Multitask Regularization for Image Aesthetic Evaluation

Corneliu Florea, Laura Florea
Image Processing and Analysis Laboratory (LAPI), University Politehnica of Bucharest
Bucharest, Romania
{corneliu.florea;laura.florea}@upb.ro

***Abstract*—Convolutional neural networks are data hungry and in cases when annotation is costly or difficult, additional information from other sets may be welcomed. In this paper, to improve the performance on the main task, we introduce a secondary one, over unlabeled data to provide better structuring. The solution falls in the theme of multiple task learning and unlabeled data is integrated next the main regression task by classification via pseudo-labeling. The method is showed to improve the baseline performance for image aesthetic assessment on the AADB benchmark.**

***Index Terms*—multi-task, self-labeling, aesthetic, entropy regularization**

I. Introduction

Interesting and aesthetically pleasant images was long ago a desiderata of the photographer. Camera manufacturers and photographic software builders have sought methods to evaluate the aesthetic quality, to suggest improvements and make adjustments to given images. Recently, this direction has gathered new momentum due to the usage of convolutional neural networks. For a broader introduction to the topic and and more detailed taxonomy we refer to the review by Athar et al. [1]. In short, the aesthetic assessment proposes qualitative evaluation and solutions to judge how visually pleasant is an image, based on photographic rules [6].

The first challenge of the problem is to establish strong ground truth. The annotation process requires that many observers to judge the visual quality of content and the average of subjective ratings, called Mean Opinion Score (MOS), will be used as ground truth. The difficulty arises from the fact that aesthetics is a subjective matter and considering many opinions, one most likely will end having always averages scores. Alternatively, only few, but highly qualified persons may be involved; yet, such persons are rare and less willing to annotate large volumes of data. Educating many average users is expensive and time consuming. Therefore the second major challenge is to acquire volumes of data, accurately annotated such that the power of deep learning may be unleashed.

In this paper, we tackle the problem of aesthetic evaluation using a small, but carefully annotated databases. We build a CNN based solution and, observing the limited information in the training set, we propose a strategy to use additional unlabeled data. Since the main task is of a regression, while the unlabeled data is used in a classification framework, the problem is treated as *multiple task learning*.

The remainder of the paper is organized as follows: in section II we review some of the relevant prior works. In section III we formalize the proposed algorithm. Details of implementation and achieved results are in section IV. The paper end with discussion and conclusions.

This work was supported by the Ministry of Innovation and Research, UEFISCDI, by the project TRANSLATE, TE 66/2020, PN-III-P1-1.1-TE-2019-0543. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

II. Related work

Because the theme of this paper is the image aesthetic evaluation and the technical contribution lies into the transfer learning, we will review relevant works in both directions.

Predicting Image Aesthetics The idea of constructing artificial measures for the pleasantness of an image have intrigued computer vision researcher for a while. Since the aesthetics are subjective, a starting point is the understanding of human neurological reaction to the beauty [10]. From a computer vision point of view, older works used classical feature descriptors and classifiers [5], [12], while the later ones tapped in the deep leaning power. Here, Kong et al. [9] tried different loss functions over features extracted from the top layers of a pre-trained CNN; Reddy et al. [14] proposed a visualization technique to inspect the CNN–found relation between attributes and global score. Codella et al. [4] also imposed constrains over previous layers in a manner similar to ours; the difference lies in the fact that they use a unique task and formulated the loss function such to affect multiple layers, while we used multiple tasks, but the loss affects directly on the the exit layer.

Transfer Learning for Multiple task Learning The limited amount of annotated data available of in a supervised learning has troubled researchers in the field for a long period. Using additional domain related, but different data has its appeals. The direction is called transfer learning and strong reviews may be followed in now classical work of Pan et al. [13] and more recently in the one by Zhuang et al. [17]. The hereby proposed solution follows the track of Caruana [3] which suggested that when data is scarce or with incomplete annotation, it is better to learn multiple tasks at once, in opposition to a single one.

