Does composition matter? A study on the impact of removing macro-level indicators of aesthetics from images by zooming in

s4804902
Radboud University, Nijmegen, Netherlands

ABSTRACT

The aim of this paper is to give insight into the question of what makes a beautiful photograph. We focus on the impact of leaving out image-wide indicators of aesthetics - such as composition and subject - by zooming in on pictures. To this end, we train a Neural Network to discern between aesthetically high - and low quality images (as judged by humans). We then compare its performance on full size images and zoomed-in fragments. We address the question of where to zoom in on the image, comparing random zoom with an attention-based zooming method. Using attention-based zoom, we find that we can reduce images by over 97% in size while only losing 1% of Neural Network performance, indicating that image-wide indicators of aesthetics might not be so important.

1 INTRODUCTION

We define the aesthetics of a photograph by how well it is liked by humans. Judging the aesthetics of a photograph will always be, to a certain extent, subjective and up to the taste of the individual. It is not entirely subjective either, because there are qualities like composition, subject, lighting, sharpness and texture which do bear some kind of objectivity. Some of these qualities, like composition and subject, can only be seen when looking at the whole image. Others, such as sharpness, and texture can not only be seen when looking at the whole image, but are also present at pixel-level. Still others, such as lightning, fall somewhere in-between these two groups, which we will call *macro-level* and *micro-level indicators of aesthetics*. If we zoom in on a picture, we lose its macro-level indicators, but we still have micro-level indicators through which we can assess its quality. I want to give an insight into the importance (or lack thereof) of macro-level indicators, by comparing full pictures with zoomed-in fragments.

So how does this comparison work? First of all we will need a dataset of images that are either good or bad aesthetically (I will use 'good' and 'bad' from now on for convenience). The only way to attain such labels is by using human judgement, so we will say that an image is good (bad) if the average score of a large group of judges is high (low). The perfect dataset for this is the AVA-dataset [4]. It consists of 100.000s of pictures sent in for a photograph competition, all judged by fellow participants and the general public. Usually images have hundreds of judgements, making the mean score an accurate reflection of their aesthetics.

Now that we have judgements for a large collection of pictures, we could conduct the following experiment: From each image, select a small fragment and show it to a group of human judges. Then we ask the question, do you think this fragment is part of a good or bad image? If the human judges are able to answer this question with high accuracy, this would indicate that the lack of macro-level indicators is not severe, and thus that these indicators might not be important. Oppositely, if the human judges fail to classify the fragments accurately, we could conclude that macro-level indicators are essential to photography. Since finding a big group of human judges is costly and time-consuming, I opt for a different approach using a Convolutional Neural Network (CNN), as they have recently emerged as excellent models for image recognition. For my work, I train a CNN to discern between good and bad images, and then compare it to a CNN trained on zoomed-in fragments of said images. The drop in performance of the CNN (measured by classification accuracy) will indicate the gravity of not having macro-level indicators.

Another question I want to address is how to select the zoomed-in fragments. If we are unlucky, we zoom in on a part that is not very important to the photograph. We especially do not want to zoom in on the background, because it will not be very informative of the aesthetics of the full picture. Therefore, we want to establish the *focus* of the picture in order to zoom in on it. This is no trivial task, but fortunately we can use the CNN trained on the full images size AVA-dataset. This is because when a CNN makes predictions, some pixel values are weighed more heavily than others. The most heavily weighed pixels form the focus of the pictures, which we can extract using a method called GRAD-Cam [5].

