ITMO UNIVERSITY

Saint Petersburg, Russia

Специализированные технологии машинного обучения / Advanced Machine learning Technologies

Lecture 6 – Advanced Quality Assessment in ML

Outline

ITMO UNIVERSITY

1. Metrics for regression and classification (repeat)
2. Losses and model's quality
3. How to evaluate the quality for non-numerical entity?
4. Metrics of quality for text generation
5. Image Quality Assessment: FR and NR

Regression metrics

ITMO UNIVERSITY

- R^2 – determination coefficient;
- MSE (RMSE) – mean squared error;
- MAE – mean absolute error;
- MAPE - mean absolute percentage error;
- SMAPE – symmetric mean absolute percentage error;
- …

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2 \qquad RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$$

$$R^2 = 1 - \frac{RSS}{TSS}$$

$$RSS = \sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n}(y_t - \hat{y}_t)^2$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t| \qquad MAPE = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|$$

$$TSS = \sum_{t=1}^{n}(y_t - \bar{y}_t)^2$$

$$SMAPE = \frac{100\%}{n}\sum_{t=1}^{n}\frac{|\hat{y}_t - y_t|}{\frac{1}{2} * (|y_t| + |\hat{y}_t|)}$$

Classification metrics (I)

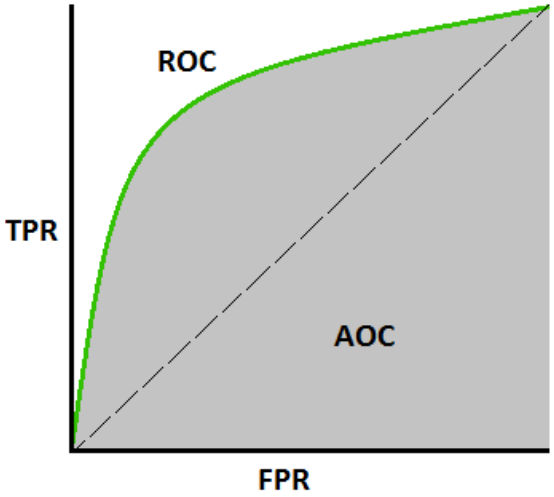
ITMO UNIVERSITY

Confusion matrix

		Predicted 0	Predicted 1
Actual	0	TN (True Negative)	FP (False Positive)
	1	FN (False Negative)	TP (True Positive)

I type error - is an error consisting in the rejection of a correct hypothesis.

II type error – is an error consisting in making a false hypothesis.



- In some problems, probabilities in their pure form are needed without their binarization by the probability threshold T.
- As the threshold value increases, it is less FP and greater FN.
- When constructing a **ROC curve (Receiver Operating Characteristic),** the binarization threshold is varied, and some values are calculated depending on the number of **FP** and **FN** errors.
- In order to characterize the curve numerically, the AUC (Area Under the Curve) metric is used - the area under the ROC curve.

Classification metrics (II)

ITMO UNIVERSITY

Metric	Formula	Evaluation Focus
Accuracy	$ACC = \frac{TP+TN}{TP+TN+FP+FN}$	Overall effectiveness of a classifier
Error rate	$ERR = \frac{FP+FN}{TP+TN+FP+FN}$	Classification error
Precision	$PRC = \frac{TP}{TP+FP}$	Class agreement of the data labels with the positive labels given by the classifier
Sensitivity	$SNS = \frac{TP}{TP+FN}$	Effectiveness of a classifier to identify positive labels
Specificity	$SPC = \frac{TN}{TN+FP}$	How effectively a classifier identifies negative labels
ROC	$ROC = \frac{\sqrt{SNS^2+SPC^2}}{\sqrt{2}}$	Combined metric based on the Receiver Operating Characteristic (ROC) space [53]
F_1 score	$F_1 = 2\frac{PRC \cdot SNS}{PRC+SNS}$	Combination of precision (PRC) and sensitivity (SNS) in a single metric
Geometric Mean	$GM = \sqrt{SNS \cdot SPC}$	Combination of sensitivity (SNS) and specificity (SPC) in a single metric

ITMOre than a UNIVERSITY

Source – Lucue et.al., 2018

External methods of clustering QA

ITMO UNIVERSITY

The Jaccard Index

$$Jaccard = \frac{TP}{TP + TN + FP}$$

Entropy measures the "purity" of class labels. If all clusters consist of objects of the same class, then the entropy is 0.

$$E = -\sum_i p_i \left(\sum_j \frac{p_{ij}}{p_i} \cdot \log \frac{p_{ij}}{p_i} \right)$$

Purity associates a cluster with the most numerous class in this cluster, is in the interval [0, 1], and the value = 1 corresponds to optimal clustering.

$$P = \sum_i p_i \cdot \max_j \frac{p_{ij}}{p_i}$$

The F-measure is the harmonic average between precision and recall.

$$F = \sum_j p_j \cdot \max_i \left[2 \cdot \frac{p_{ij}}{p_i} \cdot \frac{p_{ij}}{p_j} \Big/ \frac{p_{ij}}{p_i} + \frac{p_{ij}}{p_j} \right]$$

ITMO More than a UNIVERSITY

Internal methods of clustering QA

ITMO UNIVERSITY

Cluster Cohesion

The closer to each other the objects within the clusters are, the better the separation. It is necessary to minimize the intraclass distance, for example, the sum of the squares of the deviations:

$$WSS = \sum_{j=1}^{M} \sum_{i=1}^{|c_j|} (x_{ij} - \bar{x_j})^2, \quad \text{M – number of clusters}$$

The silhouette shows the similarity of the object to the objects of its cluster in comparison with others.

