Leveraging Self Weak-supervision for Improved VLM Performance

Shuvendu Roy, Ali Etemad

Queen’s University, Kingston, Canada

Introduction

We present SelfPrompt, a novel semi-supervised prompt-tuning approach for tuning vision-language models (VLMs) in a semi-supervised learning setup. Existing methods for tuning VLMs in semi-supervised setups struggle with efficiently using the limited label set budget, accumulating noisy pseudo-labels, and properly utilizing unlabeled data. SelfPrompt addresses these challenges by introducing (a) a weakly-supervised sampling technique that selects a diverse and representative labelled set, (b) a cluster-guided pseudo-labelling method that improves pseudo-label accuracy, and (c) a confidence-aware semi-supervised learning module that maximizes the utilization of unlabelled data by learning from high- and low-confidence pseudo-labels differently.

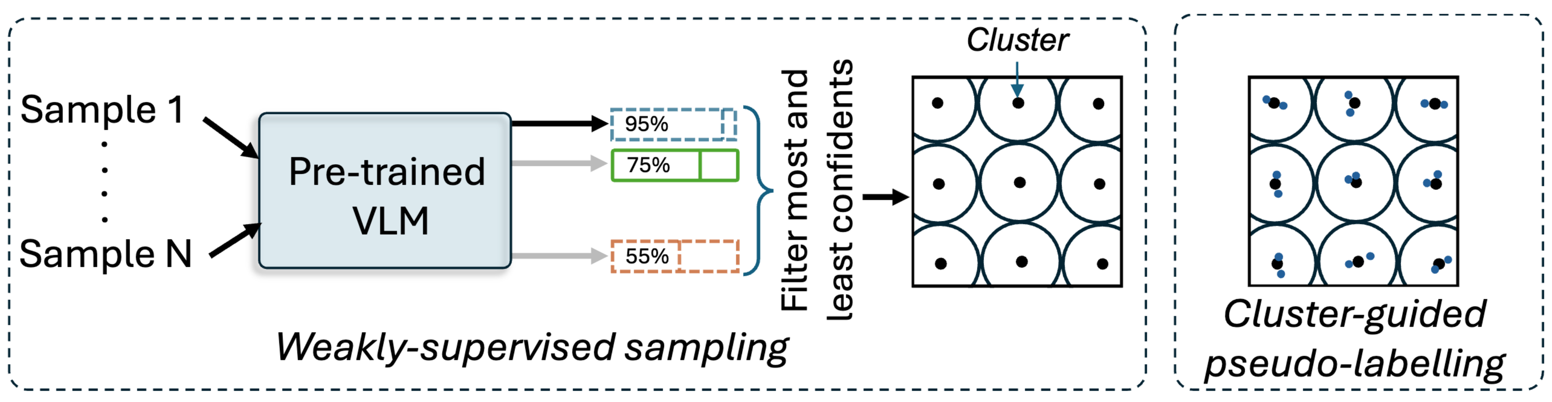

Figure 1: (left) A visual illustration of the weakly-supervised sampling module. Using predictions from the pre-trained VLM, the least and most confident samples, which are not representative of the downstream data, are filtered out. The remaining feature space is then clustered into a number of clusters equal to the labelling budget to ensure maximum diversity among the selected samples. (right) Cluster-guided pseudo-labelling assigns the same class label to samples near the cluster centers as the pseudo-label.

Method

Weakly-supervised sampling

To overcome the limitations of random selection, we introduce a weakly-supervised sampling module that selects the most diverse and representative N samples from the unlabelled set. This module operates through a two-step protocol:

Step 1: Filtering with weak supervision. We leverage the zero-shot predictions of the pre-trained VLM as weak supervision to filter the unlabelled set U . Specifically, we remove samples with both the highest and lowest confidence predictions by the VLM. To this end, we divide the sorted samples into q quantiles, $\{Q_1, Q_2, \cdots Q_q\}$, and select $\mathcal{D}_{\text{filtered}} = \bigcup_{k=2}^{q-1} Q_k$

Step 2: Diversity Sampling. We select *N* diverse samples from the filtered dataset with a cluster-based sampling technique. To this end, we apply k-means clustering to group the samples into *N* clusters and select one sample per cluster closest to the cluster center.

Cluster-guided pseudo-labelling

To improve the pseudo-label quality, especially at the beginning of the training, we propose a novel clustering-guided pseudo-labelling approach that does not utilize the zero-shot prediction from the VLM as the pseudo-label. Specifically, for each cluster $\mathcal{C}_j$, we pick the p samples closest to the cluster centers to form a pseudo-label set $\mathcal{P}_j = \{x_j^1, x_j^2, \ldots, x_j^p\}$.

