ALSA: Anchors in Logit Space for Out-of-Distribution Accuracy Estimation

Chenzhi Liu, Mahsa Baktashmotlagh, Yanran Tang, Zi Huang, Ruihong Qiu
The University of Queensland

BMVC 2025

Introduction

The reliability of machine learning models heavily relies on the assumption that the test data is in-distribution. However, out-of-distribution (OOD) data with distribution shifts can significantly degrade model performance. To ensure the reliability of deployed models, we propose ***ALSA***, a method designed to evaluate model performance on unlabeled OOD data.

Highlights:

- Theoretical analysis of logits distribution.
- Our method, ***ALSA***, proposes a general framework for evaluating vision, language, and graph models.

Method

What's an anchor? Logits are distributed along a band near a hyperplane, and an *anchor* is a point used to estimate the probability of correct predictions in a small surrounding area within this band. A set of *anchors* is identified to model prediction performance within the band in the logits space.

How it Works? (1) An anchor is a triplet $(\vec{a}, p, v)$. (2) A logit $\vec{z}$ is influenced by the anchor via a Gaussian-like formula

$$\mathrm{Infl}(\vec{z}, (\vec{a}, p, v)) = p\ exp(-v^2 dist^2(\vec{z}, \vec{a}))$$

where $dist(\vec{z}, \vec{a})$ is the cosine distance. (3) Use sigmoid function to predict probability from total influence. (4) Rectification corrects unlearned areas.

An alternative exponential-like influence function is also evaluated as follows:

$$\mathrm{Infl}(\vec{z}, (\vec{a}, p, v)) = p\ exp(-v^2 dist\ (\vec{z}, \vec{a}))$$

Method Assumption.

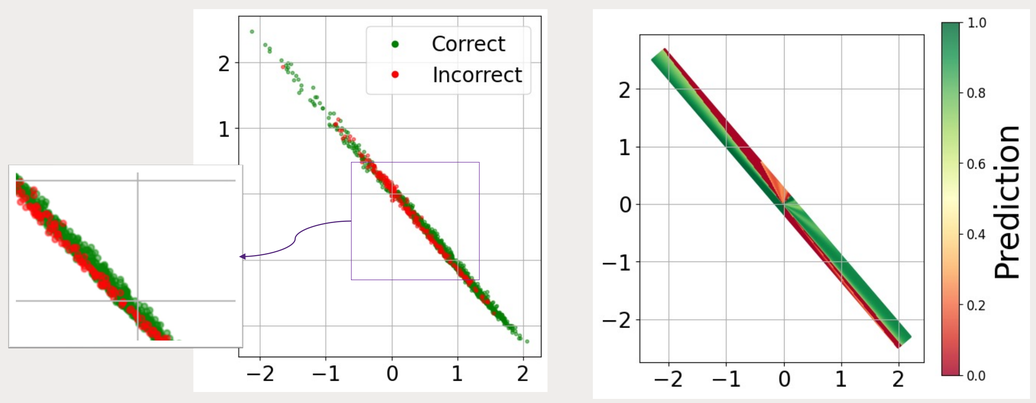
- Under the assumption of the invariance of the conditional distribution $p(y|\vec{z})$ which is a necessary condition of covariate shift assumption.
- Without making further assumptions on the model or distribution shift, predicting model performance on an unlabeled dataset has been shown to be impossible [1].

Motivation

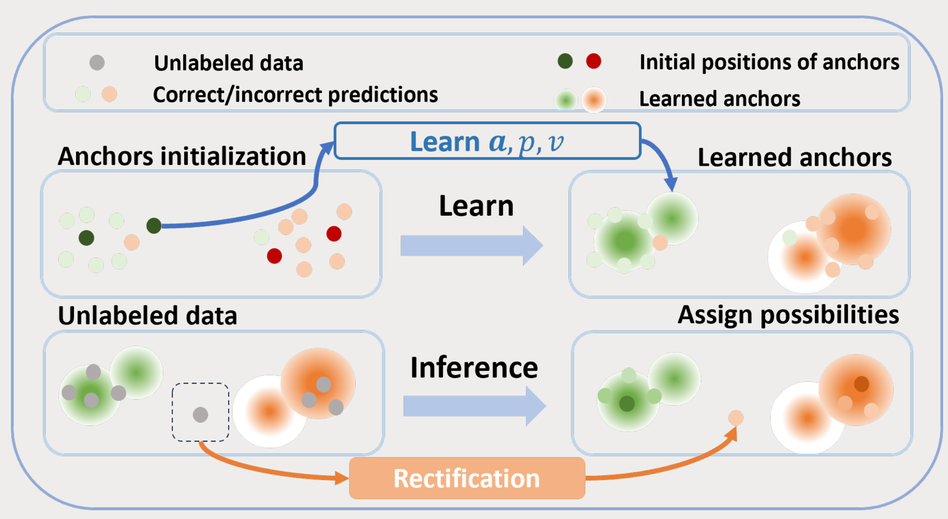
Current Problems. Prior methods focus on predicted class probabilities (softmax scores).Softmax function compresses logits, losing information useful for model performance evaluation.

