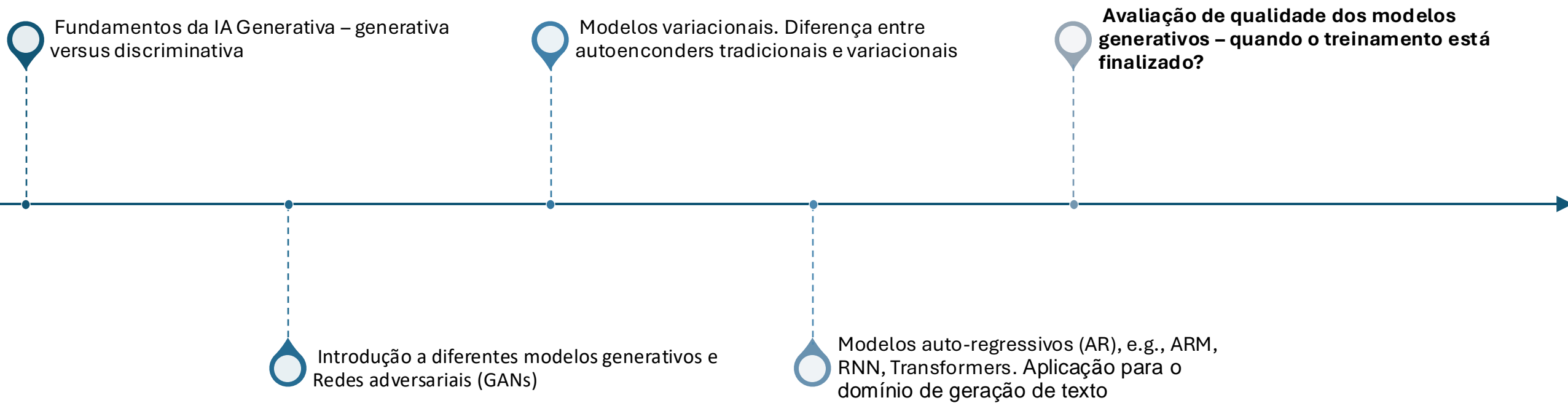


Universidade Federal do Rio Grande do Norte
Instituto Metr pole Digital
Bacharelado em Intelig ncia Artificial (BIA)

Avalia  o de qualidade em modelos generativos

Prof. Dr. Andr  Fonseca
IA Generativa (IMD3004) – Aula 05

Retrospectiva



E, agora...

Como avaliar qualidade de um modelo generativo?



Modelo discriminativo
(e.g. um classificador)

Gato
Panda-
vermelho
Cachorro
Raposa
Saco de torrada



Nessa categoria o modelo possui uma "**chave de resposta**", logo podemos comparar a performance do treinamento de maneira objetiva.

Vetor ou espaço latente



Model generativo (e.g.
GAN)

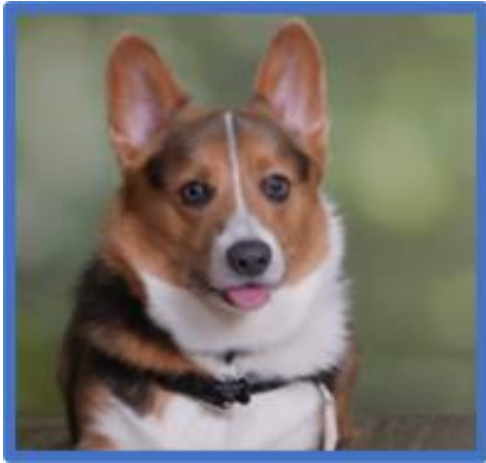
Modelo para geração de imagens
de cachorros



Como avaliar a performance do modelo?

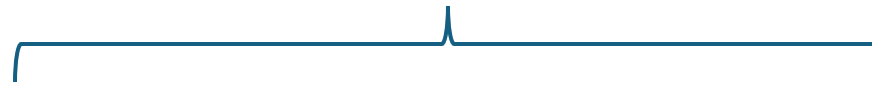
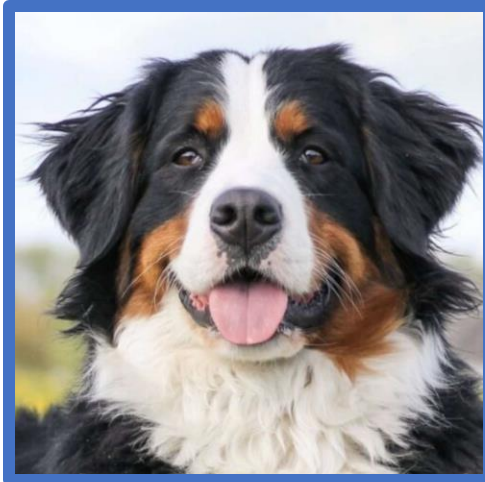
Imagens

Propriedades para avaliar:



Fidelidade

Qualidade de
imagem

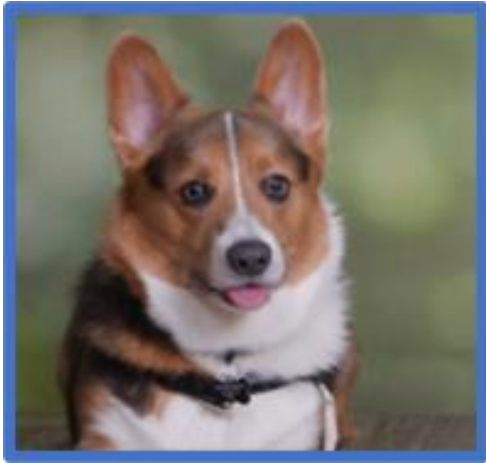


Diversidade

Variedade das imagens
(e.g., diferentes raças de
cachorro)

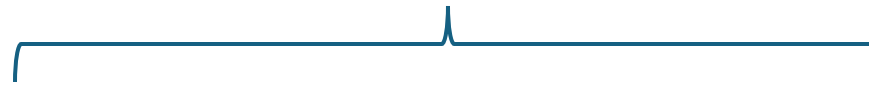
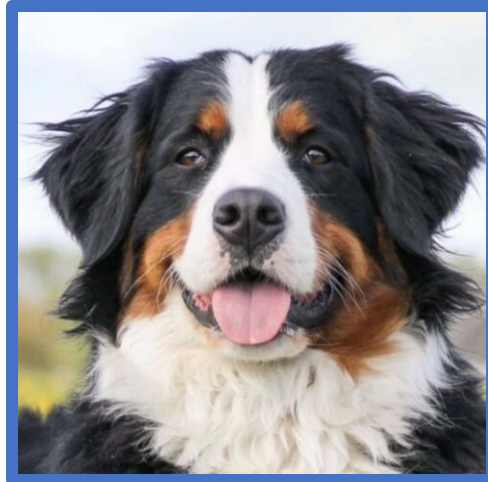
Imagens

Propriedades para avaliar:



Fidelidade

e.g., Fréchet Inception Distance, Inception Score e Structural Similarity Index



Diversidade

e.g., Fréchet Inception Distance e Diversity score

Obviamente, que tal problemática não se aplica unicamente a geração de imagens...

