Carnegie Mellon University

Losses Part 2

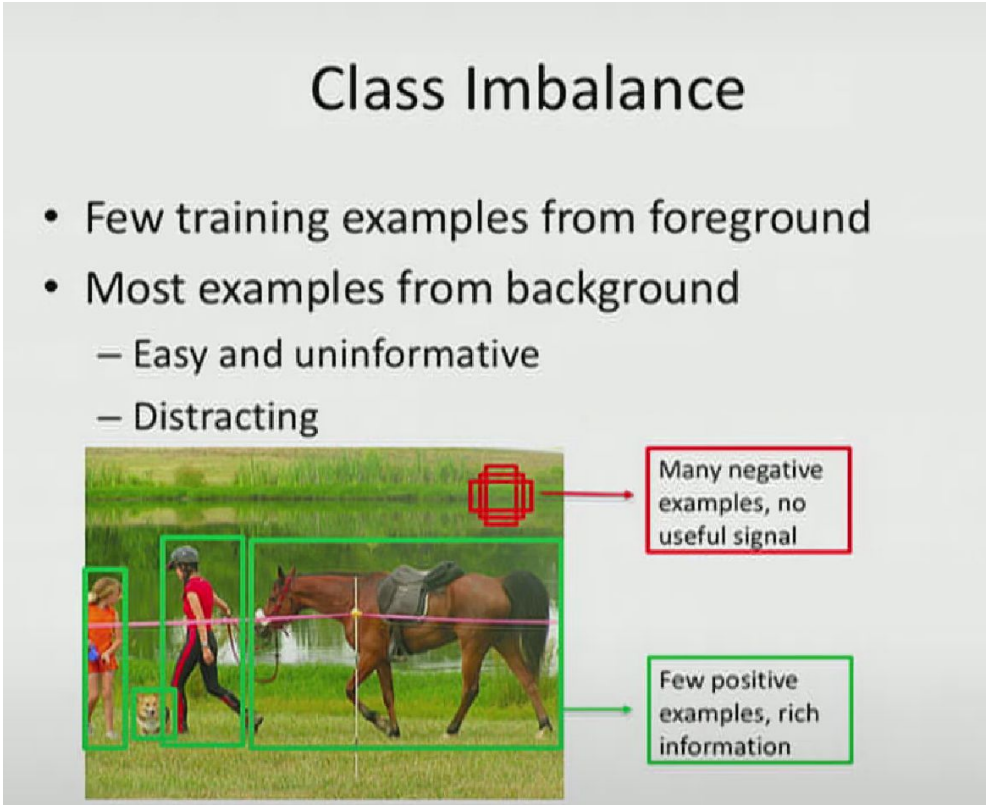
Contents

- Loss Functions for Image Classification and Segmentation
- Loss Functions for Sequence Decoding
- Loss Functions for Image Generation
- Loss Functions for Point-Clouds
- Loss Functions for LLM fine-tuning