Logit positions offer insights into prediction performance:

- Logits are distributed in a band near a hyperplane.
- Incorrect logits distribution varies by class, reflecting class-wise performance differences.
- Even within the narrow band, probabilities of correct predictions vary.



(**Left**) Distribution of logits and their corresponding prediction correctness. (**Right**) Probability of correct predictions at various positions learned by the ***Anchors*** model.



Workflow of ***ALSA***.

Experimental Results

A series of experiments were conducted to evaluate vision, language, and graph models, **assessing the robustness and generalization capabilities of *ALSA***. The results demonstrate the following:

Vision and Language Models (MAE comparison): Our model **outperforms the SOTA methods** in most cases.

Dataset	Variant	AC	DoC	IM	ATC	COT	ALSA-E	ALSA-G
CIFAR10	CIFAR-10.1	4.82	4.69	5.70	1.27	3.02	2.83	**1.22**
	CIFAR-10.2	7.97	7.84	8.74	4.59	6.20	**0.61**	*4.19*
	CIFAR-10C	6.57	6.44	7.60	**2.28**	3.18	5.09	4.18
CIFAR100	CIFAR-100C	9.46	7.80	9.02	3.85	**2.41**	7.26	6.56
A-WILD	Subpop1	3.70	2.16	2.14	5.90	3.39	*0.99*	**0.97**
	Subpop2	4.19	2.65	2.65	5.82	4.18	*1.43*	**1.17**
W-WILDS	Subpop	0.78	0.79	0.74	1.55	0.97	**0.32**	*0.54*
ImageNet	ImageNetv2	3.63	2.68	2.85	**0.75**	9.57	4.07	3.22
MNIST	MNIST-M	13.27	13.44	16.29	13.63	4.96	**2.82**	*3.09*
Office-31	domain	38.67	34.47	32.90	41.55	11.40	*8.35*	**7.77**

Graph Models: We evaluated four different models on the ogbn-arxiv dataset, and our method consistently **produced the lowest prediction error** across all models.

Model	AC	DoC	IM	GNNEval	ATC	COT	ALSA-E	ALSA-G
GCN	1.61	2.94	3.89	10.59	19.76	6.64	*1.13*	**0.99**
SAGE	1.31	1.72	2.07	10.42	25.31	5.85	2.21	**1.08**
GAT	3.24	4.16	4.60	13.34	23.89	5.06	*1.73*	**1.01**
GIN	30.09	5.86	7.17	9.54	32.28	30.79	*2.18*	**2.14**
AVG.	9.06	3.67	4.43	10.97	25.31	12.09	*1.81*	**1.30**

Statistical Validation: We analysed the **correlation between predicted and true accuracy** to validate results, with high correlation confirming our method's robustness.

Model	AC		DoC		IM		GNNEval		ATC		COT		ALSA	
	R^2	ρ	R^2	ρ	R^2	ρ	R^2	ρ	R^2	ρ	R^2	ρ	R^2	ρ
GCN	0.84	0.96	0.52	0.96	0.25	0.96	-29.28	-0.23	-17.42	0.95	-1.12	0.96	**0.94**	**0.97**
SAGE	0.85	0.97	0.74	0.98	0.61	0.98	-29.10	-0.29	-48.32	**0.98**	-2.13	0.96	**0.87**	**0.98**
GAT	-0.13	0.85	-0.50	0.84	-0.73	0.85	-44.46	-0.25	-29.58	0.55	-0.74	0.75	**0.91**	**0.99**
GIN	-162.85	0.85	-6.40	**0.85**	-8.69	0.76	-19.58	-0.16	-189.22	0.82	-170.22	0.83	**-0.32**	0.46
AVG.	-40.32	0.91	-1.41	0.91	-2.14	0.89	-30.60	-0.23	-71.13	0.83	-43.55	0.88	**0.60**	**0.85**

References

[1] Garg et al. (2022). Leveraging unlabeled data to predict out-of-distribution performance. ICLR.

Acknowledgement

This research has been partially supported by Australian Research Council Discovery Projects (DP230101196, DE250100919, CE200100025, and FT230100426).