No-Reference Face Image Quality Assessment

이어드림 24조(3p)

노치현, 마준호, 임건희
발표자 : 임건희

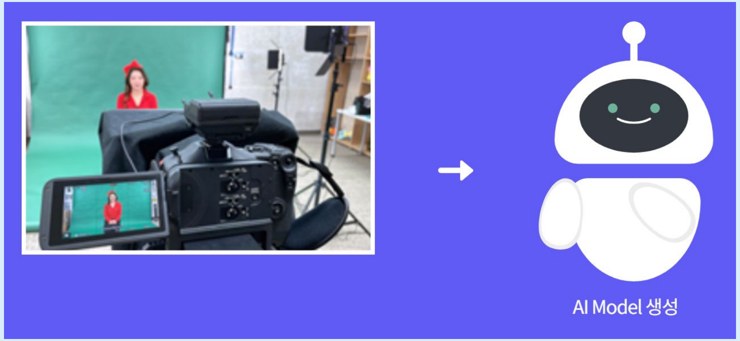
Contents

- 기획
- 진행 과정
- 결과 및 해석
- Todos

기획

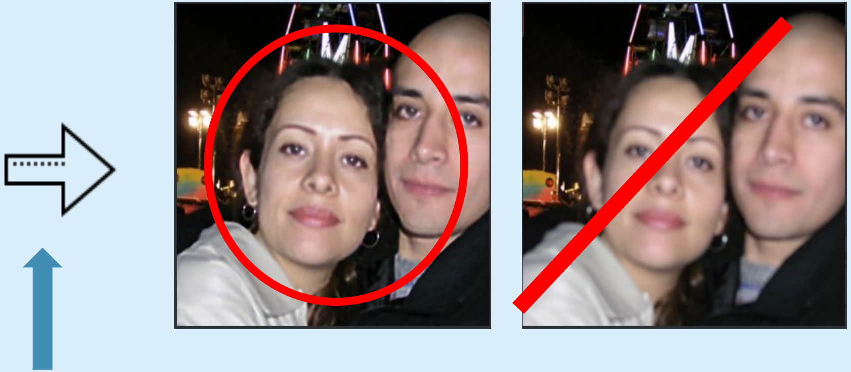
- 사업모델 및 기획
- 문제정의 및 목표
- KADID-10k Dataset
- Flow - Chart for Training
- Flow - Chart for Test

사업 모델 및 기획



AI PARK 비즈니스 모델

AI 앵커 '제나' 아바타 제작



이미지 선별 위한 작업 필요
(이미지 선별 자동화)

문제 정의 및 목표

<table>
<tr><th colspan="3">이미지 품질 평가 방법</th></tr>
<tr><td>subjective methods</td><td colspan="2">objective methods</td></tr>
<tr><td rowspan="3">MOS
DMOS
.
.
.</td><td colspan="2">FR-IQA</td></tr>
<tr><td colspan="2">RR-IQA</td></tr>
<tr><td colspan="2">NR-IQA</td></tr>
<tr><td></td><td>OU</td><td>OA</td></tr>
</table>

MOS score 와
상관성이 높은 NR-IQA를 찾자!
(OU 방법 활용)

문제 정의 및 목표

문제점 1

▪MOS 값을 제공하는 얼굴 이미지 데이터셋이 없음

1) Natural image의 mos를 제공하는 IQA database를 활용하자
→ 인물 이미지에서 안면 이미지만을 crop하면 natural image의 mos기준과 큰 차이점이 없다.

문제점 2

▪단일 Metric의 사용은 성능이 부족함

1) 여러 Metric 들을 fusion 하자
→ 영상 데이터의 경우 넷플릭스의 VMAF 가 벤치마크에서 최상권을 기록
→ FR-IQA로 동일한 시도를 한 논문 존재 (Full-Reference Image Quality Assessment Based on an Optimal Linear Combination of Quality Measures Selected by Simulated Annealing)

KADID-10k Dataset

The 81 pristine images in KADID-10k

KADID-10k

▪81개의 reference image

▪각 Distorted image별 Dmos 제공

▪총 25개 유형의 distortion 제공

▪3개 유형의 Blur distortion
(Gaussian, Lens, Motion)

KADID-10k Dataset

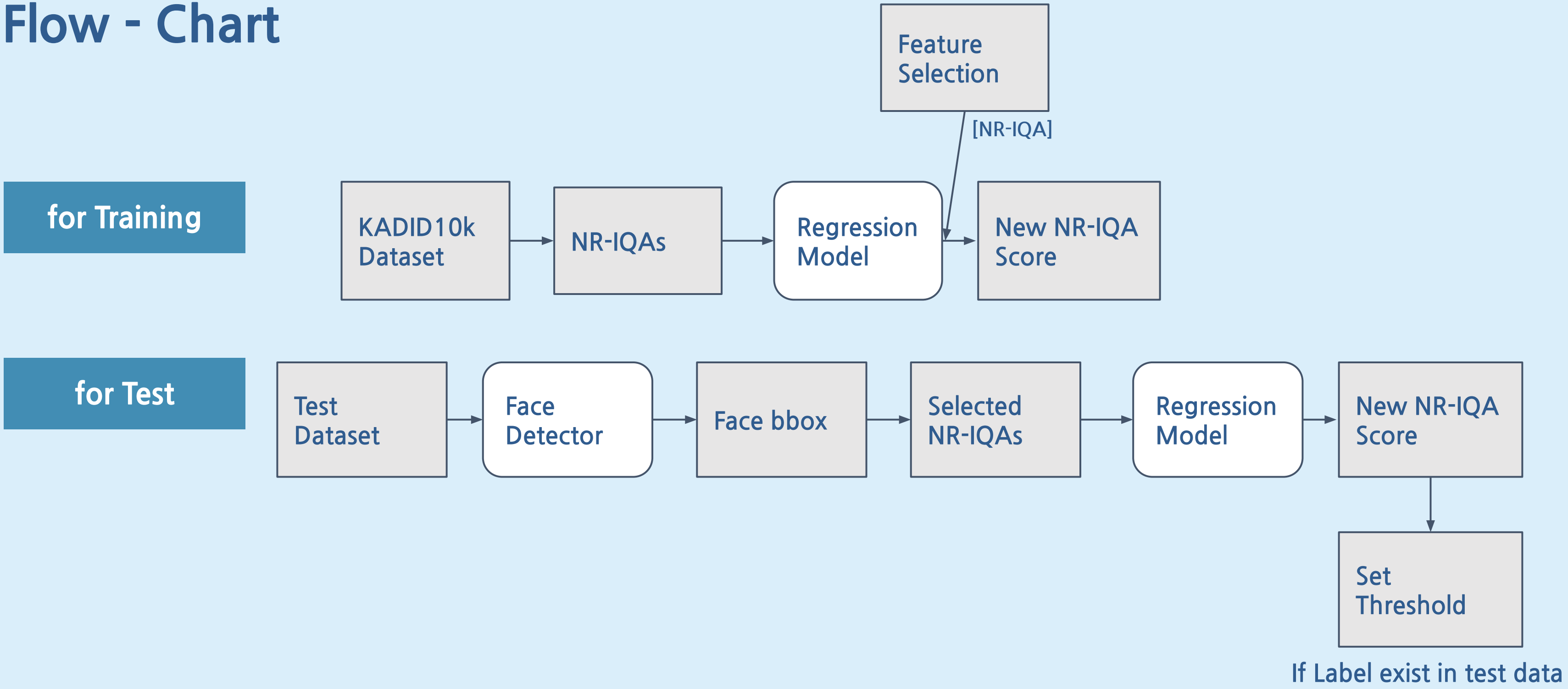
The 81 pristine images in KADID-10k

KADID-10k

- ▪81개의 reference image
- ▪각 Distorted image별 Dmos 제공
- ▪총 25개 유형의 distortion 제공
- ▪3개 유형의 Blur distortion (Gaussian, Lens, Motion)

Distortion은 blur 적용

Flow - Chart


Feature Selection
[NR-IQA]
for Training
KADID10k Dataset
NR-IQAs
Regression Model
New NR-IQA Score
for Test
Test Dataset
Face Detector
Face bbox
Selected NR-IQAs
Regression Model
New NR-IQA Score
Set Threshold
If Label exist in test data

진행 과정

- NR Image Quality Assessment
- 데이터 탐색
- Feature Selection
- 모델 학습
- Test NR-IQAs in KADID Dataset

