Training the DNN of a Single Observer by Conducting Individualized Subjective Experiments

2023 15th International Conference on Quality of Multimedia Experience (QoMEX)

Training the DNN of a Single Observer by Conducting Individualized Subjective Experiments

Pavel Majer[1], Lohic Fotio Tiotsop[2], Marcus Barkowsky[1]
[1]*Deggendorf Institute of Technology, University of Applied Sciences, Deggendorf, Germany*
[2]*Control and Computer Engineering Department, Politecnico di Torino, Torino, Italy*
pavel.majer@stud.th-deg.de, lohic.fotiotiotsop@polito.it, marcus.barkowsky@th-deg.de

***Abstract*—Predicting the quality perception of an individual subject instead of the mean opinion score is a new and very promising research direction. Deep Neural Networks (DNNs) are suitable for such prediction but the training process is particularly data demanding due to the noisy nature of individual opinion scores. We propose a human-in-the-loop training process using multiple cycles of a human voting, DNN training, and inference procedure. Thus, opinion scores on individualized sets of images were progressively collected from each observer to refine the performance of their DNN. The results of computational experiments demonstrate the effectiveness of our approach. For future research and benchmarking, five DNNs trained to mimic five observers are released together with a dataset containing the 1500 opinion scores progressively gathered from each of these observers during our training cycles.**

***Index Terms*—Individual quality perception, Artificial-intelligence-based observer, Subjectively annotated image dataset**

I. Introduction and Related Work

Several authors [1]–[3] have pointed out the need to go beyond the Mean Opinion Score (MOS) in order to achieve a more complete assessment of the end users' quality-of-experience (QoE). Therefore, approaches to predict the whole distribution of the opinion scores for a given stimulus have been proposed [4]–[6].

The distribution of opinion scores does not however tackle several QoE related questions. For instance, what are the characteristics of the end users that would not be satisfied with the quality of a given stimulus?

Authors have therefore proposed to predict the opinion scores of an individual observer [7]. The authors of [8]–[10] have trained a Neural Network (NN) that can mimic the quality perception of an individual observer. Such an NN is called an Artificial Intelligence-based Observer (AIO). The AIOs can allow for instance to address the aforementioned question, since each modeled observer has well known characteristics and their AIO can be used as a representative of all end users with these characteristics.

Individual opinion scores are known to be very noisy as compared to the MOS [11], [12]. Therefore, how to train an effective AIO is a challenging and hot research question. The literature in that sense is rather recent and limited. In [8], [9], the authors combined three subjectively annotated datasets to deal with the lack of training samples. In [10], the training of the AIOs was done in two learning steps. A deep CNN was first trained on a synthetically annotated large-scale dataset; then, the features learned by this deep CNN were refined during a second learning step to get the AIOs.

In this work, we propose a human-in-the-loop learning approach to train the AIOs. More precisely, each observer whose quality perception is to be mimicked with a deep CNN first rated a given set of images. The collected ratings were used to perform a first training of their AIO. The trained AIO was then used to make inference on a large-scale dataset, and to identify a new set of images that the observer must evaluate, i.e images for which the AIO provided a questionable prediction of the quality. The identified images were evaluated by the observer and the gathered opinion scores were used to refine the AIO during a second training process. The AIO obtained after this second learning step was again used to select a new set of images to be rated by the observer. Finally, a third and last training process was conducted with the gathered opinion scores to get the final deep CNN modeling the observer.

The obtained results show that the accuracy of the trained AIOs is comparable to that of a real observer when trying to repeat their opinion scores on a given set of images. Also, each AIO can predict the opinion scores of the observer it is mimicking with higher accuracy than the AIO of other observers. This suggests that each trained AIO did not learn only generic perceptual features, but rather features that model intrinsic characteristics of the scoring behavior of the real observer that it is mimicking.

The trained AIOs as well as the subjectively annotated dataset created for our analysis are made freely available to researchers at: *http://media.polito.it/AIOs-from-human-in-the-loop-training-process*.

II. Human Voting and AIO Training Cycles

A. Subjective Tests Setup and Paper's Notation

Our analysis aimed at training the AIO of five observers. Each of these five observers was invited in five different testing sessions. During each session, the observer evaluated the quality of 300 JPEG compressed images using the five point absolute category rating scale. The laboratory in which all sessions were conducted was prepared in accordance with

979-8-3503-1173-0/ 23/$31.00 ©2023 IEEE

Fig. 1. The picture illustrates the viewing distance, the used monitor and the lighting conditions under which the observers provided their opinion scores.

the relevant ITU recommendations [13]. The picture in Fig 1 illustrates the experimental setups.