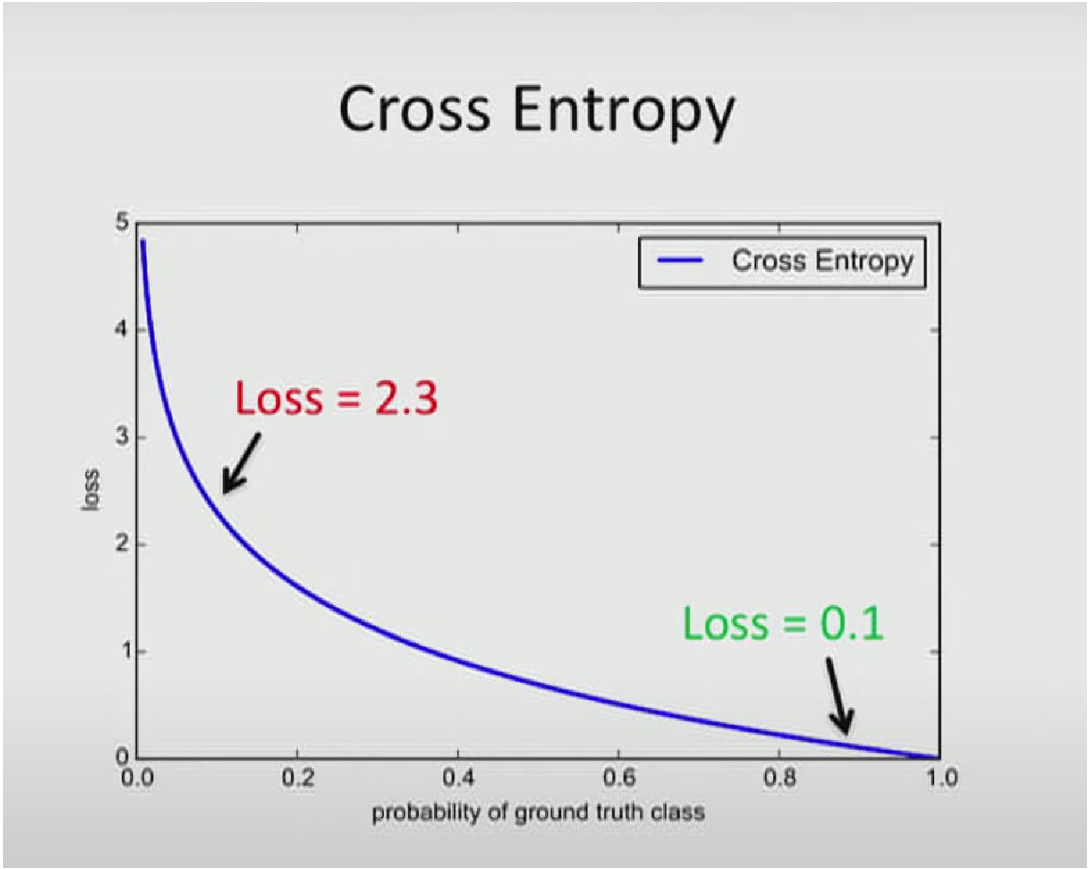
TA: Sreeharsha Paruchuri (sparuchu@cs.cmu.edu)

Loss for Object Detection: Focal Loss

- Dense Single-Shot Detectors (SSDs) predict 100s of bounding boxes per image
- These various boxes allow the network to classify different shaped objects

