Sentiment analysis on 2D images of urban and indoor spaces using deep learning architectures

Konstantinos Chatzistavros
Information Technologies Institute - CERTH
Thessaloniki, Greece
konschat@iti.gr

Theodora Pistola
Information Technologies Institute - CERTH
Thessaloniki, Greece
tpistola@iti.gr

Sotiris Diplaris
Information Technologies Institute - CERTH
Thessaloniki, Greece
diplaris@iti.gr

Konstantinos Ioannidis
Information Technologies Institute - CERTH
Thessaloniki, Greece
kioannid@iti.gr

Stefanos Vrochidis
Information Technologies Institute - CERTH
Thessaloniki, Greece
stefanos@iti.gr

Ioannis Kompatsiaris
Information Technologies Institute - CERTH
Thessaloniki, Greece
ikom@iti.gr

ABSTRACT

This paper focuses on the determination of the evoked sentiments to people by observing outdoor and indoor spaces, aiming to create a tool for designers and architects that can be utilized for sophisticated designs. Since sentiment is subjective, the design process can be facilitated by an ancillary automated tool for sentiment extraction. Simultaneously, a dataset containing both real and virtual images of vacant architectural spaces is introduced, while the SUN attributes are also extracted from the images in order to be included throughout training. The dataset is annotated towards both valence and arousal, while five established and two custom architectures, one which has never been used before in classifying abstract concepts, are evaluated on the collected data.

CCS CONCEPTS

• **Information systems → Sentiment analysis**.

KEYWORDS

Image Sentiment Analysis, Deep Learning, Vision Transformers, Data Mining, Image Processing

1 INTRODUCTION

Image sentiment analysis aims at the prediction of the sentiment that is evoked by the content of an image to its viewer. In recent years, with the rise of social networks and virtual reality applications, the study of image sentiment analysis has gradually attracted attention from academic and industrial communities. The existence of a supplementary automated tool can guide architects and urban designers, so that their creations meet the needs of the citizens and improve the aesthetics of places. Automatic understanding of the sentiment hidden behind images and videos has many applications, such as online advertisement, brand monitoring and customer feedback to name a few. Our work focuses on the sentiments that images of outdoor or indoor spaces evoke to their viewers from the architectural side of view. Moreover, a sentiment estimation from outdoor or indoor spaces is of great interest towards creating digital models based on architectural sketches. The goal is to predict the feeling a person would have if he/she was at the depicted in the image place, and use this automatic technique as a tool for architects and interior designers.

Most of the works on sentiment analysis are based on textual analysis of comments or opinions about specific topics. Recently, many works focus also on image sentiment analysis. However, only a few, like the work proposed in [9], concern the prediction of sentiment related to places. Respectively, there are no available annotated datasets that could be exploited for the training of sentiment prediction models in this domain, apart from OutdoorSent [9]. Hence, we created a novel dataset with a variety of outdoor and indoor images, along with screenshots of VR environments, oriented to urban architecture and interior design that was annotated by a targeted group of people (architect students, office workers, citizens). The novelty of our dataset is threefold: i) it contains images of outdoor and indoor places that are annotated in terms of sentiment from the architectural point of view, ii) it is a mixed dataset that contains both real photos and screenshots of VR spaces, the first to exist related to sentiment analysis and iii) the annotations concern both the valence and arousal sentiment dimensions, while we found no other dataset of similar content with arousal annotations, enabling an holistic design evaluation. Moreover, we tested this new dataset using both well-established CNN-based network architectures and a recently proposed exclusively for image sentiment analysis [36] to evaluate state-of-the-art techniques under the scope of a specific problem. The developed dataset can be exploited by the research community acting as a benchmark for the under study problem; hence, it is publicly available [1] following the required copyrights. Finally, due to their improved performance in many tasks, a modified architecture of swin transformers [21] is proposed and validated, a technique that has not been previously used in the context of image sentiment analysis.

2 RELATED WORK

There are numerous of works for sentiment analysis based on the analysis of text, but recently the study of image sentiment analysis has been intensified too, due to the evolution of deep learning and multimedia distribution though social networks. In this section, we review important works on image sentiment analysis for a variety

[1] https://m4d.iti.gr/urban-indoor-outdoor-sentiment-analysis-dataset-mindspaces/

of applications based on the sentiment model, the feature extraction and classification approaches that they follow.

