Rethinking deep active learning: Using unlabeled data at model training *Supplementary material*

Oriane Siméoni Mateusz Budnik Yannis Avrithis Guillaume Gravier
Inria, IRISA, Univ Rennes, CNRS

APPENDIX

It has been observed that most acquisition strategies do not provide a significant improvement over standard uncertainty when using deep neural networks; for instance, all strategies perform similarly on CIFAR-10 and CIFAR-100 according to [14] and [8]. To better understand the differences, the ranks of examples acquired by different strategies are compared pairwise by [14]. We make a step further in this direction, using label propagation as a tool.

A. Measuring agreement

After the classifier is trained at any cycle using any reference acquisition function a, we apply two different acquisition functions, say $a^{(1)}$ and $a^{(2)}$, followed by labeling of acquired examples and label propagation, obtaining two different sets of predicted pseudo-labels $\hat{\mathbf{y}}^{(1)}$ and $\hat{\mathbf{y}}^{(2)}$ and weights $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$ on the unlabeled examples U. We define the *weighted accuracy*

$$A_{U,\mathbf{w}}(\mathbf{z}, \mathbf{z}') = \sum_{i \in U} \eta[\mathbf{w}]_i \delta_{z_i, z'_i} \tag{13}$$

for $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^{|U|}$, where δ is the Kronecker delta function. Using the average weights $\mathbf{w} := \frac{1}{2}(\mathbf{w}^{(1)} + \mathbf{w}^{(2)})$, we then measure the weighted accuracy $A_{U,\mathbf{w}}(\mathbf{y}^{(1)}, \mathbf{y}^{(2)})$, expressing the *agreement* of the two strategies, as well as the weighted accuracy $A_{U,\mathbf{w}}(\mathbf{y}^{(k)}, \mathbf{t})$ of $a^{(k)}$ relative to the true labels $\mathbf{t}$ on U for $k = 1, 2$. More measurements include weighted accuracies relative to true labels on subsets of U where the two strategies agree or disagree. This way, assuming knowledge of the true labels on the entire set X, we evaluate the quality of pseudo-labels used in semi-supervised learning in each cycle, casting label propagation as an efficient surrogate of the learning process.

B. Results

We show results on CIFAR-10 with $b = 1000$ in this study. Following the experiments of [14], we first investigate the correlation of the ranks of unlabeled examples obtained by two acquisition functions. As shown in Figure 5(a), Uncertainty and jLP are not as heavily correlated compared to, for example, *CoreSet* and *Uncertainty* in Figure 5(b). The correlation between *jLP* and *CoreSet* is also quite low as shown in Figure 5(c).

It may of course be possible that two strategies with uncorrelated ranks still yield models of similar accuracy. To investigate this, we measure agreement as described above. Results are shown in Table IV. Uncertainty is used as a reference strategy, *i.e.* we train the model for a number of cycles using Uncertainty and then measure agreement and disagreement of another strategy to Uncertainty. After cycle 1, any two methods agree on around 80% of the pseudo-labels, while the remaining 20% have on average smaller weights compared to when the methods agree.

We reach the same conclusions from a similar experiment where we actually train the model rather than perform label propagation. Hence, although examples are ranked differently by different strategies, their effect on prediction, either by training or label propagation, is small.

In order to facilitate reproducibility, in this section we present all the detailed results in Table V and Table VI. We describe results obtained with the five methods presented before, namely Random, Uncertainty, CEAL, CoreSet and jLP. We evaluate them on CIFAR-10 with 10 and 100 labels per class (budget $b = 100$ and $b = 1000$ respectively), CIFAR-100 with $b = 1000$ in Table V. We present results obtained on MNIST with only 1 label per class ($b = 10$) and SVHN with $b = 100$ in Table VI.

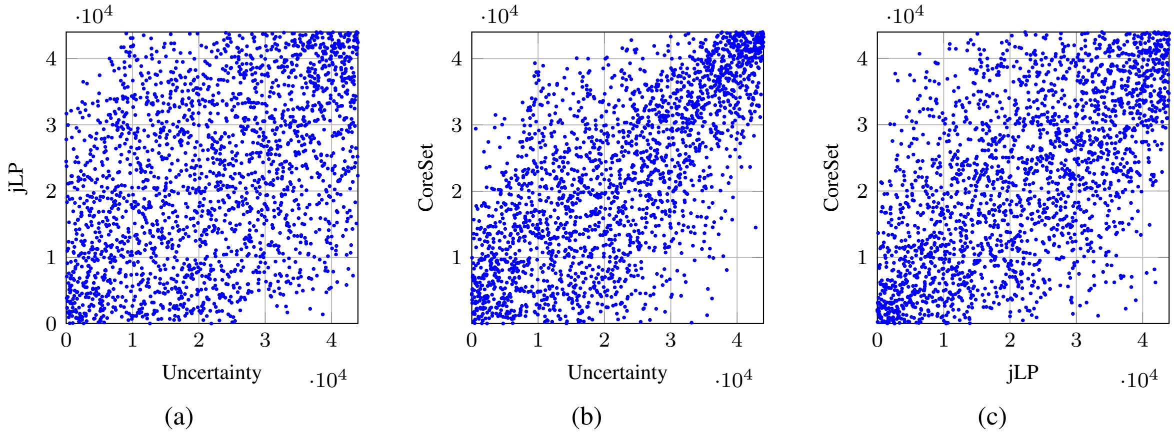

Fig. 5. Ranks of examples obtained by one acquisition strategy *vs.* the ranks of another on CIFAR-10 with $b = 1000$ after cycle 1. A random 5% subset of all examples is shown.