2 RELATED WORK

This research is belongs to the field of *Automated Image Aesthetics Assessment* (AIAA). This field concerns itself with using computers to judge images as accurately as possible based on aesthetics. Earlier works explored the use of hand-crafted features, citing as inspiration "of thumb in photography, common intuition, and observed trends in photography ratings"[2] as input for a simple classifier. Extracting these features is time consuming, and it is difficult to predict which features will perform well. Therefore, when computer vision made its advance, researchers made use of deep learning to automatically extract the most relevant features from an image, instead of having to compute them manually. An example of such an approach can be found in [3]. Some well known networks for computer vision are VGG19, Xception [6], [1]. [1]

In this work, I combine AIAA with network attention, so that we can find the focus of each picture needed to determine where we need to zoom in. Network attention comes from the notion that a Neural Network does not weigh all its inputs equally, i.e. the network pays more attention to some inputs than others. This can be compared to human vision, where we receive sensory input in the form of photons, from which only a small percentage is

TxMM'21-'22, Radboud University, Nijmegen, Netherlands

[1] although I could only find the arXiv papers, both papers have been cited thousands of times which makes me believe in their credibility

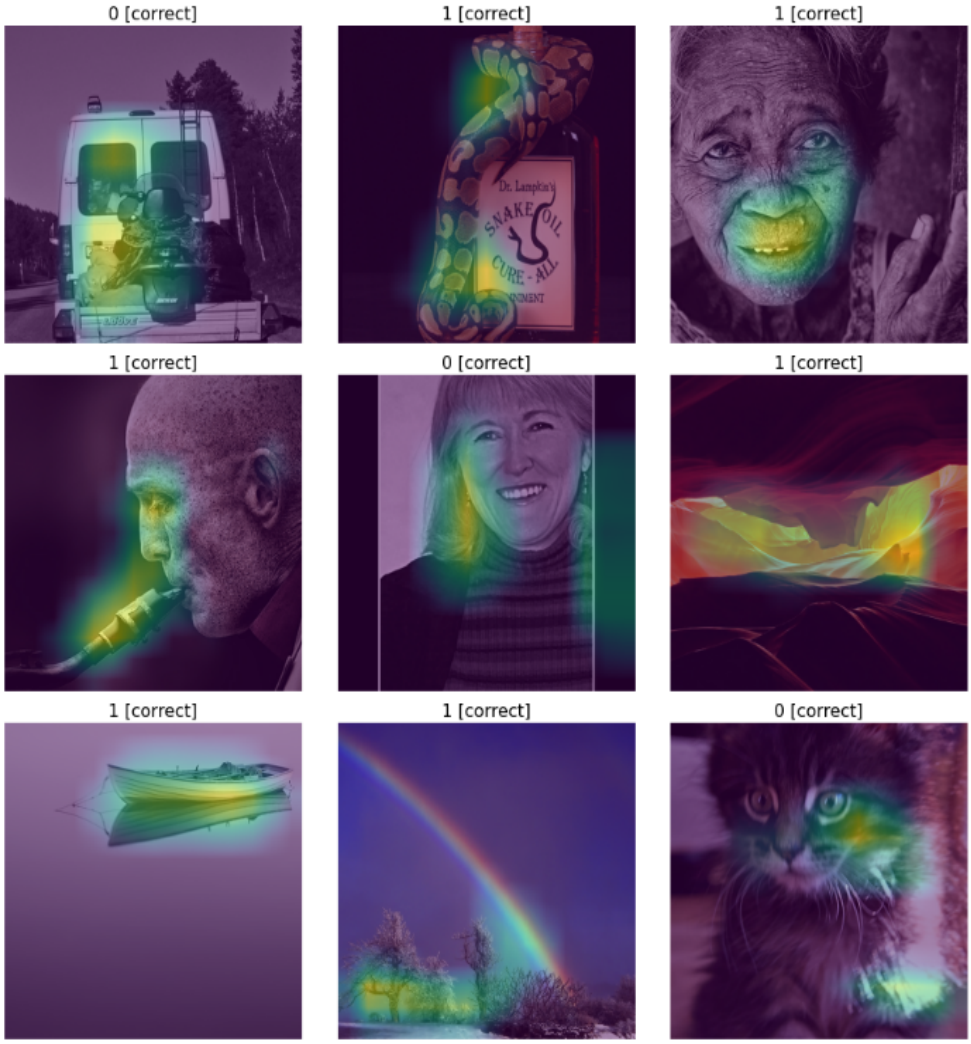

Figure 1: 9 examples of automatically finding the focus in an image. Highlighted parts are the parts that the CNN used for its decision making. The predictions are 1 for good and 0 for bad images, and the [label] indicates the correctness of the predictions

paid attention to. Attention has been used for a wide variety of purposes, and I will focus on methods that visualise attention for image recognition. One example is Grad-CAM [5], which can be applied to a wide variety of networks without having to change their implementation.

