Trustworthy machine learning

NEURAL INFORMATION PROCESSING SYSTEMS

Typicalness-Aware Learning for Failure Detection

Yijun Liu[1] **Jiequan Cui**[2] **Zhuotao Tian**[1✉] **Senqiao Yang**[3]
Qingdong He[4] **Xiaoling Wang**[1] **Jingyong Su**[1✉]
{liuyijun}@stu.hit.edu.cn
[1]Harbin Institute of Technology (Shenzhen) [2]Nanyang Technological University
[3]The Chinese University of Hong Kong [4]Tencent Youtu Lab

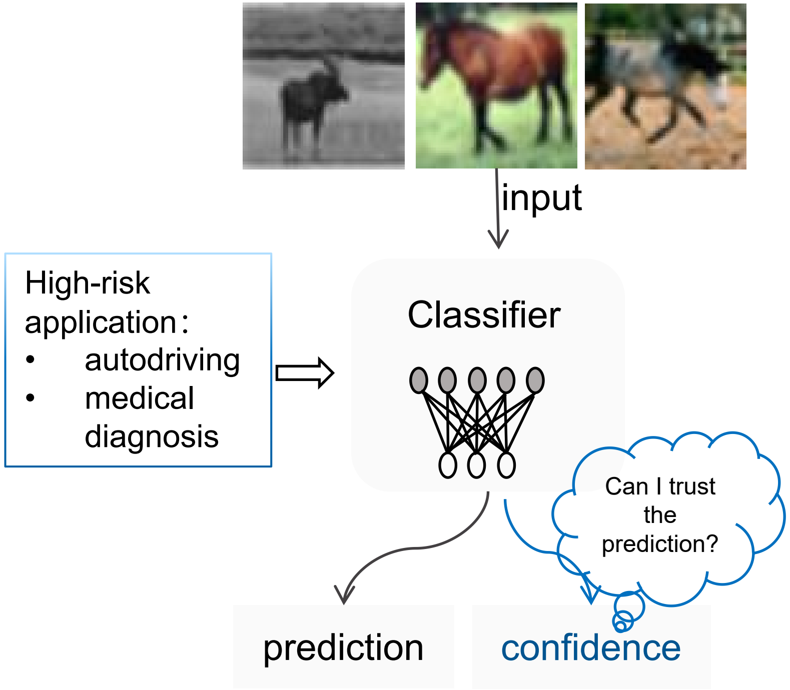
Yijun Liu

2024.11.4

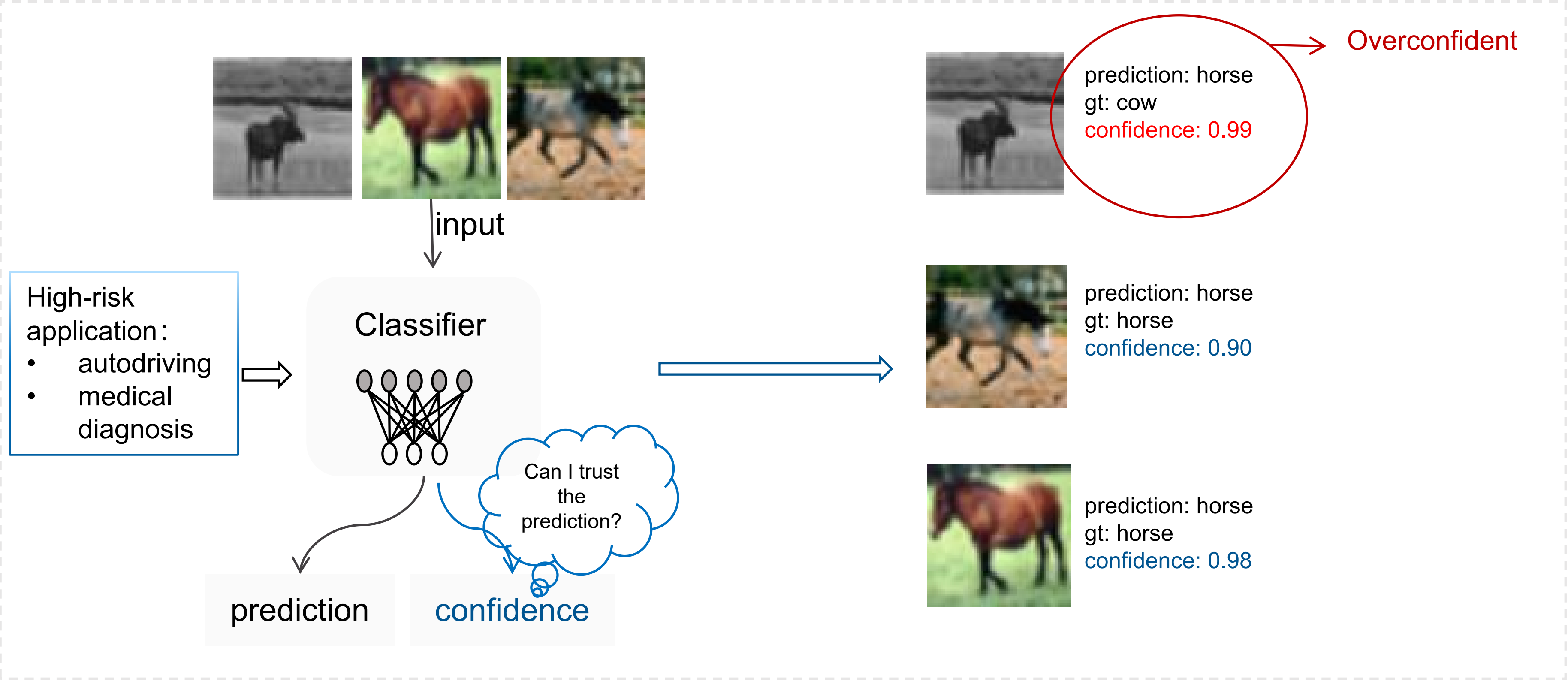
目录

content

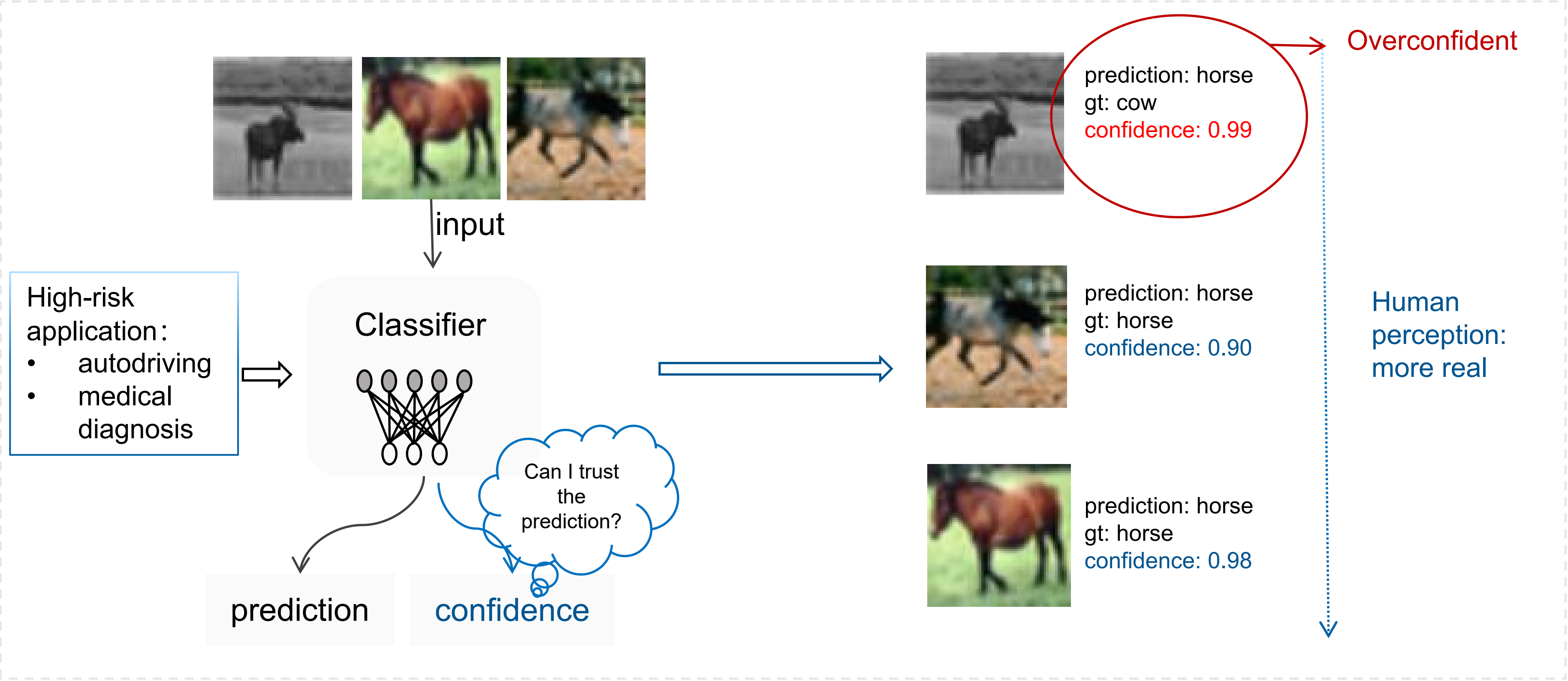
Problem Description