2.1 Sentiment Model Approaches

Different approaches of the sentiment expression were presented throughout the related literature[1]. In general, there are two main approaches for emotion modelling, the *Dimensional Model* and the *Categorical Model*. The Dimensional approach represents emotions as points in a two or three dimensional space. Emotions have three basic underlying dimensions, namely the valence, the arousal and the control (or dominance). However, as the control dimension has a small effect, most of the related works focus on a 2D emotion space. This space is obtained by considering only the arousal and the valence axis, as shown in Fig. 1. Most image sentiment analysis works take into account only the valence dimension, either by predicting the sentiment polarity in terms of two levels (positive, negative) or levels (positive, neutral, negative). There are also approaches that adopt more than three levels for the valence dimension, like the work of [32], where five polarity levels are used. The Categorical Model refers to specific emotions, such as "anger", "fear" etc and these emotions can be mapped to the Valence-Arousal-Control space. There are many studies regarding the emotion categories that should be considered, like the *Plutchnik's Wheel of Emotions* [26] and Ekman's theory [11].

In our work, we use the Dimensional Model, taking into consideration both the Valence and Arousal dimensions, using a three-level sentiment polarity.

2.2 Image Sentiment Analysis Datasets

International Affective Picture System (IAPS) [19] comprises a dataset of annotated images in terms of emotions that they evoke to their viewers, using the Valence-Arousal-Control (VAC) sentiment model approach. IAPS, which is one of the most widely used material for psychological research, is composed by 716 photos covering various scenes. The authors of [22] released the *Affective Image Classification* Dataset, which contains 1035 abstract artworks and artistic photos that were crowdsourced to be classified between eight essential emotions. The *Geneva Affective Picture Database* (GAPED) [8] includes 730 pictures labelled considering negative, positive and neutral, while all images have been rated considering the valence, arousal, and the scene coherence. In [3], the *Visual Sentiment Ontology* (VSO), a significantly larger dataset of about 0.5 million images of various topics from social media, is presented. Its images were crawled from social media and labeled with thousands of ANP (Adjective Noun Pair) concepts. In addition, the authors created a separate image dataset from Twitter for a sentiment prediction benchmark. Another dataset created for image sentiment analysis purposes is the *Emotion6* dataset [25], in which the images are labeled based on the Ekman's six basic emotion categories [11]. The *DeepSent* dataset that consists of 1,269 Twitter images annotated in terms of positive or negative sentiment was presented in [36]. These images were manually labeled by five people using the Amazon Mechanical Turk[2] (AMT) crowd-sourcing platform. The authors of [31] crawled approximately 3M tweets from July to December 2016. The collected tweets have been filtered considering only the ones including an image while accompanied by English text. The tweets' extracted sentiment has been classified using a polarity classifier based on a paired LSTM-SVM architecture. The most confident predictions have been used to determine the sentiment labels of the images in terms of positive, negative and neutral. The resulting *Twitter for Sentiment Analysis dataset (T4SA)* consists of 1M tweets and related 1.5M images. *Flickr and Instagram (FI) dataset* [37] is collected from social websites using emotion categories as query keywords. Workers from Amazon Mechanical Turk were then hired to further label the images, shaping a total of 23,308 well-labeled images with eight sentiment categories. By briefly presenting the most widely used images sentiment analysis datasets throughout the research community, the need for a dataset containing also arousal annotations on real and synthetic images for design purposes is exposed due to its absence. Hence, the expansion of the existing dataset with synthetic images will augment data availability towards improving the developed models for more accurate sentiment analysis.

[2]https://www.mturk.com/

2.3 Image Sentiment Analysis Approaches

Different image sentiment analysis approaches have been proposed in the literature involving low-level features (e.g., color, texture), semantic features, machine learning and deep learning techniques.

In [3], the concept of adjective-noun pairs (ANPs) was presented aiming to describe images in terms of emotions/sentiments, establishing a novel mid-level representation for bridging the affective gap. An alternative to low-level attributes to classify visual sentiments, named Sentribute, was presented in [38]. The authors achieved to establish an association between those attributes and the emotional sentiments evoked by the images. Another methodology was presented by the authors of [41] used the ANPs to extract image features with a support vector machine (SVM) classifier for sentiment classification achieving precision of 86% on the visual sentiment ontology (VSO) dataset. ANPs were also used for automatic emotion and sentiment extraction from images in [13], providing enhanced performance compared to low-level image descriptors. Additionally, authors of [35] combined CNNs with separate adjective and noun networks, accomplishing better results than previous works that used ANPs. The researchers of [23] extracted objective text descriptions from the images, and they trained a support vector machine (SVM) classifier to determine the sentiment polarity. They used a dataset with 47235 images and were able to achieve an accuracy of 73.96% combining text and visual features. In [33] a deep-learning-based long short-term memory model (LSTM) is proposed for image sentiment analysis on Flickr and Twitter image datasets achieving an accuracy of 84% and 75% on each one respectively. DeepSentiBank [6] introduces a visual sentiment concept classification method based on deep convolutional neural networks (CNNs) for the prediction of visual sentiment concepts in the form of adjective noun pairs (ANPs) that were automatically discovered from web photos tags. Also, in [36] a CNN architecture specifically designed for visual sentiment analysis is presented, while the authors built the aforementioned DeepSent dataset. The authors of [5] presented extensive experiments comparing several fine-tuned CNNs for visual sentiment prediction. Moreover, they provided visualizations of local patterns that the network learned to associate

