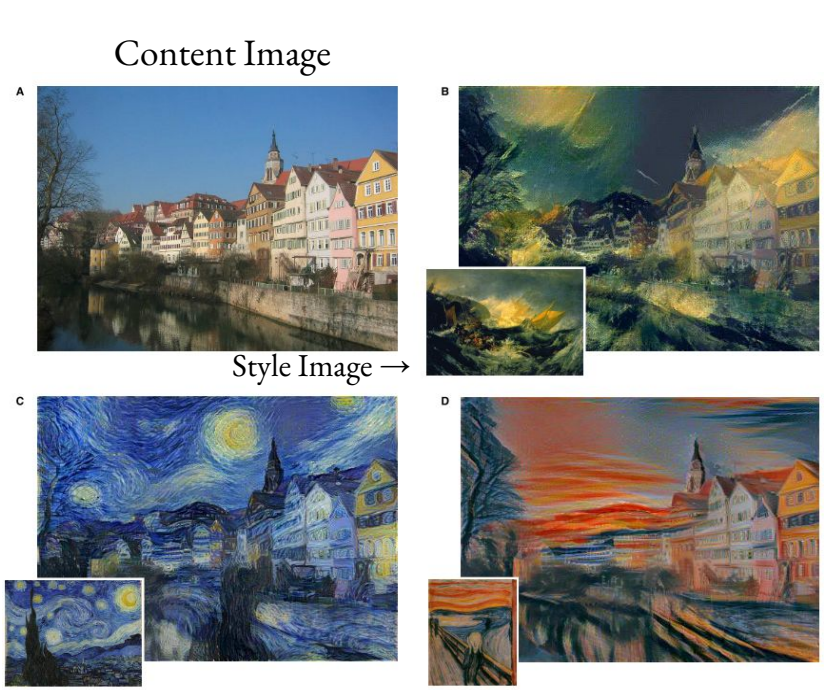
Quality Assessment for Style Transfer

Neural Style Transfer



Artistic Style Transfer

Photorealistic Style Transfer

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Style Transfer using GANs

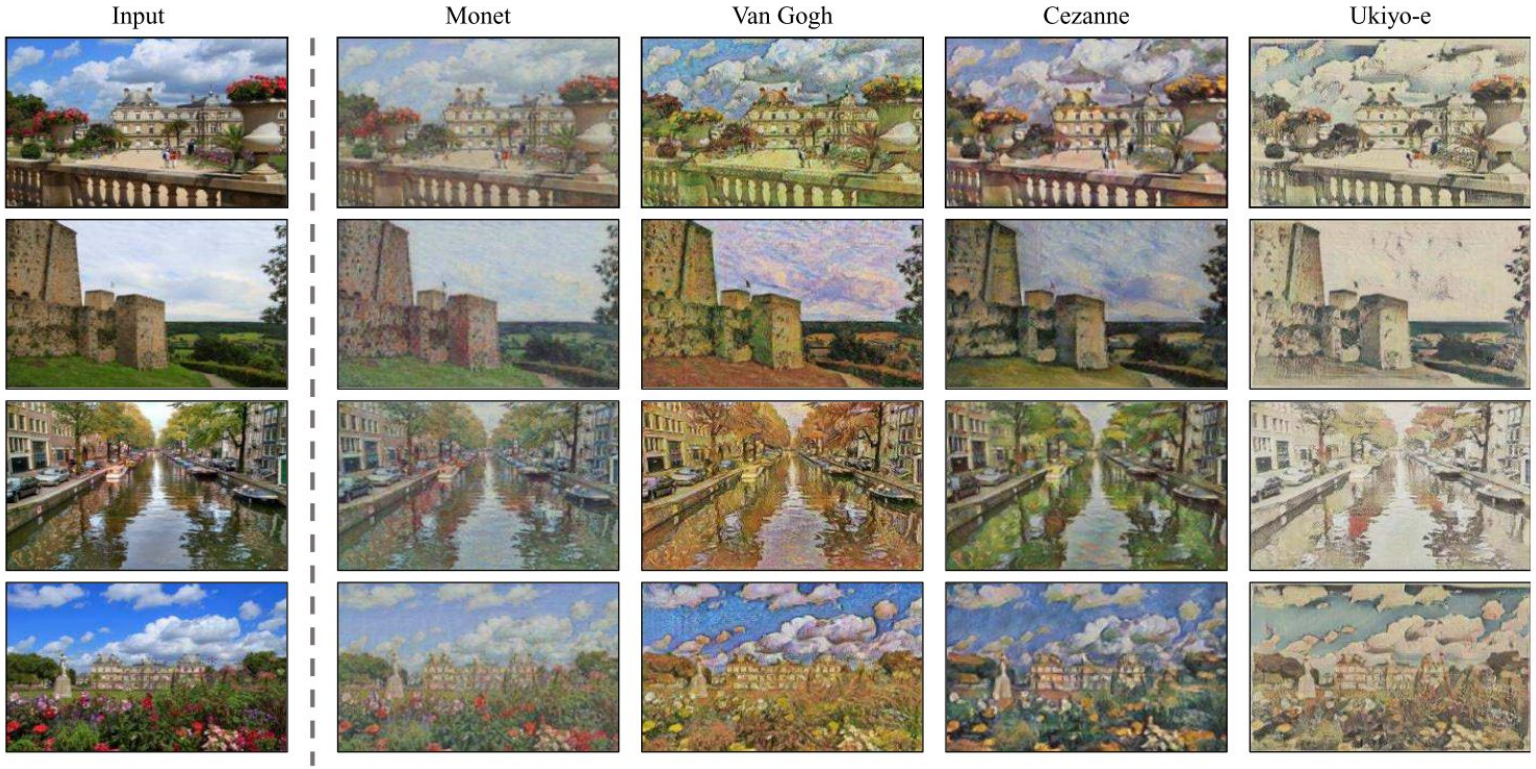


Figure 8: We transfer input images into different artistic styles. Please see our website for additional examples.

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision.

Photo-realism in Style Transfer

The quantification of what is photo-realistic is a challenging task.
Here we show three quality metric that could be used to measure the photorealism in the style transfer.

- User Study.
- Reference based quality assessment: PieAPP.
- No-reference quality assessment: NIMA.

