[23-2] UST Seminar

Detecting OOD with Fine-tuned CLIP's Class-Specific Threshold Adjustments

UST-ETRI School Hwang Jihyun
(aribae@etri.re.kr)

Table of contents

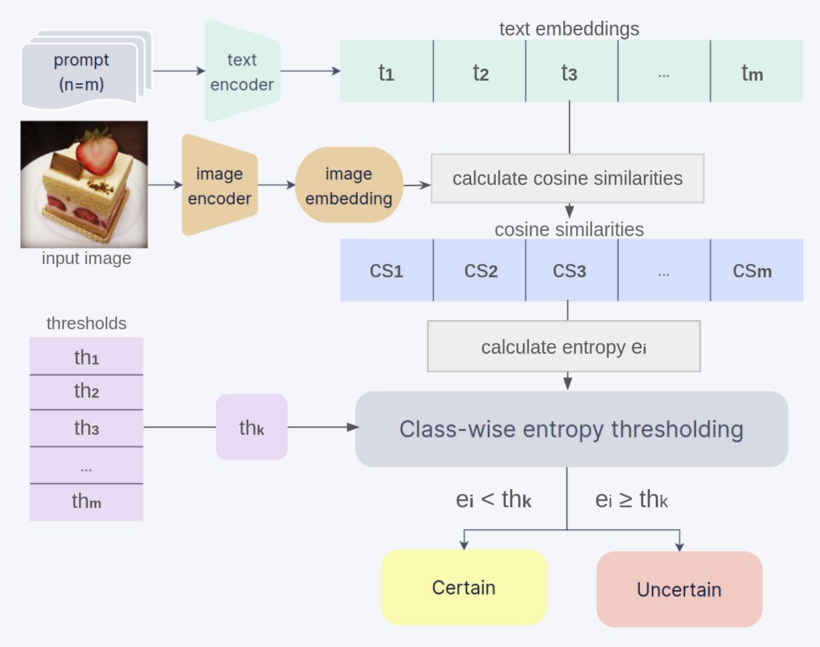
1. Recap of previous study
2. Introduction
3. Experiment & Result
4. What’s next

Recap of previous study

Introduction

Estimating uncertainty in zero-shot image classification using the vision-language model, **CLIP**.

Proposed Architecture:



Results

[Misclassification detection results]

- Class-wise entropy thresholding method shows higher performance as a result of synthesizing the entire set of results.

		Dataset	
		CIFAR10	Food101
1	Class-wise Thresholds	0.779	0.845
2	Mean of class-wise thresholds single threshold	0.456	0.367
3	Grid search single threshold	0.529	0.576

Uncertainty detection performance

Results

[Misclassification detection results]

- Out-of-Distribution (OOD): The model represents an untrained data area and is used to assess predictive uncertainty for a given model.
- The results of OOD also confirmed that class-wise entropy thresholds showed the highest performance

		Dataset	
		CIFAR10	Food101
1	Class-wise Thresholds	0.933	0.894
2	Mean of class-wise thresholds single threshold	0.689	0.475
3	Grid search single threshold	0.779	0.751