We introduce the following notation used throughout the paper. We denote by $\mathcal{O}$ the set of the five observers to be modeled; by I_o^s the set of images evaluated by the observer $o \in \mathcal{O}$ during their s-th session. We denote by OS_o^s the opinion scores provided by the observer $o \in \mathcal{O}$ during their s-th session, i.e., when evaluating the images in I_o^s. Finally, we will call AIO_o^s, the AIO of the observer $o \in \mathcal{O}$, trained on the images in I_o^s, using the opinion scores in OS_o^s as ground truth.

All the images used in all the sessions were progressively selected from a dataset of 100,000 JPEG compressed images that we will call $\mathcal{I}$. The images in $\mathcal{I}$ were generated following the procedure used in [10], i.e., compressing 20,000 pristine quality images, selected from the ImageNet dataset [14], using five different ranges of JPEG compression so that the qualities of the images in the obtained dataset cover the whole five point absolute category rating scale.

For all the observers, the first $(s = 1)$ and the second $(s = 2)$ test sessions were carried out with the same set I^{s1s2} of images, i.e.,

$$I_{o_1}^1 = I_{o_2}^1 = I_{o_1}^2 = I_{o_2}^2 = I^{s1s2} \tag{1}$$

where o_1 and o_2 represent two generic observers selected from $\mathcal{O}$.

In other words, the first two sessions for each observer consisted in a repeated evaluation of the images in I^{s1s2}. This repetition, as shown later in the result section (see Table I), is useful to benchmark the performance of the trained AIOs.

During the third and the fourth session, each observer evaluated an individualized set of images. The sets of images that each observer evaluated in these two sessions were determined by the performance of their AIO at predefined stages of the training process. Thus, we have iterated between AIOs training phases and subjective experiments.

As for the first and the second session, during the fifth session, all subjects were shown a similar set of images, thus:

$$I_{o_1}^5 = I_{o_2}^5 = I^{test} \quad \forall o_1, o_2 \in \mathcal{O}. \tag{2}$$

The images in the set I^{test} were never used during the training of the AIOs and we therefore define this set of images and the related opinion score OS_o^5 gathered from each observer $o \in \mathcal{O}$ as the test set for our analysis.

The images in the sets $I_o^s \;\; s = 2, 3, 4$ and the corresponding opinion scores $OS_o^s \;\; s = 2, 3, 4$ were instead used to perform the three-steps training process that yielded the AIO of each observer $o \in \mathcal{O}$ following the procedure described in the next section. Note that the images in I_o^1 were not considered for the training because, as already mentioned, $I_o^1 = I_o^2 \;\; \forall o \in \mathcal{O}$.

B. A Human-in-the-Loop Approach to Train the AIOs

Fig 2 summarizes our approach to train the AIO of a generic observer $o \in \mathcal{O}$. For each observer $o \in \mathcal{O}$, we first trained the AIO_o^2 by performing transfer learning from a pretrained deep CNN called JPEGResNet50 [10]. The JPEGResNet50 is a deep CNN with 52 hidden layers that was trained on a large-scale synthetically annotated dataset to make it a suitable starting point for transfer learning in image quality assessment [10]. That is the reason why we started from such a network. All the trained AIOs in this work have the same architecture, i.e., the one of the JPEGResNet50, but they obviously have different weights.

The transfer learning to obtain the AIO_o^2 was performed on the images in I_o^2 using the opinion scores in OS_o^2 as ground truth. To obtain the AIO_o^2 from the JPEGResNet50, the latter was trained for 13 more epochs, using the stochastic gradient descent with momentum algorithm. The best learning rate was experimentally found to 0.0001 and the momentum parameters was set to 0.9 as typically recommended. These settings were adopted for all the training processes.

Once we obtained the AIO_o^2, we used it to select from $\mathcal{I}$ the next set of images, i.e., I_o^3, that had to be evaluated by the observer o, so that the provided opinion scores can be used to refine the AIO_o^2 yielding the AIO_o^3.