Later, many other works coerced CNNs to approach multiple tasks, including on unlabeled data, to improve the performance on the main issue. For instance, Ge et al. [7] injected

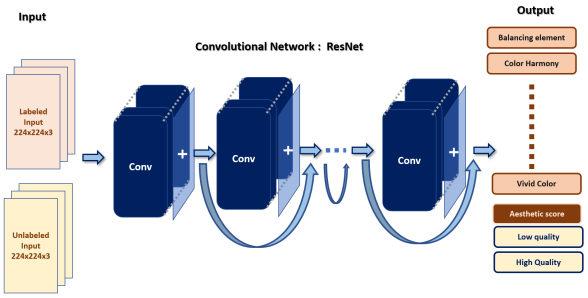

Fig. 1. The schematic of the learner in the proposed solution. While various architectures are tried, the preferred version uses residual networks.

additional data as synthetic unknown classes produced by a Generative Adversarial Network (GAN) and taught a CNN to classify those instances. Bendale and Boult [2] enforced a penultimate activations from pre–trained CNN to structure additional unlabeled data. Yoshihashi et al. [16] used a hand crafted hierarchical CNN architecture for open-set recognition by jointly training for classification and reconstruction.

Our proposal differ from prior art by the nature of task addresses (regression and classification for aesthetics) and by the actual solution used, that is to structure unlabeled data by self-labeling via entropy regularization.

III. Method

In this paper we employ a pair of databases. As learner a convolutional neural network (CNN) is employed. The core idea is that CNN can address simultaneously multiple tasks. Out of the two databases, one is labeled, and one is not. The images have the same nature, as they are photographs aiming to exhibit aesthetic qualities. The images from one database contains multiple labels describing a unique image, thus addressing the problem of multiple instance learning, while the ones from the second set no label, but in the learning process, we attach them to a binary classification problem.

In order to formalize the previously mentioned intuitive ideas, we will denote by $\{\mathcal{X}^l, \mathcal{Y}^l\} = \{(\mathbf{x}_i^l, \mathbf{y}_i^l)\}_{i=1}^N \overset{\text{iid}}{\approx} p(\mathcal{X}^l, \mathcal{Y}^l)$ the labeled set and the unlabeled by $\{\mathcal{X}^u\} = \{\mathbf{x}_j^u\}_{j=N+1}^{N+M} \overset{\text{iid}}{\approx} p(\mathcal{X}^u)$. The learned predictor is $f : X \to Y$, $f \in \mathcal{F}$ where $\mathcal{F}$ – hypothesis space. The learner will be used to produce predictions for the unlabeled set: $\hat{\mathcal{Y}}^u = f(\mathcal{X}^u)\}$; $\mathcal{Y}^u$ - results after cleaning the label space $\hat{\mathcal{Y}}^u$.

With respect to the probability density functions of the databases, both $p(\mathcal{X}^l)$, $p(\mathcal{X}^u)$ are drawn from the same distribution $p(\mathcal{X}^l)$, while the labeled space is different from the built on-line for the initially unlabeled one: $p(\mathcal{Y}^u) \neq p(\mathcal{Y}^l)$. The problem may bee see as a combination of multiple instance learning (as we learn different tasks) and transfer learning as we use information from the unlabeled domain to improve the performance in the labeled one. According to the taxonomy established by Pan et al. [13] it is a case of inductive transfer learning, implemented as *multi-task learning*.

Input: Labeled inputs $\mathbf{x}_i^l$, labels y_i^l. Unlabeled inputs $\mathbf{x}_j^u$.
Initialize: Net weights θ_0 .
for *epoch=1:*N_{ep}*:* **do**
  for *b=1:*N_{batch}*:* **do**
    Pass the labeled batch b: $(\mathbf{x}_b^l; y_b^l)$:
    a. Find predicts $y_{pred} = f_{\theta_b}(\mathbf{x}_b^l)$;
    b. Compute labeled loss $\mathcal{L}_l$;
    Pass an unlabeled batch b: $\mathbf{x}_b^u$
    c. Find predicts $\hat{y}_b^u = f_{\theta_b}(\mathbf{x}_b^u)$;
    d. Determine high confidence predicts $y_b^u = \operatorname{argmax} \hat{y}_b^u$;
    e. Compute unlabeled loss $\mathcal{L}_u$;
    Compute total loss $S(\theta)$ Compute gradient $g^u := \nabla S(\theta)$;
    Update on net parameters $\theta_b = \theta_{b-1} + g_u(\theta_{b-1})$ using SGD ;
  end
end
Result: trained network f_θ

Algorithm 1: Multi-Task Transfer algorithm.

