IMAGE EMOTION RECOGNITION USING COGNITIVE CONTEXTUAL SUMMARIZATION FRAMEWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating the perceived emotion to visual stimuli has gained significant traction in the recent years. The existing frameworks rely either on a person's presence in the image or are based on object feature extraction and low-level image features. By focusing on the person/object in the image, the existing frameworks fail to capture the context or the interaction between multiple elements in the image. Also, what if an image does not have a human subject or an object? We address this drawback by building a Cognitive Contextual Summarization (CCS) model based on an One-For-All (OFA) backbone trained on multiple tasks, including image captioning. The ability of the backbone to recognize elements in the image and generate captions helps us capture interactions through captions, which we decode using BERT for contextual understanding. The end-to-end fusion of the OFA and the BERT features helps us predict continuous human emotion (Valence, Arousal) from an image. We train our framework on the Building Emotional Machines dataset in the literature, and the experiments show that our model outperforms the State-of-the-art.

1 INTRODUCTION

With the penetration of modern devices, humans interact more with the computers than with one another. These interactions are largely influenced by emotions, which are evoked as the first response to any content being viewed (Joshi et al., 2011; Lang, 1979; Lang et al., 1998). The paradigm of affective computing has evolved for proactive estimation of emotion from a visual stimuli. In the literature, emotion models/datasets are of two types viz. categorical and dimensional. Categorical emotion models consider few basic emotion categories such as *happy*, *fear*, *sad*, etc. as given in Ekman (1992) and Mikels et al. (2005b) emotion models. The dimensional emotion models represent emotions in the 2D space of Valence-Arousal (VA), where valence measures the amount of pleasantness and unpleasantness and arousal is the intensity of the emotion from being very passive to very active. Human emotions are complex and not discrete, and hence, it is realistic to measure emotions in a continuous dimensional space (Posner et al., 2005).

Assigning a continuous score to the emotion evoked or elicited in the observer by an image is a very complex problem. In order to estimate emotion one must take into account all elements in the image which in turn depends on how well the different elements (foreground, background, objects, etc.) in the image are recognized. As opposed to this, emotion recognition of natural language is comparatively easier (Acheampong et al., 2020). In literature, image emotion estimation has been addressed by Kim et al. (2018). They propose a dataset wherein each image has a corresponding valence and arousal value. They also propose a framework to estimate image emotions by considering the object and background features in the image. Although the work is very promising the framework for estimating image emotion depends only on the presence and detection of objects in the image which limits its application. In order to rectify this, we propose a novel emotion prediction framework which takes an entire image as a input without the need for additional guidance/annotation and predicts the valence/arousal of the whole image. In our experiments we use the Building Emotional Machines (BE) dataset (Kim et al., 2018) in the literature. We leverage the recent advancements in cross-modal deep learning by exploiting both the image feature space and the language feature space (Wang et al., 2022). Our framework is inspired by the fact that contextual interpretation/perception of the image information during cognition leads to the elicited emotion (Greenaway et al., 2018). Therefore, the proposed framework utilizes the image captioning network called One-For-All (OFA)

- a Task-Agnostic and Modality-Agnostic framework (Wang et al., 2022) trained on multiple tasks, and the Bidirectional Encoder Representations from Transformers (BERT) model for context interpretation (Devlin et al., 2018) to predict the emotions. Image caption networks (ICN) are superior to object detectors because image captions not only capture the objects but also the relationship between the objects present in it. Therefore, resulting in a better contextual understanding.

To summarise, in this paper, we have described the following contributions 1) Building a continuous emotion estimation framework for images in-the-wild i.e. with no requirement of additional information (eg. Person detection, Object detection, etc.) 2) We propose a model composed of OFA (Wang et al., 2022) and BERT. OFA summarizes the visual information in the image, which is passed to BERT for contextual understanding (Devlin et al., 2018) to predict the evoked emotion. The Image Captioning OFA backbone not only captures objects in an image (visual cognition) but also captures the relationship between these objects (contextual interpretation/perception of the image). Hence, this model has been named Cognitive Contextual Summarization Model. We train and test our framework on BE dataset and report a new benchmark for the image emotion estimation task.

In Figure 1, images in a row have the same valence, which indicates that images in the same row evoke similar degree of pleasantness or unpleasantness in the mind of the viewer. Within a row the arousal of the images gradually increases from the left to right. What this means is the left most images are indicative of extreme passiveness while the right most images are indicative of extreme activity or intensity. In between there is a gradual transition. As one moves from the top row to the bottom row the valence increases, indicating that the pleasantness evoked in the viewer's mind by the images increases. In Figure 2, images in a row have the same arousal, which indicates that images in the same row evoke similar extents of passiveness or activity in the viewer's mind. However, within a row the valence of the images gradually increases from left to right. What this means is the left most images are extremely unpleasant while the right most images are very pleasant to look at. As one moves from the top row to the bottom row the arousal increases, indicating that the activity or the intensity evoked by the image rises.