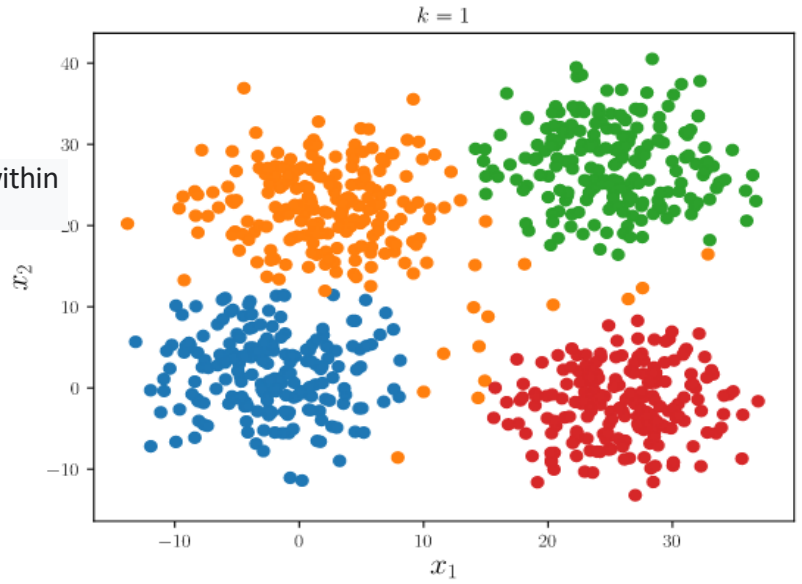
$$Sil(C) = \frac{1}{N} \sum_{c_k \in C} \sum_{x_i \in c_k} \frac{b(x_i, c_k) - a(x_i, c_k)}{max\{a(x_i, c_k), b(x_i, c_k)\}}$$

Compactness - the average distance between an object and other objects within the cluster

$$a(x_i, c_k) = \frac{1}{|c_k|} \sum_{x_j \in c_k} \|x_i - x_j\|$$

Separability - the average distance between an object and other objects from other clusters:

$$b(x_i, c_k) = \min_{c_l \in C | c_k} \left\{ \frac{1}{|c_l|} \sum_{x_j \in c_l} \|x_i - x_j\| \right\}$$



ITMO UNIVERSITY

Quality of text generation

Machine translation metrics

ITMO UNIVERSITY

“BLUE score” metric (bilingual evaluation understudy):

- correspondence between a machine's output and translation made by human (or every desired output of the model with specified label – “reference translation”);
- scores are calculated for individual translated segments—generally sentences—by comparing them with a set of good quality reference translations.
- those scores are then averaged over the whole corpus to reach an estimate of the translation's overall quality.
- *Intelligibility or grammatical correctness, order of words are not taken into account;*

$$BLUE = m/w_t$$

where:
m - is number of words from the candidate that are found in the reference,
w_t - is the total number of words in the candidate.

Main idea:

- the closer a machine translation is to a professional human translation, the better it is (this also works for other text generation tasks like text summarization, Q&A systems, conditional text generating etc.)

BLUE. Example

ITMO UNIVERSITY

$$BLUE = m/w_t$$

- m - is number of words from the candidate that are found in the reference,
- w_t - is the total number of words in the candidate.

For "(the, the , the , the , the , the)" there is a perfect score 7/7=1 , despite the fact that the candidate translation above doesn't reflect the content of either of the references.

- It can be reformulated: change **m** with $\mathbf{m}_{max}$ - the maximum total count in any of the reference translations. In that case P is changed to 2/7.

Also, we also can construct N-gramms:

→ using n-grams approach better represents the sentences in different tasks;
→ there is no guarantee that an increase in BLEU score is an indicator of improved translation quality;
→ Uncertainties…

Input: *«Le chat est sur le tapis »*

Candidate	the	the	the	the	the	the	the
Reference 1	the	cat	is	on	the	mat	
Reference 2	there	is	a	cat	on	the	mat

Comparing metrics for candidate "the the cat"

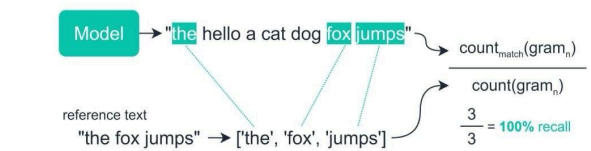
Model	Set of grams	Score
Unigram	"the", "the", "cat"	$\frac{1+1+1}{3} = 1$
Grouped Unigram	"the"*2, "cat"*1	$\frac{1+1}{2+1} = \frac{2}{3}$
Bigram	"the the", "the cat"	$\frac{0+1}{2} = \frac{1}{2}$

ROUGE metrics

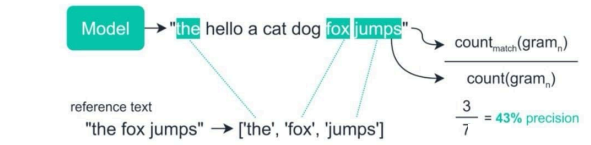
ROUGE-N
measures the number of matching ‘n-grams’ between our model-generated text and a ‘reference’.

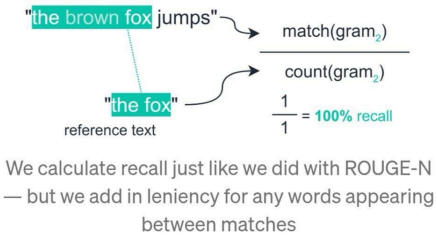
ROUGE-S:
using the skip-gram metric allows us to search for consecutive words from the reference text, that appear in the model output but are separated by one-or-more other words.

ROUGE-L
measures the longest common subsequence (LCS) between our model output and reference. All this means is that we count the longest sequence of tokens that is shared between both.

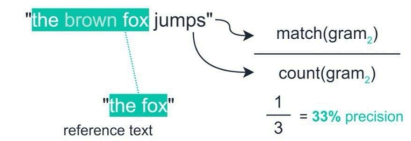


ROUGE-N



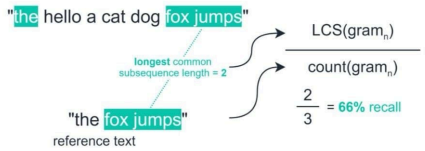


We calculate recall just like we did with ROUGE-N — but we add in leniency for any words appearing between matches

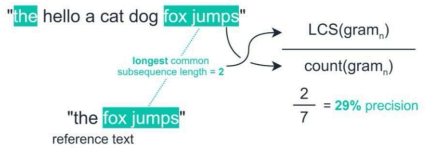


The same applies to our precision metric too

ROUGE-S



Our LCS recall calculation



Precision is much the same but we switch our total n-gram count from the reference to the model

ROUGE-L

ITMO UNIVERSITY

Image Quality Assessment

Quality?

ITMO UNIVERSITY

Image quality attributes

- Sharpness
- Noise
- Dynamic range
- Tone reproduction
- Contrast
- Color accuracy
- Distortion
- Vignetting,
- Exposure accuracy
- Lateral chromatic aberration (LCA)
- Lens flare
- Color moiré
- Artifacts

Visual Perception?

ITMO more than a UNIVERSITY

Image Quality Assessment. FR and NR.

Full-reference methods:
Evaluate difference between the input and the target and estimate the quality metric as the measure of similarity between them.

