Learning Fashion By Simulated Human Supervision

Eli Alshan Sharon Alpert Assaf Neuberger Nathaniel Bubis Eduard Oks
Amazon Lab126
{alshan, alperts, neuberg, bubis, oksed}@amazon.com

Abstract

We consider the task of predicting subjective fashion traits from images using neural networks. Specifically, we are interested in training a network for ranking outfits according to how well they fit the user. In order to capture the variability induced by human subjective considerations, each training example is annotated by a panel of fashion experts. Similarly to previous works on subjective data, the panel votes are converted to a classification or regression problem and the corresponding network is trained and evaluated using standard objective metrics. The question is which objective metric, if any, is most suitable to measure the performance of a network trained for subjective tasks? In this paper, we conducted human approval tests for outfit ranking networks trained using various objective metrics. We show that these metrics do not adequately estimate the human approval of subjective tasks. Instead, we introduce a supervising network that unlike objective metrics, is designed to capture the variability induced by human subjectivity. We use it to supervise our outfit ranking network and we demonstrate empirically, that training our outfit ranking network with the suggested supervising network achieves greater approval ratings from human subjects.

1. Introduction

Training deep networks with millions of parameters over large-scale data sets has proven itself as a game changer for computer vision, driving algorithm performance on objective, type-specific tasks to near human-level performance. One of the attributing factors to this success is the precise way in which the tasks are defined. For example, assigning a label to an object from one of n predefined class categories. Such tasks have a well-established performance metrics, often directly optimized by the network.

Due to the success of deep networks, several works have tried to apply the same methodology to solve problems influenced by human subjective considerations. A few notable examples are problems like judging the aesthetic quality of an image [17, 1] or understanding fashion [12, 13]. If we look at a typical classification or regression problem each training example is associated with a single value or a label. The goal is to reduce the label variability as much as possible to avoid miss labeled examples that makes the training process much harder. However, for subjective tasks, the labels have an inherent variability because they represent the opinions of a group or a panel of human subjects. This variability is not a miss label but an integral part of the label itself.

The most common way to train a deep convolutional neural network (CNN) using this type of labels, is to either discretize or convert the votes to scalar score. This effectively converts the subjective task into regular objective classification or regression problem which is optimized using common network losses. Since the task is now considered objective, the performance of the trained network is measured using standard classification or regression metrics like precision/recall or mean square error. In customer-facing services, these metrics are also used as an estimate to algorithms' approval rating from real customers or experts in the field. Although this approach is true for objective tasks, it does not extend to subjective tasks because the human approval rating also varies due to human subjective considerations.

In this paper, we address the problem of training and evaluating CNNs for subjective tasks. As a test case, we are interested in the subjective task of understanding fashion. Specifically, selecting the most fashionable outfit by ranking pairs of outfits. Following other approaches to learning from subjective data e.g.[12, 13, 17], we transformed the subjective labels to objective targets and trained a CNN using standard objective metrics. As opposed to other methods, we took an extra step and compared the objective metric to actual human approval rating. Our experiments show that the actual human approval differs significantly from the objective metrics. Hence, these metrics are limited in predicting the quality of an algorithm trained for subjective tasks.

Intuitively, the best way to produce an algorithm with high approval rates is to directly maximize the human approval ratings. However, this requires significant human supervision and is not practical for most applications. Instead, we suggest to sample the human approval space and train a model that mimics human supervision. We then use this ”mimicking” network to supervise the training process of a ranking network that ranks outfits according to their quality or fashionability. We demonstrate that using this training process, the resulting network achieves higher approval ratings from human subjects compared to other methods.

The main contributions of the paper include:

- Understanding the performance gap in conventional approaches for training CNNs for subjective tasks.
- A new training framework for subjective data to maximizing human approval.

2. Related works

Automatic detection of image related subjective traits has been a key research area in computer vision for quite a long time. In virtually all of the works related to subjective traits, the training labels are obtained by mapping the votes of a panel of human subjects to a discrete set of labels or a single score.