User Study

- The goal is to get the human opinion score for input images.
- The user study can be in the form of Reference based Quality Assessment or No-Reference Quality Assessment.
- A simple user study will allow to take votes or score for the input images.
- In case of voting, the image with the highest vote will be considered best.
- In case of scoring based user study, the images with high score would be considered best.

PieAPP

Reference Based Quality Assessment

- PSNR and SSIM are the example for reference based quality assessment. However, they are less useful in the setting of style transfer and photo-realism.

Introduction to PieAPP

- PieAPP stands for Perceptual Image-Error Assessment through Pairwise Preference.
- Addresses the challenge of estimating perceptual differences between images.
- Utilizes pairwise preference to align with human perception more accurately

image *A* | reference image *R* | image *B*

Which image, A or B, is more similar to the reference R?

Prashnani, E., Cai, H., Mostofi, Y., & Sen, P. (2018). Pieapp: Perceptual image-error assessment through pairwise preference. *Computer Vision and Pattern Recognition.*

Introduction to PieAPP

- The task is to predict perceptual image error like human observers.
- PieApp proposes a new, large-scale dataset labeled with the probability that humans will prefer one image over another.

image *A* | reference image *R* | image *B*

Which image, A or B, is more similar to the reference R?

Conceptual Foundation

- Compares two distorted images, A and B, against a reference image R.
- Collects human judgments on which distorted image is closer to R.
- Encodes judgments as probability p_{AB}, the likelihood of preferring A over B.

Bradley-Terry Model for Pairwise Preference

- The model provides a probabilistic framework for human preferences.
- Probability of preference, p_{AB}, is modeled using the sigmoid function:

$$p_{AB} = \frac{1}{1 + e^{(s_A - s_B)}}$$

- s_A and s_B are perceptual error scores of images A and B, respectively.

Deep Learning Implementation

- Implements f through a Deep Convolutional Neural Network (DCNN).
- Feature Extraction (FE): Computes features for input patches from images A, B, and R.
- Score Computation (SC): Uses feature differences to compute patch-wise weights and errors.

Learning and Optimization

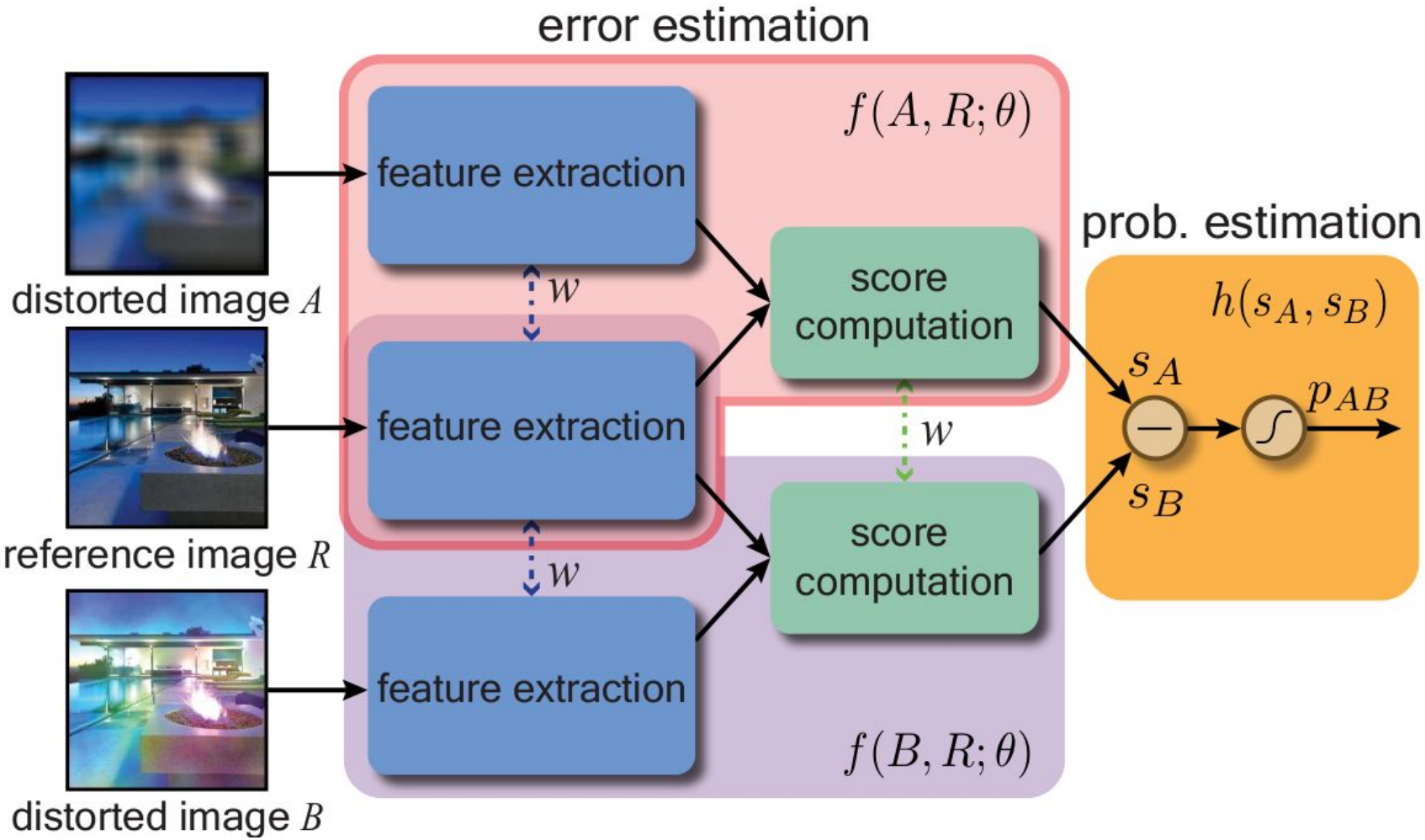
- $f(\cdot, \cdot; \theta)$ maps distorted and reference images to perceptual error scores.
- Parameters θ are learned to make f predict human-like perceptual errors.
- Employs Bradley-Terry model to estimate pairwise preferences accurately.