3 APPROACH

My approach consists of 4 steps:

(1) Data collection
(2) Training of network on full images
(3) Zooming in with network attention
(4) Training of network on zoomed-in fragments

3.1 Data collection

The data used for this task is from the AVA-dataset [4]. Pictures were collected using their dpchallenge.com ID, and have a labels that is their average rating. As per recommendation of the original authors, a parameter δ is introduced to remove aesthetically ambiguous images: only photos with an average rating of over 5.5+δ or under 5.5-δ are considered. For this work, δ is set to 1.5. This gives us only pictures rated 7 or higher (labelled with a '1') and pictures rated 4 or lower (labelled with a '0'). After this filtering, the dataset contains 11040 images.

I have mentioned that I train a network on the full image first, but this has a caveat: Neural Networks expect every input image to have the same resolution. In [3], this challenge is tackled by mixing different techniques such as cropping, padding and warping images. I opt for a simple approach: Large images get cropped and small images get padded, such that all images have a resolution of

480x480 pixels. 480x480 was chosen because many images in the dataset have a resolution of either 640x480 or 480x640.

3.2 Training the network

In this step , I train a CNN to judge whether a picture in our dataset is good (rating of 7 or higher) or bad (rating of 4 or lower). For this I use the Xception model, a CNN architecture used for image classification. I chose Xception because it is expressive (over 20.000.000 parameters) but still computationally efficient. I make use of the implementation provided by Keras [2]. In order to make Xception fit for a 2-class problem, I replace the top dense layer with a dense layer of two nodes. Dropout has been added to prevent overfitting, with a dropout rate of 0.2. The 11040 images were split into a training set of size 8970 and a validation set of size 2070. The loss function used is Categorical Cross-entropy, which is the default for image classification. Training has been sped up using TPUs provided by Kaggle. The network trains in approximately 10 minutes. The full code of this training phase is available at Kaggle [3].

3.3 Grad-CAM

In order to find the parts of the images that the CNN focuses on, I implement Grad-CAM to the previously described model. Grad-CAM monitors an input image as it goes through our model, and scores pixels according to their contribution to the final output.

Grad-CAM gives every pixel a relevance score, which can be plotted as a heat map over the original image (figure 1). This visualises which part of the image is most informative to the network when predicting the label of the image. We call the highest scoring pixel the *centre of attention*. I crop the original image to a 72x72 size patch, with the centre of attention in the middle. This is a reduction of over 97% with respect to the original 480x480 resolution. 72 was chosen because 72x72 is the smallest resolution that is accepted by Xception. If the centre of attention is more than 72 pixels removed from one of the edges of the image, a black padding is added to ensure the right dimension. This was not a very big concern, since empirically the centre of attention was usually far removed from the edge. A possible explanation is that the pixels at the edge are present in less convolutions (no padding was used in the CNN), and thus can not contribute as much to the output. Code for this phase is available at Kaggle [4]

3.4 Training the network, again

After we collect the 11040 patches, we train the exact same architecture used earlier, but now on the 72x72 fragments. Since the images are smaller, training time reduces to approximately one minute. I also train a copy of the same network on 72x72 fragments that have been cropped from the centre of each image, in order to determine the impact of the attention-based zooming. Code for this phase is available at Kaggle [5]

[2]https://keras.io/api/applications/xception/
[3]https://www.kaggle.com/code/matthijsneutelings/create-model
[4]https://www.kaggle.com/code/matthijsneutelings/create-patches
[5]https://www.kaggle.com/code/matthijsneutelings/patches-model/

Table 1: Accuracy on full vs zoomed-in images

Patch size	Validation accuracy
480x480	93.9%
72x72	92.5%

Table 2: Zooming method comparison

Crop strategy	Validation Accuracy
Centre-of-attention crop	92.5%
Middle-crop	84.0%

4 RESULTS AND ANALYSIS

I report the accuracy on the validation set for both the model trained on 480x480 and 72x72 images in table 1, where accuracy is the fraction of correctly predicted images.