A popular problem associated with subjective traits is estimating an image’s aesthetic quality. Datta et al. [2, 1] estimated the mean vote of the panel and used handcrafted features, such as color and texture, to train an SVM classifier for estimating aesthetic quality. Lu et al. [9] used a predefined threshold to binarize the panel votes into either high or low aesthetic quality and trained a CNN on a binary classification problem. Workmen et al. [17] explored several approaches like discretization, mean and panel distribution for estimating the image scenicness.

Apart from image aesthetics, predicting the mean of an annotator panel was used to predict facial attractiveness [8], evaluation of facial beauty [4] and exploring image memorability [7]. Binarization was used to estimate urban perception [11] and to study the phenomenon of image virality [3].

In this work, we focus on the domain of fashion images, which has only recently become the focus of research. Apart from more traditional objective problems like segmentation of garments [19] and style classification [16] there were also attempts to capture the aesthetics in fashion. Most notable is [12] that introduced the Fashion144K dataset that assigns a fashionability measure, ranging from 1 (not fashionable) to 10 (very fashionable) to each outfit, based on users’ votes. The authors used this data set to train deep networks to estimate the discrete fashionability score.

3. Learning fashion using CNNs

We train a CNN for ranking outfits according to how well they fit the user. Our dataset consists of image pairs showing individuals wearing different outfits. Each pair depicts the same person wearing two different outfits. As in similar subjective problems, the labels are the votes of a panel of fashion experts where each fashion expert selects the outfit he/she thinks is better. The number of experts voting for each pair varies between 1-100.

We consider the CNN as a function $R(I_A^i, I_B^i; \Theta_R)$, where I_A^i, I_B^i are the corresponding outfit images of pair i and Θ_R represents the model parameters. In light of recent works like [18, 17] we evaluate several ways to formulate the problem using objective terms. One option is to consider this problem as a binary classification task where the groundtruth represents the majority vote in the experts’ panel. Using this formulation, the output of the model is the probability of each outfit receiving the majority vote of the panel of experts. Training usually involves minimizing the typical Cross-Entropy loss:

$$\arg\min_{\Theta_R}\{-\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\}, \quad (1)$$

where $y_i, \hat{y}_i$ are the binary indicator of the majority vote in of the panel of experts and the predicted panel consensus produced by $R(I_A^i, I_B^i; \Theta_R)$, respectively. The above approach assumes an underlying binary value representing the majority vote in a winner-takes-all manner. However, for many outfits, there is high variability in the expert votes and the selected outfit might be only marginally better. In this case, the mean or the consensus of the experts panel may serve as a better representative value. The panel consensus for a pair of outfits I_A^i, I_B^i is:

$$\hat{c}_i = \frac{\#\text{votes}(I_A^i)}{\#\text{votes}(I_A^i) + \#\text{votes}(I_B^i)} \quad (2)$$

where $\hat{c}_i \in [0, 1]$. When $\hat{c}_i > 0.5$ the majority of experts voted for I_A^i and when $\hat{c}_i < 0.5$ the majority of the experts voted for I_B^i. Training a network to predict $\hat{c}$ involves some form of a regression loss like Cross-Entropy (1) or some ℓ_p metric:

$$\arg\min_{\Theta_R}\{\frac{1}{N}\sum_{i=1}^{N} |R(I_A^i, I_B^i; \Theta_R) - \hat{c}_i|^p\}. \quad (3)$$

Each loss function provides a different penalty for deviations from the ground-truth. The question is which of them correlates well with human perceived errors? To answer this , we empirically tested each method on our data set and computed the accuracy of each model. In order to make this comparison valid, each experiment has the same model (Siamese Inception-Resnet v2) with the same initialization weights. For the binary case, we trained the network using Cross-Entropy (1) and for the panel consensus we trained with Cross-Entropy (1), Mean Square Error (MSE) , Mean Absolute Error (MAE) and Huber loss.

3.1. Traning CNN's for outfit ranking

Our data set consists of images showing people wearing different outfits. Since the input to the network is a pair of images, we compose pairs of images showing the same person wearing different outfits. We allow each image to be part of several different pairs.

