VI-OOD: A Unified Representation Learning Framework for Textual Out-of-distribution Detection

Li-Ming Zhan, Bo LIU, Xiao-Ming Wu

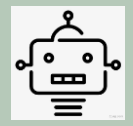
Department of Computing

The Hong Kong Polytechnic University

Problem Definition

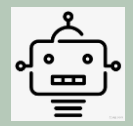
My password was changed.

Don't worry, I can help you find it!



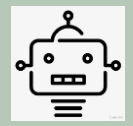
I haven't got my credit card.

Sorry for that. Your card will be arriving soon.



I recently found my card!
Could you help me re-establish it?

Sorry, I don't understand your question.



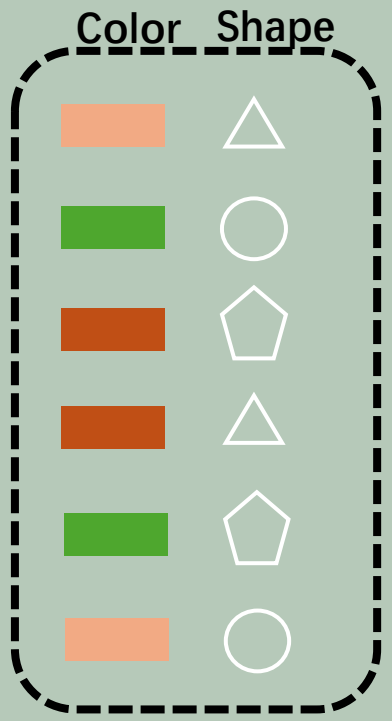
known intent classes

Unknown intent classes

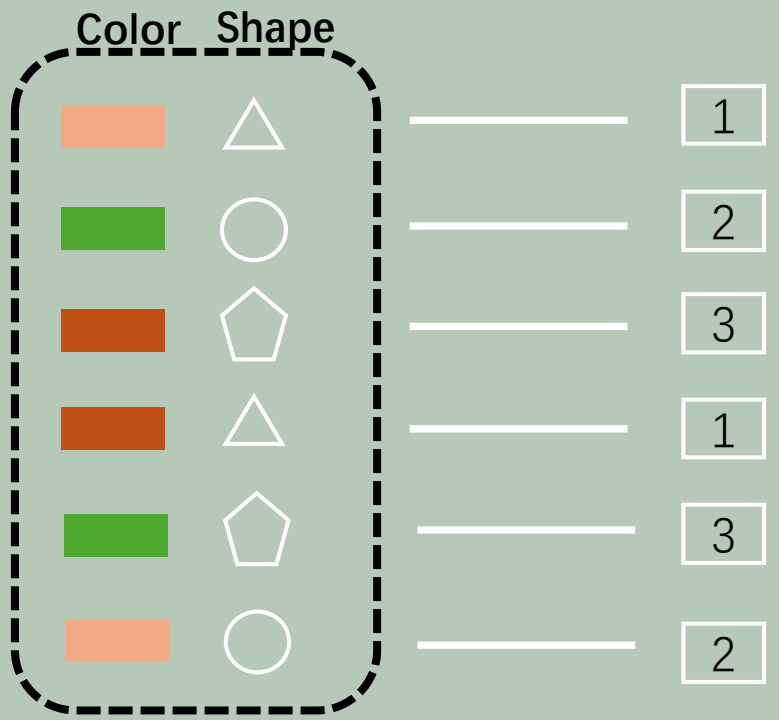
The goal of textual OOD detection is to raise an alarm or become aware of unknown inputs while keeping the known in-distributon (ID) classes correctly classified.