Confidence-aware semi-supervised learning

To make the best use of the unlabelled data, we propose a confidence-aware semi-supervised module that learns from the high-confident samples in a supervised learning setup, while learning from the low-confident samples in a weakly-supervised setting. We first predict the output distribution for each sample in the unlabelled set $\boldsymbol{p}_i = f(x_i) \in \mathbb{R}^C$. Then we incorporate the *t* most confident samples-per-class into our pseudo-label set as:

$$\mathcal{X}^+ = \mathcal{X}_P \cup \Big(\bigcup_{c=1}^{C} \text{top}_t(\{\boldsymbol{x}_i | \arg\max(\boldsymbol{p}_i) = c\}) \Big)$$

Finally, we learn from the labelled set, pseudo-labeled set, and weakly labelled set, together as follow:

$$\mathcal{L}_{final} = \frac{1}{|\mathcal{X}_L|} \sum_{(x,y)\in\mathcal{X}_L} \ell(f(x), y) + \frac{1}{|\mathcal{X}+|} \sum_{(x,y)\in\mathcal{X}+} \ell(f(x), y) + \frac{\lambda}{|\mathcal{X}_{weak}|} \sum_{(x,s)\in\mathcal{X}weak} \ell_w(f(x), \boldsymbol{s})$$

Here, ℓ_w is a partial label learning loss defined as:

$$\ell_w(f(x), \boldsymbol{s}) = -\sum_{c\in C} \boldsymbol{s}^c \log\left(p(c|x)\right)$$

Algorithm 1 SelfPrompt

Input: Unlabelled set U with M samples, label budget N, pre-trained VLM (θ, ϕ), learnable prompt P, number of sessions S, hyper-parameters t, number of clusters N, pseudo-labels per cluster p
// Filtering with weak supervision
for each $x_i \in U$ **do**
  Compute the class probability distribution using Eq. 1 as: $\boldsymbol{p}_i = [p_i^1, p_i^2, \cdots, p_i^C]$
  Compute confidence scores, $c_i = \max_{1\le c\le C}(p_i^c)$
end for
Sort samples in descending order of c_i and divide into q quantiles, $\{Q_1, Q_2, \cdots Q_q\}$.
Remove samples from first and last quantiles to get $\mathcal{D}_{\text{filtered}} = \bigcup_{k=2}^{q-1} Q_k$
// Diversity Sampling
Extract embeddings $\mathbf{z}_i = \theta(x_i)$ for $x_i \in \mathcal{D}_{\text{filtered}}$.
Perform k-means clustering on $\{\mathbf{z}_i\}$ to form N clusters $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_N\}$.
Select a sample from each cluster j, $x_j^* = \arg\min_{x_i\in\mathcal{C}_j} \|\mathbf{z}_i - \boldsymbol{\mu}_j\|^2$ for $j = 1, \ldots, N$, where $\boldsymbol{\mu}_j$ is the cluster center.
Form labelled set $\mathcal{X}_L = \{(x_1^*, y_1), (x_2^*, y_2), \ldots, (x_N^*, y_N)\}$.
// Cluster-guided pseudo-labelling
for $j = 1$ to N **do**
  Select p additional samples per cluster $\mathcal{C}_j$ nearest to the cluster center x_j^*.
  Assign cluster label to selected samples: $\mathcal{P}_j = \{(x_{jk}, y_j)\}_{k=1}^{p}$
end for
Create pseudo-label set: $\mathcal{X}_p = \bigcup_j \mathcal{P}_j$
// Confidence-aware semi-supervised learning
for $s = 1$ to S **do**
  if $s == 1$ **then continue**
  end if
  Predict probability distribution, $\boldsymbol{p}_i = f(x_i) \in \mathbb{R}^C$, for $x_i \in U$
  Update $\mathcal{X}^+$ as, $\mathcal{X}^+ = \mathcal{X}_P \cup \big(\bigcup_{c=1}^{C} \text{top}_t(\{\boldsymbol{x}_i | \arg\max(\boldsymbol{p}_i) = c\})\big)$
  Form weakly-labelled set: $\mathcal{X}_{\text{weak}} = \{(x_i, s_i) \mid x_i \in U \setminus \mathcal{X}^+\}$
  Train VLM using loss:

$$\mathcal{L}_{\text{final}} = \frac{1}{|\mathcal{X}_L|} \sum_{(x,y)\in\mathcal{X}_L} \ell(f(x), y) + \frac{1}{|\mathcal{X}^+|} \sum_{(x,y)\in\mathcal{X}^+} \ell(f(x), y) + \frac{\lambda}{|\mathcal{X}_{\text{weak}}|} \sum_{(x,s)\in\mathcal{X}_{\text{weak}}} \ell_w(f(x), s)$$