Image Quality Assessment. FR and NR.

ITMO UNIVERSITY

Full-reference methods:
Evaluate difference between the input and the target and estimate the quality metric as the measure of similarity between them.

?

An image generated by a StyleGAN

No-reference methods:
Evaluate the quality of the image as it is.

ITMO UNIVERSITY

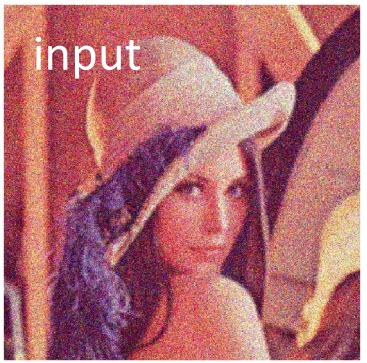
Full-reference IQA

Metrics and loss functions for FR IQA

ITMO UNIVERSITY

To train a model we should minimize the difference between target (reference) and input images:

$$\hat{\theta} = \arg\min_{\theta} \mathcal{L}(\hat{I}_y, I_y) + \lambda\Phi(\theta),$$



We already know various loss functions for that task:

$$\mathcal{L}_{\text{pixel\_l1}}(\hat{I}, I) = \frac{1}{hwc}\sum_{i,j,k}|\hat{I}_{i,j,k} - I_{i,j,k}|,$$
$$\mathcal{L}_{\text{pixel\_l2}}(\hat{I}, I) = \frac{1}{hwc}\sum_{i,j,k}(\hat{I}_{i,j,k} - I_{i,j,k})^2$$

- L1 and L2 pixel-wise losses;

$$\mathcal{L}_{\text{pixel\_Cha}}(\hat{I}, I) = \frac{1}{hwc}\sum_{i,j,k}\sqrt{(\hat{I}_{i,j,k} - I_{i,j,k})^2 + \epsilon^2},$$

- Charbonnier loss (RMSE with regularization);

$$\mathcal{L}_{\text{content}}(\hat{I}, I; \phi, l) = \frac{1}{h_l w_l c_l}\sqrt{\sum_{i,j,k}(\phi^{(l)}_{i,j,k}(\hat{I}) - \phi^{(l)}_{i,j,k}(I))^2}$$

- Content loss;

$$\mathcal{L}_{\text{texture}}(\hat{I}, I; \phi, l) = \frac{1}{c_l^2}\sqrt{\sum_{i,j}(G^{(l)}_{i,j}(\hat{I}) - G^{(l)}_{i,j}(I))^2}$$

- Style (texture) loss;

$$G^{(l)}_{ij}(I) = \text{vec}(\phi^{(l)}_i(I)) \cdot \text{vec}(\phi^{(l)}_j(I))$$

- Gram matrix fot image I;

$$\mathcal{L}_{\text{gan\_ce\_g}}(\hat{I}; D) = -\log D(\hat{I}),$$
$$\mathcal{L}_{\text{gan\_ce\_d}}(\hat{I}, I_s; D) = -\log D(I_s) - \log(1 - D(\hat{I}))$$

- Discriminator + Generator losses for GAN;

Peak Signal-to-Noise Ratio (PSNR)

ITMO UNIVERSITY

- The most popular metric for Full-Reference IQA evaluation.
- Inversely proportional to MSE;
- Is evaluated in dB;
- Absolute value depends on bitrate L (for 8-bit images L=255);

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{L^2}{\frac{1}{N}\sum_{i=1}^{N}(I(i) - \hat{I}(i))^2}\right)$$

- Good for quality measurement of lossy transformation (e.g., image compression, image inpainting)
- Very easy to compute, stable;
- The most widely used evaluation criteria
- Fails on many real-world scenes;
- Insensitive to blur;

Structural Similarity Index (SSIM)

ITMO UNIVERSITY

- SSIM – is well-known popular full-reference metric for IQA.
- SSIM measures the structural similarity between images, based on independent comparisons in terms of luminance, contrast, and structures.
- For an image I with N pixels, the luminance μ and contrast σ are estimated as the mean and standard deviation of the image intensity and then the comparisons on luminance and contrast, are calculated:

$$\mathcal{C}_l(I,\hat{I}) = \frac{2\mu_I\mu_{\hat{I}} + C_1}{\mu_I^2 + \mu_{\hat{I}}^2 + C_1}$$ - luminosity difference

$$\mathcal{C}_c(I,\hat{I}) = \frac{2\sigma_I\sigma_{\hat{I}} + C_2}{\sigma_I^2 + \sigma_{\hat{I}}^2 + C_2}$$ - contrast difference

- Structure comparison function C_s is dependent on covariance between two images and defined as:

$$\sigma_{I\hat{I}} = \frac{1}{N-1}\sum_{i=1}^{N}(I(i)-\mu_I)(\hat{I}(i)-\mu_{\hat{I}})$$

$$\mathcal{C}_s(I,\hat{I}) = \frac{\sigma_{I\hat{I}} + C_3}{\sigma_I\sigma_{\hat{I}} + C_3},$$

Since the SSIM evaluates the reconstruction quality from the perspective of the HVS, it better meets the requirements of perceptual assessment than PSNR.

But not in every case…

Final metric evaluation: $\mathrm{SSIM}(I,\hat{I}) = [\mathcal{C}_l(I,\hat{I})]^\alpha[\mathcal{C}_c(I,\hat{I})]^\beta[\mathcal{C}_s(I,\hat{I})]^\gamma$