Motivation

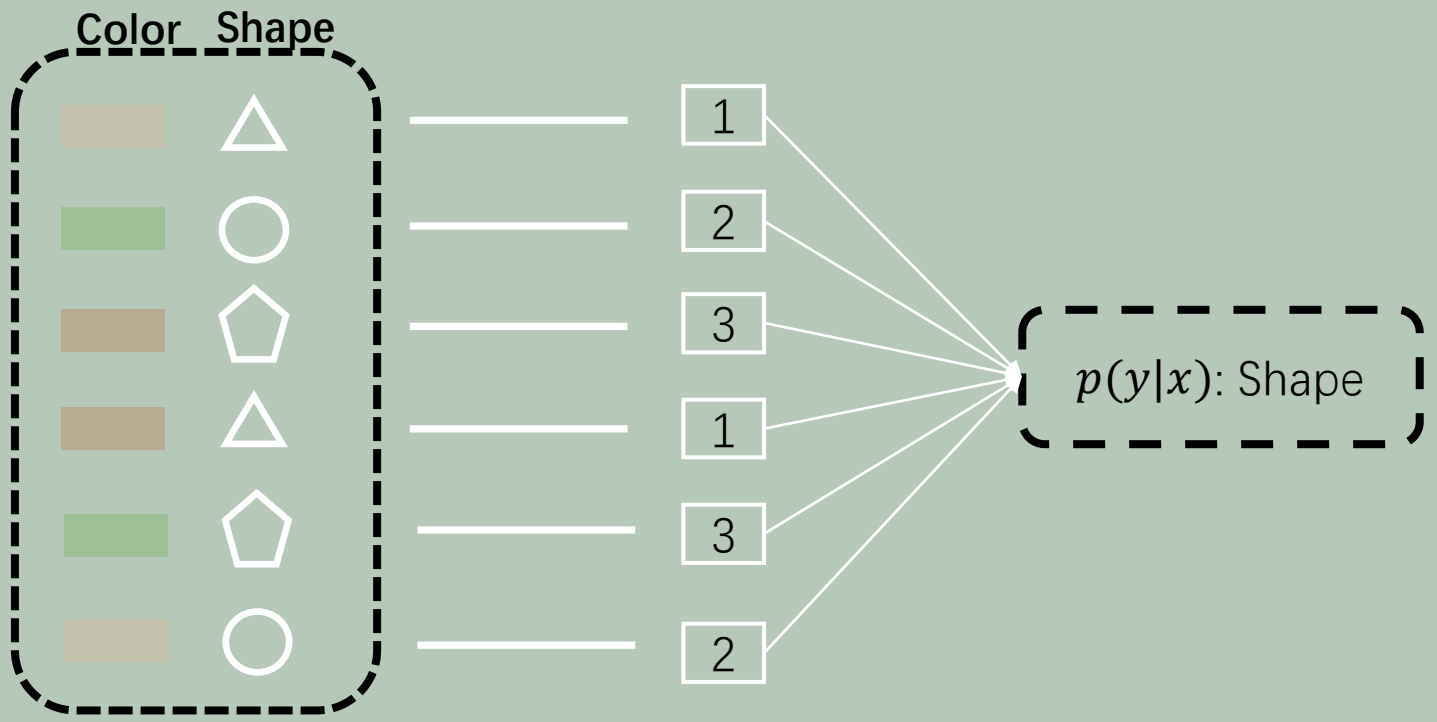
The training objective， $p(y|x)$**, is biased and does not take into account the ability to handle OOD inputs.**



