UKRAINIAN CATHOLIC UNIVERSITY

BACHELOR THESIS

On Methods of Estimation of Image Aesthethic

Author:
Iryna KOSTYSHYN

Supervisor:
Oles DOBOSEVYCH

A thesis submitted in fulfillment of the requirements for the degree of Bachelor of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2020

Declaration of Authorship

I, Iryna KOSTYSHYN, declare that this thesis titled, "On Methods of Estimation of Image Aesthethic" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Bachelor of Science

On Methods of Estimation of Image Aesthethic

by Iryna KOSTYSHYN

Abstract

Evaluation of image aesthetics has been a longstanding problem in image processing and computer vision. Nowadays this topic is discussed more than ever, researchers attempt not only to measure the aesthetic quality of images but also to find new applications of this methods. In this work, we research image assessment methods, their pitfalls. We try to propose new assessment methods for these tasks using activation maps and explore the possibility of using such neural networks for training Image Restoration GANs. This work also incorporates the study of the relevance of existing methods on the example of their work on data not specially prepared for this specific task.

Acknowledgements

I would like to thank my supervisor Oles Dobosevych for his dedicated support and guidance. I also wish to thank the Ukrainian Catholic University for my solid academic training.

Contents

List of Figures

List of Abbreviations

ANN	**A**rtifitial-**N**eural **N**etwork
CNN	**C**onvolutional-**N**eural **N**etwork

Dedicated to my dear sister

Chapter 1

Introduction

1.1 Motivation

Amount of digital photos is growing every day, so visual aesthetic assessment for images and videos become a hot topic. Aesthetic assessment tasks are always associated with judging the visual and aesthetic quality of pictures and videos. Image quality assessment tries to represent the human perception of quality according to specific rules generally agreed by human visual perception. Evaluate image aesthetics can be a very tough and challenging task. Not every person can find the same image aesthetically pleasant.

Image aesthetic assessment metrics can be used in many applications from analyzing the performance of algorithms in different fields of computer vision like image compression, image enhancement, and image processing. The industry has also become highly interested in image assessment methods as it can improve overall user experience, so it keeps users satisfied and business profitable. For example, social platforms with user generated content use this to recommend users videos or images that more aesthetically pleasing to get more regular users to the platform.

It is also a useful application on streaming platforms. For example, Netflix makes extensive use of such approaches in its product. Such algorithms evaluate the main banners for movies, the placement of subtitles, so that does not overlap important details or the text in the image, and so on.

1.2 Problem

Following the recent advances in deep convolutional neural networks, researchers have investigated various data-driven learning-based methods for aesthetic assessment and have reported remarkable results in the past few years [10, 14, 5]. To train a neural network on image aesthetic assessment avoiding subjectivism, we need a massive diverse dataset with specific labels. Datasets available for this task are usually created in an idealistic way consisting of high-resolution images. The methods trained on such datasets can be biased and do not work with real-world data.

Inspired by the recent achievements on aesthetic assessment in work "Revisiting Image Aesthetic Assessment via self-supervised Feature Learning [11]", we state a hypothesis that image aesthetics can be highly correlated with other highly currently spoken topics in Computer Vision and Deep Learning. We decide to study a correlation of aesthetic image level with the methods proposed for removing different image degradations, i.e., haze, moire, blur. To this extent, we interested if the quality of the image correlates with the beauty level of the image and how this problem can be developed with the help of existing SOTA datasets for image degradation removal tasks.

1.3 Goals of the bachelor thesis

The main objective of the thesis is to explore aesthetic assessment methods, research the weaknesses and problems of currently existing approaches. Besides the main objective, we highlight related objectives:

- To study a correlation between image distortions and aesthetic.
- Hypothesis confirmation that image aesthetic assessment can be used in image restoration tasks.

In this work we make a contribution into solving the problem stated above, by doing a comprehensive research of the current datasets and approaches situation by following the next structure:

- To make an exploration on the existing approaches to aesthetic assessment.
- To make an evaluation of image aesthetic assessment methods on real-world data.
- To use interpretability algorithms to explore neural network and make assumption about the neural network.
- To make a survey to check the hypothesis whether the image quality is necessary correlates with image attractiveness aesthetic level.

Chapter 2

Background information and theory

2.1 Neural Networks

Artificial neural networks are systems that consist of nodes and connections between them. Such nodes are usually referred to as neurons. Neurons are united into layers and have connections to other neurons in the previous layer. A neuron can have a different amount of connections. Fully connected layers are layers, each neuron of which has connections to all neurons in the previous layer. To each connection designated coefficient, so-called weight or weight coefficient, which controls an amount of information which should be transferred to another neuron. The output of Every neuron is modified by the activation function.It is also known as Transfer Function.

The Activation Functions can be of 2 types.

- Linear Activation Function
- Non-linear Activation Functions (E.g. Sigmoid or Logistic Activation Function, Tanh or hyperbolic tangent Activation Function, ReLU (Rectified Linear Unit) Activation Function, Leaky ReLU and others)

The concept of artificial neural networks was created inspired by a structure of the animal brain. Neural Networks can learn to perform tasks using preceding specific training, usually without being programmed with particular rules. To train the network means to estimate the best weights in all neurons to minimize the error between the predicted by the network outcome and correct outcome. The prediction error of a neural network is called a loss. It is used to calculate the gradients which are used to update the weights of the neural network.

2.2 Convolutional Neural Networks

Convolution neural network is a deep neural network architecture that commonly used in the computer vision field to solve various tasks as image classification, object localization, pattern recognition, etc.

Convolutional networks are neural networks that perform specific kind of linear operation - convolution in place of general matrix multiplication in at least one of their layers. Such a layer is called the Convolution Layer, and it is considered as one of the main types of layers that are essential for CNN architecture.

