Transfer learning of ImageNet Object Classification Challenge features to image aesthetic binary classification

Bruno Tinen, Jun Okamoto Junior

Department of Mechatronics and Mechanical Systems Engineering
Escola Politécnica of the University of São Paulo – São Paulo – Brazil

bruno.tinen@usp.br, jokamoto@usp.br

Abstract. *The aesthetic classification of photographies is a problem of separating aesthetically pleasing images from not pleasing images using algorithms that describe and evaluate both emotional and technical factors. Since the mass adoption of deep convolutional neural network (DCNN) models for image classification problems, different DCNN architectures have been developed due to its overall better performance, pushing the boundaries of the state-of-the-art performance of the image classification further. This paper evaluates how architectures and features that were primarily developed for the ImageNet Object Classification Challenge perform when analyzed under the aesthetic scope. A high-level transfer learning model composed of a DCNN layer and a top layer that behaves as a linear SVM is proposed and seven different DCNN architectures are trained using it. Scenarios with just transfer learning and with fine-tuning are evaluated and a model using the ResNet-Inception V2 architecture is proposed, which achieves results better than current state-of-the-art for the experimental conditions used.*

1. Introduction

In the visual perception of form, equilibrium, harmony, and clarity compose what a human being understands as aesthetic and are factors considered essential to image formation, regardless of it being a photograph, a painting or a sculpture [Filho 2009]. In photography, the quality and beauty of an image can be described using a series of factors, not only emotional ones – as the historical moment of a photography – but also technical ones – as the illumination and composition.

The aesthetic classification of photographies is a problem that can be formulated as an intrinsic binary classification of images as aesthetically pleasing or not, being subjectivity and ambiguity the greatest challenges [Marchesotti et al. 2011]. Using machine learning algorithms it is possible to evaluate the aesthetic quality of an image by the extraction and classification of descriptors that represent both emotional and technical factors. When those descriptors are extracted using learning algorithms they not always have an equivalent that would be produced by a human being or have a logical representation, but are the ones that perform the best.

The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is a reference competition of object detection and image classification that has been used to measure the progress and compare advances of new algorithms in the machine learning area. Since

ILSVRC 2012, year in which AlexNet performed considerably better than the other competing models [Krizhevsky et al. 2012], deep convolutional neural networks have gained increasing attention in image-related problems. Several deep learning architectures were conceived, such as Inception [Szegedy et al.] and VGG [Simonyan and Zisserman 2014], each one improving over previous results on the ILSVRC.

Deep learning models were also applied to the aesthetic classification problem, with approaches varying from training known network architectures from random weights [Talebi and Milanfar 2018] to fine-tuning models previously trained on the ILSVRC dataset [Jin et al. 2016a]. Deep learning models have outperformed models based on handcrafted features and models based on other forms of generic image descriptors [Deng et al. 2017].

The classifiers trained used the Aesthetic Visual Analysis (AVA) [Murray et al. 2012a] dataset, which is composed of more than 255,000 annotated examples with aesthetic scores. The AVA dataset is widely used in aesthetic classification related works and the comparisons made in this paper focus on models trained using it.

This paper proposes a high-level machine learning architecture composed by a known deep convolutional neural network (DCNN) architecture and a top layer that behaves as a linear Support Vector Machine (SVM) activation layer, replacing the DCNN top softmax activated layer. The impact of different DCNN architectures is evaluated on the aesthetic classification problem using this setup, both in transfer learning and fine-tuning scenarios. The results presented give insights on:

- the transferability of ILSVRC Object Classification Challenge features for aesthetic classification with the AVA dataset;
- how deep learning architecture impacts on aesthetic classification model's performance;
- how comparable are the results obtained for the ILSVRC with those obtained on aesthetic classification on the AVA dataset.

It is also proposed a transfer learning model with fine-tuning that performs better than current state-of-the-art models using the ResNet-Inception V2 DCNN architecture.

2. Aesthetic Descriptors and Learning Methods

Several approaches have been developed to solve the aesthetic classification problem. In order to develop a working solution, the selection and development of aesthetic descriptors and learning methods need to be addressed. On both topics, several approaches were developed, with descriptors ranging from concrete to generic ones and learning methods going from SVMs to DCNNs.