2 Related Work

2.1 Building Emotional Machines (BE) Dataset

Several efforts in the past to collect emotion estimation datasets have been made (Kosti et al., 2017; Mikels et al., 2005a; Lang et al., 1997; 2005; Kim et al., 2018). Table 1 describes all the existing datasets used for determining image emotion on a continuous scale i.e. Valence and Arousal. As observed, BE is both large in scale and is a social dataset i.e. the images in the BE dataset are collated from social media. For annotation, large number of annotators annotate each image with a Valence and Arousal score on a scale of 1-9. The score for each image is a mean of scores across annotators. Figures 1 and 2 show sample images from the BE dataset with corresponding Valence and Arousal ground truth. As observed from the images not all images have human subjects and objects and hence there is a need for an image emotion estimation framework which could cater to all kinds of images. It is worth noting the variability of images in the same scale of Valence or Arousal values.

Given the variation of images in a specific range, the complexity of image emotion prediction becomes evident. Given the nature (image variations) and scale (no. of image and annotators) of BE dataset, we test and train our framework on the BE dataset.

Dataset	**Images**	**Annotators**	**Type**
IAPS (Lang et al., 2005)	1182	100	Natural
Gaped (Lang et al., 1997)	730	60	Natural
NAPS (Marchewka et al., 2014)	1356	204	Natural
BE (Kim et al., 2018)	10766	1339	Social

Table 1: Widely used Image Emotion Datasets and their characteristics

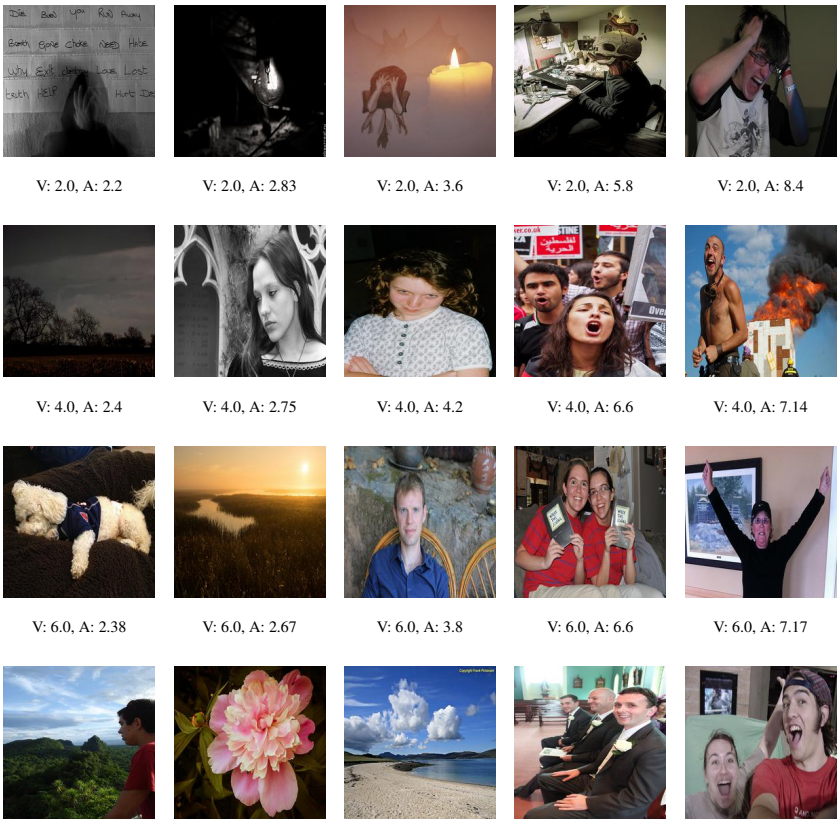


Figure 1: Images from the BE dataset (Kim et al., 2018) in the ascending order of the Arousal (left to right) in each row. Images in a row have the same Valence but different Arousal. V and A denote Valence and Arousal ground truths respectively.

2.2 Emotion Estimation Frameworks

Human emotions studied through physiological signals (Subramanian et al., 2018) are invasive and fail to achieve the scale and diversity of modern requirements. Therefore, the studies have moved towards crowd sourcing and non-invasive ground truth collection to understand the mapping between human emotions and visual stimuli.

