Worker Qualifications for Image-Aesthetic-Assessment Tasks in Crowdsourcing

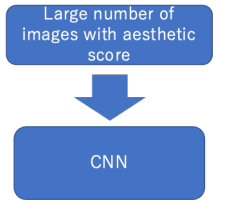
Yudai Kato Marie Katsturai
Doshisha University

Keishi Tajima
Kyoto University

HCOMP2022 Work-in-Progress

Background

- Automatic assessment of images' aesthetic quality has been actively studied
- CNN have yielded significant performance improvements over conventional visual features
- Supervised learning approaches require a large amount of data.



CNN to predict aesthetic score

We should consider how to efficiently collect aesthetic scores that are carefully assessed.

Methods used in prior studies

preprocessing method
- Qualification test
- Gold injection

postprocessing method
- Outlier detection
- Answer aggregation (e.g., majority voting)

Problems when applied to subjective tasks

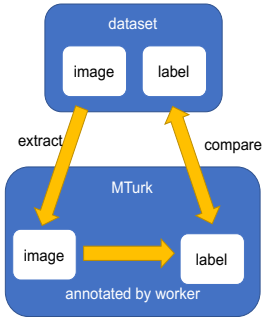
- Large variation in answers from worker to worker.
- Difficult to prepare gold standard.

Objective

In this study, we explore a strategy for setting worker eligibility requirements to stabilize the quality of the results

Task design

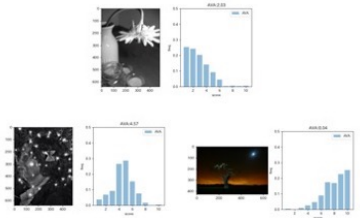
1. Extract images from an existing dataset.
2. Order tasks under various qualification conditions.
3. Compare the gold standard from the existing dataset and labels obtained in MTurk



What conditions can produce results close to the gold standard?

AVA dataset

- Aesthetic Visual analysis dataset
- 255,000 images with an average of 210 annotations



Qualification conditions

- Five qualification conditions were set.
- Expected: condition 4 achieves the best results.

Condition	Approval rate	The number of approved tasks	Remarks
1	over 95%	over 100	common in previous studies
2	under 95%	None	Bad worker
3	None	under 100	New worker
4	over 98%	over 5,000	Very strict
5	None	None	No qualifications

Result

Condition	Average		Variance		Remarks
	correlation	MAE	correlation	MAE	
1	0.32	1.06	-0.06	2.38	common in previous studies
2	0.29	1.39	0.12	2.46	Bad worker
3	0.25	1.18	0.03	3.29	New worker
4	**0.43**	**0.81**	-0.02	3.17	Very strict
5	0.29	1.55	0.12	2.39	No qualifications

- As expected, the average scores produced by workers that satisfied Condition 4 were the closest to the AVA dataset.
- Condition 1 was a relatively severe restriction; but interestingly, it did not lead to any significant differences from the other conditions, except for Condition 4
- we observed almost no linear relationship in the variances between the MTurk ratings and the AVA ratings for the test images

Conclusion

- It is effective limiting eligibility to only those workers who had been approved for thousand tasks and had a high approval rate of over 98%
- Standard criterion, which was often used in related studies, was insufficient for the target subjective task