with image sentiment giving an insight on how visual positivity (or negativity) is perceived by the model. Furthermore, in [9] a novel urban outdoor image dataset named OutdoorSent was introduced, while five different ConvNet architectures were compared, including a custom sentiment analysis specific architecture from [36], using varying datasets of images, as well as with and without combining the activation maps of the convolution layers with SUN [24] and YOLO [27] semantic attributes. Lastly, the researchers of [40] proposed a novel CNN model that learns and integrates the content information from the high layers of the deep network with the style information from the lower layers to form a more discriminative representation for emotion recognition. In addition, a new loss function is designed through the latter work by including the emotion labeling quality to optimize the proposed inference model. In conclusion, the collected images for the presented work can contribute in the validation of several algorithms towards both valence and arousal in order to create automated tools for sentiment analysis.

3 DATASET DESCRIPTION

In order to train and evaluate the produced CNNs, a total of 1064 images were selected by sampling images from various sources, such as the Places dataset [42], VR environments, and via image crawling. The motive behind selecting images from the Places dataset was its scene-centric character containing various scenes of outdoor spaces that would enrich the final dataset. In fields such as sentiment analysis and opinion retrieval applications the emotional response can be measured utilizing the Self-Assessment-Mannequin (SAM) [4] aiming to deploy a two-fold dataset for classification purposes. In this research, we adopted the three level sentiment polarity for valence and arousal respectively as seen in Fig. 1, since dominance has little effect on shaping sentiments. Additionally, some samples from the final dataset can be seen on Fig. 2 and Fig. 3. Ultimately, images were divided into batches of 30 images and later integrated to GoogleForms questionnaires, enabling us to control the non-duplication of answers on each form. Consequently, the process became less exhausting and error-prone, while the interface contained details regarding the objectives in Spanish, Catalan, English and French. Participants with selected backgrounds studying or working on architecture and design were asked to provide their emotional response via the created questionnaires. Finally, every image was annotated at least by 5 individuals, similarly to [36], while a total number of 50 people participated to the annotation task.

Nevertheless, sentimental subjectivity has significant impact on every datasets' ground truth and as it is stated in [1], visual features, annotation time and the sentiment holder are essential and affect sentiment polarity over time, therefore sentimental ground truth should be considered as something that will evolve along with major societal challenges. In the case of sentiment analysis of architectural space, globally we are on the verge of re-evaluating current practices regarding design rules, while environmental and societal requirements begin to change drastically. Also, as design and architecture are examined, the selected images tend to exclude human presence or traffic, while the vast majority of the selected images were captured by the height of human vision, as indicated

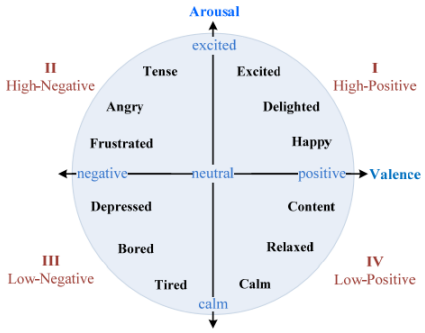

Figure 1: Two dimensional valence-arousal representation

Table 1: Initial valence - arousal distribution of images between classes.

Dataset	Positive	Neutral	Negative
Valence	508	325	231
Arousal	226	505	333

that it would be beneficial for architects. Hence, this work aims to provide a baseline towards understanding the effect of urban visual information on human sentimental stimuli.

As two datasets are formed based on the label consensus, which takes into consideration the amount of people that assigned the same sentimental stimulus to a specific image. Based on the majority voting, the created datasets are slightly unbalanced regarding the distribution of images per class as seen in Table 1. Concerning the datasets' content, 9.3% is derived from Unity virtual environments [3] of indoor spaces with multiple lighting, texture, furniture arrangement and object configurations. The number of synthetic images across the different classes is equal to approximately 8-10%, creating an equal distribution throughout the datasets' classes. Also, part of the synthetic images depict public spaces and urban cultural landmarks. Thus a variety of virtual spaces enhance its uniqueness and sparsity towards understanding the way human sentiment is triggered based on indoor and outdoor design. Additionally, to the best of our knowledge, the uniqueness of the presented dataset relies also on mixing synthetic and real images for sentiment analysis for the first time. Another potential advantage from introducing synthetic images in the presented dataset is the suitability of the produced models, to be used in a 3D design environment, guiding designers throughout the design process objectives.