Texto

Propriedades para avaliar:

Fluência

Entrada: "Descreva um cachorro brincando no parque."

Resposta: "Cachorro brincar no parquinho feliz. Saltar correr no parque ele muito alegremente."

Perplexidade

Adequação

Entrada: "Quem foi Albert Einstein?"

Resposta: "Albert Einstein era muito famoso porque adorava maçãs e trabalhava como fazendeiro."

e.g., BLEU, ROUGE

Repetição

Entrada: "Descreva uma viagem ao espaço."

Resposta: "O espaço é vasto e bonito. O espaço é vasto e bonito. O espaço é vasto e bonito."

Análise de n-gramas
únicos

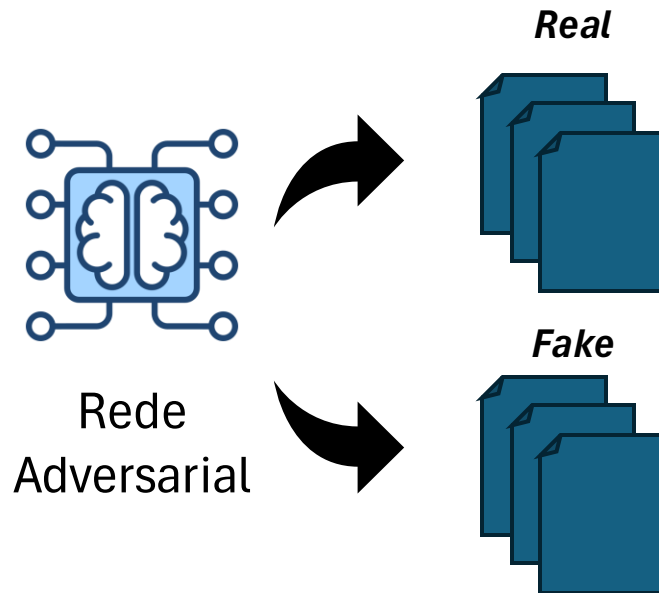
Hoje vamos falar
de...

Métodos de avaliação de **imagens!**

~ Com ênfase em GANs ~

Princípios para avaliação

1) Após o treinamento



Real



Fake



Categorização definida
pelo discriminador

1. A avaliação do desempenho do modelo depende da comparação entre amostras reais e falsas.

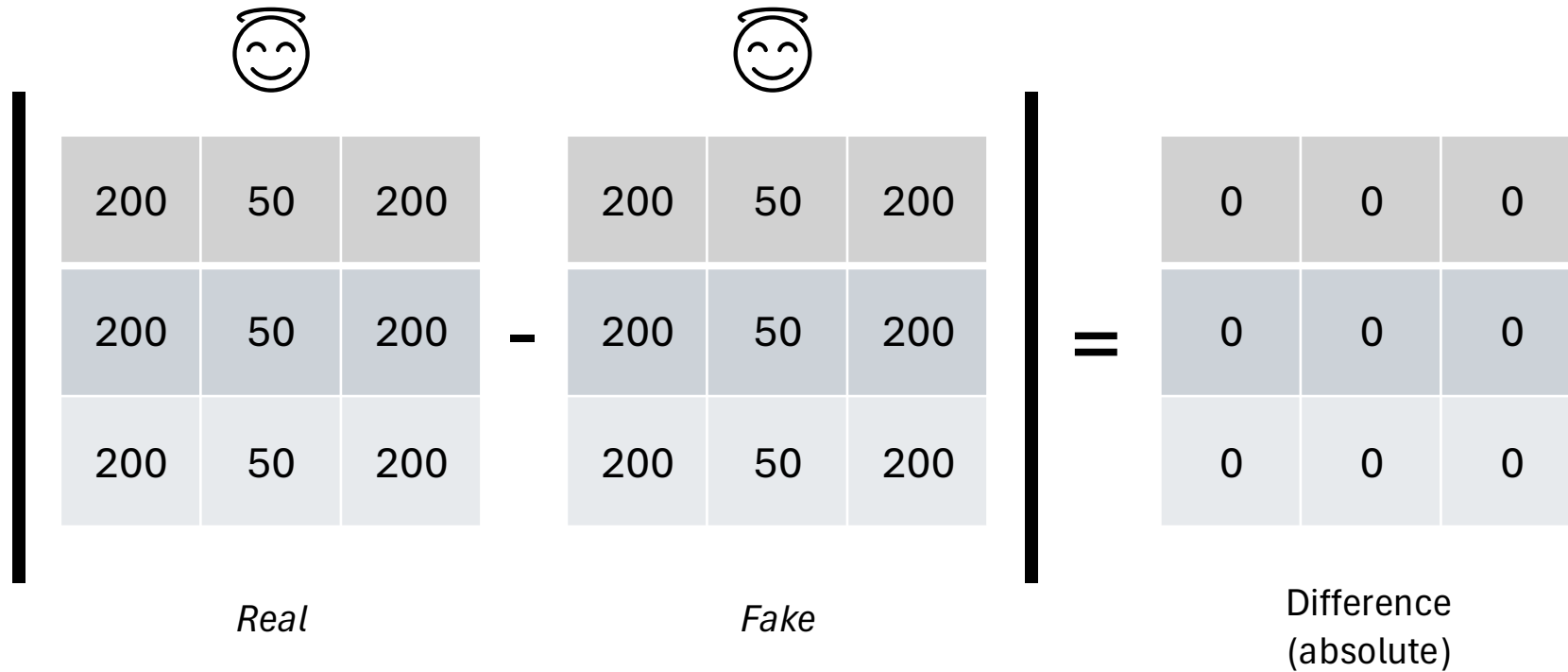
2. Ainda que intuitivo, o processo pode ser extremamente trabalhoso, sobretudo se realizado **visualmente!**