To identify the set I_o^3, we used the AIO_o^2 to predict the quality of the 100,000 images in the set $\mathcal{I}$. For each image, the softmax layer of AIO_o^2 outputs a five class probability distribution, i.e., the probability that the observer o would score the quality respectively as "Bad", "Poor", "Fair", "Good" and "Excellent". From this probabilistic output, we computed the variance of the prediction of the AIO_o^2 for each image in

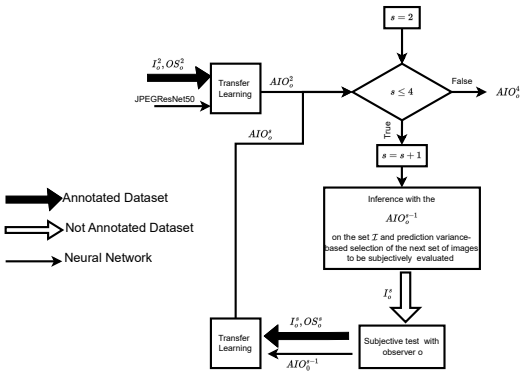

Fig. 2. The diagram summarizes our human-in-the-loop training procedure of the AIO of a generic real observer.

$\mathcal{I}$. We then obtained the set I_o^3 by selecting the 300 images whose quality prediction exhibited the highest variance.

The larger the variance of the prediction is for a given image, the lower is the confidence of the AIO on its prediction of the quality of that image. Therefore, a natural attempt to improve the AIO performance, is to ask the observer to rate these images that are critical for their AIO, so that the provided opinion scores can be used during the next training process, and thus informing the AIO on how to correctly rate the quality of these images previously considered as critical.

The observer o was invited to score the quality of the newly selected images in I_o^3. These images along with the collected opinion scores OS_o^3 were then used to perform a second transfer learning step. This time we started from the AIO_o^2 and updated its weights using as ground truth the newly gathered opinion scores OS_o^3 for the images in I_o^3, and obtained the AIO_o^3.

We adopted the same procedure to train the AIO_o^4 starting from the AIO_o^3. Thus, the images in I_o^4 were selected based on the variances of the predictions of the AIO_o^3. The selected images were evaluated by the observer o and a third transfer learning step was conducted starting from the AIO_o^3 to obtain the AIO_o^4 that we considered as the final model of the observer o. For simplicity sake, from now on the final AIO of each observer $o \in \mathcal{O}$ will be denoted by AIO_o.

III. Results

We used the performance of the five real observers considered in this work to benchmark that of their trained AIOs. In particular, we exploited the repeated evaluation of the images in I^{s1s2} to estimate the accuracy of a real observer interpreting the repetition as a prediction of their own ratings. The Table I shows the estimated accuracy for each observer. For instance, when rating for the second time the images in I^{s1s2}, the observer #1 was able to predict/repeat their first opinion score on 58% of the images.

TABLE I
ACCURACY OF REAL OBSERVERS TO PREDICT THEIR OPINIONS

Obs 1	Obs 2	Obs3	Obs 4	Obs 5
58%	57%	73%	62%	47%

TABLE II
ACCURACY OF AIOS TO PREDICT THE OPINIONS OF REAL OBSERVERS

	Obs 1	Obs 2	Obs3	Obs 4	Obs 5
AIO_1	**63**%	59%	43%	39%	50%
AIO_2	59%	**67**%	39%	36%	52%
AIO_3	39%	41%	**57**%	**58**%	57%
AIO_4	45%	42%	52%	51%	59%
AIO_5	36%	35%	55%	53%	**60**%
JPEGResNet50	42%	56%	39%	32%	43%

Since each AIO is trained to mimic their related observer, we can consider the percentages in Table I as a kind of reference or benchmark accuracy for the AIO when predicting the opinion scores of the observer it is mimicking. In other words, the AIO of the observer $o \in \mathcal{O}$ can be considered effective if it can predict the opinion scores of the observer o with an accuracy that is close to or higher than the performance of the observer o in predicting their own opinion scores.

Table II shows the accuracy of each AIO when predicting the opinion scores of the real observers on the test set. The accuracy is the fraction of images for which the predicted opinion score by the AIO is equal to the opinion score provided by the real observer. For instance, the AIO_1 correctly predicted 63% of the opinion scores of the observer it is mimicking, i.e. the observer #1, and 43% of the opinion scores of the observer #3.