The convolution operation is shown in Figure 2.1, mainly performs dot products between the filters (kernels) and local regions of the input. After applying the filter to an image a feature map is created. As shown in the image. The objective of the

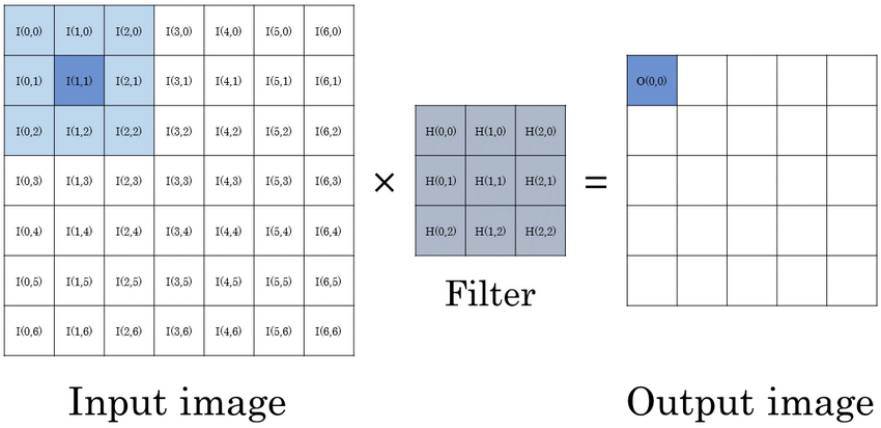


FIGURE 2.1: Operation of convolution
Source: https://www.researchgate.net

Convolution Operation is to extract the high-level features from the input image. Conventionally, the first Convolution Layer is responsible for capturing low-level features such as edges, color, gradient orientation, etc. By attaching more convolutional layers the neural network is able to capture the high-level features such as eyes, palms or wheels to identify the object.

Besides Convolution Layer, there are two more main types of layers: Pooling Layer and Fully-Connected Layer.

The main idea of applying Pooling is to reduce the spatial size of the representation, in other words, to reduce the number of parameters (features). Pooling Layer is useful for denoising and extracting dominant features. There are two types of Pooling (as shown in Figure 2.2): Max Pooling and Average Pooling. Max Pooling returns the maximum value patch on the feature map. Average Pooling returns the average of all the values from the patches on the feature map.

The main purpose of a fully connected layer is to make a linear combination of the extracted features from convolution/pooling process to obtain a prediction score.The output is flattened into a single vector of values, each representing a probability that a particular feature applies to a label.

2.3 Similarity learning

Similarity learning is an area of machine learning which is mostly used for semi-supervised setups. Similarity learning makes it possible to measure similarity between two elements of the same set. It has applications in face verification field, ranking and recommendation systems, etc.

There are several learning setups: Regression similarity learning, Classification similarity learning, Ranking similarity learning and Locality sensitive hashing (LSH). Ranking similarity learning is a commonly used approach in image/video aesthetic assessment. The idea behind this is to find out whether inputs are similar or dissimilar. The objective of the similarity learning is to learn to predict relative distances between data inputs.

There are two kinds of Ranking Losses: When we use pairs of training data or triplets of training data. Both losses setups compare distances between embeddings (representations) of input data samples. Losses that are adjusted to use pairs in the

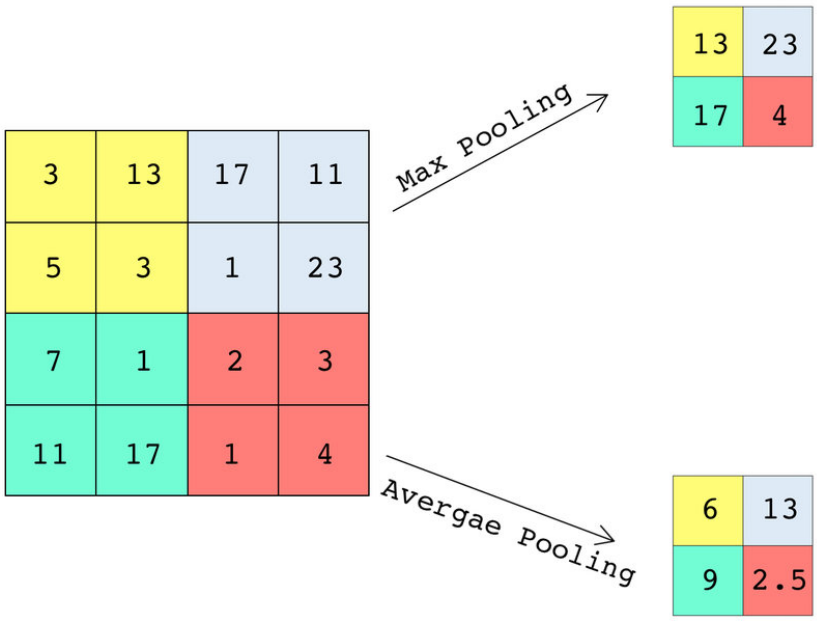


FIGURE 2.2: Types of pooling in CNN
Source: https://www.researchgate.net

training procedure called Pairwise Losses. In case of using pairwise losses, we use positive and negative pairs of training data points.

Positive pairs are composed of a sample x_a so-called anchor and a positive sample x_p and negative pairs consisting of an anchor sample x_a and a negative sample x_n. x_p and x_a should be similar in the metric we aim to learn and x_n should be dissimilar to x_a in that metric.

The aim differs depending on whether the pair is positive or negative. For positive pairs, it is to learn representations with a small distance d between them, where for negative pairs it is a greater distance than some margin value m. Pairwise Ranking Loss enforces 0 distance in representations for positive pairs, and a distance greater than a margin for negative pairs. Where r_a, r_p and r_n are the samples representations and d is a distance function, the following representation can be given:

$$L = \begin{cases} d\left(r_a, r_p\right) & \text{if} \quad \text{Positive Pair} \\ \max\left(0, m - d\left(r_a, r_n\right)\right) & \text{if} \quad \text{Negative Pair} \end{cases}$$

The loss will be 0 for negative pairs that have a distance between the two elements (their representation) from a pair that is bigger than the margin m. When that distance is not larger than a margin, the loss will be positive, and weights, so-called NN parameters, will be updated. That means the NN trains to create more distant representation for the two elements. For positive pairs, in the case when there is no distance between elements representations the loss will be zero.

If we bring y variable(boolean flag equal to 0 for a negative pair and to 1 for a positive pair) to the equation and the distance d is the euclidian distance, we write the formula like:

$$L\left(r_0, r_1, y\right) = y\left\|r_0 - r_1\right\| + (1 - y)\max\left(0, m - \left\|r_0 - r_1\right\|\right)$$

Losses that are adjusted to use pairs in the training procedure called Triplet Ranking Loss.

