Supplementary Materials: Better Fine-tuning via Instance Weighting for Text Classification

Zhi Wang, Wei Bi, Yan Wang, Xiaojiang Liu

Appendix A: Convergence rate of IW-Fit

We study the nonconvex *finite-sum* problems of the form

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} f(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{\theta}), \tag{1}$$

where both f and $f_i(i \in [n])$ may be nonconvex. We denote the class of such finite-sum Lipschitz smooth functions by $\mathcal{F}_n$. We optimize functions in $\mathcal{F}_n$ of the Importance Weighting based Fine-tuning (IW-Fit) algorithm.

Definition 1. *For $f \in \mathcal{F}_n$, IW-Fit takes an index $i \in [n]$ and a point $x \in \mathbb{R}^d$, and returns the pair $(f_i(\boldsymbol{\theta}), \nabla f_i(\boldsymbol{\theta}))$.*

Definition 2. *We say $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth if there is a constant L such that*

$$||\nabla f(\boldsymbol{\vartheta}) - \nabla f(\boldsymbol{\theta})|| \leq L||\boldsymbol{\vartheta} - \boldsymbol{\theta}||, \quad \forall \boldsymbol{\vartheta}, \boldsymbol{\theta} \in \mathbb{R}^d. \tag{2}$$

Definition 3. *A point $\boldsymbol{\theta}$ is called ϵ-accurate if $||\nabla f(\boldsymbol{\theta})||^2 \leq \epsilon$. A stochastic iterative algorithm is said to achieve ϵ-accuracy in t iterations if $\mathbb{E}[||\nabla f(\boldsymbol{\theta}^t)||^2 \leq \epsilon$, where the expectation is over the stochasticity of the algorithm.*

Definition 4. *We say $f \in \mathcal{F}_n$ has σ-bounded gradients if $||\nabla f_i(\boldsymbol{\theta})|| \leq \sigma$ for all $i \in [n]$ and $\boldsymbol{\theta} \in \mathbb{R}^d$.*

Let α_t denote the learning rate at iteration t, and w_{i_t} be the instance weight assigned to sample i by IW-Fit. By stochastic gradient descent (SGD), we have

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \alpha_t w_{i_t} \nabla f_{i_t}(\boldsymbol{\theta}^t), i \in [n]. \tag{3}$$

Definition 5. *We say the positive instance weight w in IW-Fit is bounded if there exist constants $\underline{w}$ and $\overline{w}$ such that $\underline{w} \leq w_i \leq \overline{w}$ for all $i \in [n]$.*

Theorem 1. *Suppose the loss function of IW-Fit $f \in \mathcal{F}_n$, where $\mathcal{F}_n$ is the class of finite-sum Lipschitz smooth functions, has σ-bounded gradients, and the instance weight w is clipped to be bounded by $[\underline{w}, \overline{w}]$. Let $\alpha_t = \alpha = c/\sqrt{T}$ where $c = \sqrt{\frac{2(f(\boldsymbol{\theta}^0) - f(\boldsymbol{\theta}^*))}{L\sigma^2 \underline{w}\overline{w}}}$, and $\boldsymbol{\theta}^*$ is an optimal solution. Then, the iterates of IW-Fit satisfy:*

$$\min_{0\leq t\leq T-1} \mathbb{E}[||\nabla f(\boldsymbol{\theta}^t)||^2] \leq \sqrt{\frac{2(f(\boldsymbol{\theta}^0) - f(\boldsymbol{\theta}^*))L\overline{w}}{T\underline{w}}}\sigma.$$

Proof. According to the Lipschitz continuity of ∇f, the iterates of IW-Fit satisfy the following bound:

$$\mathbb{E}[f(\boldsymbol{\theta}^{t+1})] \leq \mathbb{E}[f(\boldsymbol{\theta}^t) + \langle \nabla f(\boldsymbol{\theta}^t), \boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t \rangle + \frac{L}{2}||\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t||^2]. \tag{4}$$

