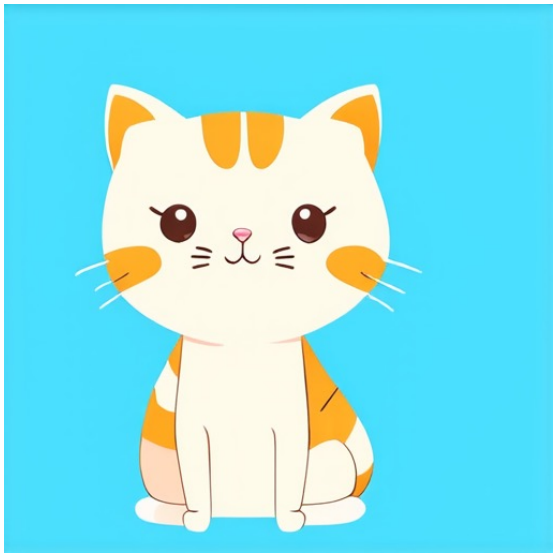
BOSTON UNIVERSITY

Project 4 Evaluation

https://dl4ds.github.io/sp2026/

Project 4 Inputs



In JSON files of same basename:

{"inputs": "minimalist vector-style cute cartoon dog, flat pastel colors, no gradients, simple face expression", "seed": 397}

https://github.com/DL4DS/synth-cute

Project 4

Build a model

- Synthetic data set
  - Built with Stable Diffusion 3.
  - 2K images each of 11 subjects
  - → 22K images total
- Build your own image diffusion model based on this data set.

Grading criteria

Auto-grader

- Inception score
- Fréchet Inception Distance

Manual-grading (subjective)

- 5% our chosen seeds
- 5% your choice of best output

Will share our favorites in Piazza

Quantifying Performance - Inception Score

Grading via another model

- Train classifier on image training set
- Usually the Inception model pretrained on ImageNet
- Want generated images to have a single very likely classification.
- But average flat classification across generated images.
- Formal formula checking KL-divergence between those on a per-generated image basis…

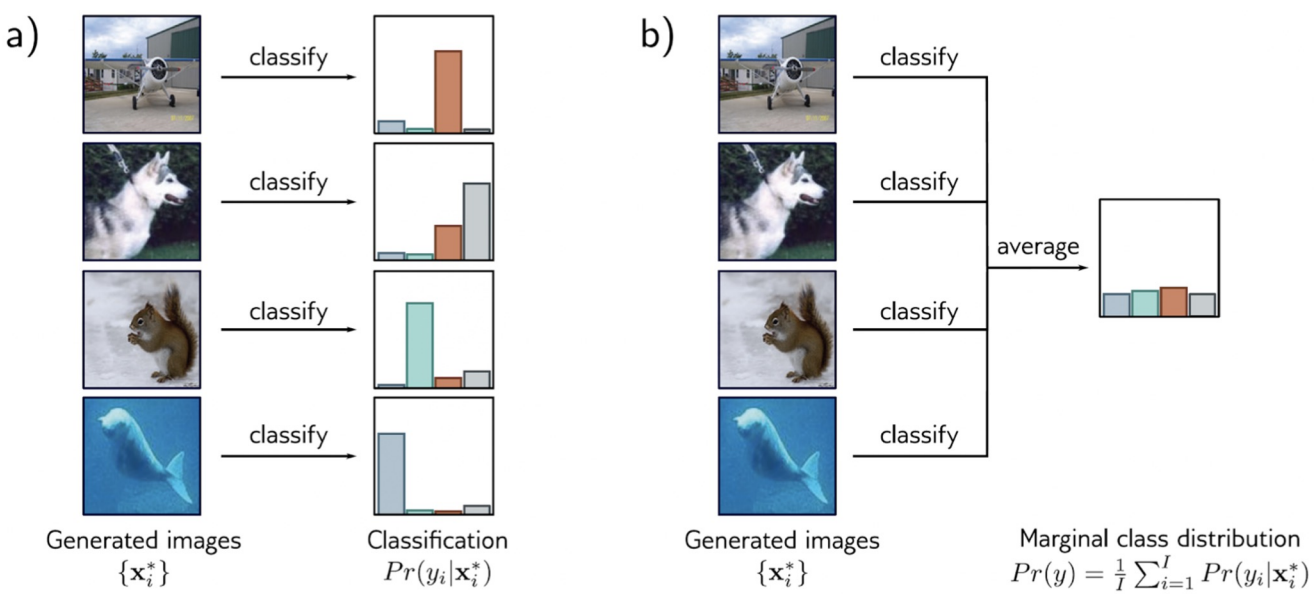


Figure 14.4 Inception score. a) A pretrained network classifies the generated images. If the images are realistic, the resulting class probabilities $Pr(y_i|\mathbf{x}_i^*)$ should be peaked at the correct class. b) If the model generates all classes equally frequently, the marginal (average) class probabilities should be flat. The inception score measures the average distance between the distributions in (a) and the distribution in (b). Images from Deng et al. (2009).

