

Appendix

Anonymous Author(s)

Affiliation

Address

email

1 Proof of Proposition

When the training set is \mathcal{X} , the prediction of NTK for the test sample x , $\hat{f}(x)$, is as in eq. 1, where \mathcal{Y} denotes labels of training samples with one-hot form and we denote $\mathcal{K}(x, \mathcal{X})\mathcal{K}(\mathcal{X}, \mathcal{X})^{-1}$ as weight matrix W . The difference between NTK predictions with true labels, $\hat{f}_y(x)$, and predictions with CPL transformed to true label classes through a label mapping function g , $g(\hat{f}_{cpl}(x))$, as shown in eq. 2. We write the product of the row vector W and the labels as the element-wise multiplication, where w_i represents the i^{th} element of W . When true labels of labeled samples are the dominant labels in their corresponding CPL clusters, $y_i = g(y_{cpl,i})$ and the result of the eq. 2 is zero, i.e., $\hat{f}_y(x) = g(\hat{f}_{cpl}(x))$.

$$\begin{aligned}\hat{f}(x) &= \hat{f}_0(x) + \mathcal{K}(x, \mathcal{X})\mathcal{K}(\mathcal{X}, \mathcal{X})^{-1}(\mathcal{Y} - f_0(\mathcal{X})) \\ &= \hat{f}_0(x) + W\hat{f}_0(x) - W\mathcal{Y}\end{aligned}\tag{1}$$

$$\begin{aligned}\hat{f}_y(x) - g(\hat{f}_{cpl}(x)) &= W\mathcal{Y} - g(W\mathcal{Y}_{cpl}) \\ &= \sum_{i \in D_C} (w_i y_i) - \sum_{i \in D_C} (w_i g(y_{cpl,i})) \\ &= \sum_{i \in D_C} w_i (y_i - g(y_{cpl,i}))\end{aligned}\tag{2}$$

Empirical evidence of assumption Based on proposition, $\hat{f}_y(x_i) = g(\hat{f}_{cpl}(x_i))$, we argue $\argmax \hat{f}_y(x_i)$ is most likely equal to $g(\argmax \hat{f}_{cpl}(x_i))$. We validate this claim on the CIFAR-10 and CIFAR-100 datasets. The validity of this claim at different numbers of annotations is shown in fig. 1. This claim is valid in most cases, especially when the dataset has few classes.

Explanation of impurity error and over-clustering error As analyzed in sec.3.3, the approximation error consists of P_{fnf} and P_{nff} . The P_{fnf} denotes the NTK prediction agrees with y but does not agree with y_{cpl} . From our claim, it can be deduced that the true label class y_i corresponds to at least two different CPL classes: $y_{cpl,i}$ and $\argmax \hat{f}_{cpl}(x_i)$, i.e., over-clustering.

The P_{nff} denotes the NTK prediction does not agree with y but agrees with y_{cpl} . Based on our claim, $\argmax \hat{f}_y(x_i)$ would be the dominant class within the CPL. But it is different from the true label, y_i , i.e., y_i is not the dominant class within this CPL cluster. So, this CPL cluster includes the impure sample.

Empirical evidence of impurity error and over-clustering error To validate the analysis of over-clustering error and impurity error, we conducted experiments on the CIFAR-10 and CIFAR-100 datasets. Specifically, we examined the ratios of impure samples within CPL in the P_{nff} term

25 and the ratios of over-clustering in the P_{fnf} term. The results are shown in fig. 2 and fig. 3. The
 26 experimental results show that impurity and over-clustering explain most of the approximation errors.

