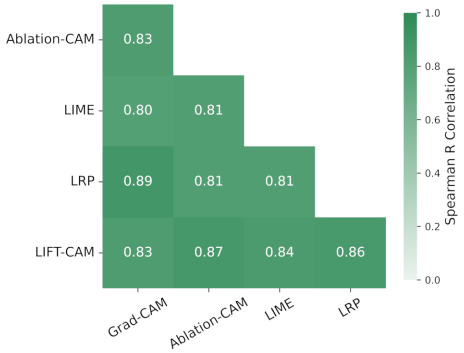

Figure 6: RSA analysis of different CAM methods.

The consistently high correlation values between all pairs indicate that variations in results obtained through different CAM methods can be expected to be minimal. We did observe both LIFT-CAM and LRP resulting in a notable association between "Pillow" and "Sexual Desire". Other than this, the differences between the methods appear to lie within the relative strengths of the associations observed, rather than the assocations themselves.

A.3 Results of ObjectPaste experiment

The barplot showing the results of the experiment described in Section 3.2 can be found in Fig. 7.

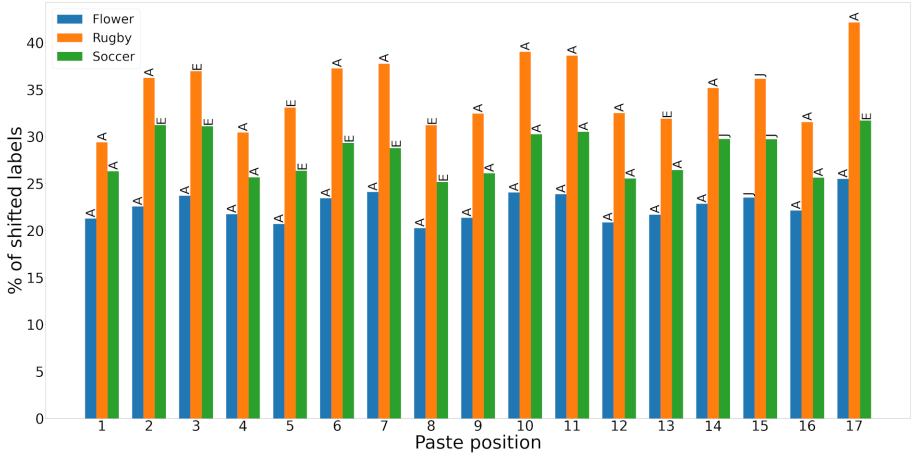

Figure 7: Percentage of images in the corpus whose predicted label changed when a specific object was pasted at a specific position. Refer to Fig. 3.B for the positions and objects. The letters 'A' for 'Admiration', 'E' for 'Excitement' and 'J' for 'Joy' atop each bar indicate the EmoNet label most often switched to.

[5]Which ordering is used does not matter, as long as it is consistently used.

A.4 Photo Credits

Following is the list of credits for the images used in the figures.

- Fig. 1: Photo by Sander Sammy on Unsplash, https://tinyurl.com/2tc69hfy, Unsplash license.
- Fig. 3: A: original photo by istolethetv, https://tinyurl.com/3r86e3w2, CC 2.0 license; Rugby ball by Peter Griffin, https://tinyurl.com/dmb77rks, CC0 license. B: Soccer ball by Jean Schecter, https://tinyurl.com/3xxkkysb, CC 4.0 BY-NC; Lotus flower at https://pngimg.com/image/69752, CC 4.0 BY-NC.
- Fig. 5: Photo by one of the authors.
- Fig. 4: Original photo by Darius Bashar on Unsplash, https://tinyurl.com/4eephaxb, Unsplash license.