After substituting (3) into (4), we have:

$$\begin{aligned}\mathbb{E}[f(\boldsymbol{\theta}^{t+1})] &\leq \mathbb{E}[f(\boldsymbol{\theta}^t)] - \alpha_t w_t \mathbb{E}[||\nabla f(\boldsymbol{\theta}^t)||^2] + \frac{L\alpha_t^2 w_t^2}{2}\mathbb{E}[||\nabla f_{i_t}(\boldsymbol{\theta}^t)||^2] \\ &\leq \mathbb{E}[f(\boldsymbol{\theta}^t)] - \alpha_t w_t \mathbb{E}[||\nabla f(\boldsymbol{\theta}^t)||^2] + \frac{L\alpha_t^2 w_t^2}{2}\sigma^2. \end{aligned} \tag{5}$$

The first inequality follows from the unbiasedness of the stochastic gradient $\mathbb{E}_{i_t}[\nabla f_{i_t}(\boldsymbol{\theta}^t)] = \nabla f(\boldsymbol{\theta}^t)$. The second inequality uses the assumption on gradient boundedness in Definition 4. Re-arranging (5) we obtain

$$\mathbb{E}[||\nabla f(\boldsymbol{\theta}^t)||^2] \leq \frac{1}{\alpha_t w_t}\mathbb{E}[f(\boldsymbol{\theta}^t) - f(\boldsymbol{\theta}^{t+1})] + \frac{L\alpha_t w_t}{2}\sigma^2. \tag{6}$$

Summing (6) from $t = 0$ to $T - 1$ and using that α_t is fixed α we obtain

$$\begin{aligned}\min_t \mathbb{E}[||\nabla f(\boldsymbol{\theta}^t)||^2] &\leq \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[||\nabla f(\boldsymbol{\theta}^t)||^2] \\ &\leq \frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{\alpha w_t}\mathbb{E}[f(\boldsymbol{\theta}^t) - f(\boldsymbol{\theta}^{t+1})] + \frac{1}{T}\sum_{t=0}^{T-1}\frac{L\alpha w_t}{2}\sigma^2 \\ &\leq \frac{1}{T\alpha\underline{w}}\left(f(\boldsymbol{\theta}^0 - f(\boldsymbol{\theta}^T)\right) + \frac{L\alpha\overline{w}}{2}\sigma^2 \\ &\leq \frac{1}{T\alpha\underline{w}}\left(f(\boldsymbol{\theta}^0 - f(\boldsymbol{\theta}^*)\right) + \frac{L\alpha\overline{w}}{2}\sigma^2 \\ &\leq \frac{1}{\sqrt{T}}\left(\frac{1}{c\underline{w}}(f(\boldsymbol{\theta}^0) - f(\boldsymbol{\theta}^*)) + \frac{Lc\overline{w}}{2}\sigma^2\right). \end{aligned} \tag{7}$$

The first step holds because the minimum is less than the average. The second step is obtained from (6). The third step follows from the assumption on instance weight boundedness in Definition 5. The fourth step is obtained from the fact that $f(\boldsymbol{\theta}^*) \leq f(\boldsymbol{\theta}^T)$. The final inequality follows upon using $\alpha = c/\sqrt{T}$. By setting

$$c = \sqrt{\frac{2(f(\boldsymbol{\theta}^0) - f(\boldsymbol{\theta}^*))}{L\sigma^2\underline{w}\overline{w}}} \tag{8}$$

in the above inequality, we get the desired result. □

Appendix B: Supplementary experimental results from *Amazon* to *Yelp*

This appendix shows the full results of the robustness study part in the first set of experiments from *Amazon* to *Yelp*, which are listed in Tables 1, 2, and 3.

Moreover, to demonstrate the performance stability of IW-Fit, we conduct another two sets of experiments based on randomly sampling from the source/target datasets, and the results are listed in Tables 4, 5, 6, and 7. It can be observed that, under various settings, IW-Fit can stably improve the classification accuracy on the fine-tuning stage.