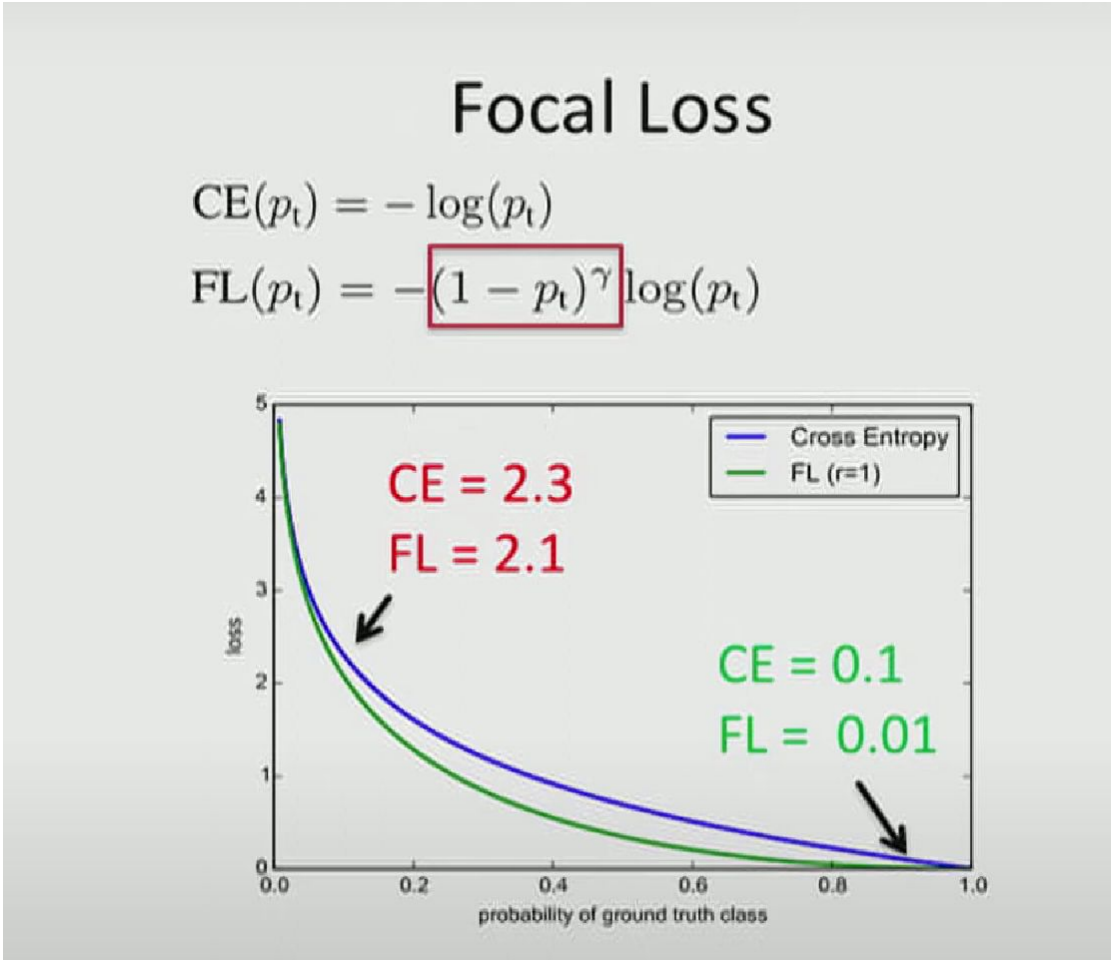


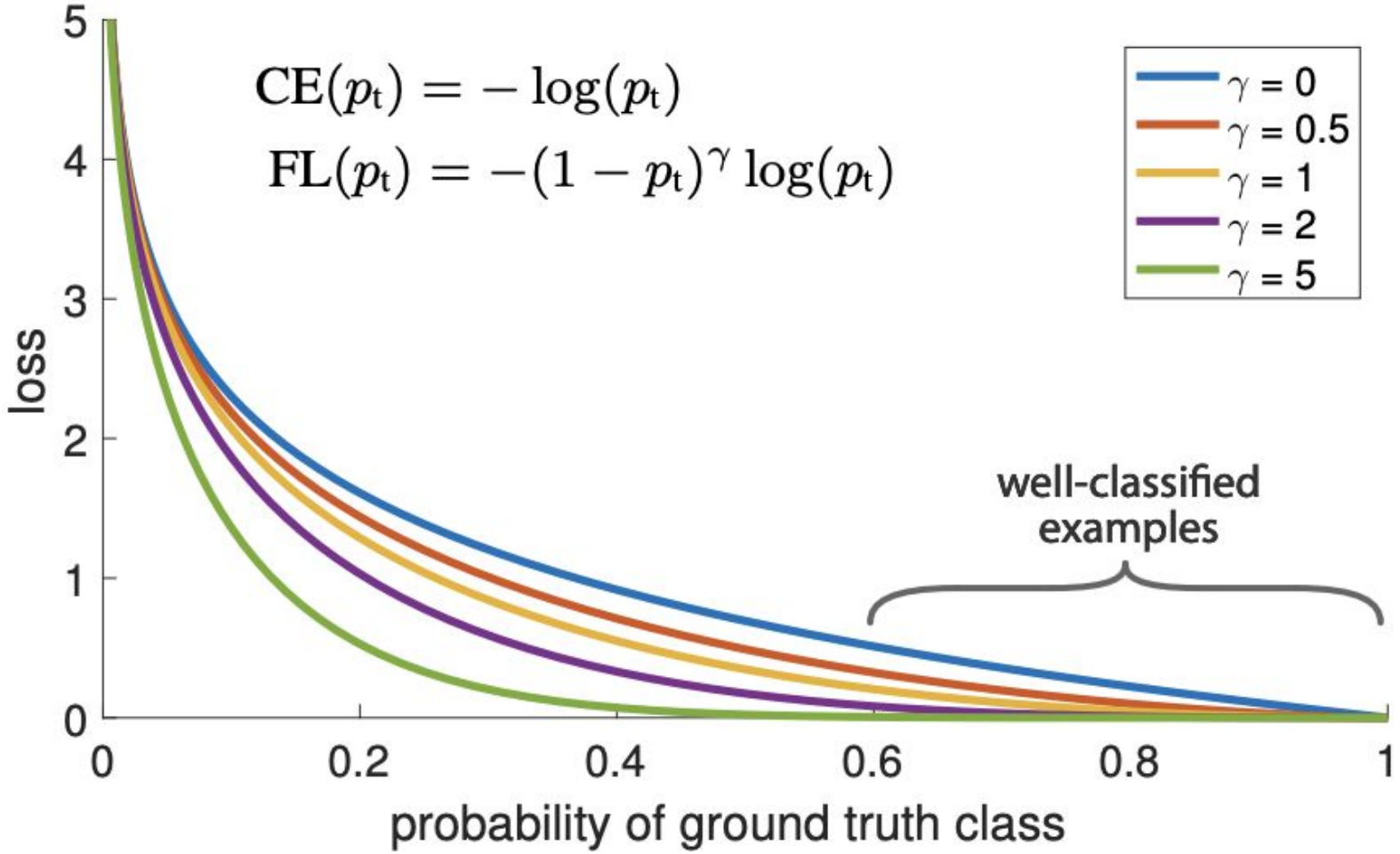
- Usually the Cross-Entropy Loss is used for image classification
- Consider we have 100 hard examples and 100000 easy examples
- **40x bigger loss** from the easy examples



