



Supervised Learning

Rowel Atienza, PhD

University of the Philippines

github.com/roatienza

2023

*A computer program is said to learn from **experience E** with respect to some class of **tasks T**, and **performance measure P**, if its performance at tasks in T , as measured by P , improves with experience E . – Tom Mitchell*

Supervised Learning

Experience

Task

Performance

Robot Learning

Learn from Experience (E) to perform Task (T) by improving the Performance (P)



<https://spectrum.ieee.org/covariant-ai-gigantic-neural-network-to-automate-warehouse-picking>

Supervised Learning

Experience: Dataset $\mathcal{D}_{train} = \{\mathbf{x}_n, y_n\}$ with N examples: input $\mathbf{x}_n \in \mathbb{R}^D$ with corresponding label $y_n \in \mathbb{R}$

Task: Classifier $y = f(\mathbf{x}, \boldsymbol{\theta}) : \mathbb{R}^D \rightarrow \mathbb{R}$ where $\boldsymbol{\theta}$ are parameters

Ground Truth Experience: Dataset $\mathcal{D}_{test} = \{\mathbf{x}_m, y_m\}$, $\hat{y}_m = f(\mathbf{x}_m, \boldsymbol{\theta})$ for $m = 1 \dots M$

Performance: Accuracy on \mathcal{D}_{test}

$$\text{Accuracy} = \frac{\# \text{ Correct Prediction}}{\text{Size of Test Dataset}} = \frac{\sum_{m=1}^M (\hat{y}_m \text{ equals } y_m)}{M}$$

Machine Learning & Generalization

Initial Test Accuracy $Accuracy_0$ of $\hat{y}_m = f(\mathbf{x}_m, \boldsymbol{\theta}^0)$, $\boldsymbol{\theta}^0$ initial parameters

Machine Learning: Finding the optimal parameters $\boldsymbol{\theta}^*$ using $\mathcal{D}_{train} = \{\mathbf{x}_n, y_n\}$: $\boldsymbol{\theta}^0 \rightarrow \boldsymbol{\theta}^*$

$$f(\mathbf{x}_n, \boldsymbol{\theta}^*) \approx y_n \quad n = 1 \dots N$$

Final Test Accuracy $Accuracy_*$ of $\hat{y}_m = f(\mathbf{x}_m, \boldsymbol{\theta}^*)$

Machine Learning: $Accuracy_* > Accuracy_0$

Generalization: $Accuracy_{*(train)} - Accuracy_{*(test)} = Gap \rightarrow 0$

Experience

The Datasets

MNIST Dataset

$\mathcal{D}_{train} = \{\mathbf{x}_n, y_n\}, N = 60,000$

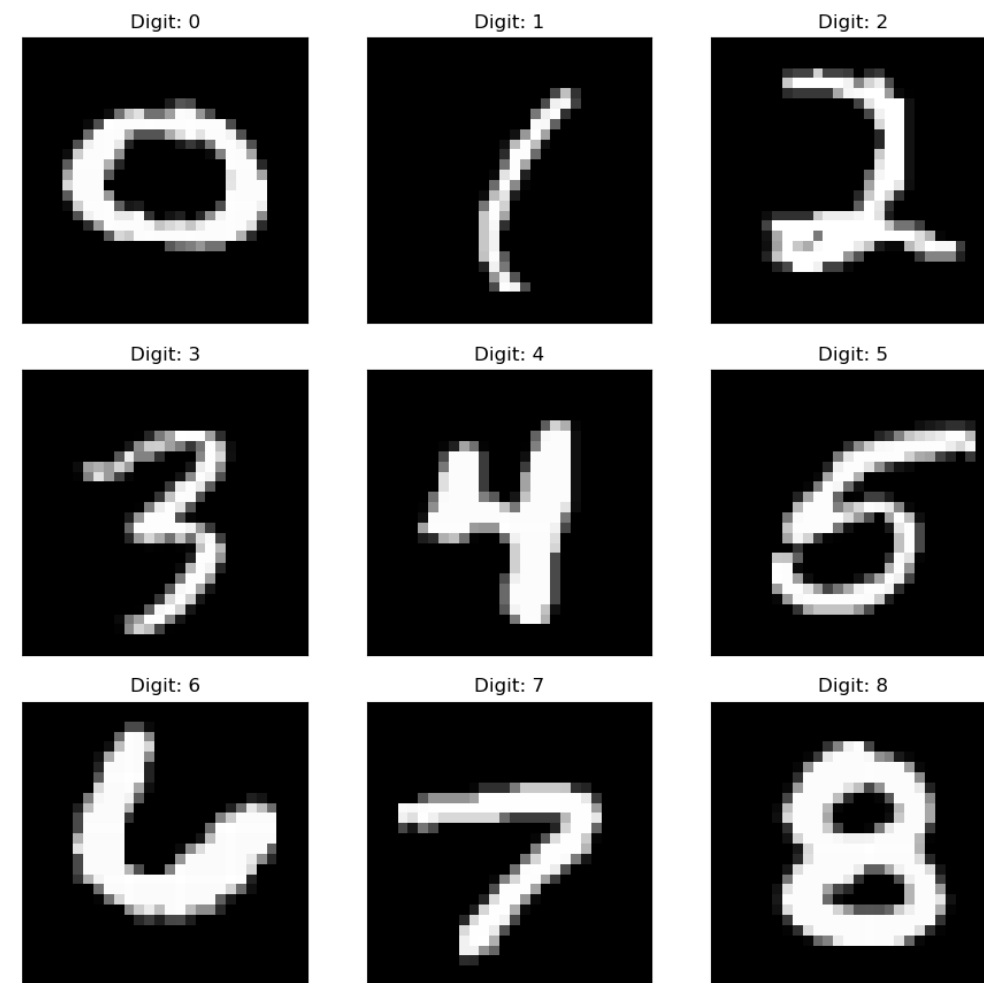
$\mathcal{D}_{test} = \{\mathbf{x}_m, y_m\}, M = 10,000$