2.1. Aesthetic descriptors

Aesthetic descriptors can be defined as features that can be extracted from photographies which can model in an approximate way a criteria used to classify images regarding its aesthetics [Marchesotti et al. 2011]. These descriptors can be related to approximation of photography principles [Aydin et al. 2015], colors and composition [Suran and K. 2015], for example. They may also have no relation to a human understandable characteristic of an image.

It is possible to classify these descriptors in two groups: concrete descriptors and generic descriptors. These two groups differ in the way they are obtained and in the scope they represent.

Concrete aesthetic descriptors are statistical models of criteria that human beings use to judge photographies. Concrete descriptors are mathematical formulations of specialists rules, that approximate human aesthetic comprehension. These descriptors can be further divided into low level [Datta et al. 2006a, Khan and David 2012], high level [Bhattacharya et al. 2010, Ke et al. 2006, Luo and Tang 2008, Dhar et al. 2011, Khan and David 2012, Lo et al. 2012, Mavridaki and Mezaris 2015] and semantic [Bhattacharya et al. 2010, Dhar et al. 2011, Murray et al. 2012b]. The first two categories are typically used together [Marchesotti et al. 2011]. It is also possible to extract structural characteristics, using graphlets [Zhang et al. 2014] or salient regions [Wong and Low 2009]. Semantic descriptors describe the contents of an image in a categorized way [Bhattacharya et al. 2010, Dhar et al. 2011, Murray et al. 2012b]. They can be used in a single photography or between photographies. With them, it is also possible to use low and high-level features to find relations between semantic categories.

Generic aesthetic descriptors implicitly model photography characteristics and can be obtained by machine learning techniques. These descriptors do not necessarily have a direct relationship with photography rules or visual factors, like the ones modeled by concrete features. From a photography set previously classified as aesthetically pleasing and not pleasing and using techniques like Convolutional Neural Networks, it is possible to find the main elements that have been used to classify the images in these categories. It is also possible to obtain generic features by using transfer learning on DCNNs that have been previously trained to classify objects, like Inception, VGG and MobileNet [Talebi and Milanfar 2018]. These transfer learned descriptors can also be used with additional convolutional layers [Jin et al. 2016b].

Differently from the concrete descriptors, there is no limitation of the description of these rules by specialized knowledge and there is the possibility to model implicit rules. Generic descriptors are capable of learning efficiently what concrete descriptors try to do explicitly [Marchesotti et al. 2011]. Generic features have been shown to be better photography descriptors, resulting in higher accuracy in classification [Talebi and Milanfar 2018]. Generic and concrete descriptors can also be used together [Marchesotti and Perronnin 2013, Marchesotti et al. 2015].

2.2. Learning methods

The use of machine learning models for aesthetic classification of photographies enables the automatic classification given a pre-built training dataset, regarding both the classification per se and the generation of generic aesthetic features.

There are two main problems to be solved when dealing with aesthetic classification. The first one is the binary classification, that has the objective of discriminating between aesthetically pleasing and not pleasing images. Particularly for binary classification other models that can be highlighted are Naive Bayes [Ke et al. 2006, Luo and Tang 2008] and Adaboost [Luo and Tang 2008, Khan and David 2012].

The second one is a multiclass problem, when a numeric value, similar to a mean grade or distribution, is to be predicted for a given photography. Support Vector Regres-

sion [Bhattacharya et al. 2010, Jiang et al. 2010, Li et al. 2010] and Bayesian Networks [Gao et al. 2015] were used when dealing with classification for this kind of problem.

Among the used models, it is possible to highlight SVMs on the classification task – both in binary [Datta et al. 2006a, Luo and Tang 2008, Dhar et al. 2011, Khan and David 2012, Lo et al. 2012, Mavridaki and Mezaris 2015, Nishiyama et al. 2011, Luo et al. 2013, Wong and Low 2009] and multiclass [Bhattacharya et al. 2010, Jiang et al. 2010, Li et al. 2010]. The use of SVMs is justified by its performance when compared to K-nearest neighbors, random forest and Adaboost [Khan and David 2012].

Another important model used in both binary and multiclass aesthetic classification are neural networks. A fully connected network can be used with softmax activation in the last layer for classification algorithms [Ma et al. 2017, Lu et al. 2015, Talebi and Milanfar 2018, Jin et al. 2016b].

2.2.1. Linear SVMs with DCNNs

It has been shown that for some DCNN architectures, using a L2-SVM objective to train DCNNs, replacing the conventional softmax top layer with a neural network that behaves as a linear SVM, offers superior performance on several classification tasks [Tang 2013].