- **Focal Loss** exponentially reduces the contribution of each individual easy sample

- In a supervised learning task, you usually have a good idea of the probabilities associated with the data distribution

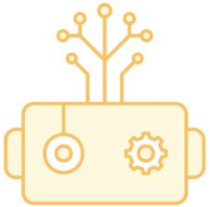



$\mathrm{CE}(p_t) = -\log(p_t)$
$\mathrm{FL}(p_t) = -(1-p_t)^{\gamma}\log(p_t)$
$\gamma = 0$
$\gamma = 0.5$
$\gamma = 1$
$\gamma = 2$
$\gamma = 5$
well-classified examples
loss
probability of ground truth class
0
1
2
3
4
5
0
0.2
0.4
0.6
0.8
1

Loss for discriminative embeddings: ArcFace Loss

11-785

Introduction to Deep Learning

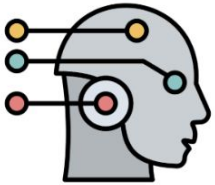


We train an AI to detect human faces

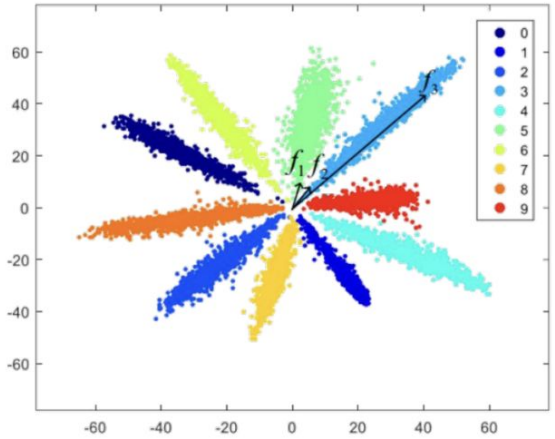
As part of HW2P2, we train a model to classify if the image has a human face or not.

How do we make it tell two faces apart?

In comes ArcFace Loss!