3. Portanto, precisamos de formas inteligentes de avaliar o modelo.

Técnicas de comparação de imagens

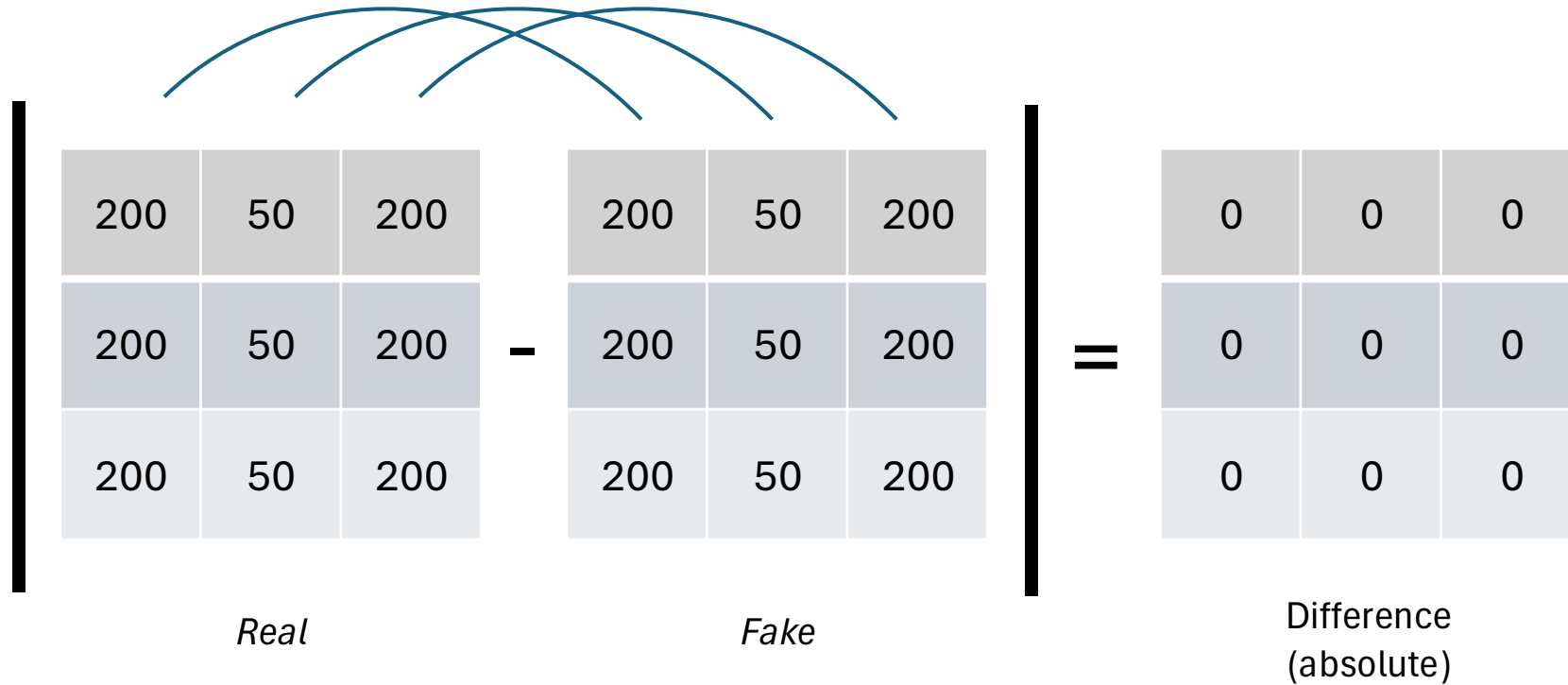
- **Pixel distance:** Esta é a abordagem mais direta, onde se mede a diferença entre os valores de pixel de duas imagens.
 - Erro quadrático médio (Mean Square Error, MSE)
- **Feature distance:** Em vez de comparar diretamente pixels, a abordagem compara as representações extraídas de imagens.
 - Utilização de redes neurais pré-treinadas (Inception-v3) para extrair recursos de alto nível das imagens (por exemplo, número de olhos).

Pixel distance



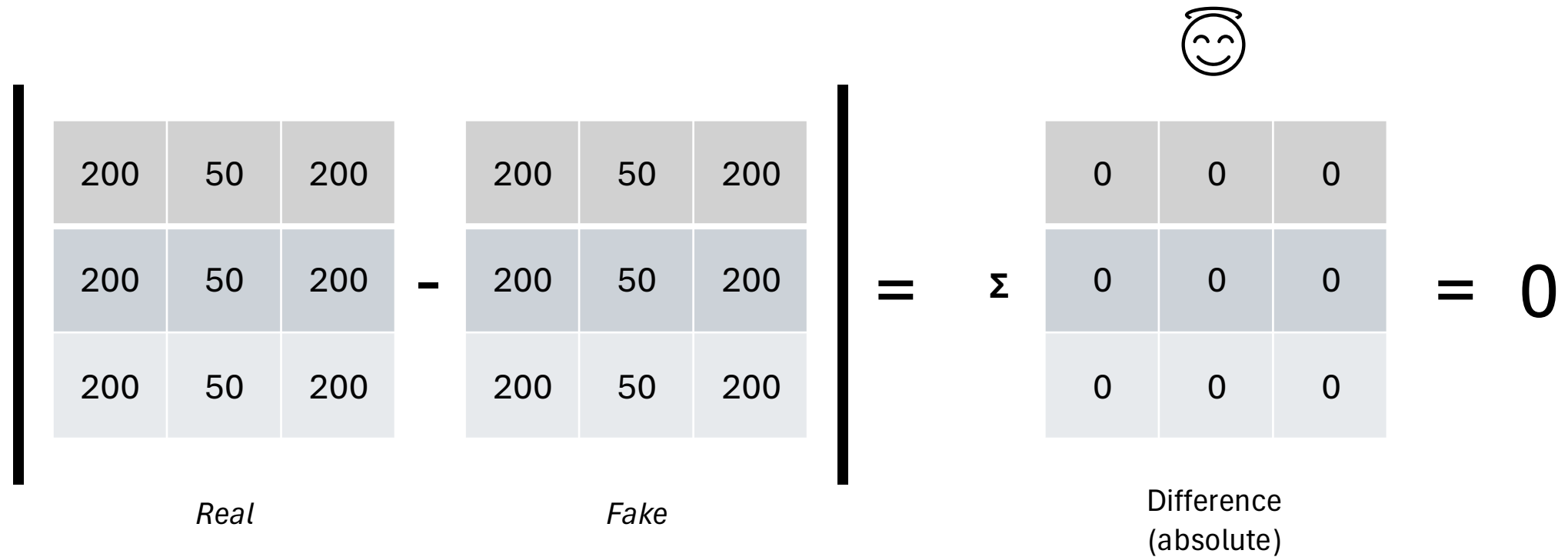
As imagens (rostos) são apresentadas como matrizes, em que os valores (0 – 255) dos *pixels* esta apresentada em cada célula .

Pixel distance



Aqui realizada uma subtração elemento a elemento.