end for

Experiments

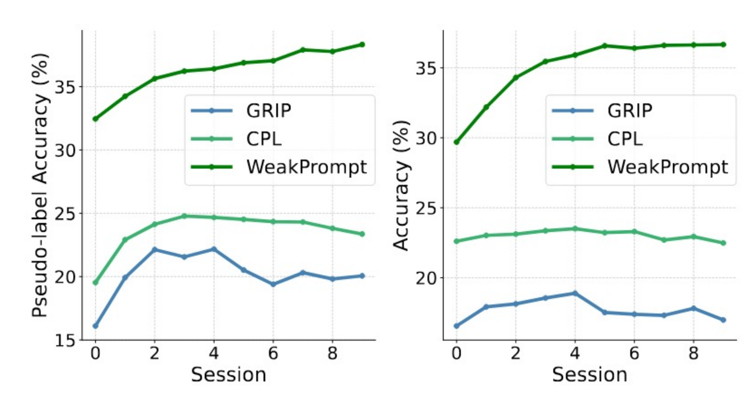

Figure 2: (left) Pseudo-label accuracy; (right) Test accuracy over training sessions.

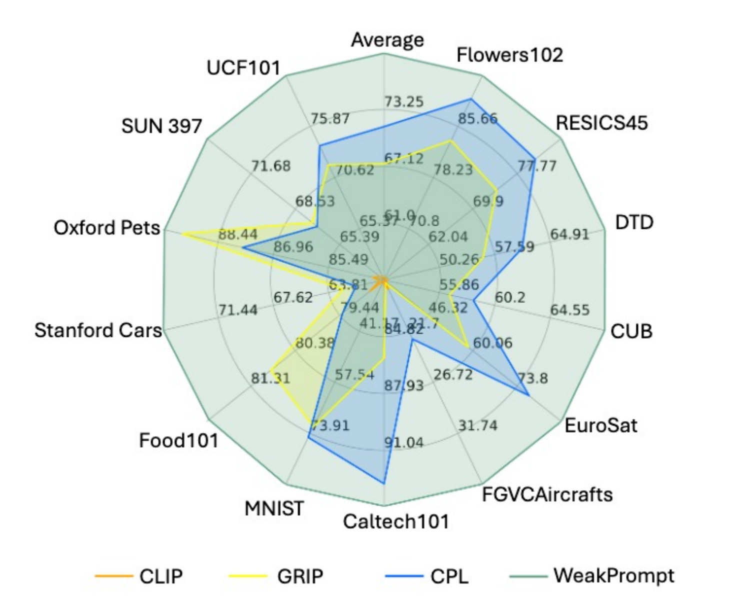

Figure 3: Performance comparison to prior works on semi-supervised tuning of VLMs.

Table 1: Comparison results of top-1 test accuracy (%) on 13 benchmarks on semi-supervised learning with textual prompt strategy.

Methods	Average	Flowers102	RESISC45	DTD	CUB	EuroSAT	FGVCAircraft
Zero-shot CLIP	55.17	$63.67_{0.00}$	$54.48_{0.00}$	$43.24_{0.00}$	$51.82_{0.00}$	$32.88_{0.00}$	$17.58_{0.00}$
CoOp	62.28	$75.96_{0.74}$	$68.13_{0.55}$	$37.10_{5.45}$	$55.29_{0.59}$	$62.05_{1.64}$	$20.02_{0.77}$
GRIP	67.40	$83.60_{0.48}$	$74.11_{0.68}$	$56.07_{0.79}$	$56.65_{0.33}$	$58.66_{2.64}$	$16.98_{0.20}$
CPL	71.41	$89.66_{0.36}$	$80.98_{0.11}$	$61.21_{0.56}$	$58.53_{0.24}$	$77.51_{0.80}$	$22.48_{0.63}$
SelfPrompt	**79.33**	$\mathbf{93.04}_{0.33}$	$\mathbf{85.58}_{0.18}$	$\mathbf{72.18}_{0.78}$	$\mathbf{68.84}_{0.16}$	$\mathbf{87.49}_{0.12}$	$\mathbf{36.71}_{0.70}$
Δ	↑ 7.92	↑ 3.38	↑ 4.60	↑ 10.97	↑ 12.31	↑ 9.98	↑ 14.23
	Caltech101	**MNIST**	**Food101**	**StanfordCars**	**OxfordPets**	**SUN397**	**UCF101**
Zero-shot CLIP	$82.01_{0.00}$	$25.10_{0.00}$	$78.81_{0.00}$	$60.29_{0.00}$	$84.32_{0.00}$	$62.54_{0.00}$	$60.42_{0.00}$
CoOp	$84.69_{1.43}$	$58.22_{1.98}$	$76.23_{1.45}$	$58.23_{2.45}$	$82.34_{1.44}$	$62.19_{1.78}$	$69.19_{1.03}$
GRIP	$85.99_{1.06}$	$71.78_{2.59}$	$80.89_{1.14}$	$62.83_{1.42}$	$89.40_{0.33}$	$67.34_{0.98}$	$71.94_{0.95}$
CPL	$92.87_{1.14}$	$75.18_{4.40}$	$79.38_{1.05}$	$61.93_{1.30}$	$87.79_{1.31}$	$66.98_{0.65}$	$73.88_{1.32}$
SelfPrompt	$\mathbf{94.10}_{0.92}$	$\mathbf{90.23}_{0.36}$	$\mathbf{82.19}_{0.17}$	$\mathbf{75.21}_{0.33}$	$\mathbf{89.86}_{0.48}$	$\mathbf{74.77}_{0.18}$	$\mathbf{81.07}_{0.44}$
Δ	↑ 1.23	↑ 15.05	↑ 2.81	↑ 13.28	↑ 2.07	↑ 7.79	↑ 7.19

