The Usefulness of Class Activation Maps for Understanding Image Aesthetic Classifiers

David Chu
New York University
New York, NY
dfc296@nyu.edu

Varun Dhuldhoya
New York University
New York, NY
vyd208@nyu.edu

Abstract

Traditionally, automated aesthetic image classification has relied on handcrafted features. This meant that the creators of the classifier by necessity had to have some insight into the classifier behaviour. As in many other image classification problems, deep neural networks have proven to be very successful in aesthetic image classification. However, they are also much more opaque. In this paper we explore the usefulness of salience maps in understanding our classifier, as well as their use for image cropping.

1 Introduction

With the explosion of the number of images available online today, it is increasingly important to have some automated method of separating aesthetically good images from bad images. However, understanding the aesthetic qualities of an image (whether images are good looking or not) is a fairly subjective process. As a result, traditional methods of automated classification normally leverage subject matter experts to create hand-crafted features. As with other image classification tasks, deep neural networks have become the favored approach towards aesthetic image classification. Unfortunately, this has also meant that the features that are learned are no longer as comprehensible to the users and creators. The neural networks learn weights between layers, but these weights may not be meaningful to us as humans.

In this paper we seek to utilize class activation maps (specifically GradCAM), to better understand the spatial aesthetics as the classifier understands them. To do so, we implemented two CNN image classification networks, and generated activation maps on their outputs. Our two models were a fine tuned VGG-16 model, and a two headed CNN model.

Furthermore we explore potential uses of these activation maps for cropping. We also examine whether some features of these activation maps can be used to give us confidence that these activation maps are in fact accurate.

2 Related Works

In this section we outline the relevant background works for image aesthetic classification.

Data

This project utilizes a sample of the Aesthetic Visual Analysis (AVA) dataset [2]. This dataset contains about 250,000 images scraped from www.dpchallenge.com. Images are submitted as part of challenges, and each image is given a score ranging from 1 to 10 by each user. In the dataset, each image is annotated with it's score distribution (number of 1's, 2's, etc... the image received). Furthermore, about 200,000 images are annotated with at least a one of 66 textual tags (Portraiture, Nature, Military, etc…).

Models

Several deep neural network models have been proposed for carrying out Aesthetic image classification. The approach from [5] utilizes a two headed CNN classifier, with one head trained to do a binary classification task of high/low aesthetic value while the other head is trained to classify the textual tag. Furthermore, this approach utilizes a per-sample weighting of the loss so that images with a very high or low rating contribute more to the training of the network. Another approach [4] attempts to learn the score distribution of images instead of a single target classification.

Class Activation maps

For generating our class activation maps, we focus on Grad-CAM [3] as a more generalized form of class activation maps. Grad-CAM removes the model constraints of the original class activation map generation technique. Furthermore, research [1] has indicated that it might be possible to use the mean pixel intensity values of our maps to derive useful information. We experiment to see whether this is the case.

3 Data

While the original dataset has about 250,000 images, due to compute constraints, we have used a smaller sample of the data based on the train/test sets used by the original authors for photographic style classification. The original train/test set used by the author contains 11,270 training images and 2809 test images. We further filtered out any images which did not have at least one textual tag, resulting in 8900 training images and 2203 test images. The scores for the training images are normally distributed around a mean value of 5.4 (see Figure 1). Additionally, textual tags are heavily skewed, with some having a very low number of occurrences (see Figure 2).

For binary classification, we made an assumption that images that had an average score greater than 5 would be considered high quality, while images less than or equal to 5 would be considered low quality. However, this resulted in a highly imbalanced dataset with 2495 low quality images and 6405 high quality images. Without any mitigation, we found that this imbalance resulted in a classifier quickly learning to always predict high quality images. Therefore we used an oversampling strategy to ensure that low quality and high quality images were chosen equally from the training set. This was done by using weighted random sampling with replacement from our training set to create training batches.

