Aesthetic Rating and Emotion Analysis using deep learning

Mohammed Suhail, Shashi Kumar , Aparna Balagopalan

Abstract

Aesthetics is defined as the set of principles concerned with the nature and appreciation of beauty. Although there are no universal features that can be used to determine the aesthetics of an image, some specific features tend to be more pleasing that others. In this project we aim to build a model that classifies the aesthetics of a given image. We have also created a large-scale emotion-dataset consisting of natural images and belonging to six differnet emotion classes.Both aesthetic rating and emotion analysis are subjective problems and require a good mapping between locally computable features to the high-level cognizance, which we aim to build using well-trained and developed deeep learning models.

1 Introduction

1.1 Aesthetic Rating

Aesthetic rating has a wide variety of applications. Aesthetic rating can be deployed in image search engines so as to rank display results with high aesthetics. Photo management and picture editing softwares can also benefit from automated aesthetic rating. These days people take lots of photos and later struggle to find the good ones and delete the rest. Image aesthetics classification can help automate this task. Automating the process of aesthetic rating is challenging for several reasons:

1. Large difference between the images belonging to the same class
2. The inherent semantic gap between low-level features and high level aesthetic rules

3. The subjective nature of the problem

Despite these challenges automated computational aesthetics has been an area of active research recently.

1.2 Emotion Classification

Right images can bring out emotions in human beings. In this sense, a person moods can be changed by showing her some images. But what is right image for any particular emotion for an individual? We are trying to explore for what could be the link between images and induced emotion in this project. This could be very helpful for people with frequent mood swing disorder or people with depression. Final aim of this project is to put up an open source website with images for different emotions to help people.

Emotion classification in natural images is defined as developing methods to be able to differentiate the emotions conveyed by an image from that of another.This topic of research has great bearing now with the huge amount of data generated in the form of images on various platforms.Analysis of the sentiments reflected in the images could lead to trend prediction in numerous scenarios.

There is no open source large dataset available for emotion classes. So, first step of this project is to crawl a large dataset and verify it's reliability in terms of concepts captured for different classes of emotions. Emotions can be very subjective so the variance of dataset must be verified. We also trained a first step rudimentary classifier to see the performance on this dataset.