4 EXPERIMENTS AND RESULTS

4.1 Learning Details

In this subsection, the selected architectures are presented, as we took advantage of some widely used Convolutional Neural Networks (CNNs) for visual analysis and opinion mining tasks, namely the VGG16 [29], InceptionV3 [30], ResNet50 [16], DenseNet169 [17], Xception [7]. At the same time, we examined the published Robust

[3] https://unity.com

Figure 2: Valence dataset sample images. On the left *Negative*, centre *Neutral*, right *Positive* deriving from the Unity design environment

Figure 3: Arousal dataset sample images. On the left *Calm*, centre *Neutral*, right *Excited* deriving from the Unity design environment

architecture [36] because of its reduced complexity compared to the aforementioned CNNs. Additionally, another lower complexity architecture never used for sentiment analysis was utilized based on advanced Vision Transformers (ViT) [21] capable of constructing hierarchical feature maps, as it will be described later. An experiment including vision transformers towards tackling a sentiment analysis task, is of great interest towards evaluating its classification ability on abstract concepts, which as far as we are concerned has never been conducted before. The aforementioned argument is reinforced as results from [21] suggest that the proposed Swin Transformers can serve as a general-purpose backbone for computer vision tasks. To speed up the training phase, architectures [7, 16, 17, 29, 30] were fine-tuned with the pre-trained ImageNet [10] weights. The aforementioned 7 architectures are included in order to test them on a common benchmark dataset aiming to determine which one gives better results and why it is better than the others.

In addition, a simplified Swin Transformer architecture was validated, receiving reshaped input images (224, 224, 3), patch size is set to (2, 2) with embedded patch contents and positions, an initial learning rate of 0.0001, and a weight decay of 0.01 are used, while training took place for 25 epochs. Concerning the self-attention parameters, those are fixed between blocks, but can also vary for larger architectures. The number of attention heads was set to 8, the embedded dimensions were set to 64 and the number of nodes to 256 in order to restrain the increase in computational cost. Finally, the shift-window parameter are, window size equal to 2, the size of shifting equal to 1 for stability and elevated precision, while the number of patches is equal to 112, in an overall attempt to conduct experiments with a low-complexity swin transformer architecture, comparable to [36].

The ConvNet Places365 [42] pre-trained on the Places2 dataset was utilized for the SUN attributes extraction. This 102-dimensional feature descriptor extract information related with materials, surface characteristics, lighting and spatial organization of a scene. Convolution maps and SUN attributes are coupled in a series of dense layers once the SUN attributes have been extracted in order to forecast the ultimate sentimental stimulus. More specifically, the coupled information is incorporated into the first fully connected layer, after each utilized CNN/ViT backbone.

All image dimensions were reshaped to (224, 224) before being fed to the models besides the Inception architecture, which has (299, 299) input shape, as recommended for use on Table 2 at [30]. For every one of the aforementioned architectures the top Dense interpretation layers were updated to cover the problem's needs. Hence, convolutional layers will work as feature extractors and the newly introduced interpretation layers will be capable of interpreting the discovered patterns in order to classify each image, see Fig. 4. The added interpretation layers after each one of the CNN and ViT backbones described, are two couples of fully connected and dropout layers with 2048 neurons and 0.2 drop rate for the first couple, and 1024 neurons and a drop rate equal to 0.2 on the second one. Afterwards, a fully connected layer with 24 neurons, as introduced on the [36] publication, is included, as seen on 4. The aforementioned dense layers were integrated with ReLU activation function, while the final layer is consisted of 3 neurons with a softmax activation function in order to provide the predicted sentimental responses.

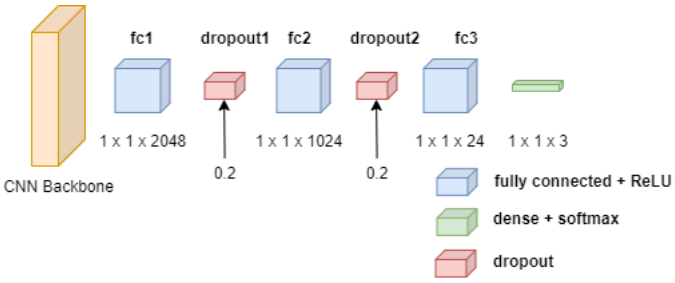

Figure 4: Newly inserted interpretation and regularization layers on top of every used CNN.

The architecture based on the Swin Transformer [21] produces hierarchical feature maps by merging image patches, whilst its backbone is shown in Fig. 5. This leads to linear computation complexity to the input image size due to computing for each local window's self-attention, in contrast with existing Vision Transformers, whose complexity is quadratic to image size due to computing the self-attention globally for each resolution. As stated in [21], the Multi-head Self Attention (MSA) module in a Transformer is replaced by the Swin Transformer consisting of a shifted window based approach. Swin Transformer computes token subsets through non-overlapping windows that are alternately moved within Transformer blocks, as opposed to other vision transformer variations that compute embedded patches (tokens) globally, making it suitable to process high-resolution images [21]. Results from [21] on ImageNet-1K for image classification [10], COCO object detection [20], and ADE20K semantic segmentation [43] verify that Swin Transformers can be utilized as a general backbone for image analysis tasks.