Emotions have been estimated in the past using color statistics, texture and also some mid level features (Kim et al., 2018) like Scale Invariant Feature Transform (SIFT) (Lowe, 2004), bag of words, etc. These models failed to capture the complexity of the emotions from the images and hence, deep learning methods gained significant traction. Most of the classical affect recognition methods focus on facial expressions (Kollias & Zafeiriou, 2018; Khaireddin & Chen, 2021) and sometimes leverage additional cues, such as body pose (Schindler et al., 2008).

All of these aforementioned works focus only on images that contain people, which is a small subset of all possible images, and thus fail to account for diversity. Therefore, in this paper we introduce a framework which works for any type of image and does not mandate the presence of any type of content in the image for emotion recognition.

3 Cognitive Contextual Summarization Framework

The Cognitive Contextual Summarization (CCS) framework is a fusion network with two branches which regresses over an input image to output a valence or arousal score. Both the branches are built upon a common One-For-All (OFA) encoder (Wang et al., 2022). Image embeddings in the first

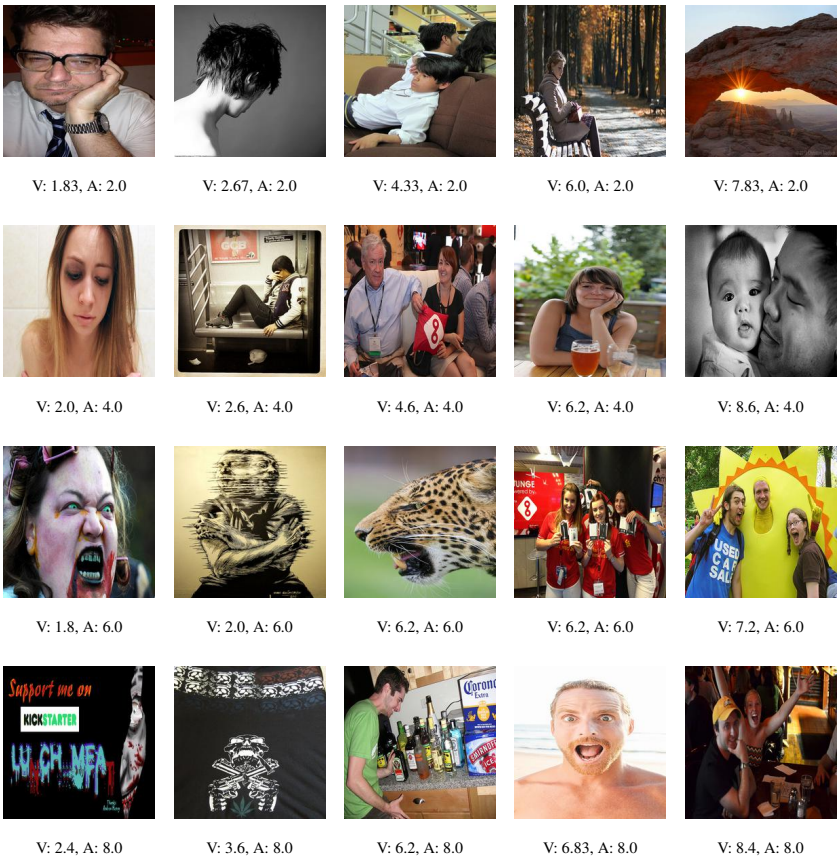


Figure 2: Images from the BE dataset (Kim et al., 2018) in ascending order of Valence (left to right) in each row. Images in a row have the same Arousal but different Valence. V and A denote Valence and Arousal ground truths respectively.

branch are obtained from the OFA encoder, while in the other branch the OFA decoder, followed by the BERT, are stacked to get the image caption embeddings of the input image. The weighted fusion of both the embeddings viz. OFA encoder embedding and the BERT (image caption) embedding is then regressed to obtain the valence/arousal scores for the input image. To better understand the framework, we first briefly review OFA and the significance of BERT in the emotional recognition context and then introduce CCS in detail.

