中国科学技术大学
University of Science and Technology of China

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

NEURAL INFORMATION PROCESSING SYSTEMS

Zero-Shot Vision Models by Label-Free Prompt Distribution Learning and Bias Correcting

Xingyu Zhu, Beier Zhu, Yi Tan, Shuo Wang, Yanbin Hao, Hanwang Zhang

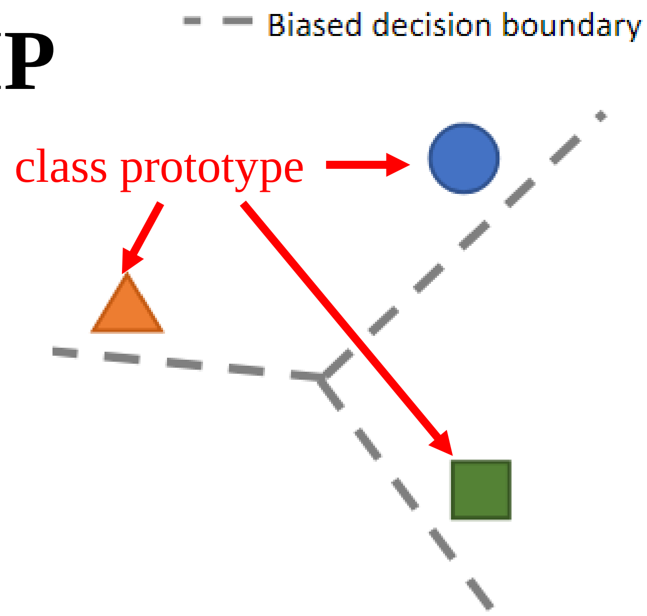
Presented by: Xingyu Zhu

xingyuzhu@mail.ustc.edu.com

1. Motivation
2. Method
3. Experiments
4. Contributions

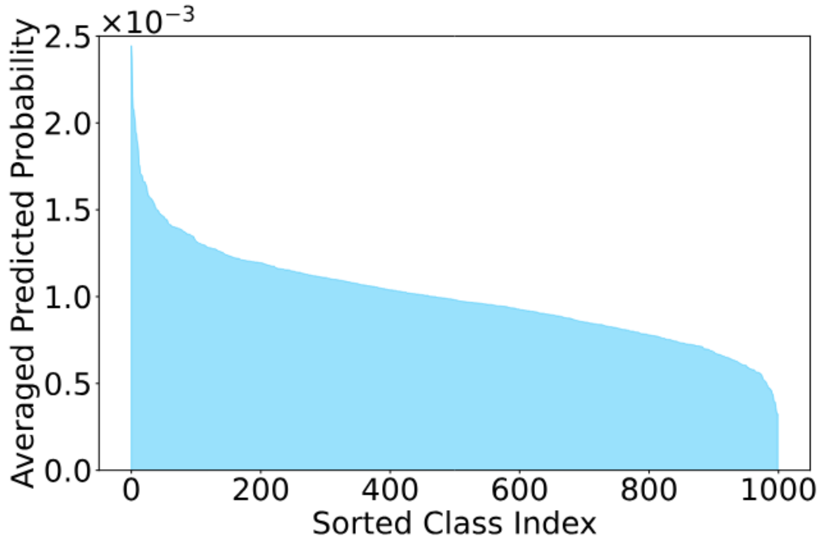
1.Motivation

Prompts optimization & Pre-trained CLIP



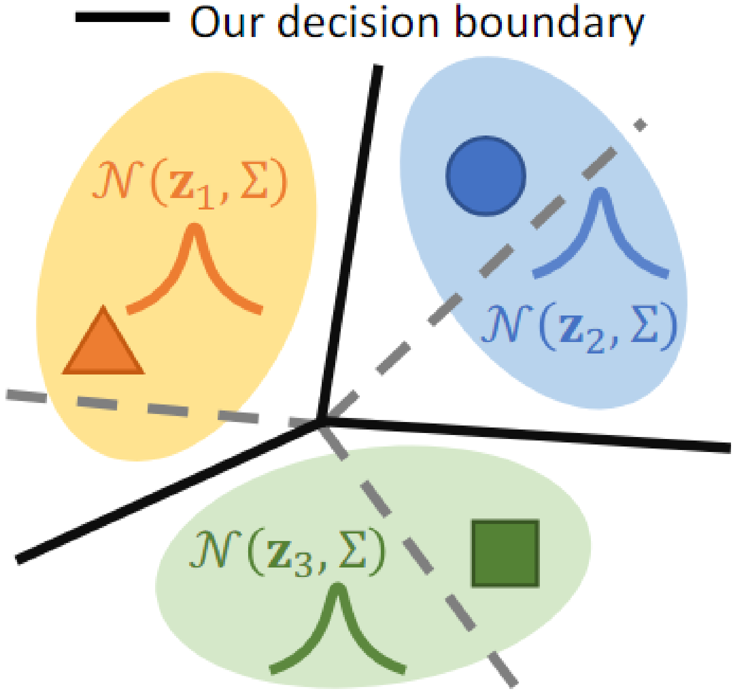
- Text prompts optimization based methods on downstream labeled data has proven effective in improving performance.
  - accuracy of ImageNet from 68.7 to 69.9 with 80 handcraft prompts.