Uncertainty detection performance In OOD dataset

Results

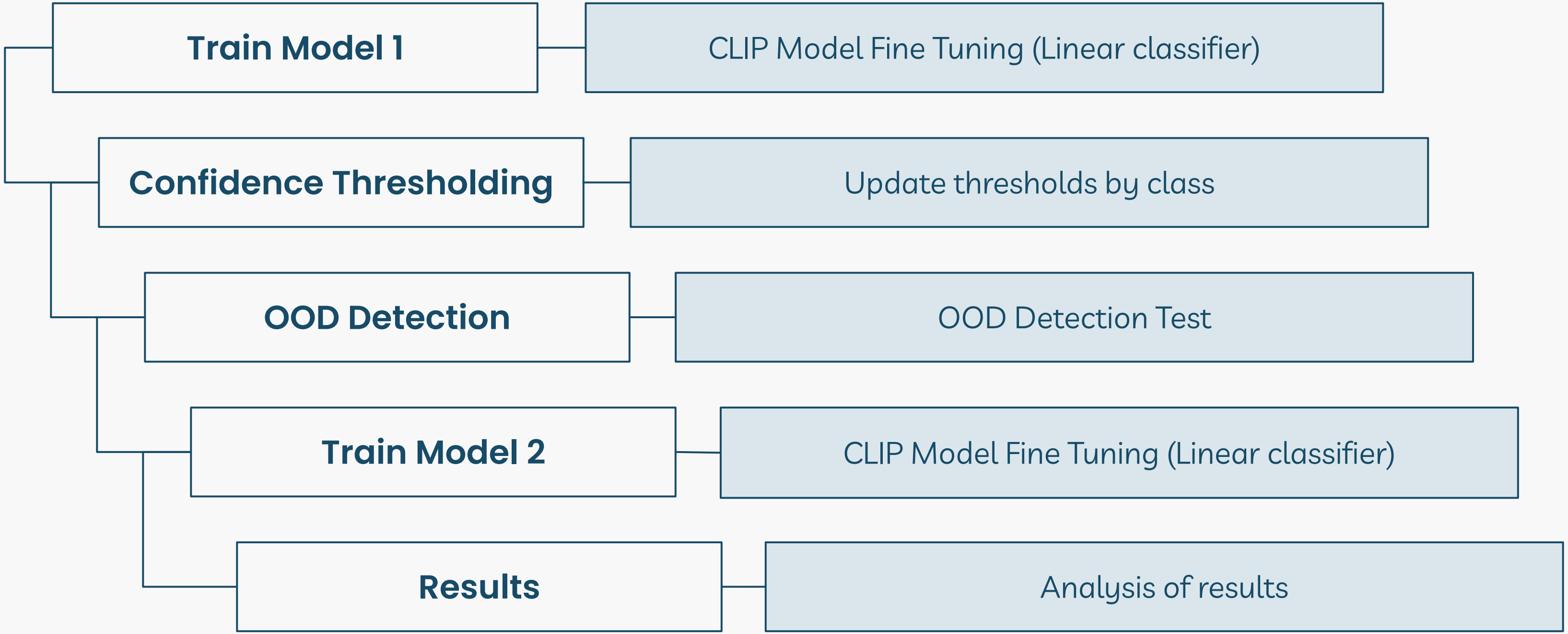
image class: cat
prediction: cat
entropy: 0.1193235
threshold: 0.8461579

image class: bird
prediction: ship
entropy: 0.8560749
threshold: 0.2974607

Uncertainty detection image sample

Introduction

Introduction



Train Model 1 — CLIP Model Fine Tuning (Linear classifier)