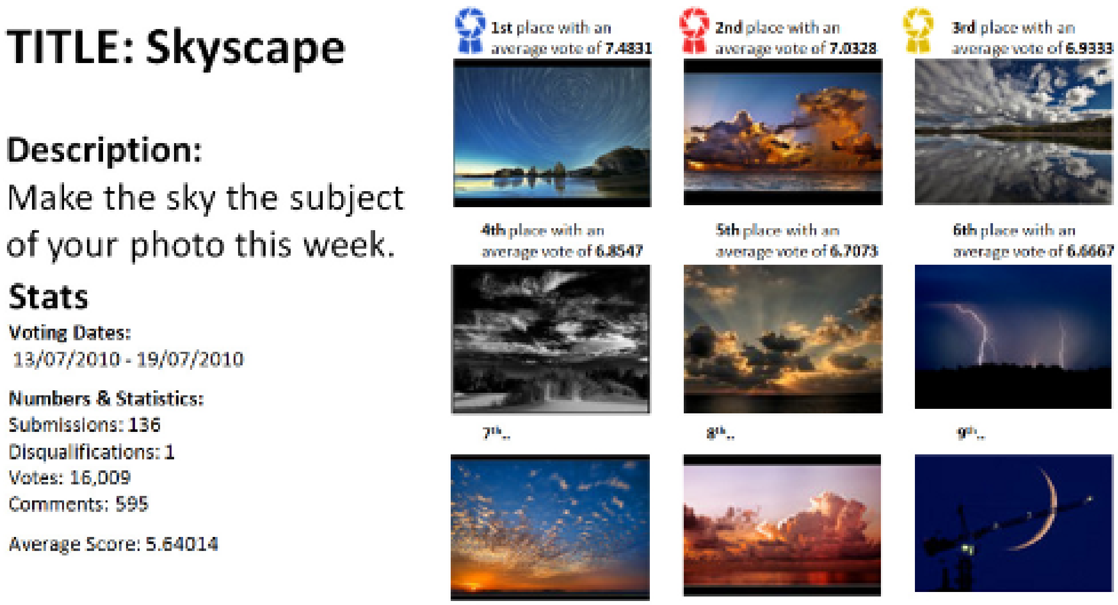
2 Datasets

2.1 AVA

Consisting of around 250,000 images, the AVA dataset has numerous annotations such an aesthetics ratings and labels for every image.The dataset, which has been built for Aesthetic Visual Analysis, has been obtained from dpchallenge.net where photographers regularly upload pictures and obtain photo ratings from other members of the photography community.As a result of this , AVA dataset has additional labels such as that for style and semantics.The images have three different kinds of attributes:

- **Aesthetic**:There are on average 200 ratings given to each image by photographers
- **Style**: Each image has tags from a set of selected 14 styles such as Rule of Thirds,Macro ,Motion Blur etc.
- **Semantic**:Approximately 60 percent of the images have some connotations such as 'Black and White','Humour' etc.There are 66 such connotations.

1



TITLE: Skyscape

Description:
Make the sky the subject of your photo this week.

Stats

Voting Dates:
13/07/2010 - 19/07/2010

Numbers & Statistics:
Submissions: 136
Disqualifications: 1
Votes: 16,009
Comments: 595

Average Score: 5.64014

1st place with an average vote of 7.4831
2nd place with an average vote of 7.0328
3rd place with an average vote of 6.9333
4th place with an average vote of 6.8547
5th place with an average vote of 6.7073
6th place with an average vote of 6.6667
7th..
8th..
9th..

Figure 1: Example of images in AVA dataset

2.2 AADB

The AADB data set consist of around 10,000 images downloaded from Flikr website. All the non-photographic images like cartoons, paintings, drawings etc were manually removed. The annotation of each image and their aesthetic score assignment were done by using the Amazon Mechanical Turk service. Each image has eleven attribute which include , shallow depth of field, motion blur,rule of thirds, balancing element, interesting content, object emphasis, good lighting, color harmony, vivid color, balancing element, repetition and symmetry. In all our experiments we split the dataset into 8500 images for training, 1000 for testing and 500 for validation. Some example images from the AADB datset along with their attributes and score are shown in the figure below 2

2.3 Emotion Dataset

Most human beings can commonly feel six types of emotions: Happy, Sad, Fear, Anger, Motivation and Solitude[8]. We crawled images from Flickr with these emotion tags. Flickr has many limitations for their API usage like only 4000 images can be crawled using one API Key in one TCP connection. So, we tried