CYCLE	1					2				
MEASURE	%agree	accuracy (13)		avg weights		%agree	accuracy (13)		avg weights	
AGREE?		$=$	$\neq$	$=$	$\neq$		$=$	$\neq$	$=$	$\neq$
Random	79.98	79.97	38.39	0.32	0.17	86.98	88.07	39.77	0.46	0.28
CoreSet	80.58	79.52	44.57	0.27	0.16	87.32	87.94	43.80	0.45	0.29
jLP (ours)	80.24	80.03	48.79	0.27	0.15	86.96	88.12	45.55	0.43	0.27

TABLE IV

AGREEMENT RESULTS BETWEEN ACQUISITION STRATEGIES ON CIFAR-10 WITH $b = 1000$ AFTER CYCLES 1 AND 2. ALL STRATEGIES ARE COMPARED TO UNCERTAINTY AS REFERENCE, WHICH IS ALSO EMPLOYED IN THE PREVIOUS CYCLES. $\%agree$ IS PERCENTAGE OF PSEUDO-LABELS AGREEING TO THE REFERENCE. ACCURACY IS WEIGHTED ACCORDING TO (13) AND WEIGHTS ARE ACCORDING TO (10). MEASUREMENTS DENOTED BY $=$ ($\neq$) REFER TO THE SET OF PSEUDO-LABELS THAT AGREE (DISAGREE) WITH THE REFERENCE.

METHOD	CIFAR-10, $b = 100$			CIFAR-10, $b = 1000$		CIFAR-100, $b = 1000$		
PRE		✓	✓				✓	✓
SEMI			✓		✓			✓
CYCLE 0	100 LABELS			1K LABELS		1K LABELS		
Random	29.17±1.62	35.20±2.26	39.84±2.63	63.61±1.42	78.85±0.86	19.63±0.99	23.71±0.86	27.46±0.52
CYCLE 1	200 LABELS			2K LABELS		2K LABELS		
Random	36.66±1.08	41.76±1.32	**50.69±2.95**	75.09±0.51	83.49±0.81	**32.44±1.69**	**34.88±0.90**	**40.65±0.63**
Uncertainty	37.59±1.93	40.56±2.21	46.04±2.78	76.22±0.68	84.94±0.35	32.09±1.50	34.54±0.70	38.88±1.11
CoreSet	**39.23±1.17**	**43.04±0.92**	48.08±1.64	76.44±0.34	**84.98±0.19**	32.05±1.40	33.95±0.57	39.63±0.70
CEAL	38.92±2.00	39.74±1.72	–	**76.52±0.73**	–	31.59±0.93	33.78±0.39	–
jLP (ours)	38.86±1.36	42.07±0.74	48.66±2.64	75.74±0.39	84.62±0.47	32.16±1.98	33.48±0.52	40.30±1.53
CYCLE 2	300 LABELS			3K LABELS		3K LABELS		
Random	42.12±1.83	46.31±1.40	**58.72±4.04**	79.45±0.56	85.33±0.42	**42.45±0.90**	**42.37±0.53**	**47.42±0.53**
Uncertainty	**43.66±1.57**	44.02±1.73	52.04±2.46	**81.26±0.30**	**87.65±0.29**	40.43±0.63	41.04±0.27	46.30±1.12
CoreSet	43.01±2.14	47.00±2.57	50.85±4.23	81.11±0.61	87.21±0.31	41.32±0.70	40.47±0.38	46.74±1.00
CEAL	41.74±1.15	44.92±2.09	–	81.37±0.54	–	41.19±0.41	41.55±0.45	–
jLP (ours)	42.30±1.61	**47.99±1.17**	51.18±1.80	80.97±0.40	87.16±0.44	40.65±1.21	40.81±0.40	47.03±0.47
CYCLE 3	400 LABELS			4K LABELS		4K LABELS		
Random	45.91±1.63	50.63±0.59	**62.37±1.41**	82.33±0.21	86.66±0.21	**47.85±0.84**	**47.54±0.63**	50.38±0.25
Uncertainty	**47.89±1.78**	50.03±1.38	55.47±2.10	**84.47±0.49**	**89.32±0.24**	47.26±0.79	46.39±0.81	50.42±0.24
CoreSet	46.75±2.41	51.40±1.99	56.93±2.90	84.27±0.36	88.75±0.45	46.22±0.39	46.34±0.92	50.85±0.32
CEAL	45.55±2.39	49.73±1.82	–	84.05±0.44	–	46.34±0.44	46.67±0.38	–
jLP (ours)	45.49±1.71	**51.54±1.24**	56.67±2.58	83.82±0.02	88.85±0.38	46.52±0.99	45.94±0.44	**50.90±0.67**
CYCLE 4	500 LABELS			5K LABELS		5K LABELS		
Random	**50.94±1.75**	**55.31±1.28**	**64.35±1.37**	84.10±0.10	87.23±0.21	**51.43±0.56**	**51.40±0.47**	53.58±0.64
Uncertainty	49.73±2.29	53.17±1.52	60.71±2.77	**86.49±0.19**	**90.42±0.28**	50.83±0.31	49.90±0.82	52.20±0.50
CoreSet	50.11±1.40	54.17±0.40	62.94±2.41	86.39±0.36	90.33±0.13	50.48±0.84	49.54±0.95	**53.67±1.29**
CEAL	48.14±1.24	53.46±1.27	–	86.31±0.23	–	50.62±0.28	50.18±0.60	–
jLP (ours)	48.93±2.22	53.89±1.42	59.83±4.02	85.94±0.38	89.91±0.28	50.24±0.93	50.20±0.44	53.37±0.64