NR Image Quality Assessment

- IL-NIQE
- NIQE
- CNN
- BRISQUE
- CPBD
- WaDIQaM / DIQaM
- HYPER

Metric	특징
IL-NIQE	• First, we enriched the feature set by introducing three new types of quality-aware NSS features • Second, instead of using a single global MVG model to describe the test image
CNN	• 먼저 a contrast normalization 를 수행하고, 겹치지 않는 패치를 샘플링을 함
BRISQUE	• BRISQUE - Blind/Referenceless Image Spatial QUality Evaluator. • 계산 복잡성이 매우 낮아 실시간 애플리케이션에 매우 적합.
CPBD	• This work presents a perceptual-based no-reference objective image sharpness metric
NIQE	• NIQE - Natural Image Quality Evaluator • 자연이미지에서 관찰되는 통계적인 정형화로부터 측정가능한 편차만 사용
WaDIQaM /DIQaM	• Weighted Average Deep Image QuAlity Measure for NR IQA (WaDIQaM-NR) • Deep Image QuAlity Measure for NR IQA (DIQaM-NR)

NR score / KADID-10k MOS 성능 확인

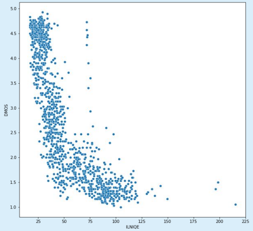
Blur 처리된 이미지만으로 진행

scatter plot
correlation

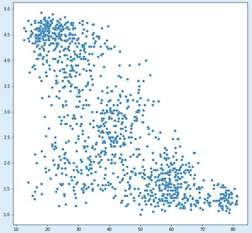
ROCC
LCC

NR Metric Scatter plot

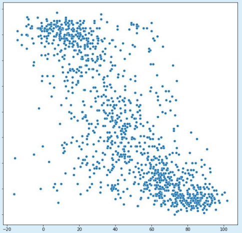
IL NIQE



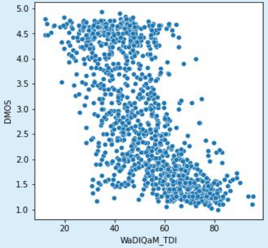
CNN



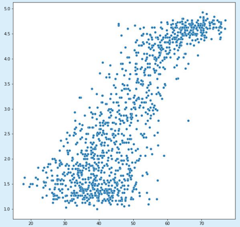
BRISQUE



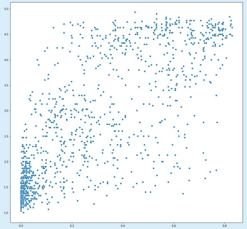
WaDIQaM



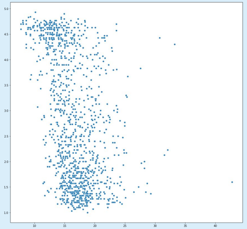
HYPER



CPBD



NIQE

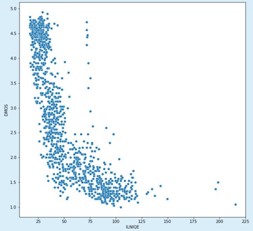


특징 있음

특징 없음

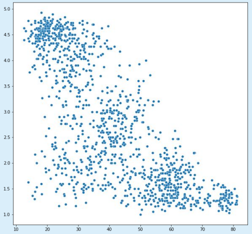
NR Metric Correlation

IL NIQE



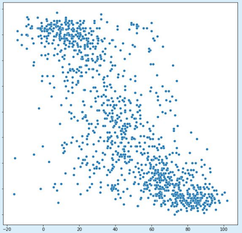
0.788

CNN



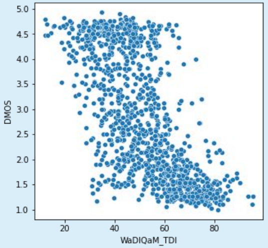
0.752

BRISQUE



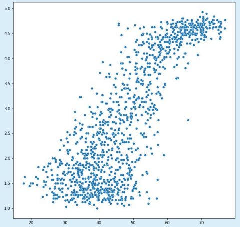
0.780

WaDIQaM



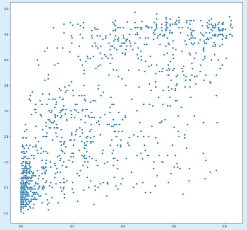
0.695

HYPER



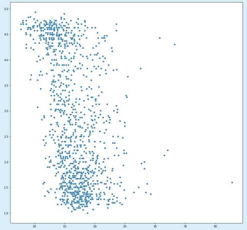
0.831

CPBD



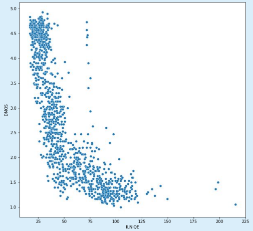
0.766

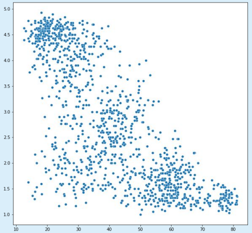
NIQE

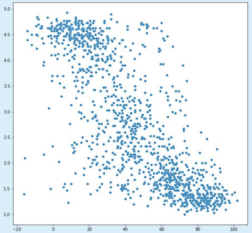


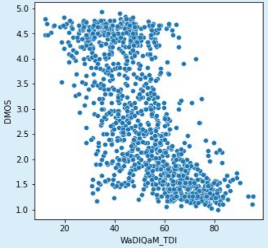
0.421

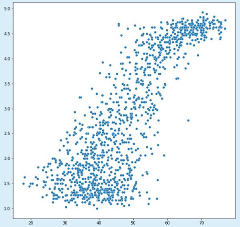
NR Metric ROCC / LCC

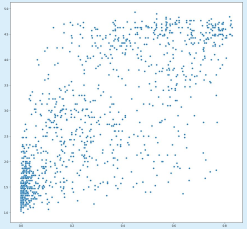


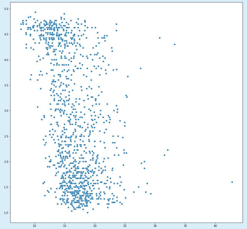








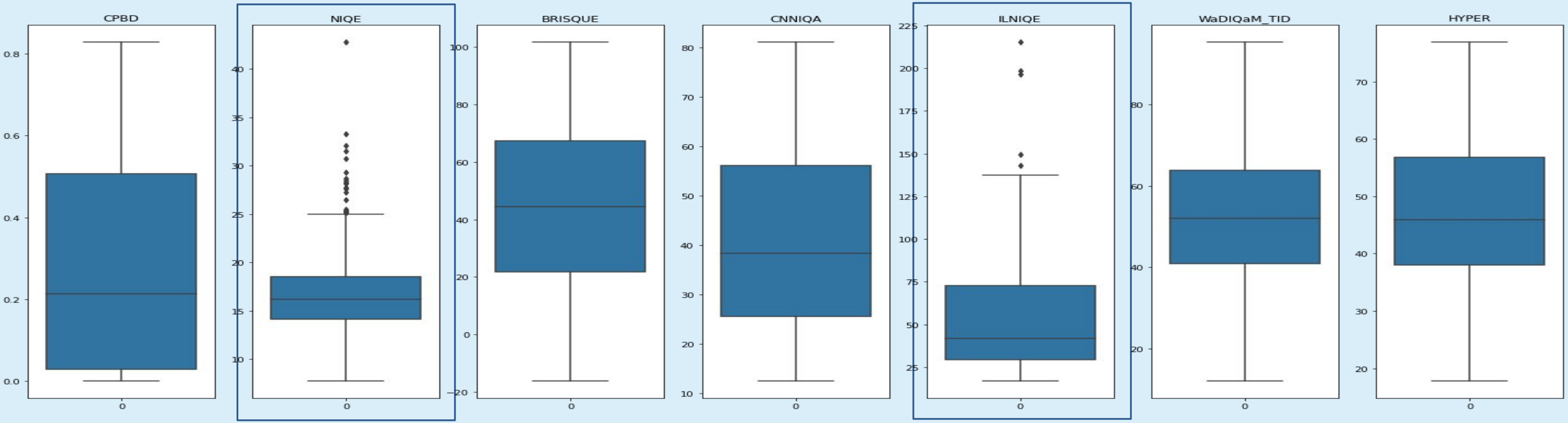




IL NIQE	CNN	BRISQUE	WaDIQaM	HYPER	CPBD	NIQE
0.869	0.764	0.795	0.716	0.784	0.790	0.460
0.770	0.752	0.780	0.696	0.831	0.766	0.421

NR Metric 데이터탐색(이상치)

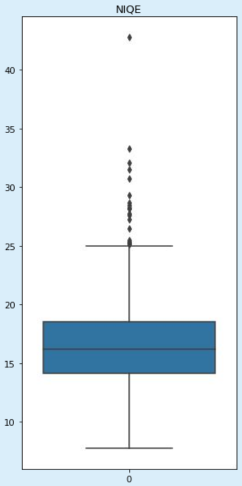
CPBD NIQE BRISQUE CNN IL NIQE WaDIQaM HYPER