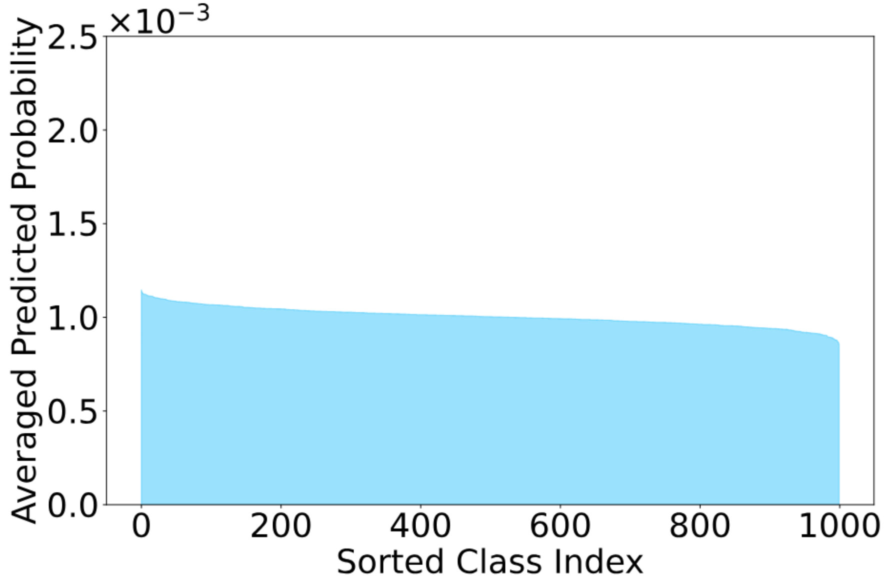
- CLIP model is pre-trained on highly imbalanced Web-scale data, it suffers from inherent label bias.
  - the highest class probability exceeds 0.002, whereas the lowest is below 0.0005.



2.Method

- A label-**F**ree p**ro**mpt distribution **l**earning and b**i**as **c**orrection framework, dubbed as **Frolic**
  - We employ Gaussian distributions to model the varied visual representations of text prototypes and adaptively fuses these with the original CLIP through confidence matching.
  - We develop a bias estimation mechanism, which transitions the sampling process from the pre-training data distribution to a class-conditional sampling from downstream distribution.





