Reliably combining quality indicators

Adriaan Barri, Ann Dooms, Peter Schelkens

Is machine learning (ML) suitable for objective quality assessment?

In recent years, machine learning (ML) has gained increased attention as a technique to improve the accuracy of objective quality measures. By incorporating ML, objective quality measures can mimic mechanisms of the human visual system (HVS) that otherwise had to be modeled explicitly. As a consequence, ML-based quality measures require fewer computations and are less affected by our limited knowledge of the HVS. On the downside, they yield less transparent quality predictions, because the ML responses are often difficult to interpret. The absence of interpretability may disguise serious vulnerabilities, such as consistency violations, unstable predictions in the high quality range, and severe false orderings. Our recently developed Locally Adaptive Fusion (LAF) method addresses these issues by imposing strict regulations on the ML behavior. This article analyzes the prediction performance of LAF by traditional validation techniques and by complementary stress tests on an unannotated image database. These tests explain the benefits of LAF and illustrate the importance of a thorough validation.

Objective quality measures based on machine learning (ML) require fewer computations and are less affected by inaccuracies in the HVS models. But they may also yield less transparent quality predictions when the ML responses are difficult to interpret. The absence of interpretability may disguise serious vulnerabilities in the design of the objective quality measure.

Locally Adaptive Fusion (LAF) when transparency is important

In contrast to traditional ML methods, Locally Adaptive Fusion (LAF) is specifically designed for objective quality assessment. The LAF method predicts quality in two steps. Firstly, the signal is subjected to multiple fusion units. Each fusion unit is a fixed weighted sum of predetermined quality indicators, which are meant for specific content or distortion types. Secondly, the calculated fusion unit values are combined through adaptive weighting, using a second set of weights that change depending on the received signal. This nonlinear response allows LAF to better mimic complex HVS mechanisms.

To ensure interpretability, the weights of LAF are directly related to the quality indicators. The strict regulations imposed by LAF come with three additional advantages: reproducibility, consistency, and computational scalability.

The behavior of LAF is strictly regulated and much easier to interpret in comparison with other nonlinear ML methods (e.g. neural networks). By design, the weights of LAF are directly related to the quality indicators. As a result, the influence of the quality indicators on the quality prediction of the received signal can be visualized (Figure 1).

The imposed regulations of LAF come with three additional advantages: reproducibility, consistency, and computational scalability. Firstly, the training phase of LAF does not require a random initialization. Unlike neural networks, re-training LAF on the same data will always produce the same weights. Secondly, the LAF response is always consistent with the input quality indicators to avoid overfitting. Thirdly, LAF can be easily configured to find the optimal trade-off between computational complexity and prediction accuracy.

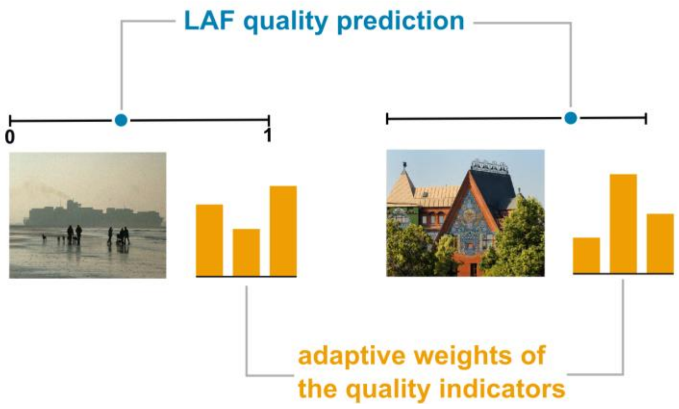

Figure 1. The weights used by LAF are directly related to the quality indicators. These weights change depending on the content and distortion type. In the above illustration, LAF predicts the quality of two still images by adaptively weighting three input quality indicators.

Validation methods adjusted to ML-based quality measures

This section compares the reliability of LAF with a one-layer feed forward neural network (FFNN). To avoid the curse of dimensionality, we limited the ML input to three simple quality indicators for still images, one of the no-reference type and the two others of the reduced-reference type. The selected quality indicators respectively measure blocking artifacts, spatial information loss, and contrast similarity. The performance of ML-based quality measures is typically tested on multiple annotated databases. However, these tests revealed no significant differences between LAF and FFNN (Table 1). For a more thorough comparison, we needed complementary stress tests on an unannotated database.