Optimization Framework

- Objective: Minimize the difference between model predictions and human judgments.
- Formulated as:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{T} \left(h\left(f(A_i, R_i; \theta), f(B_i, R_i; \theta)\right) - p_{AB_i}\right)^2$$

- T is the total number of training pairs, and p_{AB_i} is the ground-truth probability of preference.


error estimation
$f(A, R; \theta)$
feature extraction
w
feature extraction
w
feature extraction
$f(B, R; \theta)$
score computation
w
score computation
prob. estimation
$h(s_A, s_B)$
s_A
s_B
p_{AB}
distorted image A
reference image R
distorted image B

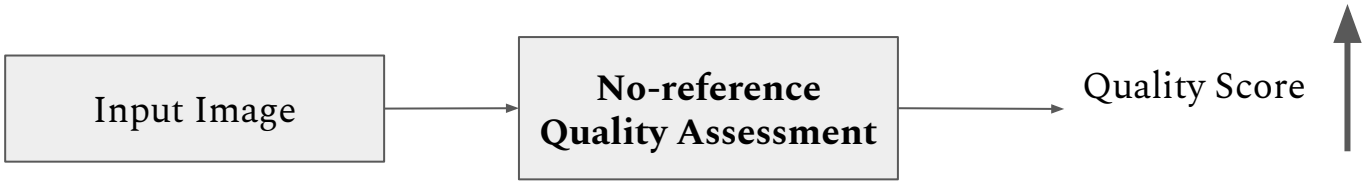
Results and Contributions

- PieAPP significantly outperforms existing metrics in predicting perceptual image error.
- Demonstrates the effectiveness of pairwise preference learning in image quality assessment.
- Offers a robust framework for future advancements in perceptual image error assessment.

NIMA

Talebi, H., & Milanfar, P. (2018). NIMA: Neural image assessment. *IEEE Transactions on Image Processing.*

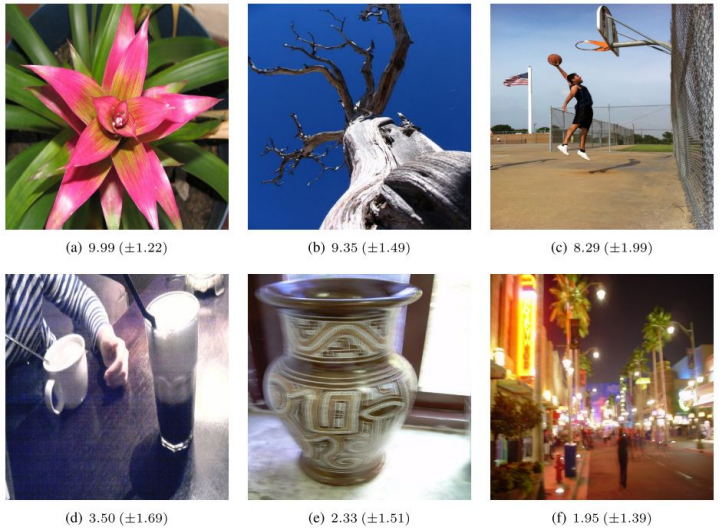
No-reference Quality Assessment



- The reference based quality assessment compares the input image with a reference image. However, in No-reference quality assessment, we have only input image and no-reference image is given.
- No-reference quality assessment is also known as blind approaches.

Introduction to NIMA

- Objective: Develop a CNN-based system to predict image quality and aesthetics.
- Key Innovation: Predicts the distribution of human opinion scores the average.
- The method should give the quality score even if the input image is corrupted, e.g., blurred image.



(a) 9.99 (±1.22) (b) 9.35 (±1.49) (c) 8.29 (±1.99)

(d) 3.50 (±1.69) (e) 2.33 (±1.51) (f) 1.95 (±1.39)

No-reference quality score (NIMA) of images.

Convolutional Neural Network (CNN) Adaptation

- Base Architectures: VGG16, Inception-v2
- Modification: Replace the final layer with 10 outputs for scores 1-10.
- Training: Pretrained on ImageNet, fine-tuned on quality assessment datasets

Model	*Million Parameters*	*Billion Flops*	*CPU Timing (ms)*	*GPU Timing (ms)*
NIMA(MobileNet)	3.22	1.29	30.45	20.23
NIMA(Inception-v2)	10.16	4.37	70.49	39.11
NIMA(VGG16)	134.30	31.62	150.34	85.76

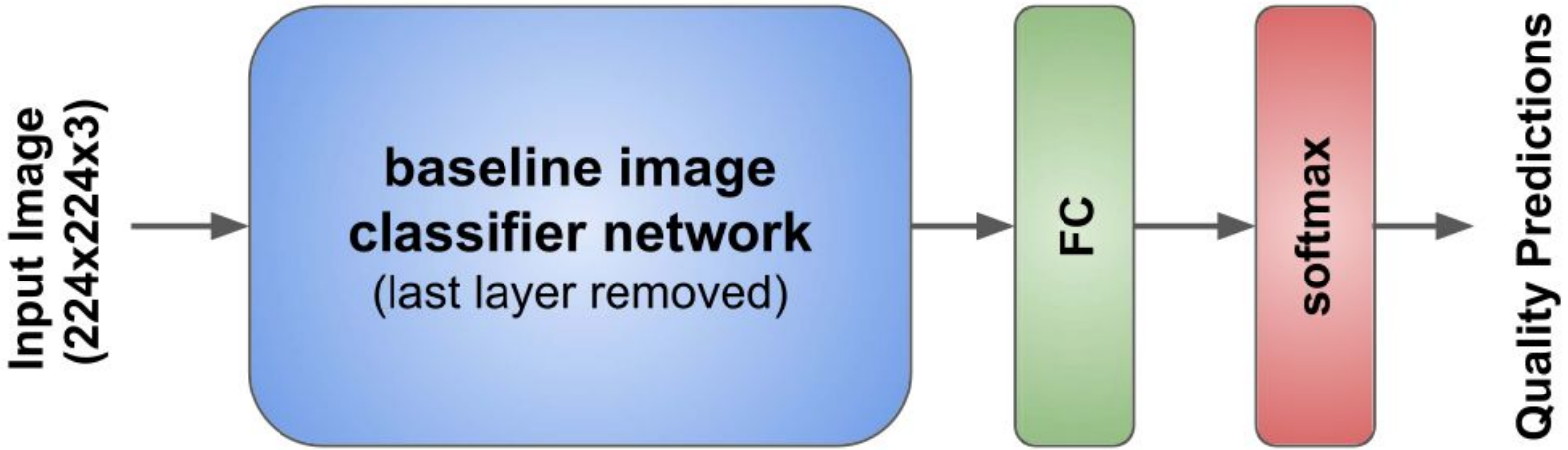
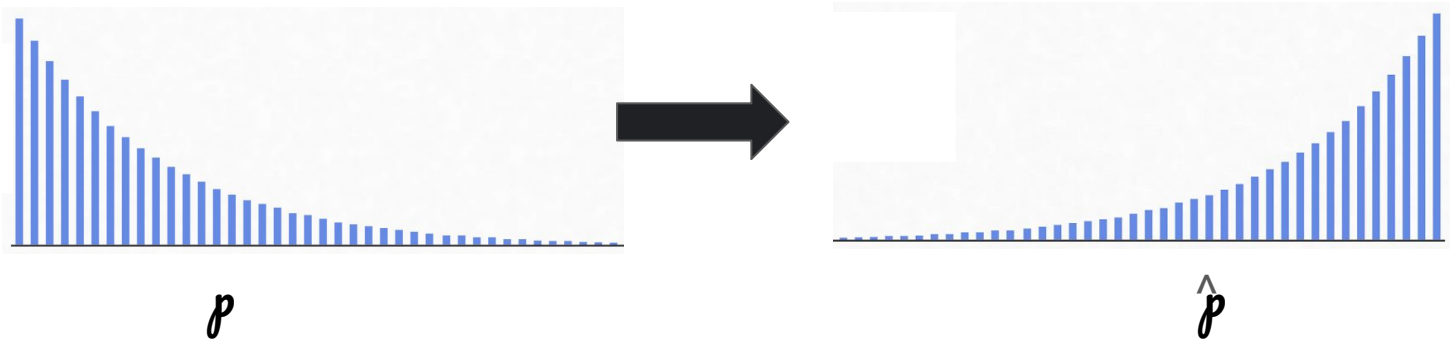