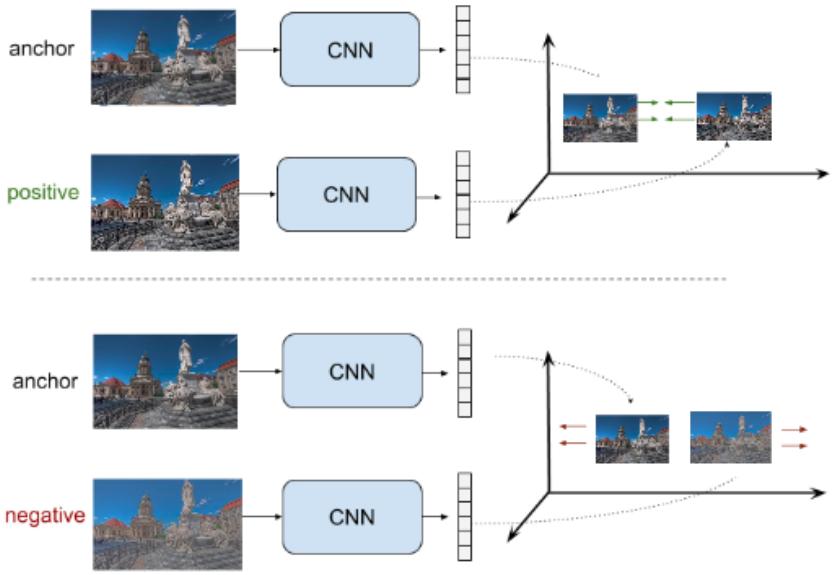


FIGURE 2.3: Pairwise ranking losses

Usually, Triplet ranking loss performs better than Pairwise loss on the same data samples. The triplets are consist of an anchor sample x_a, a positive sample x_p and a negative sample x_n.

The aim is for the distance between the anchor sample and the negative representations $d(r_a, r_n)$ to be larger (and more significant than the margin m) than the distance between the anchor and the positive representations to be $d(r_a, r_p)$. Following the notation,we can write:

$$L\left(r_a, r_p, r_n\right) = \max\left(0, m + d\left(r_a, r_p\right) - d\left(r_a, r_n\right)\right)$$

There three possible cases:

- Easy Triplets: $d(r_a, r_n) > d(r_a, r_p) + m$

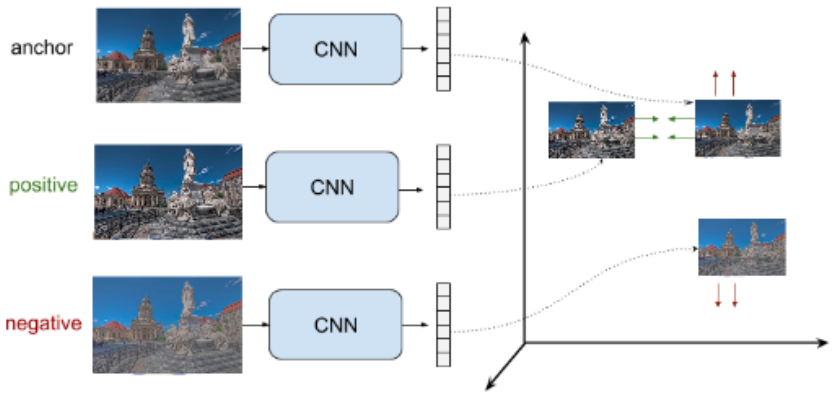


FIGURE 2.4: Triplet ranking losses

The distance between a negative sample and the anchor is bigger than the distance between a positive and the anchor, but the distance is not greater than the margin. This means that loss is 0 and weights remain the same.

- Hard Triplets: $d(r_a, r_n) < d(r_a, r_p)$
  The negative sample is closer to the anchor than the positive. The loss is positive and greater than margin m.
- Semi-Hard Triplets: $d(r_a, r_p) < d(r_a, r_n) < d(r_a, r_p) + m$
  The distance between a negative sample and the anchor is bigger than the distance between a positive and the anchor, but the distance is smaller than the margin, so the loss is still positive and smaller than m.

2.4 Pretext tasks

Pretext tasks are predesigned tasks for neural networks. Visual features are learned by learning objective functions of pretext tasks. Pretext task is used in self-supervised learning to generate useful feature representations that can be learned to give the desired results.

Pretext tasks are normally tasks that help to solve the main task. Pretext tasks can help with an understanding of input data. For instance, the task of identifying the image degradation that was applied to an image can be a pretext task while your main task is image enhancement or aesthetic assessment, etc.

2.5 Interpretability algorithms

Neural networks are often viewed as a black box, which means that the result given by it will not give you an understanding of the structure of the function. It makes the analysis harder and also causes 'trust issues' for both developers and casual users. Gaining more insight into how these models work would be hugely beneficial. Interpretability algorithms are the algorithms that help to understand a neural network and the reason for a network's prediction better, to debug, to explore the logic of a network. Interpretability algorithms is an approach to the problem of attributing the prediction of a deep network to its input features. Attribution methods are capable of indicating which words from the text were critical to assign a piece of text to a specific label or which pixels played a significant role in image classification.

We would like to highlight such algorithms as Integrated Gradients, DeepLift, and Occlusion.

2.5.1 Integrated Gradients

Integrated Gradient as the attribution method was proposed in 2017 [20]. Such methods are quite actively used to assess the quality of Image Classification[7, 1]. Integrated Gradients compute the partial derivatives of the output concerning each input entity and calculate the average gradient. The mathematical formulation can be expressed as follows:

$$IntegratedGrads_i(x) ::= \left(x_i - x_i'\right) \times \int_{\alpha=0}^{1} \frac{\partial F\left(x' + \alpha \times (x - x')\right)}{\partial x_i} d\alpha$$

Integrated Gradients along the i - th dimension of input X. Alpha is the scaling coefficient. The equations are copied from the original paper. Integrated Gradients have

a notable characteristic: the attributions sum up to the target output minus the target output evaluated at the baseline. The baseline is often chosen to be zero.

2.5.2 DeepLift

DeepLIFT [19] like Integrated gradient, is a back-propagation based approach. The objective of DeepLIFT is to highlight specific neurons that cause a difference between the inputs and the corresponding images or baselines. DeepLIFT is computed with a backward pass on the NN. Mathematical formulation of the algorithm (from original paper):

$$m_{\Delta x \Delta t} = \frac{C_{\Delta x \Delta t}}{\Delta x}$$

x is the input neuron with a difference from reference Δx, and t is the target neuron with a difference from reference Δt. C is then the contribution of Δx to Δt.