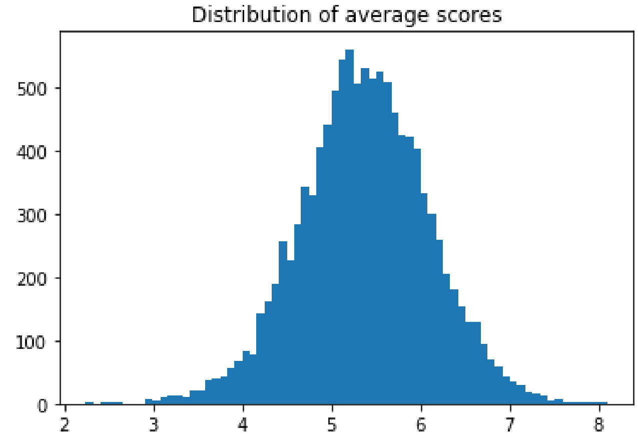

Figure 1: Average Score distribution of training set

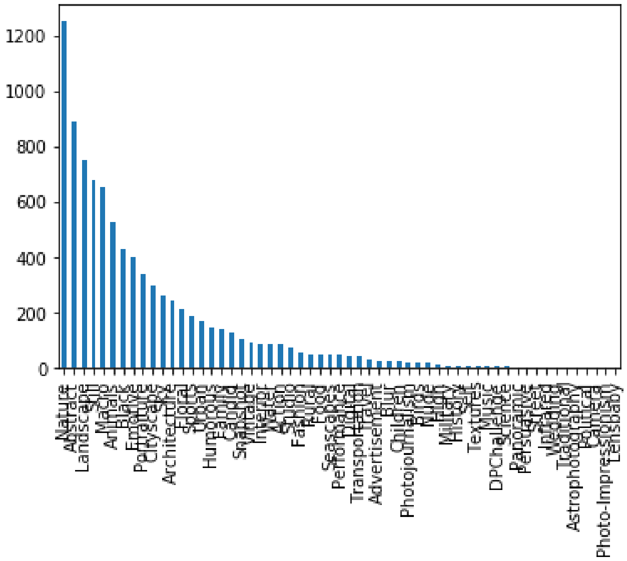

Figure 2: Textual tag distribution of training set

For the textual tags, we did not attempt to address the imbalance. In order to resolve this issue, we could try in the future to incorporate other data sets that contain images from missing categories.

Prior to being fed through our model for training, images were resized to 256x256, then a random 224x224 crop of the image was taken. For testing, we only resized the images to 224x224. While resizing images can destroy some spatial information, this was a necessity due to the feature encoders we chose. In aesthetic image classification, the norm is to not do normalization as would be expected in some other types of classification problems as that can further distort the spatial aesthetics.

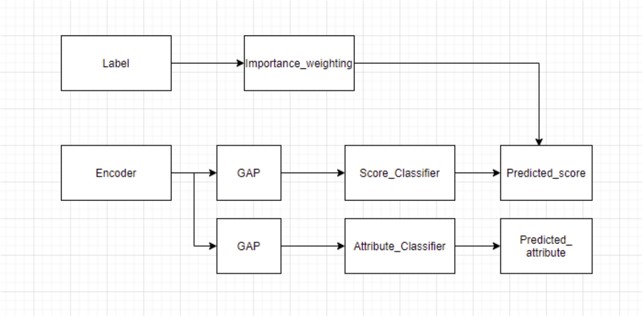

Figure 3: Training process

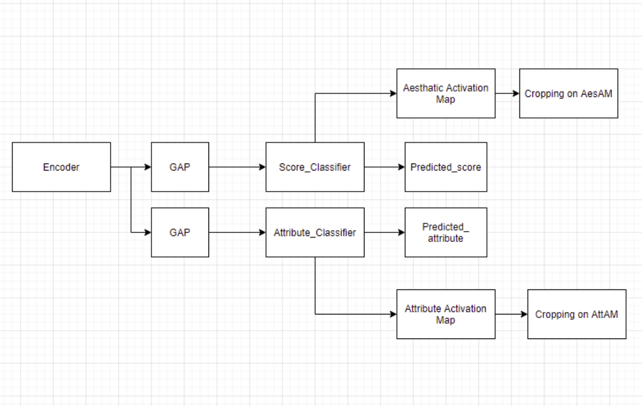

Figure 4: Testing process

4 Methods

4.1 Overview of proposed model

The network is composed of four parts: feature encoder f_{enc}, global average pooling f_{gap}, aesthetic level classifier f_{hl} and aesthetic attribute classifier f_{att}. For generating activation maps we used an implementation of GradCAM.(as shown in Figure 3 and Figure 4)

Image classification in deep learning generally includes two parts: feature extraction layers f_{enc} and classifier learning layers f_{hl} or f_{att}. In the training phase, our model includes two branches of sub-network: the aesthetic level classifier f_{hl} with high/low level supervision, and the attribute classifier f_{att} with attribute category supervision. Given an input image I, the network first extracts the features using θ_{enc}, where θ_{enc} is the feature encoding parameter. Then the f_{gap} vectorizes the output from the last convolutional feature map by summing each channel's map. Finally this vectorized result is passes to the the two classifiers f_{hl} and f_{att}

4.2 Re-weighting the samples

As we can see for the AVA dataset(see Figure 1) more than 80% of the samples have their average score lying in the intermediate range(between 4.5 and 6.5). Hence sample specific weight needs to be assigned to each sample to balance it's importance in the training phase.