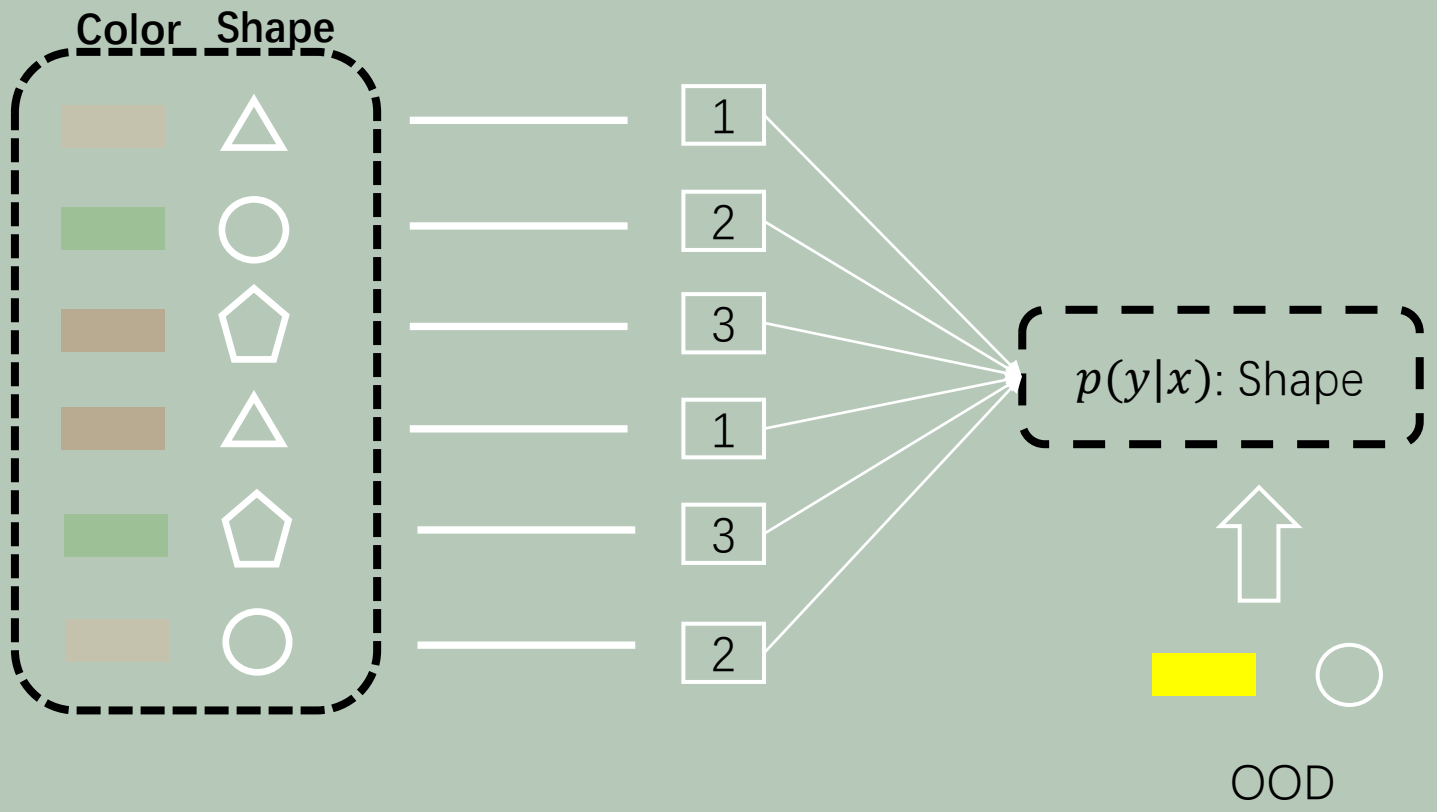
The training objective，p(y|x), is biased and does not take into account the ability to handle OOD inputs.



The training objective，p(y|x), is biased and does not take into account the ability to handle OOD inputs.



The training objective，p(y|x), is biased and does not take into account the ability to handle OOD inputs.