3.1 OFA

OFA is a sequence-to-sequence framework which is task-agnostic and modality-agnostic. OFA can be used for vision-only, language-only, and vision and language tasks. It can achieve this by projecting data of various modalities to a unified space, i.e. by discretizing the text, image and object and representing them with tokens in a unified vocabulary. OFA has an encoder-decoder architecture where both encoders and decoders are a stack of transformer layers. A transformer layer in the encoder consists of self-attention and a feed-forward network (FFN). On the other hand, a transformer layer in the decoder comprises a self-attention, a FFN and cross-attention for building the connection between the decoder and the encoder output representations. OFA is pre-trained for cross-modal tasks, including visual grounding, grounded captioning, image-text matching, image captioning, and visual question answering. The task-specific pre-training equips OFA with the ability to localize different components in the image for the corresponding natural language input and vice versa. This training paradigm helps OFA capture local feature signatures across an image, and the finetuning on tasks like image captioning helps the model learn the relationship between different visual components, i.e. summarizing the context. Hence, the output from the OFA encoder forms one branch in our CCS framework.

3.2 BERT - CAPTURING LANGUAGE CONTEXT

As discussed above, the OFA pre-training incorporates vision-language pairs. Given an input image for a VQA task, OFA's encoder and decoder help map the vision features to the text features and generate captions. These captions are generated with the help of a unified vocabulary. Since the generated captions come from the same vocabulary used for images, the OFA decoder's feature representation is an extension of the OFA encoder embeddings. Hence, to capture the language context, we do not directly extract OFA decoder features but instead stack BERT after the OFA decoder, which takes the caption as input and gives feature embeddings for the input caption. Given the huge corpus of language-only data that the BERT is trained on, we expect it to capture language context better. The same is validated through our experiments in the ablation study section.

3.3 CCS ARCHITECTURE

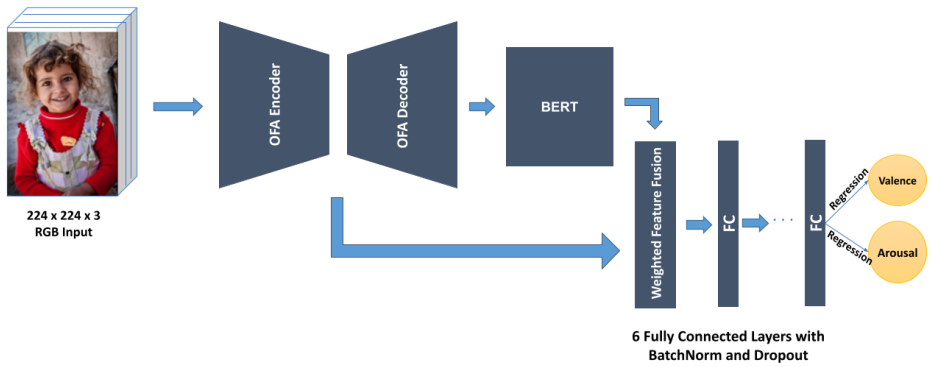


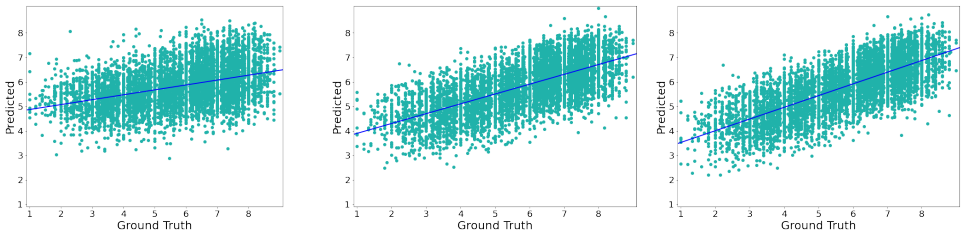
Figure 3: Cognitive Contextual Summarization (CCS) Model

Figure 3 shows the architecture of the Cognitive Contextual Summarization model. At first an image is sent as input to the OFA network. The caption generated by the OFA network is further passed into BERT and the embedding obtained from BERT which is of size 768x1 is the output of the first branch as shown in 3. At the caption generation step, the modality changes from image to text and hence, there could be some loss of information. Therefore, to capture image features the encoding obtained from the last layer of the last transformer in the OFA encoder which is of size 968x1024 is extracted. This is the output of the second branch as shown in 3. We weight the fusion of the outputs from both the branches to assign weightage to the contribution made by each branch. The weighted fusion results into a 1D vector which is then fed to Fully-Connected (FC) layers. After each FC layer we stack BatchNorm layers. To avoid overfitting we also introduce two dropout layers each after the 2nd FC layer and the 4th FC layer. The choice of the number of FC layers has been justified in an ablation study in the supplementary material.