We experiment with two types of sampling

- Assigning weight based on the class labels
  We assign labels to each sample(1 if their avg score is >5 and 0 if its <=5) . We count the total number of samples for each class and reweight the samples as weight for each class W_i = 1/(number of samples belonging to class i)
- Assigning weight based on the average score
  We assign weights to the samples based on their average score. If the average score lies between 4.5 and 6.5 we assign 1 to the sample but if the average score of the sample lies outside this range we assign 7 to it

4.3 Models

4.3.1 VGG16

Our baseline model is a fine-tuned VGG16 network with a binary classification layer. We freeze the CNN layers and only train the classifier layers to do high/low aesthetic classification. We use a weighted binary Cross Entropy loss as described above. Samples with a score less than 4.5 and greater than 6.5 have their loss multiplied by 7 to ensure that the network learns more from highly/lowly rated images than from images with an ambiguous score.

4.3.2 MobileNet V2

MobileNet V2 is a significant improvement over its predecessor MobileNet V1 which uses depthwise separable convolutions. It significantly reduces the number of parameters when compared to the network with regular convolutions with the same depth in the nets. This results in lightweight deep neural networks. We use the MobileNet V2 as a baseline model and freeze the CNN layers and only trained the classifiers. We add two classifiers to the baseline model:

- Aesthetic classifier that predicts the probability distribution of scores from 1-10 using softmax activation(based on [4])
- Attribute Classifier with Relu activation to predict attribute associated from the image(from 0-66)

For the Attribute classifier we use cross entropy loss function but for the Aesthetic classifier instead of using cross entropy we use Earth Mover's Distance loss since it is proven to work on ordered buckets.

4.3.2.1 Earth Mover's Distance Loss

Earth Mover's Distance is defined as the minimum cost to move the mass of one distribution to another. Soft-max cross-entropy is widely used as training loss in classification tasks.

However, in the case of ordered-classes (e.g. aesthetic and quality estimation), cross-entropy loss lacks the inter-class relationships between score buckets. Training on datasets with intrinsic ordering between classes can benefit from EMDbased losses.

These loss functions penalize mis-classifications according to class distances.

We calculate EMD loss as

$$EMD(p,\hat{p}) = (\frac{1}{N}\sum_{k=1}^{N}(CDF_p k - CDF_{\hat{p}} k)^r)^{1/r}$$

where CDF is the cumulative density function.

4.4 Experiments and Visualizations

4.4.1 Model Training

For VGG16, we fine tune the pretrained pytorch VGG16 model using stochastic gradient descent and weighted cross entropy loss. We get a validation accuracy of 72%. We then generate class activation maps using GradCAM[3].

For MobileNet v2 we run the model for 50 epochs using Adam's optimizer with a learning rate of 3×10^{-6}. We get a validation accuracy of 75% using this model. We then use this model to generate class activation maps for both the Aesthetic and Attribute classifiers using GradCAM[3]. After this we generate bounding boxes for image cropping based on the Attribute and Aesthetic class activation maps.

4.4.2 Aesthetic class activation maps

The activation map for the highest rated image in the validation dataset. As we can see a large portion of the image is included in the image crops

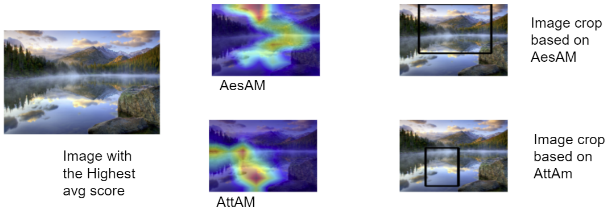

Figure 5: Activation maps for the highest rated image

The activation map for the lowest rated image in the validation dataset. As we can see only a small portion of the image is included in the image crops

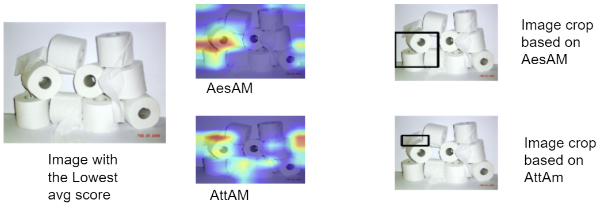

Figure 6: Activation maps for the lowest rated image

4.4.3 Image Cropping as an application

One of the potential application of generating bounding boxes is to generate a aesthetically pleasing crop. This can be done in two ways. The first is to generate a cropped image based on the

Aesthetic bounding box alone. Another method is to calculate the average bounding box based on both the attribute and aesthetic bounding boxes, but The 66 attributes have very complex relation. The scope of attribute information is very general, in which many attributes come from life experience or from different point of view such as science and technique, macro, Emotive, Abstract, History, Humorous, Political, and this may cause some of the results to be unintuitive.