Appendix C: Supplementary experimental results from *Yelp-2015* to *Yelp-2016*/*Yelp-2017*

This appendix shows the full results of the robustness study part in the first set of experiments from *Yelp-2015* to *Yelp-2016*/*Yelp-2017*, which are listed in Tables 8, 9, and 10.

For the other two sets of experiments, testing accuracies on *Yelp-2016* are listed in Tables 11, 12, 13 and 14; and testing accuracies on *Yelp-2017* are listed in Tables 15, 16, 17 and 18. The observations are consistent with those on the task *Amazon* to *Yelp*.

Fixed parameters	encoder	embedding	None
Uniform	47.98 ± 0.04	51.00 ± 0.05	52.46 ± 0.06
Gradient	$\mathbf{48.23 \pm 0.03}$	51.16 ± 0.08	52.80 ± 0.09
Variance	47.98 ± 0.03	51.25 ± 0.08	53.20 ± 0.07
Hard-mixing	$\mathbf{48.23 \pm 0.03}$	51.06 ± 0.10	53.44 ± 0.08
Soft-mixing	$\mathbf{48.23 \pm 0.03}$	$\mathbf{51.53 \pm 0.10}$	$\mathbf{53.54 \pm 0.07}$

Table 1: Testing accuracies by varying fine-tuning parameters from *Amazon* to *Yelp* (1st set).

Percentage of target set	25%	50%	75%
TgtOnly	40.92 ± 0.10	45.32 ± 0.10	48.68 ± 0.10
SrcOnly	46.30 ± 0.10	46.30 ± 0.09	46.30 ± 0.10
All	49.68 ± 0.09	50.23 ± 0.10	50.80 ± 0.11
Uniform	49.72 ± 0.09	50.90 ± 0.10	51.68 ± 0.10
Gradient	49.72 ± 0.09	51.04 ± 0.11	51.88 ± 0.11
Variance	$\mathbf{50.46 \pm 0.09}$	51.36 ± 0.09	51.80 ± 0.10
Hard-mixing	50.24 ± 0.08	51.38 ± 0.10	51.78 ± 0.10
Soft-mixing	$\mathbf{50.46 \pm 0.09}$	$\mathbf{51.50 \pm 0.09}$	$\mathbf{52.36 \pm 0.10}$

Table 2: Testing accuracies by varying the target fine-tuning data size from *Amazon* to *Yelp* (1st set).

Percentage of source set	25%	50%	75%
TgtOnly	50.22 ± 0.11	50.22 ± 0.11	50.22 ± 0.11
SrcOnly	42.70 ± 0.10	47.86 ± 0.10	49.25 ± 0.11
All	50.60 ± 0.09	50.90 ± 0.08	51.30 ± 0.09
Uniform	50.30 ± 0.09	51.15 ± 0.08	52.20 ± 0.07
Gradient	50.30 ± 0.10	51.18 ± 0.08	52.20 ± 0.07
Variance	$\mathbf{51.20 \pm 0.09}$	51.60 ± 0.07	52.63 ± 0.11
Hard-mixing	50.90 ± 0.08	51.55 ± 0.05	52.48 ± 0.10
Soft-mixing	$\mathbf{51.20 \pm 0.09}$	$\mathbf{51.78 \pm 0.08}$	$\mathbf{52.70 \pm 0.10}$

Table 3: Testing accuracies of varying the source pre-training data size from *Amazon* to *Yelp* (1st set).

