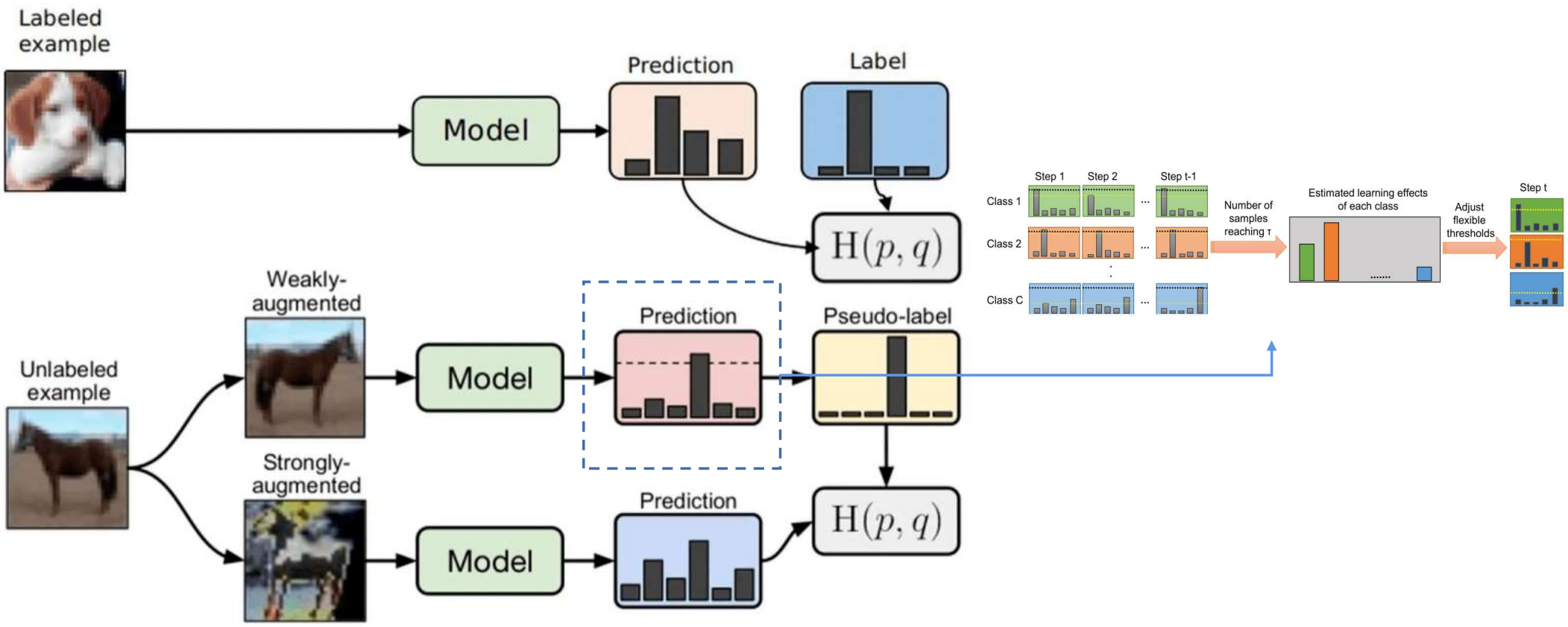
MarginMatch: Improving Semi-Supervised Learning with Pseudo-Margins

Tiberiu Sosea Cornelia Caragea
University of Illinois Chicago
tsosea2@uic.edu cornelia@uic.edu