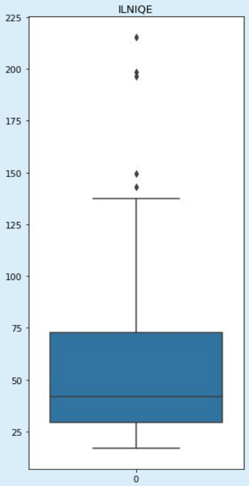


NR Metric 데이터탐색(이상치)

NIQE

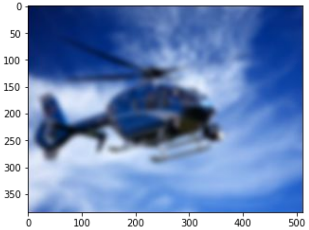
IL NIQE





ILNIQE > 140 5건

5건 ILNIQE mean score : 180 , mean mos : 1.3
blur 5단계 ILNIQE mean score : 103 , mean mos : 1.3
→ 이상치 데이터 모두 정상으로 판단
→ 5건 모두 blur 레벨 4,5에서 발생 (ILNIQE 신뢰가능)

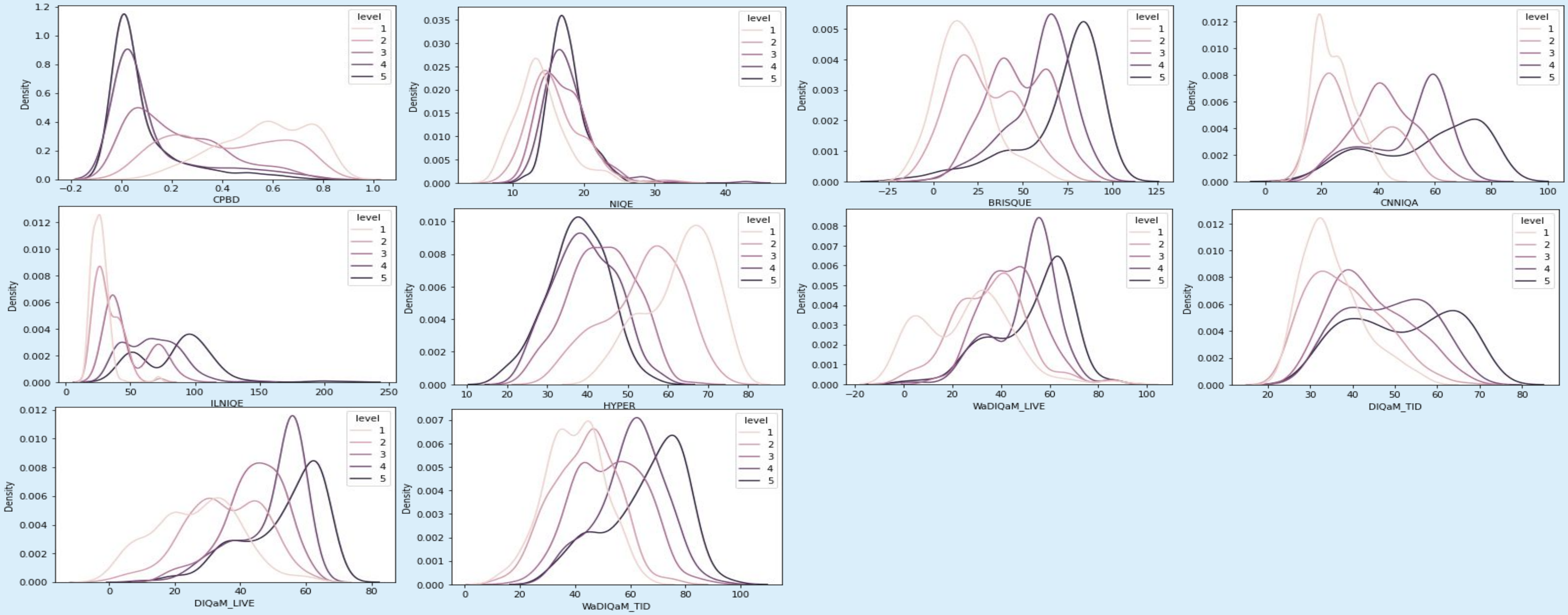


NIQE > 25 20건

20건 NIQE mean score : 28.8 mean mos : 2.3
blur 5단계 NIQE mean score : 17.7 , mean mos : 1.3
→ 20건 중 12건이 이미지가 좋으나 NIQE 평가는 나쁨
→ INQE 신뢰불가능 → NR Metric 에서 제외

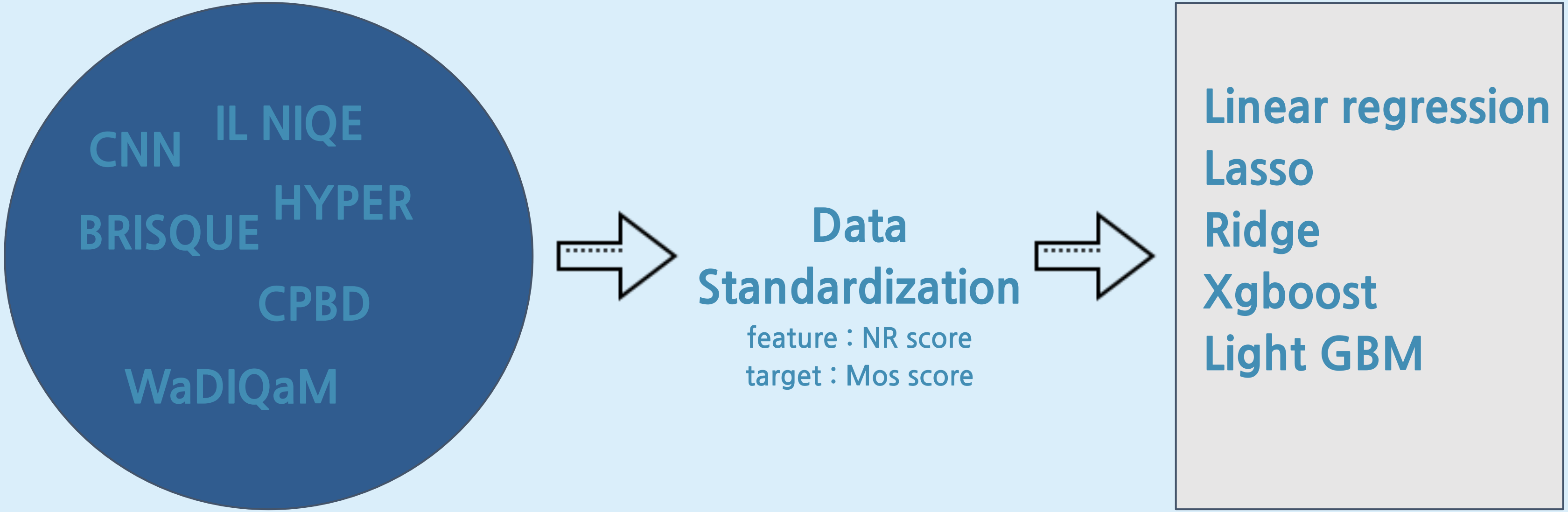
NR Metric 데이터탐색

KADID-10k score Distribution for each NR-IQAs



NR Metric Selection

Model fitting



CNN

IL NIQE

BRISQUE

HYPER

CPBD

WaDIQaM

Data Standardization

feature : NR score

target : Mos score

Linear regression

Lasso

Ridge

Xgboost

Light GBM

NR Metric ROCC / LCC with MOS

10개의 NR Metric 을 활용한 ROCC / LCC

Metric	ROCC	LCC
ILNIQE	0.880	0.788
CNN	0.764	0.752
BRISQUE	0.795	0.780
CPBD	0.790	0.766
NIQE	0.460	0.421

Metric	ROCC	LCC
WaDIQaM_LIVE	0.595	-0.576
DIQaM_LIVE	0.711	0.704
WaDIQaM_TID	0.715	0.695
DIQaM_TID	0.560	0.552
HYPER	0.784	0.831

Model fitting

Train : Val = 3 : 1, default_hyperparameter, no_scaling

	Train set	Val set		
Models	MSE	MSE	ROCC	LCC
Linear Reg	0.220	0.293	0.920	0.910
Lasso	0.240	0.303	0.918	0.910
Ridge	0.220	0.293	0.920	0.910
XGB	0.052	0.173	0.929	0.944
LGBM	0.011	0.181	0.922	0.941

How to choose NR Metric

D-test, L-test, t-Test

D-test(pristine/distorted image discrimination test)