Table 2: Comparison with existing SOTA on base-to-novel generalization in a 2-shot training setup.

ViT-B/16	ul.	Base	Novel	HM
CLIP	✗	69.3	74.2	71.7
Co-CoOp	✗	71.9	73.4	72.6
MaPLe	✗	74.9	73.3	74.0
PromptSRC	✗	78.1	74.7	76.3
CoPrompt	✗	74.2	72.4	73.1
PromptKD	✓	79.7	76.8	78.1
SelfPrompt	✓	**85.6**	**80.8**	**83.0**
Δ		↑ 5.9	↑ 4.0	↑ 4.9

Table 3: Ablation Study

W.S.S.	C.G.P.	C.A.SSL	Accuracy
✓	✓	✓	79.33
✗	✓	✓	76.12
✓	✗	✓	74.39
✓	✓	✗	78.01
✗	✗	✓	73.49
✗	✓	✗	75.67
✓	✗	✗	73.08
✗	✗	✗	71.41

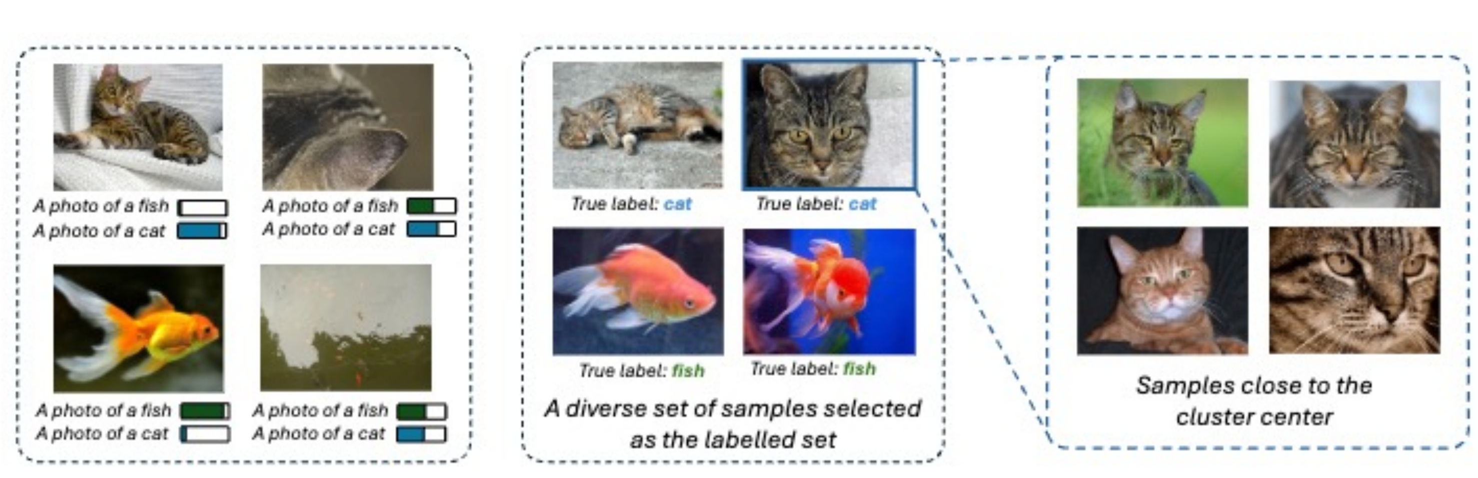

Figure 4: Qualitative analysis of weakly-supervised sampling and cluster-guided pseudo-labelling with two classes (fist and cat). (left) Illustrations of the most confident samples, which provide minimal information gain, alongside the least confident samples, which are less representative of their respective classes. (middle) Examples of selected samples demonstrating high semantic diversity. (right) Samples close to the cluster centers) exhibit high visual and semantic similarity.