

Copyright Notice

These slides are distributed under the Creative Commons License.

DeepLearning.AI makes these slides available for educational purposes. You may not use or distribute these slides for commercial purposes. You may make copies of these slides and use or distribute them for educational purposes as long as you cite DeepLearning.AI as the source of the slides.

For the rest of the details of the license, see

<https://creativecommons.org/licenses/by-sa/2.0/legalcode>



deeplearning.ai

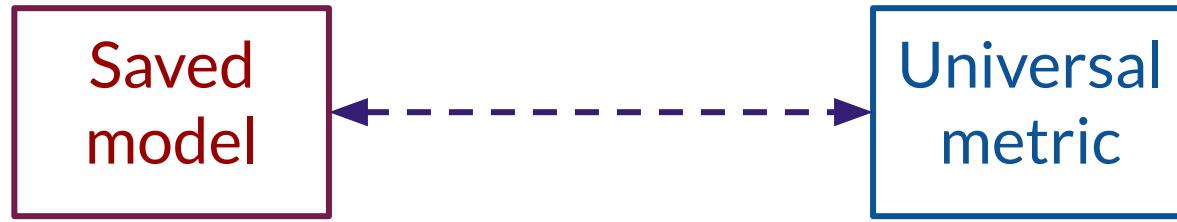
Evaluation

Outline

- Why evaluating GANs is hard
- Two properties: fidelity and diversity



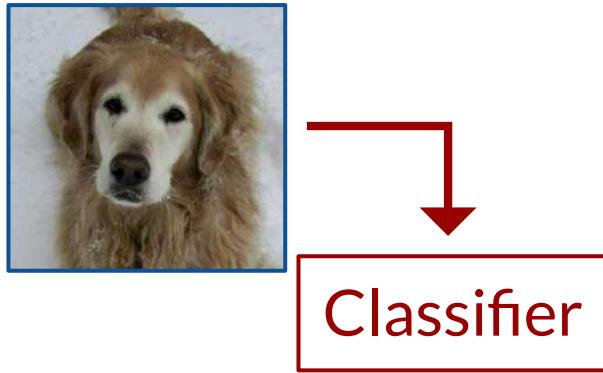
Why is evaluating GANs hard?



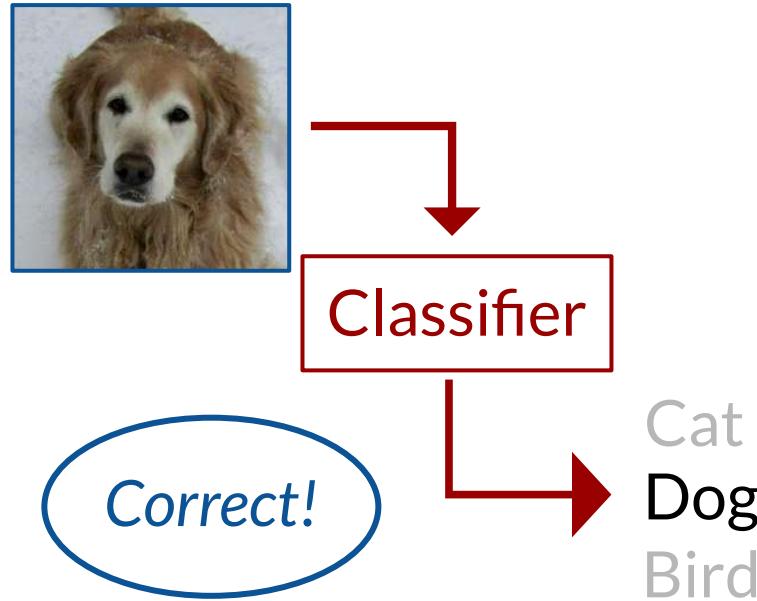
Why is evaluating GANs hard?

Classifier

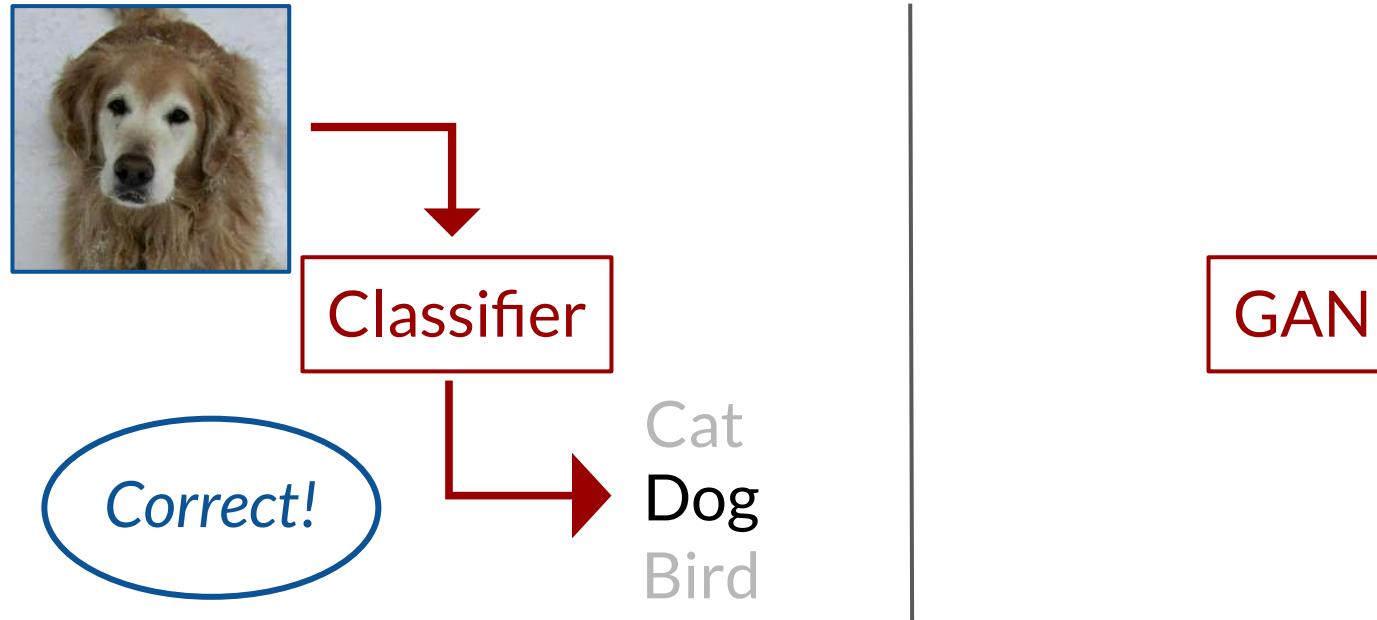
Why is evaluating GANs hard?



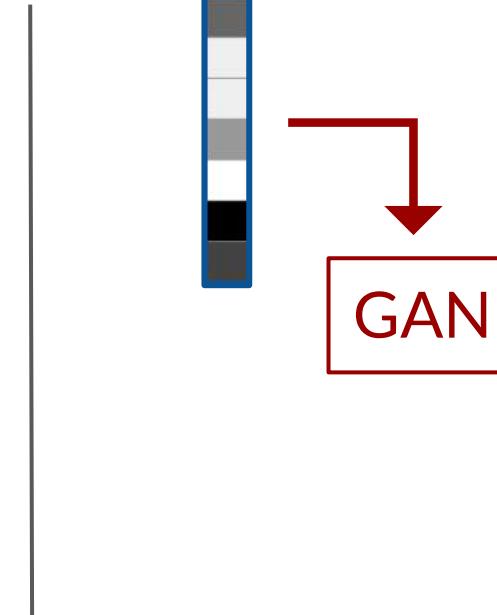
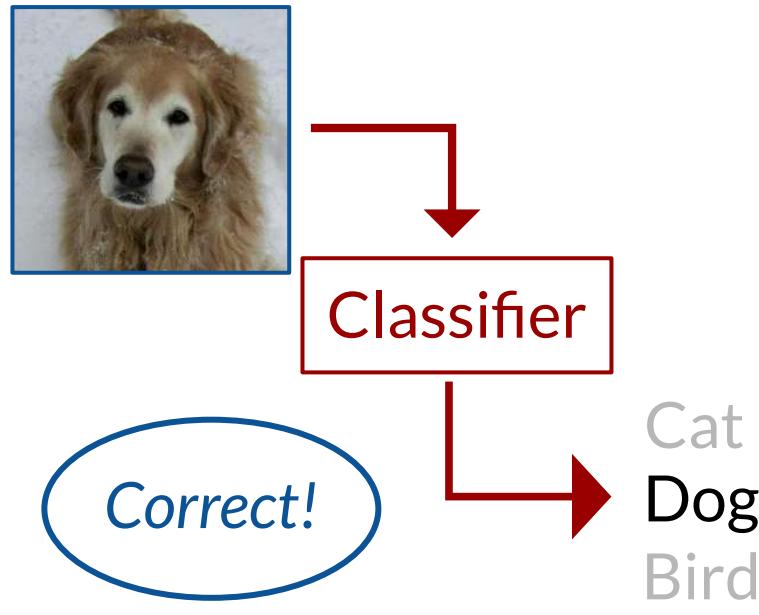
Why is evaluating GANs hard?



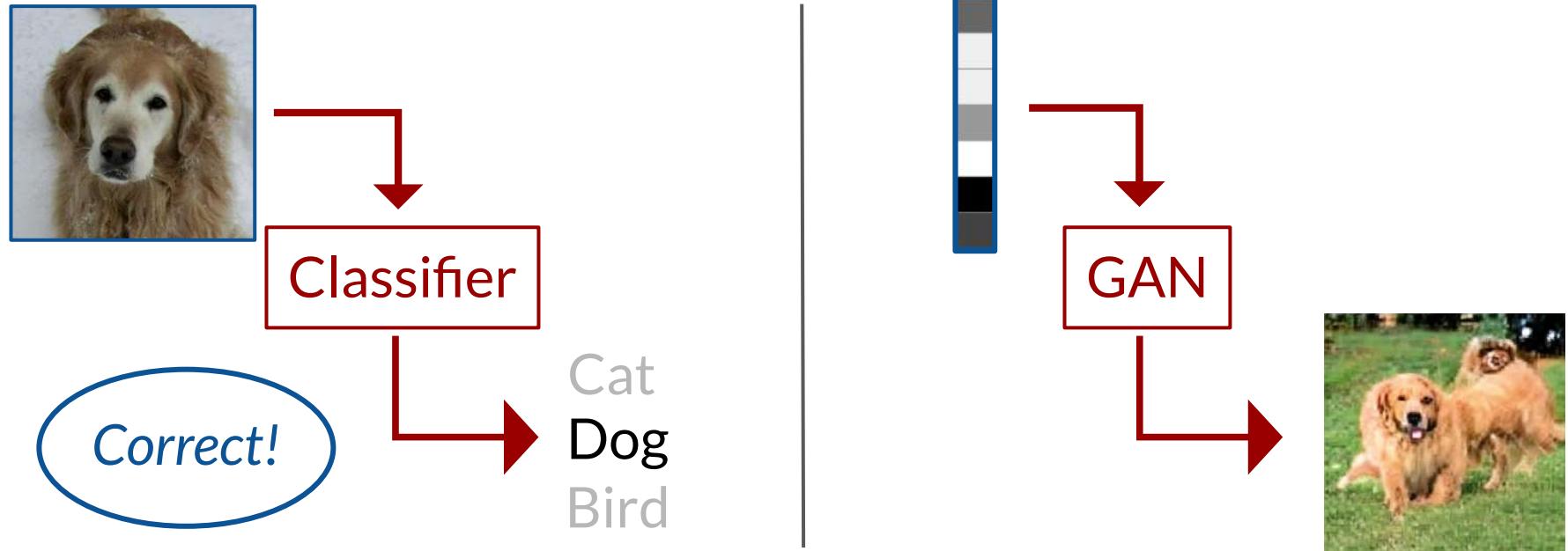
Why is evaluating GANs hard?