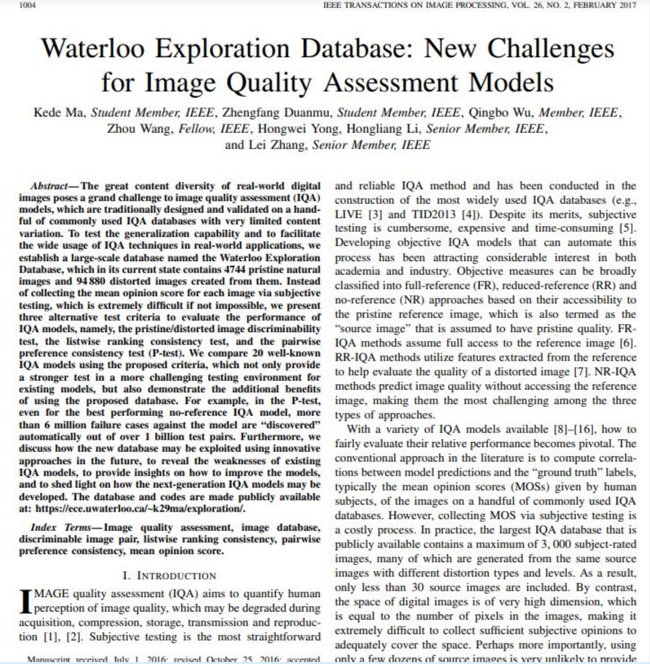
IQA 모델이 왜곡 이미지로 부터 원본 이미지를 잘 분리해 내는지를 시험

L-test(listwise ranking consistency test)

IQA 모델이 같은 콘텐츠 및 같은 왜곡 유형이지만 다른 정도로 왜곡된 이미지들의 순위를 잘 매길 수 있는지를 시험

P-test(pairwise preference consistency test)

품질차이를 느낄 수 있는 이미지쌍(quality-discriminable image pair, DIP)을 제시했을 때 IQA모델이 더 나은 품질의 것을 잘 선택 할 수 있는지를 시험



1004 IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 26, NO. 2, FEBRUARY 2017

Waterloo Exploration Database: New Challenges for Image Quality Assessment Models

Kede Ma, *Student Member, IEEE*, Zhengfang Duanmu, *Student Member, IEEE*, Qingbo Wu, *Member, IEEE*, Zhou Wang, *Fellow, IEEE*, Hongwei Yong, Hongliang Li, *Senior Member, IEEE*, and Lei Zhang, *Senior Member, IEEE*

***Abstract*—The great content diversity of real-world digital images poses a grand challenge to image quality assessment (IQA) models, which are traditionally designed and validated on a handful of commonly used IQA databases with very limited content variation. To test the generalization capability and to facilitate the wide usage of IQA techniques in real-world applications, we establish a large-scale database named the Waterloo Exploration Database, which in its current state contains 4744 pristine natural images and 94 880 distorted images created from them. Instead of collecting the mean opinion score for each image via subjective testing, which is extremely difficult if not impossible, we present three alternative test criteria to evaluate the performance of IQA models, namely, the pristine/distorted image discriminability test, the listwise ranking consistency test, and the pairwise preference consistency test (P-test). We compare 20 well-known IQA models using the proposed criteria, which not only provide a stronger test in a more challenging testing environment for existing models, but also demonstrate the additional benefits of using the proposed database. For example, in the P-test, even for the best performing no-reference IQA model, more than 6 million failure cases against the model are "discovered" automatically out of over 1 billion test pairs. Furthermore, we discuss how the new database may be exploited using innovative approaches in the future, to reveal the weaknesses of existing IQA models, to provide insights on how to improve the models, and to shed light on how the next-generation IQA models may be developed. The database and codes are made publicly available at: https://ece.uwaterloo.ca/~k29ma/exploration/.**

***Index Terms*—Image quality assessment, image database, discriminable image pair, listwise ranking consistency, pairwise preference consistency, mean opinion score.**

I. Introduction

IMAGE quality assessment (IQA) aims to quantify human perception of image quality, which may be degraded during acquisition, compression, storage, transmission and reproduction [1], [2]. Subjective testing is the most straightforward and reliable IQA method and has been conducted in the construction of the most widely used IQA databases (e.g., LIVE [3] and TID2013 [4]). Despite its merits, subjective testing is cumbersome, expensive and time-consuming [5]. Developing objective IQA models that can automate this process has been attracting considerable interest in both academia and industry. Objective measures can be broadly classified into full-reference (FR), reduced-reference (RR) and no-reference (NR) approaches based on their accessibility to the pristine reference image, which is also termed as the "source image" that is assumed to have pristine quality. FR-IQA methods assume full access to the reference image [6]. RR-IQA methods utilize features extracted from the reference to help evaluate the quality of a distorted image [7]. NR-IQA methods predict image quality without accessing the reference image, making them the most challenging among the three types of approaches.

With a variety of IQA models available [8]–[16], how to fairly evaluate their relative performance becomes pivotal. The conventional approach in the literature is to compute correlations between model predictions and the "ground truth" labels, typically the mean opinion scores (MOSs) given by human subjects, of the images on a handful of commonly used IQA databases. However, collecting MOS via subjective testing is a costly process. In practice, the largest IQA database that is publicly available contains a maximum of 3, 000 subject-rated images, many of which are generated from the same source images with different distortion types and levels. As a result, only less than 30 source images are included. By contrast, the space of digital images is of very high dimension, which is equal to the number of pixels in the images, making it extremely difficult to collect sufficient subjective opinions to adequately cover the space. Perhaps more importantly, using only a few dozens of source images is very unlikely to provide

Manuscript received July 1, 2016; revised October 25, 2016; accepted

How to choose NR Metric

D-test, L-test, t-Test

KADID-10k dataset

D-test(pristine/distorted image discrimination test)

IQA 모델이 왜곡 이미지로 부터 원본 이미지를 잘 분리해 내는지를 시험

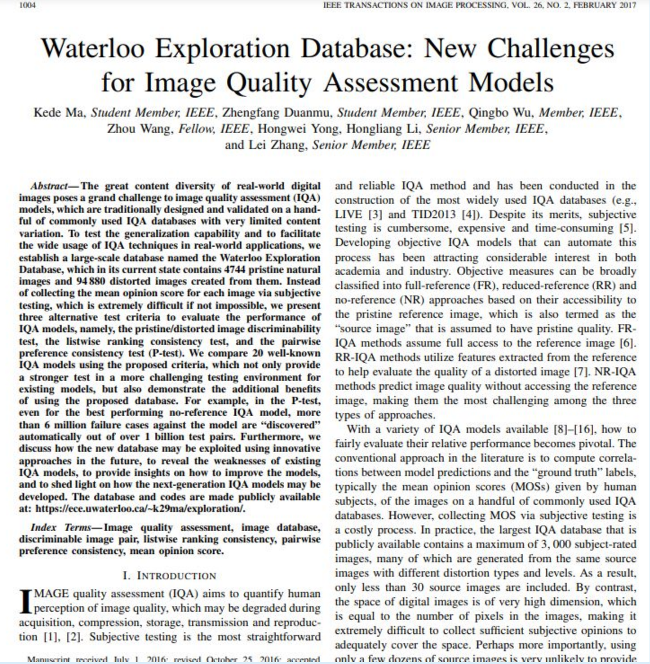
L-test(listwise ranking consistency test)

IQA 모델이 같은 콘텐츠 및 같은 왜곡 유형이지만 다른 정도로 왜곡된
이미지들의 순위를 잘 매길 수 있는지를 시험

P-test(pairwise preference consistency test)

품질차이를 느낄 수 있는 이미지쌍(quality-discriminable image pair, DIP)을
제시했을 때 IQA모델이 더 나은 품질의 것을 잘 선택 할 수 있는지를 시험

Test dataset



1004 IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 26, NO. 2, FEBRUARY 2017

Waterloo Exploration Database: New Challenges for Image Quality Assessment Models

Kede Ma, *Student Member, IEEE*, Zhengfang Duanmu, *Student Member, IEEE*, Qingbo Wu, *Member, IEEE*, Zhou Wang, *Fellow, IEEE*, Hongwei Yong, Hongliang Li, *Senior Member, IEEE*, and Lei Zhang, *Senior Member, IEEE*