2.Method

Label-Free Prompt Distribution Learning：

- Gaussian distribution is effective to model the distribution of the CLIP features , but require extra labeled training data.

$$\mathbb{P}(\mathbf{x}) = \sum_{j=1}^{K} \boxed{\pi_j} \mathcal{N}(\mathbf{x}; \mathbf{z}_j, \boxed{\Sigma}), \quad \mathcal{N}(\mathbf{x}; \mathbf{z}_j, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\{-\frac{1}{2}(\mathbf{x} - \mathbf{z}_j)^\top \Sigma^{-1} (\mathbf{x} - \mathbf{z}_j)\}$$

image data $\{x_i\}_{i=1}^{N}$

the class description $\{z_j\}_{j=1}^{K}$

$$\mathbf{x}_i = \Phi_{\mathsf{v}}(x_i); \quad \mathbf{z}_j = \Phi_{\mathsf{t}}(z_j),$$

$$\boxed{\begin{gathered} y = \underset{j}{\operatorname{argmax}}\, f_{\mathrm{g}}(\mathbf{x})_j = \underset{j}{\operatorname{argmax}}\, \mathbf{w}_j^\top \mathbf{x} + b_j \\ \mathbf{w}_j = \hat{\Sigma}^{-1} \mathbf{z}_j \quad b_j = -\tfrac{1}{2} \mathbf{z}_j^\top \mathbf{w}_j \end{gathered}}$$

$$y = \underset{j}{\operatorname{argmax}}\, f_{\mathrm{c}}(\mathbf{x})_j = \underset{j}{\operatorname{argmax}}\, \mathbf{z}_j^\top \mathbf{x},$$

zero-shot CLIP

2.Method

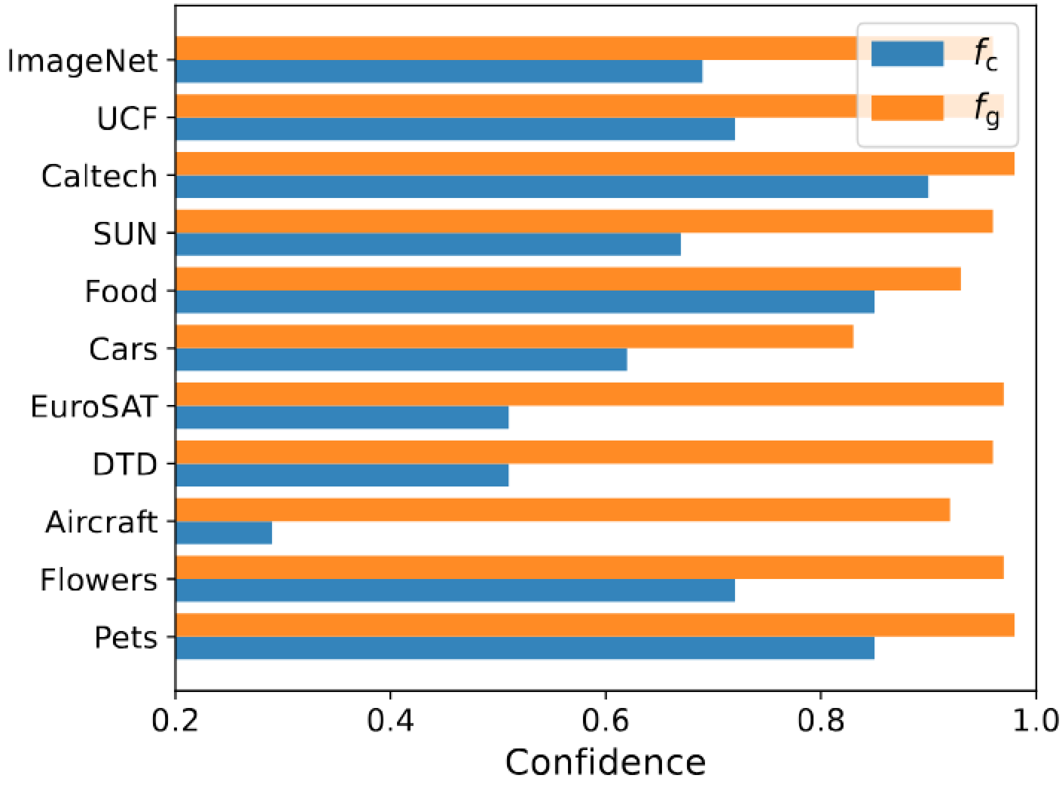
Prediction Fusion via Adaptive Calibration.:

- Combining the zero-shot predictions with the ones from the learned model can further improve performance.

$$f(\mathbf{x}) = f_{\mathrm{c}}(\mathbf{x}) + \boxed{\alpha} f_{\mathrm{g}}(\mathbf{x})$$

$$\begin{cases} \mathrm{conf}(f, \tau) = \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} \max_j \mathrm{softmax}(f(\mathbf{x}_i)/\tau)_j \\ \tau_{\mathrm{g}} = \underset{\tau_{\mathrm{g}}}{\mathrm{argmin}} \left| \mathrm{conf}(f_{\mathrm{g}}, \tau_{\mathrm{g}}) - \mathrm{conf}(f_{\mathrm{c}}, \tau_{\mathrm{c}}) \right| \end{cases}$$

$$\boxed{f_{\mathrm{f}}(\mathbf{x}) = f_{\mathrm{g}}(\mathbf{x})/\tau_{\mathrm{g}} + f_{\mathrm{c}}(\mathbf{x})/\tau_{\mathrm{c}}}$$



2.Method

Correcting Pre-training Label Bias:

- Pre-training datasets typically exhibit a long-tailed concept distribution, leading to biased performance in zero-shot models

$$f_{\mathsf{d}}(\mathbf{x})_y = f_{\mathsf{f}}(\mathbf{x})_y - \boxed{\ln \beta_y}, \quad \beta_y = \mathbb{P}(y)$$

$$s(\mathbf{x}) = \mathrm{softmax}(f_{\mathsf{f}}(\mathbf{x}))$$

$$\boldsymbol{\beta}^0 = [1/K, ..., 1/K]^\top, \; f_{\mathsf{d}}^0 = f_{\mathsf{f}}, \; \mathbf{s}_j^0 = \frac{1}{|\mathcal{C}_j^0|} \sum_{\mathbf{x} \in \mathcal{C}_j^0} s(\mathbf{x}), \text{ and } S^0 = [\mathbf{s}_1^0, ..., \mathbf{s}_K^0]$$

$$\frac{\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^{t-1}\|_1}{\|\boldsymbol{\beta}^{t-1}\|_1} = \|\boldsymbol{\beta}^t - \boldsymbol{\beta}^{t-1}\|_1 < \epsilon, \quad \|\boldsymbol{\beta}^{t-1}\| = 1 \text{ by definition}$$

2.Method

Pipeline of our Frolic & Estimation of β:

Algorithm 1 Pipeline of our Frolic

Given: Unlabeled data $\{\mathbf{x}_i\}_{i=1}^{N}$,
prototypes $\{\mathbf{z}_j\}_{j=1}^{K}$ and τ_{c}

Build $f_{\mathsf{c}}(\mathbf{x})_y = \mathbf{z}_y^\top \mathbf{x}$

Compute $\hat{\Sigma} = \hat{M} - \frac{1}{K}\sum_j \mathbf{z}_j \mathbf{z}_j^\top$
where $\hat{M} = \frac{1}{N}\sum_i \mathbf{x}_i \mathbf{x}_i^\top$

Compute $\mathbf{w}_j = \hat{\Sigma}^{-1}\mathbf{z}_j$, $b_j = -\frac{1}{2}\mathbf{z}_j^\top \mathbf{w}_j$

Build $f_{\mathsf{g}}(\mathbf{x})_y = \mathbf{w}_y^\top \mathbf{x} + b_y$

Search τ_{g} by Eq. (9)

Build $f_{\mathsf{f}}(\mathbf{x}) = f_{\mathsf{g}}(\mathbf{x})/\tau_{\mathsf{g}} + f_{\mathsf{c}}(\mathbf{x})/\tau_{\mathsf{c}}$

Compute $\hat{\beta}$ by Algorithm 2

return $f_{\mathsf{d}}(\mathbf{x}) = f_{\mathsf{f}}(\mathbf{x}) - \ln \hat{\beta}$

Algorithm 2 Estimation of β

Given: Unlabeled data $\{\mathbf{x}_i\}_{i=1}^{N}$,
predictor $f_{\mathsf{f}}(\cdot)$ and tolerance ϵ.