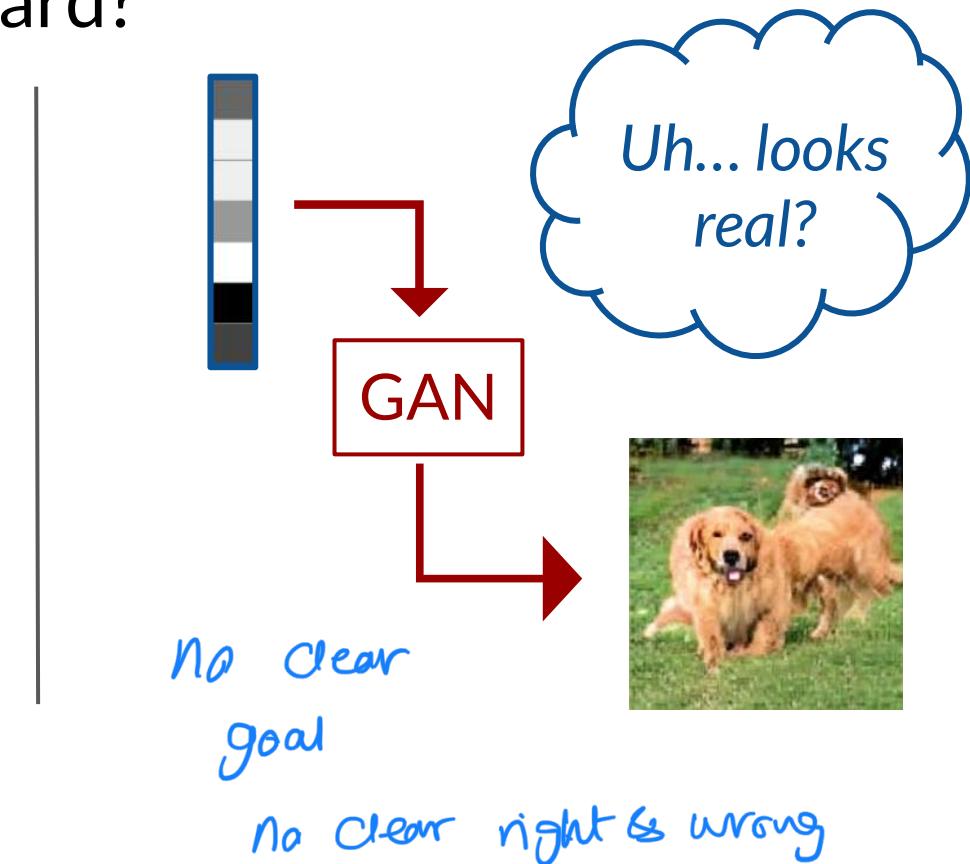
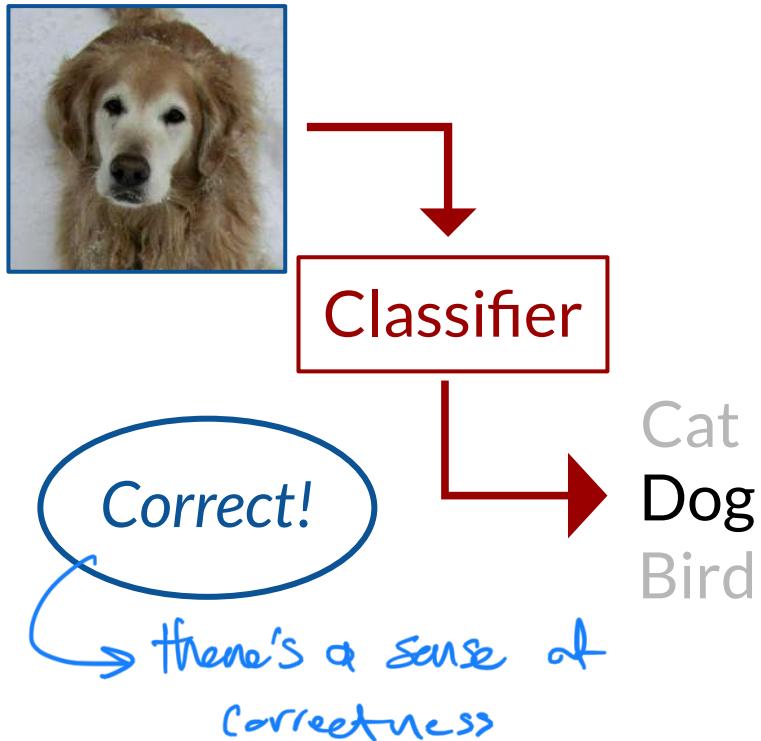
Why is evaluating GANs hard?



Why is evaluating GANs hard?



Why is evaluating GANs hard?



Two Important Properties

Fidelity:
quality of images



(Left) Available at: <https://github.com/NVlabs/stylegan>

Two Important Properties

Fidelity:
quality of images



Diversity:
variety of images



(Left) Available at: <https://github.com/NVlabs/stylegan>

Fidelity

Fidelity

7 777

Fake

how realistic the generated images are

Fidelity

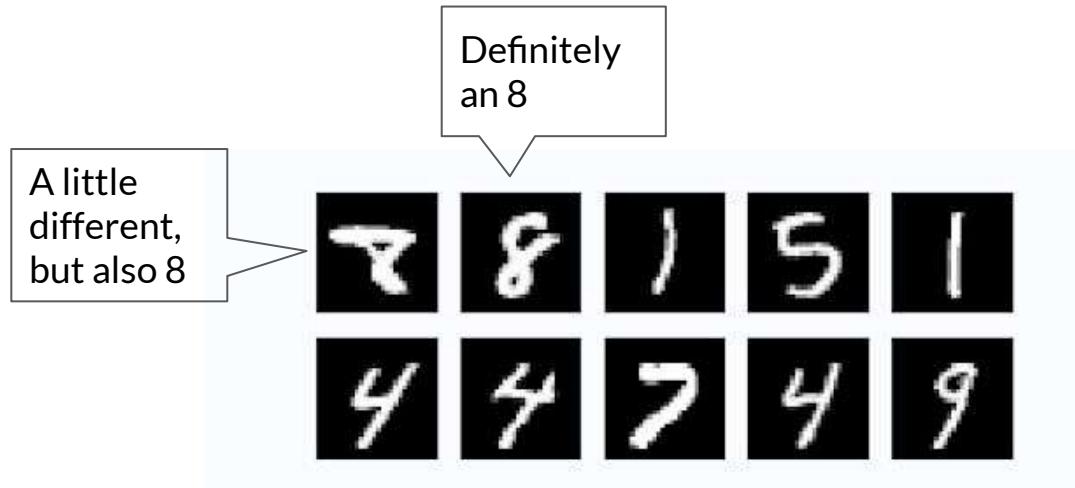


Real

Fake

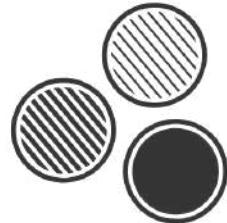
Realistic vs. Unrealistic

Diversity of distribution



Summary

- No ground-truth = challenging to evaluate
- Fidelity measures image quality and diversity measures variety
- Evaluation metrics try to quantify fidelity & diversity





deeplearning.ai

Comparing Images

Outline

- Pixel distance
- Feature distance



Pixel Distance

Subtract pixel values.

200	50	200
200	50	200
200	50	200

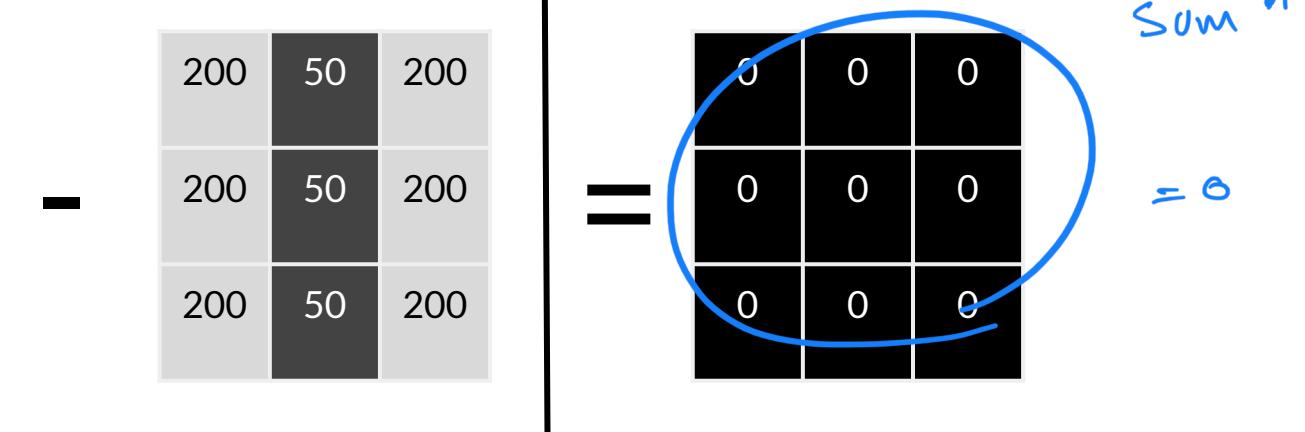
Real

200	50	200
200	50	200
200	50	200

Fake

0	0	0
0	0	0
0	0	0

Absolute
difference



not good.

Pixel Distance

200	50	200
200	50	200
200	50	200

Real

50	200	200
50	200	200
50	200	200

Fake

only one pixel shift → huge difference!

0	0	0
0	0	0
0	0	0

Absolute
difference

*will be
huge*

higher level features instead of pixels.

Feature Distance less sensitive to shifts.

Feature Distance

Real



Fake



Feature Distance

Real



2 eyes,
2 droopy ears,
1 nose, ...

Fake



2 eyes,
1 droopy ear,
5 legs,
1 nose, ...

Feature Distance

Real

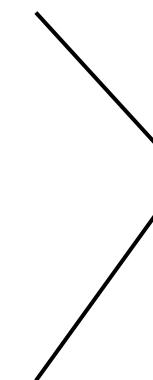


2 eyes,
2 droopy ears,
1 nose, ...

Fake



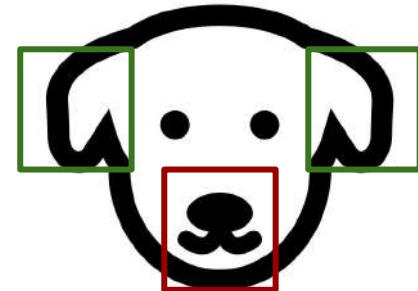
2 eyes,
1 droopy ear,
5 legs,
1 nose, ...



Compare
features!