Pixel distance


$$\begin{array}{|c|c|c|} \hline 200 & 50 & 200 \\ \hline 200 & 50 & 200 \\ \hline 200 & 50 & 200 \\ \hline \end{array} - \begin{array}{|c|c|c|} \hline 200 & 50 & 200 \\ \hline 200 & 50 & 200 \\ \hline 200 & 50 & 200 \\ \hline \end{array} = \sum \begin{array}{|c|c|c|} \hline 0 & 0 & 0 \\ \hline 0 & 0 & 0 \\ \hline 0 & 0 & 0 \\ \hline \end{array} = 0$$

Real *Fake* Difference (absolute)

Nesse cenário, as duas imagens são iguais. Portanto, a distância entre as matrizes é igual a zero.



Qual o problema óbvio dessa metodologia?

Pixel distance

$$\left| \begin{array}{|c|c|c|} \hline 200 & 50 & 200 \\ \hline 200 & 50 & 200 \\ \hline 200 & 50 & 200 \\ \hline \end{array} \right| - \left| \begin{array}{|c|c|c|} \hline \mathbf{50} & 200 & 200 \\ \hline \mathbf{50} & 200 & 200 \\ \hline \mathbf{50} & 200 & 200 \\ \hline \end{array} \right| = \Sigma \begin{array}{|c|c|c|} \hline 150 & 150 & 0 \\ \hline 150 & 150 & 0 \\ \hline 150 & 150 & 0 \\ \hline \end{array} = 900!!!$$

Real *Fake* Difference
(absolute)

Em um cenário de diferenças mínimas entre imagens a métrica se torna pouco confiável.



Solução:

Comparar imagens através de atributos
informativos!

Feature distance

Real



Fake



Extração de atributos

4 Patas
2 Olhos
2 Orelhas
1 Focinho
1 Cauda

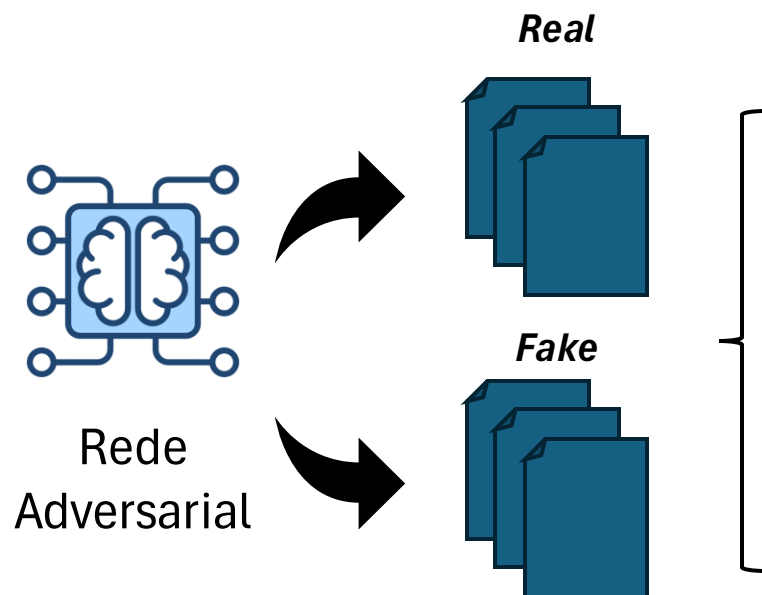
6 Patas
2 Olhos
1 Orelhas
1 Focinho
0 Cauda

Cálculo de distância
e.g.,
Euclideana ou
Cossenos

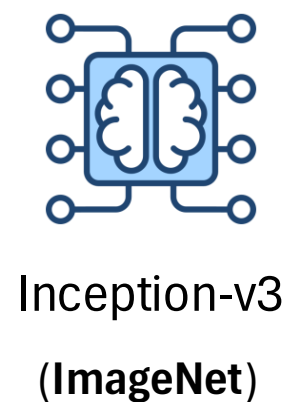
Legal, mas...

Como extrair atributos de um modelo
generativo?