Initialize β^0, f_{d}^0 and S^0 by Eq. (13)

$t = 0$

repeat

$t = t + 1$

Update β^t by solving $(S^{t-1} - I)\beta^t = \mathbf{0}$

Update $f_{\mathsf{d}}^t = f_{\mathsf{f}} - \beta^t$

Update S^t from $\mathbf{s}_j^t = \frac{1}{|\mathcal{C}_j^t|}\sum_{\mathbf{x}\in\mathcal{C}_j^t} s(\mathbf{x})$,
where $\mathcal{C}_j^t$ is assigned by f_{d}^t

until $\|\beta^t - \beta^{t-1}\|_1 < \epsilon$

return $\hat{\beta} = \beta^t$

3.Experiments

Main Results

Table 1: Comparison of accuracy (%) on 10 datasets for CLIP ViT-B/16 and ViT-L/14.

	Method	Pets	Flowers	Aircraft	DTD	EuroSAT	Cars	Food	SUN	Caltech	UCF	Average
ViT-B/16	CLIP [28]	88.9	70.4	24.8	44.3	47.7	65.2	86.1	62.5	92.9	66.7	64.9
	TPT [31]	87.7	68.9	24.7	47.7	42.4	66.8	84.6	65.5	94.1	68.0	65.0
	PromptAlign [30]	90.7	72.3	24.8	47.2	47.8	68.5	86.6	67.5	94.0	69.4	66.8
	SuS-X-SD [34]	90.5	73.8	28.6	54.5	57.4	66.1	86.0	67.7	93.6	66.5	68.4
	TDA [15]	88.6	71.4	23.9	47.4	58.0	67.2	86.1	67.6	94.2	70.6	67.5
	GPT4-Prompt [40]	91.0	74.5	28.0	48.5	48.8	66.8	86.3	65.5	94.6	72.0	67.6
	CuPL-CLIP [26]	92.0	73.2	27.7	54.3	52.7	66.4	86.2	68.5	94.6	70.7	68.6
	Frolic	**92.9**	**74.8**	**31.5**	**56.1**	**58.5**	**69.1**	**87.2**	**70.8**	**95.2**	**75.2**	**71.1**
	InMaP [27]	92.9	71.8	28.4	48.0	64.1	70.6	87.7	70.5	93.1	74.0	70.1
	+ Frolic	**93.6**	**74.3**	**31.8**	**58.0**	**65.3**	**71.7**	**88.2**	**72.8**	**95.4**	**75.9**	**72.7**
ViT-L/14	CLIP [28]	93.5	79.3	32.4	53.0	58.0	76.8	91.0	67.5	94.8	74.2	72.0
	TPT [31]	93.6	76.2	31.9	55.2	51.8	77.7	88.9	70.2	95.5	74.9	71.5
	TDA [15]	93.5	80.5	34.7	56.7	64.1	78.3	90.9	71.5	95.9	76.6	74.2
	GPT4-Prompt [40]	94.1	81.5	36.3	54.8	54.1	77.9	91.4	70.3	96.2	80.6	73.7
	CuPL-CLIP [26]	94.3	79.8	35.5	62.7	61.2	78.0	91.3	72.4	96.7	75.9	74.7
	Frolic	**94.9**	**82.4**	**40.0**	**64.1**	**66.2**	**80.8**	**91.8**	**74.5**	**97.2**	**80.0**	**77.1**
	InMaP [27]	95.2	80.7	37.6	60.2	70.6	82.5	92.2	75.0	94.9	80.4	76.9
	+ Frolic	**95.4**	**81.8**	**42.1**	**66.9**	**71.0**	**83.5**	**92.4**	**77.3**	**97.3**	**82.2**	**78.9**

3.Experiments

Main Results

Table 2: Comparison of accuracy (%) on ImageNet and its variants for CLIP ViT-B/16 and ViT-L/14.