DeepLIFT is highly correlated with Integrated Gradients, but some experiments showed that DeepLIFT is faster and can even cover those cases where Integrated Gradients may mislead.

2.5.3 Occlusion

Occlusion [21] is a perturbation based attribution approach. To compute attributions, it removes or replaces each adjacent rectangular region with a given baseline, and measures the difference with the original output.

Chapter 3

Related Works

In this chapter, briefly described other works that are touching upon the same topics. This way, we will gain a better understanding of how this paper contributes to the field.

3.1 Will People Like Your Image?

The authors approach the problem using similarity learning on data from Flikr's (online platform)[4]. Flikr has a feature called the favorite list. Each user can create their collection of images they want to remember by adding an image to the favorite list. The aesthetic score described in the paper is based on time-independent data ('faves' and 'views') for each photo. The authors collect a massive dataset (AROD) of 380k images with the meta-information. To judge the pleasingness of an image, they examine the correlation between the "views" (number of visits) and the "faves" (number of clicks that favor image) as a criterion for visual aesthetics. Both these landmarks are highly dependent on visual aesthetics and encode the visual quality in all its facets.

The authors assume that number of "faves" and "views" followed exponential growth. Based on that it is possible to express $F(i)$ as r_F^t and $V(i)$ as r_V^t for any arbitrary image $i \in \mathcal{I}$ where exponential growth rate $r_{(\cdot)} > 0$ and time $t \in \mathbb{N}$. This allows to approximate the score $S(i)$ of the image time $t-$ -independentely by

$$S(i) \sim \frac{\log F(i)}{\log V(i)}$$

Considering the score $S(i)$ gives a criteria to rank images $i \in \mathcal{I}$, which can be learned by neural networks.

The approach they came up with is to optimize relative distances

$$\delta : \mathcal{I} \times \mathcal{I} \to \mathbb{R}, \quad (i, j) \mapsto \left\| \Phi_i - \Phi_j \right\|_2$$

between encodings Φ_i, Φ_j from pairs of images (i, j). Then instead of the training of a convolution neural network on human-labeled visual aesthetic scores, they train ANN on these encodings using only information about aesthetic similarity. To optimize the desired metric, they adapt the technique called triplet loss, and directional triplet loss:

$$L(a, p, n) = L_e(a, p, n) + L_d(a, n)$$

$$L_e(a, p, n) = \left[m + \left\| \Phi_a - \Phi_p \right\|_2^2 - \left\| \Phi_a - \Phi_n \right\|_2^2 \right]_+$$

for images a, p, n and some margin m. Here, $[x]_+$ is the non-negative part of x.

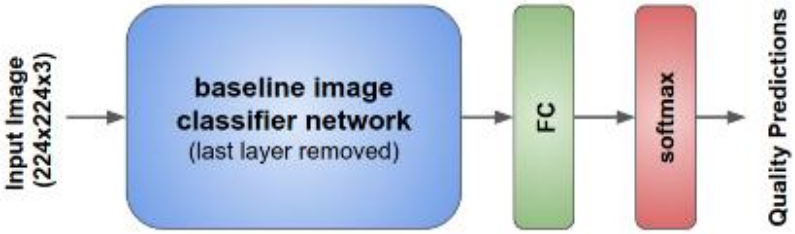


FIGURE 3.1: Modified baseline image classifier network used in NIMA framework.
Source: NIMA: Neural IMage Assessment paper[3]

They add a directional term to the loss function $L_e(a, p, n)$.

$$L_d(a, n) = sign(s(n) - s(a)) \cdot [\|\Phi_a\| - \|\Phi_n\| + \tilde{m}]_+$$

That leads to increasing the norms of more pleasing images and redusing the norms of less attractive ones.

To decide whether two images are similarly aesthetic or not the score $S(i)$ used.

$$D = \left\{ (a, p, n) \text{ with } \alpha < \frac{|S(a) - S(p)|}{|S(\bullet) - S(n)|} < \beta \right\},$$

with $a, p \in \bullet$ and $\alpha, \beta \in \mathbb{R}$.

3.2 NIMA: Neural IMage Assessment

NIMA is a Neural IMage Assessment approach presented at 2018. The uniquness of an approach is that authors predict the distribution of human opinion scores(aesthetics scores) using a convolutional neural network. Because authors aim to predict with higher correlation with human assessments, rather than classifying the images to low/high score, the distribution of scores is predicted as a histogram.

The architecture of the NIMA aesthetic and quality predictor based on an image classifier (shown in 3.1). In experiments they used such image classifier architectures as VGG-18, Inception-2, MobileNet. Last layer of classifier network is replaced by a fullyconnected layer to output 10 classes of quality scores. Baseline network are trained on ImageNet dataset.

The aim of this approach can be simplified to this formulation: to find the probability mass function $\mathbf{p}$ that is an accurate estimate of $\hat{\mathbf{p}}$.

In this approach instead of cross-entropy loss Earth Mover's Distanc(EMD) based loss[8] was used. These loss functions penalize misclassifications according to class distances.

In the case of aesthetic and quality estimation we can write the ordering as follows: $s_1 < \cdots < s_N$. EMD is defined as the minimum cost to move the mass of one distribution to another. Given the ground truth and estimated probability mass functions $\mathbf{p}$ and $\hat{\mathbf{p}}$, with N ordered classes of $r-$ norm distance $\left\|s_i - s_j\right\|_r$, where $1 \leq i, j \leq N$, the normalized Earth Mover's Distance can be written as:

$$EMD(\mathbf{p}, \hat{\mathbf{p}}) = \left(\frac{1}{N} \sum_{k=1}^{N} \left| \text{CDF}_{\mathbf{p}}(k) - CDF_{\hat{\mathbf{p}}}(k) \right|^r \right)^{1/r},$$

where $CDF_{\mathbf{p}}(k)$ is the cumulative distribution function as $\sum_{i=1}^{k} \mathbf{p}_{s_i}$. Ground truth distribution of human opinion scores can be represented as an empirical probability mass function $\mathbf{p} = [p_{s_1}, \ldots, p_{s_N}]$ with $s_1 \leq s_i \leq s_N$ where s_i is the i th score bucket, and N denotes the total number of score buckets.