\mathbf{x} : 28×28 grayscale images of digits 0 to 9

y : class label

Available:

`torchvision.datasets.MNIST()`



LJSpeech Dataset

$\mathcal{D}_{train} = \{\mathbf{x}_n, \mathbf{y}_n\}, N = 12,228$

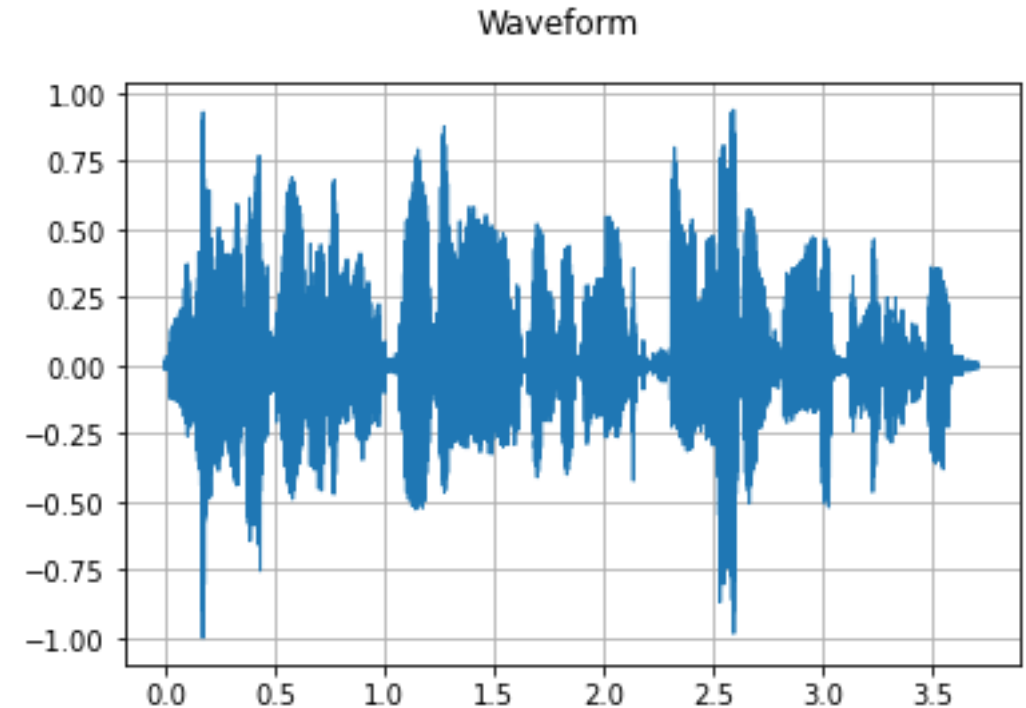
$\mathcal{D}_{test} = \{\mathbf{x}_m, \mathbf{y}_m\}, M = 523$

\mathbf{x} : speech

\mathbf{y} : text transcript

Available:

`torchaudio.datasets.LJSPEECH()`



the association was organized under the most promising auspices

Stanford Sentiment Treebank Dataset

$$\mathcal{D}_{train} = \{\mathbf{x}_n, y_n\}, N = 67,349$$

$$\mathcal{D}_{dev} = \{\mathbf{x}_p, y_p\}, P = 872$$

$$\mathcal{D}_{test} = \{\mathbf{x}_m, y_m\}, M = 1,821$$

\mathbf{x} : phrases

y : sentiment (+ or 1.0 /- or 0.0)

Available:

`torchtext.datasets.SST2()`

\mathbf{x} : The gorgeously elaborate “The Lord of the Rings” ...
 y : 0.833

LAION5B: 5B Image-Text Pairs

$$\mathcal{D}_{train} = \{\mathbf{x}_n, \mathbf{y}_n\}, N = 5B$$

\mathbf{x} : text

\mathbf{y} : image

Available:

<https://laion.ai/blog/laion-5b/>

Backend url:

<https://knn5.laion>

Index:

laion_5B

french cat



[Clip retrieval](#) works by converting the text query to a CLIP embedding, then using that embedding to query a knn index of clip image embeddings

Display captions ☒

Display full captions ☐

Display similarities ☐

Safe mode ☒

Hide duplicate urls ☒

Hide (near)

duplicate images ☒

Search over

image

Search with multilingual clip ☐



french cat



french cat



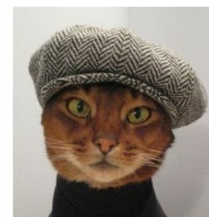
How to tell if your feline is french. He wears a b...



イケメン猫モデル「トキ・ナンタケット」がかっこいい - NAVER まとめ



Hilarious pics of funny cats! funnycatsgif.com



Hipster cat



網友挑戰「加幾筆畫出最創意貓咪圖片」，笑到岔氣之後我也手...