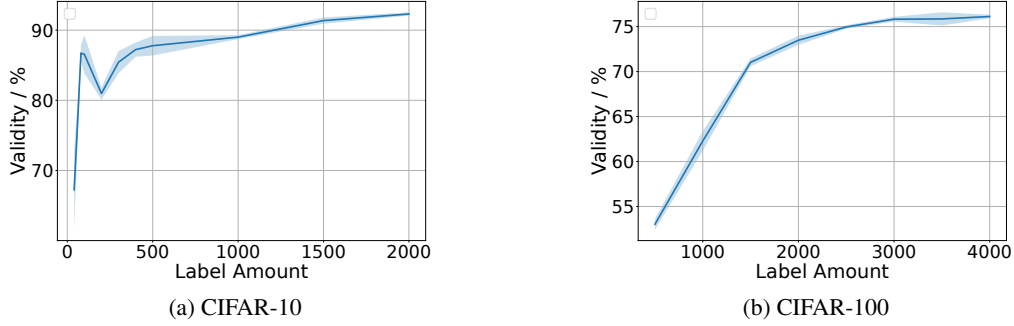


Figure 1: The validity of the claim ($\argmax \hat{f}_y(x_i) = g(\argmax \hat{f}_{cpl}(x_i))$) at different numbers of annotations. The shaded area represents std.

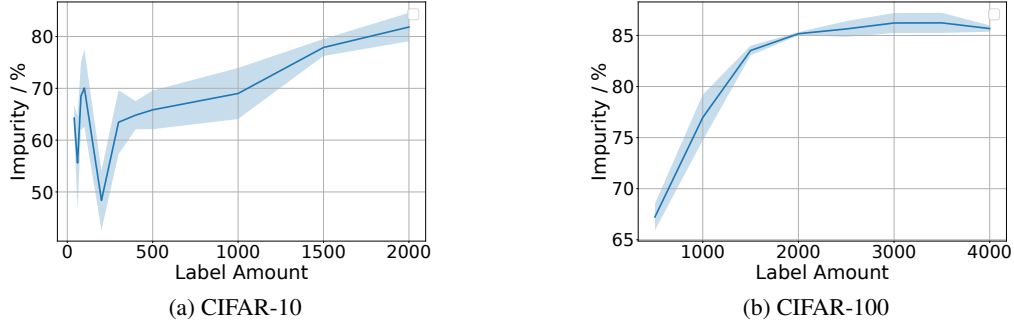


Figure 2: Proportion of approximation error caused by impure samples within CPL to the P_{nff} term. The shaded area represents std.

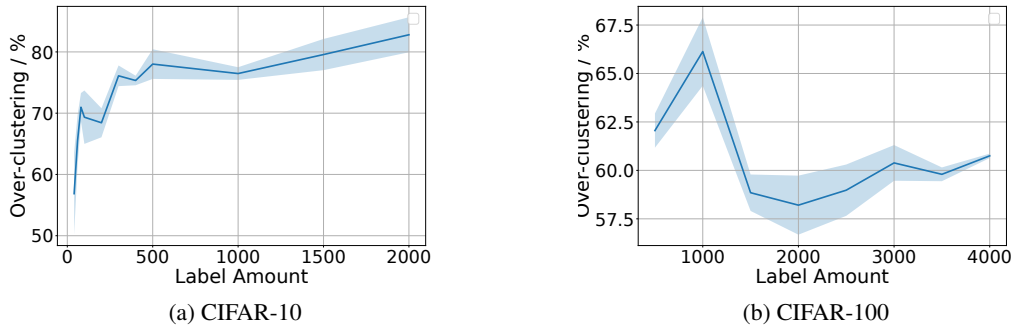


Figure 3: Proportion of approximation error caused by over-clustering of CPL to the P_{fnf} term. The shaded area represents std.

27 2 Dataset Description

28 CIFAR-10 and CIFAR-100 datasets both consist of 50,000 training samples and 10,000 testing
 29 samples, with a resolution of 32x32. CIFAR-10 is composed of 10 classes, while CIFAR-100 consists
 30 of 100 classes. Similarly, SVHN dataset also contains 10 classes, with 73,257 training samples and
 31 26,032 testing samples. Its resolution is 32x32. ImageNet-100 is a subset of the ImageNet dataset,

comprising 100 classes, with 128,545 training samples and 5,000 testing samples. Oxford-IIIT Pet dataset includes 37 classes, with 3,680 training samples and 3,669 testing samples. For both ImageNet-100 and Oxford-IIIT Pet dataset, all samples were resized to a resolution of 224x224 following the method described in [14].