Considering the slightly imbalanced datasets, image augmentation techniques such as rotation, width/height shift and zooming were applied. We employed 5-fold cross-validation utilizing 60% of

the images for training, 20% for validation and 20% for testing. Furthermore, we applied a weighted optimization during the training phase to tackle the problem of class imbalance, where the significance of each sample is expressed by assigning the minority class with increased weight and vice versa. More specifically, the classes' weights are assigned based on calculating the quotient of the minimum number of samples appearing in a class by the number of samples in each subclass.

Concerning the hyper-parameter tuning needed for the training process, batch size was set to 32. Adam optimizer was chosen [18], a method suited for large parameter problems, which is a computationally efficient method with minor memory requirements, and an initial learning rate of 0.0001. A learning rate scheduler was also implemented for reducing the loss at the last stages of training, gradually decreasing the learning rate as the training process is almost complete, down to 0.000001.

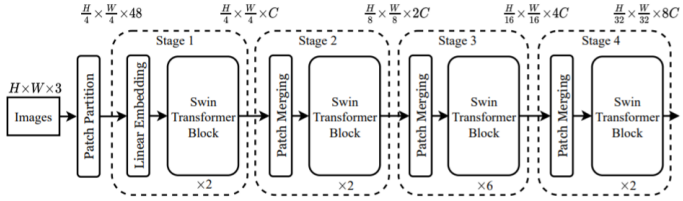

Figure 5: Swin Transformer architecture used as a backbone for sentiment analysis [21]

4.2 Results

Having discussed the deployed architectures throughout this research effort, the utilized dataset and the approach followed for sentiment analysis we proceed towards analyzing the experimental results. A total of 28 experiments were conducted including the SUN attributes insertion, by training separately the valence and arousal models, because a hierarchical model would pose disadvantages, such as the complexity of the overall combined model and therefore expensiveness of computational resources. Work from [9] focuses on sentiment understanding of urban outdoor images, while claiming that indoor images do not influence the sentiment classification of outdoor images taking into consideration the computational cost trade off. This work aims towards evaluating indoor and outdoor space in real and virtual 3D environments, therefore a relevance gap occurs, but some results from [9] can serve as a baseline for comparison.

Another experiment performed was the use of the SUN semantic attribute database [24] in order to extract the scene attributes as low-dimensional features. The extracted features are used to capture high-level context and semantics in scenes. A direct comparison with [9] results can be conducted, however, YOLO attributes were not included as the majority of images contained vacant spaces in an attempt to avoid inserting noisy tensors, due to irrelevant classes in the PASCAL VOC2012 object detection dataset [12]. Following this methodology enables us to evaluate the added value of scene attributes.

In Table 2, results for three valence classes are presented, while performance fluctuates from 43.55% up to 70.31%. Performance levels are enhanced by 9.73% for [21], 5.11% for [36] and 6.1% for [16], while they are reduced by 0.94% for [17], 1.7% for [7], 0.24% for [30] and 4.15% for [29] when the SUN attributes are introduced throughout training. The aforementioned performance reduction leads to the conclusion that high complexity architectures do not to benefit from the introduction of the SUN attributes. The SUN attributes tend to impact positively simple architectures, as [9] states, and our experiments confirm this claim as the Robust and Swin based architectures are both benefited. This claim is coherent and lucid for smaller deep CNNs compared to [7, 16, 17, 29, 30] as the increased number of convolutional layers with different filter and stride size extract diverse sets of features that are later fed towards the interpretation layers. Hence, a tensor decorated with the SUN attributes can offset the reduced number of extracted features between shallower CNNs. At the same time, the same does not seem to apply for all of the rest architectures on Table 2.

Having discussed results from Table 2, we move forward to the experimental evidence of Table IV from [9], where five architectures [16, 17, 29, 30, 36] were trained on the OutdoorSent dataset on the same three valence classes. Compared to results from Table IV from [9], the rest of the architectures report an increased accuracy performance by 5.68% for the [36], 26.07% for [29], 5.22% for [30], 2.71% for [16] and 6.27% for [17] without the SUN attributes. Respectively, when introducing the SUN attributes we observed an enhanced accuracy by 0.66% for [36], 5.34% for [30], 8.7% for [16] and 6.88% for [17] and a decrease by 3.34% for [29]. Experiments outperformed the corresponding ones from [9] and that can be an indicator that this dataset is suitable and robustly annotated towards the task of sentiment analysis of 3D spaces, while we also performed experiments with [7].