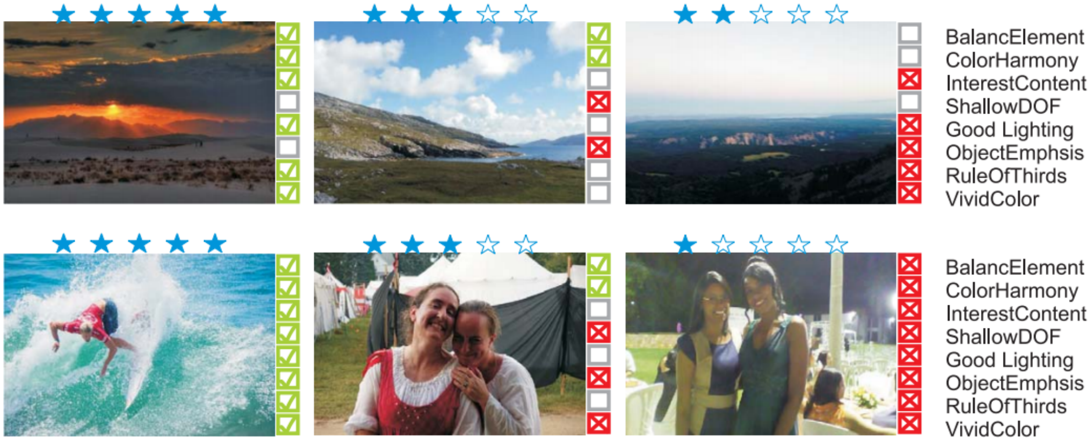


Figure 2: Images from the AADB dataset. Each image has fields corresponding to 11 different attributes and an aesthetic score associated with it

to download 4000 images yearly starting from 2005. For all the emotion tags around 300,000 images were downloaded. There might be repetition among these images because an image might have several tags for which We exclusively queried for images which have single emotion tag. In case of Fear, tag 'scary' is very similar, we also queried for these types of synomyms but in an exclusive manner. Images are of different dimensions and resolutions but we didn't scale or crop them, it can be done by the users according to their needs and requirements. But for rudimentary classifier in this project, images are scaled to 256x256 and then center cropped to 224x224 and works reasonably good. Images for each type of emotion is shown in fig 3

In the fig 3 different aspects of emotions can be seen. Facial description of the lady depicts that she's sad, a concept of workout motivation, embracing solitude amidst beautiful sun, a littered dark room with scary scratches on the wall, angry man going to hit someone with axe and a happy faced guy. These concepts can be very subjective, so the emotion dataset should include large number of concepts to capture subjectiveness.

3 Previous Methods

3.1 Rating Pictorial Aesthetics using Deep Learning

In this paper [1] the authors conduct systematic evaluation of the single column DCNN and develop a deep convolutional neural network that learn features jointly fron heterogeneous inputs simultaneously. They

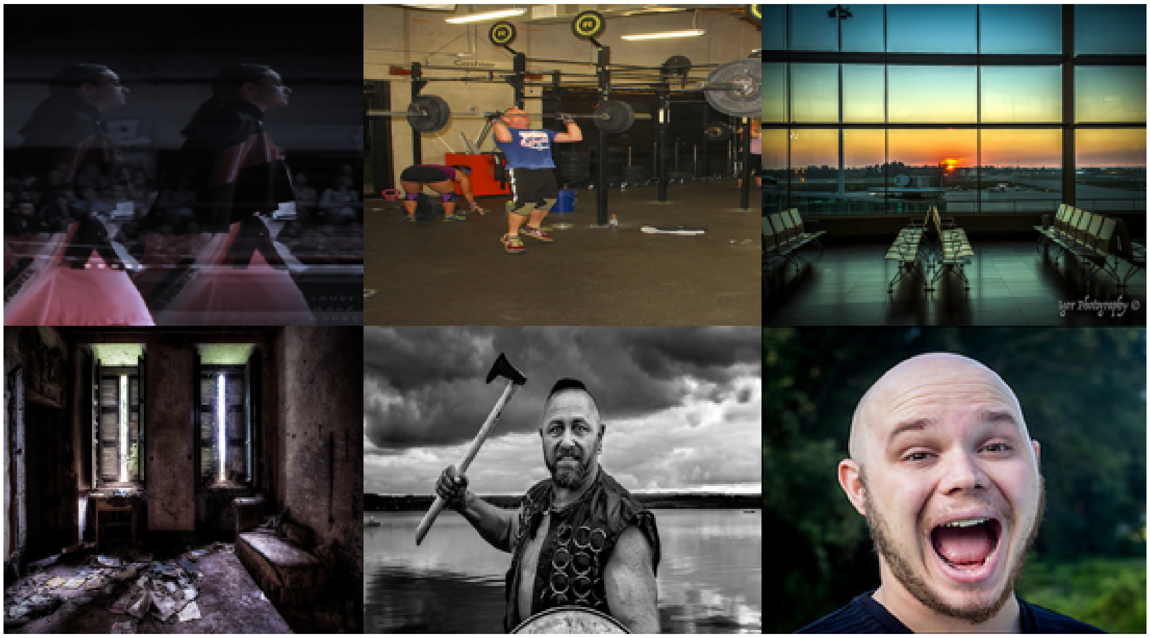


Figure 3: Images from crawled emotion dataset. Sad, Motivation, Solitude, Fear, Anger and Happy emotion respectively.