The L2-SVM learning optimization problem, which minimizes the squared hinge loss, can be represented as (1).

$$\min_{w} \quad \frac{1}{2} w^T w + C \sum \max(1 - w^T x_n t_n, 0)^2 \tag{1}$$

Replacing the input $x$by the penultimate activation $h$and differentiating the objective function $l(w)$ for the L2-SVM, (2) is obtained.

$$\frac{\partial l(w)}{\partial h_n} = -2C t_n w(\max(1 - w^T h_n t_n, 0)) \tag{2}$$

With this, back propagation algorithm can be used exactly the same as in softmax-based activation layers. On a two-neuron layer as the output layer, the predicted class can is given by (3), where $a_k(x)$ is the output activation of the k-th neuron on the layer.

$$\arg\max_{k} a_k(x) \tag{3}$$

3. Methodology

To make it possible to transfer features from the ILSVRC to the aesthetic classification problem a single high-level learning architecture is proposed in order to make the obtained results comparable. Both source domain and target domain of this transfer learning problem can be considered the same, as both datasets - ILSVRC Object Classification Challenge and AVA dataset - are image datasets. On the other hand, the source task is

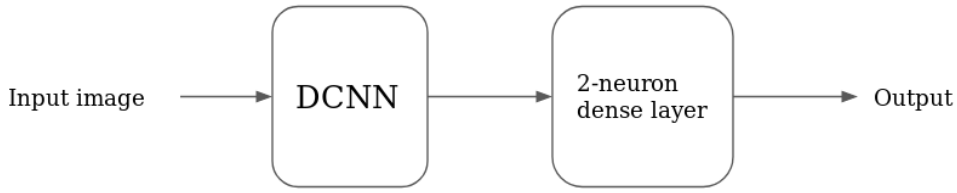

Figure 1. High-level model architecture proposed.

the object classification while the target task is the aesthetic classification. They are different but related tasks. Using the knowledge of which objects are in an image can help in the aesthetic judgment. Therefore, it is possible to classify this transfer learning problem as an inductive transfer learning [Pan and Yang 2010]. More specifically, transferring knowledge of the parameters is what is going to be explored.

The target classification problem being considered is a binary one, discriminating between aesthetically good and bad photographies. The output layer of a binary classification can be modeled as a two-neuron one-hot encoded layer, where each neuron represents one of the classes of the problem. The output class is the one represented by the neuron with the highest activation value.

The higher-level classification model is composed by a deep convolutional neural network (DCNN) architecture, based on literature existing architectures, with the top layers of the network, which would typically be a dense 1000-neuron layer for the object classification problem, replaced by a two-neuron dense layer. The high-level model architecture is shown in Figure 1. The DCNN is pre-trained with the ILSVRC Object Classification Challenge dataset and its output parameters represent the knowledge that is being transferred. These DCNN parameters can also be fine-tuned using the target dataset, but their initial values are entirely based on the ILSVRC Object Classification Challenge dataset. While the DCNN acts as a parameter extractor from the input image, the two-neuron dense layer is responsible for the classification, using the features generated by the DCNN and outputting a hot-encoded output.

The input image is always considered with a size of 224x224x3, which is the input expected in the majority of the DCNN architectures used. In architectures that expected larger input images, a padding was added accordingly. A dropout with a 0.5 ratio is used over the features output from the DCNN in order to prevent overfitting. The value 0.5 leads to the maximum amount of regularization in a linear case [Baldi and Sadowski 2013], but tuning this parameter might lead to a better performance of the models.

The loss used during training and evaluation is the quadratic hinge loss, making the two-neuron top layer of the classifier behave as a linear SVM using the approach described in subsection 2.2.1. To further approximate the top layer to a soft-margin SVM classifier, a L2 regularization is added to the two-neuron dense layer with a factor of 0.001.

The Adam optimizer was used during training. This optimizer was initialized with an initial learning rate, lr, of 0.005. The initial value for the learning rate is 5 times higher

Table 1. DCNN top-1 and top-5 accuracy over ILSVRC Object Classification Challenge validation dataset.