Fig. 8. Modified baseline image classifier network used in our framework. Last layer of classifier network is replaced by a fully-connected layer to output 10 classes of quality scores. Baseline network weights are initialized by training on ImageNet dataset [15], and the added fully-connected weights are initialized randomly.

Interesting Problem



Aim is to transport pile of sand with minimum effort.

Calculating Cumulative Distribution Functions

- CDF Definition:
  - $CDF_p(k) = \sum_{i=1}^{k} p_i$
  - $CDF_{\hat{p}}(k) = \sum_{i=1}^{k} \hat{p}_i$

Earth Mover's Distance (EMD) Loss Function

- Measure minimal work to transform one distribution into another.
- Key Formula:

$$EMD(p, \hat{p}) = \left(\frac{1}{N} \sum_{k=1}^{N} |CDF_p(k) - CDF_{\hat{p}}(k)|^r \right)^{\frac{1}{r}}$$

- Notations:
  - $p, \hat{p}$: Ground truth and predicted distributions.
  - N: Number of score buckets.
  - CDF: Cumulative distribution function.
  - r-norm distance

Squared EMD for Simplification

- Squared EMD Loss:

$$\text{Squared EMD}(p, \hat{p}) = \frac{1}{N} \sum_{k=1}^{N} (CDF_p(k) - CDF_{\hat{p}}(k))^2$$

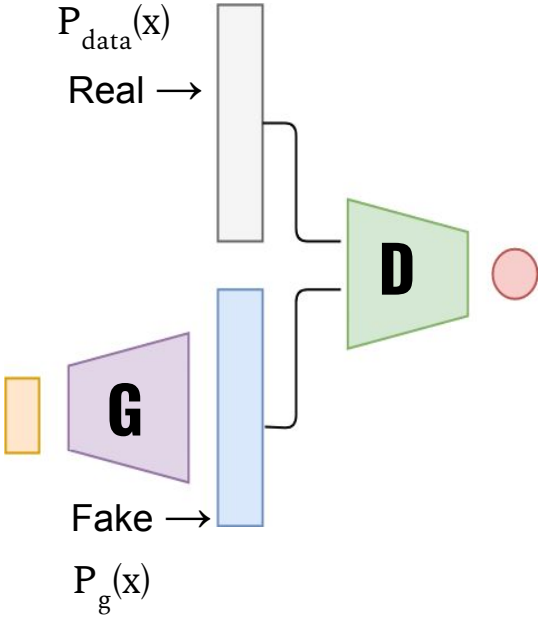
- Advantages:
  - Simplifies optimization.
  - Penalizes the difference between predicted and ground truth distributions.

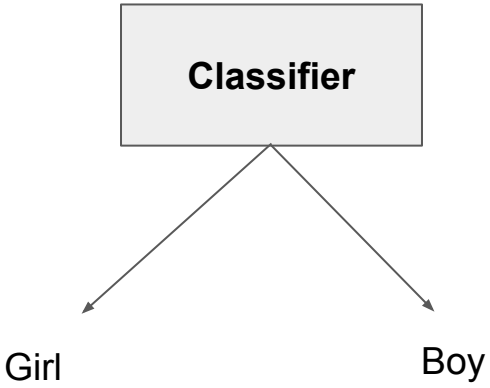
Conclusion and Future Work

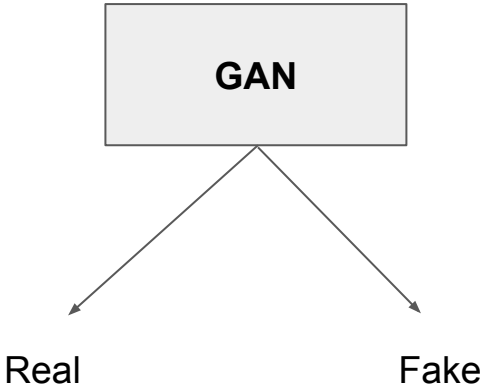
- NIMA Achievements:
  - Accurately predicts human opinion distribution on image quality.
  - Uses EMD to effectively capture the nuances of human perception.
- Future Directions:
  - Explore real-time applications.
  - Extend to other image enhancement applications.

Quality Assessment for Generative Models

Why Evaluation of Generative models is hard?