To obtain the groundtruth labels each pair of outfit images is annotated by 1 to 15 fashion experts, where the median number of annotations per pair is 9. Our data set is divided into three non-overlapping sets: train, validation and test. The sets are divided such that the pairs showing the same person can only appear in one set. To avoid sampling errors [10], all the outfit pairs in the test set were annotated by a panel of at least 40 fashion experts.

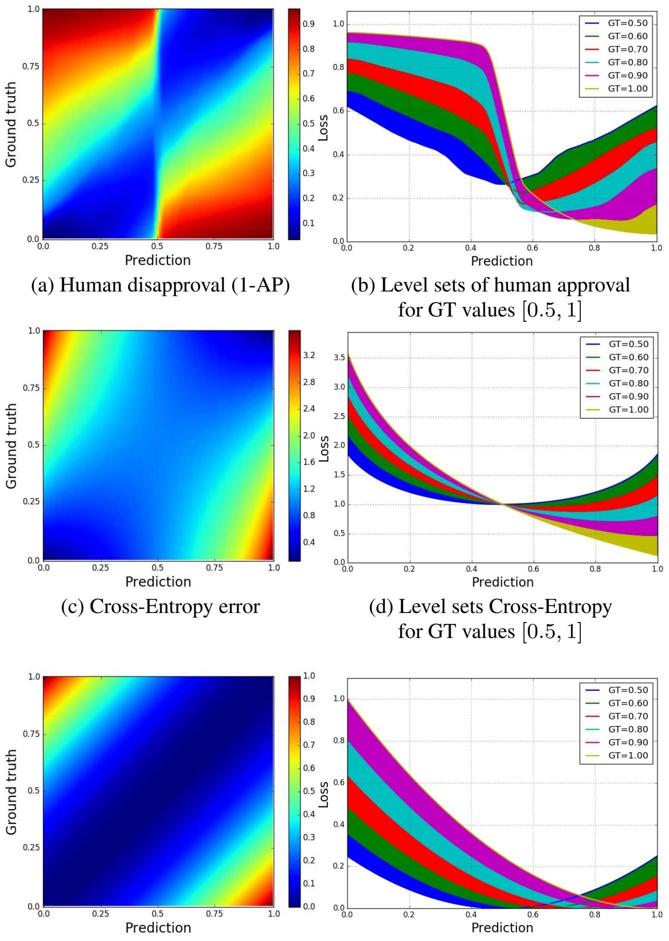
For each label definition, we trained a network using several loss functions. Two quantitative metrics are used for evaluation each network variant: (1) Binary accuracy - measures only the accuracy in predicting the "winning" outfit , and (2) rMSE - measures the deviation in the consensus prediction. Table 1 summarize the performance of the various models.

3.2. Comparison with Human Ratings

In order to evaluate the compatibility of the approaches above with human subjective perception of error, we conducted a human rating test. The test was performed by presenting a human subject, in our case a fashion expert, with a pair of outfits together with the algorithm prediction. The human subject is asked to either approve or disapprove (0 or 1) the algorithm prediction. Our test set contains the rating of 62 fashion experts for 3.7K outfit pairs. For each approach, we compute the average approval on the entire test set. Since we don't expect that any approach would reach a perfect 100% approval rate we computed an upper bound for the performance by considering the average approval rate of the perfect consensus estimator (prediction equals groundtruth). This approach achieves only 75.5% approval rate, due to subjective considerations of the fashion experts.

We can see that all of the approaches achieved comparable binary accuracy. However, a network trained using binary labels scored 2.6% less in human approval than the same network trained on panel consensus labels. This shows that the errors induced by the binary labels approach were perceived by the fashion experts as more significant than the other methods. In addition, we can see some correlation between the rMSE metric and the human approval ratings, although the Huber loss achieved comparable human approval with slightly bigger rMSE.