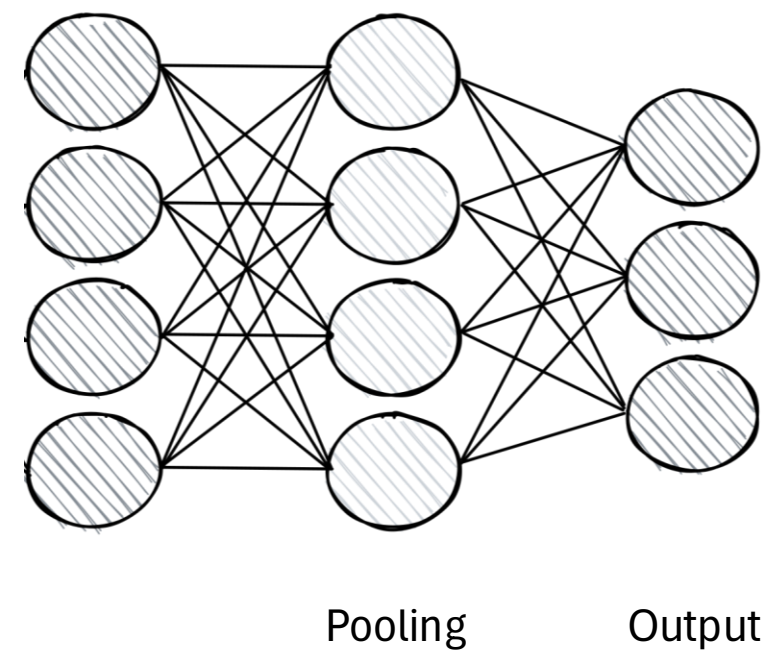
1) Depois do treinamento



2) Introdução de modelo pré-treinado

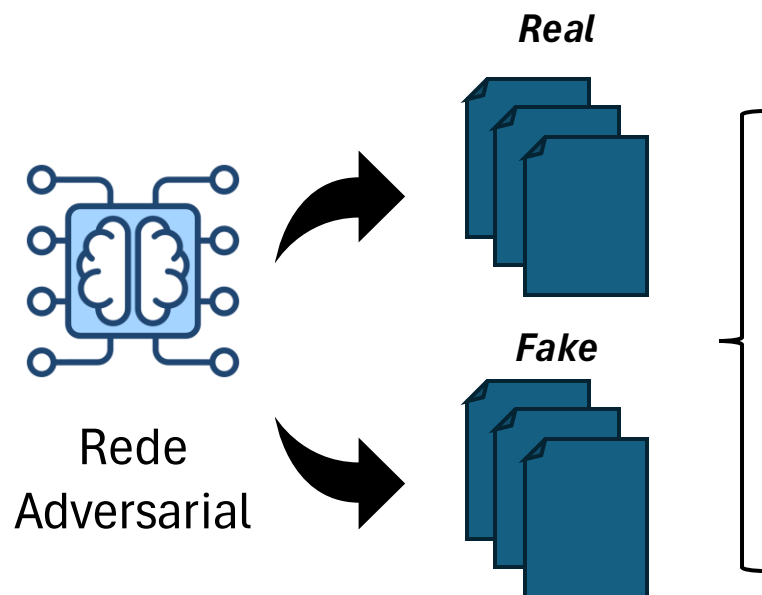


3) Extração dos atributos

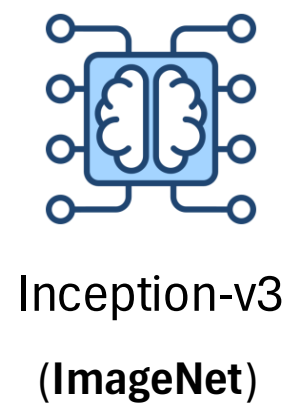


As amostras (imagens) são introduzidas em um modelo de classificação, Inception-v3. Tal modelo foi treinado utilizando o banco ImageNet – melhor referência para aplicações de imagem.

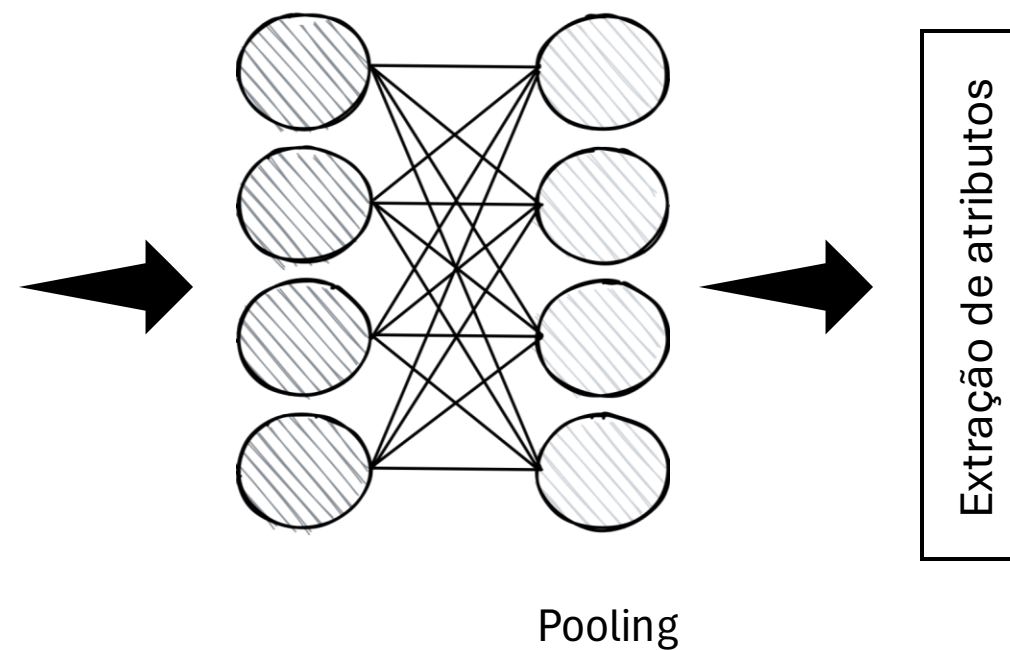
1) Após o treinamento



2) Introdução de modelo pré-treinado



3) Extração dos atributos



A última camada que faz a classificação do modelo é removida. Dessa forma, os atributos são extraídos da camada de "**pooling**"

Feature distance

Real



Fake



Extração de atributos

Real

Embedding = [0.8, -0.3, 1.5,
..., até $n = 2048$]

Fake

Embedding = [0.2, 0.54,
0.49, ..., até $n = 2048$]

Cálculo de distância

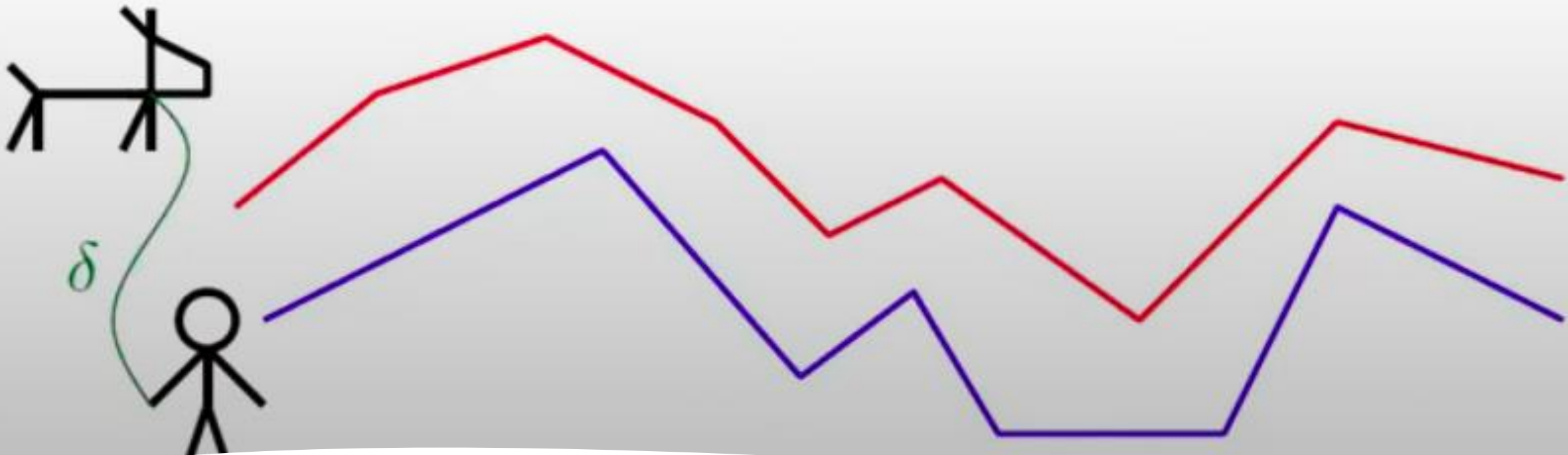
e.g.,
Euclideana ou
Cossenos

**Fréchet Inception
Distance**

Conclusão

Agora podemos comparar imagens a nível de atributos, logo sendo mais robusto a pequenas mudanças!





Fréchet Inception Distance (FID)

- O FID é uma métrica que avalia a "**qualidade**" e a "**diversidade**" de imagens geradas por um modelo pré-treinado.
- Originalmente, desenvolvido para medir distância entre curvas – *dog walker problem*.