Model	Top-1 Accuracy	Top-5 Accuracy	Trainable Parameters
VGG16	71.3%	90.1%	138,357,544
MobileNet	70.4%	89.5%	4,253,864
ResNet50	74.9%	92.1%	25,636,712
NASNetMobile	74.4%	91.9%	5,326,716
MobileNetV2	71.3%	90.1%	3,538,984
InceptionV3	77.9%	93.7%	23,851,784
InceptionResNetV2	80.3%	95.3%	55,873,736

than the one commonly used in Adam (0,001) in order to increase the initial speed of learning. The learning rate is updated at the end of every batch training iteration using (4). A decay value of 0.003 was used at every learning rate update.

$$lr_{new} = \frac{lr_{old}}{1 + decay * iteration} \tag{4}$$

In total 7 different DCNN architectures were evaluated: VGG16 [Simonyan and Zisserman 2014], MobileNet [Howard et al. 2017], ResNet50 [He et al. 2015], NASNetMobile [Zoph et al. 2017], MobileNet V2 [Sandler et al. 2018], Inception V3 [Szegedy et al. 2015] and Inception-ResNet V2 [Szegedy et al. 2016], each with its own layer and neuron setup. The performance of such architectures on the ILSVRC Object Classification Challenge are shown in Table 1 and are used later for comparisons with the results obtained in this paper.

4. Experiments

Two sets of experiments were run using the hig- level model described in section 3 for each one of the seven DCNN architectures used. The first experiment performs transfer learning on ILSVRC Object Classification Challenge features only by training the top layer of the network. The second experiment expands the first one by fine-tuning the DCNNs layers in conjunction with the training of the top layer. Both experiments were run on the AVA dataset and results were compared with related works that have similar experiments setup.

4.1. Dataset

The AVA (Aesthetic Visual Analysis) [Murray et al. 2012a] dataset is a dataset created to supply examples for works on aesthetic photography classification, containing approximately 255,000 examples, each of them with three kinds of annotation: 1 to 10 scores given by users, semantic category and style. All the data was collected from the photo competition website `http://www.dpchallenge.com`. A comparison of the AVA dataset with other datasets, such as Photo.net [Datta et al. 2006b] and CUHK [Ke et al. 2006], shows that the first has a greater number of properly annotated examples whereas the later ones have less examples with fewer metadata.

Table 2. Results obtained from transfer learning only models.

	Accuracy	MCC
VGG16	73,92%	0,478
MobileNet	70,52%	0,437
ResNet50	49,88%	0
NASNetMobile	63,94%	0,329
MobileNet V2	71,10%	0,427
Inception V3	60,52%	0,055
Inception-ResNet V2	62,96%	0,085

From the available labels of AVA dataset, only the mean calculated from the 1 to 10 scores was considered. This mean was later used to create the binary label (aesthetically pleasing and not pleasing) based on the mean distribution of the scores on the whole dataset. Being $\bar{x}$ the mean score of all the dataset images, σ^2 the standard deviation and x_i the mean of a single example, the examples with $x_i > \bar{x} + \sigma^2$ were considered aesthetically pleasing images and the examples with $x_i < \bar{x} + \sigma^2$ were considered aesthetically not pleasing images. All the examples in between this interval were not considered.

4.2. Experimental Settings

A subset of the AVA dataset was used in all experiments. There was a total of 38,306 negative examples and 39,577 positive examples, totaling 77,883 examples. From these examples, 5,000 were separated to the validation set and 5,000 for the test set, each of them containing 2,500 examples for each class, approximately 6.5% of the total examples for each set. The reason for this was to keep the training set with as many examples as possible while having validation and test sets that could properly evaluate the model's performance. All images were redimensioned to 224x224x3, padding with black pixels when necessary to save on storage space.

Training was done in batches of 50 images and 1,360 iterations of the algorithm are considered as an epoch of training. After each epoch, a validation step was run over 34 steps each one on a step of 50 images as well. Experiments considering only transfer learning were run for 20 epochs and for the fine-tuning model experiments 2 epochs for fine-tuning were done before training the algorithm for more 20 epochs. This weight initialization of the classification layer improves the initial model convergence, as the last layer is firstly initialized with weights that are specific for the aesthetic classification, without the influence of the other parameters of the network [Talebi and Milanfar 2018][Deng et al. 2017]. All tests were run on a Google Cloud Platform instance with 4 Intel Broadwell CPUs, 16GBs RAM and one Nvidia Tesla K80.

4.3. Experimental Results and Analysis

The results for both experiments were obtained using the models which had the best performance on the validation dataset.

Results in Table 2 refer to transfer learning only experiments. From the DCNN architectures chosen only ResNet50 was unable to generate an acceptable model just with transfer learning, yielding a model that classifies all pictures in only one of the classes. The top accuracy obtained from transfer learning only models was using VGG16, with

Table 3. Results obtained from fine tuning models.