Model learned to represent faces



Each color is a unique face

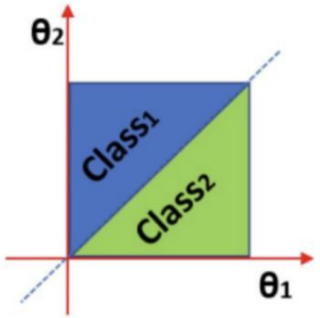
Cosine Similarity?

Won't work well

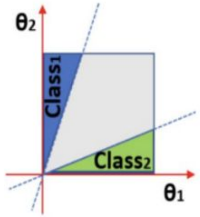
Due to magnitude F1, F2 more similar than F2, F3

Let's focus more on the angles rather than the magnitude? That's the intuition behind **ArcFace Loss**

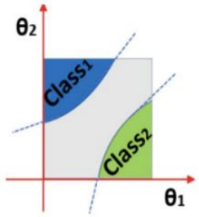
Softmax learns to separate



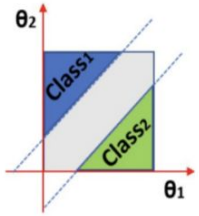
Softmax



SphereFace



CosFace



ArcFace

We need to add margin in between the separated classes!

Loss for Sequence Alignment and Decoding:
Connectionist Temporal Classification

11-785
Introduction to Deep Learning

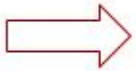
Multimodal Learning
is all the rage?

Let's solve this
video-to-text task

Input

Output

2025



We have frames that
represent a digit

Predict digit for each frame,
Merge together duplicates

We Solve It!

Multimodal Learning is all the rage?

Let's solve this video-to-text task

Predict digit for each frame

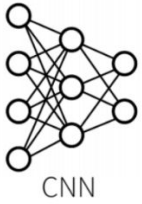
2 2 2 2 2 0 0 0 2 2 2 2 5 5 5 5

Merge duplicates

2 0 2 5

Loss for Point Clouds: Chamfer Loss

Input Image



Predicted points

Sampled GT points

$$d_{CD}(S_1, S_2) = \sum_{x \in S_1} \min_{y \in S_2} \|x - y\|_2^2 + \sum_{y \in S_2} \min_{x \in S_1} \|x - y\|_2^2$$

Can handle different number of points in (prediction, GT)

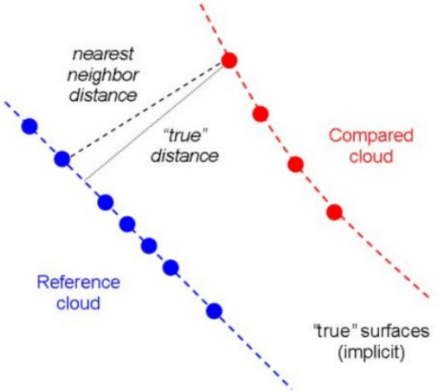
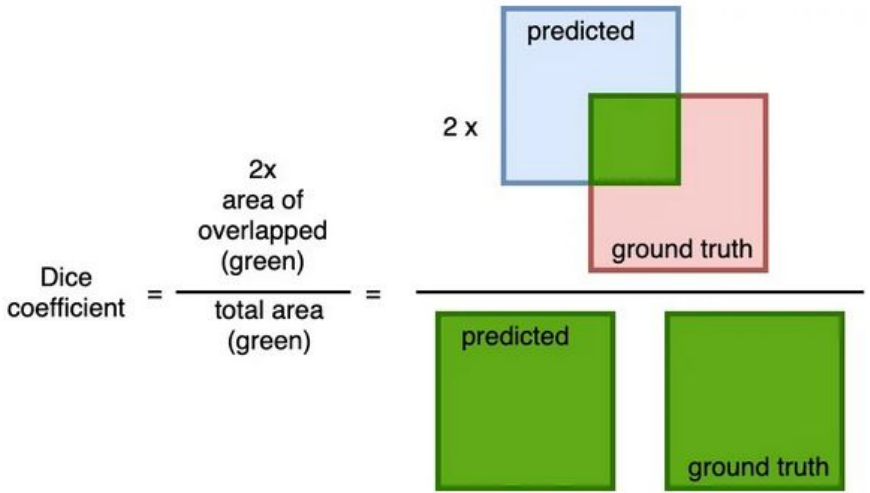