We see that even with only a fraction of the image, we can predict the aesthetic value. A reduction of over 97% in image size lead to an accuracy decrease of little over 1%. This suggests that even without the macro-level indicators, images are separable on aesthetics by micro-level indicators.

Furthermore, I investigate the usefulness of the zooming strategy by comparing its attained accuracy with another zooming method: taking the 72x72 middle pixels from every image. Results are posed in table 2, and suggest that my zooming method does have a big effect (losing 1.4% accuracy instead of 9.9%). An illustration on why my zooming technique works better is the bottom left image in figure 1. The centre-of-attention crop will be on the boat, which contains a lot of information on the aesthetics of the picture. On the other hand, if we would have selected the middle patch, we would have a patch full of background water that makes it difficult for our model to decide its aesthetics.

5 DISCUSSION

I have shown that we can assess the aesthetics of images only by their micro-level indicators, with only a small fraction that gets misclassified. On the other hand, any photographer will likely tell you that macro-level indicators like composition and subject do matter. What this work really shows is that there is a lot of information on image aesthetics available at micro-level, almost as much as in the full image when we look at the numeric results.

5.1 limitations

No work is without limitations, and this work is no exception. I encountered the following limitations during my research:

- Even though this work talks about 'full' images, the nature of CNNs makes it so that every image has to have the same resolution. Since the original resolutions are not equal, I used cropping and padding in order to make all the images have the same 480x480 resolution. Nevertheless, I would argue that the 9 images in figure 1 are large enough to call them 'full' images.