4.4.4 Activation maps intensity values

Taking an idea from [1], we examine whether a value can be derived from class activation map itself that corresponds to the aesthetic value of the image. The intuition behind this idea is that highly rated images should also correspond to highly activated maps with respect to the high aesthetic class. Conversely, lowly rated images should have lowly activated maps. We generate a value by taking the scalar mean of the pixel intensity (a value between 0 and 1) of the class activation map. Plotting this value with the avg score, we see a very minor positive relationship (see Figure 7). However, when we take this value and check its Pearson correlation score with the avg score of the image seen in Table 1, the results show that the correlation is quite weak. Furthermore, attempting to use the value as split point for a simple classifier where images with a mean map intensity value greater than the split point are classified as high quality and all others as low quality leads to only 49.7% accuracy.

These results casts some doubt onto the validity of the class activation maps being generated. Further research needs to be done to compare different visualization techniques with respect to explaining aesthetic image classification.

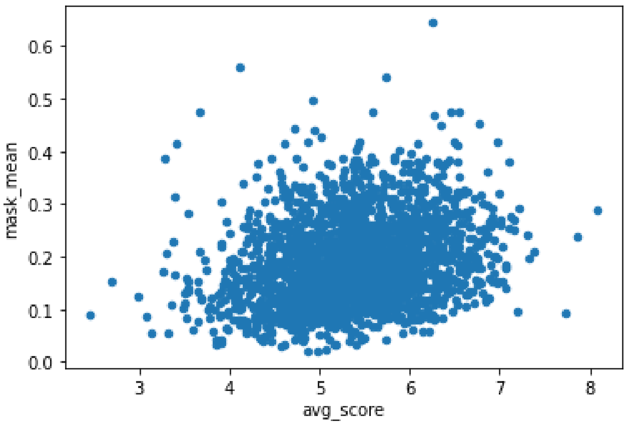

Figure 7: Scatter plot showing relationship between average score and mean pixel intensity of class activation map

	All Images	Images with score <4.5 or >6.5
Pearson's r	0.24	0.34

Table 1: Pearson's correlation coefficient between average intensity of class activation map and average score

4.4.5 Activation maps for score classification

We generated activation maps for each score from 1-10 in the Aesthetic classifier to try to explain which areas of the image lead to certain scores for the image. What we observed was that for the images with scores in the midrange(from 4.5 to 6.5) the activation maps were not clear but for images with high or low scores the activation maps gave indications as to why the image had a particular score.

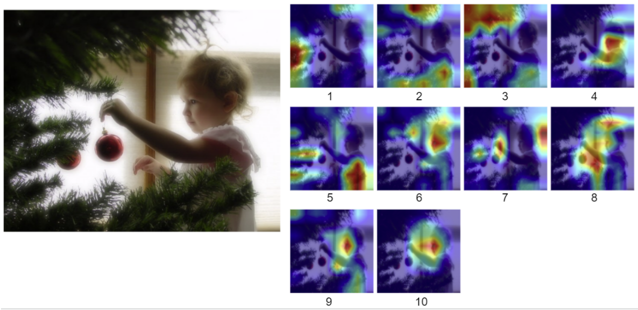

Figure 8: Activation maps for scores from 1-10

5 Conclusion

In this study we examine the usefulness of class activation maps for explaining the aesthetic values of an image. We then attempt to quantify the correctness of the class activation maps by generating a value that should be correlated with score. Finally we propose a potential application of class activation maps for image cropping.

Activation maps are an intuitive way for people to understand what a classifier is "looking" at when it comes up with its classification. By overlaying the activation maps on top of the original image, our eyes are naturally drawn to the hot spots. However, related research as well as our own experimentation with pixel intensity and score correlation suggests that we should be careful to verify that the maps being generated are meaningful.

References

[1] Xukun Li, Huaiyu Zhang, Doina Caragea, and Muhammad Imran. 2018. Localizing and Quantifying Damage in Social Media Images. *CoRR* abs/1806.07378 (2018). arXiv:1806.07378 http://arxiv.org/abs/1806.07378

[2] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2408–2415. https://doi.org/10.1109/CVPR.2012.6247954

[3] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 618–626. https://doi.org/10.1109/ICCV.2017.74

[4] Hossein Talebi and Peyman Milanfar. 2018. NIMA: Neural Image Assessment. *IEEE Transactions on Image Processing* 27, 8 (2018), 3998–4011. https://doi.org/10.1109/TIP.2018.2831899

[5] Chao Zhang, Ce Zhu, Xun Xu, Yipeng Liu, Jimin Xiao, and Tammam Tillo. 2018. Visual aesthetic understanding: Sample-specific aesthetic classification and deep activation map visualization. *Signal Processing: Image Communication* 67 (2018), 12–21. https://doi.org/10.1016/j.image.2018.05.006