How to derive a metric for Evaluation for Generative models

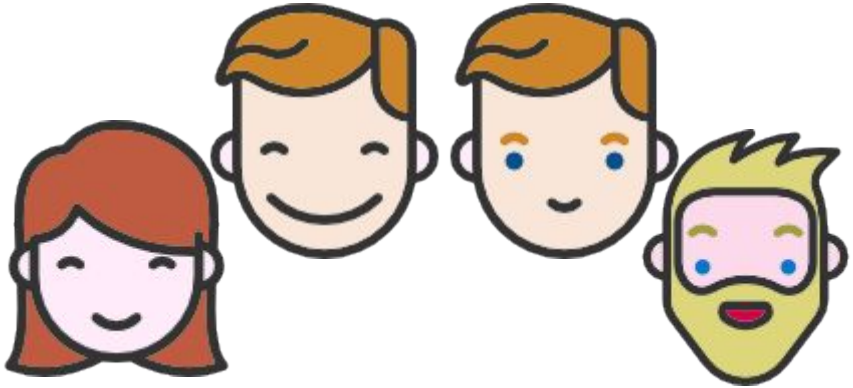
Evaluation Properties

Fidelity

- Quality if images produced by the model

Diversity

- Variety of images produced by the model



https://github.com/NVlabs/stylegan

Properties for GANs Evaluation

Fidelity

(measures image quality)

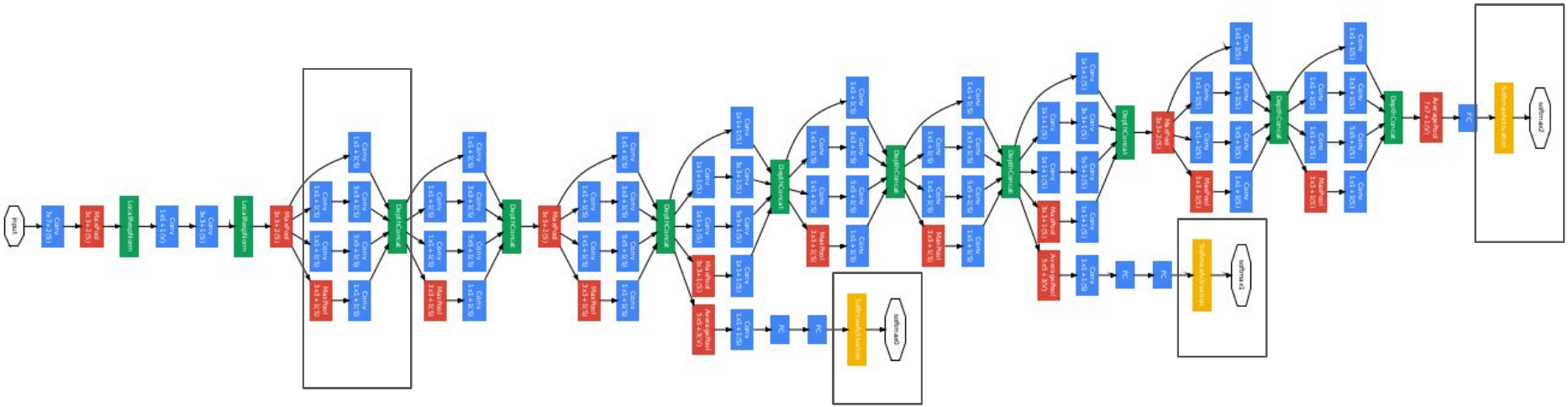
Real

Fake

Diversity
(measures variety)

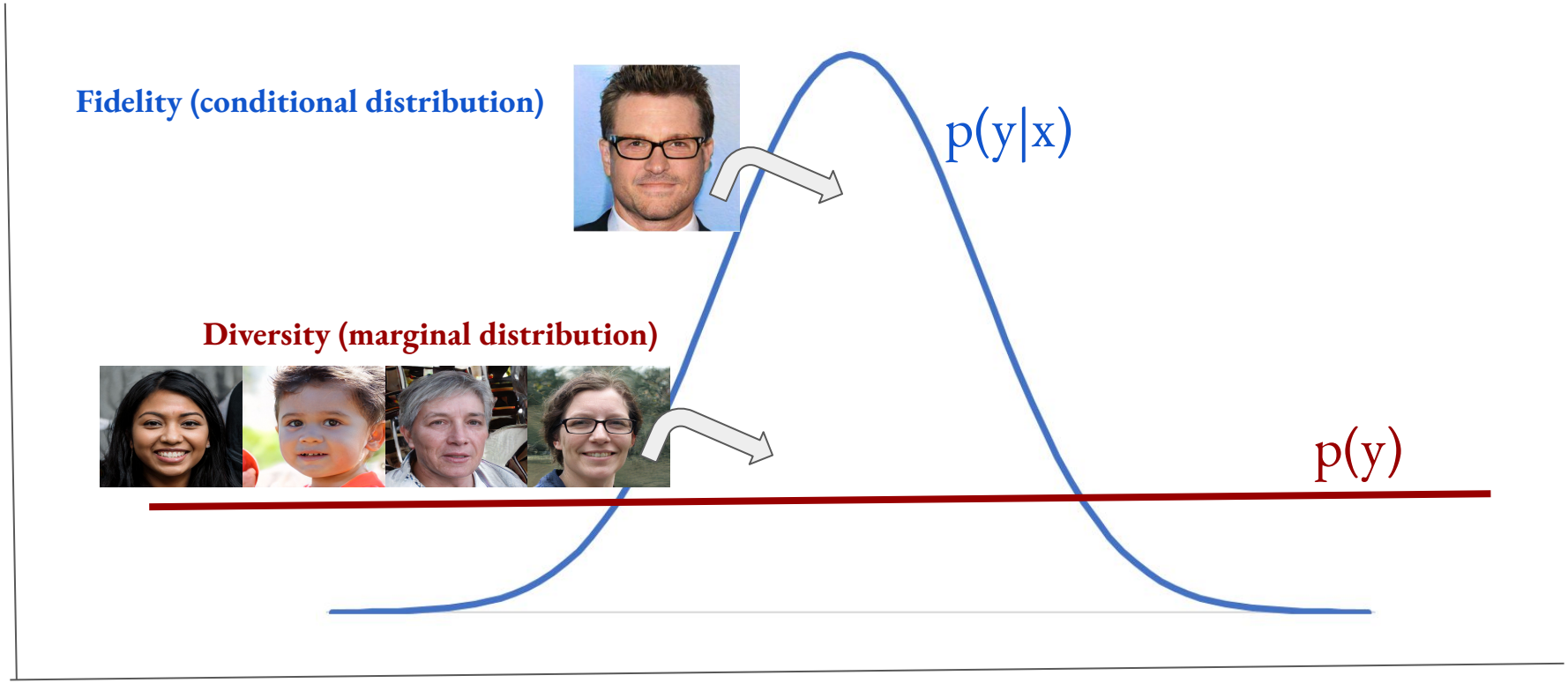
Inception Score

Inception Net (Main Components)



- **Inception Layer**: The multi-pathway convolutional blocks that enable the network to learn complex features using fewer parameters.
- **Auxiliary classifiers**: At intermediate layers of the network to encourage intermediate feature learning.