- Em modelos generativos, FID calcula a distância entre distribuições - **normais** e **multivariadas**.
- As distribuições são baseadas no *embeddings* das amostras reais (r) e geradas (g), em que são deduzidos:
 - Media (μ) e Covariância (Σ).
- Em seguida a distância é calculada através da equação:

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

- Em modelos generativos, FID calcula a distância entre distribuições - **normais** e **multivariadas**.
- As distribuições são baseadas no *embeddings* das amostras reais (r) e geradas (g), em que são deduzidos:
 - Media (μ) e Covariância (Σ).
- Em seguida a distância é calculada através da equação:

$$FID = \underbrace{\|\mu_r - \mu_g\|^2}_{\text{Mede a distância entre os centros das distribuições (media)}} + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2})$$

Mede a distância entre os
centros das distribuições
(media)

- Em modelos generativos, FID calcula a distância entre distribuições - **normais** e **multivariadas**.
- As distribuições são baseadas no *embeddings* das amostras reais (r) e geradas (g), em que são deduzidos:
 - Media (μ) e Covariância (Σ).
- Em seguida a distância é calculada através da equação:

$$FID = \underbrace{\|\mu_r - \mu_g\|^2}_{\text{Mede a distância entre os centros das distribuições (media)}} + \underbrace{Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2})}_{\text{Mede a diferença entre as formas e dispersões das distribuições (covariância).}}$$

Mede a distância entre os centros das distribuições (media)

Mede a diferença entre as **formas** e **dispersões** das distribuições (covariância).

- Em modelos generativos, FID calcula a distância entre distribuições - **normais** e **multivariadas**.
- As distribuições são baseadas no *embeddings* das amostras reais (r) e geradas (g), em que são deduzidos:
 - Media (μ) e Covariância (Σ).
- Em seguida a distância é calculada através da equação:

$$FID = \underbrace{\|\mu_r - \mu_g\|^2}_{\text{Mede a distância entre os centros das distribuições (media)}} + \underbrace{Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2})}_{\text{Mede a diferença entre as formas e dispersões das distribuições (covariância).}}$$

O método supõe uma distribuição normal para cálculo das estatísticas descritivas, o que reduz o **tempo computacional**. No entanto, não é uma aproximação realista.

- Em modelos generativos, FID calcula a distância entre distribuições - **normais** e **multivariadas**.
- As distribuições são baseadas no *embeddings* das amostras reais (r) e geradas (g), em que são deduzidos:
 - Media (μ) e Covariância (Σ).
- Em seguida a distância é calculada através da equação:

$$FID = \underbrace{\|\mu_r - \mu_g\|^2}_{\text{Mede a distância entre os centros das distribuições (media)}} + \underbrace{Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2})}_{\text{Mede a diferença entre as formas e dispersões das distribuições (covariância).}}$$

Mede a distância entre os centros das distribuições (media)

Mede a diferença entre as **formas** e **dispersões** das distribuições (covariância).

Por que **multivariada**? A métrica não avalia atributos independentemente, pelo contrário ela busca capturar a relação entre atributos (covariância). Tal aspecto é considerado para todos os *embeddings* da imagem.

- Em modelos generativos, FID calcula a distância entre distribuições - **normais** e **multivariadas**.
- As distribuições são baseadas no *embeddings* das amostras reais (r) e geradas (g), em que são deduzidos:
 - Media (μ) e Covariância (Σ).
- Em seguida a distância é calculada através da equação:

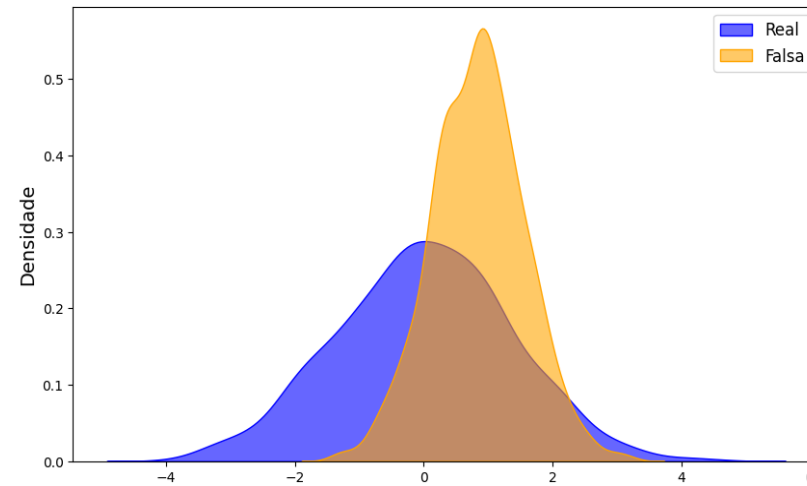
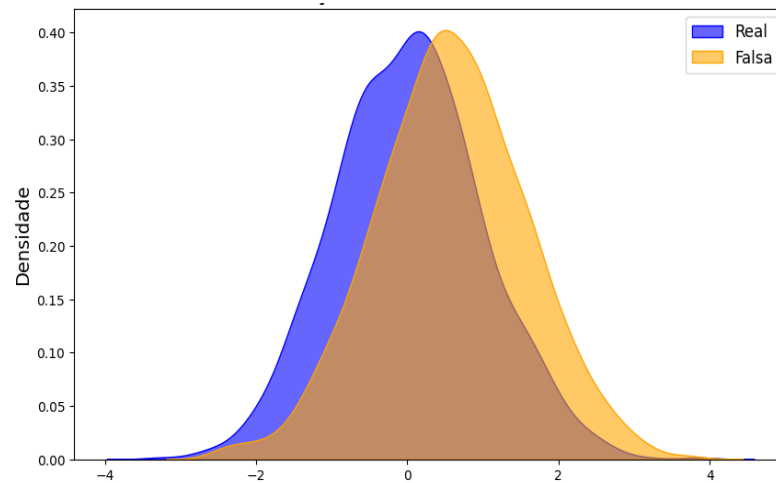
$$FID = \underbrace{\|\mu_r - \mu_g\|^2}_{\text{Mede a distância entre os centros das distribuições (media)}} + \underbrace{Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2})}_{\text{Mede a diferença entre as formas e dispersões das distribuições (covariância).}}$$

Mais importante, por que FID captura tanto **fidelidade** (qualidade) quanto a **diversidade**?

$$FID = \underbrace{\|\mu_r - \mu_g\|^2}_{\text{Mede a distância entre os centros das distribuições (media)}} + \underbrace{T_r(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2})}_{\text{Mede a diferença entre as formas e dispersões das distribuições (covariância)}}$$

Mede a distância entre os centros das distribuições (media)