***Abstract*—The great content diversity of real-world digital images poses a grand challenge to image quality assessment (IQA) models, which are traditionally designed and validated on a handful of commonly used IQA databases with very limited content variation. To test the generalization capability and to facilitate the wide usage of IQA techniques in real-world applications, we establish a large-scale database named the Waterloo Exploration Database, which in its current state contains 4744 pristine natural images and 94 880 distorted images created from them. Instead of collecting the mean opinion score for each image via subjective testing, which is extremely difficult if not impossible, we present three alternative test criteria to evaluate the performance of IQA models, namely, the pristine/distorted image discriminability test, the listwise ranking consistency test, and the pairwise preference consistency test (P-test). We compare 20 well-known IQA models using the proposed criteria, which not only provide a stronger test in a more challenging testing environment for existing models, but also demonstrate the additional benefits of using the proposed database. For example, in the P-test, even for the best performing no-reference IQA model, more than 6 million failure cases against the model are "discovered" automatically out of over 1 billion test pairs. Furthermore, we discuss how the new database may be exploited using innovative approaches in the future, to reveal the weaknesses of existing IQA models, to provide insights on how to improve the models, and to shed light on how the next-generation IQA models may be developed. The database and codes are made publicly available at: https://ece.uwaterloo.ca/~k29ma/exploration/.**

***Index Terms*—Image quality assessment, image database, discriminable image pair, listwise ranking consistency, pairwise preference consistency, mean opinion score.**

I. INTRODUCTION

IMAGE quality assessment (IQA) aims to quantify human perception of image quality, which may be degraded during acquisition, compression, storage, transmission and reproduction [1], [2]. Subjective testing is the most straightforward and reliable IQA method and has been conducted in the construction of the most widely used IQA databases (e.g., LIVE [3] and TID2013 [4]). Despite its merits, subjective testing is cumbersome, expensive and time-consuming [5]. Developing objective IQA models that can automate this process has been attracting considerable interest in both academia and industry. Objective measures can be broadly classified into full-reference (FR), reduced-reference (RR) and no-reference (NR) approaches based on their accessibility to the pristine reference image, which is also termed as the "source image" that is assumed to have pristine quality. FR-IQA methods assume full access to the reference image [6]. RR-IQA methods utilize features extracted from the reference to help evaluate the quality of a distorted image [7]. NR-IQA methods predict image quality without accessing the reference image, making them the most challenging among the three types of approaches.

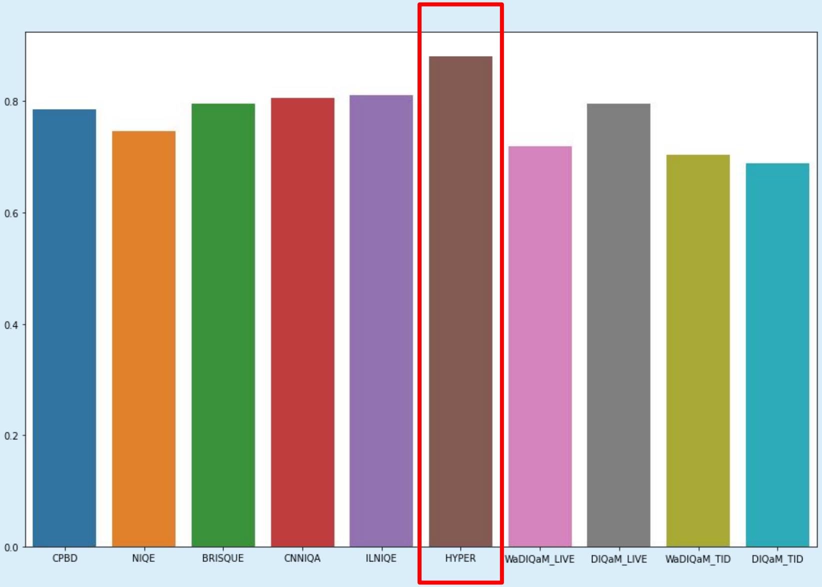
With a variety of IQA models available [8]–[16], how to fairly evaluate their relative performance becomes pivotal. The conventional approach in the literature is to compute correlations between model predictions and the "ground truth" labels, typically the mean opinion scores (MOSs) given by human subjects, of the images on a handful of commonly used IQA databases. However, collecting MOS via subjective testing is a costly process. In practice, the largest IQA database that is publicly available contains a maximum of 3, 000 subject-rated images, many of which are generated from the same source images with different distortion types and levels. As a result, only less than 30 source images are included. By contrast, the space of digital images is of very high dimension, which is equal to the number of pixels in the images, making it extremely difficult to collect sufficient subjective opinions to adequately cover the space. Perhaps more importantly, using only a few dozens of source images is very unlikely to provide