Proceeding, Table 3 aggregates the results for all experiments on the arousal dataset. On that account, [7, 17, 29, 30] networks indicate a slight performance reduction, while [16] records a slight enhancement of 0.47% when trained with the SUN attributes. Once again, [36] and [21] are benefited from introducing the SUN attributes during training and a performance enhancement by 9.4% and 5.68% is observed respectively. Indicatively, in Fig. 6 the predictive ability of Densenet is visualized via the confusion matrix trained for 10 epochs on the valence dataset when inserting the SUN attributes. Regarding the predictive ability of Densenet, Fig. 6 indicates that the majority of the testing samples per class are accurately classified. The rest of the architectures exhibit a similar but declining behaviour while classifying the test images towards their potential sentimental response. Hence, the validated architectures, especially the presented Densenet, are able to generate reliable predictions in order to be a complementary tool for architects and designers throughout the design process. Unfortunately, to our knowledge there are no datasets and algorithms deployed for classifying the arousal of architectural 3D spaces, therefore an evaluation or comparison of results is infeasible. Nevertheless, research described in study [34] gives some insights on the relationship between arousal and crucial spatiotemporal aspects of form for affect-driven architectural design. Findings demonstrate that characteristics such as curved or complicated spaces are strongly associated with higher arousal, a dominant characteristic on many samples of the 'Excited' label, in contrast with the 'Calm' one.

Table 2: CNNs performance on the valence dataset in terms of accuracy(%)

Architecture	Without Attributes	With Attributes
Densenet [17]	**68.37**	**70.31**
InceptionV3 [30]	65.93	65.69
ResNet50 [16]	63.01	69.1
VGG-16 [29]	58.4	54.25
Robust [36]	43.55	48.66
Xception [7]	65.2	63.5
Swin based [21]	44.28	54.01

Table 3: CNNs performance on the arousal dataset in terms of accuracy(%)

Architecture	Without Attributes	With Attributes
Densenet [17]	**69.87**	**70.61**
InceptionV3 [30]	69.62	64.19
ResNet50 [16]	66.42	66.9
VGG-16 [29]	53.58	50.37
Robust [36]	40.72	50.12
Xception [7]	68.72	62.22
Swin based [21]	46.17	51.85

As stated, there is evidence of published research effort regarding evaluating the arousal stimulus in various computer vision tasks. Nevertheless, there is no related research effort towards evaluating architectural spaces, as the majority of the publications utilize social media data, such as twitter, in order to examine user's valence and arousal. Arousal, again, is not included in most publications regarding sentiment analysis whilst more works evaluate EEG signals [14, 39], images [36], whilst more common is multimedia opinion mining systems analyzing images and social media posts [32], [2]. It is clear that the relevance gap between different tasks requires the need to extract totally different features in order to train robust algorithms.

4.3 Computational Cost

We implement the presented DCNN models combining Keras, Tensorflow frameworks [15] and Python 3 on a Windows X86-64 machine with Intel(R) Core(TM) i5-7600K @3.8GHz, 8 GB RAM and NVIDIA GeForce RTX 2070. The described system was used throughout all training processes. Specifically, Xception [7] was trained for 8 epochs, Densenet [17] and InceptionV3 [30] were trained for 10 epochs, VGG-16 [29] for 20 epochs, Robust [36] and the Swin based [21] for 25 epochs and ResNet50 [16] for 27 epochs. On inference mode there is not much variation between the produced architectures as approximately 1.5 seconds per image are required in order to produce valence and arousal predictions.

5 COMPARISON AND CONCLUSIONS

In this research, we validated several CNNs on a novel dataset consisting of real and virtual images of vacant 3D indoor/outdoor spaces. The goal was to investigate its potential for the task of visual

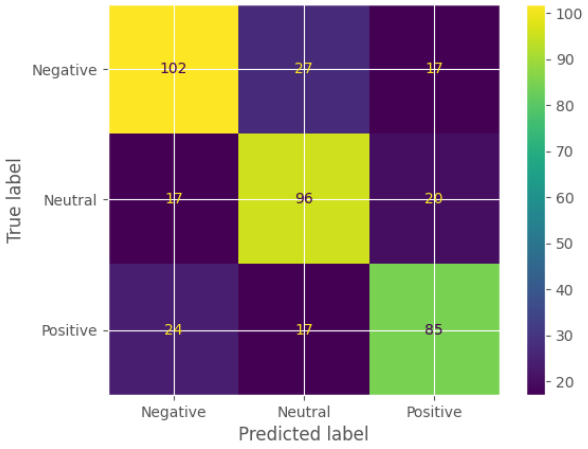

Figure 6: Densenet confusion matrix trained on the valence dataset for 10 epochs including the SUN attributes.

sentiment extraction, and the impact of inhomogeneous images during training state-of-the-art algorithms. Results suggest that performance of high complexity architectures can reach adequate levels. Regarding lower complexity architectures, as described in previous sections, certain limitations occur compared to the rest of the established architectures [7, 16, 17, 29, 30]. Another finding concerning the proposed Swin based architecture [21], is that it outperforms the Robust [36] for both valence and arousal, but it cannot be safely recommended for sentiment analysis tasks as more research is required, confirming that they are inferior trained from scratch on a mid-sized dataset such as ImageNet, compared to well-established CNNs. Pursuant to the results section, while compared to [9], the introduced SUN attributes seem to have a positive effect by enhancing the performance of lower complexity CNN architectures to a greater extent than the widely established CNNs.