where α, β, γ - are control parameters for adjusting the relative importance

Problems with PSNR and SSIM

ITMO UNIVERSITY

Gap Between IQA Metric and Human Judgment

Ground Truth
PSNR / SSIM

23.52 / 0.7056
Good in PSNR, SSIM

19.86 / 0.5530
Preferred by Human

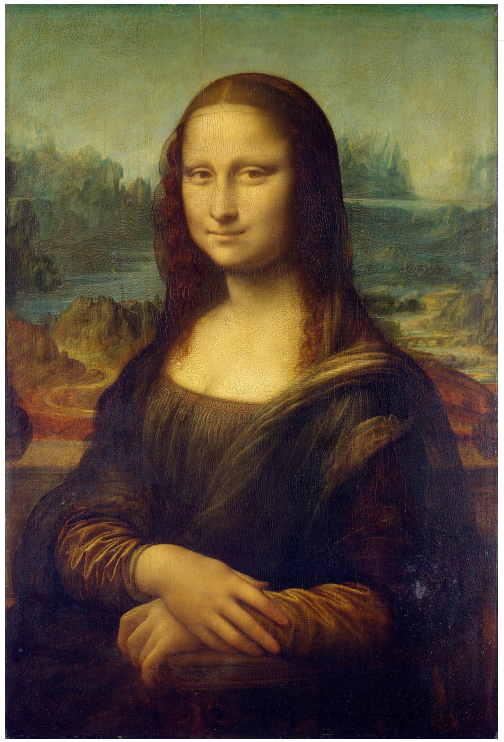
Generated by GAN

IQA and human perception

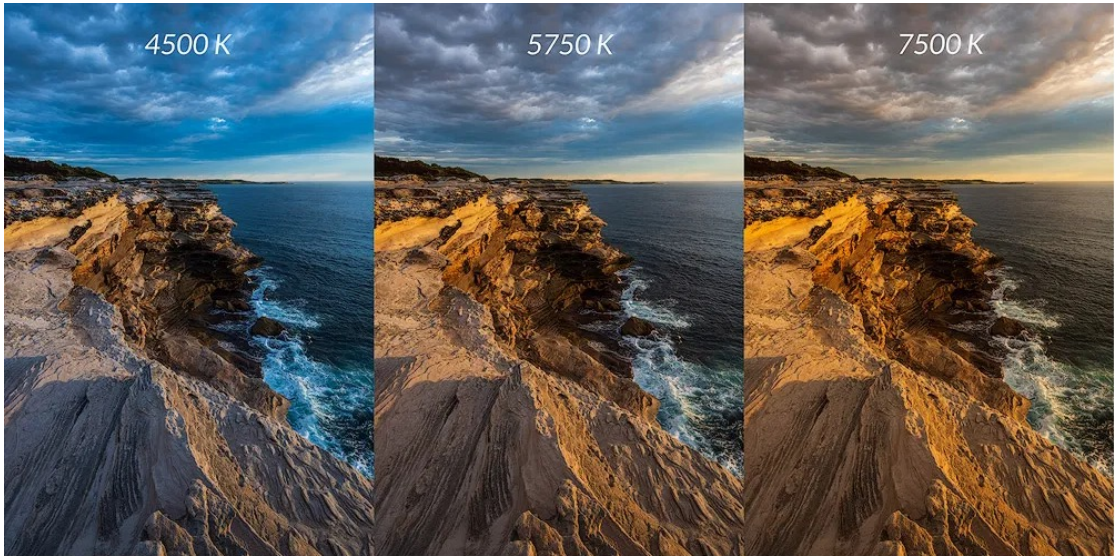
ITMO UNIVERSITY

Boke

Under-exposed



Painting/art

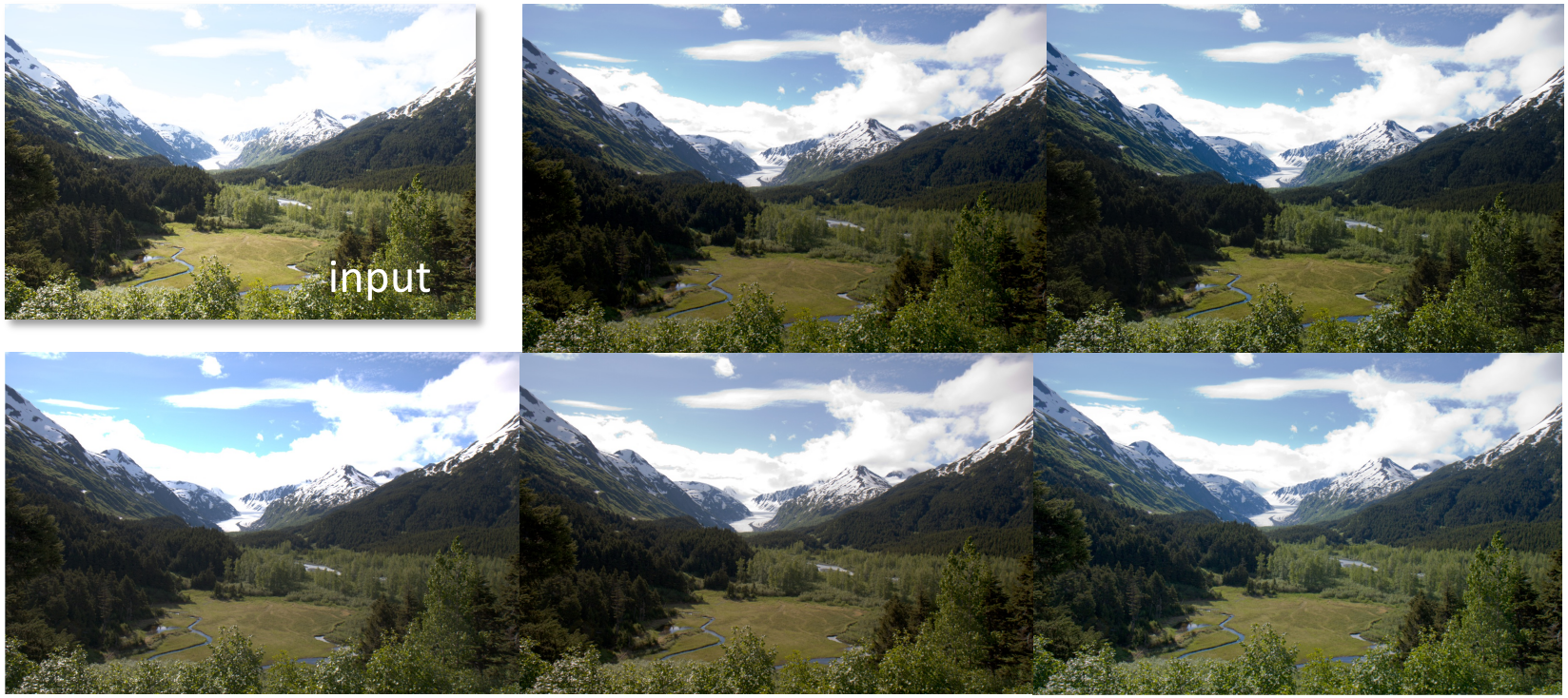


White balance (color temperature)