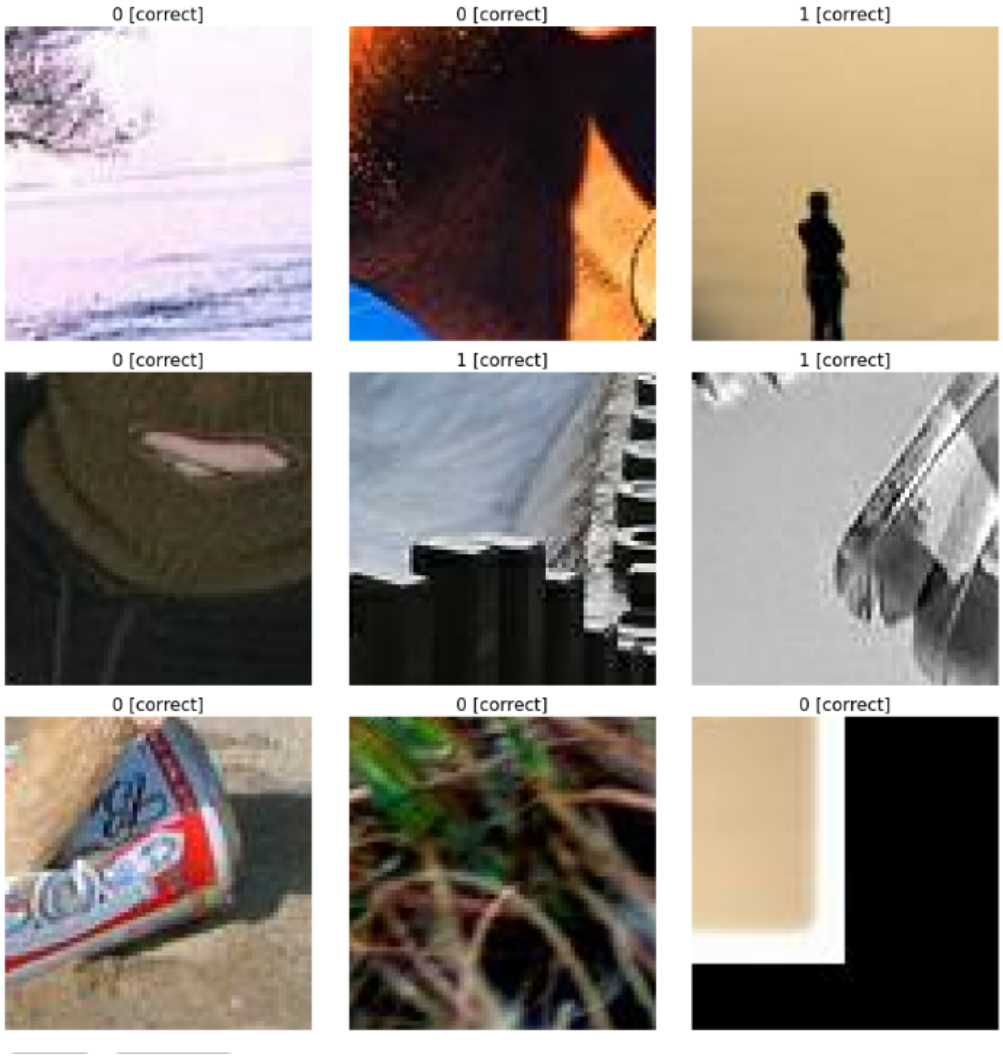

Figure 2: 9 examples of zoomed-in fragments with the attention method. Notice that for most images, any notion of composition or subject is gone. The photo on the bottom right had to be padded. As in figure 1, a 1 indicates a predicted good image and a 0 a predicted bad image, with the [label] indicating the label's correctness

- Some of the ratings of images in the dataset are skewed, because images can get judged on other things than their aesthetics. In particular, people also judge images based on whether they fit the theme of the challenge they were submitted to. This lead to images that were labelled as bad, when in reality they were of average or even good aesthetics. However, this seems to apply only to a small fraction of the images in the dataset.
- setting δ to 1.5 ensured we could only use about 10% of the full AVA dataset. Choosing a lower delta value would have lead to a substantial increase in training data. This would probably have reduced the accuracy, since there would be more 'average' images in the dataset.
- possibly the biggest limitation is the minimum resolution of 72x72 required by Xception. Even though this is already a substantial reduction of the full image size, performance on even smaller patches could not be tested.

6 CONCLUSION AND OUTLOOK

In this work, I have shown that the removal of macro-level indicators of aesthetic quality from images by zooming in on them has very little impact on automated image aesthetics assessment. This shows that even small fractions of images contain enough information to make an accurate prediction of their quality. Furthermore, I have shown that we can not get the same results for any zooming strategy, and that zooming in on the focus of the image (as opposed to the background) is crucial in maintaining accuracy. To this end, I have shown that gradient-based attention can be used to automatically find the focus of a picture.

6.1 future work

For future work, it would make sense to see what the results are when reducing patch size even further. With a CNN, VGG has a minimum convolution size of 32x32, resulting in patches over 5 times smaller than the ones we used here. If one wants to zoom in even further, a different architecture will have to be used. It might also be interesting to see if we can discern images based on single pixels.

REFERENCES

[1] François Chollet. 2017. Xception: Deep Learning with Depthwise Separable Convolutions. arXiv:1610.02357 [cs.CV]

[2] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. 2006. Studying Aesthetics in Photographic Images Using a Computational Approach. In *Computer Vision – ECCV 2006*, Aleš Leonardis, Horst Bischof, and Axel Pinz (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 288–301.

[3] Xin Lu, Zhe L. Lin, Hailin Jin, Jianchao Yang, and James Zijun Wang. 2014. RAPID: Rating Pictorial Aesthetics using Deep Learning. *Proceedings of the 22nd ACM international conference on Multimedia* (2014).

[4] 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

[5] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2019. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* 128, 2 (Oct 2019), 336–359. https://doi.org/10.1007/s11263-019-01228-7

[6] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs.CV]