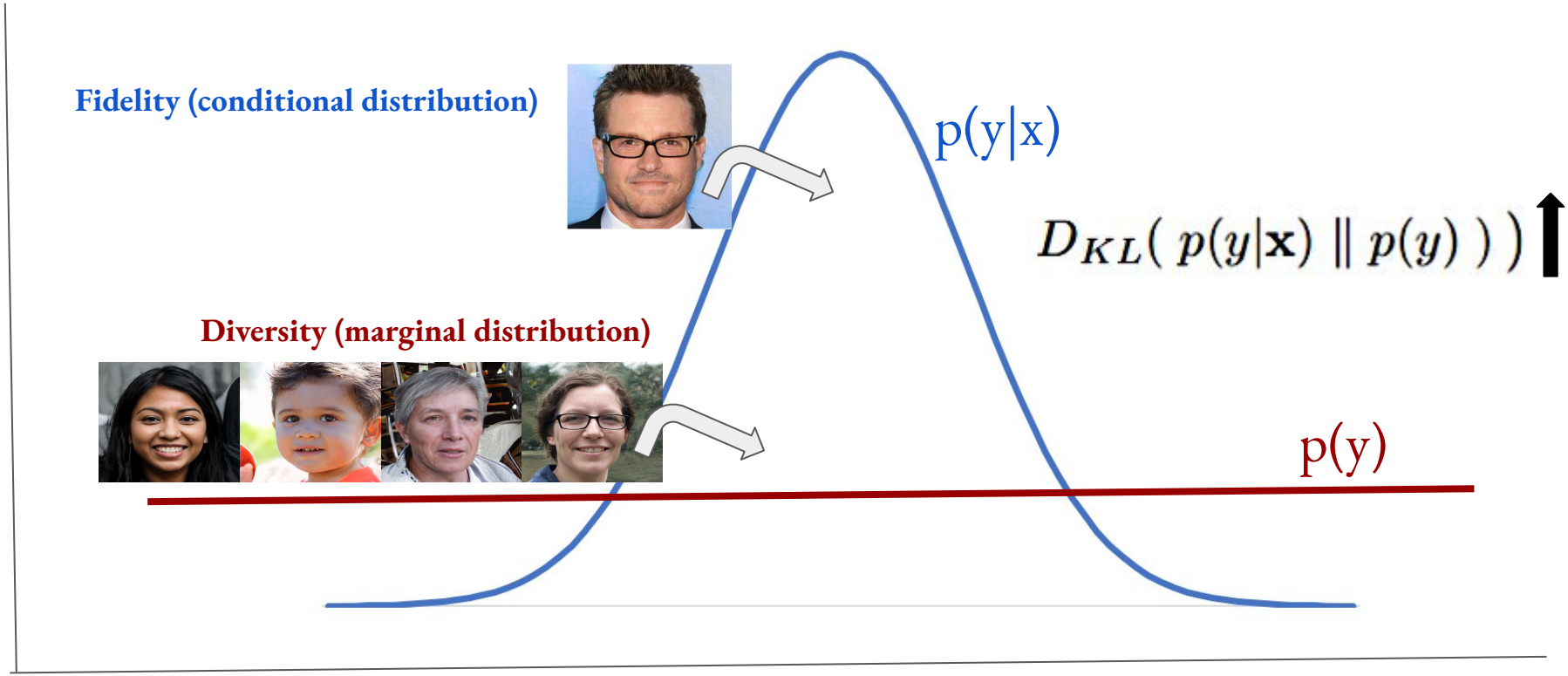
Intuition


Fidelity (conditional distribution)
p(y|x)
Diversity (marginal distribution)
p(y)

Detailed Components of the Inception Score

- Conditional Distribution ($p(y|x)$): How confidently a pre-trained classifier (e.g., Inception model) can assign a class to each generated image.
- Marginal Distribution ($p(y)$): Aggregate class distribution across the entire set of generated images, calculated by averaging $p(y|x)$ over all images.

Intuition


Fidelity (conditional distribution)
p(y|x)
$D_{KL}(p(y|\mathbf{x}) \| p(y)))$
Diversity (marginal distribution)
p(y)

Detailed Components of the Inception Score

- Conditional Distribution ($p(y|x)$): How confidently a pre-trained classifier (e.g., Inception model) can assign a class to each generated image.
- Marginal Distribution ($p(y)$): Aggregate class distribution across the entire set of generated images, calculated by averaging $p(y|x)$ over all images.
- Expectation and KL Divergence: The expectation quantifies the average discrepancy (via KL divergence) between each image's classifiability and the overall class distribution.

Inception Score

- The inception score (IS) is the most widely used scoring algorithm for GANs.
- IS is computed using a pre-trained inception V3 network (trained on Imagenet) to extract the features of both generated and real images.

$$\mathrm{IS}(G) = \exp\left(\mathbb{E}_{\mathbf{x}\sim p_g} D_{KL}(p(y|\mathbf{x}) \parallel p(y)) \right)$$

Here, x represents a sample, sampled from distribution P_g.
P(y|x) is the conditional distribution
P(y) is the marginal distribution.

- A perfect generative model will have IS equal to the number of classes.

Calculating Inception Scores

- Step-by-Step Process: Generate images, classify each using the Inception model, calculate $p(y|x)$ and $p(y)$, compute the KL divergence for each image, and finally calculate the score by taking the exponential of the average KL divergence.

- Interpretation: A higher IS indicates that on average, images are both more distinct (high D_{KL}) and more varied across classes.

The Significance of High Inception Scores

- Indicator of Model Performance: High IS suggests that the model produces diverse, yet recognizable images.
- Quality vs. Diversity: A balance between high confidence in image classification and a wide spread of generated classes.
- Use in Model Development: Provides targets for GAN developers to optimize towards for better image generation.