CVPR 2023

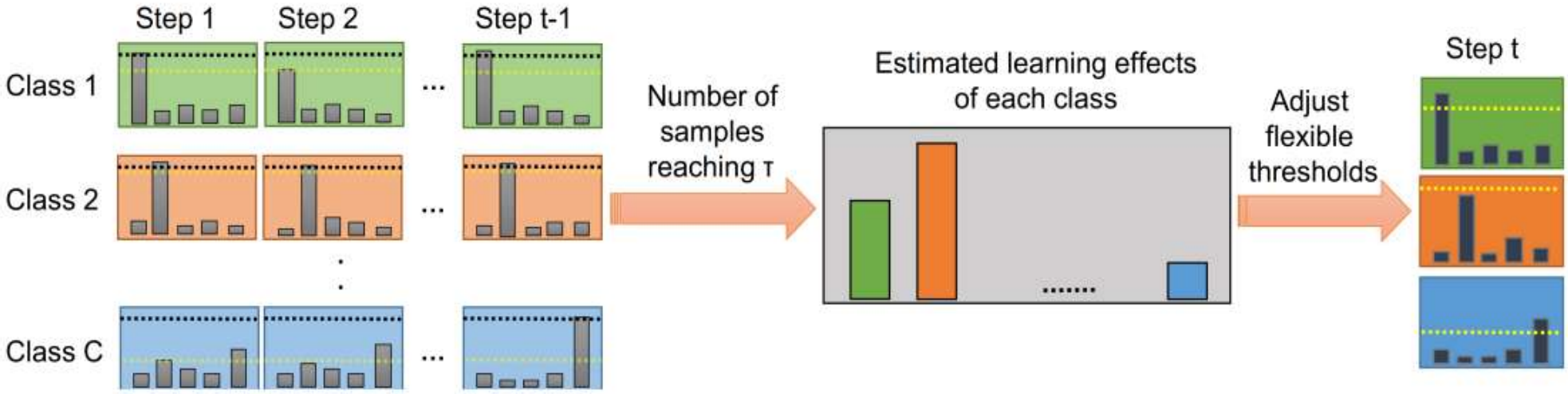
Background



FixMatch → FlexMatch

Background

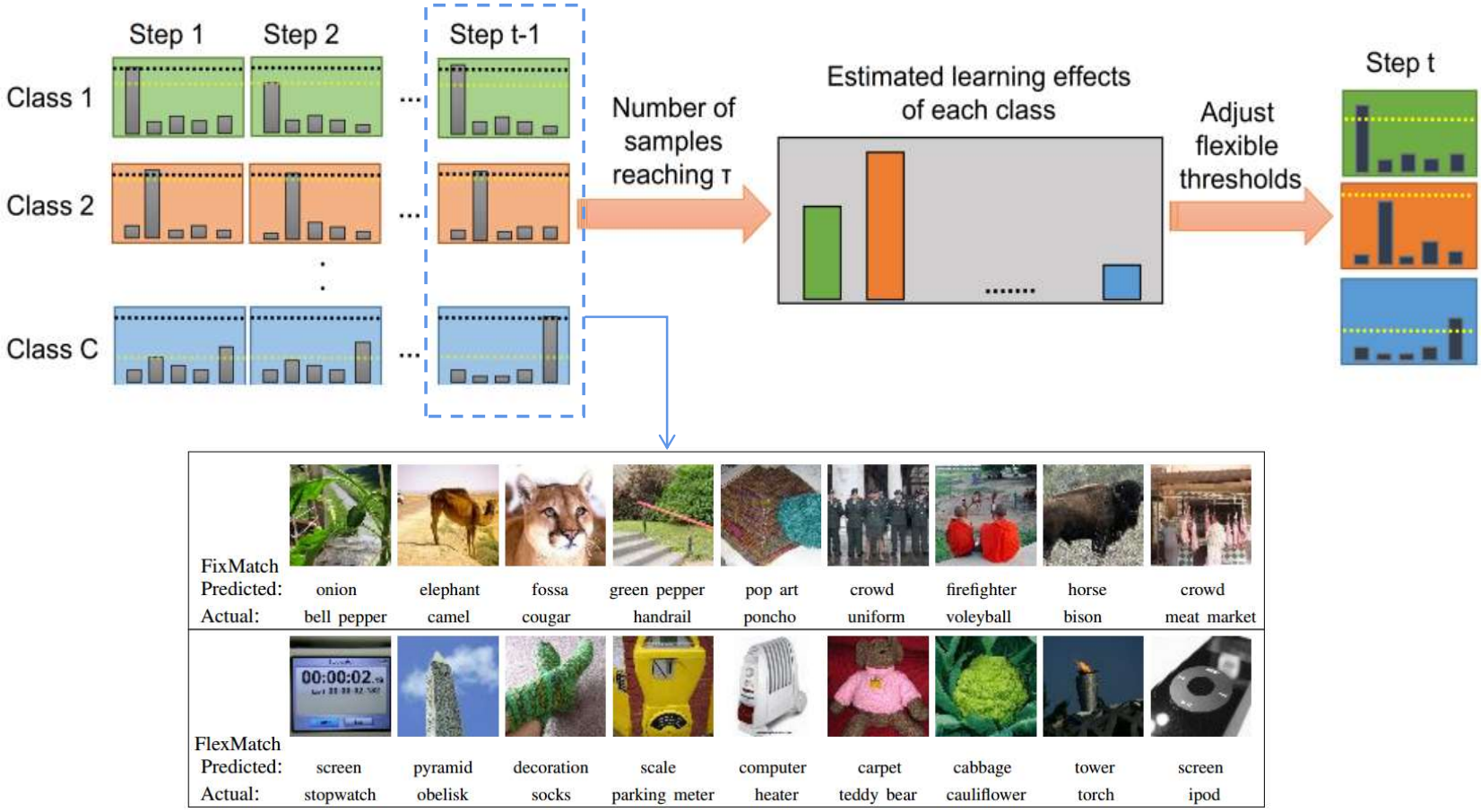
FlexMatch



$$\alpha_c = \sum_{i=1}^{n} \mathbb{1}(\max(p_\theta(y|\pi(\hat{x}_i))) > \tau)\mathbb{1}(\hat{p}_\theta(y|\pi(\hat{x}_i)) = c)$$

$$\mathcal{T}_c = \frac{\alpha_c}{\max_c(\alpha_c)} \times \tau$$

Background

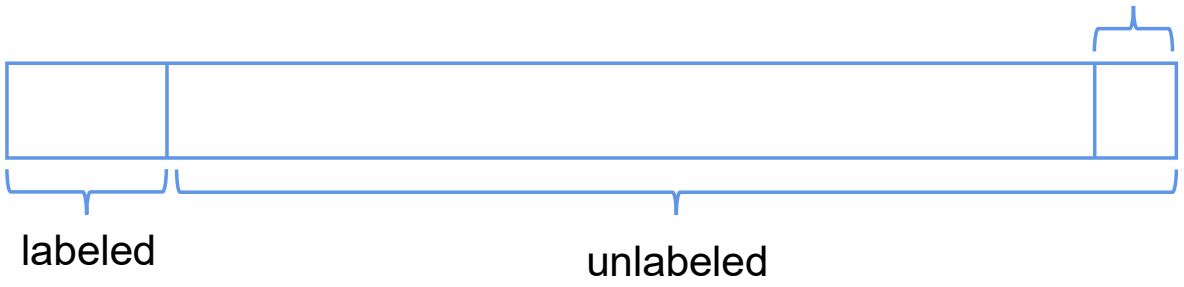

Step 1
Step 2
Step t-1
Class 1
Class 2
Class C
Number of samples reaching τ
Estimated learning effects of each class
Adjust flexible thresholds
Step t
FixMatch
Predicted: onion elephant fossa green pepper pop art crowd firefighter horse crowd
Actual: bell pepper camel cougar handrail poncho uniform volleyball bison meat market
FlexMatch
Predicted: screen pyramid decoration scale computer carpet cabbage tower screen
Actual: stopwatch obelisk socks parking meter heater teddy bear cauliflower torch ipod

Method

Pseudo-Margins

$$\mathrm{PM}_c^t(\hat{x}) = z_c - max_{c!=i}(z_i)$$

$$\mathrm{APM}_c^t(\hat{x}) = \frac{1}{t}\sum_{j=1}^{t}\mathrm{PM}_c^j(\hat{x})$$