To understand the root cause of this outcome, we collected the raw data of the human approval test and estimated the approval/disapproval surface. We discretize the space into 10K equally spaced bins of predicted and groundtruth consensuses $\hat{c}_{pred}, \hat{c}_{gt} \in [0, 1]$. For each bin, we computed the average approval (AP) and disapproval (1-AP) of the fashion experts votes. Figure 1 shows the disapproval (1-AP) approval surface for the fashion experts votes (a) compared to Cross-Entropy and RMSE (b& c). Below each surface, we plot the corresponding level sets for ground-truth consensus values $[0.5, 1]$.



(a) Human disapproval (1-AP)

(b) Level sets of human approval for GT values $[0.5, 1]$

(c) Cross-Entropy error

(d) Level sets Cross-Entropy for GT values $[0.5, 1]$

(e) Root Mean Square error

(f) Level sets Root Mean Square for GT values $[0.5, 1]$

Figure 1. Human subjects disapproval surface (a) and its corresponding level set (b). For comparison we've added the error surfaces of CE loss (c) and rMSE loss (e), their corresponding surface level sets are shown in (d & f). Note how different the Human subjects disapproval surface from both loss surfaces).

It is interesting to see that the human disapproval surface is quite different from both Cross-Entropy and rMSE. The correlation between rMSE and the human approval shown in the empirical evaluation (Table 1) does not imply causation. If we look at the disapproval surface, one can see that for large groundtruth consensus values (> 0.8 or < 0.2), the human approval has an inflections point on 0.5. As the predicted consensus moves towards the ground true value there is only a minor increase in the approval rating. This

Labels	Loss	Accuracy	rMSE	MAE	Human approval
Binary	Cross-Entropy	0.682	0.263	0.208	0.668
Panel consensus	Cross-Entropy	**0.689**	0.249	0.200	0.694
Panel consensus	MSE (ℓ_2)	0.686	**0.243**	0.199	**0.699**
Panel consensus	MAE (ℓ_1)	0.685	0.248	**0.198**	0.693
Panel consensus	Huber	0.679	0.246	0.201	**0.699**

Table 1. Performance comparison between objective loss functions for the outfit ranking task.

behavior is similar to binary classification where only a change in the winning outfit and not the value of the consensus triggers a change in the approval. On the other hand, we see a very different behavior in consensus values close to 0.5, where the absolute distance between the predicted and groundtruth value trigger changes in the approval.

It is clear from our evaluation. that human disapproval is quite different from losses like CE and rMSE. It is therefore natural to ask whether we can design a loss function that captures more accurately human perceived errors. We next show a how we utilize it for training a CNN for outfit ranking.

4. Fashion Ranking using Simulated Supervision Network

Our goal is to train a CNN for outfit ranking. From the previous section, it is clear that we need a better way to estimate the approval/disapproval manifold of our fashion experts. Our proposed approach uses a CNN to estimate the approval/disapproval manifold, this network is then used to supervise the outfit ranking network by providing gradients to update the ranking network weights. We argue, that this supervising network is able to capture more accurately the variations induced by subjectivity and therefore can provide more accurate supervision to the ranking network. First, we show how the simulated supervision network (denoted SN) is constructed and discuss several different variants of this network. Finally, we show how this network serves as a building block for the fashion ranking network optimization.

4.1. Simulated Supervision Network

We construct a network that performs a simulation of the human approval test. If we recall Section 3.2, the human approval test is conducted by presenting the fashion experts a pair of outfits together with the algorithm's prediction $\hat{c}_{pred} \in [0, 1]$. The fashion expert is then asked to either approve or disapprove (0 or 1) the algorithm's prediction.

The suggested supervising network (SN) shares a similar structure to the human approval test. Given an input return a binary value that either approve or disapprove the input. To collect data to train our supervising network, we conducted a similar approval test as above, but instead of the algorithm's prediction $\hat{c}_{pred}$ we show the fashion experts a uniform random prediction $\hat{c}_{rand} \sim \text{Uniform}(0, 1)$. The human annotators were not aware that the predictions were randomly generated.