3 Hyperparameters of Training

Classifier is a 2-layer MLP with the architecture: Linear + BatchNorm + ReLU + Linear. The dimension of the output of the first linear layer is shown in table 1. Total number of training epochs is 100. Data augmentation includes random crops and horizontal flips. The remaining hyperparameters are shown in table 1.

Table 1: Hyperparameters

Dataset	Backbone	Classifier	Learning Rate	Momentum	Weight Decay	Batch Size
CIFAR-10	Resnt18	MLP64	0.3	0.9	0.0003	100
CIFAR-100	WRN-28-8	MLP512	0.3	0.9	0.0003	100
ImageNet-100	Resnet50	MLP4096	0.1	0.9	0	512
SVHN	Resnt18	MLP512	0.3	0.9	0.0003	100
Oxford-IIIT Pet	Resnet50	MLP512	0.1	0.9	0	128

4 Computation time

We compare the computation time of our method with that of LookAhead [26] (an active learning method based on NTK) as shown in fig. 4. Our method demonstrates a comparable sample selection time to LookAhead. When the dataset with few classes, our method takes slightly longer than LookAhead. Conversely, when the dataset with many classes, our method takes slightly less time than LookAhead.

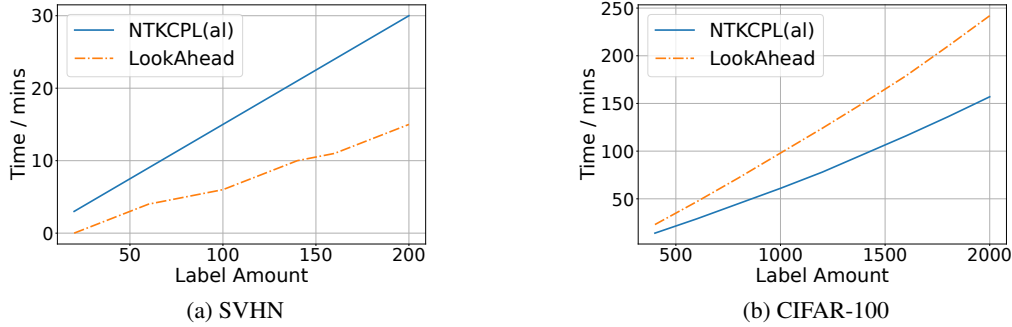


Figure 4: Comparison of cumulative sample selection time on single RTX 3090 GPU.

5 Experiment Results

The results of the experiment are shown in table 2, table 3, table 4 and table 5, where we report the average accuracy and std. over 5 runs.

6 Broader Impacts

This paper focuses on active learning. The goal of research in this field is to help reduce the cost of sample annotation. Specifically, this paper investigates active learning on top of self-supervised models. By aiming to more accurately estimate the model’s empirical risk across the entire active learning pool, we propose an active learning method that operates effectively over a wider range of annotation quantities. We do not anticipate any potential negative social impacts arising from this work.

Table 2: Comparison of accuracy of different active learning strategies on CIFAR-100 with budget step size 500. All results are averages over 5 runs. The best results are shown in red and the second-best results are shown in blue.