employ global and local transformations of images to improve the performance of the model.They also proposed a regularized DCNN using the style tags that improve their testing accuracy on the AVA Dataset. The final cost function used is the negative maximum likelihood cost function of the aesthetic attribute given the input to the network and the style attributes.

3.1.1 Single Column Convolutional Network

There are a large variety of compositional principles that contribute toward the aesthetics of an image. This make it difficult to hand-craft these features manually to train a model. Instead the authors leverage the power of CNN's to automatically detect and identify the useful features and patterns. However it is difficult to train the CNN directly to learn for aesthetics. This issue is addressed by the authors through the use of a global and local view. Each image is normalized using 4 different transformations namely center crop, wrap, padding and random crop. These four images are then used to single column CNN's independently to compare how well each transformation performs with respect to the aesthetic rating problem. Experimental results revealed that the best performance was obtained with the random crop transformation followed by the wrap transformation.

3.1.2 Double Column Convolutional Neural Network

The authors of the papers also propose a Double column CNN that can automatically learn features from a heterogeneous input i.e. a global view and a local view. The DCNN takes the wrap image and the random crop image as the inputs. A 256 dimensional column vector is extracted from each column which jointly forms a fully connected layer. This fully connected layer is jointly trained on the input. The architecture of the DCNN is shown in Figure 4 The binary label of the images provide a weak supervision for the DCNN to learn the features that are relevant for aesthetics. In order to provide a better supervision for the learning process the authors propose a regularized DCNN in which another column is added onto the earlier network 4 so as to incorporate the style attribute of the image in the learning process. The performance of the architecture is listed in the table 1

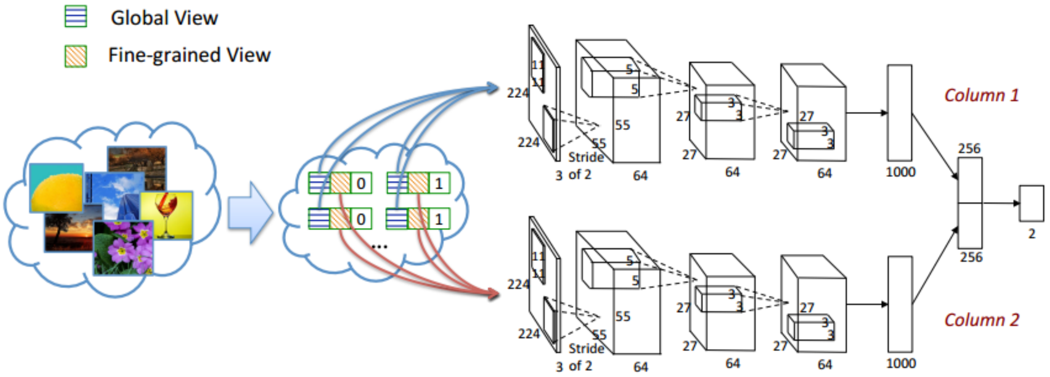


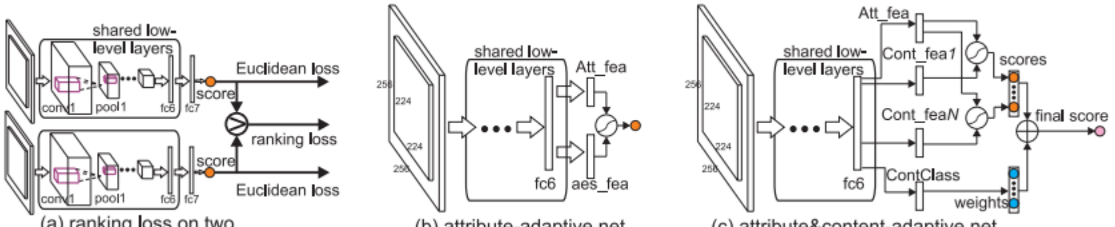
Figure 4: Architecture of DCNN used for aesthetic rating. The network takes a global input and a local input to decide the class of the image[1]