4.2. Dataset for training the Simulated Supervision Network

We collected 7,289 different outfit pairs, generating 60,000 training examples. Each example $(\hat{c}_{gt}, \hat{c}_{rand})$ is associated with a corresponding binary value indicating approval/disapproval. The pairs were generated by discretizing the space spanned by all potential $\hat{c}_{gt}, \hat{c}_{pred}$ combinations into 100 bins. We sampled enough points such that the standard error of the average approval rate in each bin is below 0.01. The ground-truth value for each pair $\hat{c}_{gt}$ was estimated by computing the mean votes of a panel of 20 fashion specialists. We optimize the supervising network parameters, Θ_{SN}, by maximizing the approval prediction accuracy using Cross-Entropy loss

$$\arg\min_{\Theta_{SN}} \{-\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\}. \quad (4)$$

where $y_i, \hat{y}_i$ are the groundtruth expert approval and the predicted approval produced by $SN(\hat{c}^i_{gt}, \hat{c}^i_{pred}; \Theta_{SN})$, respectively.

We train two variants of the supervising network. The first only considers the approval/disapproval as a function of the groundtruth and the predicted value. The second also considers the two outfit images as inputs to the network. We discuss the performance of both networks in the Section 5.

4.3. Training the Ranking Network

We train a network, denoted R, for outfit ranking. We train R to maximize the expected approval rating of the supervising network SN. During the training process, the ranking network R performs a simulated approval test on the mini-batch using the current model parameters. The simulated network SN in-turn provides gradients for R to improve its approval rating.

Formally, given a simulated supervision network SN with parameters Θ_{SN} the ranking network R attempts to maximize the approval:

$$\arg\max_{\Theta_R}\{\frac{1}{N}\sum_{i=1}^{N} SN(\hat{c}^i_{gt}, R(I^i_A, I^i_B;\Theta_R);\Theta_{SN})\}. \tag{5}$$

5. Experiments

We trained an outfit ranking network using the supervising network. In our experiments, we consider two variants of the supervising network. The first is *consensus only*, denoted by SN_c. The network receives two inputs, the ground-truth consensus $\hat{c}_{gt}$ and the algorithm prediction $\hat{c}_{pred}$ and outputs a binary value - approve/disapprove. The network is modeled by two fully connected layers with ReLU activations. The second layer outputs two logits that, after a soft-max normalization, represent $Pr(approval)$ and $Pr(disapproval)$.

We require that network should be invariant to the order of input images. Formally, given a supervising network SN, we require that the following relation holds:

$$SN_c(\hat{c}_{gt}, \hat{c}_{pred}) = SN_c(1-\hat{c}_{gt}, 1-\hat{c}_{pred}). \tag{6}$$

To enforce this symmetry, we generate a second input, $1-\hat{c}_{gt}$ and $1-\hat{c}_{pred}$, in addition to the original $\hat{c}_{gt}$, $\hat{c}_{pred}$. We pass both symmetrical inputs through the network and average the output logits.

The second network, denoted SN_I, is an *image based* network. The network inputs are a pair of images I_A and I_B, the ground-truth consensus $\hat{c}_{gt}$, and the algorithm prediction $\hat{c}_{pred}$. As with SN_c, the network outputs a binary value - approve/disapprove. For each image I_A, I_B we compute a visual descriptor using the global pooling layer of a Resnet50 network [5] trained on fashion domain images. We reduce the descriptor dimensions using a single fully connected layer resulting in two d-dimensional descriptors, f^d_A and f^d_B corresponding to images I_A and I_B. The input to the network SN_I is the concatenation of f^d_A, f^d_B, $\hat{c}_{gt}$ and $\hat{c}_{pred}$. Similarly to SN_c, we require that the network should be invariant to the order of input images. In this case, for each $I_A, I_B, \hat{c}_{gt}, \hat{c}_{pred}$ we require:

$$SN_I(I_A, I_B, \hat{c}_{gt}, \hat{c}_{pred}) = SN_I(I_B, I_A, 1-\hat{c}_{gt}, 1-\hat{c}_{pred}) \tag{7}$$

To enforce this symmetry, we generate a second feature vector by concatenating f^d_B, f^d_A, $1-\hat{c}_{gt}$ and $1-\hat{c}_{pred}$. Both symmetrical inputs are passed through the network, the output logits are averaged to form a symmetry input. The rest of the the network is modeled by two fully connected layers with ReLU activations. The final output is two logits normalized by a soft-max function to estimate approval probability.