	Method	IN	IN-V2	IN-Sketch	IN-A	IN-R	ObjectNet	Average
ViT-B/16	CLIP [28]	68.7	62.2	48.3	50.6	77.7	53.5	60.1
	TPT [31]	68.9	63.4	47.9	54.7	77.0	55.1	61.1
	TDA[15]	69.5	64.6	50.5	60.1	80.2	55.1	63.3
	GPT4-Prompt [40]	68.7	62.3	48.2	50.6	77.8	53.7	60.2
	CuPL-CLIP [26]	69.9	64.4	49.4	59.7	79.5	53.7	62.7
	Frolic	**70.9**	**64.7**	**53.3**	**60.4**	**80.7**	**56.6**	**64.4**
	InMaP [27]	72.5	62.3	49.4	52.2	79.2	54.5	61.6
	+ **Frolic**	**73.3**	**63.8**	**52.9**	**52.8**	**79.6**	**56.4**	**63.1**
ViT-L/14	CLIP [28]	75.9	70.2	59.7	70.9	87.9	65.5	71.6
	TPT [31]	75.5	70.0	59.8	74.7	87.9	68.0	72.6
	TDA[15]	76.3	71.5	61.3	77.9	89.8	67.0	73.9
	GPT4-Prompt [40]	75.3	70.3	59.9	71.2	87.8	65.7	71.7
	CuPL-CLIP [26]	76.2	71.9	60.7	77.9	89.6	65.7	73.6
	Frolic	**77.4**	**72.5**	**63.1**	**78.9**	**90.3**	**68.7**	**75.1**
	InMaP [27]	79.3	72.1	65.1	62.5	84.8	71.0	72.4
	+ **Frolic**	**79.7**	**73.1**	**65.7**	**64.0**	**85.9**	**71.7**	**73.3**

3.Experiments

Ablation Study

Table 3: Accuracy (%) of different models on 10-datasets, ImageNet and its five variant datasets.

	Model		ViT-B/16			ViT-L/14		
			10-datasets	ImageNet	IN-Variants	10-datasets	ImageNet	IN-Variants
Original CLIP	(1)	f_c	65.1	68.7	58.5	72.0	75.9	72.3
	(2)	$f_c - \ln \beta$	68.4	69.7	61.2	75.1	76.2	73.4
Prompt Distribution	(3)	f_g	68.8	69.8	61.3	74.7	76.0	73.1
	(4)	$f_c + f_g$	66.3	68.9	59.1	72.5	76.1	72.4
Confidence Matching	(5)	$f_f = f_c/\tau_c + f_g/\tau_g$	70.4	69.8	61.9	75.5	76.9	73.9
	(6)	$f_d = f_f - \ln \beta$	**71.1**	**70.9**	**63.1**	**77.2**	**77.4**	**77.4**

3.Experiments

Ablation Study

Table 4: Comparison of accuracy (%) between our Frolic and other label bias correcting methods.

Model	Pets	Flowers	Aircraft	DTD	EuroSAT	Cars	Food	SUN	Caltech	UCF	ImageNet	Avg.
CLIP [28]	89.1	71.4	24.8	44.3	47.7	65.2	86.1	62.5	92.9	66.7	68.7	65.4
TDE [33]	84.1	65.8	27.4	49.8	55.3	60.3	84.6	65.5	91.6	68.2	65.9	65.3
Implicit	91.4	71.4	30.6	54.2	56.8	66.0	86.6	69.5	93.5	72.6	69.8	69.3
Frolic	**92.9**	**74.8**	**31.4**	**56.1**	**58.5**	**69.1**	**87.1**	**70.8**	**95.1**	**75.2**	**70.9**	**70.9**
Oracle Frolic	93.1	77.5	32.2	57.3	59.8	69.8	87.4	71.2	95.7	76.3	71.5	71.9