Shortcomings of Inception Score

- Using a pre-trained inception model, which may not capture all features.
- Computationally heavy (need pre trained classifier).
- IS shows a good level of accuracy even when the model generates one image per class, which means the model may lacks diversity.
- Inception score is like a No-reference quality assessment, as we do not use the real data samples when measuring the quality of the generated images.

FID Score

What is FID Score?

- FID Score Explained: A statistical measure to compare the similarity between two distributions of images – real and generated.
- Significance: Offers a quantitative measure for both the diversity (variety) and fidelity (realism) of generated images.
- Benchmark: Lower FID scores indicate better performance, signifying that generated images are both realistic and varied.

Extracting Features for FID

- Role of Inception-v A deep convolutional neural network used to extract meaningful features from images, avoiding manual feature specification.
- Feature Layer: The choice of layer (typically the pooling layer before the final classification) in Inception-v3 is crucial for capturing high-level abstractions of images.
- Process: Both sets of images (real and generated) are passed through the Inception-v3 model to obtain their feature vectors, setting the stage for further statistical analysis.

Statistical Modeling of Image Features

- Assumption: The feature vectors extracted from the images are modeled as being drawn from multivariate Gaussian distributions.
- Parameters Estimation: Calculation of the mean (μ) and covariance (Σ) of these distributions for both real and generated images, which encapsulate the core characteristics of each dataset.
- FID is computed between means and covariances matrices of activations from an **intermediate layer** of a pretrained Inception network.

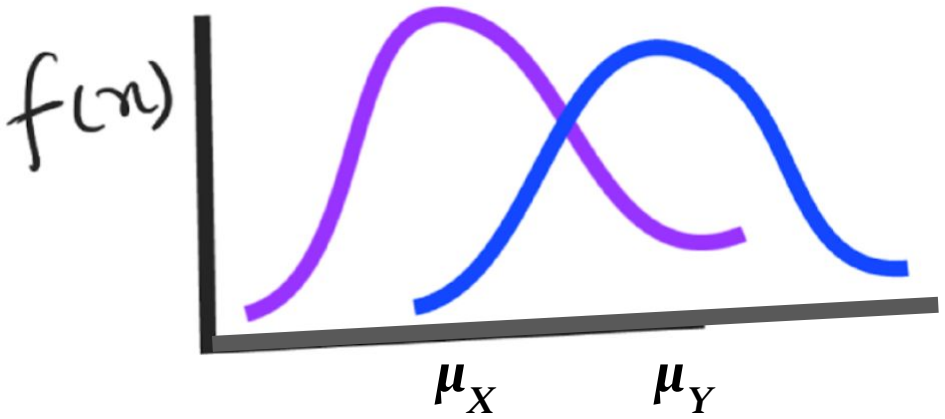
Normal Distributions

$$X \sim N(\mu, \sigma^2)$$

A normal distribution is a type of continuous probability distribution defined as follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

μ (mu) is the mean.
σ (sigma) is its standard deviation.
σ^2 (sigma^2) variance of the distribution.



Frechet Distance for Normal Distributions

$$X \sim N(\mu, \sigma^2)$$

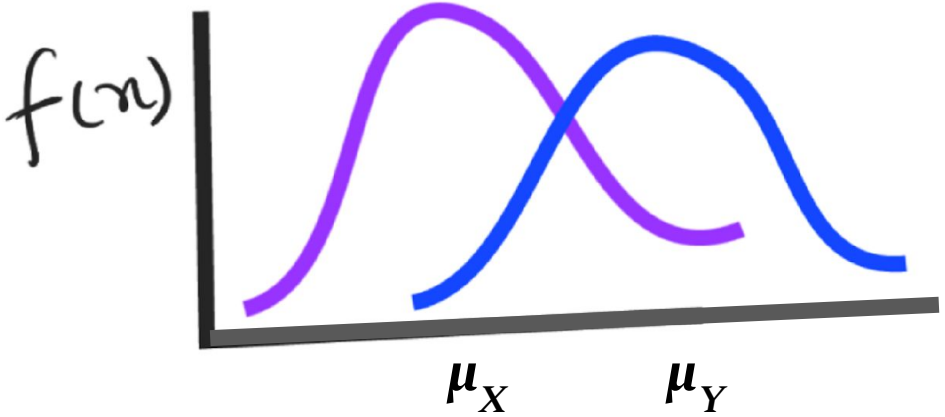
A normal distribution is a type of continuous probability distribution defined as follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

μ (mu) is the mean.
σ (sigma) is its standard deviation.
σ^2 (sigma^2) variance of the distribution.

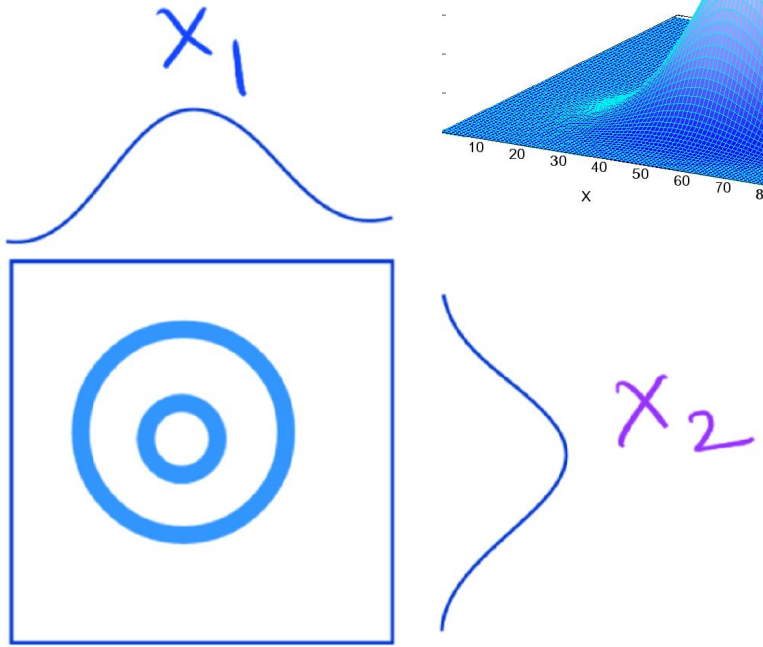
Univariate:

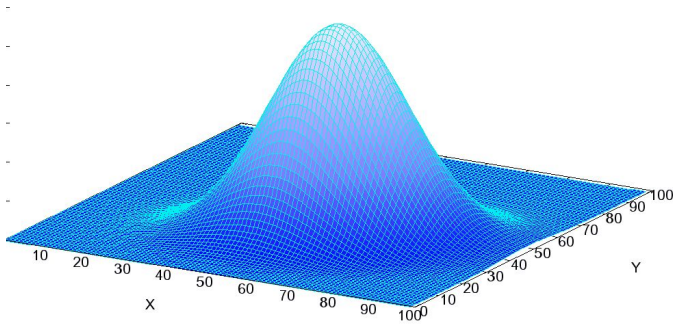
$$d(x, y) = (\mu_x - \mu_y)^2 - (\sigma_x - \sigma_y)^2$$



Bivariate Normal Distribution

To understand the multivariate normal distribution it is helpful to look at the **bivariate normal** distribution.



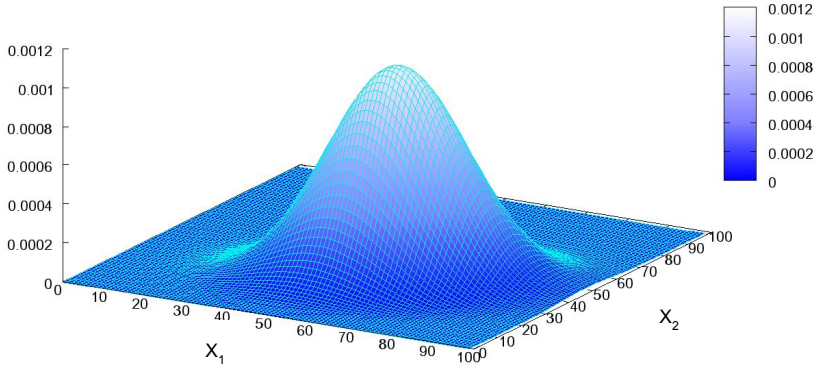


https://en.wikipedia.org/wiki/Multivariate_normal_distribution

Bivariate Normal Distribution (Parameters)

The probability density function of the univariate normal distribution contained **two parameters**: μ and σ.
With two variables, say X_1 and X_2, the function will contain **five parameters**: two means μ_1 and μ_2, two standard deviations σ_1 and σ_2 and the product moment correlation between the two variables, ρ.

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right]$$



https://en.wikipedia.org/wiki/Multivariate_normal_distribution

Bivariate Normal Distribution

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right]$$

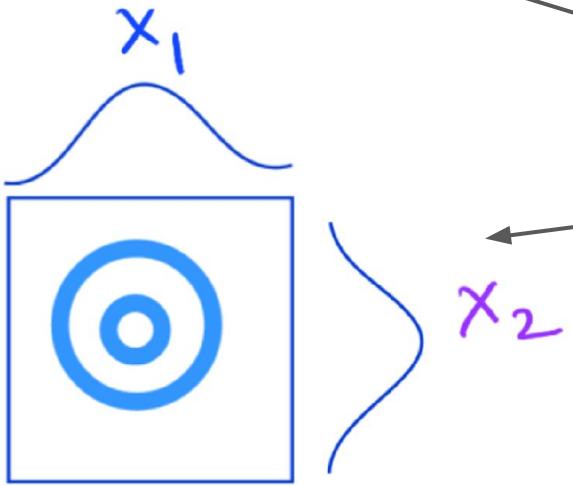
Covariance Matrix

$$\Sigma = \begin{array}{c} \\ X_1 \\ X_2 \end{array} \begin{array}{c} \begin{array}{cc} X_1 & X_2 \end{array} \\ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{array}$$

Covariance Matrix

Here, we have the variances for the two variables on the diagonal and on the off-diagonal we have the covariance between the two variables.

This covariance is equal to the correlation times the product of the two standard deviations.



X_1 and X_2 are independent variables

FID Score

Univariate: $FID = (\mu_x - \mu_y)^2 - (\sigma_x - \sigma_y)^2$

Suppose we have two random variables X_r and X_g parameterized as:

$$X_r \sim N(\mu_r, \Sigma_r), X_g \sim (\mu_g, \Sigma_g)$$

Multivariate ?

FID Score

Univariate: $FID = (\mu_x - \mu_y)^2 - (\sigma_x - \sigma_y)^2$

Suppose we have two random variables X_r and X_g parameterized as:

$$X_r \sim N(\mu_r, \Sigma_r), X_g \sim (\mu_g, \Sigma_g)$$

Multivariate FID is a distance measure between means and covariances matrices. Therefore given the two multivariate normal distributions for X_r and X_g the FID is defined as follows.

$$FID = \|\mu_r - \mu_g\|^2 + T_r(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2})$$

Note that X_r and X_g models the distributions captured from an intermediate layer of a pretrained Inception network for real samples ‘r’ and generated samples ‘g’.

FID Score

$$FID = \|\mu_r - \mu_g\|^2 + T_r(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2})$$

- Interpretation: This formula integrates both the first-order (mean difference) and second-order (covariance difference) moments of the distributions, offering a comprehensive view of similarity.

FID Score

The FID between the real images, r, and generated images, g, is defined as follows:

$$FID = \|\mu_r - \mu_g\|^2 + T_r(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2})$$

- The lower the FID score, the better the model, and the more able it is to generate more diverse images with higher quality.
- Intuitively, by computing the means and covariances, the FID is fitting a multivariate Gaussian to the activations.

FID Score

- Similar to the Inception score, FID uses a pretrained Inception network.
- FID is computed on both real data and generated data.
- A perfect generative model will have an FID score of zero.

Shortcomings of FID

- Using a pre-trained inception model, which may not capture all features.
- Compare distributions using only mean and covariance (limited)
- Computationally heavy (need pre trained classifier)
- We need to train on many samples to be able to be sure on the quality assessment from FID

Inception Score and FID Score

- FID score is preferable over the Inception score because it is robust to noise and that it can easily measure the diversity of the images.
- The FID can be used to detect a low variety within each mode of the distribution of images produced by the generator. While the Inception score will be high for generators that, for example, output one sample per mode but lack within the mode variety, the score will be low, that is, a large distance, for the FID.

References

- Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play. By David Foster. O'Reilly Media.