The solution employed to give labels to the unlabeled data lies within the *self-learning* paradigm [15] as it depends strictly on the learner. The hereby choice, introduced by Lee [11] is called pseudo-labels; its main ide is that from the prediction, with respect to each class, it chooses the class with most confidence. Pseudo-labels is developed from the entropy minimization [8] by forcing the learner to be more assertive over the prediction space.

The overall problem can be approached by solving :

$$\begin{aligned} f_\theta(\mathbf{x}) &= \operatorname{argmin}_\theta S(\theta) \\ S(\theta) &= \mathcal{L}_l(\mathbf{y}^l; \mathbf{x}^l, \theta) + \lambda \mathcal{L}_u(\mathbf{x}^u, \theta) + R(\theta) \end{aligned} \tag{1}$$

Here $R(\theta)$ is a regularizer, implemented as the standard L_2 over weights (i.e. $R(\theta) = \alpha \sum_\theta \|\theta\|^2$). $\mathcal{L}_l(\cdot)$ is the loss function over the labeled part; here, various solutions are attempted but the preferred version uses L1 regression. λ is a parameter set empirically in this work to 0.5. $\mathcal{L}_l(\cdot)$ is the loss over the unlabeled set and implemented as cross-entropy :

$$\mathcal{L}_u(\mathbf{x}^u, \theta) = \sum_{j=N+1}^{N+M} -y_j \log \hat{y}_j - (1 - y_j) \log(1 - \hat{y}_j) \tag{2}$$

where $\hat{y}_j = f(\mathbf{x_j}$ is the corrupted version of the prediction of the CNN f over an unlabeled input, while y_j is the cleaned version. Following the pseudo–label principle, y_j is obtained by rounding $\hat{y}_j$ to be 100% assigned to a single class chosen as maximum from activations for the SoftMax layers in the CNN.

Algorithm 1 summarizes the proposed solution.

IV. Implementation and Results

A. Implementation

We have implemented the proposed method in Python using the Pytorch library. The code has been accelerated using Titan X GPU. Image resolution is 227×227 and batches for both labeled and unlabeled are of 32. The optimization has been carried using Stochastic Gradient Descent. We have trained for 150 epochs for small architectures (AlexNet, ResNet-34, ResNet-18) and 100 for the Resnet-50. For the first third of the training period, the learning was $5 \cdot 10^4$, followed by decimation upon every consecutive third. An epoch, for the complete version, took between 1m30sec to 2min30sec, depending on the architecture.

B. Databases

The largest image database for the study of aesthetics is The Aesthetic Visual Analysis (AVA) dataset [12]. AVA amounts to approximately 250,000 images, which that initially have been retrieved from the DPChallenge.com followed by a crowd-sourcing based annotation: each image received between 78 and 90 votes with a score ranging from 1 to 10. Yet the broad nature of the subjectivness in annotation brought its limitation: 80% of the images have a MOS between 4.6 a 6.2; thus, they are considered "average" and the standard deviation of the user score is not small either. The database is not highly helpful to teach excellence in aesthetics. In this paper, we use a subset of 50.000 randomly chosen images. Inspired by several task attempted on this database, we enforce the classifier to separate the chosen images betwen *high quality* and *low quality* (i.e. binary classification).

Better isolation of aesthetic attributes lies in the smaller AADB (Aesthetics and Attributes DataBase) by Kong et al. [9]. AADB creators collected photographic images within a broad thematic and composition frames from the Flickr website with a Creative Commons license. They further cleansed the set by eliminating non-photographic images such as cartoons, drawings, paintings etc. Followed the advice from professional photographers, they have selected 11 attributes relevant to the appraisal of aesthetic value: **interesting content**, *object emphasis, good lighting, color harmony, vivid color, shallow depth of field, motion blur, rule of thirds, balancing element, repetition, and symmetry*. Each image was annotated with a score between -1 and 1 for these attributes; the average is taken as the global aesthetic score. Overall, the AADB dataset contains 10,000 images, out of which 1000 are in test and the rest in training and validation. By averaging the opinion of multiple users, each attribute has small variation value. The strength of using carefully selected and annotated images, allowed AADB creators to obtain a highly competitive score on AVA database, by training the network mainly on the AADB [9].

Images from the two databases may be followed in figure 2. For images originated in AADB database, we show also the annotated attributes.