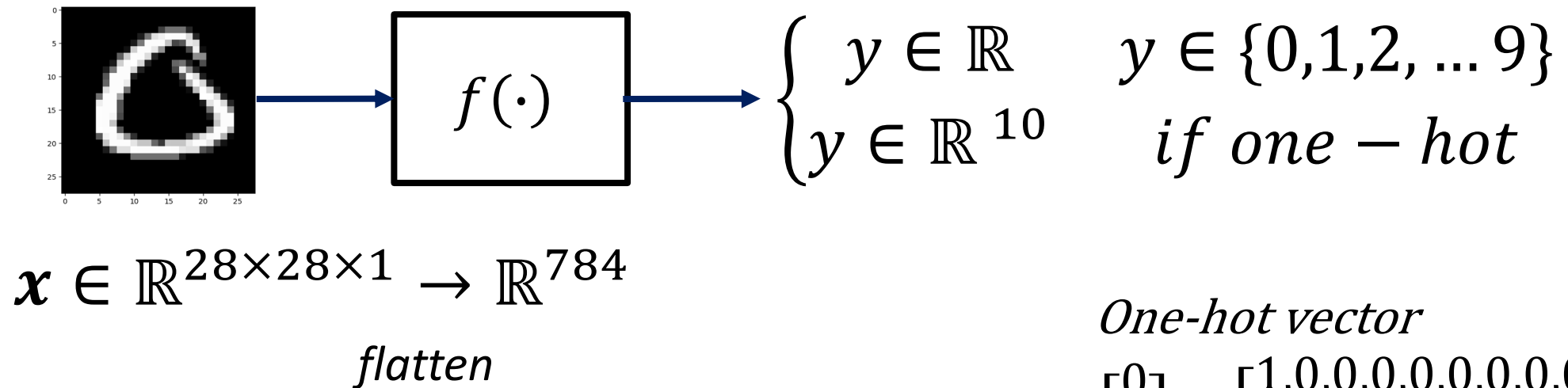
cat in a suit Georgian sells tomatoes



French Bread Cat Loaf Metal Print

Task

Multi-label Classification (Recognition)



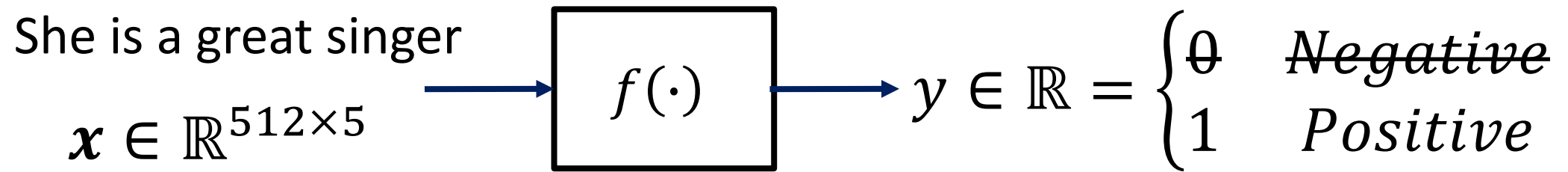
Also known as Multi-class Logistic Regression

One-hot vector

$$\begin{bmatrix} 0 \\ 1 \\ 2 \\ \vdots \\ 9 \end{bmatrix} \rightarrow \begin{bmatrix} 1, 0, 0, 0, 0, 0, 0, 0, 0, 0 \\ 0, 1, 0, 0, 0, 0, 0, 0, 0, 0 \\ 0, 0, 1, 0, 0, 0, 0, 0, 0, 0 \\ \vdots \\ 0, 0, 0, 0, 0, 0, 0, 0, 0, 1 \end{bmatrix}$$

Task: Binary Classification

Task: Sentiment Classification



Assuming embedding size is 512

Also known as Binary Logistic Regression

Task: Sequence to Sequence

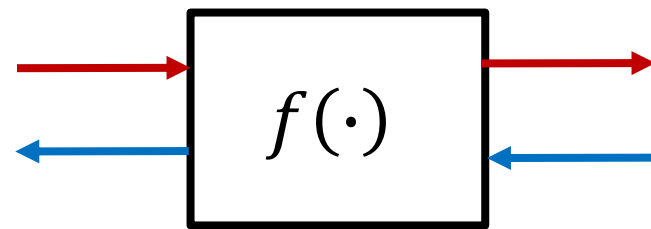
TEXT: {the association
was organized under the
most promising auspices}

→

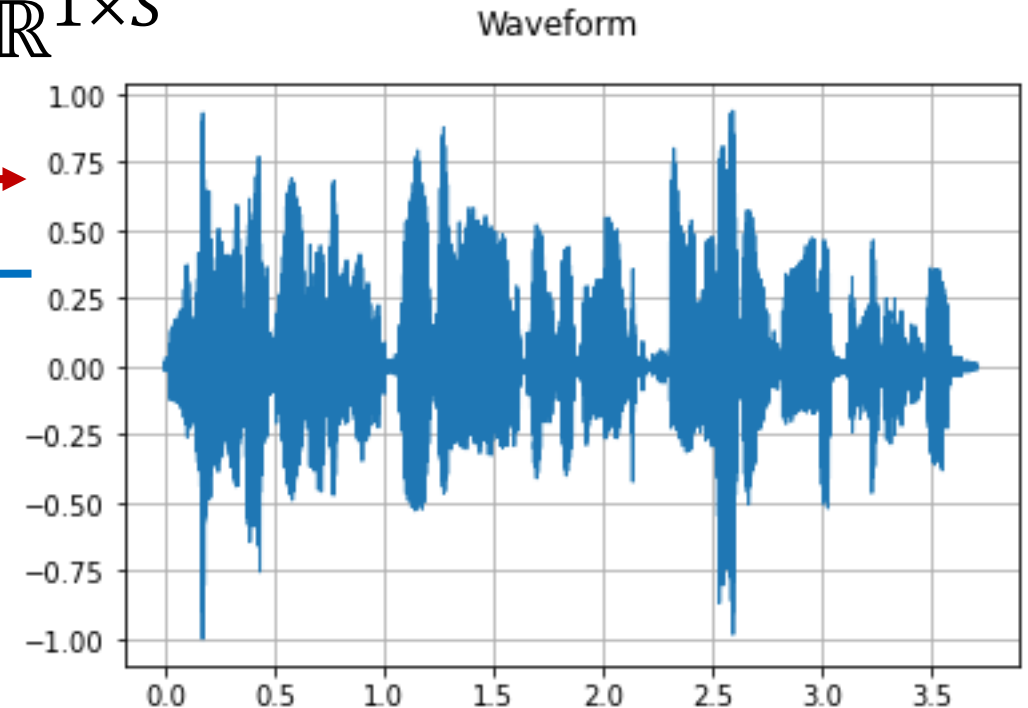
PHONEME: {DH IY0 AH0
S OW2 S IY0 EY1 SH AH0
N W AH0 Z AO1 R G AH0
N AY2 Z D AH1 N D ER0
DH AH0 M OW1 S T P R
AA1 M AH0 S IH0 NG
AO1 S P IH0 S IH0 Z}