Manuscript received July 1, 2016; revised October 25, 2016; accepted

How to choose NR Metric

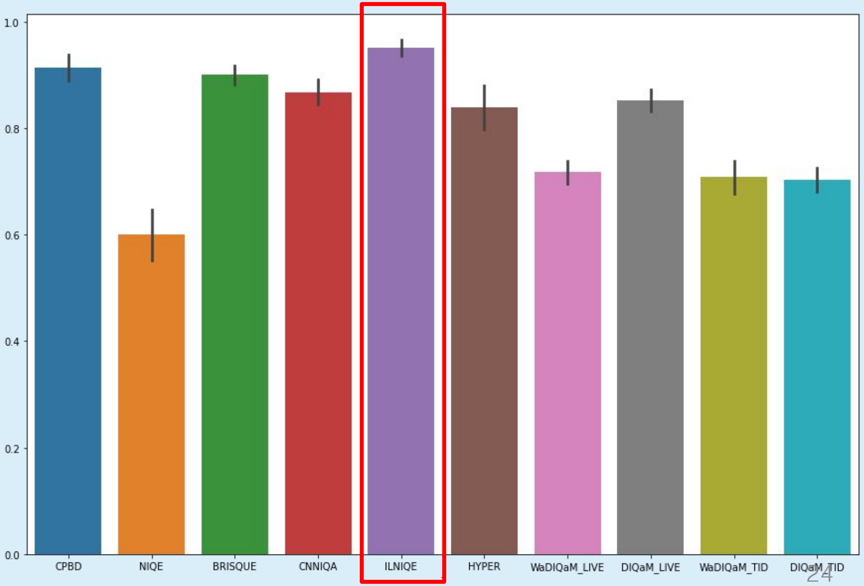
D-test for NR-IQAs in KADID-10k

D = 0.880



L-test for NR-IQAs in KADID-10k

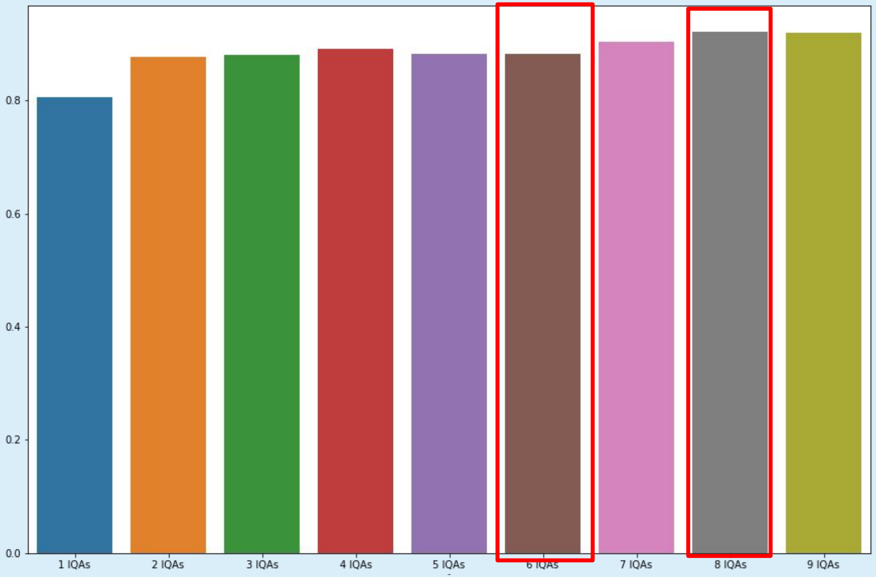
L_s = 0.966, L_k = 0.937



How to choose NR Metric

D-test for New_NR-IQA in KADID-10k Val set

D = 0.883

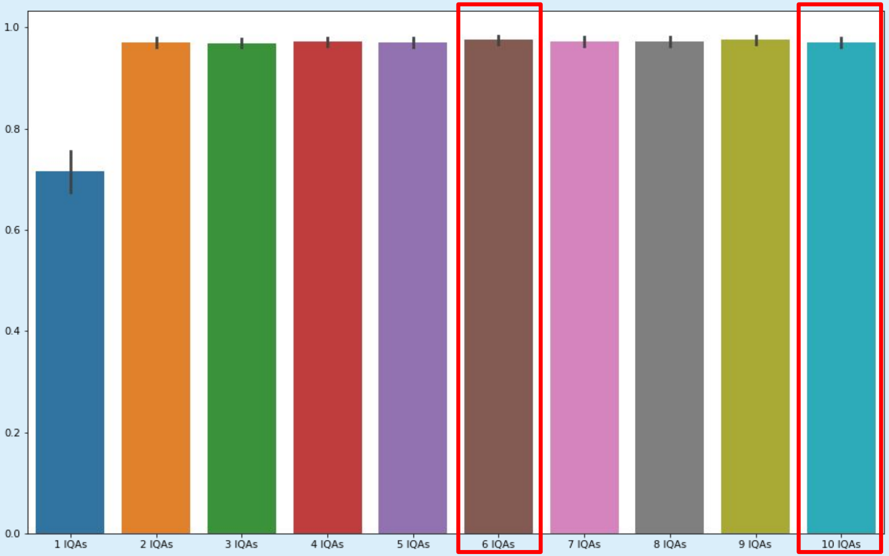


D=0.921

How to choose NR Metric

L-test for New_NR-IQA in KADID-10k Val set

L_s = 0.983, L_k = 0.967

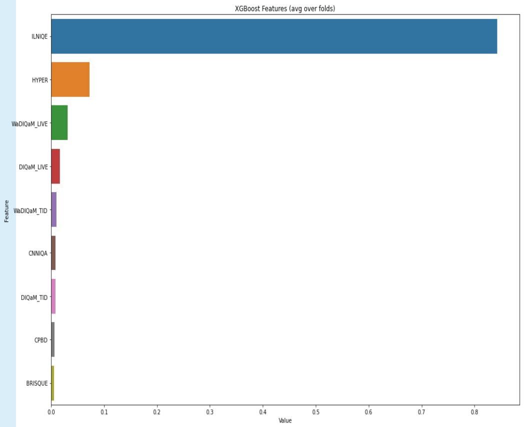


L_s = 0.991, L_k = 0.981

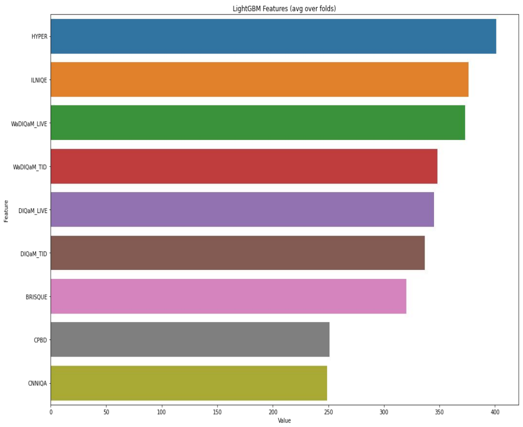
How to choose NR Metric

Feature Importance

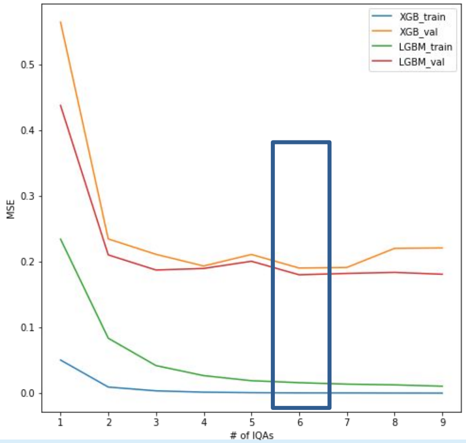
XGBoost



LGBM



실험결과

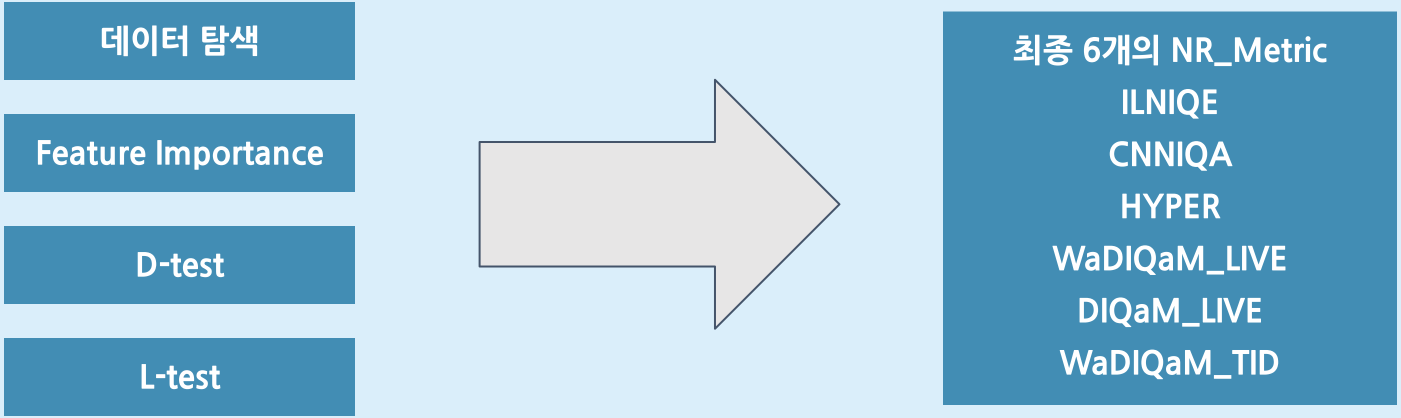


Feature Importance 가 높은 순서대로 Metric 조합

→ 6개의 NR_Metric 사용할 때 가장 좋은 mse 값이 나옴 mse : 0.1799