Summary

- Pixel distance is simple but unreliable
- Feature distance uses the higher level features of an image, making it more reliable



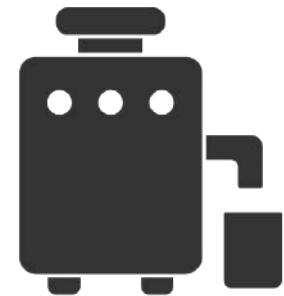


deeplearning.ai

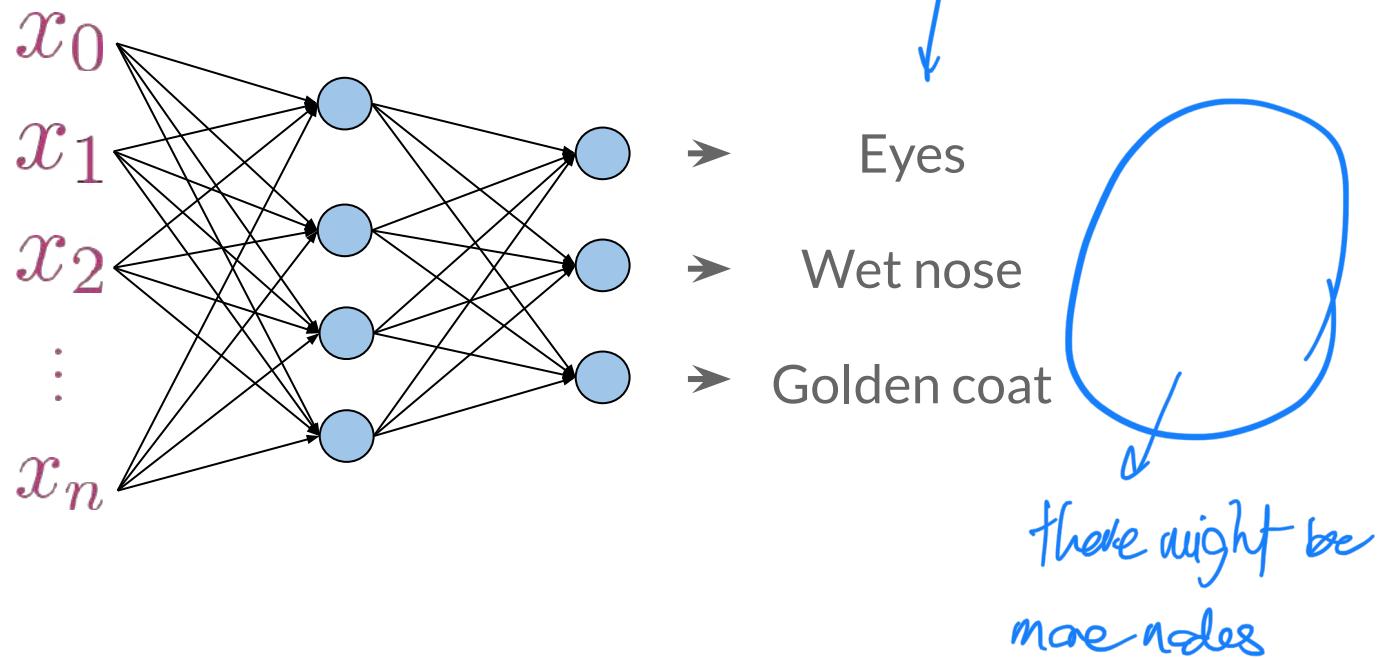
Feature Extraction

Outline

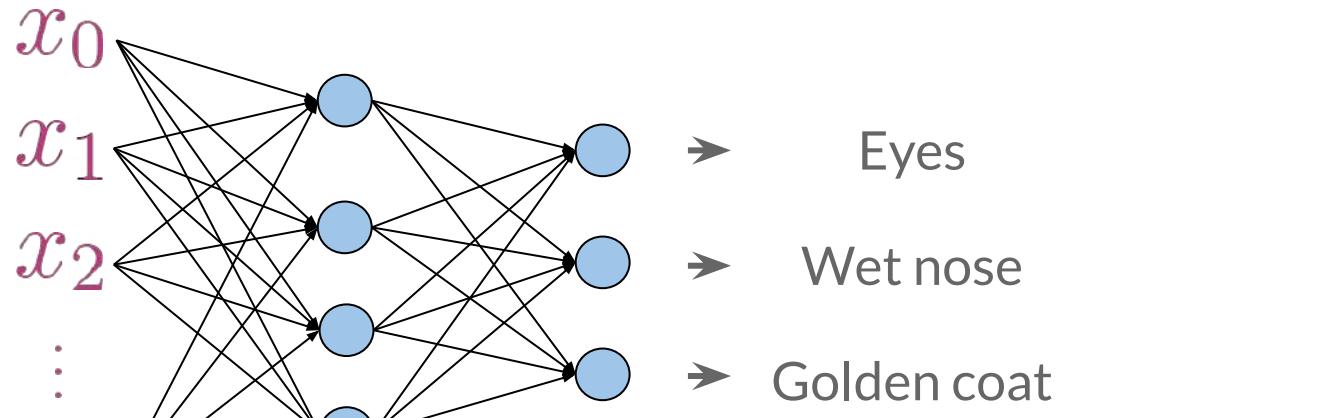
- Feature extraction using pre-trained classifiers
- ImageNet dataset



Classifier → Feature Extractor



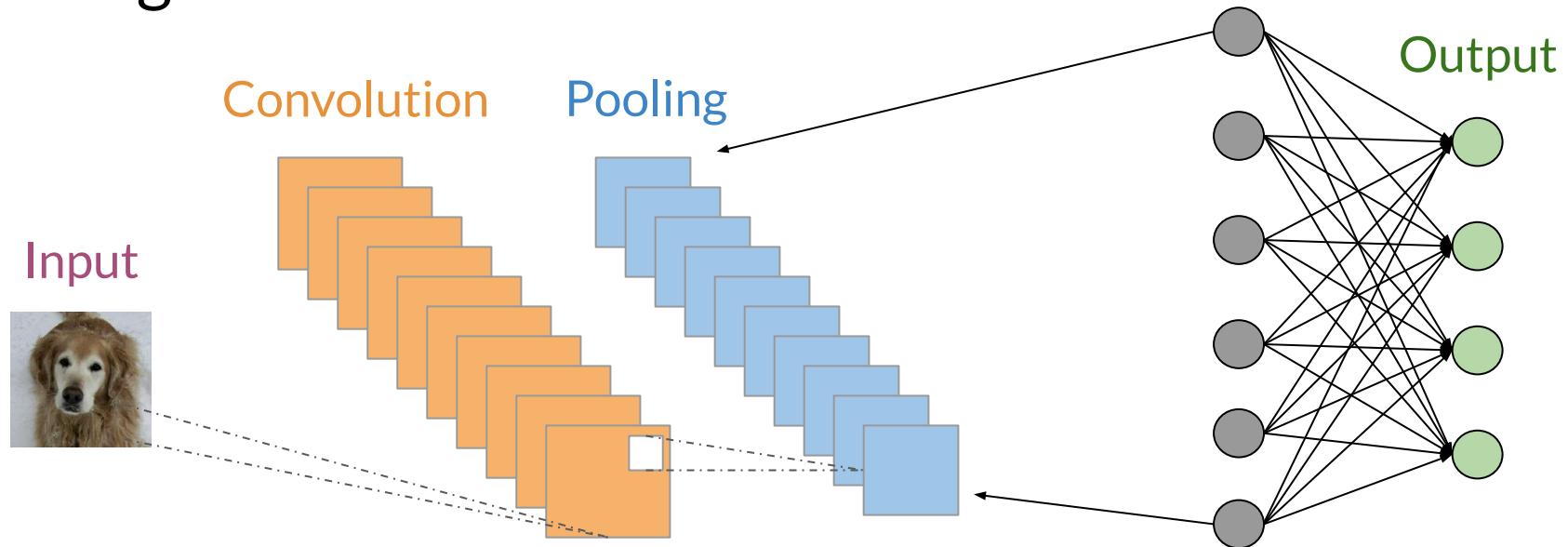
Classifier → Feature Extractor



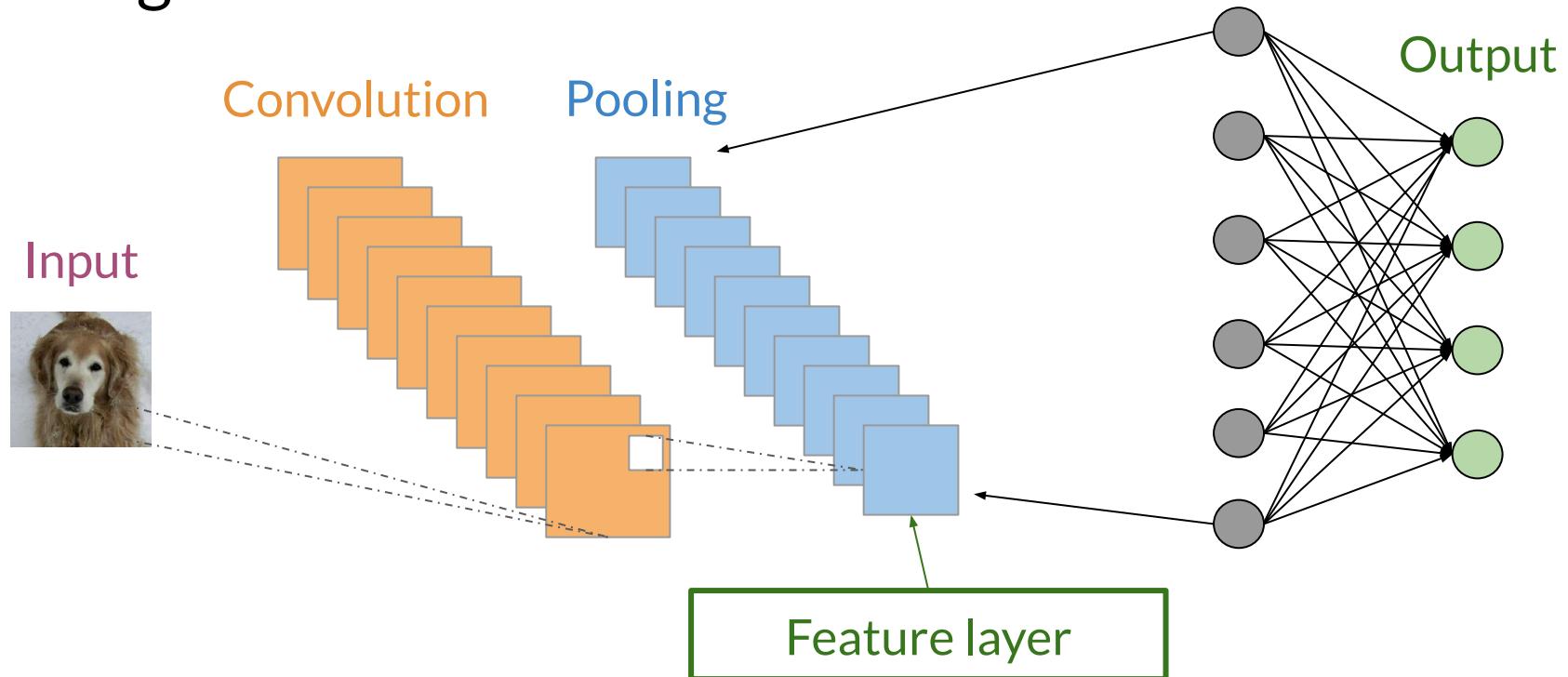
Extensively pre-trained
classifiers available to use