3.Experiments

Ablation Study

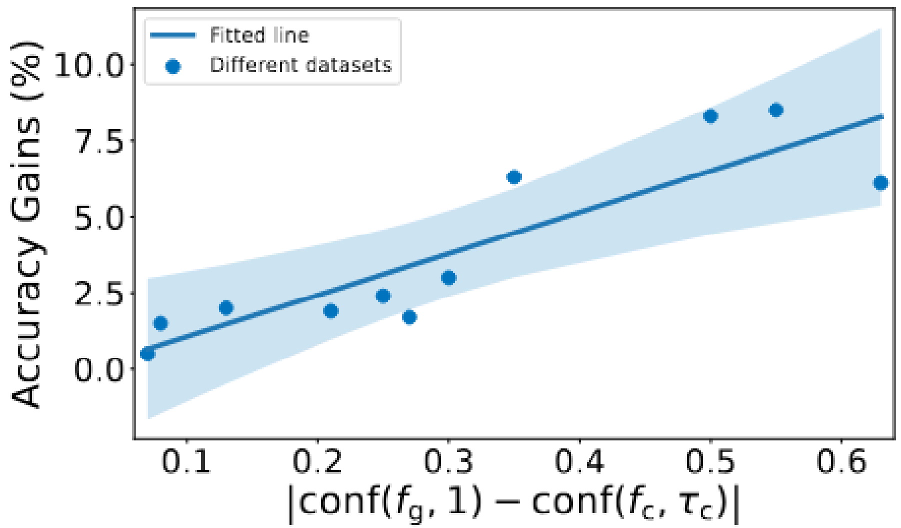


Figure 3: Relation between gains and confidence differences.

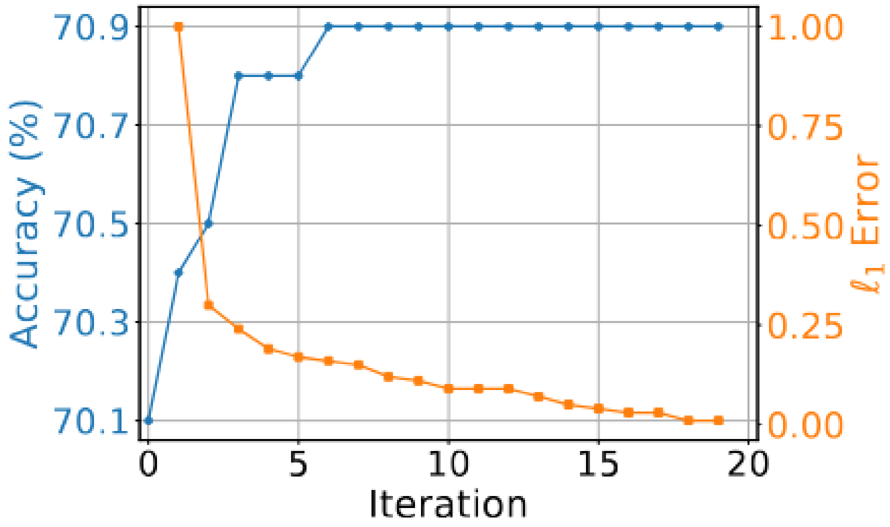


Figure 4: Convergence of accuracy and ℓ_1 error of on ImageNet.

4.Contributions

- We enhance zero-shot performance by estimating a distribution over prompt prototypes to capture the variance in visual appearances. We demonstrate that this process can be implemented entirely without labels.
- We propose a confidence matching technique that fuses the original CLIP model with a Gaussian distribution-based model to further enhance zero-shot performance.
- We develop an unsupervised method to correct pre-training label bias. Unlike existing methods that require access to pre-training data.

中国科学技术大学
University of Science and Technology of China

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

NEURAL INFORMATION PROCESSING SYSTEMS

Thanks All!
Q & A

Presented by: Xingyu Zhu

xingyuzhu@mail.ustc.edu.com