Concluding, this research can prove beneficial to a broad area of study such as psychology, virtual reality in architecture, social science, and even towards designing sophisticated multimodal sentiment analysis systems for real life and virtual environments. We are ambitious towards enhancing the percentage of VR related images on future work, including both indoor/outdoor content from virtual environments and secondary real images. That would facilitate testing the produced algorithms on design environments, while evaluating the produced architectures on external task compatible datasets would be exemplary. Apart from its exploitation in sentiment analysis, the introduced dataset could complement Synthia dataset [28], which originally used for semantic segmentation of urban scenes, providing also sentiment analysis metadata facilitating the deployment of multimodal systems on virtual environments. Hence, future research efforts can contribute towards reinforcing our understanding and mapping of future urban environments.

ACKNOWLEDGMENTS

This work was supported by the EC-funded Horizon 2020 Research and Innovation Programmes Mindspaces under Grant Agreement No.825079 and the xR4DRAMA, under Grant Agreement No 952133.

REFERENCES

[1] Alessandro Ortis Giovanni Farinella Sebastiano Battiato. 2019. An Overview on Image Sentiment Analysis: Methods, Datasets and Current Challenges. *16th International Conference on Signal Processing and Multimedia Applications* 54, 2 (01 2019), 10 pages. https://doi.org/10.5220/0007909602900300

[2] Ashutosh Bhawsar, Devashish Katoriya, Ninad Kapadnis, Bhushan Shilawat, and Anand Kolapkar. 2020. Automated Sentiment Analysis of Web Multimedia.

[3] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-Scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs. In *Proceedings of the 21st ACM International Conference on Multimedia* (Barcelona, Spain) *(MM '13)*. Association for Computing Machinery, New York, NY, USA, 223–232. https://doi.org/10.1145/2502081.2502282

[4] Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25, 1 (Jan. 1994), 49–59. https://doi.org/10.1016/0005-7916(94)90063-9

[5] Victor Campos, Brendan Jou, and Xavier Giro-i Nieto. 2016. From Pixels to Sentiment: Fine-tuning CNNs for Visual Sentiment Prediction. https://doi.org/10.48550/ARXIV.1604.03489

[6] Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. 2014. DeepSentiBank: Visual Sentiment Concept Classification with Deep Convolutional Neural Networks. https://doi.org/10.48550/ARXIV.1410.8586

[7] François Chollet. 2016. Xception: Deep Learning with Depthwise Separable Convolutions. https://doi.org/10.48550/ARXIV.1610.02357

[8] Elise Dan-Glauser and Klaus Scherer. 2011. The Geneva affective picture database (GAPED): A new 730-picture database focusing on valence and normative significance. *Behavior research methods* 43 (03 2011), 468–77. https://doi.org/10.3758/s13428-011-0064-1

[9] Wyverson Bonasoli de Oliveira, Leyza Baldo Dorini, Rodrigo Minetto, and Thiago H. Silva. 2020. OutdoorSent. *ACM Transactions on Information Systems* 38, 3 (Jun 2020), 1–28. https://doi.org/10.1145/3385186

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[11] Paul Ekman, Wallace Friesen, Maureen O'Sullivan, A. Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William LeCompte, Tom Pitcairn, and Pio Ricci Bitti. 1987. Universals and Cultural Differences in the Judgments of Facial Expressions of Emotion. *Journal of personality and social psychology* 53 (11 1987), 712–7. https://doi.org/10.1037/0022-3514.53.4.712

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2010. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision* 88, 2 (June 2010), 303–338.

[13] Delia Fernandez, Alejandro Woodward, Victor Campos, Xavier Giro i Nieto, Brendan Jou, and Shih-Fu Chang. 2017. More Cat than Cute?. In *Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes*. ACM. https://doi.org/10.1145/3132515.3132520

[14] Divya Garg and Gyanendra Verma. 2020. Emotion Recognition in Valence-Arousal Space from Multi-channel EEG data and Wavelet based Deep Learning Framework. *Procedia Computer Science* 171 (01 2020), 857–867. https://doi.org/10.1016/j.procs.2020.04.093

[15] Aurélien Géron. 2019. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. https://arxiv.org/abs/1412.6980

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. https://doi.org/10.48550/ARXIV.1512.03385

[17] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 2261–2269. https://doi.org/10.1109/CVPR.2017.243

[18] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. https://doi.org/10.48550/ARXIV.1412.6980

[19] Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, et al. 1999. International affective picture system (IAPS): Instruction manual and affective ratings. *The center for research in psychophysiology, University of Florida* (1999).