usually applicable to real world
photos

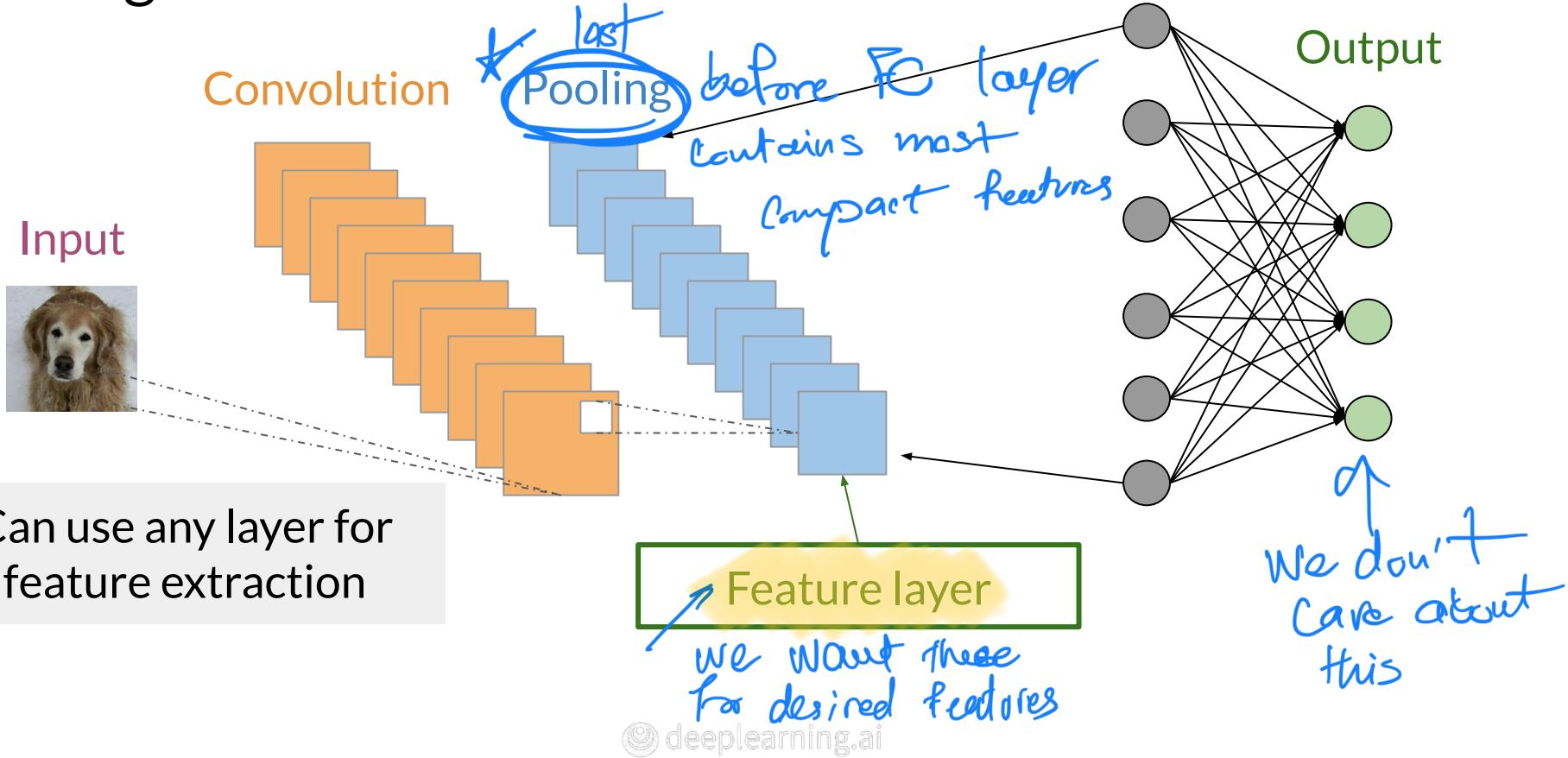
Using a Pre-trained Classifier



Using a Pre-trained Classifier

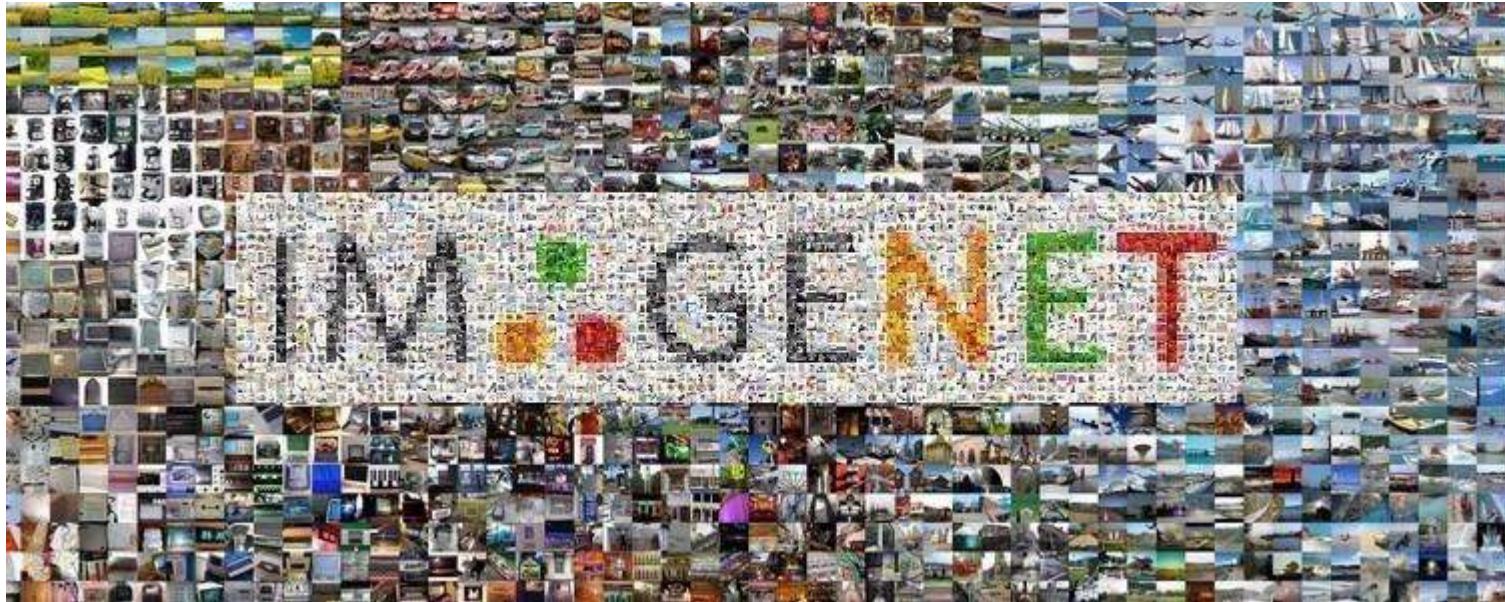


Using a Pre-trained Classifier



Can use any layer for
feature extraction

ImageNet



© 2016
Stanford
Vision Lab

ImageNet Attributes

- > 14 million images
- > 20,000 categories

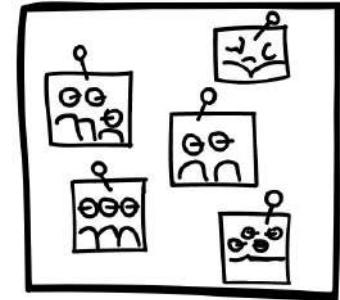
1. ImageNet embeddings / features
2. ^{pretrained} model



© 2016
Stanford
Vision Lab

Summary

- Classifiers can be used as feature extractors by cutting the network at earlier layers
- The last pooling layer is most commonly used for feature extraction
- Best to use classifiers that have been trained on large datasets—ImageNet





deeplearning.ai

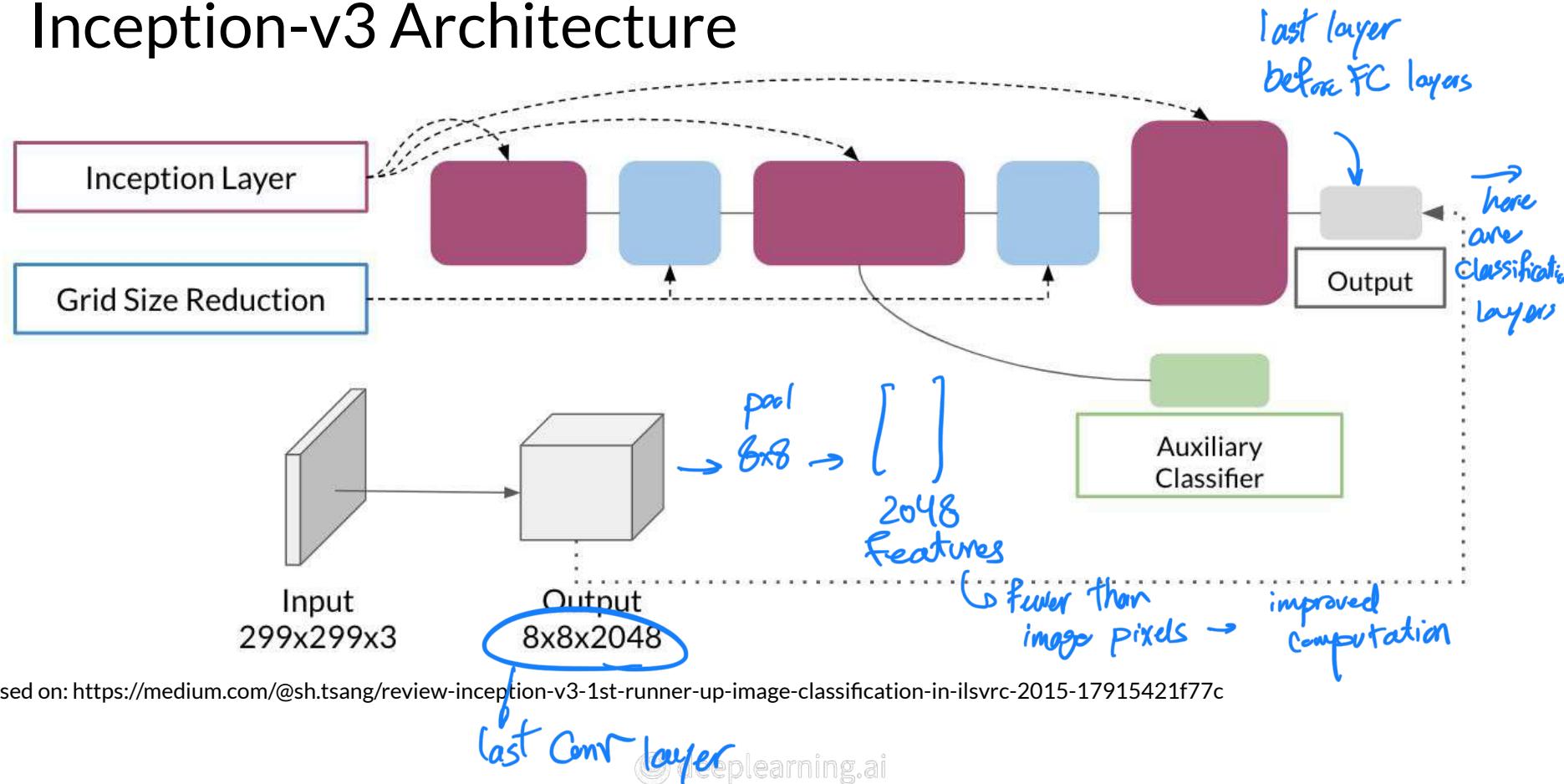
Inception-v3 and Embeddings

Outline

- Inception-v3 architecture
- Comparing extracted feature embeddings



Inception-v3 Architecture



Based on: <https://medium.com/@sh.tsang/review-inception-v3-1st-runner-up-image-classification-in-ilsvrc-2015-17915421f77c>

Embeddings

from those 2048



2 eyes,
2 droopy ears,
1 nose, ...

A blue circle highlights the list of features: "2 eyes, 2 droopy ears, 1 nose, ...".