MIT Adobe FiveK Dataset

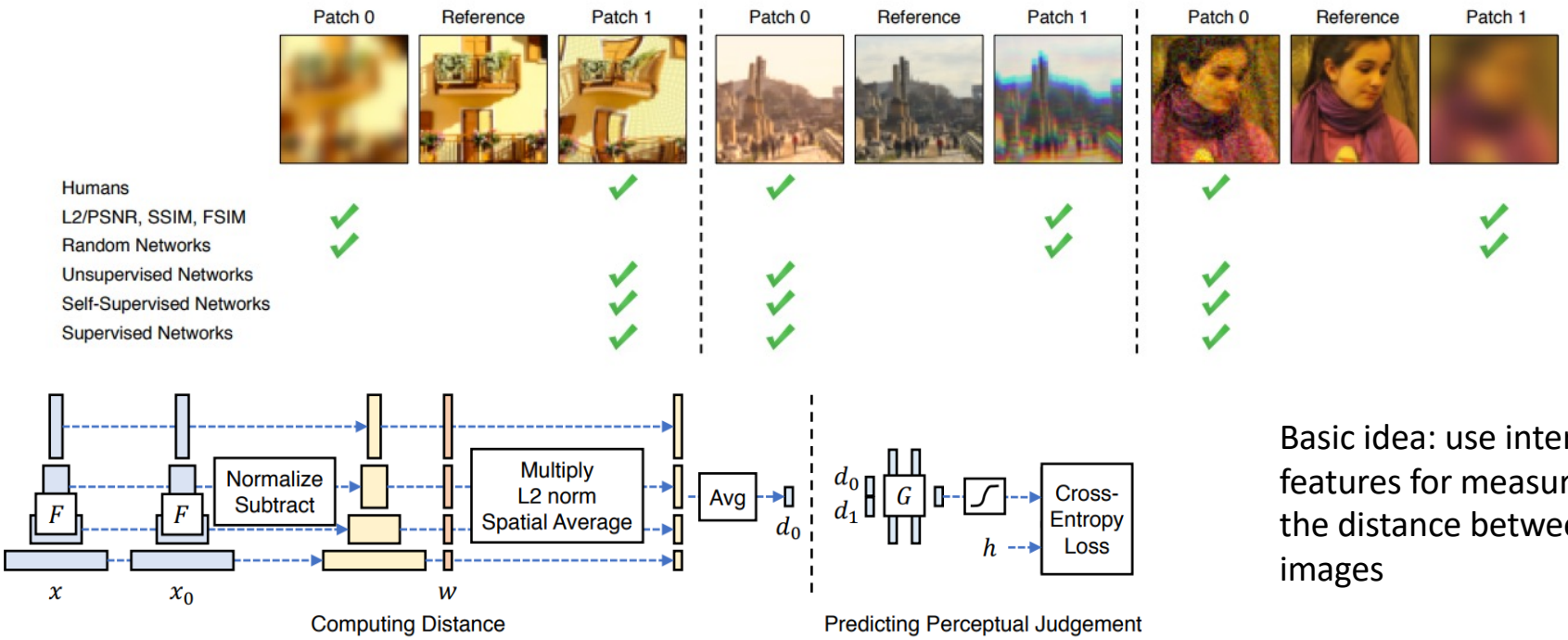
ITMO UNIVERSITY

MIT Adobe FiveK Dataset consists of 5,000 photographs taken with SLR cameras by a set of different photographers. Each of them retouched all the 5,000 photos using a software dedicated to photo adjustment (Adobe Lightroom) on which they were extensively trained. Retouchers were asked to achieve visually pleasing renditions, akin to a postcard.



Learned Perceptual Image Patch Similarity (LPIPS) metric

ITMO UNIVERSITY



Basic idea: use internal features for measuring the distance between images

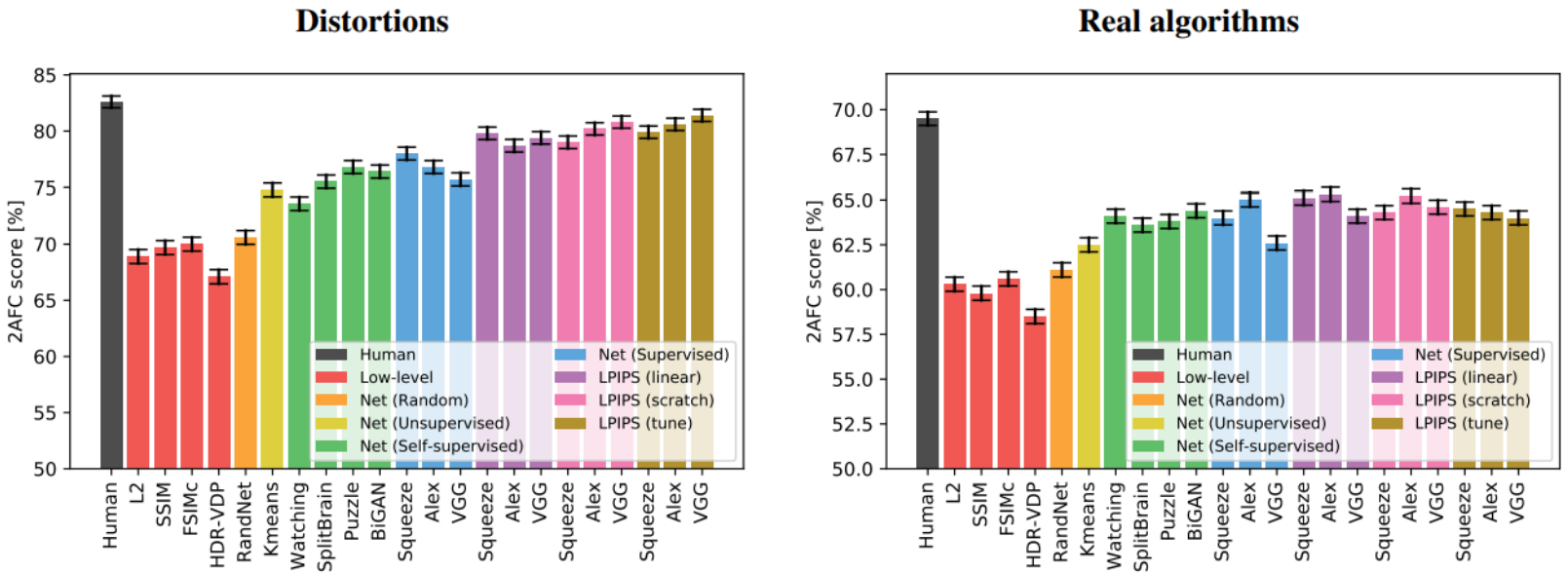
$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} ||w_l \odot (\hat{y}^l_{hw} - \hat{y}^l_{0hw})||_2^2$$ - evaluating distance between reference and target