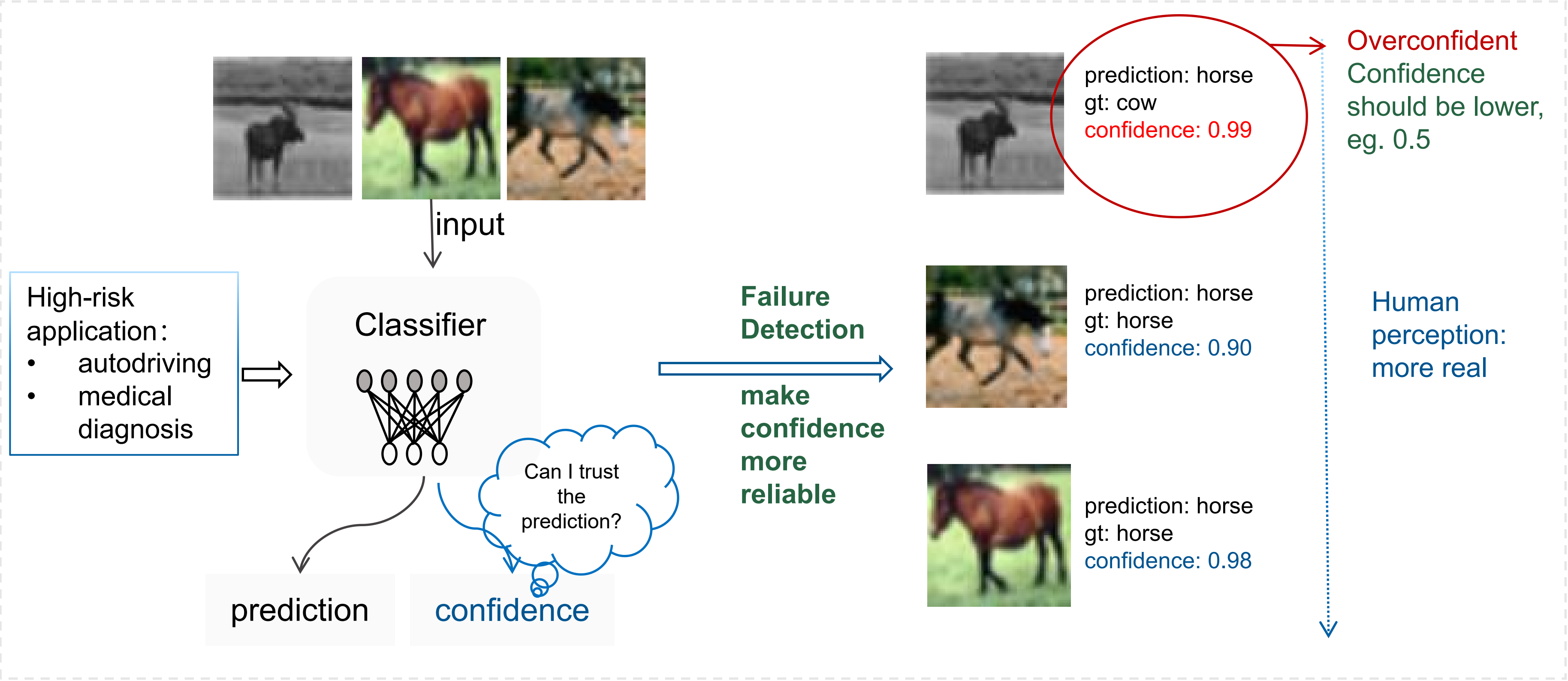
Problem Description



Problem Description



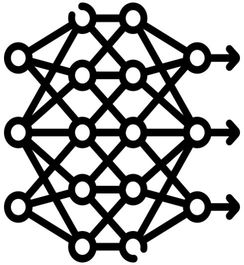
Problem Description



Key Insight & Motivation

Atypical Sample

Deep Neural Network



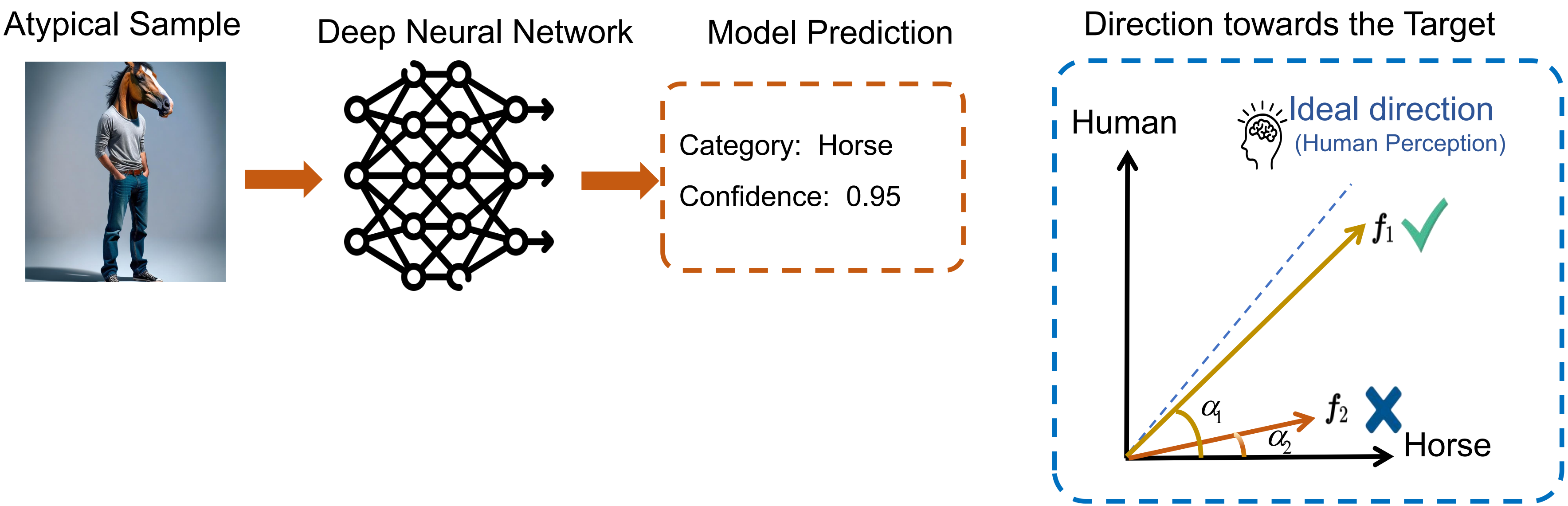
Model Prediction

Category: Horse

Confidence: 0.95

* Whether this image is labeled as Human or Horse, neither label is accurate

Key Insight & Motivation

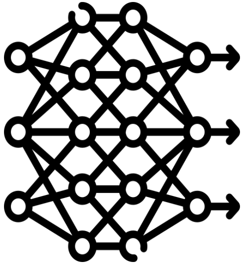


* Whether this image is labeled as Human or Horse, neither label is accurate

Key Insight & Motivation

Atypical Sample