Table 1. The validation tests on LIVE, CSIQ, and TID revealed no significant performance differences. More details are in (Barri A. et al., 2014).

Tests on annotated databases (Pearson correlation)	LAF	FFNN
Repeated cross-validation on the LIVE database	0.96	**0.965**
Database independence *Training set:* LIVE – *Test set:* CSIQ	**0.967**	0.959
Robustness for unknown distortions *Training set:* LIVE – *Test set:* TID	**0.822**	0.790

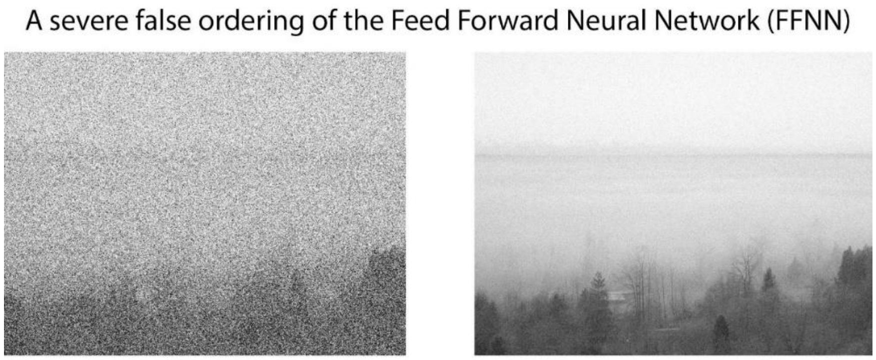
A severe false ordering of the Feed Forward Neural Network (FFNN)

Figure 2. Even when ML-based quality measures obtain high correlation values on annotated quality assessment databases, they may produce severe false orderings on larger, unannotated test databases. In the above illustration, the FFNN prefers the quality of the left image. Such severe false orderings confirm the importance of complementary stress tests during validation.

We evaluated the ML-based quality measures on an unannotated stress test database containing 650 reference and 26,000 distorted images. We acquired three new insights:

- **Traditional ML is often inconsistent.** Given two signals, suppose all input quality indicators systematically give a higher rate to the first signal. Then we proved the LAF method will always agree with the preference of the indicators. Traditional ML tends to ignore the indicators to better fit the training data. For FFNN, we discovered more than 100,000 of these consistency violations.
- **Traditional ML is unstable in the high quality range.** For the quality predictions of barely distorted images, LAF will optimize the weights of the indicators to the high quality range. The FFNN will still employ the no-reference blocking indicator, but this yields unstable quality predictions due to the low visibility of the artifacts.
- **Traditional ML may produce severe false orderings.** The quality predictions should decrease when the distortion rate is gradually increased. On the stress test database, LAF produces fewer false orderings than FFNN (6 vs. 119). Moreover, the false orderings of FFNN were often very severe (Figure 2).

Adriaan Barri is a member of the iMinds research center of Flanders, and the department of electronics and informatics (ETRO) at the Vrije Universiteit Brussel, Belgium. He holds a PhD bursary from the agency for Innovation by Science and Technology (IWT). His research focuses on machine learning and quality assessment.

Ann Dooms is a member of iMinds and holds a professorship at ETRO, Vrije Universiteit Brussel. Dooms leads a research team in Multimedia Forensics, which studies the lifecycle of a multimedia item to answer forensic questions ranging from authenticity and traitor tracing over perceptual quality and compressed sensing to digital painting analysis.

Peter Schelkens is research director at iMinds and holds a professorship at ETRO, Vrije Universiteit Brussel,. In 2010, he joined the board of councillors of the Interuniversity Microelectronics Institute (IMEC), Belgium. In 2013, he obtained an ERC Consolidator Grant focusing on digital holography. He is an elected member of the IEEE Technical Committees IVMSP and MMSP, and is participating in the JPEG and MPEG standardization activities.

What have we learned?

Not all vulnerabilities of ML-based quality measures can be detected by traditional validation methods. Most vulnerabilities can be reduced or even avoided when more interpretable ML methods are used, such as LAF. We firmly believe that LAF is more reliable than other ML solutions for real-life applications. More information can be found in the referenced paper and at www.locally-adaptive-fusion.com.

References

Barri A., Dooms A., Jansen B., Schelkens P. (2014),
"A Locally Adaptive System for the Fusion of Objective Quality Measures,"
IEEE Transactions on Image Processing, Vol. 23, No. 6, pp. 2446-2458.