Task: **Text to Speech**
or **Speech to Text**

$$\mathbf{x} \in \mathbb{R}^{512 \times P} \quad \mathbf{y} \in \mathbb{R}^{1 \times S}$$



P: Phoneme Length
S: #Speech Samples
*Assuming embedding
size is 512*

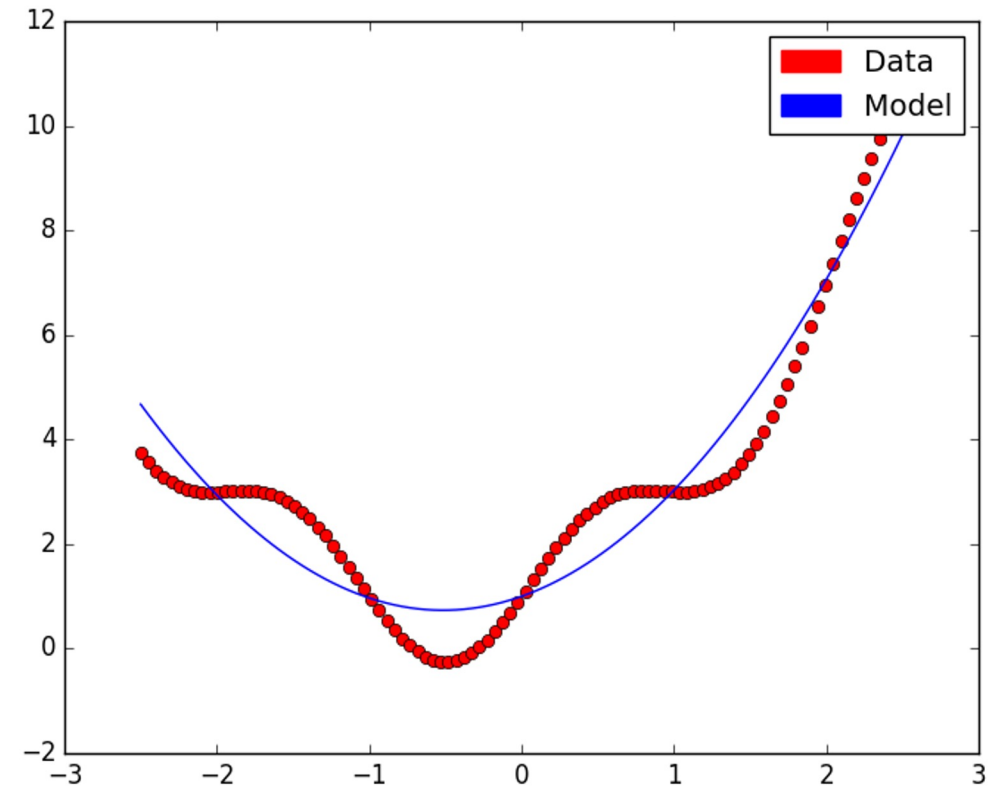


Task: Curve Fitting

$$\mathcal{D}_{train} = \{x_n, y_n\}$$

$$\mathcal{D}_{test} = \{x_m, y_m\}$$

Given: Data points



Also known as Linear Regression

Task: Image Synthesis

Input: “Anime scene of
Rihanna as a space pilot”
by @hardmaru

Output



A long, straight wooden pier made of weathered planks extends from the bottom center towards the middle of the frame, leading the eye into a calm, blue-grey lake. The water's surface is smooth, reflecting the sky and the surrounding landscape. In the background, a dense forest of tall evergreen trees lines the shore. Beyond the forest, a range of mountains is visible, with the highest peaks covered in snow and partially shrouded by soft, white clouds. The sky above is a pale blue with scattered, wispy clouds. The overall atmosphere is peaceful and serene.

Input: Image and user query
by LLaVA project

Supervised Learning

Deep Learning

```
pg  
[2023-07-29 18:32:19,906] [INFO] [real_accelerator.py:110:get_accelerator] Setting ds_acc  
elevator to cuda (auto detect)  
Loading checkpoint shards: 100%|███████████| 3/3 [00:22<00:00, 7.38s/it]  
USER:
```

Performance

Accuracy (Classification) ↑

Accuracy: Classification Performance Score on $\mathcal{D}_{test} = \{(\mathbf{x}_m, y_m)\}$

$$Accuracy = \frac{\# \text{ Correct Prediction}}{\text{Size of Test Dataset}} = \frac{\sum_{m=1}^M (\hat{y}_m \text{ equals } y_m)}{M}$$

Word Error Rate (WER) in ASR

$$WER = \frac{\textit{Substitution} + \textit{Insertion} + \textit{Deletion}}{\textit{Number of Spoken Words}}$$

Ground Truth: **It is the shining armor. Clear as night and day.**

Substitution: Replaced word(s) (e.g. *night* by *knight*)

Insertion: Added word(s) that is not there (e.g. instead of *the shining armor*, model transcribed it as *the knight and shining armor*)

Deletion: Omitted word(s) (e.g. instead of *clear as night and day*, model transcribed it as *clear night and day*)

Generative Model (Voice/Video) ↑

Mean Opinion Score (MOS) is a numerical measure of the human-judged overall quality of an event or experience.

