Supplementary Material: Semi-Supervised Active Learning with Temporal Output Discrepancy

Siyu Huang[1] Tianyang Wang[2] Haoyi Xiong[1] Jun Huan[3] Dejing Dou[1]
[1]Baidu Research [2]Austin Peay State University [3]Styling AI
{huangsiyu,xionghaoyi,doudejing}@baidu.com wangt@apsu.edu lukehuan@shenshangtech.com

A. Proofs

Theorem 1 *With an appropriate setting of learning rate η,*

$$D_t^{\{1\}}(x) \leq \eta\sqrt{2L_t(x)}\|\nabla_w f(x;w_t)\|^2. \tag{1}$$

Proof. We apply one-step gradient descent to w_t then using first-order Taylor series,

$$\begin{aligned} D_t^{\{1\}}(x) &\overset{\text{def}}{=} \|f(x;w_{t+1}) - f(x;w_t)\| \\ &= \|f(x;w_t - \eta\nabla_{w_t}L_t(x)) - f(x;w_t)\| \\ &= \|f(x;w_t) - \eta\nabla_w f(x;w_t)^{\mathrm{T}}\nabla_w L_t(x) - f(x;w_t)\| \\ &= \|-\eta\nabla_w f(x;w_t)^{\mathrm{T}}\nabla_w L_t(x)\|. \end{aligned} \tag{2}$$

Recall that

$$\nabla_w L_t(x) = (y - f(x;w_t)) \cdot \nabla_w f(x;w_t). \tag{3}$$

By substituting Eq. 3 into Eq. 2,

$$\begin{aligned} D_t^{\{1\}}(x) &= \eta\|(y - f(x;w_t)) \cdot \nabla_w f(x;w_t)^{\mathrm{T}}\nabla_w f(x;w_t)\| \\ &\leq \eta\|(y - f(x;w_t))\| \cdot \|\nabla_w f(x;w_t)\|^2 \\ &= \eta\sqrt{2L_t(x)}\|\nabla_w f(x;w_t)\|^2. \end{aligned} \tag{4}$$

Corollary 1 *With an appropriate setting of learning rate η,*

$$D_t^{\{T\}}(x) \leq \sqrt{2}\eta\sum_{\tau=t}^{t+T-1}\left(\sqrt{L_\tau(x)}\|\nabla_w f(x;w_\tau)\|^2\right). \tag{5}$$

Proof.

$$\begin{aligned} D_t^{\{T\}}(x) &\overset{\text{def}}{=} \|f(x;w_{t+T}) - f(x;w_t)\| \\ &\leq \sum_{\tau=t}^{t+T-1}\|f(x;w_{\tau+1}) - f(x;w_\tau)\| \\ &\leq \sqrt{2}\eta\sum_{\tau=t}^{t+T-1}\left(\sqrt{L_\tau(x)}\|\nabla_w f(x;w_\tau)\|^2\right). \end{aligned} \tag{6}$$

Remark 1 *For a linear layer $\phi(x;W)$ with ReLU activation, the Lipschitz constant $\mathcal{L}(W) \leq \|x\|$.*

Proof.

$$\begin{aligned} &\|\phi(x;W+r) - \phi(x;W)\| \\ =\ &\|\max(0,(W+r)^{\mathrm{T}}x+b) - \max(0,W^{\mathrm{T}}x+b)\| \\ \leq\ &\|r^{\mathrm{T}}x\| \\ \leq\ &\|x\| \cdot \|r\|. \end{aligned} \tag{7}$$

Therefore, the Lipschitz constant $\mathcal{L}(W) \leq \|x\|$.