Embeddings



Inception-v3

2 eyes,
2 droopy ears,
1 nose, ...

inception
feature
extractor

= $\phi(x)$

Embedding of x

We want
 $\phi(x)$ to be
similar & close (between
real/fake)

[] 2048

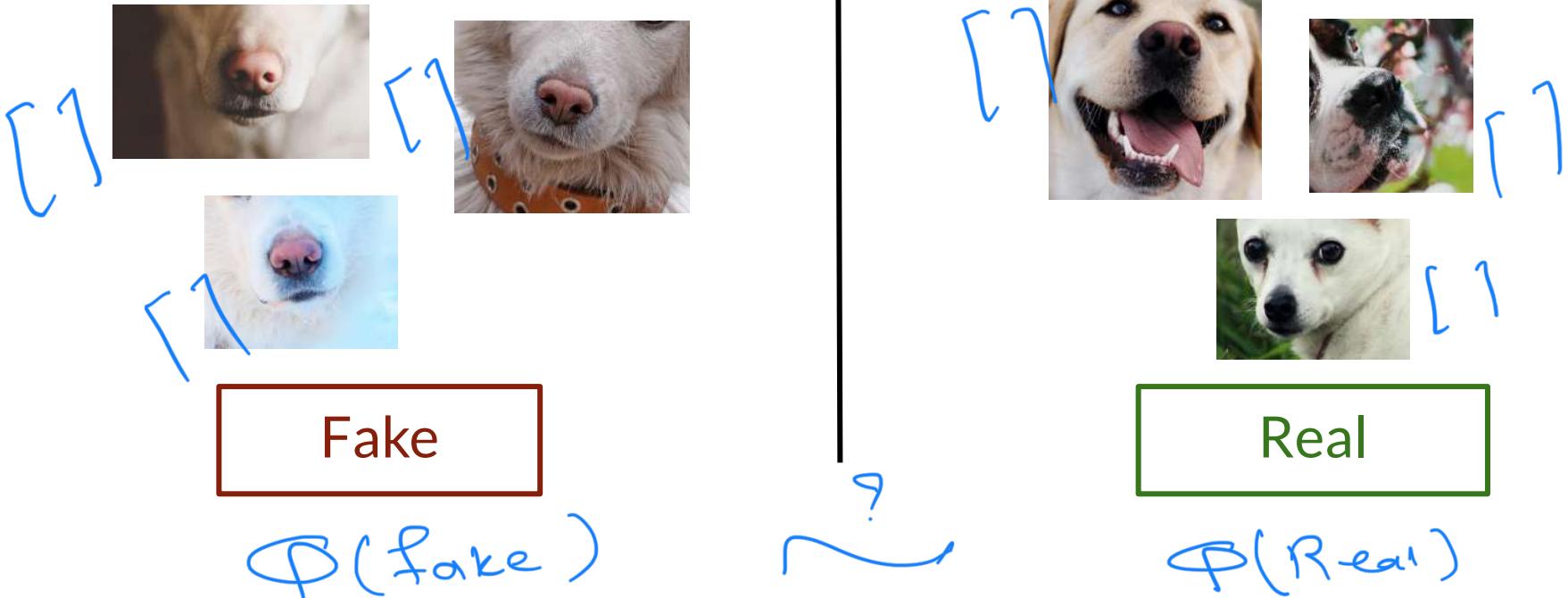
Comparing Embeddings



Fake

Can Subtract []'s.
use means

Comparing Embeddings



Summary

- Commonly used feature extractor: Inception-v3 classifier, which is pre-trained on ImageNet, with the output layer cut off
- These features are called embeddings
- Compare embeddings to get the feature distance





deeplearning.ai

Fréchet Inception Distance (FID)

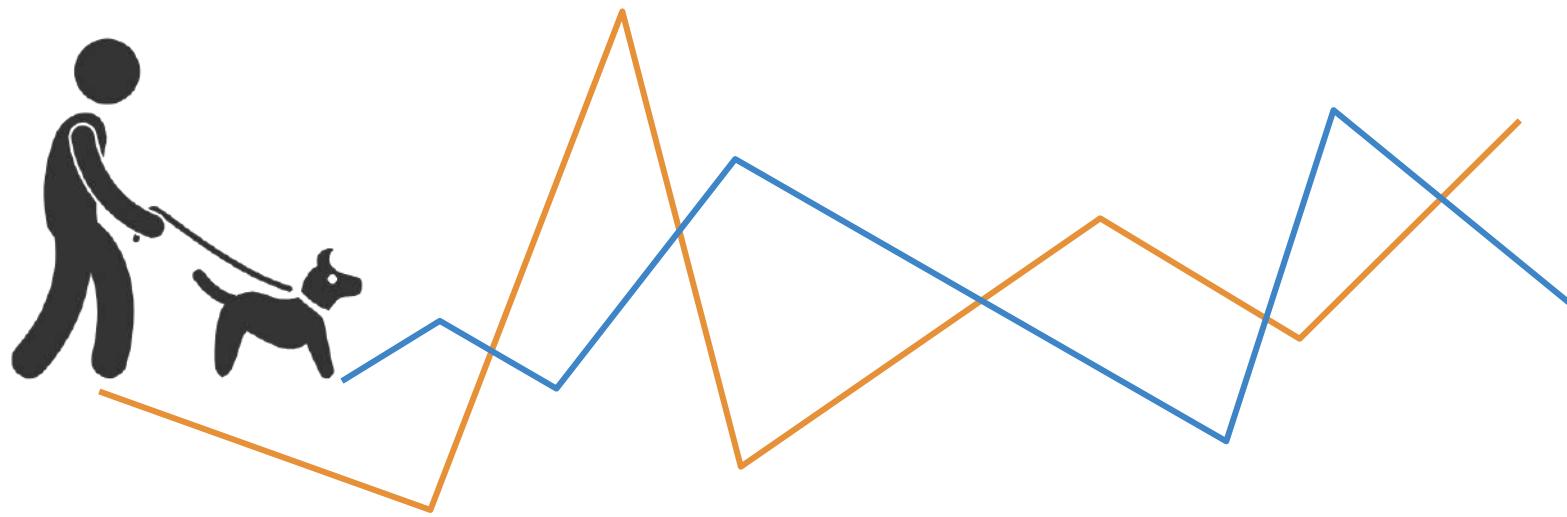
most popular metric for
feature distance between
real and fake

Outline

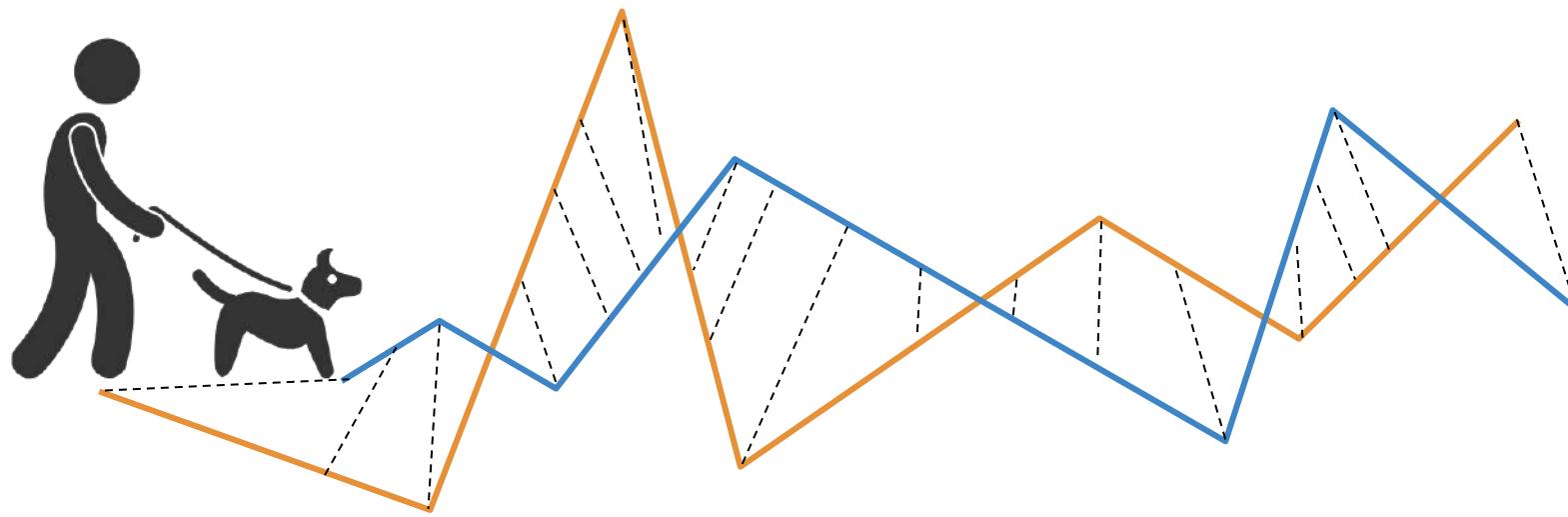
- Fréchet distance
- Evaluation method: Fréchet Inception Distance (FID)
- FID shortcomings