Deep Neural Network



Model Prediction

Category: Horse

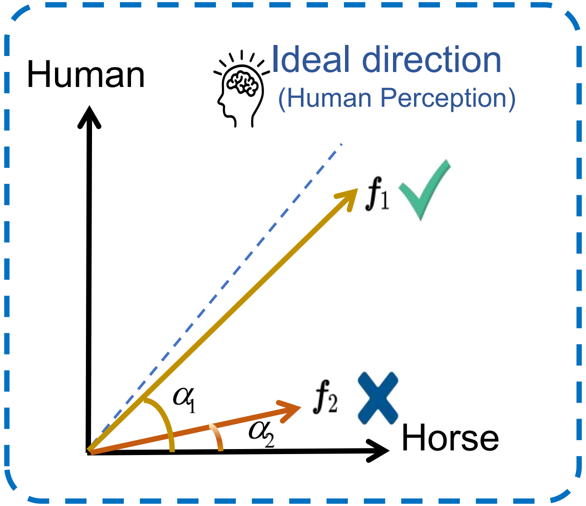
Confidence: 0.95

$$\mathcal{L}_{\mathrm{CE}}(f(\boldsymbol{x};\theta),y) = -\log \frac{e^{\|\boldsymbol{f}\|\cdot \hat{f}_y}}{\sum_{i=1}^{k} e^{\|\boldsymbol{f}\|\cdot \hat{f}_i}}$$

$$\alpha = < \hat{\boldsymbol{f}}, \hat{\boldsymbol{f}}_y >$$

① $\|\boldsymbol{f}\| \uparrow$ ② $\alpha \downarrow$

Direction towards the Target



* Whether this image is labeled as Human or Horse, neither label is accurate

Method