Image courtesy: cloudcompare

Link to documentation

Loss for Image Segmentation: Dice Loss / IoU

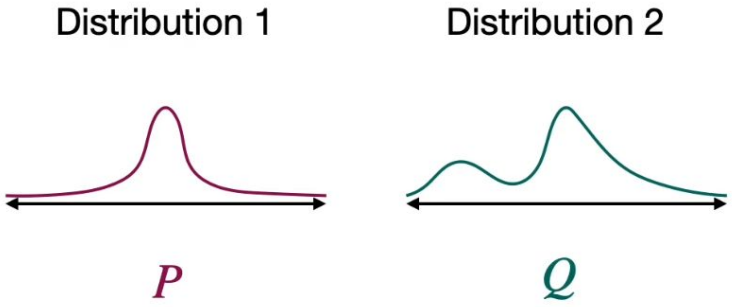
- Can consider this to be a **continuous F1 score** as it handles class imbalance
- It is differentiable, unlike using the Intersection over Union (IoU) directly



$$\mathcal{L}_{\text{Dice}} = 1 - \text{DiceCoef}$$

Link to documentation

Loss for Image Generation: Perceptual, KL-Divergence, ELBO

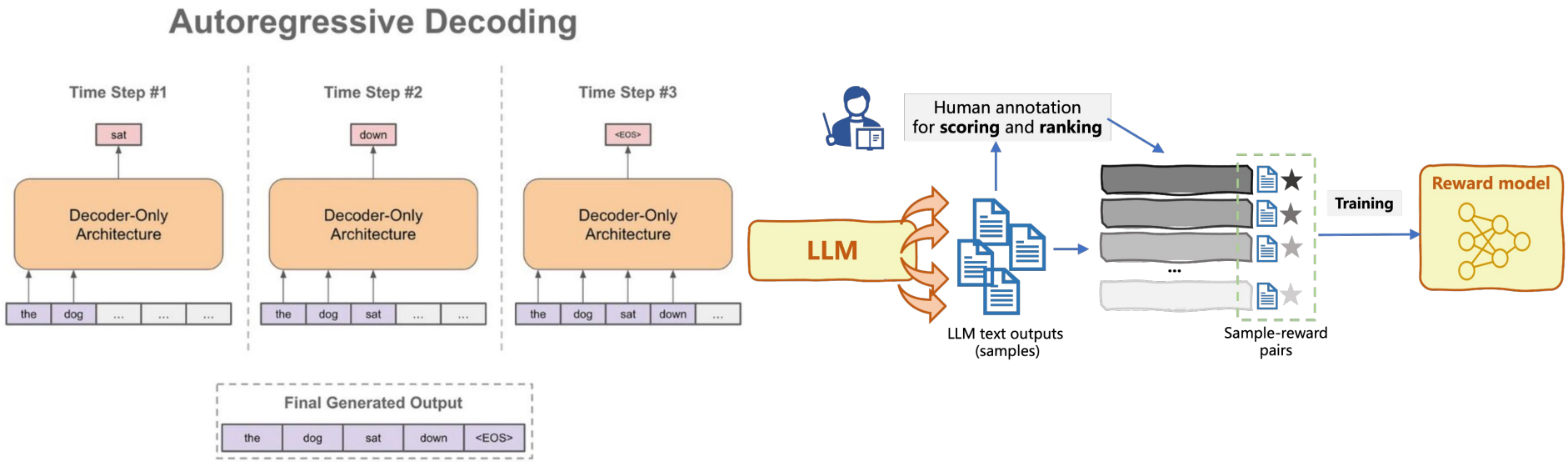


$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

$$D_{KL}(P||Q) = \int P(x) \log \frac{P(x)}{Q(x)} dx$$

Link to documentation

Loss for (one stage of) LLMs: RLHF



Unsupervised pre-training via **next token prediction**

Human preferences are used to learn structure

Link to documentation

Are we **training AI** or is **AI training us?**