To train the end-to-end aesthetic and quality assessment model the researches used datasets such as AVA [15], TID2013 [18] and LIVE [6].

Chapter 4

Datasets

This chapter provides information about datasets used to train aesthetic assessment and datasets that we gather for our hypothesis verification. All of these datasets can be accessed freely.

4.1 AROD

The source of the dataset is a Flikr. Flikr is an image hosting service that was created for people to share their images. The Flikr has the ability to 'save' images you like to your favorite list. Each photo that is published at Flikr has information about how many users viewed the photo and how many users added the photo to their favorite list.

The dataset is publicly available as a list of links to images and correspondent 'faves' and 'views' to it. Originally dataset contained 380k images with meta information needed for a training procedure.

4.2 AVA

AVA:A Large-Scale Database for Aesthetic Visual Analysis is a popular dataset for image aesthetic assessment. The AVA dataset includes about 255,000 images, rated based on aesthetic qualities by amateur photographers along with semantic labels for over 60 categories as well as labels related to photographic style and aesthetic score. The image ratings distributed from 1 to 10, when 1 is the lowest aesthetic score for a sample image. The mean scores are mostly around 5.5.

4.3 TID2013

Tampere Image Database 2013 (TID2013) contains 3000 images, from 25 original (anchor) images, 24 types of distortions with five levels for each. Before-mentioned distortions can be united into compression artifacts, noise, blur and color artifacts groups. Anchor images are obtained by cropping from Kodak Lossless True Color Image Suite [12].

The data collection procedure was performed as follows: participants in the experiment are shown random pairs of distorted images and asked to rate which one is better. The best image gets 1 point; otherwise, it gets 0 points. Each image hits these pairs nine times. In the end, the scores are summed; these amounts are the result and are used as a quality assessment. Accordingly, the rating for each image may vary from 0 to 9.

4.4 LIVE

The database contains 1,162 images that have over 350,000 opinion scores overall. All the data was collected using the usual mobile devices, because of that images have complex distortions (and mixtures of distortions), which are not the synthetic modeled distortions. That gives the LIVE database a big privilege. For each image from LIVE scores distributed from 1 to 10, when 10 is the highest aesthetic score.

4.5 GoPro

It is a common benchmark for image motion blurring. The dataset was proposed in Deep Multi-scale Con-volutional Neural Network for Dynamic Scene Deblurring original paper [16] in 2016. The dataset consists of 3 214 blurry images along with 3 214 ground truth (clear) images corresponding to them. Data was obtained by filming 240 frames per second (fps) video sequences on the GoPro Hero 4 camera. Blurred images was generated through averaging consecutive short-exposure frames.

4.6 NH-Haze

NH-Haze [2] is the first dehazing dataset that contains nonhomogeneous haze scenes. The dataset consists of 110 photos representing 55 scenes; each scene is depicted in two pictures, one with the presence of haze, the other without. All scenes are outdoor. Dataset was used in the NTIRE 2020 NonHomogeneous Dehazing image challenge [17].

4.7 Demoireing Single Image

This is a private dataset that was used in the NTIRE 2020 Demoireing - Track 1 Single image image challenge [17]. The dataset containing 11 000 moire/clear image pairs. The dataset has the following structure: 11000 image pairs divided into: 10000 for training, 500 for validation, 500 for testing Data. One input image is sequences of 7 input frames and the output image is a moire-free image corresponding to the middle input frame.

Chapter 5

Experiments

5.1 Aesthetic quality assessment and Image Restoration

The first experiment aims to evaluate if image aesthetic assessment approaches are valid and can be used as an essential component in loss function that uses in image restoration approaches.

Additionally, the experiment aims to evaluate existing methods on real-world data.

In NIMA [3], it was demonstrated that the obtained model could be used to assess image quality, so reasonable to ask whether NIMA can be used to improve the approaches for Image Restoration. For example, as an additional component in a loss function. For instance, Lee Fei Fei's work [9] shows that the corresponding difference in VGG layers can be used as a loss component, and it improves the overall results. To answer this question, we took real data and calculated the following metrics.

Knowing the performance models on real data we can assess how these approaches can be used as an additional component in loss. Firstly, we pass the real-world data through the trained neural networks. Then, we calculate the accuracy for datasets to make a hypothesis verification. To measure performance, we calculate the accuracy as follows:

$$Acc = \frac{1}{N}\sum_{i=1}^{n} G(S(I_s), S(I_b)) \tag{5.1}$$

$$G(s,b) = \begin{cases} 1 \text{ if } & s \geq b \\ 0 \text{ if } & s < b \end{cases} \tag{5.2}$$

where $S(I)$ is the score for the image I, N is a number of image pairs. So we can say that accuracy is the number of image pairs that Neural Network predicted right (score of the sharp image is higher than the score of blurred image) divided by the total number of image pairs. The data used in experiments is described in the Datasets chapter.

NH-Haze, GoPro and Demoireiring datasets are collected for the task of restoration of images, namely dehazing, deblurring and demoireing.

We set up all the experiments on NIMA and "Will People Like Your Image?" implementations.

5.1.1 Deblurring

Motion blur is a blurring of an image due to the movement of the subject or camera. This leads to a decrease in image sharpness. Motion blurred images can be described

FIGURE 5.1: Results on a NH-Haze dataset. Where score 1 is NIMA score, score2 is "Will People Like Your Image?" score

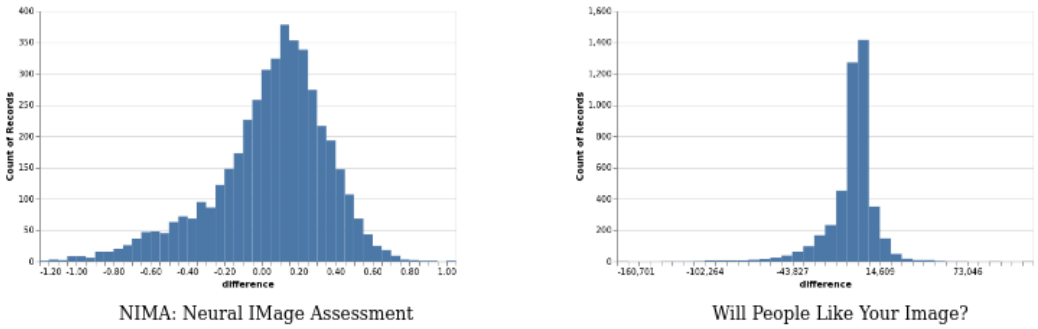


FIGURE 5.2: Distribution of score differences for the GoPro dataset

as following:

$$I_B = k(M) * I_S + N$$

where I_B is a blurred image, $k(M)$ are unknown blur kernels determined by motion field M.I_S is the sharp latent image, $*$ denotes the convolution, N is an additive noise. The formula is copied from DeblurGAN [13] paper.