We train both SN networks on 60,000 examples generated from 7,289 different outfit pairs (see Section 4.2) with 10% of the pairs left for evaluation. We evaluate three variants of the supervising network.

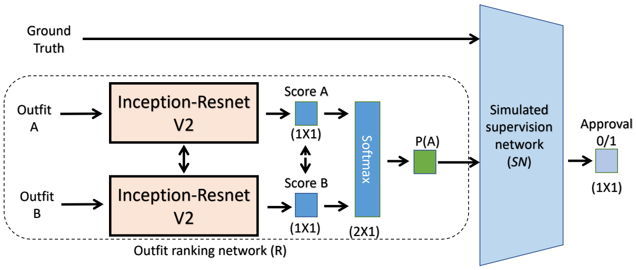


Figure 2. Our network architecture - the ranking network R is an Inception-Resnet v2 Siamese network. The supervising network SN is used to provide gradients for the training of R

1. SN_c - a supervising network that considers the approval ratings only as a function of the predicted consensus and the groundtruth consensus.

2. SN_{I8} - a supervising network that considers the approval ratings only as a function of the outfit images and the groundtruth consensus. The size of the image descriptor is 8.

3. SN_{I32} - a supervising network that considers the approval ratings only as a function of the outfit images and the groundtruth consensus. The size of the image descriptor is 32.

Potentially the length of the image descriptor can the size of the Resnet50 global pooling layer output (2048). However, due to a small number of training images, the networks with larger descriptor over-fit very quickly. Therefore, we reduced the descriptor's size by adding a fully-connected layer with output sizes of 8 and 32. The approval rate prediction accuracy of different variants of the SN networks is shown in Table 2.

Network variant	Accuracy
SN_c	0.698
SN_{I8}	0.693
SN_{I32}	0.688

Table 2. Accuracy of the different variants of the SN networks on the data collected from sampling the approval space of the fashion specialists.

5.1. Outfit ranking network

We train the outfit tanking network R to estimate the consensus of the panel of fashion experts $\hat{c}_{pred}$ given a pair of outfit images I_A, I_B. The training is performed by maximizing the approval rate using Eq.(5) for each variant of the supervising network SN. We implemented R as a Siamese neural net [6]. The network branches are modeled by Inception Resnet v2 [14] with a single output fully connected layer added after the global pooling layer. The pair of outfit images is fed into the respective branches of R resulting in

Labels	Loss	Accuracy	rMSE	MAE	Human approval
Panel consensus (best network from Table 1)	Huber	0.679	0.246	0.201	0.699
Panel consensus	$SN_c(\hat{c}_{gt}, \hat{c}_{pred})$	0.679	0.248	0.210	**0.715**
Panel consensus	$SN_{I32}(I_A, I_B, \hat{c}_{gt}, \hat{c}_{pred})$	0.666	0.258	0.221	0.704
Panel consensus	$SN_{I8}(I_A, I_B, \hat{c}_{gt}, \hat{c}_{pred})$	0.680	0.264	0.227	0.690

Table 3. Performance comparison between subjective loss functions for the outfit ranking task.

a score for each outfit. The outfit scores are normalized by a soft-max function to the estimated panel consensus $\hat{c}_{pred}$.

We trained all the layers of R without changing the parameters of the supervising network SN. Throughout our experiments, we used a learning rate of 0.003, decayed every two epochs using an exponential rate of 0.9. The network is optimized using RMSProp [15] with decay 0.9, momentum 0.9 and epsilon 0.1.