	Accuracy	MCC
VGG16	81,92%	0,639
MobileNet	82,90%	0,658
ResNet50	83,50%	0,671
NASNetMobile	84,18%	0,684
MobileNet V2	84,28%	0,686
Inception V3	84,82%	0,697
Inception-ResNet V2	86,06%	0,721
[Zhang et al. 2014]	83.24%	-
[Dong et al. 2015]	83.52%	-
[Tian et al. 2015]	80.38%	-
[Wang et al. 2016]	84.88%	-
[Jin et al. 2017]	85.53%	-

results comparable with the ones obtained by early models in the aesthetic classification problem literature.

Besides accuracy, the Mathews Correlation Coefficient (MCC) is also used to analyze the behavior of the model setups. Using MCC allows for a better understanding of the ratio of false positives and false negatives on the test set. This is clear when comparing the NASNetMobile model with Inception ones. Even though the accuracy seems comparable the MCC is very different indicating that Inception based models are more unbalanced regarding false positives and false negatives distribution. This kind of behavior is an indicator of a model that has a bias to one of the classes.

This first experiment shows that it is possible to use transfer learning only to create an aesthetic classification model with the downside of not having optimal performance. It is important to choose correctly a proper DCNN architecture otherwise features might not be able to generalize to the aesthetic problem. This choice can not be done based only on the accuracy metric; metrics like MCC can be used to avoid architectures that have a heavy bias towards one of the model classes.

Results in Table 3 refers to fined-tuned models. Comparing with results from Table 2 it is clear that the obtained models are superior for all architectures.

Fine-tuning has different results in each of the DCNN architectures. ResNet50, for instance, that could not generate a usable model with just transfer learning, outperformed VGG16 after fine-tuning. This result shows that the issue was not with the architecture itself but with how the features from ILSVRC Object Classification Challenge translated to aesthetic classification on the AVA dataset.

Overall performance of the networks had also different results for each architecture. Even though VGG16 was the best one considering only transfer learning, fine-tuning the model did not improve the model as much as in other architectures, even tough performance increases where noticeable. The fine-tuned model InceptionResNetV2 showed the best performance, following as the best result that the network also had in the ILSVRC, as shown in Table 1.

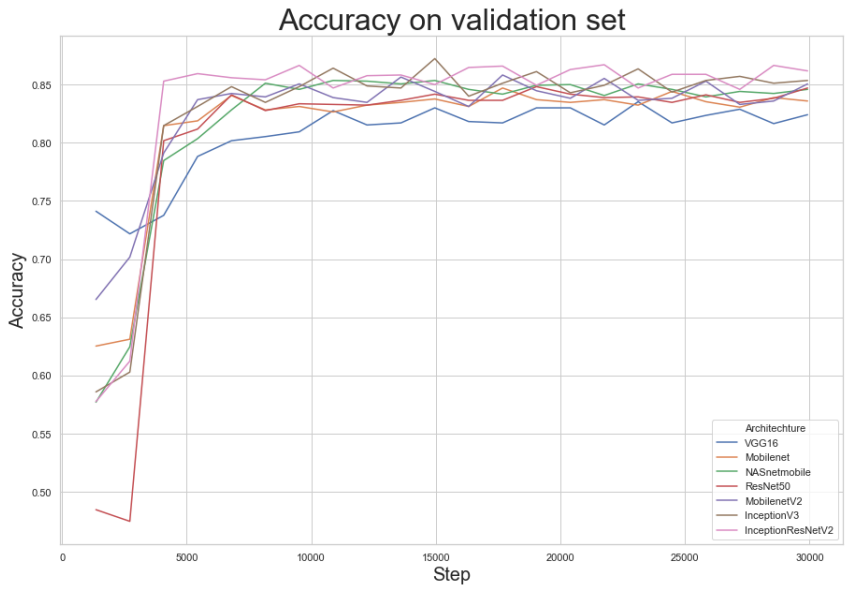

Figure 2. Accuracy over validation set during training.

Figure 2 shows the accuracy computed on the validation set after each epoch of training during the fine tuning experiment. The first two steps correspond to the top layer weights initialization and this is reflected on the lower accuracy obtained in each of them. Following the fine-tuning of the DCNN layers, it can be observed that after 5 epochs of training all the DCNN architectures have the model performance stabilized. In some cases, like Inception-ResNet V2 the stabilization occurs even earlier, with only one epoch of fine-tuning being enough to generate a model with performance equivalent with the best one obtained.