Table 1: RAPID Performance Table

δ	DCNN	RDCNN
0	73.25	74.46
1	73.05	73.70

3.2 Photo Aesthetics Ranking Network with Attributes and Content Adaptation

In this paper [2] the authors approached the aesthetic rating problem in a different manner. Rather than considering aesthetic rating as a binary class problem they posed is as a regression-based problem where the goal is to assign an aesthetic score to an images on the scale of one to ten. The architecture of the network used is shown 5



(a) ranking loss on two identical regression net

(b) attribute-adaptive net

(c) attribute&content-adaptive net

Figure 5: Architecture of different models used [2]

In this paper the authors first fine tune alex-net on the AADB dataset.However instead of the softmax loss in alex-net the use the euclidean loss function on the absolute rating (scaled between 0 and 1) and the predicted rating.

The euclidean loss function however does not account for the relative ranking of the images that have similar average aesthetic scores. The authors then proposed a Siamese network with pairwise ranking loss which exploits the relative ranking of the images in the dataset.

The authors also proposed a attribute adaptive model and a content adaptive model.

3.2.1 Attribute adaptive model and Content adaptive model

Rather than training the networks solely on the basis of aesthetic rating the authors have integrated attribute loss into the training of the network. This attribute label bases training can be considered as a side-information of deep supervision. The content of the image and its photographic attributes are strongly co-related.The model is made compatible to predict the category label by fine tuning the last two layer of alex-net using a softmax loss on the category labels.

With the proposed model they were able to achieve a training accuracy of 77.3% on the AVA dataset and a spearman's score of .6782 on the AADB dataset.

3.3 Deep Image Aesthetics Classification using Inception Modules and Fine-tuning Connected Layer

This is one of the most recent work [4] on automated aesthetic rating. In this paper the authors addressed the issue of aesthetic classification using a recently popular method called the inception module. In an inception module the size of the convolution filters are restricted to be either 1*1 or 3*3 or 5*5.

In the ILGNet for aesthetic rating proposed by the author inception modules are stacked one on top of another. The first two layer are considered to extract the local image features and the third one extracts the global features. This approach is similar to the one in RAPID in which the network was forced to learn local and global feature by using a double column CNN with two inputs. The architecture used is shown in the figure 6 The network was initially trained on image net and the connected layer was later fine tuned on the AVA dataset. This method is currently the state of the art method and achieves the highest accuracy of 79.85% on the AVA dataset.