Confidence Thresholding — Update thresholds by class

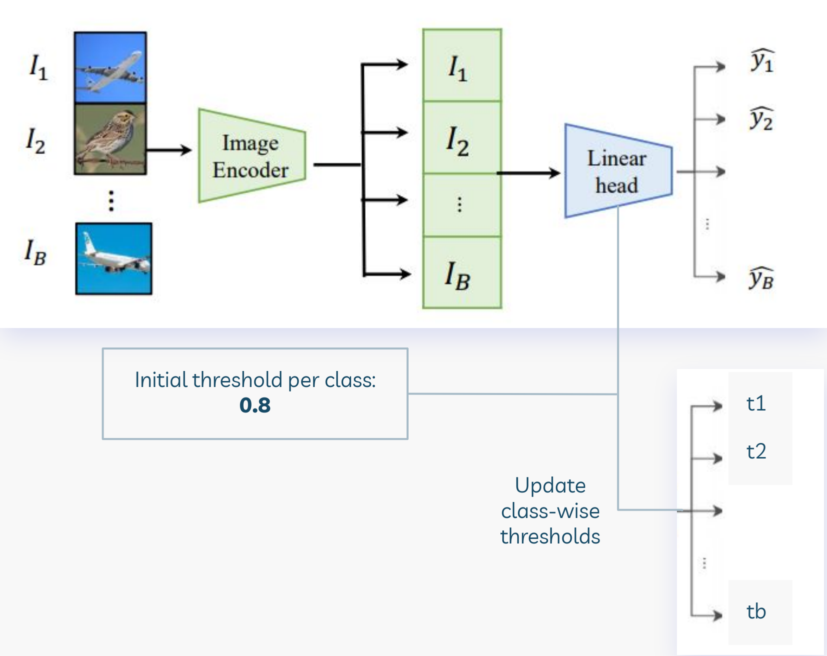
OOD Detection — OOD Detection Test

Train Model 2 — CLIP Model Fine Tuning (Linear classifier)

Results — Analysis of results

Overview

[Fine-tuning Open AI's CLIP]



Initial threshold per class:
0.8

Update class-wise thresholds

t1

t2

tb

Freeze CLIP's Image Encoder:

- Freeze clip image encoder trained with large amounts of data as encoder

Train Linear head:

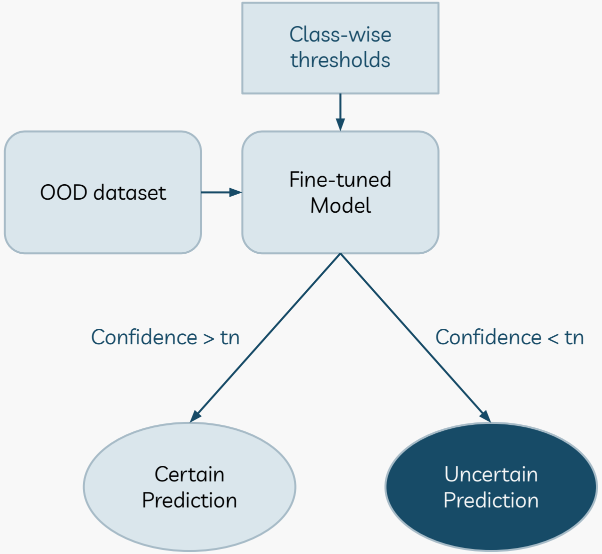
- Train the model using the linear classifier as the head (nn.linear)

Class-wise confidence thresholding:

- Update class-wise thresholds with validation data at the end of each epoch
- Set the initial threshold for each class to 0.8

Overview

[OOD detection Using Class-wise Confidence Thresholding]



Classify OOD dataset:

- Classify the OOD dataset as input into the fine-tuned clip model

Applying class-wise thresholding:

- Apply class-wise thresholding and classify them as uncertain results if they are lower than the classified class's threshold

Evaluation of OOD detection performance:

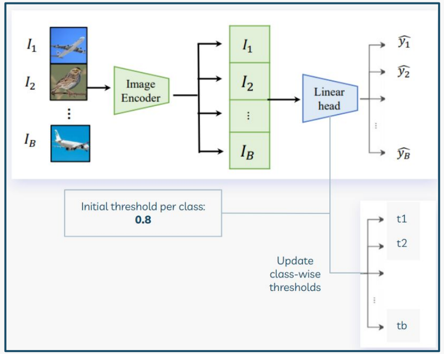
- Evaluate how much OOD data is detected by results deemed uncertain based on class-wise thresholds

Overview

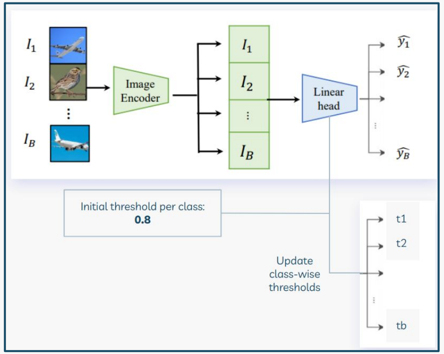
[Applying Ensemble Method]

Voting:

- Vote by gathering predictions from different models, and determine the final prediction with the most voted class or value
- In this study, **adopting the result with a higher confidence value** among the two classified results as the final result