# Labels	Random	Entropy	Coreset(self)	Coreset(al)	BADGE	TypiClust	NTKCPL(al)
500	44.76 \pm 0.77	44.79 \pm 0.71	42.88 \pm 0.73	44.56 \pm 1.16	45.06 \pm 1.08	50.06\pm0.79	49.20 \pm 0.66
1000	52.51 \pm 0.36	49.50 \pm 0.56	50.44 \pm 0.58	51.55 \pm 0.41	52.40 \pm 0.66	54.12\pm0.31	53.99 \pm 0.27
1500	55.70 \pm 0.51	52.19 \pm 0.41	54.25 \pm 0.47	54.38 \pm 0.29	55.43 \pm 0.40	55.89\pm0.27	56.52\pm0.24
2000	57.12 \pm 0.62	54.30 \pm 0.17	56.03 \pm 0.45	56.02 \pm 0.54	57.34\pm0.36	56.67 \pm 0.23	57.80\pm0.49
2500	58.34 \pm 0.42	55.32 \pm 0.38	57.22 \pm 0.27	57.19 \pm 0.70	58.53\pm0.35	57.35 \pm 0.13	59.00\pm0.56
3000	59.22 \pm 0.38	56.69 \pm 0.42	58.47 \pm 0.38	58.44 \pm 0.68	59.64\pm0.39	57.75 \pm 0.12	59.92\pm0.57
3500	59.97 \pm 0.30	57.67 \pm 0.36	59.06 \pm 0.42	59.13 \pm 0.54	60.29\pm0.16	58.25 \pm 0.20	60.53\pm0.42
4000	60.70 \pm 0.25	58.42 \pm 0.32	59.73 \pm 0.40	59.73 \pm 0.46	61.29\pm0.20	58.55 \pm 0.27	61.16\pm0.34

Table 3: Comparison of accuracy of different active learning strategies on ImageNet-100. All results are averages over 5 runs. The best results are shown in red and the second-best results are shown in blue.

# Labels	Random	Entropy	Coreset(self)	BADGE	TypiClust	LookAhead	NTKCPL(self)	NTKCPL(al)
200	50.89 \pm 3.33	52.21 \pm 1.17	49.19 \pm 2.54	52.64 \pm 2.21	62.47 \pm 1.73	50.75 \pm 2.71	63.98\pm1.39	65.88\pm1.00
400	62.98 \pm 1.81	57.13 \pm 1.59	59.94 \pm 1.03	65.68 \pm 1.95	67.81 \pm 1.31	63.82 \pm 0.56	69.69\pm0.78	72.18\pm1.36
600	74.30 \pm 1.09	63.09 \pm 0.97	70.13 \pm 1.21	75.31 \pm 0.60	74.45 \pm 0.45	72.82 \pm 1.06	77.03\pm0.77	76.84\pm0.72
800	76.17 \pm 0.68	67.55 \pm 0.95	73.39 \pm 1.09	77.50 \pm 0.53	76.30 \pm 0.17	74.20 \pm 1.00	78.28\pm0.58	78.81\pm0.71
1000	77.56 \pm 0.59	70.56 \pm 1.85	74.96 \pm 0.87	79.13 \pm 0.27	77.80 \pm 0.55	75.27 \pm 1.22	79.34\pm0.44	80.17\pm0.64
1500	80.78 \pm 0.16	75.82 \pm 0.76	78.71 \pm 0.45	81.28 \pm 0.23	80.63 \pm 0.49	78.23 \pm 0.41	81.41\pm0.31	81.84\pm0.40
2000	81.98 \pm 0.28	76.60 \pm 0.38	80.30 \pm 0.47	81.71 \pm 0.27	81.76 \pm 0.25	79.73 \pm 0.76	82.42\pm0.56	82.77\pm0.21

Table 4: Comparison of accuracy of different active learning strategies on SVHN. All results are averages over 5 runs. The best results are shown in red and the second-best results are shown in blue.