Looking at Table I and Table II, it can be noticed that 3 AIOs (AIO_1, AIO_2 and AIO_5) out of 5 predicted the opinion scores of the related observers with an accuracy that is higher than their expected reference accuracy in Table I. The AIO_3 and AIO_4 that showed lower accuracy than the related observers however guaranteed an accuracy greater than 50%. Thus, their performance is still comparable to that of a real observer. In fact, the observer #5, as it can be seen from Table I, correctly predicted the first rating only in 47% of cases.

It is also very interesting to notice from Table II that, except for the AIO_4, each AIO predicted the ratings of the observer it is mimicking better than any other AIO on the test set. This suggests that, similar to real observers, the AIOs are different, and thus, they did not learn only generic features valid for any subject, instead they probably can mimic some individual characteristics of the scoring behavior of the real observer they are modeling.

A closer look at the gathered ratings revealed that the observer #4 never chose 5 as an opinion score when rating the images used in session 2 and 3. This might explain the fact that their AIO did not predict their opinion scores better than all the other AIOs on the test set.

The last row of Table II shows the performance of the JPEGResNet50 in predicting the five real observers. It can be noticed that, at the end of the proposed training procedure, the performance of each AIO in predicting the related observer on a set of images never seen during the training strictly overcomes that of the JPEGResNet50 that was used as starting point for the training of the AIOs. For instance, for the observer #1, the proposed training process has allowed to achieve an accuracy of 63% starting from 42% offered by the JPEGResNet50. This further highlights the effectiveness of our training procedure.

IV. Conclusions

In this work we described an approach that iterates between human voting, transfer learning and inference procedures to train the deep CNN mimicking the quality perception of an individual subject. The current results are very promising, since in general each trained AIO can mimic the related real observer with an accuracy that is comparable to the one that a real subject would achieve in a repeated evaluation of a given set of images. A non-individualized prediction method would not allow for such a result. The complexity and the convergence of our approach will be investigated in a future contribution.

REFERENCES

[1] T. Hoßfeld, R. Schatz, and S. Egger, "SOS: The MOS is not enough!" in *Third International Workshop on Quality of Multimedia Experience (QoMEX)*. Mechelen, Belgium: IEEE, Sep 2011, pp. 131–136.

[2] M. Seufert, "Fundamental advantages of considering quality of experience distributions over mean opinion scores," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, June 2019, pp. 1–6.

[3] L. Fotio Tiotsop, E. Masala, A. Aldahdooh, G. V. Wallendael, and M. Barkowsky, "Computing quality-of-experience ranges for video quality estimation," in *Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. Berlin, Germany: IEEE, Jun 2019, pp. 1–3.

[4] D. Varga, D. Saupe, and T. Szirányi, "Deeprn: A content preserving deep architecture for blind image quality assessment," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. San Diego, CA, USA: IEEE, 2018, pp. 1–6.

[5] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

[6] Y. Gao, X. Min, W. Zhu, X.-P. Zhang, and G. Zhai, "Image quality score distribution prediction via alpha stable model," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2022.

[7] J. Korhonen, "Assessing personally perceived image quality via image features and collaborative filtering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8169–8177.

[8] L. F. Tiotsop, T. Mizdos, M. Uhrina, P. Pocta, M. Barkowsky, and E. Masala, "Predicting single observer's votes from objective measures using neural networks," in *Proceedings of Human Vision and Electronic Imaging conference (HVEI)*, Jan 2020.

[9] L. F. Tiotsop, T. Mizdos, M. Barkowsky, P. Pocta, A. Servetti, and E. Masala, "Mimicking individual media quality perception with neural network based artificial observers," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 18, no. 1, 2022.

[10] L. F. Tiotsop, A. Servetti, M. Barkowsky, P. Pocta, T. Mizdos, G. Van Wallendael, and E. Masala, "Predicting individual quality ratings of compressed images through deep cnns-based artificial observers," *Signal Processing: Image Communication*, vol. 112, p. 116917, 2023.

[11] L. Janowski and M. Pinson, "The accuracy of subjects in a quality experiment: A theoretical subject model," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2210–2224, 2015.

[12] L. F. Tiotsop, A. Servetti, M. Barkowsky, and E. Masala, "Regularized maximum likelihood estimation of the subjective quality from noisy individual ratings," in *2022 14th International Conference on Quality of Multimedia Experience (QoMEX)*, 2022, pp. 1–4.

[13] ITU-T Rec. P.910, "Subjective video quality assessment methods for multimedia applications," Apr. 2008.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.