Method		*Amazon* to *Yelp* (2nd set)	*Amazon* to *Yelp* (3rd set)
Baselines	TgtOnly	51.10 ± 0.11	50.67 ± 0.09
	SrcOnly	47.74 ± 0.10	48.04 ± 0.11
	All	52.37 ± 0.08	51.16 ± 0.10
Fine-tined	Uniform	52.60 ± 0.04	51.98 ± 0.06
	Gradient	52.63 ± 0.04	52.37 ± 0.10
	Variance	53.00 ± 0.03	52.22 ± 0.09
	Hard-mixing	53.13 ± 0.04	52.31 ± 0.10
	Soft-mixing	$\mathbf{53.40 \pm 0.07}$	$\mathbf{52.63 \pm 0.10}$

Table 4: Testing accuracies of transfer learning from *Amazon* to *Yelp*.

Fixed parameters	*Amazon* to *Yelp* (2nd set)			*Amazon* to *Yelp* (3rd set)		
	encoder	embedding	None	encoder	embedding	None
Uniform	50.25 ± 0.08	52.85 ± 0.10	52.60 ± 0.04	48.03 ± 0.03	51.80 ± 0.10	51.98 ± 0.06
Gradient	50.38 ± 0.10	$\mathbf{53.43} \pm 0.11$	52.63 ± 0.04	48.06 ± 0.03	52.11 ± 0.09	52.37 ± 0.10
Variance	50.33 ± 0.06	52.90 ± 0.09	53.00 ± 0.03	48.08 ± 0.03	51.94 ± 0.10	52.22 ± 0.09
Hard-mixing	50.40 ± 0.07	52.98 ± 0.10	53.13 ± 0.04	48.16 ± 0.08	52.00 ± 0.10	52.31 ± 0.10
Soft-mixing	$\mathbf{50.63 \pm 0.06}$	$\mathbf{53.43 \pm 0.11}$	$\mathbf{53.40 \pm 0.07}$	$\mathbf{48.19 \pm 0.04}$	$\mathbf{52.29 \pm 0.10}$	$\mathbf{52.63 \pm 0.10}$

Table 5: Testing accuracies of varying fine-tuning parameters from *Amazon* to *Yelp*.

Percentage of target set	*Amazon* to *Yelp* (2nd set)			*Amazon* to *Yelp* (3rd set)		
	25%	50%	75%	25%	50%	75%
TgtOnly	41.83 ± 0.09	47.37 ± 0.08	47.90 ± 0.09	38.50 ± 0.11	46.05 ± 0.11	47.95 ± 0.10
SrcOnly	47.74 ± 0.09	47.74 ± 0.09	47.74 ± 0.09	48.04 ± 0.11	48.04 ± 0.11	48.04 ± 0.11
All	48.00 ± 0.07	49.50 ± 0.08	51.03 ± 0.08	48.32 ± 0.09	49.62 ± 0.08	49.88 ± 0.08
Uniform	51.53 ± 0.03	51.67 ± 0.04	52.43 ± 0.05	49.31 ± 0.09	50.55 ± 0.06	51.70 ± 0.10
Gradient	51.53 ± 0.03	51.67 ± 0.04	52.97 ± 0.06	49.84 ± 0.10	50.78 ± 0.10	51.89 ± 0.06
Variance	52.23 ± 0.04	51.80 ± 0.03	52.80 ± 0.07	49.47 ± 0.09	50.95 ± 0.09	52.25 ± 0.07
Hard-mixing	52.00 ± 0.06	51.73 ± 0.03	52.67 ± 0.08	49.65 ± 0.10	51.08 ± 0.10	52.33 ± 0.11
Soft-mixing	$\mathbf{52.23 \pm 0.04}$	$\mathbf{51.90 \pm 0.06}$	$\mathbf{53.13 \pm 0.10}$	$\mathbf{49.91 \pm 0.07}$	$\mathbf{51.16 \pm 0.09}$	$\mathbf{52.34 \pm 0.07}$

Table 6: Testing accuracies of varying target fine-tuning data size from *Amazon* to *Yelp*.