# Labels	Random	Entropy	Coreset(self)	BADGE	TypiClust	LookAhead	NTKCPL(self)	NTKCPL(al)
20	32.30 \pm 3.44	31.41 \pm 2.00	25.57 \pm 1.42	30.70 \pm 4.61	33.42\pm3.28	29.88 \pm 3.86	31.40 \pm 5.76	33.87\pm3.42
40	47.19 \pm 3.06	42.84 \pm 5.89	33.41 \pm 4.35	47.28 \pm 6.95	50.18\pm2.79	49.27\pm4.70	46.51 \pm 3.26	48.63 \pm 2.49
60	58.84 \pm 1.72	54.40 \pm 6.78	48.32 \pm 2.23	60.48 \pm 4.86	58.47 \pm 2.79	60.34\pm4.21	56.73 \pm 5.38	61.16\pm4.52
80	67.15 \pm 1.46	59.42 \pm 7.84	57.33 \pm 1.99	67.27 \pm 4.54	67.77\pm2.63	65.66 \pm 3.13	67.73 \pm 6.35	74.26\pm2.71
100	69.22 \pm 1.97	64.59 \pm 8.29	63.91 \pm 2.22	72.83 \pm 5.21	72.33 \pm 2.37	69.85 \pm 3.48	72.80\pm4.11	77.87\pm2.33
120	70.18 \pm 2.40	67.29 \pm 5.88	65.77 \pm 2.04	74.61 \pm 1.90	74.68 \pm 1.40	72.24 \pm 2.04	76.85\pm2.34	78.39\pm1.75
140	74.43 \pm 1.44	70.60 \pm 5.67	72.49 \pm 1.69	78.16 \pm 1.89	78.38\pm2.06	75.33 \pm 2.76	78.01 \pm 1.40	81.50\pm1.13
160	75.84 \pm 2.26	72.54 \pm 5.88	75.93 \pm 1.56	81.00 \pm 1.84	79.08 \pm 1.71	77.28 \pm 6.27	81.20\pm1.40	83.48\pm1.64
180	78.13 \pm 2.26	74.53 \pm 7.23	77.26 \pm 2.14	82.70\pm2.10	80.03 \pm 1.97	79.05 \pm 3.51	81.77 \pm 1.83	84.83\pm1.36
200	80.35 \pm 2.16	74.88 \pm 5.02	78.83 \pm 2.15	82.78 \pm 1.72	82.21 \pm 0.88	80.83 \pm 3.95	84.40\pm1.40	85.08\pm0.91

Table 5: Comparison of accuracy of different active learning strategies on Oxford-IIIT Pet. All results are averages over 5 runs. The best results are shown in red and the second-best results are shown in blue.

<i># Labels</i>	<i>Random</i>	<i>Entropy</i>	<i>Coreset(self)</i>	<i>Coreset(al)</i>	<i>BADGE</i>	<i>TypiClust</i>	<i>NTKCPL(self)</i>	<i>NTKCPL(al)</i>
40	38.60±2.33	36.35±3.78	43.73±3.70	36.40±2.39	37.37±1.95	51.99±2.34	52.19±2.40	52.64±2.79
80	52.57±2.82	46.86±3.56	54.35±2.03	52.41±2.55	54.13±1.31	63.45±1.47	60.27±1.71	60.59±1.86
120	59.91±2.56	54.47±4.94	60.54±2.43	63.69±1.16	64.21±1.81	68.05±1.57	66.42±1.36	66.35±2.44
160	64.16±1.49	58.96±4.69	65.97±1.54	66.66±1.66	66.39±0.74	69.82±0.91	68.63±1.31	71.27±1.95
200	68.57±1.86	64.66±2.67	70.21±1.71	71.57±0.98	71.20±1.49	72.74±1.48	71.92±1.43	73.43±1.06
240	71.73±1.47	68.37±2.29	73.35±1.12	74.73±0.84	74.15±0.85	74.60±0.74	75.07±1.84	76.35±1.09
280	72.60±1.63	68.37±1.94	72.84±0.95	74.29±1.44	73.63±0.87	74.39±0.92	74.75±1.57	76.15±1.34
320	77.20±0.84	72.73±1.58	75.12±1.15	75.88±0.83	77.24±1.91	77.78±1.00	77.22±1.13	78.54±0.51
360	78.28±0.64	74.97±1.68	76.35±1.17	77.97±0.36	78.22±0.64	78.60±0.92	78.44±1.21	79.67±0.62
400	77.01±1.22	73.70±1.72	75.38±0.64	76.62±0.88	77.35±0.88	77.27±1.09	77.25±1.38	79.05±0.17