For both labeled and unlabeled datasets: $y_i \in \{1,2,\ldots C\}$

For erroneous datasets: $y_i = C+1$

$$\mathcal{L}_u = \sum_{i=1}^{\nu B} \mathbb{1}(\mathrm{AM}^t_{\hat{p}_\theta(y|\pi(\hat{x}_i))}(\hat{x}_i) > \gamma^t) \times \mathbb{1}(\max(p_\theta(y|\pi(\hat{x}_i))) > \mathcal{T}^t_{\hat{p}_\theta(y|\pi(\hat{x}_i))}) \times H(\hat{p}_\theta(y|\pi(\hat{x}_i)), p_\theta(y|\Pi(\hat{x}_i)))$$

$$\mathcal{L}_e = \sum_{i=1}^{B} H(C+1, p_\theta(y|\Pi(\tilde{x}_i)))$$

Method

Algorithm 1 MarginMatch

Require: Labeled data L; unlabeled data U; erroneous examples E; maximum number of iterations T; number of classes $C+1$ (C original classes plus one virtual class of erroneous examples); θ model; π weak augmentations; Π strong augmentations.

Initialize the Average Pseudo-Margin (APM) threshold γ^1 at the first iteration to a small value (e.g., $\gamma^1 = -\infty$).
for t = 1 to T **do**
  Estimate learning status α_c (using Eq. 2) and calculate the class-wise flexible thresholds $\mathcal{T}_c^t$ (using Eq. 3) for each class c.
  while U not exhausted **do**
    Labeled batch $L_b = \{(x_1, y_1), ..., (x_B, y_B)\}$, unlabeled batch $U_b = \{\hat{x}_1, ..., \hat{x}_{\nu B}\}$, erroneous (or mislabeled) batch $E_b = \{(\tilde{x}_1, C+1), ..., (\tilde{x}_B, C+1)\}$
    for $x \in U_b \cup E_b$ **do**
      Compute logits z_c for each class c after applying weak augmentations when $x \in U_b$ and strong augmentations when $x \in E_b$.
      Calculate pseudo-margin PM_c^t (using Eq. 5) and update Average PM_c^t (using Eq. 6) for each $c = 1$ to $C+1$.
    end for
    Minimize $\mathcal{L} = \mathcal{L}_s + \lambda(\mathcal{L}_u + \mathcal{L}_e)$
      $\mathcal{L}_s = \frac{1}{B}\sum_{i=1}^{B} H(y_i, p_\theta(y|\pi(x_i)))$
      $\mathcal{L}_u = \sum_{i=1}^{\nu B} \mathbb{1}(\mathrm{AM}^t_{\hat{p}_\theta(y|\pi(\hat{x}_i))}(\hat{x}_i) > \gamma^t) \times \mathbb{1}(\max(p_\theta(y|\pi(\hat{x}_i))) > \mathcal{T}^t_{\hat{p}_\theta(y|\pi(\hat{x}_i))}) \times H(\hat{p}_\theta(y|\pi(\hat{x}_i)), p_\theta(y|\Pi(\hat{x}_i)))$
      $\mathcal{L}_e = \sum_{i=1}^{B} H(C+1, p_\theta(y|\Pi(\tilde{x}_i)))$
  end while
  Update γ^{t+1} as the 95^{th} percentile erroneous sample APM^t_{C+1}.
end for

$$\mathcal{L}_u = \sum_{i=1}^{\nu B} \mathbb{1}(\mathrm{AM}^t_{\hat{p}_\theta(y|\pi(\hat{x}_i))}(\hat{x}_i) > \gamma^t) \times \mathbb{1}(\max(p_\theta(y|\pi(\hat{x}_i))) > \mathcal{T}^t_{\hat{p}_\theta(y|\pi(\hat{x}_i))}) \times H(\hat{p}_\theta(y|\pi(\hat{x}_i)), p_\theta(y|\Pi(\hat{x}_i)))$$

$$\mathrm{APM}_c^t(\hat{x}) = \mathrm{PM}_c^t(\hat{x}) * \frac{\delta}{1+t} + \mathrm{APM}_c^{t-1}(\hat{x}) * (1 - \frac{\delta}{1+t})$$