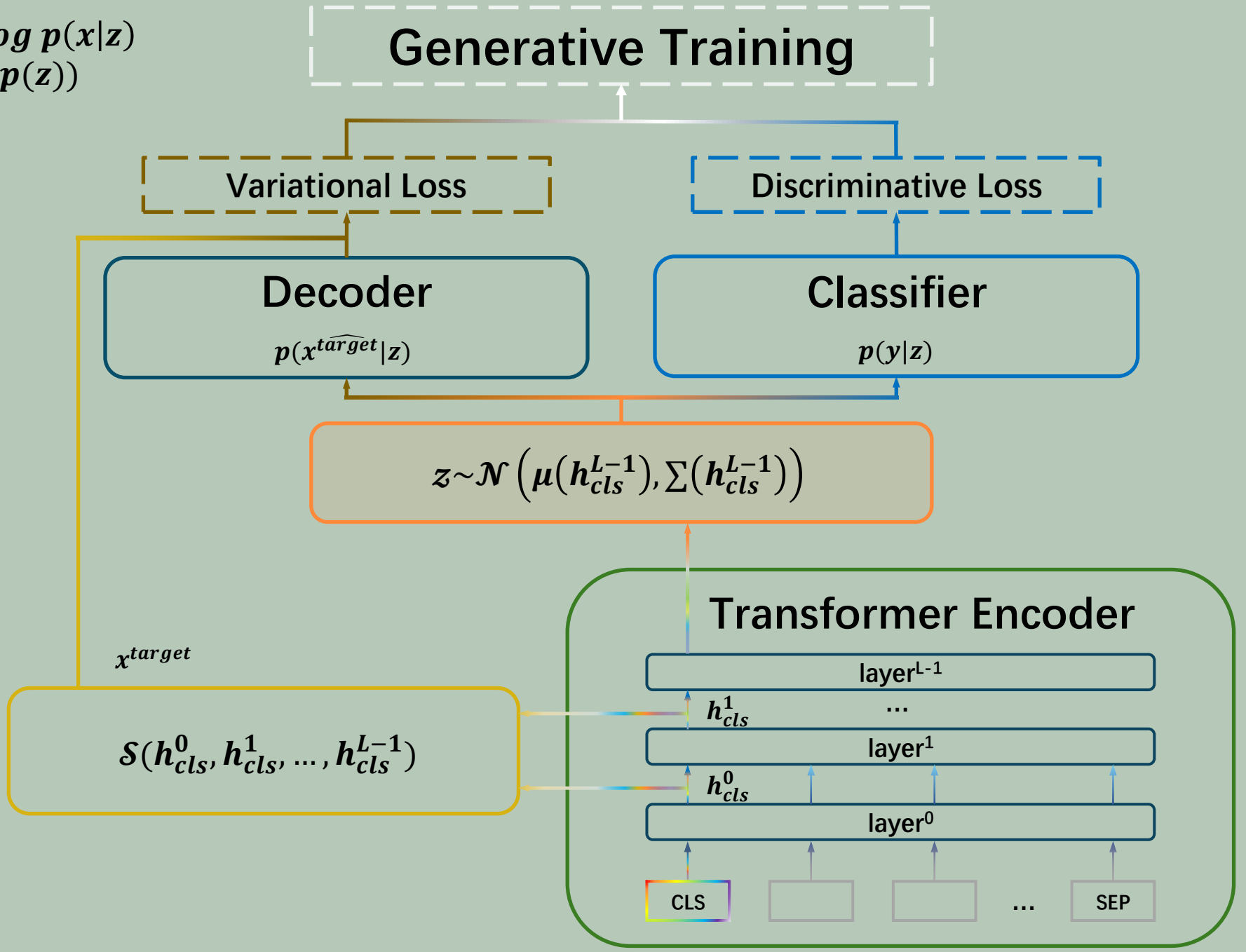


Methodology

$$p(y|x) \rightarrow p(y,x) = p(y|x)p(x)$$

ELBO: $$p(y,x) > log\ p(y|z) + log\ p(x|z) - D_{KL}(q(z|x)|p(z))$$

$$log\ p(y|z) + log\ p(x|z) - D_{KL}(q(z|x)|p(z))$$



Results

Evaluation

Metrics:

- Threshold-related：Accuracy, F1-score
- Threshold-free: AUROC, FPR@95, AUPR

Benchmarks: diverse NLP datasets from natural language understanding tasks

- Sentiment analysis: IMDB, SST-2.
- Topic classification: 20 Newsgroups, Trect-10
- Machine translation: English-German WMT16\ Multi30K
- Natural language inference: RTE, MNLI

Examples:

Taking IMDB as ID, then 20 Newsgroups, Trect-10… MNLI will be OOD

Encoder-based Transformer

Methods	SST-2 AUROC ↑	SST-2 FAR@95 ↓	SST-2 AUPR ↑	IMDB AUROC ↑	IMDB FAR@95 ↓	IMDB AUPR ↑
MSP	89.85	66.20	86.40	94.30	41.90	98.80
$\text{MSP}_{\text{Contrast}}$	85.04	63.42	69.34	94.51	44.69	98.89
MSP_{VI}	**92.85**	**51.58**	**89.72**	**95.95**	**28.03**	**99.12**
Maha	97.98	11.50	97.30	99.67	0.70	99.95
$\text{Maha}_{\text{Contrast}}$	**99.42**	**2.98**	**98.73**	99.89	**0.05**	**99.97**
Maha_{VI}	99.33	3.62	98.52	**99.90**	0.21	**99.97**
Cosine	95.65	22.65	94.68	99.50	1.53	99.88
$\text{Cosine}_{\text{Contrast}}$	98.38	8.64	96.36	**99.87**	1.93	**99.96**
$\text{Cosine}_{\text{VI}}$	**98.87**	**6.62**	**98.06**	99.57	**1.43**	99.88
Energy	89.80	67.00	86.53	93.30	56.70	98.63
$\text{Energy}_{\text{Contrast}}$	84.93	63.16	69.29	94.44	44.46	98.86
$\text{Energy}_{\text{VI}}$	**92.79**	**51.25**	**89.26**	**96.05**	**27.97**	**99.12**