We train our network on the same database shown in Section 3.1. We show quantitative results in Table 3 on three metrics: binary accuracy, rMSE, and human approval. All the metrics were computed in the same manner as shown in Section 3. All the networks are based on the same Inception-Resnet v2 architecture with identical initialization weights. Table 3 shows the performance of the raking network using the supervision of each variant of SN, for reference we included the best performing network from Table 1.

Our algorithm achieved better human approval ratings than the other approaches. Our best performing network achieved approval rating of 71.5%. The 1.6% improvement over the objective loss is not negligible since the performance upper bound is 75.5% (see Section 3). We see that the network with the highest human approval rating had a higher rMSE than the network trained with MSE loss. This contradicts the trend shown in Table 1 where, for objective tasks, low rMSE was correlated with a high approval rating. We also notice that the best network in terms of human approval had one of the lowest accuracies among all the tested networks. In fact, the second best network, in terms of human approval, had the lowest accuracy compared to all the other networks. It is interesting to point out, that the supervising network based on images was less effective than the consensus only supervising network. We think that reasons for that the relatively low number of images (7K) used for training. We intend to extend the data set to further improve its performance.

The empirical findings shown in this section indeed indicate that for subjective tasks, standard objective metrics do not adequately predict how humans would perceive the network results. We demonstrated that for subjective tasks, besides ground-truth labels, it's highly beneficial to also collect data on how humans perceive or approve the networks' predictions. Having this data allows to better model the variations induced by human subjective considerations and provide better gradients for training networks for subjective tasks.

6. Conclusions

In this paper, we have shown that metrics used for classification and regression are a poor estimate for human approval when applied for subjective tasks. To better estimate the human approval, we suggested a scheme in which we first learn a supervising network that better estimates the errors as perceived by human subjects. Then, this auxiliary network is used to supervise the network trained on the subjective task. We have demonstrated the effectiveness of our method by applying it to the subjective task of outfit ranking. An empirical evaluation showed that our approach was able to achieve higher human approval rating than standard metrics used for classification and regression.

References

[1] 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. Springer, 2006.

[2] R. Datta, J. Li, and J. Z. Wang. Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 105–108. IEEE, 2008.

[3] A. Deza and D. Parikh. Understanding image virality. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1818–1826, 2015.

[4] D. Gray, K. Yu, W. Xu, and Y. Gong. Predicting facial beauty without landmarks. In *European Conference on Computer Vision*, pages 434–447. Springer, 2010.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] E. Hoffer and N. Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2015.

[7] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 145–152. IEEE, 2011.

[8] A. Kagian, G. Dror, T. Leyvand, D. Cohen-Or, and E. Ruppin. A humanlike predictor of facial attractiveness. *Advances in Neural Information Processing Systems*, 19:649, 2007.

[9] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 457–466. ACM, 2014.

[10] A. Neuberger, E. Alshan, G. Levi, S. Alpert, and E. Oks. Learning fashion traits with label uncertainty. In *Proceedings of KDD workshop Machine learning meets fashion*, 2017.

[11] L. Porzi, S. Rota Bulò, B. Lepri, and E. Ricci. Predicting and understanding urban perception with convolutional neural networks. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 139–148. ACM, 2015.

[12] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun. Neuroaesthetics in Fashion: Modeling the Perception of Fashionability. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[13] E. Simo-Serra and H. Ishikawa. Fashion Style in 128 Floats: Joint Ranking and Classification using Weak Data for Feature Extraction. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, page 12, 2017.

[15] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

[16] S. Vittayakorn, K. Yamaguchi, A. C. Berg, and T. L. Berg. Runway to realway: Visual analysis of fashion. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 951–958. IEEE, 2015.

[17] S. Workman, R. Souvenir, and N. Jacobs. Understanding and mapping natural beauty. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5590–5599. IEEE, 2017.

[18] L. Xie and S. Newsam. Im2map: deriving maps from georeferenced community contributed photo collections. In *Proceedings of the 3rd ACM SIGMM international workshop on Social media*, pages 29–34. ACM, 2011.

[19] W. Yang, P. Luo, and L. Lin. Clothing co-parsing by joint image segmentation and labeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3182–3189, 2014.