Corollary 2 *With appropriate settings of a learning rate η and a constant C,*

$$D_t^{\{T\}}(x) \leq \sqrt{2T}\eta C\sqrt{\sum_{\tau=t}^{t+T-1}L_\tau(x)}. \tag{8}$$

Proof. By substituting $\|\nabla_w f\|^2 \leq C$ into Corollary 1 then applying Cauchy–Schwarz inequality, we have

$$\begin{aligned} D_t^{\{T\}}(x) \leq\ &\sqrt{2}\eta C\sum_{\tau=t}^{t+T-1}\sqrt{L_\tau(x)} \\ \leq\ &\sqrt{2T}\eta C\sqrt{\sum_{\tau=t}^{t+T-1}L_\tau(x)}. \end{aligned} \tag{9}$$

B. Experimental Details

B.1. Image Classification

Datasets We evaluate the active learning methods on four common image classification datasets, including Cifar-10 [5], Cifar-100 [5], SVHN [6], and Caltech-101 [2]. CIFAR-10 and CIFAR-100 consist of 50,000 training images and 10,000 testing images with the size of 32×32. CIFAR-10 has 10 categories and CIFAR-100 has 100 categories. SVHN consists of 73,257 training images and 26,032 testing images with the size of 32×32. SVHN has 10 classes

Table 1. The summary of datasets used in the experiments. '#classes' and 'image size' are characters after pre-processing. 'image size' is the size of images used for training.

dataset	task	content	#classes	image size	train	val	test
Cifar-10	image classification	natural images	10	32×32	45,000	5,000	10,000
Cifar-100	image classification	natural images	100	32×32	45,000	5,000	10,000
SVHN	image classification	street view house numbers	10	32×32	65,931	7,326	26,032
Caltech-101	image classification	natural images	101	224×224	7,316	915	915
Cityscapes	semantic segmentation	driving video frames	19	688×688	2,675	300	500

Table 2. The summary of implementation details on each dataset. 'start' is the number of initially labeled samples and 'budget' is the number of newly annotated samples in each cycle. 'cycle' is the number of active learning cycles. α is the EMA decay rate and λ is the weight for unsupervised loss.

dataset	start	budget	cycle	optimizer	lr	momentum	decay	epochs	batch	α	λ
Cifar-10	10%	5%	7	SGD	0.1	0.9	5×10^{-4}	200	128	0.999	0.05
Cifar-100	10%	5%	7	SGD	0.1	0.9	5×10^{-4}	200	128	0.999	0.05
SVHN	10%	5%	7	SGD	0.1	0.9	5×10^{-4}	200	128	0.999	0.05
Caltech-101	10%	5%	7	SGD	0.01	0.9	5×10^{-4}	50	64	0.999	0.05
Cityscapes	10%	5%	7	Adam	5×10^{-4}	-	-	40	4	0.999	0.05

of digit numbers from '0' to '9'. For training on Cifar-10, Cifar-100, and SVHN, we randomly crop 32×32 images from the 36×36 zero-padded images. Caltech-101 consists of 9,146 images with the size of 300×200. Caltech-101 has 101 semantic categories as well as a background category that there are about 40 to 800 images per category. By following [9] we use 90% of the images for training and 10% of the images for testing. On Caltech-101, we resize the images to 256×256 and crop 224×224 images at the center. Random horizontal flip and normalization are applied to all the image classification datasets. We summarize the details of the datasets in Table 1.

Implementation details We employ ResNet-18 [3] as the image classification model. On all the image classification datasets, the labeling ratio of each active learning cycle is 10%, 15%, 20%, 25%, 30%, 35%, and 40%, respectively. In an cycle, The model is learned for 200 epochs using an SGD optimizer with a learning rate of 0.1, a momentum of 0.9, a weight decay of 5×10^{-4}, and a batch size of 128. After 80% of the training epochs, the learning rate is decreased to 0.01. We summarize the implementation details in Table 2.

B.2. Semantic Segmentation

Dataset We evaluate the active learning methods for semantic segmentation on the Cityscapes dataset [1]. Cityscapes is a large scale driving video dataset collected from urban street scenes of 50 cities. It consists of 2,975 training images and 500 testing images with the size of 2048×1024. By following [7], we convert the dataset from the original 30 classes into 19 classes. We crop 688×688 images from the original images for training. Random horizontal flip and normalization are applied to the images.