Train **Model 1** with Dataset 1



Train **Model 2** with Dataset 2

Experiment & Result

Fine-tune CLIP model

[Dataset Used]

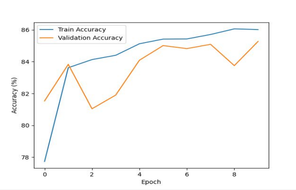
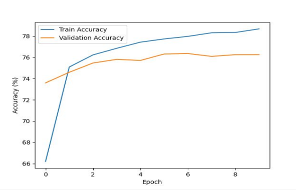
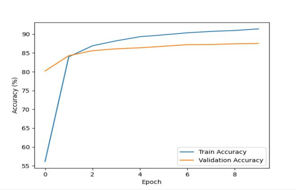
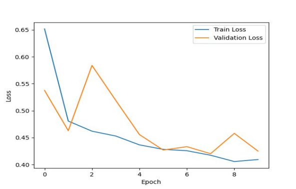
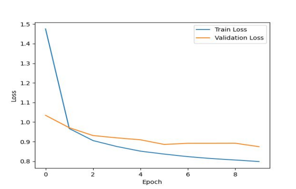
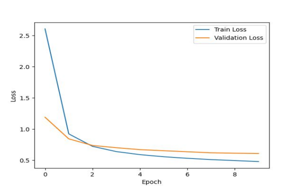
Dataset	CIFAR-10		Food 101		Caltech 256	
Data split	Train	test	Train	test	Train	test
# of classes	10		101		256	
# of images	50,000	10,000	80,800	20,200	25,000	5,000

[Hyperparameters]

Epoch	Batch size	Optimizer	Learning rate	Momentum
10	16	SGD	0.001	0.99

Train result

[Fine-tuning accuracy & loss]

Dataset	CIFAR 10	Food 101	Caltech 256
Accuracy(%)			
loss			

Train result

[Comparing the accuracy of **CLIP base mode**l and **Fine-tuned model**]

	Train dataset			Test dataset		
Dataset	CIFAR 10	Food 101	Caltech 256	CIFAR 10	Food 101	Caltech 256
CLIP (base)	**92.21**	68.75	61.55	**93.10**	66.72	62.08
Fine-tuned	86.02	**91.41**	**78.68**	84.74	**88.34**	**73.09**

Consideration of performance reduction in CIFAR 10 dataset:

- CLFAR-10 dataset that the clip model classifies well in most cases did not see an increase in performance.
- Need to experiment with more diverse datasets and hyperparameter tuning

OOD detection

[Applying Class-wise confidence Thresholds]

- Verify that the 88~94% of OOD dataset is detected by class-wise confidence thresholding

In-dist (model)	OOD	# of OOD images	# of Uncertain predictions	Detection rate(%)
CIFAR 10	Food 101	20,200	19,132	94.71
	Caltech 256	5,000	4,486	89.72
Food 101	CIFAR 10	10,000	9,205	92.05
	Caltech 256	5,000	4,113	82.26
Caltech256	CIFAR 10	10,000	7,899	78.99
	Food 101	20,200	18,722	92.68

OOD detection Using Ensemble method

[With Ensemble Method]

- Two models are used for classification, and the other dataset that is not used for model training is OOD
- Similar or significantly increased performance compared to using only one model

In-dist (model)	OOD	# of OOD images	# of Uncertain predictions	Detection rate(%)
Food 101 Caltech 256	CIFAR 10	1,000	8,838	88.38
CIFAR 10 Caltech 256	Food 101	20,200	19,006	94.09
CIFAR 10 Food 101	Caltech 256	5,000	4,514	90.28

What’s next

What’s next (ing~)

Train a new model by classifying classified uncertain samples:

- Sampling data detected as uncertain and using it as new model learning data
- Apply data clustering approach (DBSCAN)

Evaluate the model using performance indicators for OOD detection:

- Previous studies have measured performance with indicators such as AUROC, TNR of TPR 95%
- It is also important to detect OOD data, but objective indicators are needed

Thank you