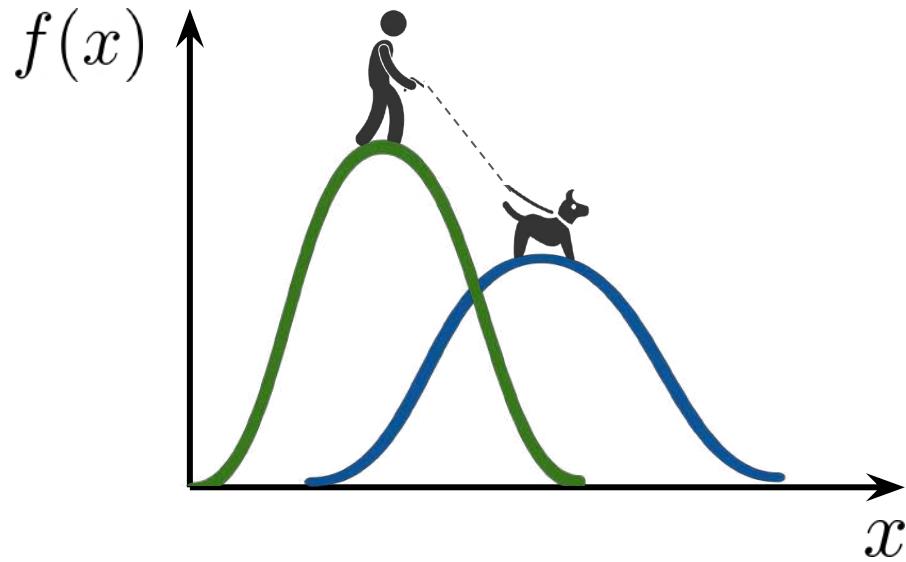
Fréchet Distance



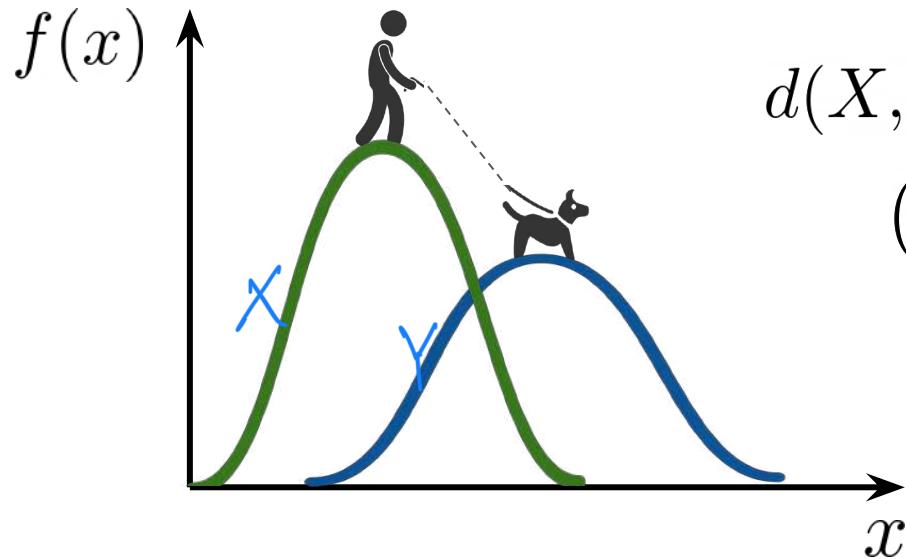
Fréchet Distance



Fréchet Distance Between Normal Distributions



Fréchet Distance Between Normal Distributions



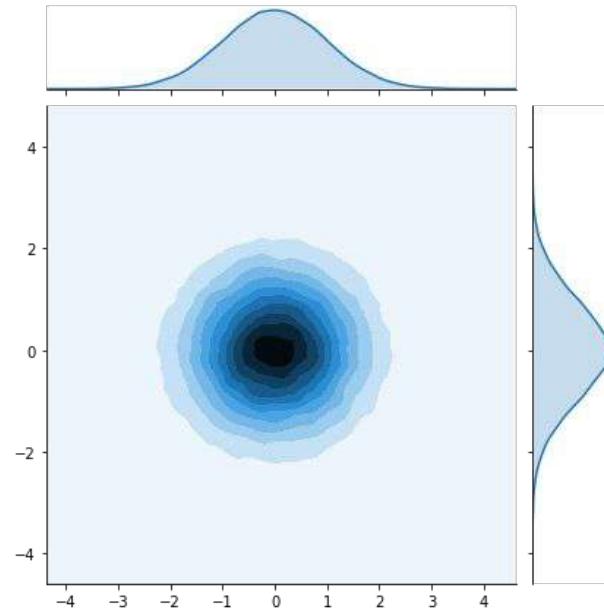
$$d(X, Y) =$$

$$(\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2$$

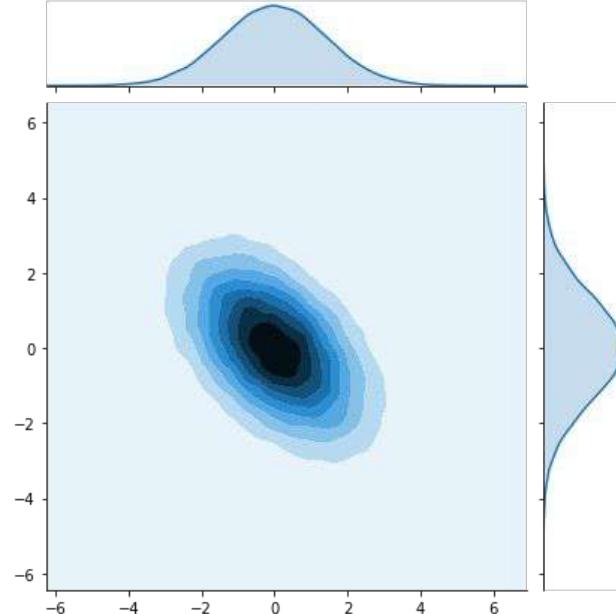
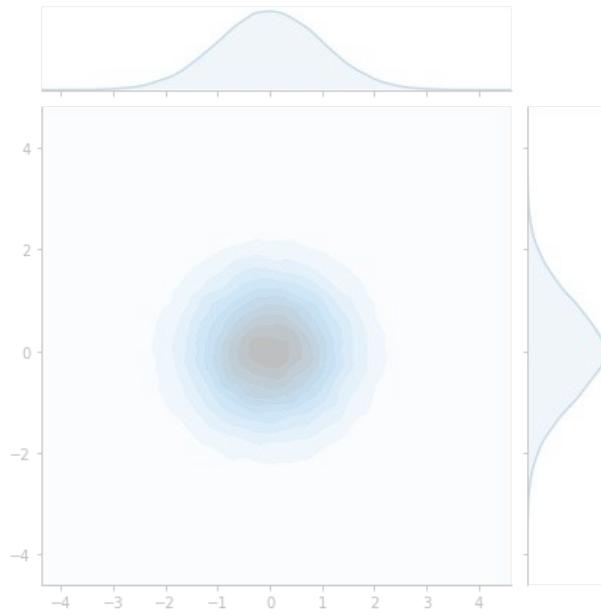
Mean

Standard deviation

Multivariate Normal Distributions

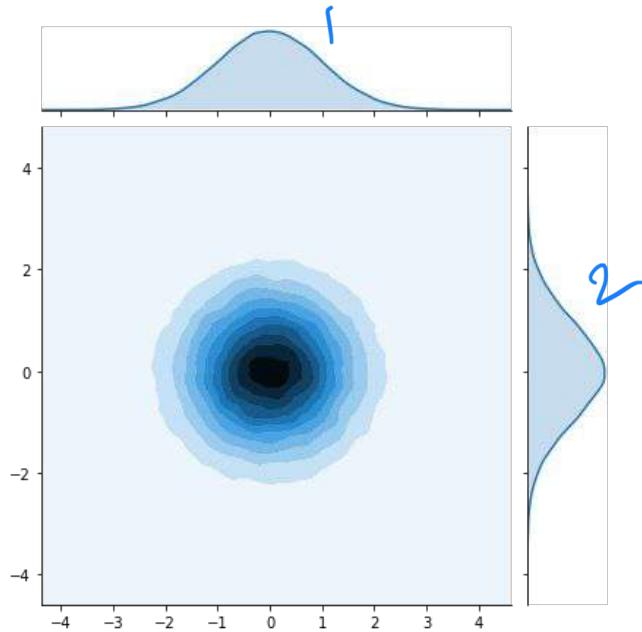


Multivariate Normal Distributions



Multivariate Normal Distributions

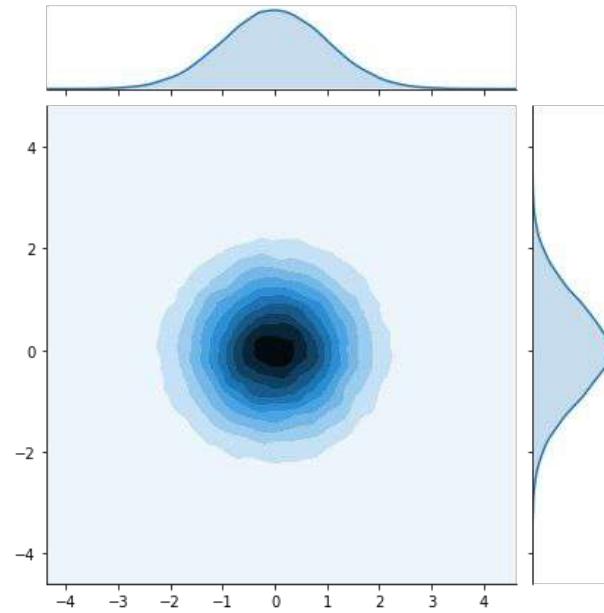
σ^2 variance



$$\Sigma = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$$

Covariance
matrix

Multivariate Normal Distributions



0's everywhere but the diagonal =
all dimensions are *independent*

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

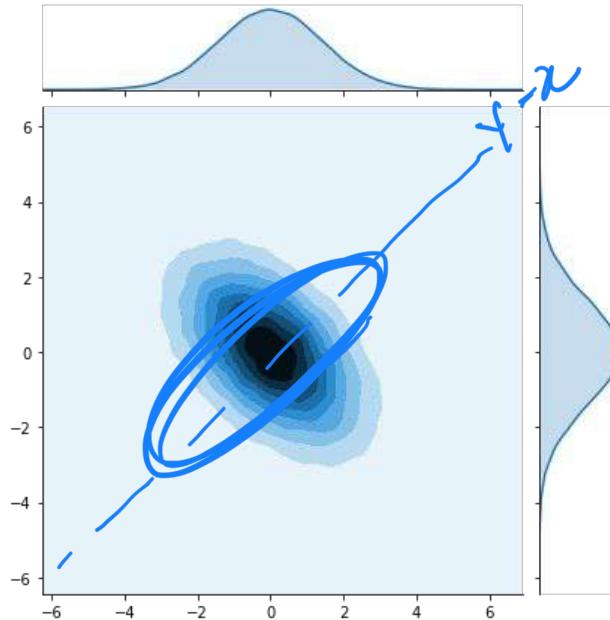
Covariance
matrix

Multivariate Normal Distributions

Non-0's not on the diagonal =
dimensions **covary**

$$\Sigma = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$$

Covariance
matrix



Multivariate Normal Fréchet Distance

Univariate Normal Fréchet Distance =

$$(\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2$$

Multivariate Normal Fréchet Distance =

$$\|\mu_X - \mu_Y\|^2 + \text{Tr} \left(\Sigma_X + \Sigma_Y - 2\sqrt{\Sigma_X \Sigma_Y} \right)$$

Can matrix instead of signs
because of "multi" variant

Multivariate Normal Fréchet Distance

Univariate Normal Fréchet Distance =

$$(\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2$$

Multivariate Normal Fréchet Distance =

$$\|\mu_X - \mu_Y\|^2 + \text{Tr} \left(\Sigma_X + \Sigma_Y - 2\sqrt{\Sigma_X \Sigma_Y} \right)$$

Multivariate Normal Fréchet Distance

Univariate Normal Fréchet Distance =

$$(\mu_X - \mu_Y)^2 +$$

$$(\sigma_X^2 + \sigma_Y^2 - 2\sigma_X\sigma_Y)$$

$$(\sigma_X - \sigma_Y)^2$$

↑
expant

Multivariate Normal Fréchet Distance =

$$\|\mu_X - \mu_Y\|^2 + \text{Tr} \left(\Sigma_X + \Sigma_Y - 2\sqrt{\Sigma_X \Sigma_Y} \right)$$

0
similar

Multivariate Normal Fréchet Distance

Univariate Normal Fréchet Distance =

$$(\mu_X - \mu_Y)^2 + (\sigma_X^2 + \sigma_Y^2 - 2\sigma_X\sigma_Y)$$

Multivariate Normal Fréchet Distance =

$$\|\mu_X - \mu_Y\|^2 + \text{Tr} \left(\Sigma_X + \Sigma_Y - 2\sqrt{\Sigma_X \Sigma_Y} \right)$$

trace ~~tr $\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$~~ = 4

Fréchet Inception Distance (FID)

FID =

$$\|\mu_X - \mu_Y\|^2 + \text{Tr} \left(\Sigma_X + \Sigma_Y - 2\sqrt{\Sigma_X \Sigma_Y} \right)$$

Real and fake embeddings are two
multivariate normal distributions

Fréchet Inception Distance (FID)

FID =

$$\|\mu_X - \mu_Y\|^2 + \text{Tr} \left(\Sigma_X + \Sigma_Y - 2\sqrt{\Sigma_X \Sigma_Y} \right)$$

Lower FID = closer
distributions

Real and fake embeddings are two
multivariate normal distributions

Fréchet Inception Distance (FID)

γ : fake
 α : real

FID =

$$\|\mu_X - \mu_Y\|^2 + \text{Tr} \left(\Sigma_X + \Sigma_Y - 2\sqrt{\Sigma_X \Sigma_Y} \right)$$

Real and fake embeddings are two multivariate normal distributions

Real embed

[] $\times 50000$

vs.

not specific range like $[0,1]$ for FID!

Lower FID = closer
distributions

Use **large sample size** to reduce noise

fake embed

[] $\times 50000$

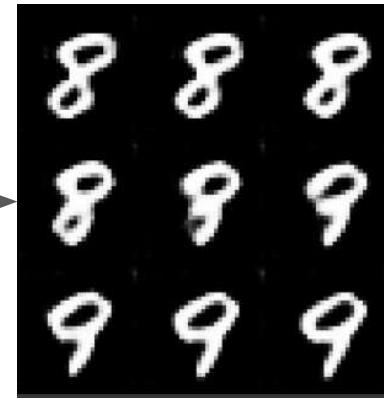


Shortcomings of FID

- Uses pre-trained Inception model, which may not capture all features



?