How to choose NR Metric

- 4가지 케이스를 토대로 결정



데이터 탐색

Feature Importance

D-test

L-test

최종 6개의 NR_Metric

ILNIQE

CNNIQA

HYPER

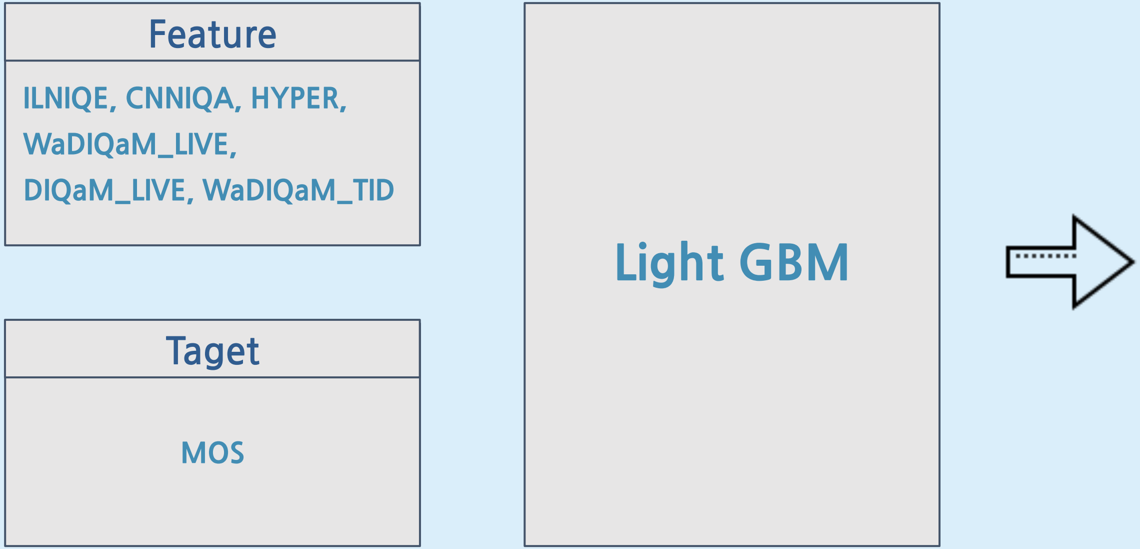
WaDIQaM_LIVE

DIQaM_LIVE

WaDIQaM_TID

Model Learning

- LGBM Regressor



Feature

ILNIQE, CNNIQA, HYPER, WaDIQaM_LIVE, DIQaM_LIVE, WaDIQaM_TID

Taget

MOS

Light GBM

MSE

train : 0.016

val : 0.179

ROCC : 0.919

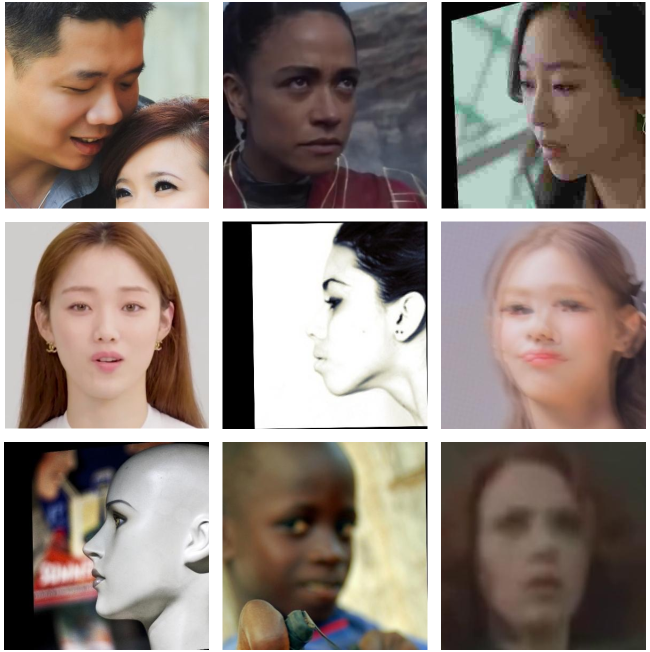
LCC : 0.936

결과 및 해석

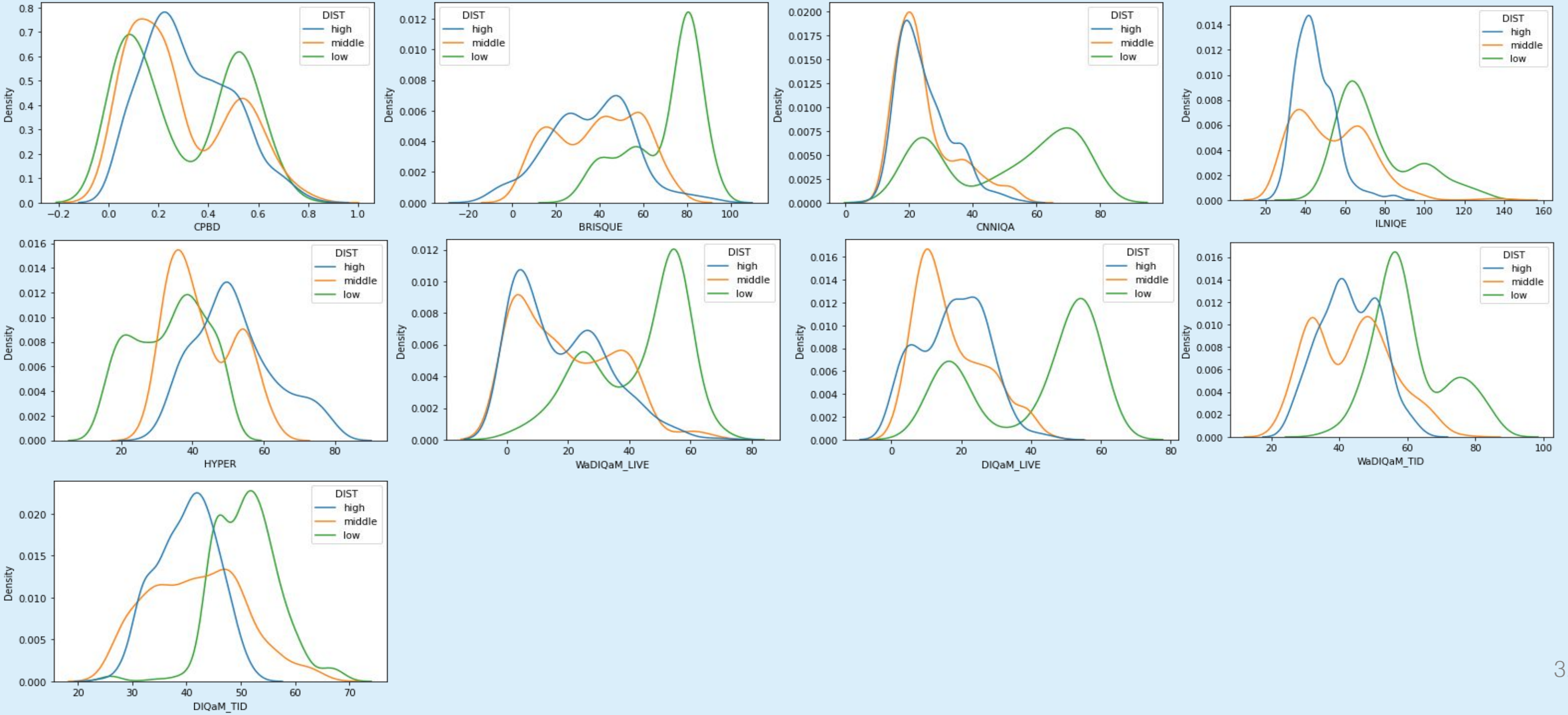
- Test Set Exploration
- Test Set score for each NR-IQAs
- Metric 계산시 긴 시간 소요

Test Set Exploration

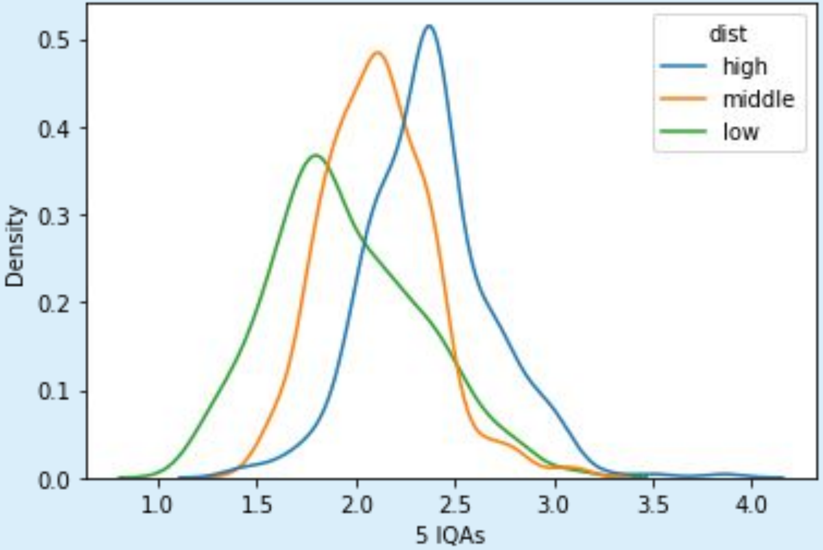
- 선명, 살짝흐린, 많이흐린 3가지 카테고리로 구성
- 384x384 size
  - 총 1000장 : 선명 334, 살짝 흐린 333, 많이 흐린 333 장
- 배경이 포함된 크롭되지 않은 얼굴 이미지
- Blur 외 Distortion 및 동영상 캡쳐샷 다수
- 둘 이상의 인물이 포함된 사진도 존재
- 정면, 좌측, 우측, 마네킹 등 다양한 데이터가 존재