No direct correlation between the number of trainable network parameters and network performance was found. Networks with lesser parameters performed better than ones with more trainable parameters on both experiments.

Finally, to make fair comparisons with results from recent works, only results obtained over the AVA dataset using the 10% subset were considered, as these have a similar dataset set up as the one used on this paper. Considering only the accuracy metric, the only model obtained that has been shown to outperform state-of-the-art models is the one based on the InceptionResNetV2 architecture. All the other obtained results are also comparable with state-of-the-art ones and can serve as a future baseline when comparing aesthetic classifiers results.

5. Conclusion

This paper proposes a high-level model for using DCNN architectures developed for the ILSVRC for the aesthetic classification problem on the AVA dataset with a two-neuron layer that behaves as a linear SVM as the top layer instead of a softmax activated one.

It is shown that transfer learning parameters without fine-tuning produce acceptable results when dealing with the image aesthetic classification problem. If sufficient computational power is not available, using transfer learning can lead to acceptable results for baseline models. It is possible to further improve over these models using fine-

tuning techniques, leading to results comparable with state-of-art performance. From all architectures in which experiments were run Inception-ResNet V2 had the best results, with better accuracy than other state-of-the-art models. The results obtained with the other models are also comparable with state-of-art results and they can serve as a baseline when comparing aesthetic classifier models.

Future work includes expanding the current methodology for the full AVA dataset. Also, it is possible to use the presented methodology to evaluate these architectures on the multiclass aesthetic classification, inferring score distributions for a given image.

References

Aydin, T. O., Smolic, A., and Gross, M. (2015). Automated aesthetic analysis of photographic images. *IEEE Transactions on Visualization and Computer Graphics*, 21:31–42.

Baldi, P. and Sadowski, P. J. (2013). Understanding dropout. pages 2814–2822.

Bhattacharya, S., Sukthankar, R., and Shah, M. (2010). Automated aesthetic analysis of photographic images. *Proceedings of the international conference on Multimedia MM '10*, pages 271–280.

Datta, R., Joshi, D., Li, J., and Wang, J. Z. (2006a). Studying aesthetics in photographic images using a computational approach. *Computer Vision ECCV 2006*, pages 288–301.

Datta, R., Joshi, D., Li, J., and Wang, J. Z. (2006b). Studying aesthetics in photographic images using a computational approach. *ECCV*, pages 7–13.

Deng, Y., Loy, C. C., and Tang, X. (2017). Image aesthetic assessment: An experimental survey. *IEEE Signal Process. Mag. 34*, pages 80–106.

Dhar, S., Ordonez, V., and Stony, T. L. B. (2011). High level describable attributes for predicting aesthetics and interestingness. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1657–1664.

Dong, Z., Shen, X., Li, H., and Tian, X. (2015). Photo quality assessment with dcnn that understands image well. *International Conference on Multimedia Modeling*.

Filho, J. G. (2009). *Gestalt do objeto*. Escrivuras Editora, Sao Paulo, SP, Brasil.

Gao, Z., Wang, S., and Ji, Q. (2015). Multiple aesthetic attribute assessment by exploiting relations among aesthetic attributes. *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval - ICMR '15*, pages 575–578.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861.

Jiang, W., Loui, A. C., and Cerosaletti, C. D. (2010). Automatic aesthetic values assessment in photographic images. *IEEE International Conference on Multimedia and Expo (ICME), 2010*, pages 920 – 925.

Jin, X., Chi, J., Peng, S., Tian, Y., Ye, C., and Li, X. (2016a). Deep image aesthetics classification using inception modules and fine-tuning connected layer. *8th International Conference on Wireless Communications & Signal Processing (WCSP)*.

Jin, X., Wu, L., He, Z., Chen, S., Chi, J., Peng, S., Li, X., and Ge, S. (2017). Efficient deep aesthetic image classification using connected local and global features. *arXiv:1610.02256v2 [cs.CV]*.

Jin, X., Wu, L., Li, X., Zhang, X., Chi, J., Peng, S., Ge, S., Zhao, G., and Li, S. (2016b). Ilgnet: Inception modules with connected local and global features for efficient image aesthetic quality classification using domain adaptation. *8th International Conference on Wireless Communications & Signal Processing, WCSP 2016*.