Zhang et. Al., The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, 2018

LPIPS experimental results

ITMO UNIVERSITY

2AFC – “two alternatives forced choice results



- Network trained on **any** task performs better than classic approaches;
- Deep embeddings are more sensitive to blur distortions;
- Humans are not ideal in 2AFC (~82%);

Zhang et. Al., The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, 2018

DISTS – Deep Image Structural and Texture similarity

ITMO UNIVERSITY

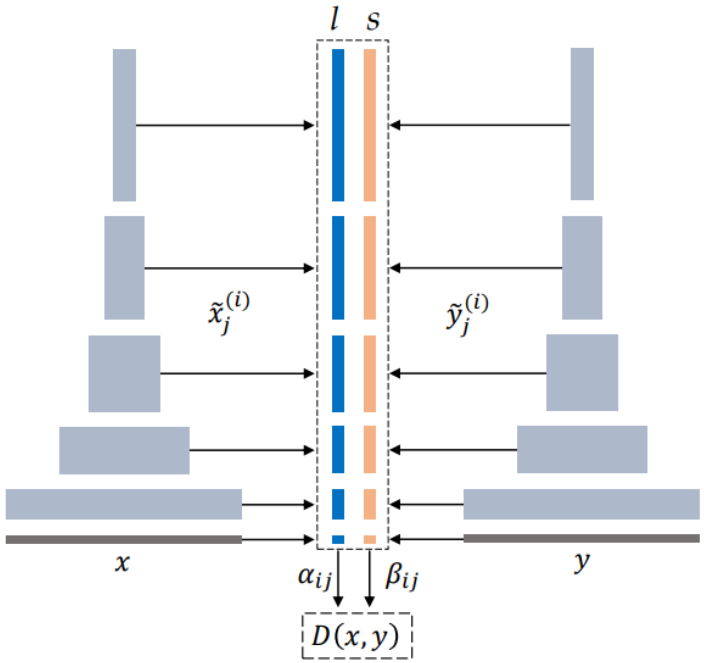
- combines texture similarity metric, structural similarity and their correlation;
- robust metric
- highly sensitive to the texture distortions;
- can be easily applied for the texture restoration tasks as a loss-function;

$$D(x, y; \alpha, \beta) = 1 - \sum_{i=0}^{m} \sum_{j=1}^{n_i} \left(\alpha_{ij} l(\tilde{x}_j^{(i)}, \tilde{y}_j^{(i)}) + \beta_{ij} s(\tilde{x}_j^{(i)}, \tilde{y}_j^{(i)}) \right)$$

$$l(\tilde{x}_j^{(i)}, \tilde{y}_j^{(i)}) = \frac{2\mu_{\tilde{x}_j}^{(i)} \mu_{\tilde{y}_j}^{(i)} + c_1}{\left(\mu_{\tilde{x}_j}^{(i)}\right)^2 + \left(\mu_{\tilde{y}_j}^{(i)}\right)^2 + c_1}$$

$$s(\tilde{x}_j^{(i)}, \tilde{y}_j^{(i)}) = \frac{2\sigma_{\tilde{x}_j \tilde{y}_j}^{(i)} + c_2}{\left(\sigma_{\tilde{x}_j}^{(i)}\right)^2 + \left(\sigma_{\tilde{y}_j}^{(i)}\right)^2 + c_2}$$

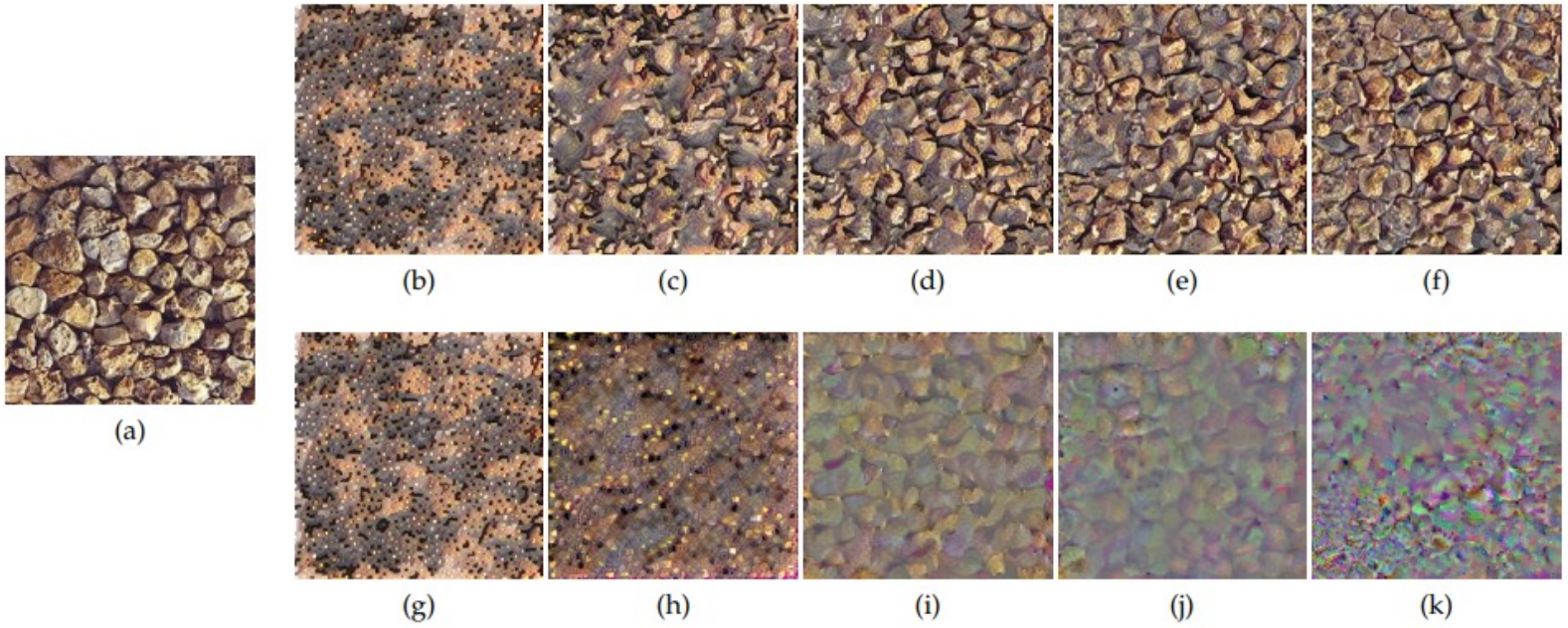
x(i) – reference image's feature maps;
y(i) – test (distorted) image feature maps;
m=5 - # of CNN layers to obtain feature maps;
n_i - # of feature maps in the i-th layer.