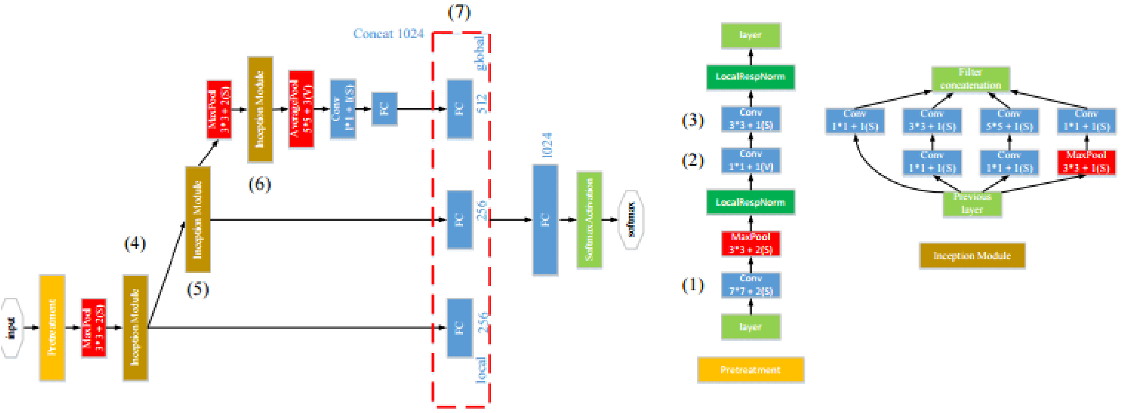


Figure 6: Architecture of ILGNet [4]

4 Proposed Methodology

4.1 Aesthetics Classification

Since the task of aesthetics classification is subjective and involves high-order perception , the most important criteria which would determine the performance of the learned model would be the subset of features selected.In order to perform this accurately , we propose two alternative routes:

4.1.1 Recurrent Visual Attention Model

Recent researches have suggested that the way humans perceive the aesthetics of an image depend on how they gaze through the image. In order to replicate this behaviour in our model we trained a visual attention based recurrent neural network on the aesthetic dataset. The visual attention model posses the classification problem as a sequential decision problem. The recurrent attention model consist of a glimpse sensor which is a part of the glimpse network and an agent that decides how to deploy the sensor on the image. The sensor looks into a part of the environment and classifies the patch. Based on the classification of the glimpse network a reward is assigned for the action of classification which then affect the agents decision of where to deploy the sensor in the next time step. The architecture of the model used is shown in figure 7 At each time step the agnet specifies the glimpse sensor where to look at. The glimpse sensor then creates a retina like representation of the patch. The retina-like representation is a vector which consist of images that are of progressive lower resolution. The glimpse network the processes the retina-like representation along with the original patch to create the glimpse vector. The glimpse vector along with the hidden state from the previous time step form the next hidden state through a linear combination followed by a non-linear Relu activation. This hidden state is then fed into the action network which then decided the class which the patch belongs to. If the agent classifies the image correctly then it recieves an award of 1 and 0 otherwise. The goal of the agent is to maximize the sum of the rewards obtained in all time steps.

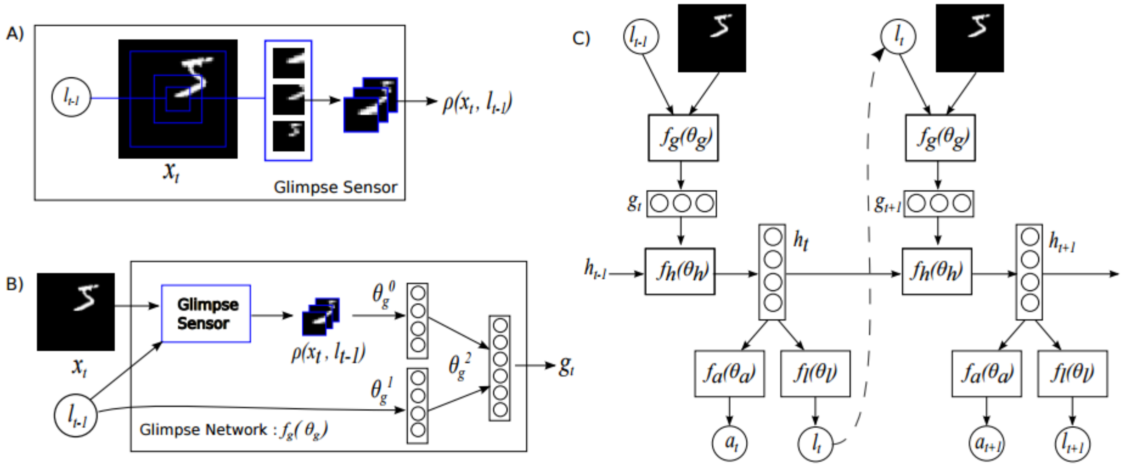


Figure 7: Architecture of recurrent attention based model [3]

4.1.2 Experiments

The RAM was trained on the AADB dataset using the torch framework. The size of the patch was set to 16x16. The scale factor used to create the retina-like representation was set to 2 along with a value of 7 for the number of patches that were inspected in an image. The dataset was split into 8500 images for training, 1000 for testing and 500 for validation. The model after training for 50 epoch achieved a highest accuracy of 65.2% on the testing set.