• Typical samples are those that exhibit similarity to a majority of other samples at the semantic level. These samples possess typical features that are easier for deep neural networks to learn and generalize.

• Atypical samples, on the other hand, differ significantly from other samples at the semantic level. They pose a challenge for the model to generalize due to their uniqueness. These samples are often located near the decision boundary.

Typical samples；ID；Fish

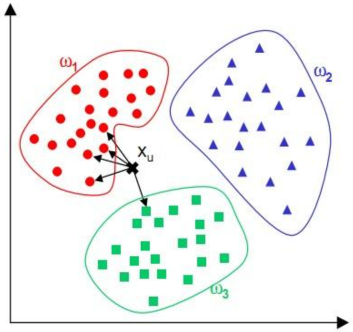
Atypical samples；ID；Covarite Shift；Fish

Atypical samples；OOD；Semantic Shift；Texture

[1] Beyond confidence: Reliable models should also consider atypicality. NeurIPS 2023
[2] Unleashing mask: Explore the intrinsic out-of-distribution detection capability. ICML2023.

Method

Measurement of typicalness



The nearest neighbor distance between sample features and the training set feature collection

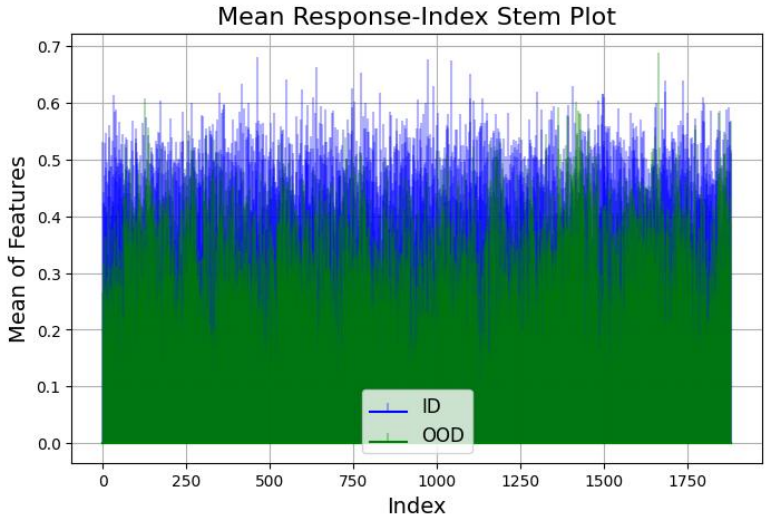
- High feature dimensionality of samples
- Large number of training samples
- Time and resource consuming

[1] Beyond confidence: Reliable models should also consider atypicality. NeurIPS 2023
[2] Unleashing mask: Explore the intrinsic out-of-distribution detection capability. ICML2023.