The examples of algorithms results are shown in the Figure. 5.1

The accuracy we obtained on the motion deblurring dataset GoPro for NIMA is 0.637%; the accuracy obtained for "Will People Like Your Image?" approach is 0.454%. (due to 5.1, 5.2)

The distribution of differences between sharp images scores and blurred images scores shown in the Figure 5.2.

5.1.2 Dehazing

Haze is a term that can be used to describe an atmospheric phenomenon that can influences the clarity of vision such things as dust, smoke, etc. Images captured under haze can be classified as images of poor quality because of color shifting, contrast, and clarity of an image. This representation is commonly used to explain hazed image:

$$I(x) = J(x)t(x) + A(1 - t(x)),$$

FIGURE 5.3: Results on a NH-Haze dataset. Where score 1 is NIMA score, score2 is "Will People Like Your Image?" score

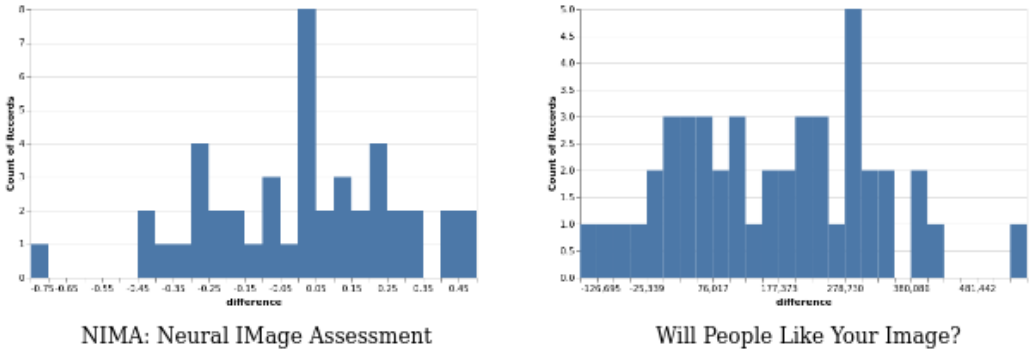


FIGURE 5.4: Distribution of score differences for the NH-Haze dataset

where $I(x)$ is the observed hazy image, and $J(x)$ is the clean image without haze. The parameter A denotes the global atmospheric light, and $t(x)$ is the transmission matrix defined as:

$$t(x) = e^{-\beta d(x)},$$

where β represents the scattering coefficient, and $d(x)$ is the distance between the object and camera. Image dehazing aims to remove haze from an image and make it clear.

The example of how both approaches work on dehazing data shown in Figure5.3

The distribution of differences between original images scores and images with haze scores shown in the Figure 5.4.

The accuracy we get passing dehazing data through NNs is 0.86% for "Will People Like Your Image?" and 0.6% for NIMA (5.1, 5.2).

5.1.3 Demoireing

Moire patterns are often an artifact of images taken by various digital cameras or created by graphic design techniques. Moire degrades the quality and resolution of

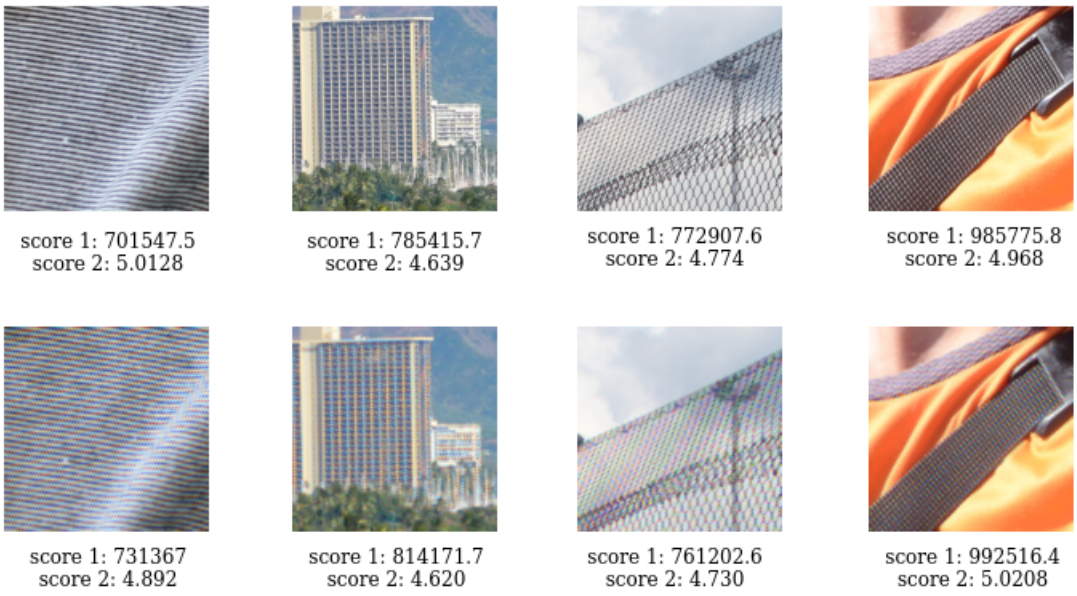
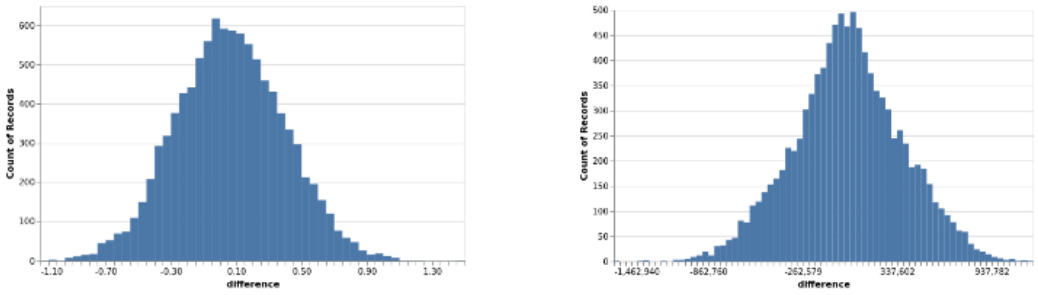


FIGURE 5.5: Results on a moire dataset. Where score 1 is NIMA score, score2 is "Will People Like Your Image?" score



NIMA: Neural IMage Assessment

Will People Like Your Image?