5 Excellent

4 Good

3 Fair

2 Poor

1 Bad

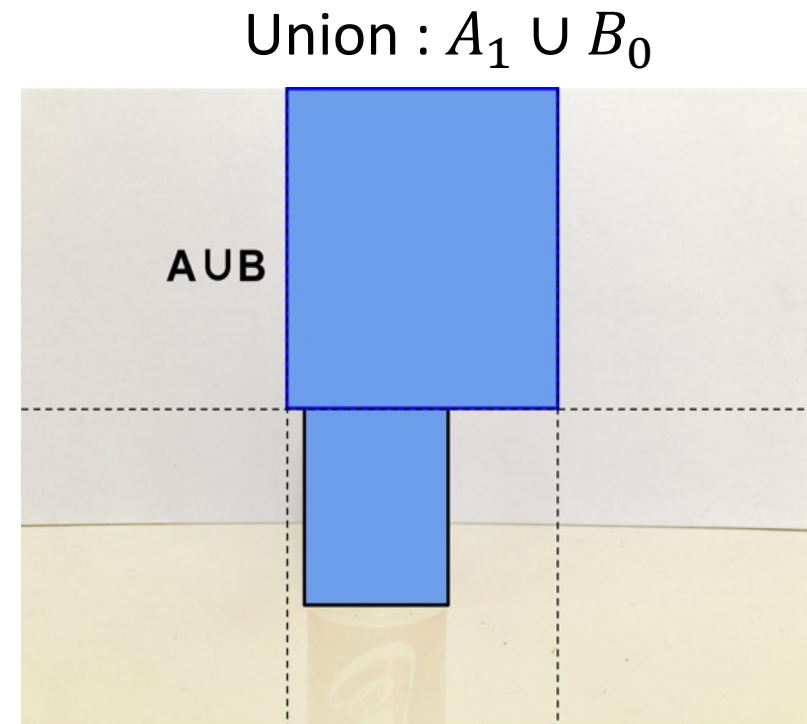
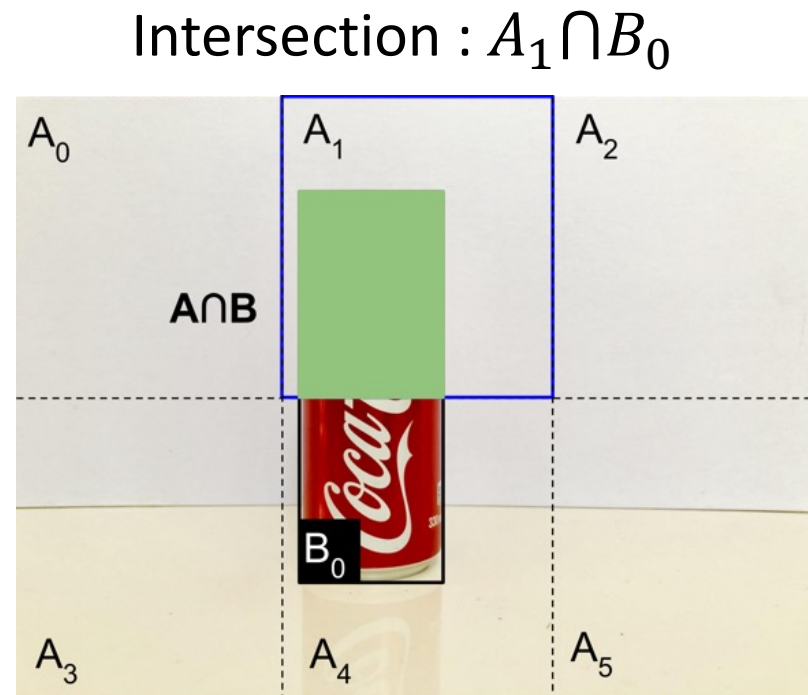
Somewhere around 4.3 - 4.5 is considered an excellent quality target. Video quality becomes unacceptable below a MOS of roughly 3.5.

<https://www.twilio.com/>

Object Detection ↑

Intersection over Union (IoU)

IoU is also known as *Jaccard index*:
$$IoU = \frac{A \cap B}{A \cup B}$$



Object Detection & Classification ↑

Precision (Bad guys out): ↑

$$precision = \frac{tp}{tp + fp} = \frac{True_Positive}{True_Positive + False_Positive}$$



Bad guys in

Recall (Good guys in): ↑

$$recall = \frac{tp}{tp + fn} = \frac{True_Positive}{True_Positive + False_Negative}$$