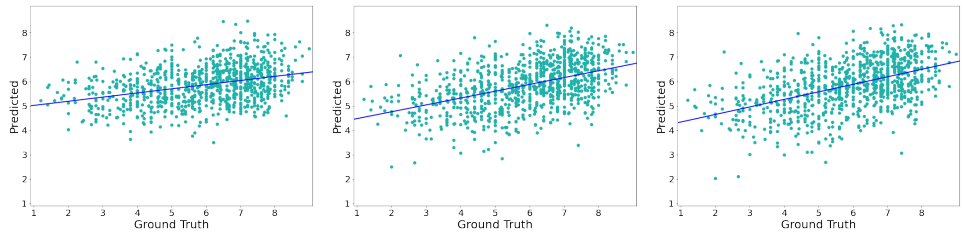
4 EXPERIMENTS & ABLATION STUDY

To validate our CCS architecture's ability to learn the contextual understanding of the image and hence, predict emotions has been verified by extensive experimentation carried out on the BE dataset. At first, we evaluate the BE model by training and testing on BE dataset. Figure 4 shows the scatter plots of ground truth and predicted valence values across training (4(a), 4(b) & 4(c)) and validation (4(d), 4(e) & 4(f)) epochs. It is clearly seen from the scatter plots and the correlation coefficients between the ground truth and the predicted values that the BE model moderately learns valence representations across images. In other words, it is able to predict the valence of the image fairly or

in other words the goodness of the linear fit (learning ability) increases across epochs while training and the model performs comparatively less during validation.

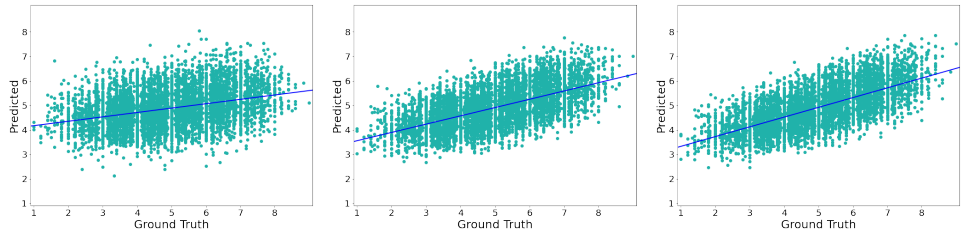


(a) Epoch 1 - Corr. Coeff.: 0.41 (b) Epoch 20 - Corr. Coeff.: 0.68 (c) Epoch 40 - Corr. Coeff.: 0.75

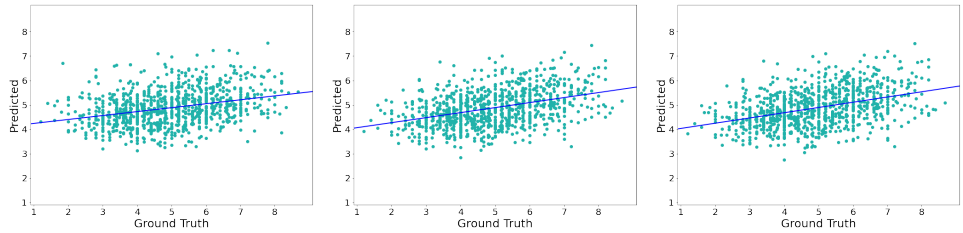


(d) Epoch 1 - Corr. Coeff.: 0.36 (e) Epoch 20 - Corr. Coeff.: 0.47 (f) Epoch 40 - Corr. Coeff.: 0.48

Figure 4: Scatterplot of Ground Truth and Predicted valence values for BE model trained on BE dataset. Top-row: Training plots. Bottom-row: Validation plots



(a) Epoch 1 - Corr. Coeff.: 0.35 (b) Epoch 20 - Corr. Coeff.: 0.62 (c) Epoch 40 - Corr. Coeff.: 0.69

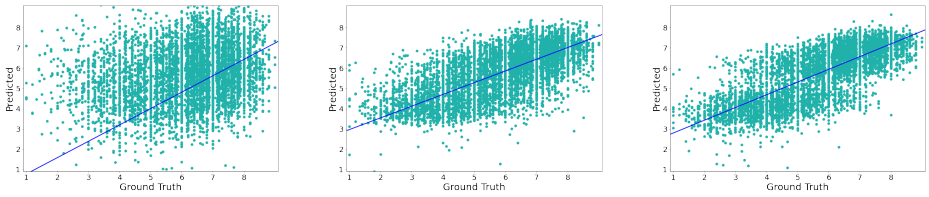


(d) Epoch 1 - Corr. Coeff.: 0.31 (e) Epoch 20 - Corr. Coeff.: 0.39 (f) Epoch 40 - Corr. Coeff.: 0.40