FIGURE 5.6: Distribution of score differences for the Demoireing Single Image dataset

graphic images. In some rare cases moire can be aesthetically pleasing. Demoireing is an image restoration task the objective of which is to eliminate moire effect and to keep an image undamaged.

The distribution of differences between original images scores and images with haze scores shown in the Figure 5.6.

The example of neural nets performances shown in Figure 5.5. The accuracy calculated for moire data using "Will People Like Your Image?" and NIMA is 0.53%and 0.57%, respectively(5.1, 5.2).

5.1.4 Results

The quantitative results make the point that both of the proposed trained approaches failed with real datasets. One can argue that one of them showed a comparatively good result with a high metric score, but it fair to take into account the sizes of datasets. That database has only 55 image pairs. That is why it is not entirely clear how confident we can be with this positive result. Based on that, we can assume that

it is hard for these methods to capture image distortions and degradations. That concludes to the point that it would be misleading to include aesthetic score as an additional score to the image degradation removal. It would not give any advantages or even worse make the training process unstable.

5.2 Exploration of neural network using interpretability algorithms

We had the intent to understand what is essential for the NIMA approach in the image, making the decision based on it. NIMA was chosen because it not only aesthetic assessment approach; it considers the quality of images as well. It clear from the previous section that model do not work well with real-world data. So we decided to investigate what its focus is. For the accomplishment of this task, we decided to use model interpretability algorithms. We chose to use attribution algorithms that use gradients of different layers to highlight the zones of the image that highly correlate with the resultant value, aesthetic score in our particular task. Also, we use the Occlusion algorithm, which is a perturbation-based approach to the attribution problem.

5.2.1 Results

You can see the results of these techniques on several images from haze and moire datasets in Figure 5.7. From Occlusion method is very clear that the neural network pays the most attention to the corners of the images. It is most likely that it was over-trained(overfit) to the features that are most likely to be spotted at those locations in the training dataset's average image and gave good results at those datasets. Contrary, for the moire effect image pairs NIMA approach does not capture any difference between two images. It can be explained that moire is not noticeable enough, and the neural network just does not catch it. On the other side, it may occur for the approach that the moire images can also be valid and do not have any defects or bad aesthetics. On the other hand, the same neural network does not find any differences between motion-blurred images and clear images. We consider motion blur as distortion that leads to severe quality loss (we described it more in detail in the next section). At this point, we can say that NIMA is not able to objectively identify the quality of an image. Probably NIMA is studying some compression artifacts instead of quality and can't be used on real image data.

5.3 Correlation between quality and aesthetic

Even we used to think that image with distortions and visual corruptions is, by default, considered as an image of poor quality and low aesthetic quality, this opinion may be wrong. Effects such as bokeh (the out-of-focus parts of an image) blur, Gaussian smoothing, smoke, fog can make the photo aesthetic. Moreover, now there is a trend in social networks for blurred photos (motion blur) and corrupt photos (with effects).

In order to study the correlation of image distortions and aesthetics, we created a survey to evaluate our hypothesis. We showed 35 pairs of images and asked participants to choose which of two images have better quality and which of two images is more pleasing to observe.

FIGURE 5.7: Eight examples where we applied different attribution methods. (a) is the original image, (b) is a visualization of attributes extracted using Occlusion method, (c) all attributes extracted using DeepLIFT overlayed on the image, (d) positive attributes extracted using DeepLIFT, (e) all attributes extracted using Integrated Gradients overlayed on the image, (f) positive attributes extracted using DeepLIFT overlayed on the image.

FIGURE 5.8: Hazed images with high scores

As image pairs, we used distorted/clear image pairs that were predicted incorrectly in the previous experiment.

We interviewed 30 people who agreed to participate in the experiment. The investigation participants are people of different genders and age groups (aged 7 to 50 years). Ten of them are professional or amateur photographers.

All photos that were distorted by a motion blur were rated as photos of lower quality and less attractive. About 20% of respondents consider a photo with moire more attractive, but only 8% said that a photo with moire is better. The reason respondents chose a moire image over a moire-less is that moire added color gradients and lines to the texture. As we mentioned earlier, haze can make the scene more attractive. The results showed that over 33% of respondents find images with haze more attractive. They justified their choice by the fact that smoke or fog adds a photo of mystery and creates a more interesting composition for the viewer. Another interesting fact that respondents do not always associate clarity of image with high quality. Images, where a scene was only partially hazed, get a higher quality score in 50%-60% cases. Images with over 50 percent of haze or have haze layered over photo rated as less attractive in all cases. The example of edge cases (when over 23% of respondents rated image labeled as distorted better than clear) shown in Figure 5.8

It noteworthy that respondents rate most of images (75%) as less attractive because of the image quality in the first place.

5.3.1 Results

The obtained data shows that there is a significant positive correlation between the aesthetic level of the image and its quality in some cases of degradation. That means our hypothesis about the relationship is between the aesthetic of the image, and its quality was right. The other exciting conclusion is that some of the degradations were found as more aesthetically pleasing than the photos without them. The example of it is a little haze in the image that doesn't cover the center depicted object.

Chapter 6

Conclusions

6.1 Results summary

In this work, we consider aesthetic assessment methods and the weaknesses that methods have. We researched the existing datasets for this problem and whether they assume real data.

In the course of our research on image evaluation methods, we identified the following hypothesis: Aesthetic metrics can be an auxiliary component in the systems' training procedure, aiming to restore the corrupt image.

Based on the results of two experiments, we can state that the hypothesis of using aesthetic image assessment methods as an additional loss function or an additional component in the loss function for training neural networks on image restoration tasks is wrong. It would instead make training procedures more complex and can be a factor that fools the neural network.