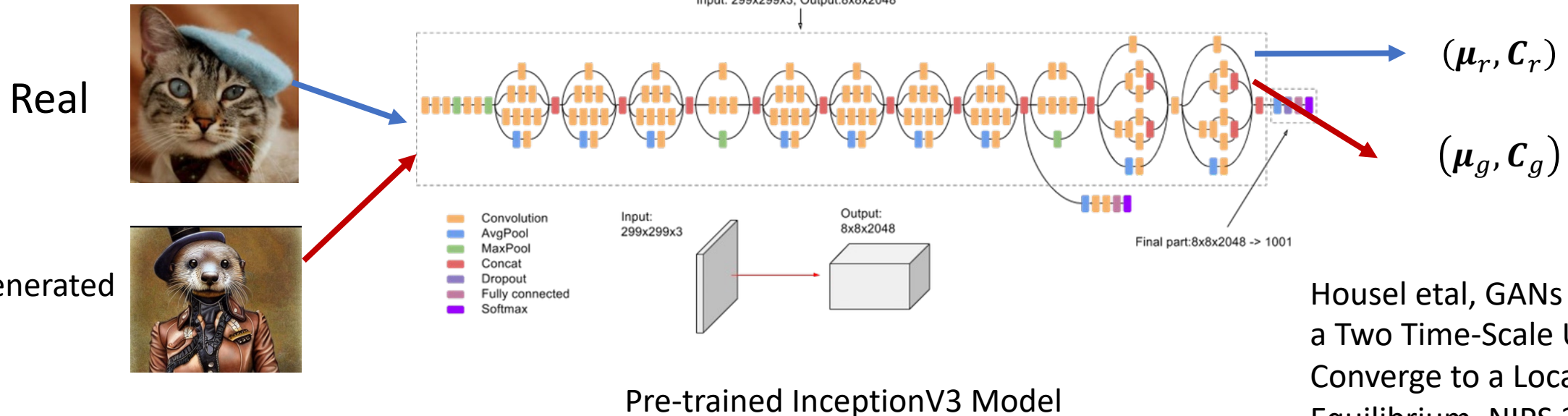
Good guys out

Related Performance Measures: Average Precision, F1 Score

Frechet Inception Distance - d

$$d^2 = \|\boldsymbol{\mu}_r - \boldsymbol{\mu}_g\|_2^2 + \text{Tr}\left(\mathbf{C}_r + \mathbf{C}_g - 2\sqrt{\mathbf{C}_r\mathbf{C}_g}\right)$$

Feature vectors Gaussian mean and covariance of real and generated images: $(\boldsymbol{\mu}_r, \mathbf{C}_r)$, $(\boldsymbol{\mu}_g, \mathbf{C}_g)$



Housel et al, GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, NIPS 2017

Evaluating Assistants

- Multimodal QA: GPT4 evaluates the helpfulness, relevance, accuracy, and level of details of the responses from the assistants, and give an overall score on a scale of 1 to 10
- ScienceQA: Accuracy

Learning

Learning by Minimizing a Loss Function

Ground truth: y_n

Prediction: \hat{y}_n

Loss Function: $\mathcal{L} = l(y_n, \hat{y}_n)$

Machine Learning: Use dataset, $\mathcal{D}_{train} = \{(\mathbf{x}_n, y_n)\}$, $n = 1 \dots N$ to estimate the parameters $\boldsymbol{\theta}$ by minimizing $L = l(y_n, \hat{y}_n)$ using an **optimizer**

Assumption: $(\mathbf{x}_1, y_1) \dots (\mathbf{x}_n, y_n)$ are IID

IID: Independent and identically distributed

Empirical Risk Minimization (ERM)

$$R_{emp}(f, \mathbf{X}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^n l(y_n, \hat{y}_n)$$

Where $\mathbf{X} := [x_1, \dots, x_n]^T \in \mathbb{R}^{N \times D}$ and $\mathbf{y} := [y_1, \dots, y_n]^T \in \mathbb{R}^N$

Least Squares Loss

$$\mathcal{L} = l(y_n, \hat{y}_n) = (y_n - \hat{y}_n)^2$$

$$R_{emp}(f, \mathbf{X}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^n (y_n - \hat{y}_n)^2 = \frac{1}{N} \sum_{i=1}^n (y_n - f(\mathbf{x}_n, \boldsymbol{\theta}))^2$$

Loss Functions

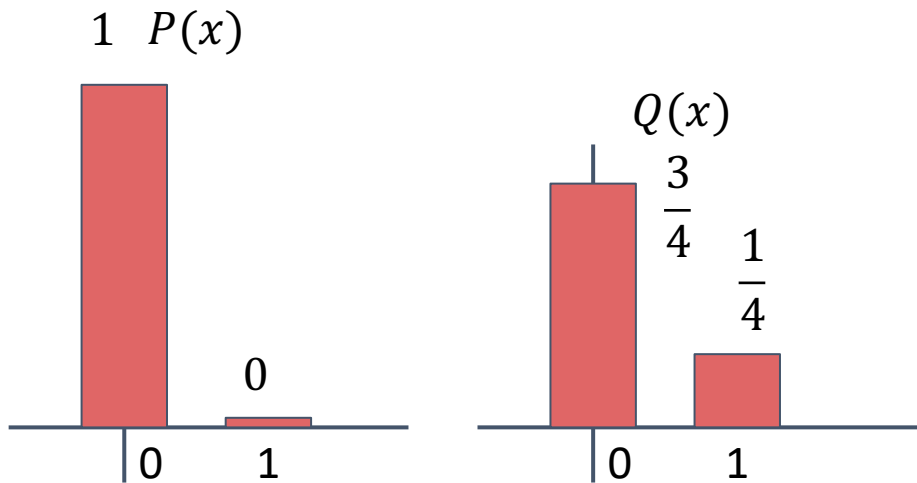
Loss Function	Equation
Mean Squared Error (MSE)	$\sum_{i=1}^{categories} (y_i^{label} - y_i^{prediction})^2$
Mean Absolute Error (MAE)	$\sum_{i=1}^{categories} y_i^{label} - y_i^{prediction} $
Categorical Cross Entropy (CE)	$- \sum_{i=1}^{categories} y_i^{label} \log y_i^{prediction}$
Binary Cross Entropy (BCE)	$-y_1^{label} \log y_1^{prediction} - (1 - y_1^{label}) \log(1 - y_1^{prediction})$

Categorical Cross-Entropy (CE)

For discrete distribution, Categorical Cross-Entropy is:

$$CE = H(P, Q) = - \underbrace{\sum_i P(x_i)}_{\text{Empirical Label}} \underbrace{\log Q(x_i)}_{\text{Predicted Label}}$$