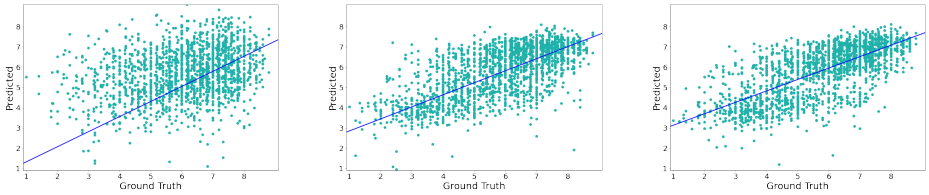
Figure 5: Scatterplot of Ground Truth and Predicted arousal values for BE model trained on BE dataset. Top-row: Training plots. Bottom-row: Validation plots

Similarly, Figure 5 shows the scatter plots of ground truth and predicted arousal values across training (5(a), 5(b) & 5(c)) and validation (5(d), 5(e) & 5(f)) epochs. It is clearly seen from the scatter plots and the correlation coefficients between the ground truth and predicted values that the BE model learns the Arousal representation very minimally and the validation scatter plots depict poor performance.

We then train our CCS model and test it on BE dataset. Figure 6 shows the scatter plots of ground truth and predicted valence values across training (6(a), 6(b) & 6(c)) and validation (6(d), 6(e) & 6(f)) epochs. It is clearly seen that the CCS model learns the valence representation across images. In other words, it is able to predict the valence of the image very well and the goodness of the linear fit (learning ability) increases across epochs during both training and validation.



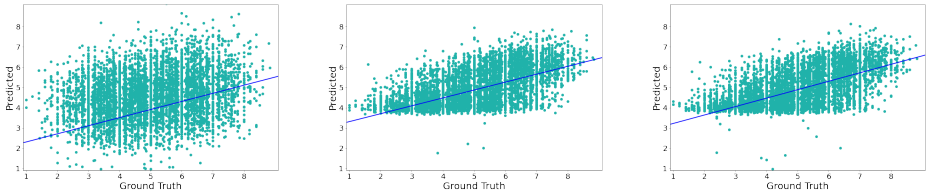
(a) Epoch 1 - Corr. coeff.: 0.52 (b) Epoch 20 - Corr. coeff.: 0.67 (c) Epoch 40 - Corr. coeff.: 0.78



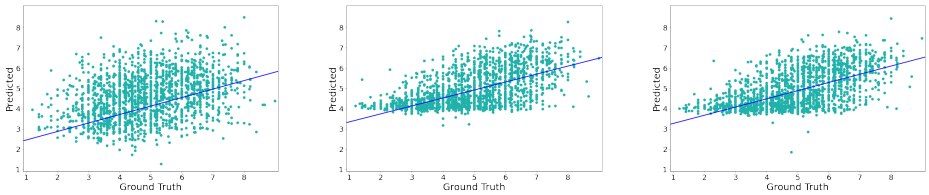
(d) Epoch 1 - Corr. coeff.: 0.53 (e) Epoch 20 - Corr. coeff.: 0.65 (f) Epoch 40 - Corr. coeff.: 0.71

Figure 6: Scatterplot of Ground Truth and Predicted valence values for CCS model trained on BE dataset. Top-row: Training plots. Bottom-row: Validation plots

Similarly, Figure 7 shows the scatter plots of ground truth and predicted arousal values across training (7(a), 7(b) & 7(c)) and validation (7(d), 7(e) & 7(f)) epochs. It is clearly seen that the CCS model learns the Arousal representation across images much better than the BE model shown in Figure 5.



(a) Epoch 1 - Corr. coeff.: 0.29 (b) Epoch 20 - Corr. coeff.: 0.55 (c) Epoch 40 - Corr. coeff.: 0.58



(d) Epoch 1 - Corr. coeff.: 0.32 (e) Epoch 20 - Corr. coeff.: 0.54 (f) Epoch 40 - Corr. coeff.: 0.54