Experiment

Dataset	CIFAR-10			CIFAR-100			SVHN			STL-10		
#Labels/Class	4	25	400	4	25	100	4	25	100	4	25	100
Pseuso-Labeling	$74.61_{0.26}$	$46.49_{2.20}$	$15.08_{0.19}$	$87.45_{0.85}$	$57.74_{0.28}$	$36.55_{0.24}$	$64.61_{5.60}$	$25.21_{2.03}$	$9.40_{0.32}$	$74.68_{0.99}$	$55.45_{2.43}$	$32.64_{0.71}$
UDA	$10.79_{3.75}$	$5.32_{0.06}$	$4.41_{0.07}$	$48.95_{1.59}$	$29.43_{0.21}$	$23.87_{0.23}$	$5.34_{4.27}$	$4.26_{0.39}$	$1.95_{0.01}$	$37.82_{8.44}$	$9.81_{1.15}$	$6.81_{0.17}$
MixMatch	$45.24_{2.15}$	$12.76_{1.14}$	$7.13_{0.34}$	$62.15_{2.17}$	$41.51_{1.19}$	$28.16_{0.24}$	$46.18_{1.78}$	$3.98_{0.17}$	$3.5_{0.13}$	$34.15_{1.54}$	$8.95_{0.32}$	$10.41_{0.73}$
ReMixMatch	$5.27_{0.19}$	$4.85_{0.13}$	$4.04_{0.12}$	$47.15_{0.76}$	$27.14_{0.23}$	$23.78_{0.12}$	$4.23_{0.31}$	$3.18_{0.04}$	$1.94_{0.06}$	$31.51_{0.75}$	$8.54_{0.48}$	$6.19_{0.24}$
FixMatch	$7.8_{0.28}$	$4.91_{0.05}$	$4.25_{0.08}$	$48.21_{0.82}$	$29.45_{0.16}$	$22.89_{0.12}$	$3.97_{1.18}$	**$3.13_{1.03}$**	$1.97_{0.03}$	$38.43_{4.14}$	$10.45_{1.04}$	$6.43_{0.33}$
FlexMatch	$5.04_{0.06}$	$5.04_{0.09}$	$4.19_{0.01}$	$39.99_{1.62}$	$26.96_{0.08}$	$22.44_{0.15}$	$8.19_{3.20}$	$7.78_{2.55}$	$6.72_{0.30}$	$29.15_{1.32}$	$8.23_{0.13}$	$5.77_{0.12}$
MarginMatch	**$4.91_{0.07}$**	**$4.73_{0.12}$**	**$3.98_{0.02}$**	**$36.97_{1.32}$**	**$23.71_{0.13}$**	**$21.39_{0.12}$**	**$3.75_{1.20}$**	$3.14_{1.17}$	**$1.93_{0.01}$**	**$25.37_{3.58}$**	**$7.31_{0.35}$**	**$5.52_{0.15}$**

Table 1. Test error rates on CIFAR-10, CIFAR-100, SVHN, and STL-10 datasets. Best results are shown in **blue**.

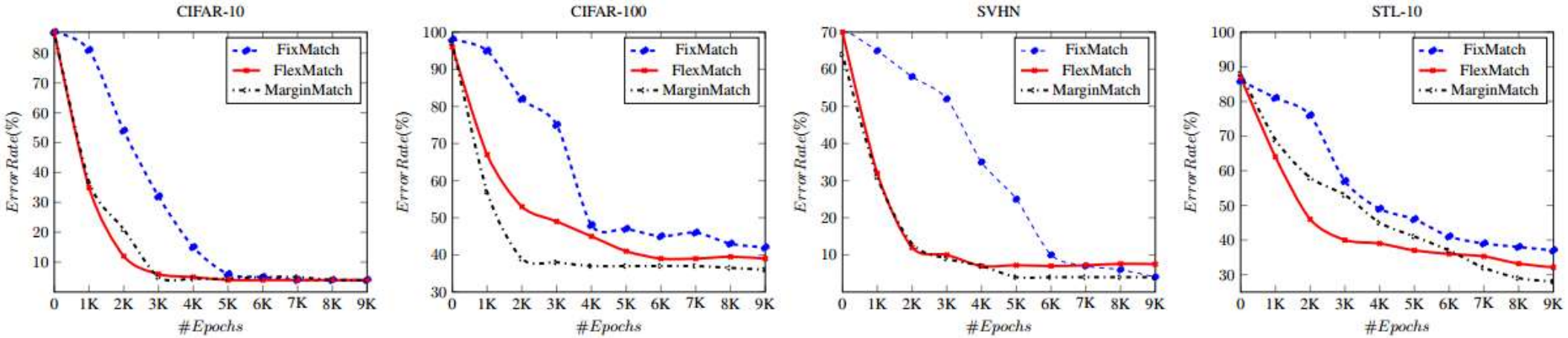


Figure 2. Convergence speed of MarginMatch against FixMatch and FlexMatch with 4 labels per class.