Empirical Label Predicted Label



Supervised Learning

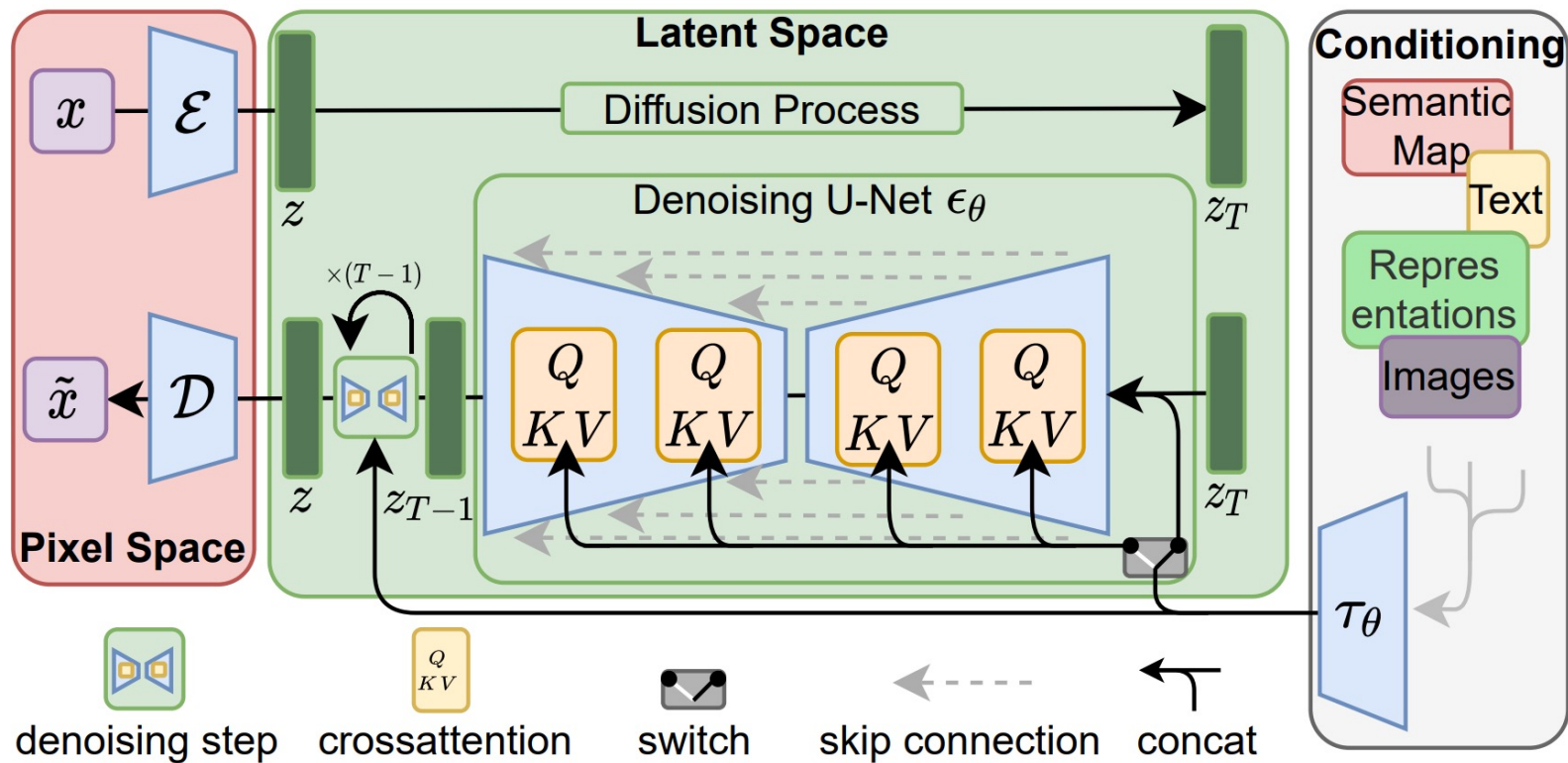
What is CE?

$$CE = - \left(1 \log \frac{3}{4} + 0 \log \frac{1}{4} \right) = 0.29$$

Minimizing $H(P, Q)$ minimizes the distance of prediction model Q from the empirical model P