TABLE V

AVERAGE ACCURACY AND STANDARD DEVIATION FOR DIFFERENT LABEL BUDGET b AND CYCLE ON CIFAR-10 AND CIFAR-100. FOLLOWING ALGORITHM 1, WE SHOW THE EFFECT OF UNSUPERVISED PRE-TRAINING (PRE) AND SEMI-SUPERVISED LEARNING (SEMI) COMPARED TO THE STANDARD BASELINE.

Method	MNIST, $b = 10$		SVHN, $b = 100$		
PRE				✓	✓
SEMI		✓			✓
Cycle 0	10 Labels		100 Labels		
Random	26.83±4.15	70.06±12.87	18.00±2.47	23.83±4.63	19.01±5.61
Cycle 1	20 Labels		200 Labels		
Random	51.68±2.72	**90.89±4.84**	**45.95±1.97**	**53.87±5.43**	**81.25±4.82**
Uncertainty	53.18±5.88	76.12±11.07	31.63±8.75	51.52±2.36	37.84±21.00
CoreSet	**57.94±7.16**	86.59±10.98	35.39±7.16	52.49±5.76	51.80±10.62
CEAL	51.57±3.18	–	38.21±2.70	44.04±4.56	–
jLP (ours)	48.60±3.15	89.16±5.53	34.04±4.75	46.78±5.18	54.88±22.90
Cycle 2	30 Labels		300 Labels		
Random	**67.31±5.19**	**91.86±3.89**	**62.05±3.23**	64.88±4.93	**89.05±2.07**
Uncertainty	63.55±2.67	80.05±13.29	44.09±13.49	63.85±3.55	64.14±6.36
CoreSet	63.66±3.84	76.28±15.38	52.59±9.20	**67.23±3.01**	73.88±13.94
CEAL	56.62±7.05	–	51.53±5.93	63.58±2.80	–
jLP (ours)	62.71±2.82	80.23±4.11	44.74±17.50	58.43±9.82	66.68±13.91
Cycle 3	40 Labels		400 Labels		
Random	**71.05±1.66**	**93.38±3.99**	**70.28±1.67**	**72.50±2.05**	**90.69±0.73**
Uncertainty	67.87±3.26	93.03±4.88	66.21±3.68	70.90±2.48	56.60±5.69
CoreSet	69.79±3.36	86.93±7.62	63.53±6.34	71.79±3.58	75.88±6.95
CEAL	65.24±7.43	–	66.48±2.80	68.95±2.06	–
jLP (ours)	65.55±4.01	90.75±5.76	63.33±9.59	71.20±2.93	73.28±11.69
Cycle 4	50 Labels		500 Labels		
Random	**76.81±2.19**	**95.20±3.61**	**75.78±1.90**	**77.93±1.55**	**91.44±0.80**
Uncertainty	72.88±5.82	83.42±5.93	68.04±6.58	76.70±1.11	55.42±10.49
CoreSet	75.76±3.93	87.04±6.44	66.17±16.11	75.11±3.40	72.51±9.99
CEAL	72.02±7.96	–	66.14±14.42	74.48±1.98	–
jLP (ours)	73.36±4.43	92.37±5.38	60.12±20.06	75.33±1.44	72.98±12.01

TABLE VI

Average accuracy and standard deviation for different label budget b and cycle on MNIST and SVHN. Following Algorithm 1, we show the effect of unsupervised pre-training (PRE) and semi-supervised learning (SEMI) compared to the standard baseline.