Ding et.al., Image Quality Assessment: Unifying Structure and Texture Similarity, 2020

DISTS experimental results example

ITMO UNIVERSITY



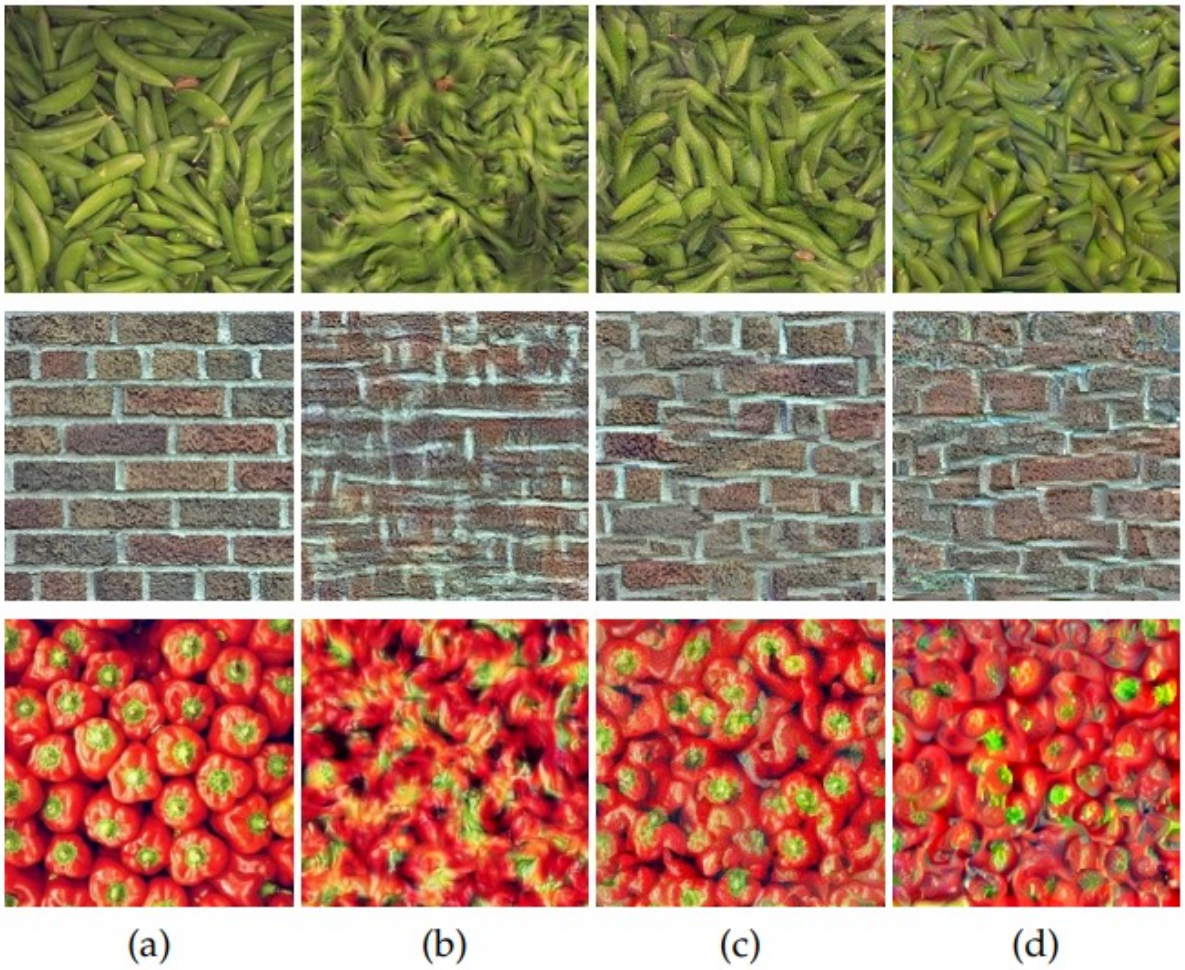
Cumulative images with correspondingly added feature maps from 1 to 5 layers of VGG16

Images corresponding to the 1-5 layers of VGG16

Ding et.al., Image Quality Assessment: Unifying Structure and Texture Similarity, 2020

DISTS experimental results example

ITMO UNIVERSITY



Images synthesized using different texture models.

a) Reference textures
b) Portilla & Simoncelli model
c) Gatys et. al. model
d) DISTS model

Ding et.al., Image Quality Assessment: Unifying Structure and Texture Similarity, 2020

ITMO UNIVERSITY

No-Reference (“blind”) IQA

No-reference – “blind” metrics for IQA

ITMO UNIVERSITY

How to evaluate image quality without a reference, as it is?

Ideas:

- Compute the variance of the pixels:

$$\mathcal{L}_{\mathrm{TV}}(\hat{I}) = \frac{1}{hwc} \sum_{i,j,k} \sqrt{(\hat{I}_{i,j+1,k} - \hat{I}_{i,j,k})^2 + (\hat{I}_{i+1,j,k} - \hat{I}_{i,j,k})^2}$$

- Compute sharpness/curtosis/skewness…

- Compute characteristics of the natural good looking images and evaluate the similarity between characteristics for the target image and estimated set.

What characteristics do we need to include in this set?

Natural scene statistics (NSS) approach

ITMO UNIVERSITY

Main idea: estimate parameters of the distorted images on different subbands using **Generalized Gaussian Distribution (GGD)** model after **Wavelet transformation**:

$$f(x|\alpha,\beta,\gamma) = \alpha e^{-(\beta|x-\mu|)^{\gamma}}$$

- the univariate generalized Gaussian density, where **μ** is the mean, γ is the shape parameter, and **α** and **β** are the normalizing and scale parameters given by:

$$\alpha = \frac{\beta\gamma}{2\Gamma(1/\gamma)} \qquad \Gamma(z) = \int_0^{\infty} t^{z-1}e^{-t}dt$$

$$\beta = \frac{1}{\sigma}\sqrt{\frac{\Gamma(3/\gamma)}{\Gamma(1/\gamma)}}$$

where σ is the standard deviation, and denotes the gamma function

Multivariate case:

$$f(x|\alpha,\beta,\gamma) = \alpha e^{-(\beta(\mathbf{x}-\mu)^{\mathbf{T}}\Sigma^{-\mathbf{1}}(\mathbf{x}-\mu))^{\gamma}}$$

Evaluate image NSS parameters (parameters of GGD distribution) in order to check how close the tested image's parameters are to desired set.