Method

Distinguishing typical samples from atypical samples

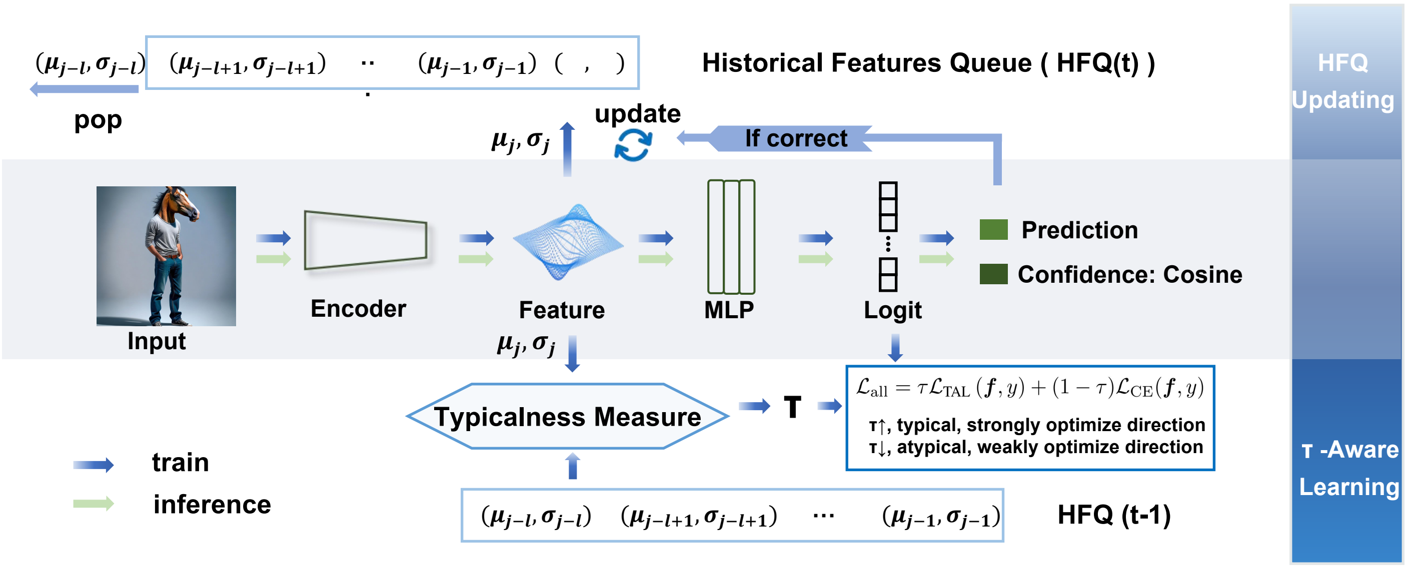
- **using ID and OOD samples as examples**



- X-axis shows the sample index
- Y-axis shows the mean responses across channels.
- ID shows higher positive responses compared to OOD

Method

Typicalness-Aware Learning



Method

Calculate Typicalness

$$Q = \{(\mu_i, \sigma_i^2) \mid \hat{y_i} = y\}$$

Add mean and varience of correct prediction to Quene

$$d = \min_{(\mu_j, \sigma_j^2) \in Q} W((\mu_{new}, \sigma_{new}^2), (\mu_j, \sigma_j^2))$$

Get minimal distance

$$\tau = 1 - \frac{d - d_{min}}{d_{max} - d_{min}}.$$

Distance normanlization

$$T(\tau) = T_{\min} + (1 - \tau) \times (T_{\max} - T_{\min})$$

Calculate dynaminc magnitude

Method

How to design the loss function?