Test Set score Distribution for each NR-IQAs



Test Set score Distribution for New NR-IQA



Test Set score Distribution

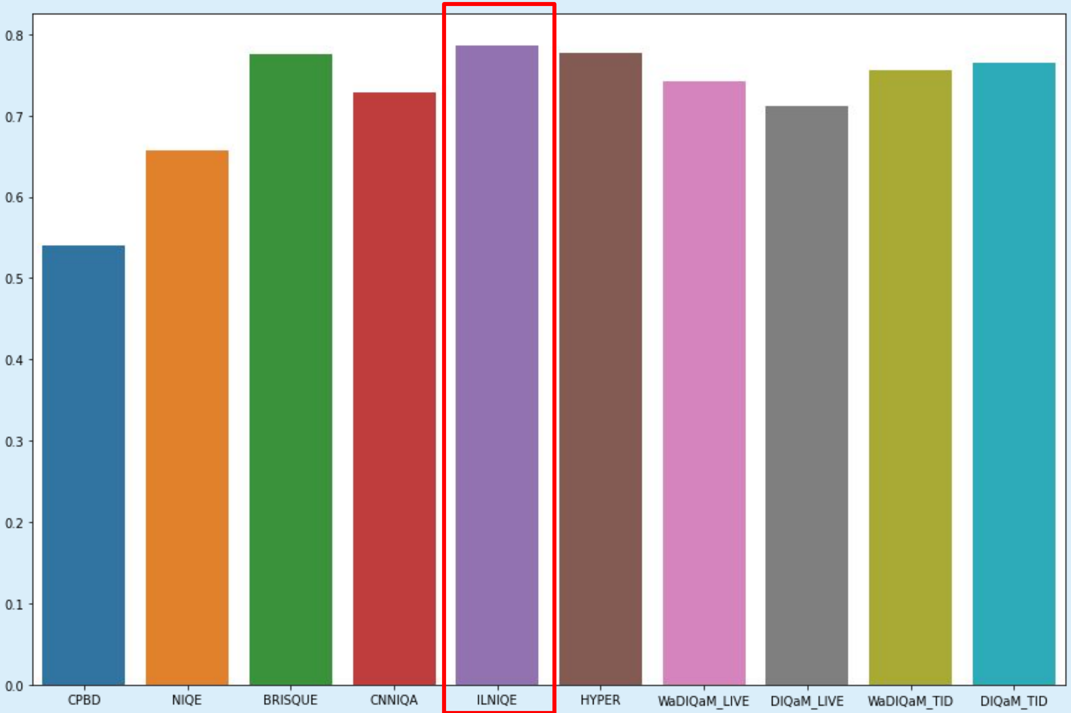
- 선명, 조금흐림, 많이흐림 각 단계에 맞게 밀집도가 형성되어 있음

- 가장 높은 밀집도 기준으로 구별이 되는 것을 알 수 있음

- 점수 분포 흐름도 각 단계에 맞게 형성되어 있음
  → high score 가 가장 높고, middle, low 순서로 잘 나옴

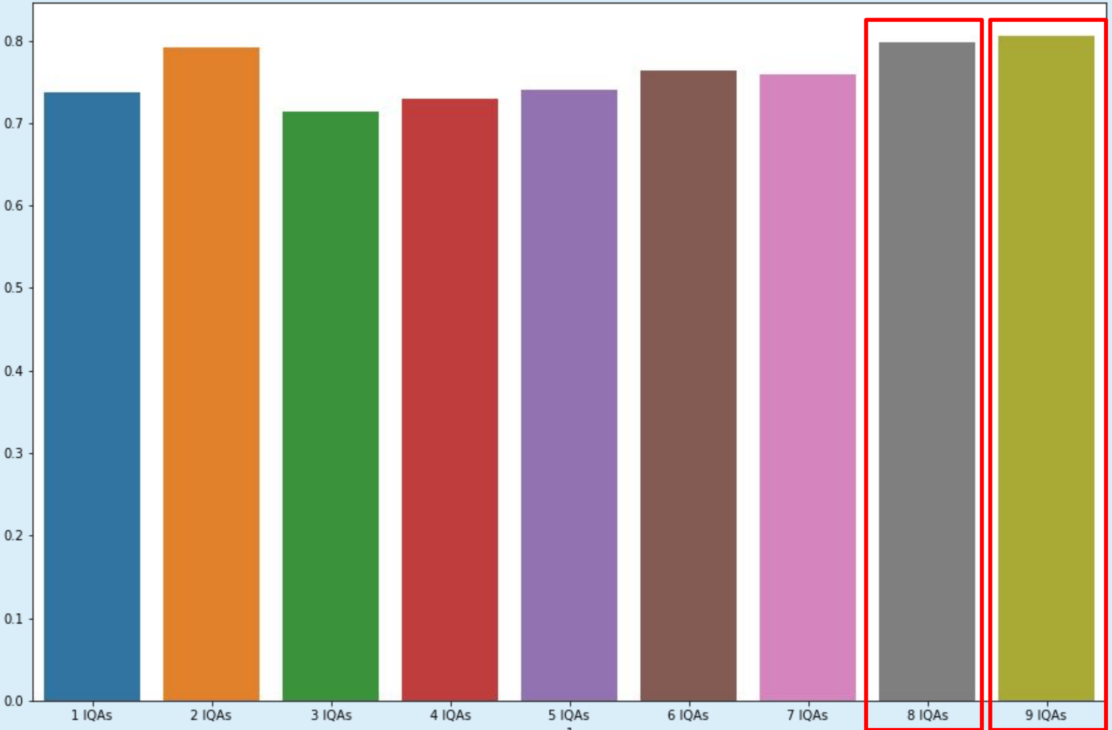
P-test for NR-IQAs in TestSet

p = 0.786



P-test for New_NR-IQA in TestSet

p = 0.798



p = 0.805

고찰

성능

- ▪개별 NR-IQA의 score 보다 New NR-IQA score 분포가 좋음
  → 개별 NR-IQA의 score분포가 울퉁불퉁한 이유가 distortion 유형의 차이 때문
  → distortion 유형에 더욱 robust한 특성이 있음

- ▪ 학습하지 않은 Blur (test set) 에도 구별가능한 성능이 확인됨

- ▪Metric들의 Fusion 이 더 나은 성능을 보여줌

- ▪Test set의 유형 및 종류에 따라 NR Metric의 개수와 종류를 다르게 Fusion 가능

- ▪Kadid set 의 MOS 를 학습함으로써, 학습된 모델을 통해 산출되는 결과값에 대한 객관성이 향상됨
  → MOS가 1~5점 사이이므로, MOS를 학습한 New NR-IQA score 가 1~5점으로 산출됨

- ▪Metric 특성으로 인해 성능 예측이 가능 → 블랙박스 현상이 줄어듬

고찰

범용성

- Flow가 매우 간단하여 적용 및 응용이 용이
- 개별 NR-IQA에 대한 이해 없이도 활용 가능
- Face IQA가 아닌 일반 IQA를 사용하여, 인물의 포지션, 머리카락, 장신구 등과 무관하게 평가 가능

유지관리 및 비용

- 최신버전의 NR-Metric 또는 필요한 Metric 으로 교체하기 쉬움 → 유지보수 하기가 좋음
- 각 NR-IQA에 대한 이해 없이도 활용 가능
- ML 모델을 사용함으로써 학습에 걸리는 시간을 줄임 → 모델의 교체에 많은 시간을 소요하지 않음
- 학습시 feature들이 모델에 미치는 영향을 산출 할 수 있음 → 블랙박스 현상이 줄어듬
- 컴퓨팅 파워가 많이 필요하지 않음

Todos

- 특성에 맞는 다양한 NR-IQA 학습 - jpeg 압축 등의 다양한 distortion
- 더 정밀한 fusion을 위한 실험체계 구축
- ML 모델이 아닌 DNN 모델 학습에서의 성능평가

감사합니다.

Q & A

참조 용어 설명

▪이미지 품질 평가(IQA)
블러링, 노이즈, 압축 등의 이미지 왜곡 현상으로 인한 이미지 품질의 손실 또는 열화 정도를 평가하는 과제

▪주관적 품질 평가(Subjective methods)
IQA가 이미지 퀄리티에 대한 인간의 인식을 컴퓨터로 예측한다는 점에서 가장 신뢰성 있는 지표

▪MOS(Mean opinion score)
주관적 평가 방식 중 가장 보편적으로 쓰이는 지표

▪객관적 품질 평가(Objective methods)

▪Full-reference method
고화질 원본 이미지와 비교하여 품질 평가

▪No-reference method
원본 이미지에 대한 참조 없이 이미지의 품질을 평가

▪OU(Opinion unaware)
라벨이 없고 따라서 학습 과정이 필요 없는 방법

참조 Distortion 종류

▪Blurs
  # 01 Gaussian blur: filter with a variable Gaussian kernel
  # 02 Lens blur: filter with a circular kernel
  # 03 Motion blur: filter with a line kernel

▪Color distortions
  # 04 Color diffusion: Gaussian blur the color channels (a and b) in the Lab color-space
  # 05 Color shift: randomly translate the green channel, and blend it into the original image masked by a gray level map: the normalized gradient magnitude of the original image
  # 06 Color quantization: convert to indexed image using minimum variance quantization and dithering with 8 to 64 colors
  # 07 Color saturation 1: multiply the saturation channel in the HSV color-space by a factor
  # 08 Color saturation 2: multiply the color channels in the Lab colorspace by a factor

▪Compression
  # 09 JPEG2000: standard compression
  # 10 JPEG: standard compression

▪Sharpness and contrast
  # 11 High sharpen: over-sharpen image using unsharp masking
  # 12 Contrast change: non-linearly change RGB values using a Sigmoid-type adjustment curve

참조 Distortion 종류

▪Noise

# 13 White noise: add Gaussian white noise to the RGB image

# 14 White noise in color component: add Gaussian white noise to the YCbCr converted image (both to the luminance ‘Y‘ and the color channels ‘Cb‘ and ‘Cr‘)

# 15 Impulse noise: add salt and pepper noise to the RGB image

# 16 Multiplicative noise: add speckle noise to the RGB image

# 17 Denoise: add Gaussian white noise to RGB image, and then apply a denoising DnCNN to each channel separately

▪Brightness change

# 18 Brighten: non-linearly adjust the luminance channel keeping extreme values fixed, and increasing others

# 19 Darken: similar to brighten, but decrease other values

# 208 Mean shift: add constant to all values in image, and truncate to original value range

▪Spatial distortions

# 21 Jitter: randomly scatter image data by warping each pixel with random small offsets (bicubic interpolation)

# 22 Non-eccentricity patch: randomly offset small patches in the image to nearby locations

# 23 Pixelate: downsize image and upsize it back to the original size using nearest-neighbor interpolation in each case

# 24 Quantization: quantize image values using N thresholds obtained using Otsus method

# 25 Color block: insert homogeneous random colored blocks at random locations in the image