Reference image (DMOS = 0)

Blur distorted image (DMOS = 73.45)

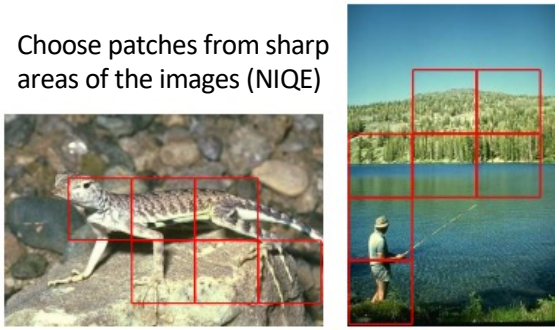
Saad et. Al., Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain, 2012

BRISQUE / NIQE

ITMO UNIVERSITY

➢ **Blind/Referenceless Image Spatial QUality Evaluator** (BRISQUE) utilizes an NSS model framework and locally normalized luminance coefficients and quantifies 'naturalness' using the parameters of the NSS model.
➢ **BRISQUE** introduces a new model of the statistics of pair-wise products of neighboring (locally normalized) luminance values.
➢ The parameters of this model further quantify the naturalness of the image (according to its similarity to NSS) and quantify quality in the presence of distortion.
➢ **NIQE** – improvement of the BRISQUE, based on more diverse dataset (distortion un-aware), use only sharp patches of the images for evaluation of parameters (need only ~100 images for convergence).

Choose patches from sharp areas of the images (NIQE)



	JP2K	JPEG	WN	Gblur	All
PSNR	0.825	0.876	0.918	0.934	0.870
SSIM	0.963	0.935	0.817	0.960	0.902
BRISQUE	*0.832*	*0.924*	*0.829*	*0.881*	*0.896*

Correlations between model's predictions and human opinion.
That's how to measure the metric's quality.

Multivariate Gaussian model (MVG)

$$f_X(x_1,\ldots,x_k) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\nu)^T\Sigma^{-1}(x-\nu)\right)$$

Difference between MVG distributions:

$$D(\nu_1,\nu_2,\Sigma_1,\Sigma_2) = \sqrt{\left((\nu_1-\nu_2)^T\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{-1}(\nu_1-\nu_2)\right)}$$

Quality is measured as the difference between parameters evaluated from the image and reference parameters from NSS.

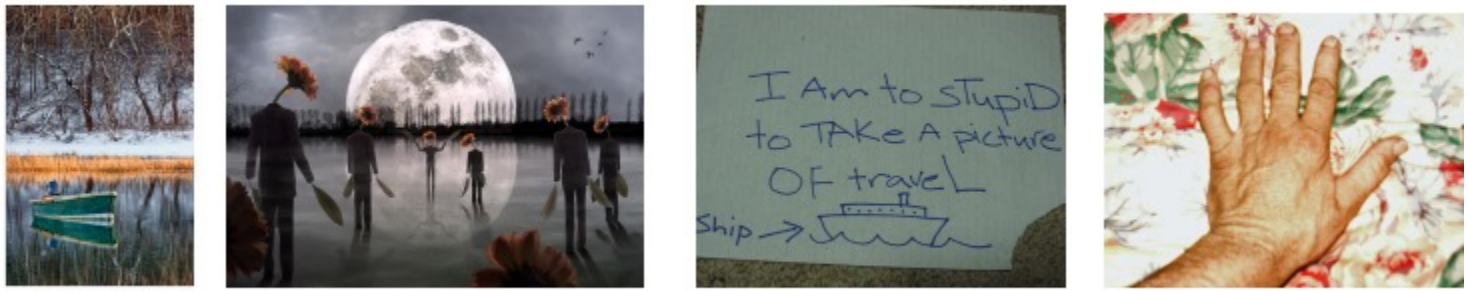
Mittal et.al, No-Reference Image Quality Assessment in the Spatial Domain 2012

NIMA - Neural Image Assessment

Deep CNN from Google Brain that is trained to predict which images a typical user would rate as looking good **(technically)** or attractive **(aesthetically)** – two separate models**.**

Trained on AVA, TID2013 and LIVE datasets with **subject rankings** of the images (>250 000 photos); Each photo was ranked with marks from 1 to 10;

Talebi et.al., NIMA: Neural Image Assessment, 2018



(a) 6.36 (±1.04) (b) 7.84 (±2.08) (c) 2.62 (±2.15) (d) 3.12 (±1.28)

Video Quality Assessment methods

ITMO UNIVERSITY

- One can use IQA for the frames of the video sequence independently in NR or FR (if there exist reference) manner
- But: large motions decrease the performance of classic IQA approaches;
- Most of the methods for VQA use motion estimation info for the fusion within the frames;
- **MMF – Multi-Metrics Fusion** approach perform best comparing to others (like meta-learning approaches)
  - get scores from different IQA (or VQA) metrics;
  - get scores from humans;
  - compute the differences;
  - take these differences as input features and train the regressor model.

Also:

- **Pixel-based approaches** (PSNR-HVS (PSNR + Human Visual System), PSNR+HVS+M (masking), VSSIM (SSIM for frames weighted according to the motions);
- **VMAF** – **Netflix Video Multi-Method Assessment Fusion** – full-reference metric; considers up/down scaling and relies on Visual Information Fidelity (VIF) and Detailed Loss metric (DLM) + temporal difference between consecutive frames. The final score is from SVM regressor, trained on subjective test by Netflix.
- Psycho-physiological approaches, etc.

Conclusion

ITMO UNIVERSITY

1. Quality assessment for generative/restoration models is tough..
2. Deep-learning based metrics usually perform better than statistics based approaches;
3. “Quality of quality metric” is usually measured as correlation coefficient between metrics predictions and human opinion scores on the benchmark datasets;
4. For IQA there is no “ideal” metric that would be good in every case – metrics works better in distortion-aimed regime;
5. No-reference IQA is usually based on NSS model;
6. “Aesthetics” of the image/video is expert-aware and subjective parameter.


ITMO UNIVERSITY
Saint Petersburg, Russia
Thank you!