Implementation details we employ the 22-layer dilated residual network (DRN-D-22) [8] as the semantic segmentation model. The labeling ratio of each active learning cycle is 10%, 15%, 20%, 25%, 30%, 35%, and 40%, respectively. In an cycle, the model is learned for 40 epochs using an Adam optimizer [4] with a learning rate of 5×10^{-4} and a batch size of 4.

C. More Experimental Results

C.1. Study on Hyper-Parameters

The unsupervised learning plays a vital role in training the task model in active learning. Here we further investigate the hyper-parameter selection for our proposed unsupervised learning method. The hyper-parameters include the unsupervised loss weight λ and the EMA decay rate α. We conduct empirical studies using our full active learning pipeline on Cifar-100 dataset to investigate the performance variation with different λ and α. Fig. 1 shows the results on labeling budgets of 10%, 20%, 30%, 40%, respectively. In most of the cases, $\lambda = 0.05$ and $\alpha = 0.999$ achieve the best performance. Therefore, we adopt $\lambda = 0.05$ and $\alpha = 0.999$ for all the experiments in this paper, wherever EMA is involved.

C.2. Evaluating Active Data Selection Strategies

As a supplementary to Section 5.2 of the paper, we comprehensively compare the active data selection strategies by training the task model with and without the unsupervised loss, respectively. Fig. 2 shows that our method achieves superior performances on most of the datasets and settings

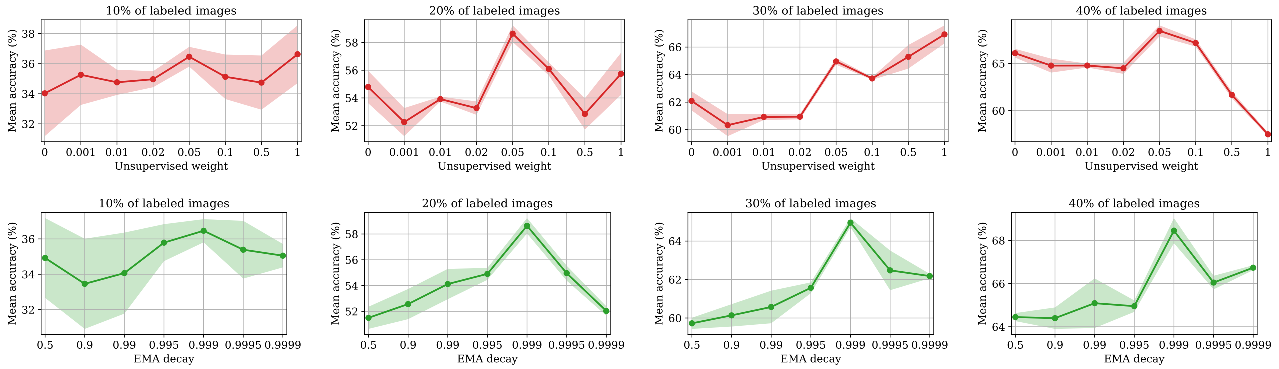

Figure 1. Empirical study on unsupervised loss weight λ and EMA decay rate α. The study is conducted on Cifar-100 with 10%, 20%, 30%, and 40% of labeled images, respectively. $\lambda = 0.05$ and $\alpha = 0.999$ achieve the best performances.