Methods	TREC-10 AUROC ↑	TREC-10 FAR@95 ↓	TREC-10 AUPR ↑	20NG AUROC ↑	20NG FAR@95 ↓	20NG AUPR ↑
MSP	97.94	8.43	89.26	**93.89**	30.49	**87.39**
$\text{MSP}_{\text{Contrast}}$	98.43	4.06	**91.19**	93.19	28.00	83.17
MSP_{VI}	**98.91**	**2.77**	90.39	93.29	**25.61**	80.09
Maha	98.99	4.87	95.11	98.39	7.77	95.91
$\text{Maha}_{\text{Contrast}}$	**99.57**	0.97	**98.59**	98.78	5.89	97.29
Maha_{VI}	99.46	**0.79**	97.67	**99.80**	**0.61**	**98.93**
Cosine	98.89	3.96	94.54	97.73	10.84	88.71
$\text{Cosine}_{\text{Contrast}}$	99.14	1.42	93.34	98.03	8.86	95.27
$\text{Cosine}_{\text{VI}}$	**99.36**	**1.19**	**96.09**	**99.39**	**2.92**	**97.19**
Energy	97.19	10.07	82.16	95.76	17.93	**88.71**
$\text{Energy}_{\text{Contrast}}$	98.45	4.73	**91.18**	**96.04**	**15.70**	88.62
$\text{Energy}_{\text{VI}}$	**99.21**	**2.84**	90.84	94.34	17.04	79.67

Average	AUROC ↑	FAR@95 ↓	AUPR ↑
avg. (MSP / Maha / Cosine / Energy)	94.00 / 98.78 / 97.94 / 94.01	36.76 / 6.21 / 9.75 / 37.93	**90.46** / 97.07 / 94.45 / 89.01
$\text{avg.}_{\text{Contrast}}$ (MSP / Maha / Cosine / Energy)	92.79 / 99.17 / 98.86 / 93.47	35.04 / 3.93 / 5.21 / 32.01	85.65 / 97.43 / 96.23 / 86.99
avg._{VI} (MSP / Maha / Cosine / Energy)	**95.25** / **99.62** / **99.30** / **95.60**	**27.00** / **1.31** / **3.04** / **24.78**	89.83 / **98.77** / **97.81** / **89.72**

Table 1: Main results of our proposed framework. MSP, Maha, Energy, and Cosine are baseline methods trained with the discriminative loss, while each corresponding method with the *VI* subscript denotes the model trained with our VI framework. The *Contrast* subscript denotes the method proposed by Zhou et al. (2021b). The best result is marked in bold. At the bottom row, averaged results across four ID datasets are included. All the reported results are presented in percentage values.

decoder-based Transformer

Methods	SST-2			TREC-10		
	AUROC ↑	FAR@95 ↓	AUPR ↑	AUROC ↑	FAR@95 ↓	AUPR ↑
MSP	78.22	70.40	74.34	**99.29**	**0.64**	**98.79**
MSP_{VI}	**83.05**	**65.32**	**69.54**	98.69	5.69	94.48
Maha	46.08	84.41	47.34	84.41	71.94	64.13
Maha_{VI}	**95.41**	**12.47**	**85.38**	**99.96**	**0.12**	**99.39**
Cosine	89.24	26.52	79.47	98.24	11.06	93.05
$\text{Cosine}_{\text{VI}}$	**95.95**	**7.21**	**85.70**	**100**	**0**	**99.92**
Energy	70.13	67.78	64.87	**99.86**	**0.14**	**99.27**
$\text{Energy}_{\text{VI}}$	**80.16**	**66.90**	**65.76**	98.64	7.40	92.44

Table 2: Results of our proposed framework on LLaMA-2-7B, fine-tuned with a classification head. The best result is marked in bold. All the reported results are presented in percentage values.

Thank you very much!