C. Results

The performance of the proposed algorithm is evaluated by Spearman Correlation Coefficient, ρ and Mean Absolute Error for the average global score and Mean Square Error for all attributes and global score. In the experiments, first we establish a supervised baseline by using only the AADB database and seeking the best loss function and respectively architecture. Next, experiments using images from both databases, in the multi-task transfer learning proposed framework, do follow.

Metric in supervised learning. The first assumed task is the identification of loss function used to train, in purely supervised manner, the CNN. Attempted losses have been L1, L2 and modified Hellinger coefficient (in the sense that here we have normalized the labels to be between 0 and 1). The basic network architecture was ResNet-18. Results may be followed in table I and one may see that L1 loss provided the best results and it will be further used.

Architecture. The second task was to see the influence of the architecture over the performance. We have tried AlexNet, ResNet-18, ResNet-34 and ResNet-50. Following the results from table I, one may conclude that using larger networks does not help due to limited amount of data and ResNet-18 suffices. The baseline is thus composed by ResNet-18 architecture trained with L1 regression loss.

Multi-task learning Different works experimenting on the AADB database used varying metrics. The performance of the proposed solution and comparison with prior art may be followed in table II. The baseline for our solution is supervised training or a ResNet 18 architecture with L1 loss. One may easily notice that using the additional task, although vaguely defined, improves significantly over baseline, thus arguing for the power of our solution. However, when compared to other works, this solution does not reach state of the art performance. For instance, when comparing to the original work of Kong et al. [9], they have carefully experimented with all attributes and showed that only 6 are advantageous; using the full set, as we did, does not present an advantage and was omitted in [9].

V. Discussion

In this paper we have addresses the problem of aesthetic evaluation on the short but carefully built AADB database. We have have successfully improved the baseline performance by introducing a supplementary task to separate good images from bad ones from an aesthetic point of view. The self-prediction was based on the pseudo-labeling concept that is simple but efficient The small size of the database lead to the fact the smaller architectures are better than larger ones, as the latter tend to overfit.

References

[1] S. Athar and Z. Wang. A comprehensive performance evaluation of image quality assessment algorithms. *IEEE Access*, 7, 2019.

[2] A. Bendale and T. E. Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016.

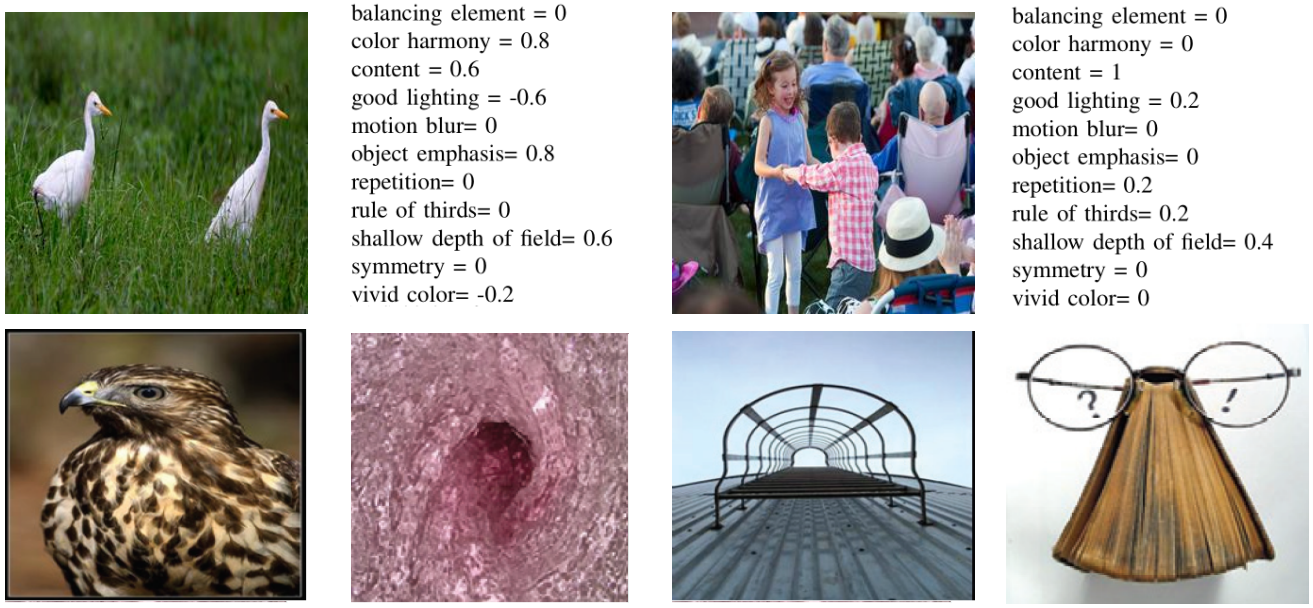

Fig. 2. Examples of images from the AADB (top row) with attributes marked and warped to the square size used by CNN and respectively from AVA, which are used as unlabeled in this work