ImageNet doesn't
have digits

Shortcomings of FID

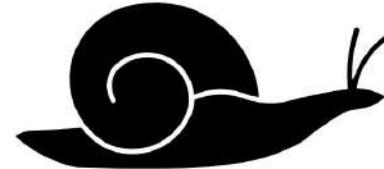
- Uses pre-trained Inception model, which may not capture all features
- Needs a large sample size



© 2016
Stanford
Vision Lab

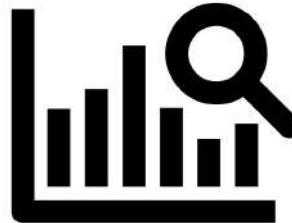
Shortcomings of FID

- Uses pre-trained Inception model, which may not capture all features
- Needs a large sample size
- Slow to run



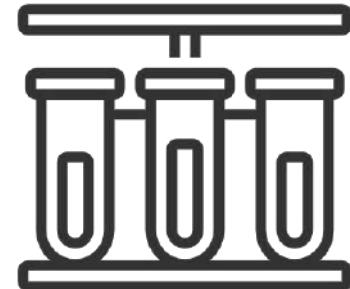
Shortcomings of FID

- Uses pre-trained Inception model, which may not capture all features
- Needs a large sample size
- Slow to run
- Limited statistics used: only mean and covariance



Summary

- FID calculates the difference between reals and fakes
- FID uses the Inception model and multivariate normal Fréchet distance
- Sample size needs to be large for FID to work well





deeplearning.ai

Inception Score

another way to calculate
dist between real & fake

Now is largely replaced by FID

Outline

- Another evaluation metric: Inception Score (IS)
 - Intuition, notation, shortcomings

Use Inception classifier
not middle features



Inception Model Classification



0.30 Cat

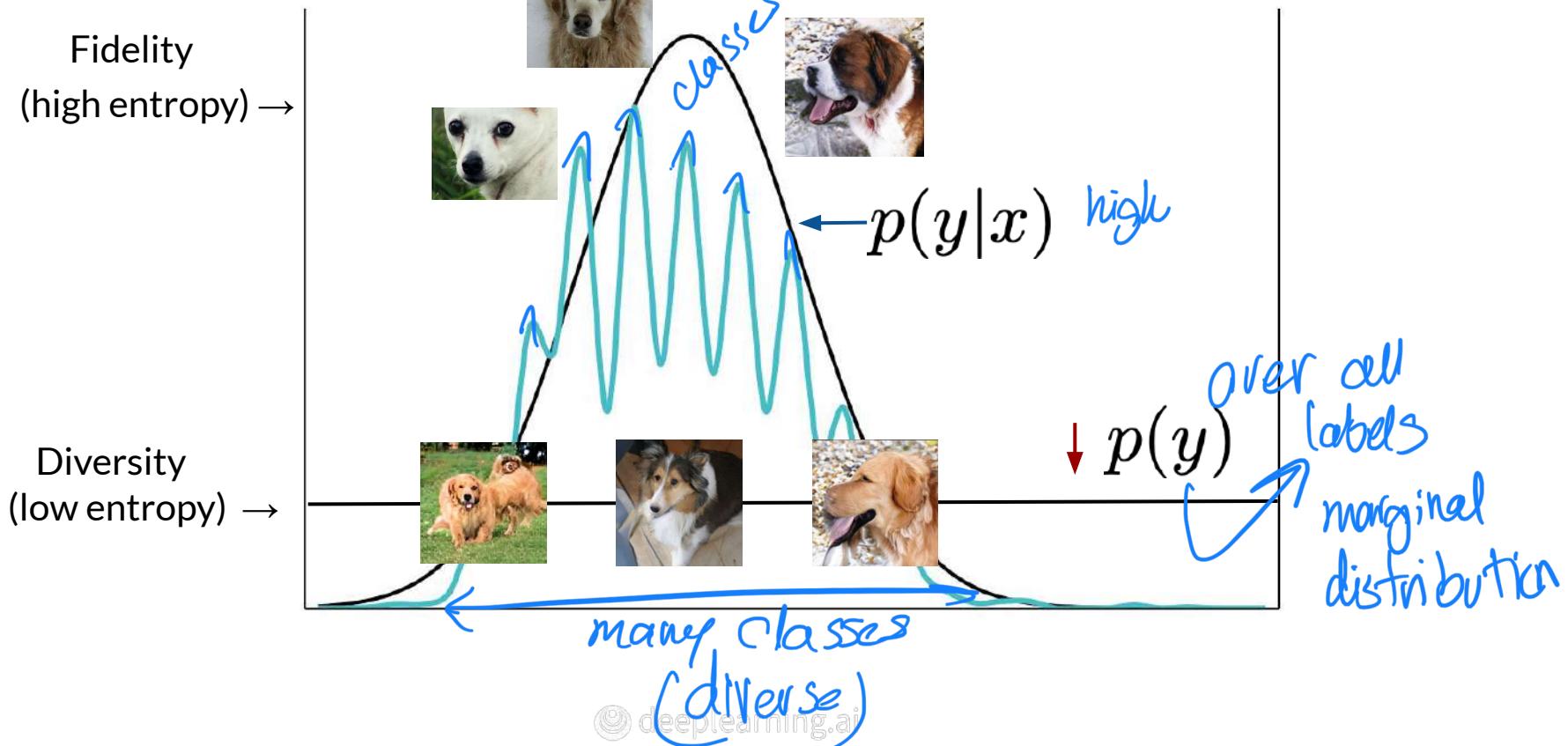
0.60 Dog

0.10 Bird

high
fidelity

Fake

Entropy



KL Divergence

$$D_{KL}(p(y|x) \| p(y)) =$$

$$p(y|x) \log \left(\frac{p(y|x)}{p(y)} \right)$$

Conditional distribution
(fidelity)

Marginal distribution
(diversity)

$p(y|x)$ good classes
 $p(y)$ many classes

high KL
when dist
are high

Inception Score (IS)

$$\text{IS} = \exp(\mathbb{E}_{x \sim p_\varepsilon} D_{KL}(p(y | x) || p(y)))$$

nice
human
readable
Score !

avg over
all classes

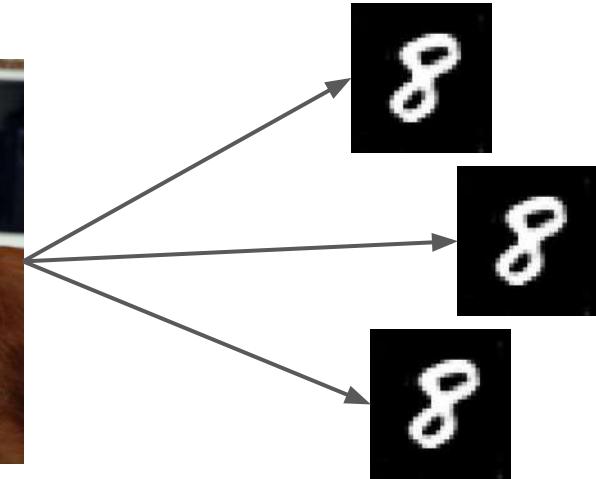
KL Divergence

Shortcomings of IS

- Can be exploited or gamed
 - Generate one realistic image of each class

but we can do better
→ ideal score!

This is the
only 8 I make



Shortcomings of IS

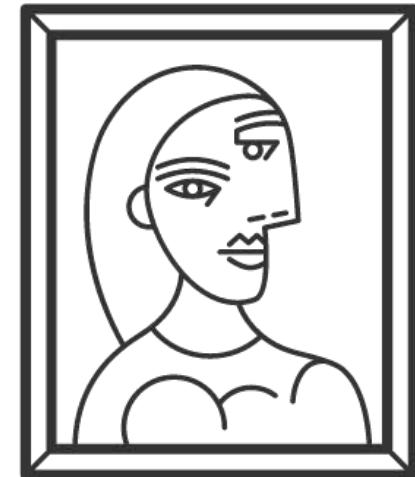
IS doesn't look
at real images

- Can be exploited or gamed
 - Generate one realistic image of each class
- Only looks at fake images
 - No comparison to real images

$$p(y|x) \quad p(y)$$

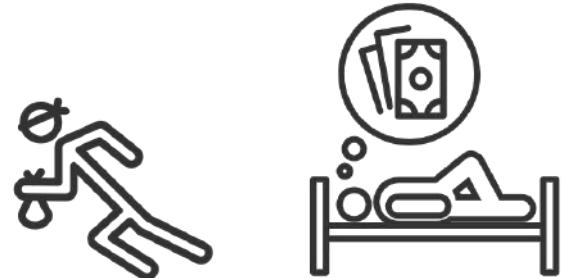
Shortcomings of IS

- Can be exploited or gamed
 - Generate one realistic image of each class
- Only looks at fake images
 - No comparison to real images
- Can miss useful features
 - ImageNet isn't everything



Summary

- Inception Score tries to capture fidelity & diversity
- Inception Score has many shortcomings
 - Can be gamed too easily
 - Only looks at fake images, not reals
 - ImageNet doesn't teach a model all features
- Worse than Fréchet Inception Distance





deeplearning.ai

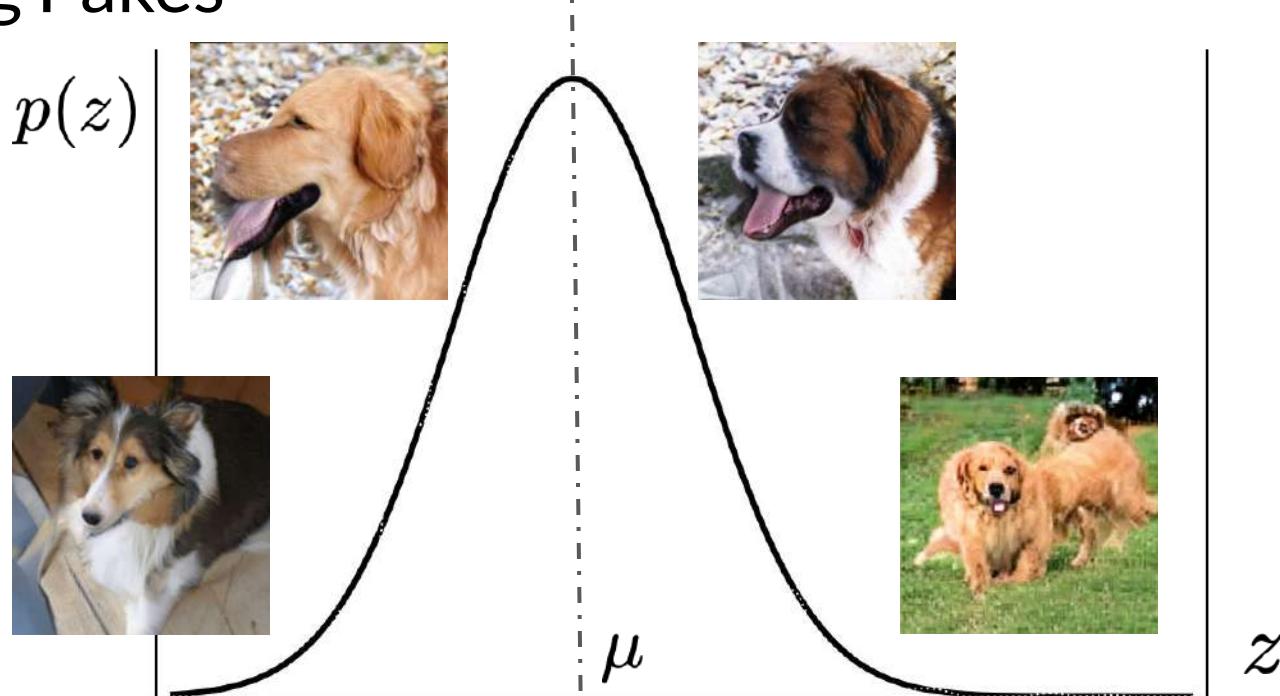
Sampling and Truncation

Outline

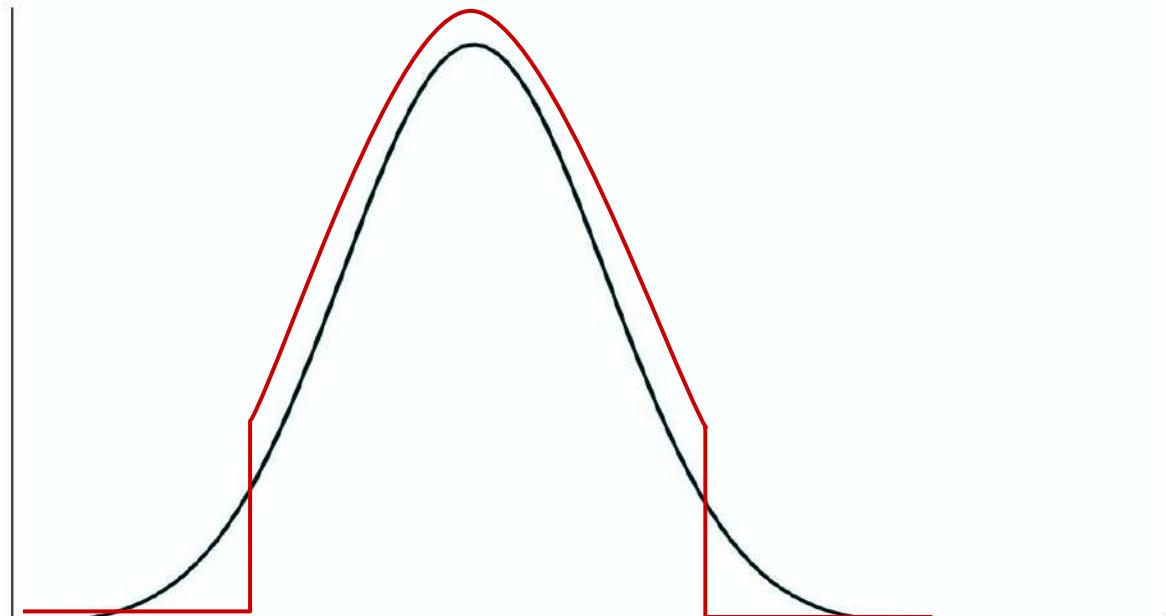
- Sampling reals vs. fakes
- The truncation trick
- HYPE!