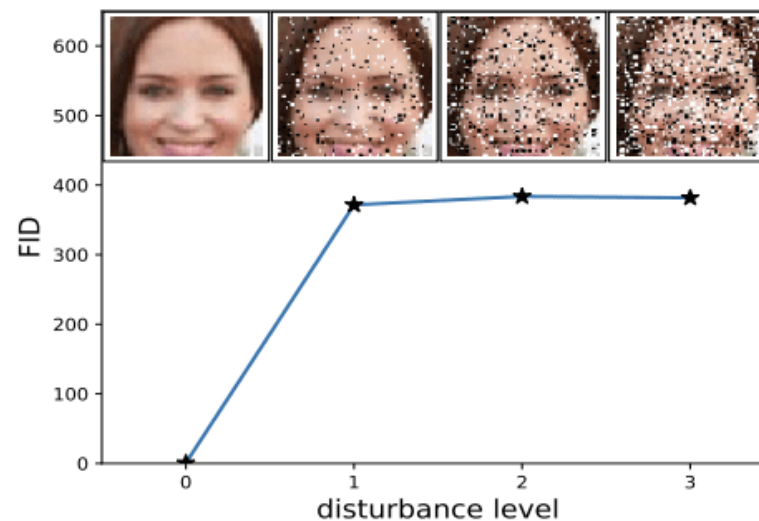
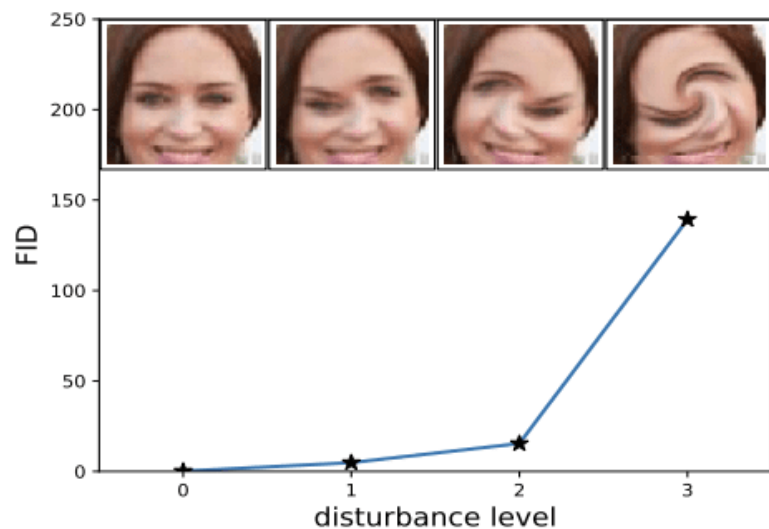
Mede a diferença entre as **formas e dispersões** das distribuições (covariância).



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

- **Valor de FID baixo:** as distribuições são muito semelhantes, ou seja, as imagens geradas estão próximas das imagens reais em termos de características.
- **Valores altos de FID:** As distribuições são diferentes, indicando que as imagens geradas não parecem tão reais.

- **Valor de FID baixo:** as distribuições são muito semelhantes, ou seja, as imagens geradas estão próximas das imagens reais em termos de características.
- **Valores altos de FID:** As distribuições são diferentes, indicando que as imagens geradas não parecem tão reais.



Conclusão

FID traz um relativo progresso em relação a outras métricas (e.g., Euclidiana, Cosseno e **Inception Score**).



Ainda assim, existem limitações inerentes a técnica!

- Suposição irrealista de distribuição Gaussiana.
- Dependência de um modelo pré-treinado limitado ao ImageNet.
- Necessidade de grandes **amostras** para confiabilidade.
- Incapacidade de capturar detalhes como outliers e artefatos.
- Escala não intuitiva e dificuldade de interpretação.

Será que existem outros
métodos de avaliação?

**Sim! Podemos utilizar métricas "tradicionais" para
avaliar imagens de um modelo – precision e recall!**

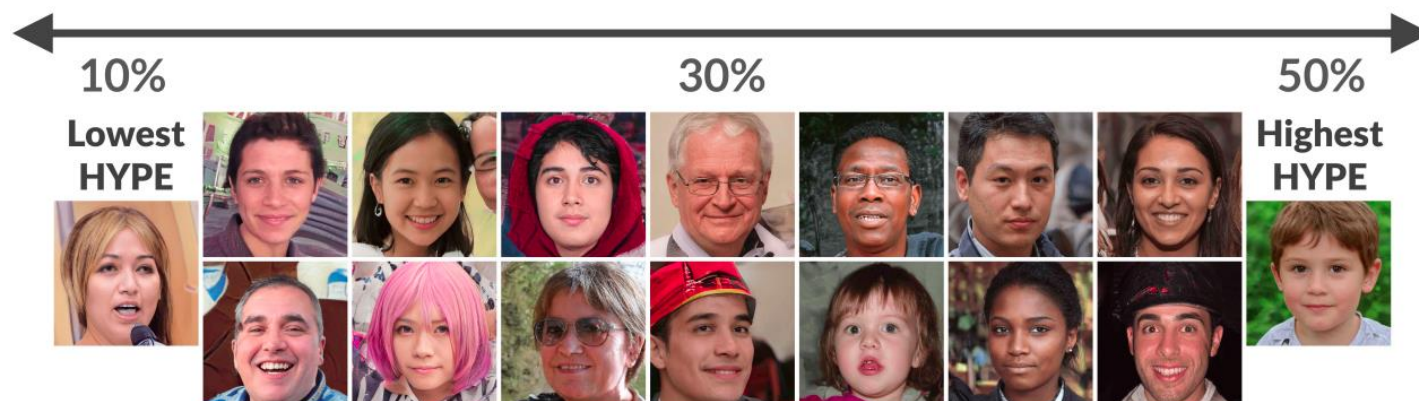
(Próxima aula)