- **fully optimize in the direction of typical samples, while not approaching infinity for atypical samples**

$$\mathcal{L}_{\mathrm{TAL}}(\boldsymbol{f}, y) = -\log \frac{e^{\hat{\boldsymbol{f}}_y * \boxed{T}(\tau)}}{\sum_{i=1}^{k} e^{\hat{\boldsymbol{f}}_i * \boxed{T}(\tau)}}.$$

Dynamic

Typicalness	Prediction	Magnitude *T*	Loss	Explanation
Atypical	Correct	↑	↓	After correct prediction, add small force to approach label direction
Typical	Correct	↓	↑	After correct prediction, add large force to approach prediction direction
——	Incorrect	↑	↑	No action for wrongly predicted samples due to avoid impact on feature extraction
——	Incorrect	↓	↓	

Experimental Results

Table 1: Evaluation results of the proposed TAL on CIFAR100.

Architecture	Method	Old setting FD			OOD Detection			New setting FD			ID-ACC
		AURC↓	FPR95↓	AUROC↑	AURC↓	FPR95↓	AUROC↑	AURC↓	FPR95↓	AUROC↑	
CIFAR100 vs. SVHN											
ResNet110 [13]	MSP[14]	99.83	67.49	84.07	293.44	83.41	74.55	376.42	66.92	84.00	72.01
	Cosine [47] [47]	96.53	65.15	84.42	271.13	78.30	79.31	361.87	56.23	86.93	72.01
	Energy [23]	135.85	74.66	77.20	275.39	83.18	77.78	387.44	66.96	83.21	72.01
	MaxLogit[14]	133.19	72.33	77.96	275.85	82.53	77.73	385.81	65.08	83.56	72.01
	Entropy [33]	100.05	66.28	84.12	287.62	81.20	75.93	373.49	61.33	84.73	72.01
	Mahalanobis [4]	114.21	73.48	80.41	263.49	72.70	80.55	368.55	58.74	85.74	72.01
	Gradnorm [16]	369.86	98.82	35.30	490.21	98.17	49.26	679.48	98.69	42.76	72.01
	SIRC [40]	100.56	66.37	84.01	287.93	81.03	75.90	374.12	61.29	84.65	72.01
	LogitNorm [39]	125.59	72.87	79.71	235.50	73.23	83.35	356.88	55.80	87.80	70.34
	OpenMix [49]	85.66	63.82	85.25	342.16	87.03	69.27	406.80	70.37	80.25	73.68
	TAL	90.60	64.84	85.36	259.64	76.37	80.28	347.72	54.39	87.89	72.45
	FMFP [48]	69.83	62.17	87.15	284.13	81.77	74.98	345.37	62.99	84.86	75.18
	TAL w/ FMFP	73.16	64.82	85.51	245.62	78.61	81.59	320.73	55.22	88.48	75.59

Experimental Results

Table 2: Evaluation results of the proposed TAL on ImageNet.

Architecture	Method	Old setting FD			OOD Detection			New setting FD			ID-ACC
		AURC↓	FPR95↓	AUROC↑	AURC↓	FPR95↓	AUROC↑	AURC↓	FPR95↓	AUROC↑	
		Imagenet vs. Textures									
ResNet50	MSP [14]	72.73	63.95	86.18	301.27	46.01	87.21	351.26	49.64	86.99	76.13
	Cosine [47]	102.98	69.93	79.49	298.35	50.64	87.54	359.74	54.43	86.17	76.13
	Energy [23]	118.66	76.33	75.81	279.16	35.64	90.47	351.93	43.69	87.74	76.13
	MaxLogit [14]	113.35	72.11	77.29	278.52	34.1	90.57	349.3	41.59	88.1	76.13
	Entropy [33]	74.61	67.07	85.48	292.54	38.3	88.92	344.73	43.95	88.27	76.13
	Mahalanobis [4]	208.22	96.19	54.23	288.17	57.61	86.51	397.18	65.34	80.22	76.13
	Residual [40]	238.18	97.01	49.0	316.1	57.77	83.89	431.12	65.55	77.12	76.13
	Gradnorm [16]	206.99	89.66	57.88	272.83	30.21	91.55	385.97	42.45	84.89	76.13
	SIRC [40]	72.91	63.67	86.11	295.13	38.88	88.53	346.42	43.82	88.03	76.13
	TAL	64.66	64.93	87.11	290.5	47.66	87.51	338.45	50.11	88.29	76.43
	TAL+SIRC	64.55	63.66	87.15	288.23	46.91	87.88	336.56	49.68	88.35	76.43

THANKS