UDL Book §14.3

Inception Score Math

Inception Score:

$$\mathrm{IS} = \exp\left[\frac{1}{\mathrm{I}}\sum_{\mathrm{i}=1}^{\mathrm{I}} \mathrm{D}_{\mathrm{KL}}[\Pr(\mathrm{y}|\boldsymbol{x}_{\mathrm{i}}^{*}) \,||\Pr(\mathrm{y})]\right]$$

where

$$\Pr(\mathrm{y}) = \frac{1}{\mathrm{I}}\sum_{\mathrm{i}=1}^{\mathrm{I}} \Pr(\mathrm{y}|\boldsymbol{x}_{\mathrm{i}}^{*})$$

Decoder ring

- $p(y)$ = probability of class y
- $p(y|x)$ = conditional probability of class y
- $KL(P(x)||Q(x)) = \sum_{x\in X} P(x)\log\frac{P(x)}{Q(x)} =$ Kullback-Leibler divergence.
  - Interpretation is comparing approximate $Q(x)$ to true $P(x)$.

Improved Techniques for Training GANs (2016)

Interpretation of Inception Score

$$\mathrm{IS} = \exp\left[\frac{1}{\mathrm{I}}\sum_{\mathrm{i}=1}^{\mathrm{I}} \mathrm{D}_{\mathrm{KL}}[\Pr(\mathrm{y}|\boldsymbol{x}_{\mathrm{i}}^{*}) \,||\Pr(\mathrm{y})]\right]$$

- $\mathrm{D}_{\mathrm{KL}}[\Pr(\mathrm{y}|\boldsymbol{x}_{\mathrm{i}}^{*}) \,||\Pr(\mathrm{y})] = 0$ when distributions are same
- So $\exp[0] = 1$ is a minimum
- $\mathrm{D}_{\mathrm{KL}} \geq 0$ always
- High values correspond to classifier being able to distinguish classes of generated images.

Drawbacks of Inception Score

- Scores high even if generated images are exactly training images
- Scores high if generates same image for each class (mode collapse)

- Does not reward diversity across classes and withing classes

Project 4 IS Scoring:
Linearly scaled between IS=1 (0 pts) and IS=9 (30 pts)

Quantifying Performance – Fréchet Inception Distance

Another visual similarity metric based on Inception model (others can be used).

- Map generated images to distribution of Inception features.
- Model the distribution of Inception features as a multivariate normal distribution.
- Compare two such distributions with the Wasserstein distance.
  - Also called “earth mover’s distance”
  - Smaller is better.
  - Closed form solution from multivariate normal assumption.

Fréchet Inception Distance Math

- Let f map images to the features (last hidden outputs) of a classification model.
  - Originally used Inception model. Project 4 will use a new classifier.
- Apply f to real data set and generated data set.
  - Compute mean μ and covariance Σ of f outputs for each data set.
  - Use these statistics to specify multivariate normal distribution approximating the distribution of the features of each data set.
- Compute Fréchet distance d between two multivariate normal distributions has closed form:

$$d^2 = |\mu_G - \mu_R|^2 + \mathrm{tr}\left(\Sigma_G + \Sigma_R - 2(\Sigma_G\Sigma_R)^{1/2}\right)$$

- The closer the means and covariance matrices, the lower the score

GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium (2017)

Fréchet Inception Distance Math

Project 4 FID Scoring:
Linearly scaled between FID=1400 (0 pts) and FID=500 (30 pts).

Class Diversity Score (30 pts)

Measures whether the model generates images across all 11 animal classes, not just a few:

- Runs generated images through the classifier and computes the average predicted class probability p(y).
- For each class, checks if its probability is at least 5% (≥ 0.05). Classes below 5% are flagged as LOW.
- $\text{Score} = \sum \min(p_y, 0.05) / 0.55$, rewarding models that spread generation evenly across all 11 classes.
- Full 30 pts if the model generates roughly uniform coverage across all classes.