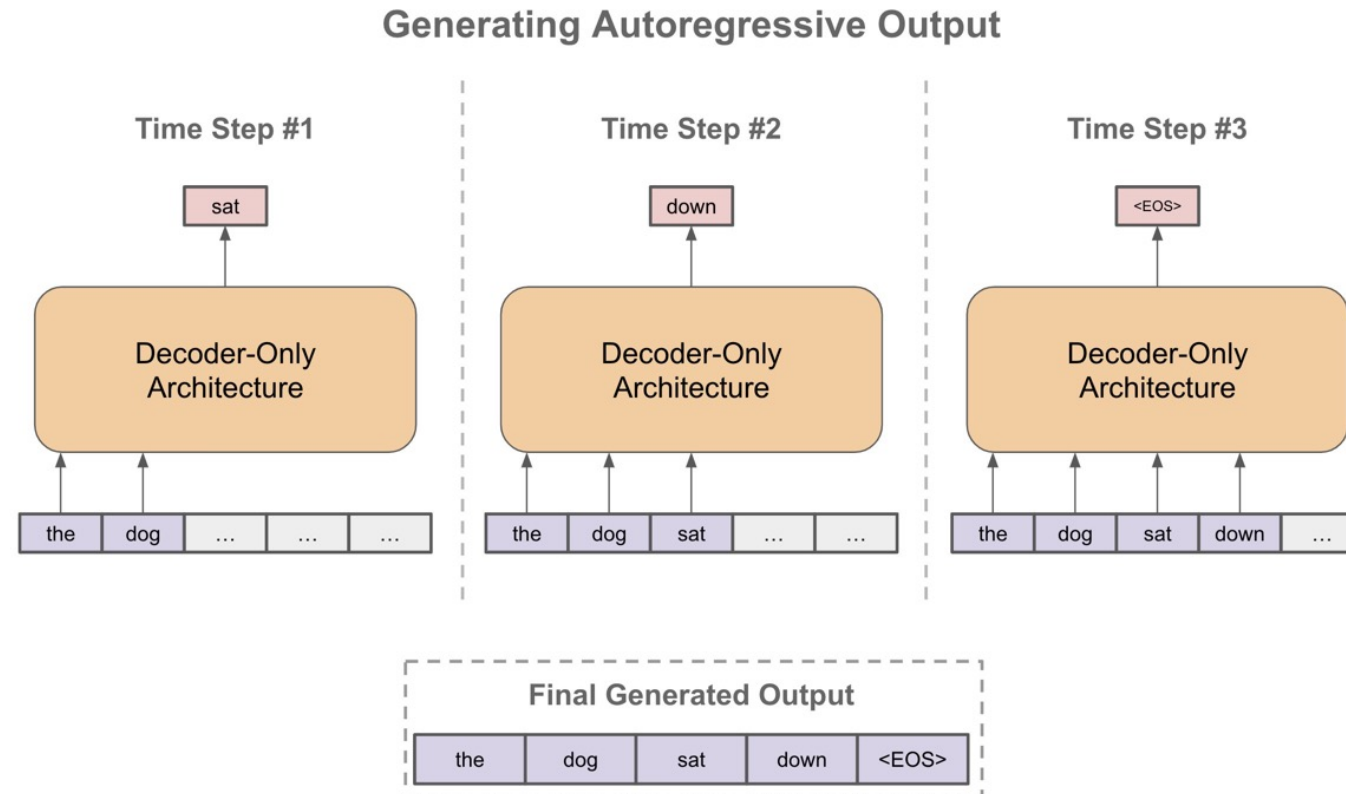
Latent Diffusion Model (LDM) Loss

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0,1), t} \left[\|\epsilon - \epsilon_{\theta}(z_t, t, \tau_{\theta}(y))\|_2^2 \right]$$



Rombach et al, High-Resolution Image Synthesis with Latent Diffusion Models, CVPR2021

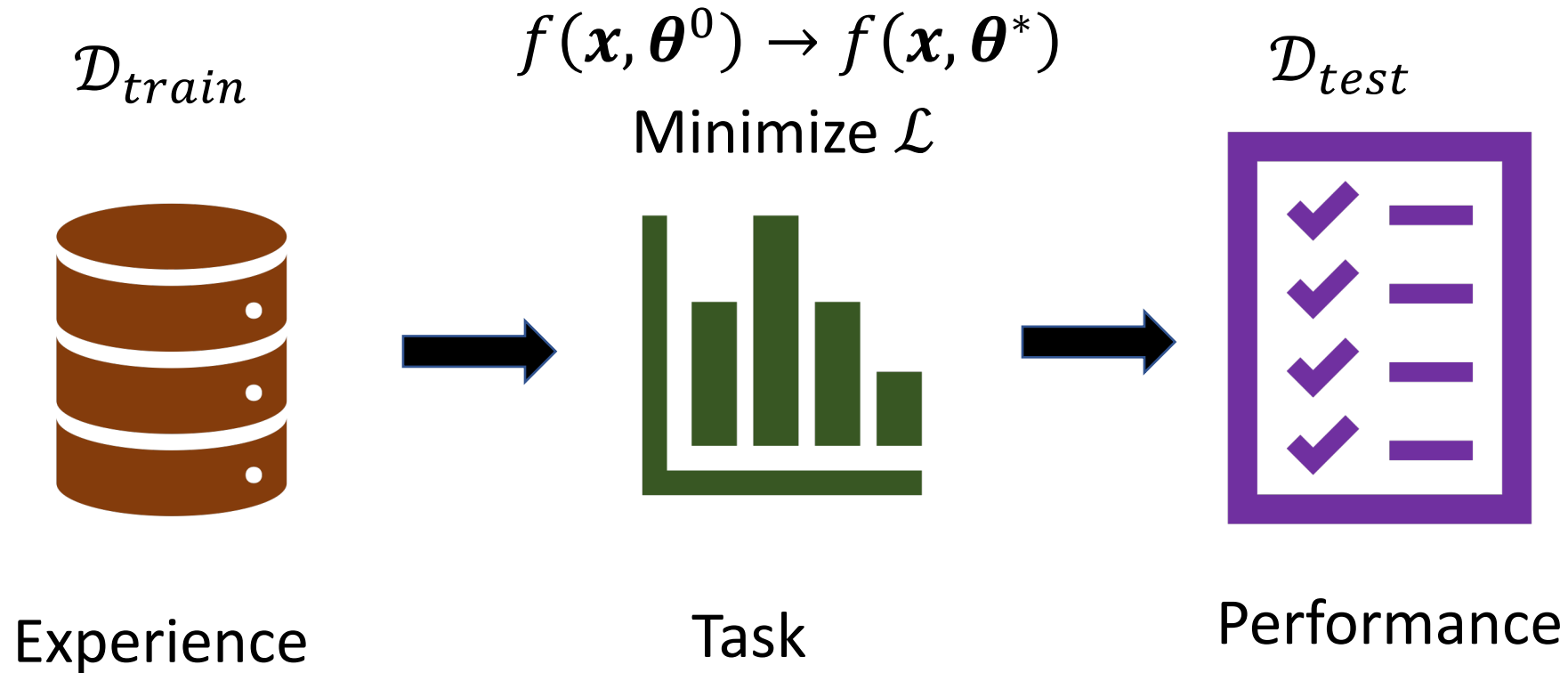
GPT CE Loss



GPT2: Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

GPT3: Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

Machine Learning Pipeline



End