TABLE I

MEAN SQUARE ERROR ON AADB WHIEL USING VARIOUS COMBINATIONS OF ARCHITECTURE AND LOSS FUNCTION. ACRONYMS STAND AS FOLLOWS: BE–BALANCING ELEMENT, CH–COLOR HARMONY, CO–CONTENT, GL–GOOD LIGHTING, MB–MOTION BLUR, OE–OBJECT EMPHASIS, RE–REPETITION, RT–RULE OF THIRDS, DF–SHALLOW DEPTH OF FIELD, SY–SYMMETRY, VC–VIVID COLOR.

Loss	**Architecture**	BE	CH	CO	GL	MB	OE	RE	RT	DF	SY	VC	All
L2	ResNet-18	0.044	0.074	0.202	0.048	0.107	0.011	0.189	0.019	0.054	0.008	0.090	0.029
Hellinger	ResNet-18	0.041	0.064	0.178	0.050	0.094	0.010	0.171	0.021	0.043	0.009	0.082	0.022
L1	ResNet-18	0.038	0.058	0.163	0.044	0.082	0.009	0.137	0.019	0.048	0.008	0.074	0.021
L1	AlexNet	0.044	0.074	0.204	0.047	0.107	0.011	0.191	0.019	0.054	0.008	0.085	0.028
L1	ResNet-50	0.043	0.075	0.202	0.047	0.113	0.011	0.189	0.016	0.055	0.008	0.084	0.024
L1	ResNet-34	0.043	0.073	0.194	0.045	0.102	0.012	0.185	0.015	0.057	0.008	0.082	0.024

TABLE II

COMPARISON BETWEEN THE PROPOSED SOLUTION AND OTHER WORKS. ρ STANDS FOR SPEARMAN CORRELATION COEFICIENT

Solution	**MAE**	**MSE**	ρ
Baseline	0.129	0.029	0.605
Proposed	0.125	0.021	0.619
Codella et al. [4]	0.136	–	–
Reddy et al. [14]	–	–	0.469
Kong et al. [9] AlexNet	–	–	0.59
Kong et al. [9] solution	–	–	0.678

[3] R. Caruana. Multitask learning: A knowledge-based source of inductive bias. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 41–48, 1993.

[4] N. C. Codella, M. Hind, K. N. Ramamurthy, M. Campbell, A. Dhurandhar, K. R. Varshney, D. Wei, and A. Mojsilovic. Teaching meaningful explanations. *arXiv preprint arXiv:1805.11648*, 2018.

[5] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *European conference on computer vision*, pages 288–301, 2006.

[6] Y. Deng, C. C. Loy, and X. Tang. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine*, 34(4):80–106, 2017.

[7] Z. Ge, S. Demyanov, Z. Chen, and R. Garnavi. Generative openmax for multi-class open set classification. In *British Machine Vision Conference*, 2017.

[8] Y. Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. In *NIPS*, pages 529 – 536, 2005.

[9] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *European Conference on Computer Vision*, pages 662–679, 2016.

[10] H. Leder, B. Belke, A. Oeberst, and D. Augustin. A model of aesthetic appreciation and aesthetic judgments. *British journal of psychology*, 95(4):489–508, 2004.

[11] D. Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshops*, 2013.

[12] 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*, pages 2408–2415, 2012.

[13] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[14] G. Viswanatha Reddy, S. Mukherjee, and M. Thakur. Measuring photography aesthetics with deep cnns. *IET Image Processing*, 14(8):1561–1570, 2020.

[15] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196, 1995.

[16] R. Yoshihashi, W. Shao, R. Kawakami, S. You, M. Iida, and T. Naemura. Classification-reconstruction learning for open-set recognition. In *EEEE Computer Vision and Pattern Recognition*, pages 4016–4025, 2019.

[17] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*,

109(1):43–76, 2020.