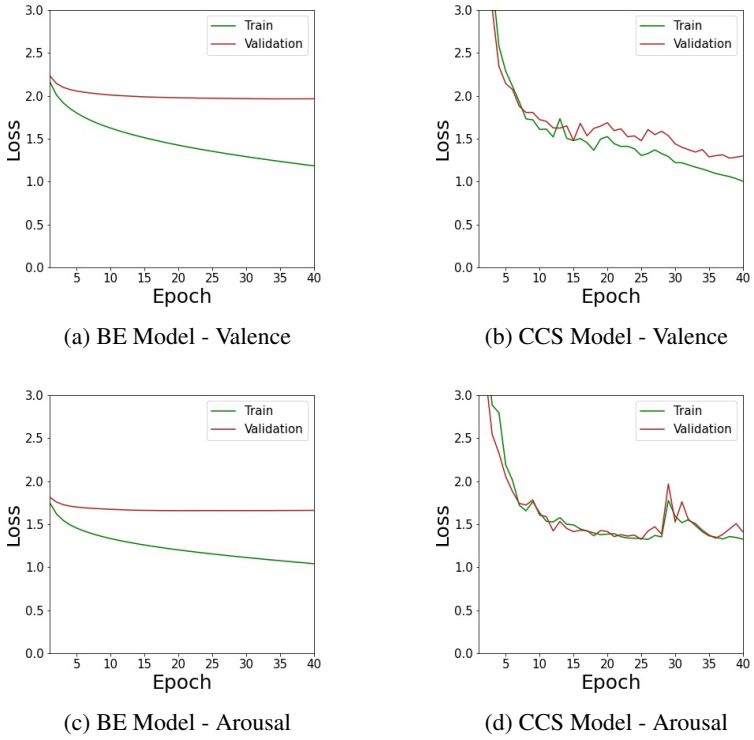
Figure 7: Scatterplot of Ground Truth and Predicted arousal values for CCS model trained on BE dataset. Top-row: Training plots. Bottom-row: Validation plots

We measured the constant mean prediction MSEs for Valence and Arousal to be 2.59 and 1.91 respectively for the BE dataset. From the Table 2, the performance of the BE model over the BE dataset is better by 36% and 23% than the constant mean prediction for Valence and Arousal respectively, indicating that the model is learning. However, the learning of the BE model is not sufficient to capture the complexity in the emotion recognition domain. This has been proven from the Table 2

which clearly depicts that the CCS model which captures the contextual understanding outperforms the BE model for both Valence and Arousal estimation.

Model	Valence	Arousal
BE	1.64	1.47
CCS (Ours)	**1.44**	**1.28**

Table 2: Performance (MSE) of CCS and BE models trained and tested on Building Emotional Machines (BE) dataset



(a) BE Model - Valence

(b) CCS Model - Valence

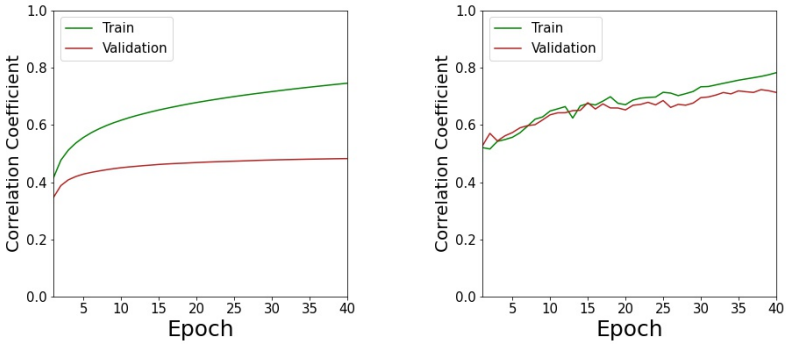
(c) BE Model - Arousal

(d) CCS Model - Arousal

Figure 8: MSE Loss vs. Epoch curves for BE and CCS models on BE dataset. Top-row: Valence plots. Bottom-row: Arousal plots

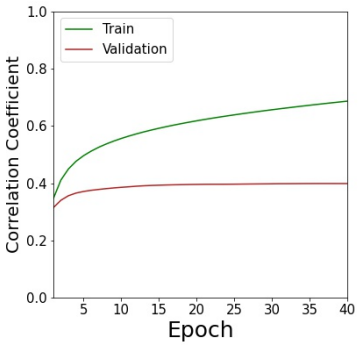
The training and validation loss plots and correlation coefficient plots in the Figures 8 and 9 further validate the relevance of the CCS model in the emotion recognition domain. The convergence across epochs is much better for the CCS model than the BE model. The huge gap in the training and validation plots for BE shows that the model attains saturation beyond which point rendering it incapable of learning the valence and arousal representations across images.

As described in Section 3.3, Cognitive Contextual Summarization (CCS) model comprises of two branches viz. the OFA encoder module and the OFA decoder and BERT module. To understand the contribution of each module with respect to the Valence and Arousal prediction on the BE dataset, we perform an ablation study. In these experiments, we present regression results of valence and arousal on OFA only module of the CCS model and the Fusion module. Table 3 contains MSE performance for valence and arousal prediction of OFA encoder compared with the CCS model. The performance of OFA encoder drops indicating that just the encoder is not enough to capture image emotion features and also validates contribution of BERT and the fusion towards the final results.