4.2 Fine tune VGG net

The next approach that we tried was to fine tune VGG-net to be used for aesthetic rating. For this the pre tained VGG-16 model was used. The final convolutional layer along with the fully connnected layer we then trained using the AADB dataset to achieve an accuracy of 56.8%. The low accuracy obtained is due to the small size of the dataset used for fine tuning the model

4.2.1 Residual Nets with Path Multiplicity

Residual Nets [7] have enhanced performance because of their depth.However, recent research has shown that another factor for the higher level of accuracy obtained using ResNets is due to the multiplicity of paths introduced due to the residual connections.It is interpereted that at every residual connection a decision to flow along it or not is made leading to exponentially many possible paths.On replicating the same number of effective paths in a Convolutional Neural Network which was not very deep, similar performance was obtained.This thus intoduces ’path multiplicity ’ as another factor along with depth and width of a Convolutional Neural Network that could influence its performance.

Multi-residual Connections

It has been observed that multiple-residual connections improve the performance of a deep Convolutional Neural Network on the CIFAR-10 dataset.We hypothesize that using similar tranformations of previous stage inputs and outputs as well as future outputs, Residual Nets could perform well on the aesthetic classfication data.

Experiments

A pretrained Residual Network with 50 layers was finetuned on the Photonet images.Photonet dataset consists of around 17,300 images on a photographers site with aesthetics ratings given by photographers , similar to the dpchallenge.net dataset. This was tried in order to observe how a vanilla ResNet performed on the aesthetic image data. With a training set consisting of 10000 images ,4000 validation images and 3269 test images, an accuracy of 54.3% .This low accuracy could be because of the low size of the training set.

4.3 Emotion Prediction

To validate the enrichness of the crawled dataset, we propose object detection, scene classification based framework to support our claim of capturing required subjectivity. Emotion concepts in an image is highly correlated with salient objects depicted and their interaction with the remaining image. Large variance in object classes for each emotion type ensures large number of concepts covered.

For object classification, pre-trained ResNet-34 on ImageNet dataset is used. ImageNet[9] is a very large scale dataset for object classification and has 1000 object classes as per ILSVRC challenge. ResNet won ILSVRC 2015 with 26.73% top 1 error and 8.74% top 5 error. Images were scaled to 256x256 with adjusted aspect ratio and center cropped to 224x224. Number of object classes for emotion classes are listed in table 2

Table 2: Number of object classes for Emotion types

Emotion	Objects	Number of Images
Happy	968	47000
Sad	991	42980
Fear	993	72250
Anger	951	32100
Motivation	884	35350
Solitude	989	73100

Most Popular classes for each type of emotion are listed in table 3

For scene classification we plan to use pre-trained VGG16 Net on Places365 dataset. Places365 is a large scale scene-centric dataset which contains 365 scene classes. Classes include many indoor, outdoor, natural etc scenes. As a part of Places2 challenge, pre-trained CNNs were released by the organizers.