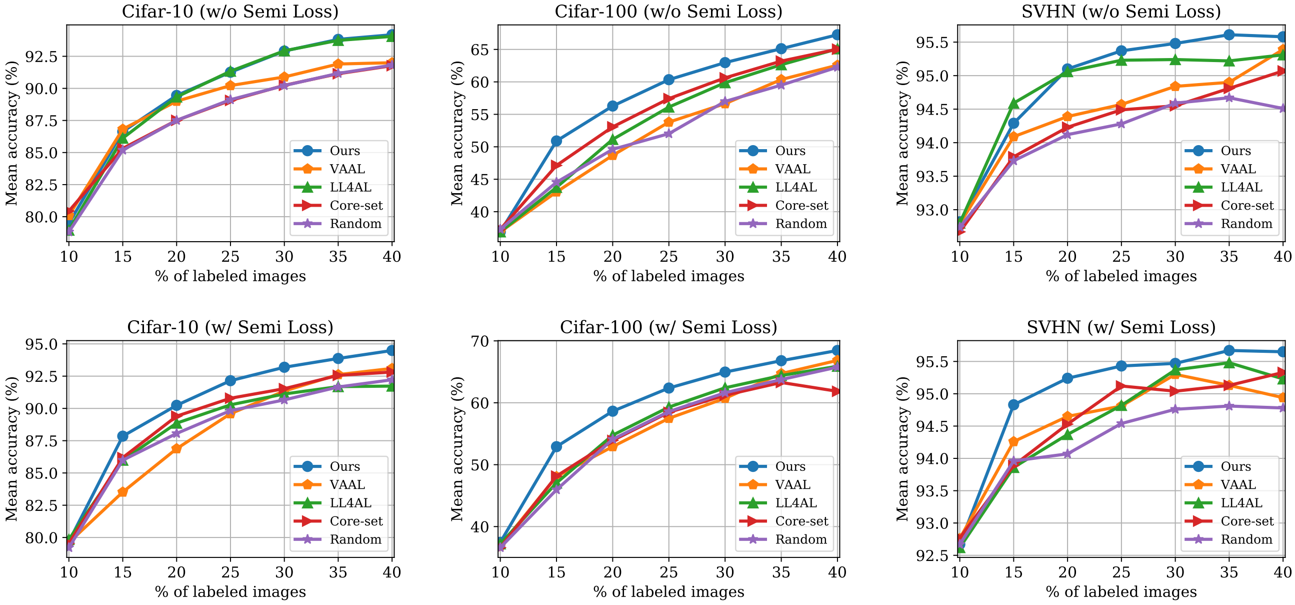

Figure 2. The performance of benchmark active learning methods trained *without* (top) and *with* (bottom) the unsupervised loss on three datasets.

(either with or without unsupervised loss), demonstrating its effectiveness in active data selection.

C.3. TOD with Different Gradient Descent Steps

Fig. 3 shows the loss estimation performances of TOD using different numbers of gradient descent (GD) steps. More GD steps may bring a better loss estimation performance especially when there are fewer sampling images.

References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2

[2] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE TPAMI*, 28(4):594–611, 2006. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 2

[5] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Technical Report*, 2009. 1

[6] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. 1

[7] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Varia-

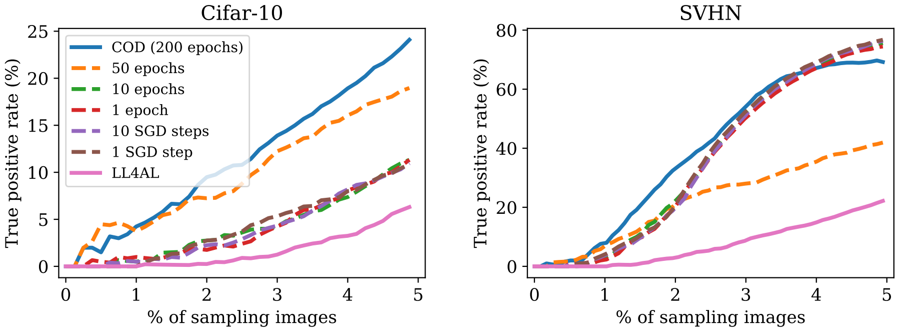

Figure 3. The effects of number of GD steps used in TOD.

tional adversarial active learning. In *ICCV*, pages 5972–5981, 2019. 2

[8] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *CVPR*, pages 472–480, 2017. 2

[9] Beichen Zhang, Liang Li, Shijie Yang, Shuhui Wang, Zheng-Jun Zha, and Qingming Huang. State-relabeling adversarial active learning. In *CVPR*, pages 8756–8765, 2020. 2