During our experiments, we investigate that systems that are declared one of the highest scores is not ready to face the natural data. They are fitted to the data that was collected specifically for the aesthetic assessment.

In order to complete our second experiment (with the activation maps), we proved that NIMA, which declared that its methods are capable of evaluating whether an image has good image quality and aesthetic quality, is not capable of doing it. NIMA also showed an accuracy of 63% on the GoPro dataset, 60% on the NH-Haze dataset, and 57% on the Demoireing SingleTrain dataset.

Also, examining the validity of approaches to image evaluation on data collected for image restoration, we encountered that one of the datasets, namely NH-Haze, has photos with a smoke present on them that look more attractive. This observation led us to assume that there is a correlation between image quality and image aesthetics, but, we cannot be sure whether it is always correct to say that poor quality is equivalent to the less attractiveness. Because of that, so we tested this concept on real respondents and realized that severe degradation is the cause of negative aesthetic evaluation. However, haze, fog, and smoke can simply not effect quality from the observer's point of view, and what is appealing can be perceived more positively by a person.

To summarize, applying conventional techniques often leads to fitting to a specific dataset and ends up inapplicable to a wide variety of problems. Therefore, it is the best practice to evaluate such methods using Inpretability Algorithms as CAM (Class Activation Mapping) and evaluation on real data, especially when it comes to quality assessment, as in NIMA.

6.2 Future work

Possible ideas for future work:

- To collect a diverse dataset that will contain data with a variety of distortion and corruption, such as radial distortion, raindrops, Watermarking, etc. Along with the images to collect several scores for different parameters, E.g., total score, quality score, aesthetic score, etc.
- To design a neural network architecture that can be used as an aesthetic metrics and will pay attention to image quality.

Bibliography

[1] Pengguang Chen. *GridMask Data Augmentation*. Jan. 2020.

[2] Ancuti Codruta, Cosmin Ancuti, and Radu Timofte. "NH-HAZE: An Image Dehazing Benchmark with Non-Homogeneous Hazy and Haze-Free Images". In: (2020).

[3] Hossein Talebi Esfandarani and Peyman Milanfar. "NIMA: Neural Image Assessment". In: *CoRR* abs/1709.05424 (2017). arXiv: 1709.05424. URL: http://arxiv.org/abs/1709.05424.

[4] *Flickr is an image hosting service and video hosting service.* URL: https://www.flickr.com/.

[5] D. Ghadiyaram and A. C. Bovik. "Massive Online Crowdsourced Study of Subjective and Objective Picture Quality". In: *IEEE Transactions on Image Processing* 25.1 (2016), pp. 372–387.

[6] D. Ghadiyaram and A.C. Bovik. *LIVE In the Wild Image Quality Challenge Database*. 2015. URL: http://live.ece.utexas.edu/research/ChallengeDB/index.html.

[7] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V. Le. "DropBlock: A regularization method for convolutional networks". In: *CoRR* abs/1810.12890 (2018). arXiv: 1810.12890. URL: http://arxiv.org/abs/1810.12890.

[8] Le Hou, Chen-Ping Yu, and Dimitris Samaras. "Squared Earth Mover's Distance-based Loss for Training Deep Neural Networks". In: *CoRR* abs/1611.05916 (2016). arXiv: 1611.05916. URL: http://arxiv.org/abs/1611.05916.

[9] Justin Johnson, Alexandre Alahi, and Fei-Fei Li. "Perceptual Losses for Real-Time Style Transfer and Super-Resolution". In: *CoRR* abs/1603.08155 (2016). arXiv: 1603.08155. URL: http://arxiv.org/abs/1603.08155.

[10] Yueying Kao, Ran He, and Kaiqi Huang. "Visual Aesthetic Quality Assessment with Multi-task Deep Learning". In: *CoRR* abs/1604.04970 (2016). arXiv: 1604.04970. URL: http://arxiv.org/abs/1604.04970.

[11] Menglei Chai Guohui Wang Peng Zhou Feiyue Huang Bao-Gang Hu Rongrong Ji Chongyang Ma Kekai Sheng Weiming Dong. "Revisiting Image Aesthetic Assessment via Self-Supervised Feature Learning". In: *Proceedings of AAAI Conference on Articial Intelligence 2020* (2019). URL: https://arxiv.org/abs/1911.11419.

[12] *Kodak lossless true color image suite,* URL: http://r0k.us/graphics/kodak/.

[13] Orest Kupyn et al. "DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks". In: *CoRR* abs/1711.07064 (2017). arXiv: 1711.07064. URL: http://arxiv.org/abs/1711.07064.

[14] X. Lu et al. "Rating Image Aesthetics Using Deep Learning". In: *IEEE Transactions on Multimedia* 17.11 (2015), pp. 2021–2034.

[15] N. Murray, L. Marchesotti, and F. Perronnin. "AVA: A large-scale database for aesthetic visual analysis". In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2012, pp. 2408–2415.

[16] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. "Deep Multi-scale Convolutional Neural Network for Dynamic Scene Deblurring". In: *CoRR* abs/1612.02177 (2016). arXiv: 1612.02177. URL: http://arxiv.org/abs/1612.02177.

[17] *NTIRE 2020 image challenges*, URL: https://data.vision.ee.ethz.ch/cvl/ntire20/.

[18] Nikolay Ponomarenko et al. "Image database TID2013: Peculiarities, results and perspectives". English. In: *Signal Processing: Image Communication* 30 (Jan. 2015), pp. 57–77. ISSN: 0923-5965. DOI: 10.1016/j.image.2014.10.009.

[19] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. "Learning Important Features Through Propagating Activation Differences". In: *CoRR* abs/1704.02685 (2017). arXiv: 1704.02685. URL: http://arxiv.org/abs/1704.02685.

[20] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic Attribution for Deep Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. International Convention Centre, Sydney, Australia: PMLR, 2017, pp. 3319–3328. URL: http://proceedings.mlr.press/v70/sundararajan17a.html.

[21] Matthew D. Zeiler and Rob Fergus. "Visualizing and Understanding Convolutional Networks". In: *CoRR* abs/1311.2901 (2013). arXiv: 1311.2901. URL: http://arxiv.org/abs/1311.2901.