To verify performance on this dataset, we trained a modified VGG16[10] type Net with BatchNormalization[11] and Spatial Local Response Normalization(LRN)[12] layers. Our network consisted of four repeating blocks of Spatial Convolution Layers followed by ReLU, LRN and Spatial BatchNormalization layer. These blocks were followed by Fully three fully connected layers and after that one final output layers. Dropout[13] with probability 0.5 is used in first fully connected layers, and dropout of 0.6 is used in further

Table 3: Most Popular classes for emotion types

Happy	Sad	Fear	Anger	Motivation	Solitude
Groom	Window Screen	Electric Locomotive	Tench	Comic Book	Park Bench
Steam Locomotive	Band Aid	Crossword	Milk Can	Menu	Valley
Grand Piano	Matchstick	Microwave	Bannister	Revolver	Promontory
Palace	Prison	Projectile Missile	Matchstick	Academic Robe	Bubble
Half Track	Parking Bench	Matchstick	Steam Locomotive	Scuba Diver	Volcano

fully connected layers. Out of 300,000 images from the dataset, 150,000 images are used as training images, 50,000 images as validation images and rest 100,000 images as test images. As loss function we used standard Cross Entropy loss for multi-class classification with starting learning rate of 0.1, momentum of 0.9 and Stochastic Gradient Descent optimization. Learning rate is reduced to its half every 5 epochs. We were able to achieve an accuracy of 55% on test images. For a rudimentary classifier, this accuracy is decent and significantly larger than a random prediction 16%.

5 Future Work

Through our work we have been able to build modules which will be part of the final architecture. Since aesthetics and emotion are affected by past experiences we intend to incoprate memory into our model. This can be achieved by replacing the RNN in RAM with an LSTM[14]. Also the action network which is currently a linear layer followed by non-linear activation can be replaced with a small sized CNN which will help in the better classification of the patches read by the sensor. We intend to make our model more biologically inspired by having numerous parallel network with each network specialized towards a particular feature. For emotion dataset, our plan is to make it open source after we verify its variance in terms of number of scene classes. We also intend to come up with new architectures and models for emotion classification on this dataset. Our final motive is to make an open source website which can show images for different types of emotion to an individual based on her subjective experiences.

References

[1] Xin Lu,Zhe Lin,Hailin Jin,Jianchao Yang, James Z. Wang RAPID: Rating Pictorial Aesthetics using Deep Learning *MM '14 Proceedings of the 22nd ACM international conference on Multimedia*

[2] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, Charless Fowlkes ,*Deep Understanding of Image Aesthetics tion. In NIPS 2016*

[3] Mnih,Volodymyr,*Recurrent models of visual attention tion. In NIPS 2016*

[4] Xin Jin, Jingying Chi, Siwei Peng *Deep Image Aesthetics Classification using Inception Modules and Fine-tuning Connected Layer*

[5] Naila Murray, Luca Marchesotti, Florent Perronnin, *AVA: A Large-Scale Database for Aesthetic Visual Analysis*

[6] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba and A. Oliva, *Places: An Image Database for Deep Scene Understanding* arXiv:1610.02055.

[7] Kaiming He and Xiangyu Zhang and Shaoqing Ren and Jian Sun, *Deep Residual Learning for Image Recognition* arXiv:1512.03385

[8]Xin Lu Poonam Suryanarayan Reginald B. Adams, Jr. Jia Li Michelle G. Newman James Z. Wang, *On Shape and the Computability of Emotions* ACMMM 2012.

[9]Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei, *ImageNet Large Scale Visual Recognition Challenge* IJCV 2015.

[10] K. Simonyan, A. Zisserman, *Very Deep Convolutional Networks for Large-Scale Image Recognition* arXiv:1409.1556

[11]Sergey Ioffe, Christian Szegedy, *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift* arXiv:1502.03167

[12]Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, *ImageNet Classification with Deep Convolutional Neural Networks* NIPS 2012.

[13]Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever and Ruslan Salakhutdinov, *Dropout: A Simple Way to Prevent Neural Networks from Overfitting* Journal of Machine Learning Research 15 (2014) 1929-1958

[14]Sepp Hochreiter, Jurgen Schmidhuber, *LONG SHORT-TERM MEMORY* Neural Computation 9(8):17351780, 1997