Experiment

Dataset	ImageNet		WebVision	
	TOP-1	TOP-5	TOP-1	TOP-5
Supervised	48.39	25.49	49.58	26.78
FixMatch	43.66	21.80	44.76	22.65
FlexMatch	42.02	19.49	43.87	22.07
MarginMatch	**41.05**	**18.28**	**43.08**	**21.13**

Table 2. Test error rates on the ImageNet and WebVision datasets. Best results are shown in **blue**.

δ	0.95	0.99	0.995	0.997	0.999	1
ERR RATE	38.13	38.05	37.92	**37.91**	39.12	39.72

Table 3. Error rates obtained on CIFAR-100 with four examples per class and various smoothing values δ. Best result is in **blue**.

Dataset	CIFAR-10			CIFAR-100			SVHN			STL-10		
#Labels/Class	4	25	400	4	25	100	4	25	100	4	25	100
Avg Confidence	$23.87_{2.73}$	$14.21_{1.37}$	$7.54_{0.78}$	$41.23_{2.15}$	$31.49_{1.48}$	$24.11_{2.36}$	$8.99_{4.27}$	$6.54_{0.39}$	$4.73_{0.01}$	$31.67_{8.44}$	$14.87_{1.15}$	$7.59_{0.17}$
Avg Entropy	$8.58_{0.41}$	$6.18_{0.15}$	$5.85_{0.12}$	$45.10_{0.91}$	$26.02_{1.11}$	$22.13_{0.25}$	$15.69_{1.25}$	$12.74_{0.78}$	$9.33_{0.05}$	$29.54_{3.51}$	$10.63_{1.35}$	$10.84_{0.47}$
Avg Margin	$7.25_{0.29}$	$5.38_{0.76}$	$4.73_{0.09}$	$39.72_{1.52}$	$25.21_{0.52}$	$23.18_{0.17}$	$18.45_{1.36}$	$11.29_{0.93}$	$8.40_{0.04}$	$28.45_{4.28}$	$9.34_{1.34}$	$7.59_{0.21}$
EMA Confidence	$4.91_{0.45}$	$4.74_{0.09}$	$3.99_{0.06}$	$38.67_{0.74}$	$25.61_{0.12}$	$21.48_{0.17}$	$3.84_{0.23}$	$3.25_{0.03}$	$1.93_{0.09}$	$25.9_{0.81}$	$7.6_{0.42}$	$5.74_{0.57}$
EMA Entropy	$6.4_{0.43}$	$8.34_{0.12}$	$4.21_{0.09}$	$41.63_{0.76}$	$36.84_{0.13}$	$22.52_{0.07}$	$3.81_{1.26}$	$3.17_{0.87}$	$2.14_{0.04}$	$27.21_{4.05}$	$8.28_{1.01}$	$6.79_{0.27}$
EMA Margin	**$4.91_{0.07}$**	**$4.73_{0.12}$**	**$3.98_{0.02}$**	**$36.97_{1.32}$**	**$23.71_{0.13}$**	**$21.39_{0.12}$**	**$3.75_{1.20}$**	**$3.14_{1.17}$**	**$1.93_{0.01}$**	**$25.37_{3.58}$**	**$7.31_{0.35}$**	**$5.52_{0.15}$**

Table 4. Test error rates comparing pseudo-margin with confidence and entropy. Best results are shown in **blue**.