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 740–755.

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. https://doi.org/10.48550/ARXIV.2103.14030

[22] Jana Machajdik and Allan Hanbury. 2010. Affective image classification using features inspired by psychology and art theory. 83–92. https://doi.org/10.1145/1873951.1873965

[23] Alessandro Ortis, Giovanni Farinella, Giovanni Torrisi, and Sebastiano Battiato. 2018. Visual Sentiment Analysis Based on on Objective Text Description of Images. 1–6. https://doi.org/10.1109/CBMI.2018.8516481

[24] Genevieve Patterson, Chen Xu, Hang Su, and James Hays. 2014. The SUN Attribute Database: Beyond Categories for Deeper Scene Understanding. *International Journal of Computer Vision* 108, 1-2 (2014), 59–81. https://doi.org/10.1007/s11263-013-0695-z

[25] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew Gallagher. 2015. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 860–868. https://doi.org/10.1109/CVPR.2015.7298687

[26] Robert. Plutchik. 1980. *Emotion, a psychoevolutionary synthesis / Robert Plutchik*. Harper Row, New York.

[27] Joseph Redmon and Ali Farhadi. 2016. YOLO9000: Better, Faster, Stronger. https://doi.org/10.48550/ARXIV.1612.08242

[28] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. 2016. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3234–3243. https://doi.org/10.1109/CVPR.2016.352

[29] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. https://doi.org/10.48550/ARXIV.1409.1556

[30] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the Inception Architecture for Computer Vision. https://doi.org/10.48550/ARXIV.1512.00567

[31] Lucia Vadicamo, Fabio Carrara, Andrea Cimino, Stefano Cresci, Felice Dell'Orletta, Fabrizio Falchi, and Maurizio Tesconi. 2017. Cross-Media Learning for Image Sentiment Analysis in the Wild. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*. 308–317. https://doi.org/10.1109/ICCVW.2017.45

[32] Can Xu, Suleyman Cetintas, Kuang-Chih Lee, and Li-Jia Li. 2014. Visual Sentiment Prediction with Deep Convolutional Neural Networks. https://doi.org/10.48550/ARXIV.1411.5731

[33] Jie Xu, Feiran Huang, Xiaoming Zhang, Senzhang Wang, Chaozhuo Li, Zhoujun Li, and Yueying He. 2019. Sentiment analysis of social images via hierarchical deep fusion of content and links. *Applied Soft Computing* 80 (04 2019). https://doi.org/10.1016/j.asoc.2019.04.010

[34] Emmanouil Xylakis, Antonios Liapis, and Georgios N. Yannakakis. 2021. Architectural Form and Affect: A Spatiotemporal Study of Arousal. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. https://doi.org/10.1109/acii52823.2021.9597420

[35] Jufeng Yang, Dongyu She, Ming Sun, Ming-Ming Cheng, Paul L. Rosin, and Liang Wang. 2018. Visual Sentiment Prediction Based on Automatic Discovery of Affective Regions. *Trans. Multi.* 20, 9 (sep 2018), 2513–2525. https://doi.org/10.1109/TMM.2018.2803520

[36] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2015. Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks. (09 2015), 381–388.

[37] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Building a Large Scale Dataset for Image Emotion Recognition: The Fine Print and The Benchmark. https://doi.org/10.48550/ARXIV.1605.02677

[38] Jianbo Yuan, Sean Mcdonough, Quanzeng You, and Jiebo Luo. 2013. Sentribute: image sentiment analysis from a mid-level perspective. https://doi.org/10.1145/2502069.2502079

[39] Qiang Zhang, Xianxiang Chen, Qingyuan Zhan, Ting Yang, and Shanhong Xia. 2017. Respiration-based emotion recognition with deep learning. *Comput. Ind.* 92-93 (2017), 84–90.

[40] Wei Zhang, Xuanyu He, and Weizhi Lu. 2020. Exploring Discriminative Representations for Image Emotion Recognition With CNNs. *IEEE Transactions on Multimedia* 22, 2 (2020), 515–523. https://doi.org/10.1109/TMM.2019.2928998

[41] Ziyuan Zhao, Huiying Zhu, Zehao Xue, Zhao Liu, Jing Tian, Matthew Chua, and Maofu Liu. 2019. An image-text consistency driven multimodal sentiment analysis approach for social media. *Information Processing Management* 56 (08 2019). https://doi.org/10.1016/j.ipm.2019.102097

[42] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 6 (2018), 1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009

[43] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2016. Semantic Understanding of Scenes through the ADE20K Dataset. https://doi.org/10.48550/ARXIV.1608.05442