Vale a pena fazer uma
menção honrosa

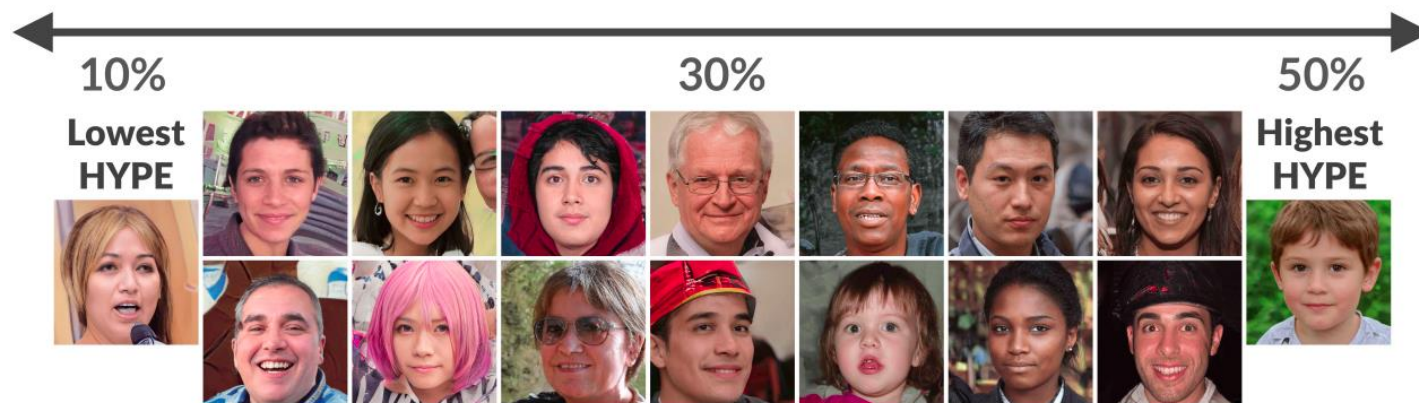
Métodos de avaliação com base na percepção
humana.

Human eYe Perceptual Evaluation (HYPE)

- **Crowdsourcing:** Participantes **humanos** avaliam as imagens geradas em plataformas como Amazon Mechanical Turk – reais ou geradas.
 - A avaliação é limitada a um tempo pré-determinado.
- **Taxa de Confusão Humana:** É calculada a porcentagem de imagens geradas classificadas incorretamente como reais.

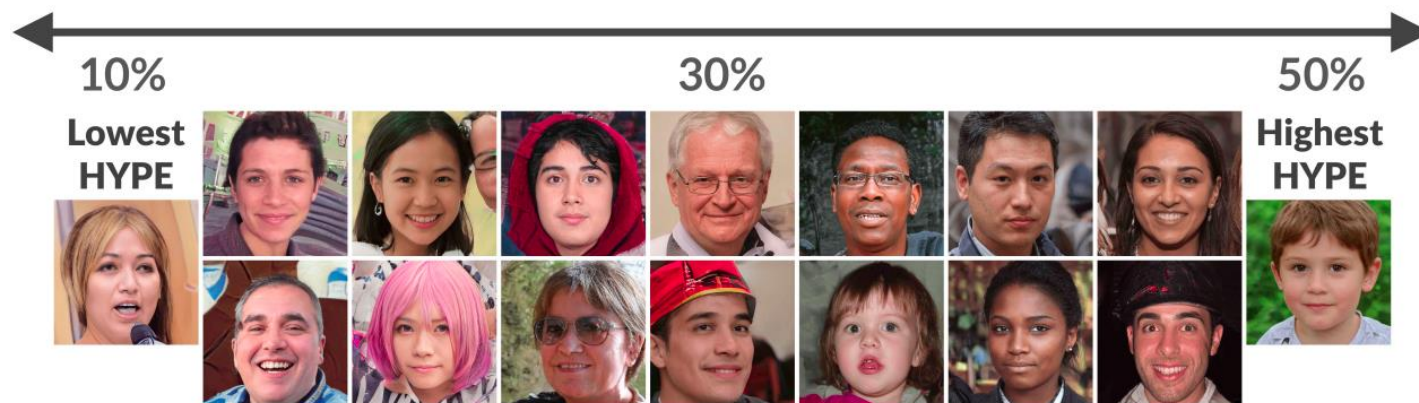


- **Crowdsourcing:** Participantes **humanos** avaliam as imagens geradas em plataformas como Amazon Mechanical Turk – reais ou geradas.
 - A avaliação é limitada a um tempo pré-determinado.
- **Taxa de Confusão Humana:** É calculada a porcentagem de imagens geradas classificadas incorretamente como reais.



Porcentagens altas representam que as imagens confundiram os avaliadores. Lembrando, todas as imagens apresentadas foram geradas por uma rede adversarial!

- **Crowdsourcing:** Participantes **humanos** avaliam as imagens geradas em plataformas como Amazon Mechanical Turk – reais ou geradas.
 - A avaliação é limitada a um tempo pré-determinado.
- **Taxa de Confusão Humana:** É calculada a porcentagem de imagens geradas classificadas incorretamente como reais.



Ainda que seja uma abordagem interessante, é impraticável para maior partes dos estudos. Além disso, depende da subjetividade humana no critério de avaliação.

Takeaway message

- Métricas de qualidade variam conforme domínios específicos – alguns problemas requerem métricas completamente novas;
- Extração de atributos é um método robusto para avaliação de qualidade em imagens;
- Fréchet Inception Distance (FID) permite capturar tanto fidelidade quanto a diversidade das amostras produzidas;
- Em alguns domínios, **avaliação humana** ainda é necessária, sobretudo para geração de dados complexos;

Homework!

- No repositório da disciplina está disponível um notebook (`exercicios/Aula_05A_-_Avaliacao_GenIA.ipynb`).
 - Os alunos devem executa-lo e debater as implicações praticas das métricas de qualidade no processo de treinamento do modelo.
 - Observe as questões distribuídas no notebook.
- A compreensão dos conceitos será fundamental para execução do trabalho final da disciplina – **o prazo está chegando!**

Referências

Livro

1. JAKUB, V. B. (2019). GANs in Action: Deep learning with Generative Adversarial Networks (1st ed.). Manning Publications.
2. FOSTER, David. Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play. 2nd Edition. O'Reilly Media. 2023

Artigos

1. BORJI, Ali. Pros and Cons of GAN Evaluation Measures: New Developments. 2021. Disponível em: <https://arxiv.org/abs/2103.09396>. Acesso em: 22 jan. 2025.
2. KYNKÄÄNNIEMI, Tuomas; KARRAS, Tero; LAINE, Samuli; LEHTINEN, Jaakko; AILA, Timo. Improved Precision and Recall Metric for Assessing Generative Models. 2019. Disponível em: <https://arxiv.org/abs/1904.06991>. Acesso em: 22 jan. 2025.
3. SAUER, Axel; CHITTA, Kashyap; MÜLLER, Jens; GEIGER, Andreas. Projected GANs Converge Faster. 2021. Disponível em: <https://arxiv.org/abs/2111.01007>. Acesso em: 22 jan. 2025.
4. ZHOU, Sharon; GORDON, Mitchell L.; KRISHNA, Ranjay; NARCOMNEY, Austin; FEI-FEI, Li; BERNSTEIN, Michael S. HYPE: A Benchmark for Human eYe Perceptual Evaluation of Generative Models. 2019. Disponível em: <https://arxiv.org/abs/1904.01121>. Acesso em: 22 jan. 2025.



Perguntas?