Experiment

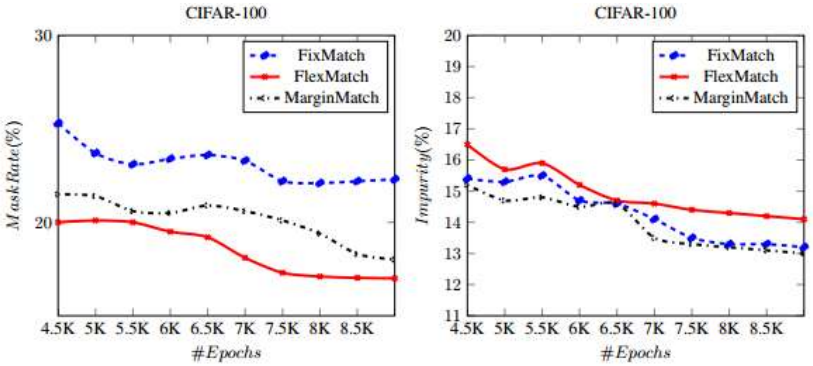


Figure 3. Mask rate and impurity on CIFAR-100 with 4 labeled examples per class.

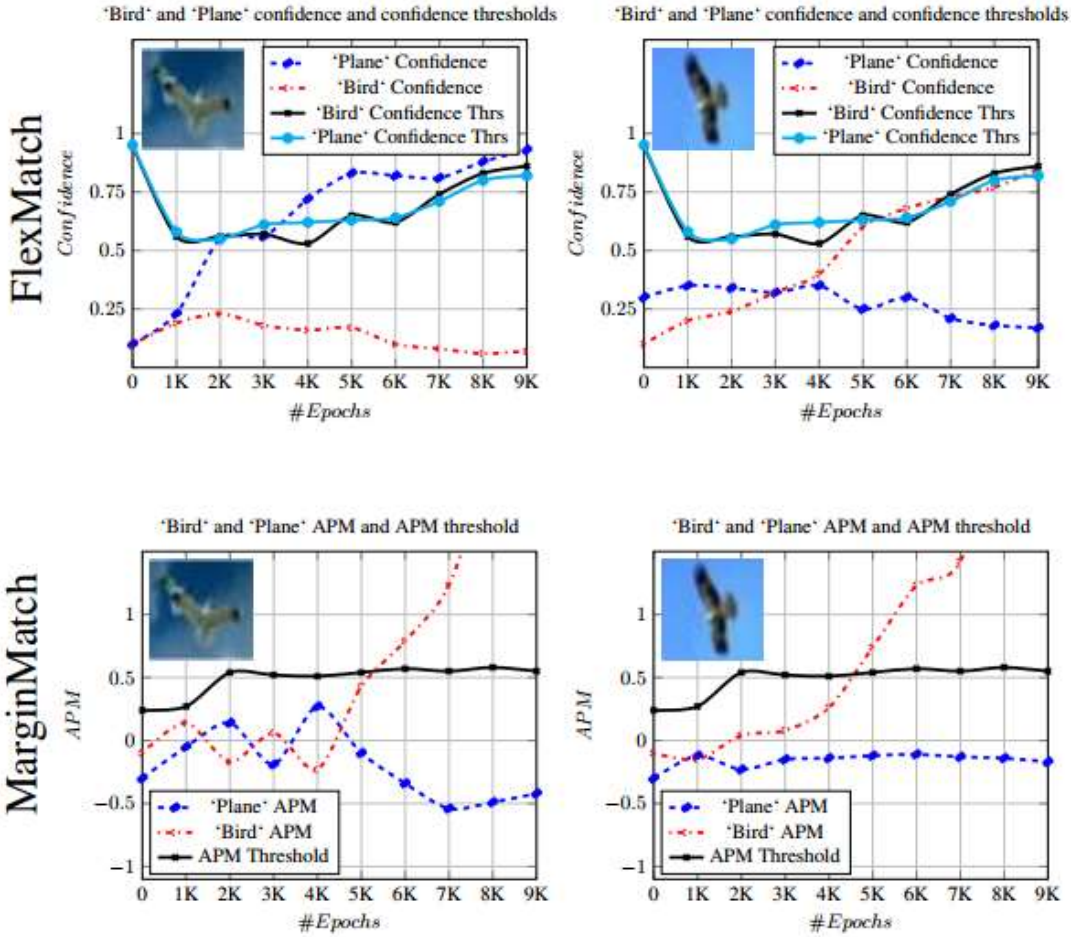


Figure 4. Confidence thresholding vs. APM Thresholding on two images from the CIFAR-10 dataset.

Thanks