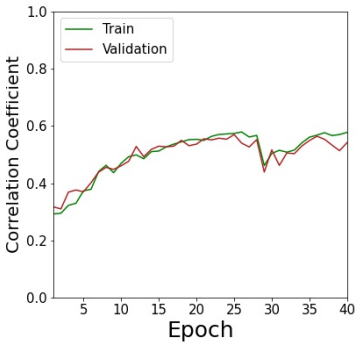


(a) BE Model - Valence

(b) CCS Model - Valence



(c) BE Model - Arousal



(d) CCS Model - Arousal

Figure 9: Correlation coefficient (between ground truth and predicted values) vs. Epoch curves for BE and CCS models on BE dataset. Top-row: Valence plots. Bottom-row: Arousal plots

Model	**Valence**	**Arousal**
OFA encoder	1.59	1.36
CCS (OFA+BERT Fusion)	1.44	1.28

Table 3: Ablation Study of the OFA model in terms of MSE on the BE dataset.

5 Conclusion

Estimating the emotions of an image is the first step towards building models for emoting through visual stimuli. In this work, we propose a novel image emotion estimation framework that, given any image, outputs the valence and arousal of the image. Our framework does away with the need for any additional annotation, unlike those used by the previous works in the literature. Untying emotion estimation from the presence of objects or humans in the image has large scale and widespread applications. It opens endless possibilities for estimating emotions from any visual content.

References

Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL `https://arxiv.org/abs/1810.04805`.

Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

Katharine H. Greenaway, Elise Katherine Kalokerinos, and Lisa A. Williams. Context is everything (in emotion research). *Social and Personality Psychology Compass*, 12, 2018.

Dhiraj Joshi, Ritendra Datta, Elena Fedorovskaya, Quang-Tuan Luong, James Z. Wang, Jia Li, and Jiebo Luo. Aesthetics and emotions in images. *IEEE Signal Processing Magazine*, 28(5):94–115, 2011. doi: 10.1109/MSP.2011.941851.

Yousif Khaireddin and Zhuofa Chen. Facial emotion recognition: State of the art performance on FER2013. *CoRR*, abs/2105.03588, 2021. URL https://arxiv.org/abs/2105.03588.

Hye-Rin Kim, Yeong-Seok Kim, Seon Joo Kim, and In-Kwon Lee. Building emotional machines: Recognizing image emotions through deep neural networks. *IEEE Transactions on Multimedia*, 20(11):2980–2992, 2018.

Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *CoRR*, abs/1811.07770, 2018. URL http://arxiv.org/abs/1811.07770.

Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1667–1675, 2017.

Peter J Lang. A bio-informational theory of emotional imagery. *Psychophysiology*, 16(6):495–512, 1979.

Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, et al. International affective picture system (iaps): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention*, 1(39-58):3, 1997.

Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert. Emotion, motivation, and anxiety: Brain mechanisms and psychophysiology. *Biological psychiatry*, 44(12):1248–1263, 1998.

Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, et al. *International affective picture system (IAPS): Affective ratings of pictures and instruction manual*. NIMH, Center for the Study of Emotion & Attention Gainesville, FL, 2005.

David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

Artur Marchewka, Łukasz Żurawski, Katarzyna Jednoróg, and Anna Grabowska. The nencki affective picture system (naps): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database. *Behavior research methods*, 46(2):596–610, 2014.

Joseph Mikels, Barbara Fredrickson, Gregory Samanez-Larkin, Casey Lindberg, Sam Maglio, and Patricia Reuter-Lorenz. Emotional category data on images from the international affective picture system. *Behavior research methods*, 37:626–30, 12 2005a. doi: 10.3758/BF03192732.

Joseph Mikels, Barbara Fredrickson, Gregory Samanez-Larkin, Casey Lindberg, Sam Maglio, and Patricia Reuter-Lorenz. Emotional category data on images from the international affective picture system. *Behavior research methods*, 37:626–30, 12 2005b. doi: 10.3758/BF03192732.

Jonathan Posner, James A. Russell, and Bradley S. Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17:715 – 734, 2005.

Konrad Schindler, Luc Van Gool, and Beatrice de Gelder. Recognizing emotions expressed by body pose: A biologically inspired neural model. *Neural Networks*, 21(9):1238–1246, 2008. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2008.05.003. URL https://www.sciencedirect.com/science/article/pii/S0893608008000944.

Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu L. Vieriu, Stefan Winkler, and Nicu Sebe. Ascertain: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, 9(2):147–160, 2018. doi: 10.1109/TAFFC.2016.2625250.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework, 2022.