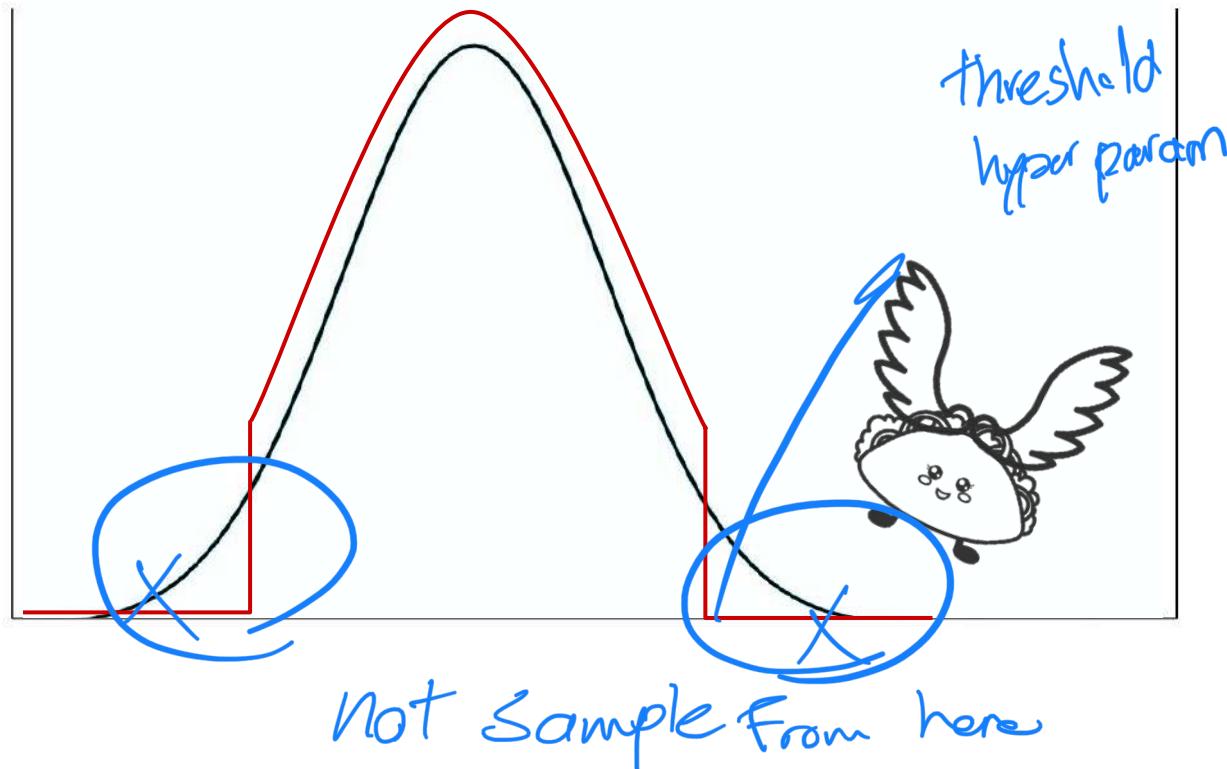
Sampling Fakes



Truncation Trick



Truncation Trick



HYPE and Human Evaluation

- Crowdsourced evaluation from Amazon Mechanical Turk
- $\text{HYPE}_{\text{time}}$ measures time-limited perceptual thresholds
- HYPE_{∞} measures error rate on a percentage of images
- Ultimately, evaluation depends on the type of downstream task



Available from: <https://arxiv.org/abs/1904.01121>

Summary

- Fakes are sampled using the training or prior distribution of z
- Truncate more for higher fidelity, lower diversity
- Human evaluation is still necessary for sampling





deeplearning.ai

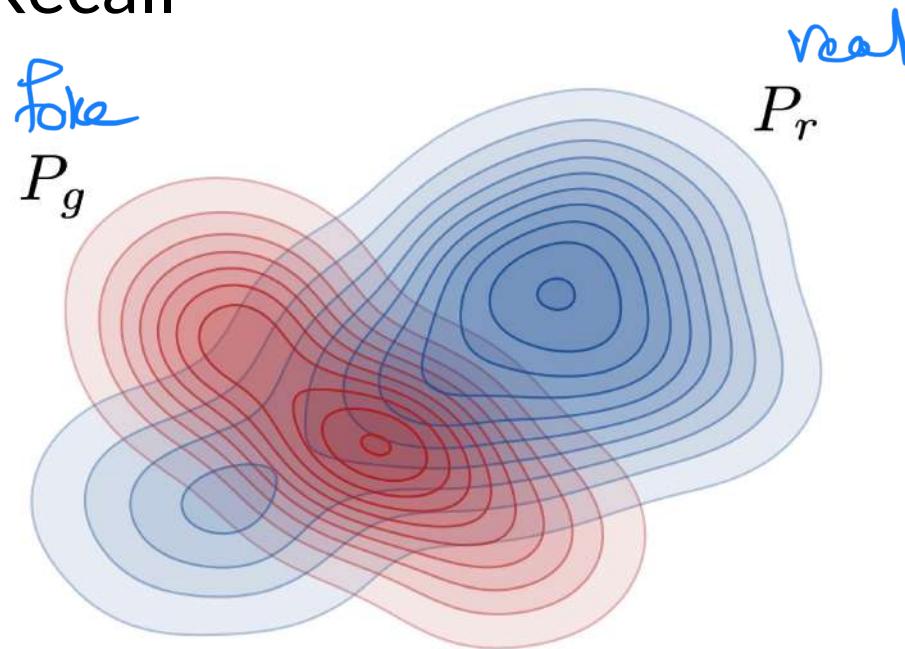
Precision and Recall

Outline

- Precision and recall in GANs evaluation
- Relating precision and recall to fidelity and diversity



Precision and Recall

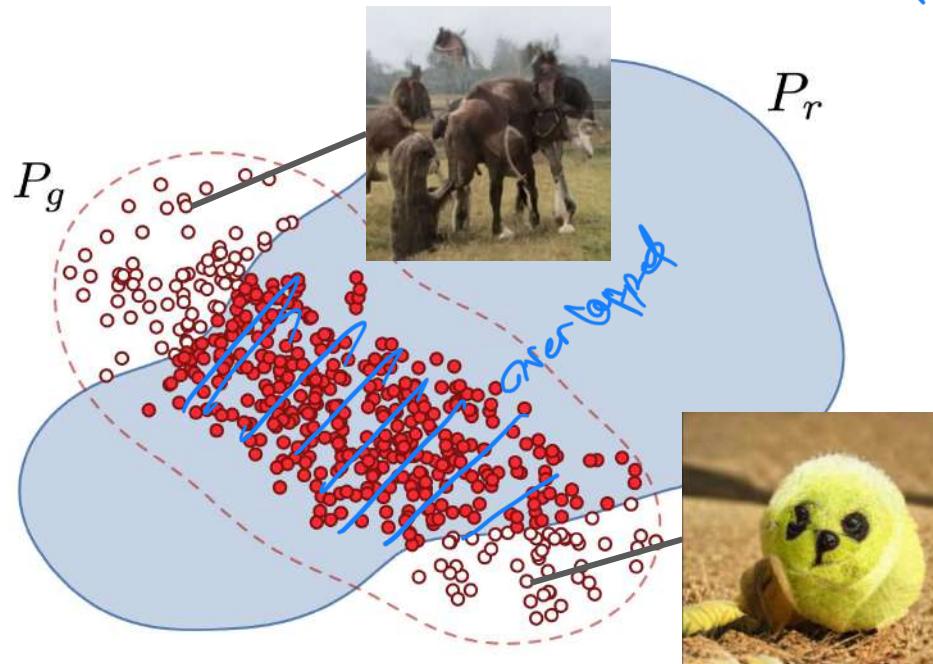


We want these two
to overlap completely

Available at: <https://arxiv.org/abs/1904.06991>

Precision

Fidelity



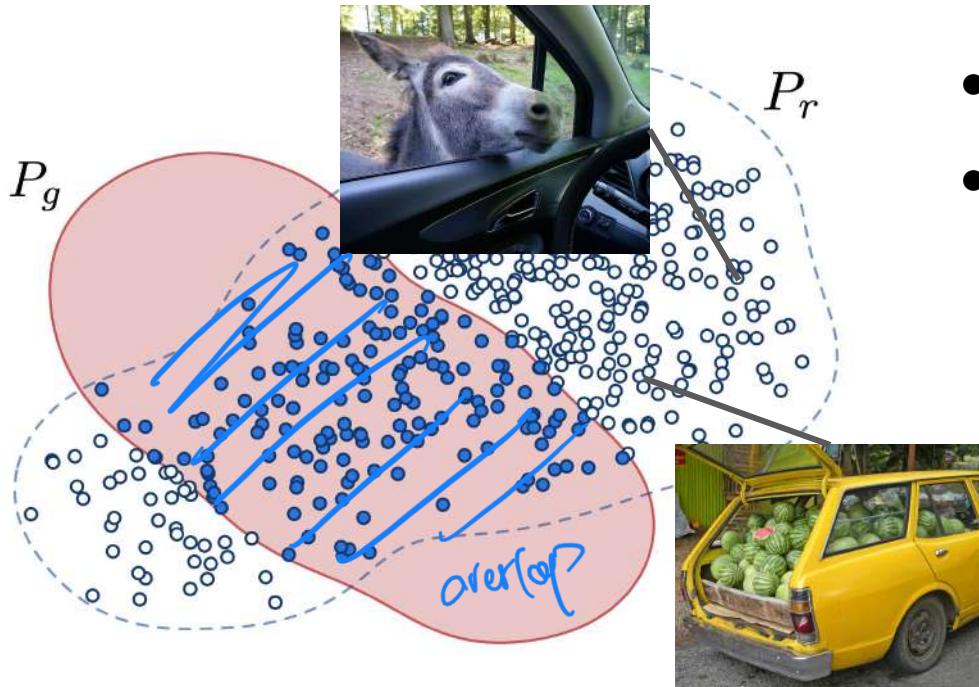
- Relates to fidelity
- Looks at overlap between reals and fakes, over how much extra gunk the generator produces (non-overlap red)

$$\text{precision} = \frac{\text{overlap part}}{\text{all fake (call } c\text{)}}$$

P_g

Diagram available at: <https://arxiv.org/abs/1904.06991>; Tennis dog available at: <https://arxiv.org/abs/1809.11096>

Recall



- Relates to diversity
- Looks at overlap between reals and fakes, over all the reals that the generator cannot model (non-overlap blue)

$$\text{recall} = \frac{\text{overlap}}{\text{all real}}$$

all real
all P_g

Diagram available at: <https://arxiv.org/abs/1904.06991>

Summary

often recall is high
& precision is low

- Precision is to fidelity as to recall is to diversity
- Models tend to be better at recall
- Use truncation trick to improve precision