Ke, Y., Tang, X., and Jing, F. (2006). The design of high-level features for photo quality assessment. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 419–426.

Khan, S. S. and David, D. V. (2012). Evaluating visual aesthetics in photographic portraiture. *Proceedings of the Eighth Annual Symposium on Computational Aesthetics in Graphics, Visualization, and Imaging*, pages 55–62.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*.

Li, C., Loui, A. C., Chen, T., and Gallagher, A. (2010). Aesthetic quality assessment of consumer photos with faces. *Proceedings - International Conference on Image Processing, ICIP*, pages 3221–3224.

Lo, K.-Y., Liu, K.-H., and Chen, C.-S. (2012). Assessment of photo aesthetics with efficiency. *International Conference on Pattern Recognition (ICPR), 2012*, pages 2186–2189.

Lu, X., Lin, Z., Shen, X., Mech, R., and Wang, J. Z. (2015). Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), ICCV '15*, pages 990–998.

Luo, W., Wang, X., and Tang, X. (2013). Content-based photo quality assessment. *IEEE Transactions on Multimedia*, 15:1930–1943.

Luo, Y. and Tang, X. (2008). Photo and video quality evaluation: Focusing on the subject. *Computer Vision ECCV 2008*, pages 1 – 14.

Ma, S., Liu, J., and Chen, C. W. (2017). A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '17)*, pages 722–731.

Marchesotti, L. and Perronnin, F. (2013). Learning beautiful (and ugly) attributes. *British Machine Vision Conference*, pages 1–11.

Marchesotti, L., Perronnin, F., Larlus, D., and Csurka, G. (2011). Assessing the aesthetic quality of photographs using generic image descriptors. *Proceedings of the IEEE International Conference on Computer Vision*, pages 1784–1791.

Marchesotti, L., Perronnin, F., and Murray, N. (2015). Discovering beautiful attributes for aesthetic image analysis. *International Journal of Computer Vision*, 113:246–266.

Mavridaki, E. and Mezaris, V. (2015). A comprehensive aesthetic quality assessment method for natural images using basic rules of photography. *Proceedings - International Conference on Image Processing, ICIP*, pages 887–891.

Murray, N., Marchesotti, L., and Perronnin, F. (2012a). Ava: A large-scale database for aesthetic visual analysis. *CVPR 2012*, pages 2408–2415.

Murray, N., Marchesotti, L., and Perronnin, F. (2012b). Learning to rank images using semantic and aesthetic labels. *British Machine Vision Conference*, pages 1–10.

Nishiyama, M., Okabe, T., Sato, I., and Sato, Y. (2011). Aesthetic quality classification of photographs based on color harmony. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 33–40.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. on Knowl. and Data Eng.*, 22(10):1345–1359.

Sandler, M., Howard, A. G., Zhu, M., Zhmoginov, A., and Chen, L. (2018). Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *CoRR*, abs/1801.04381.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Suran, S. and K., S. (2015). Aesthetic quality assessment of photographic images: A literature survey. *International Journal of Computer Applications*, 132:11–15.

Szegedy, C., Ioffe, S., and Vanhoucke, V. (2016). Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2015). Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567.

Talebi, H. and Milanfar, P. (2018). Nima: Neural image assessment. *IEEE Transactions on Image Processing*, pages 3998 – 4011.

Tang, Y. (2013). Deep learning using linear support vector machines. *CoRR*, abs/1306.0239.

Tian, X., Dong, Z., Yang, K., and Mei, T. (2015). Query-dependent aesthetic model with deep learning for photo quality assessment. *IEEE Transactions on Multimedia*, 17:1–1.

Wang, W., Zhao, M., Wang, L., Jiexiong, H., Cai, C., and Xu, X. (2016). A multi-scene deep learning model for image aesthetic evaluation. *Signal Processing: Image Communication*, 47.

Wong, L.-K. and Low, K.-L. (2009). Saliency-enhanced image aesthetics class prediction. *Proceedings - International Conference on Image Processing, ICIP*, pages 997–1000.

Zhang, L., Gao, Y., Zimmermann, R., Tian, Q., and Li, X. (2014). Fusion of multichannel local and global structural cues for photo aesthetics evaluation. *IEEE Transactions on Image Processing*, 23:1419–1429.

Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V. (2017). Learning transferable architectures for scalable image recognition. *CoRR*, abs/1707.07012.