Percentage of source set	*Amazon* to *Yelp* (2nd set)			*Amazon* to *Yelp* (3rd set)		
	25%	50%	75%	25%	50%	75%
TgtOnly	51.10 ± 0.11	51.10 ± 0.11	51.10 ± 0.11	50.67 ± 0.09	50.67 ± 0.09	50.67 ± 0.09
SrcOnly	45.33 ± 0.09	46.97 ± 0.09	47.17 ± 0.09	43.40 ± 0.08	46.54 ± 0.10	47.76 ± 0.08
All	51.21 ± 0.08	51.43 ± 0.08	51.80 ± 0.08	50.90 ± 0.07	50.53 ± 0.09	52.11 ± 0.09
Uniform	51.43 ± 0.08	52.15 ± 0.11	52.77 ± 0.04	51.30 ± 0.08	51.56 ± 0.08	51.84 ± 0.06
Gradient	$\mathbf{52.17 \pm 0.10}$	52.15 ± 0.11	52.97 ± 0.11	51.75 ± 0.11	52.10 ± 0.11	52.10 ± 0.09
Variance	51.47 ± 0.08	52.85 ± 0.06	53.07 ± 0.11	51.39 ± 0.07	51.91 ± 0.08	52.30 ± 0.06
Hard-mixing	51.50 ± 0.06	$\mathbf{53.20 \pm 0.09}$	53.00 ± 0.05	51.33 ± 0.09	51.86 ± 0.06	52.28 ± 0.07
Soft-mixing	$\mathbf{52.17 \pm 0.08}$	53.00 ± 0.04	$\mathbf{53.43 \pm 0.05}$	$\mathbf{51.77 \pm 0.09}$	$\mathbf{52.21 \pm 0.06}$	$\mathbf{52.48 \pm 0.05}$

Table 7: Testing accuracies of varying source pre-training data size from *Amazon* to *Yelp*.

Fixed parameters	*Yelp2015* to *Yelp-2016*			*Yelp2015* to *Yelp-2017*		
	encoder	embedding	None	encoder	embedding	None
Uniform	56.96 ± 0.02	56.78 ± 0.02	57.18 ± 0.09	53.46 ± 0.04	54.93 ± 0.06	55.08 ± 0.08
Gradient	57.68 ± 0.03	57.00 ± 0.02	57.82 ± 0.10	54.80 ± 0.03	55.20 ± 0.04	55.08 ± 0.08
Variance	56.96 ± 0.02	$\mathbf{57.60 \pm 0.04}$	57.60 ± 0.08	53.46 ± 0.04	55.30 ± 0.10	55.66 ± 0.10
Hard-mixing	57.44 ± 0.03	57.46 ± 0.06	58.13 ± 0.09	54.62 ± 0.02	55.35 ± 0.10	55.78 ± 0.10
Soft-mixing	$\mathbf{57.68 \pm 0.03}$	$\mathbf{57.60 \pm 0.04}$	$\mathbf{58.30 \pm 0.09}$	$\mathbf{54.84 \pm 0.04}$	$\mathbf{55.48 \pm 0.07}$	$\mathbf{55.90 \pm 0.10}$

Table 8: Testing accuracies of varying fine-tuning parameters from *Yelp-2015* to *Yelp-2016* / *Yelp-2017* (1st set).

Percentage of target set	*Yelp2015* to *Yelp-2016*			*Yelp2015* to *Yelp-2017*		
	25%	50%	75%	25%	50%	75%
TgtOnly	42.07 ± 0.09	45.57 ± 0.08	49.33 ± 0.10	43.36 ± 0.08	46.78 ± 0.11	49.60 ± 0.10
SrcOnly	56.43 ± 0.10	56.43 ± 0.10	56.43 ± 0.10	54.45 ± 0.10	54.45 ± 0.10	54.45 ± 0.10
All	56.35 ± 0.07	56.63 ± 0.09	56.78 ± 0.11	53.98 ± 0.07	53.96 ± 0.08	54.40 ± 0.07
Uniform	56.58 ± 0.03	56.98 ± 0.02	57.15 ± 0.05	54.40 ± 0.03	54.52 ± 0.06	54.90 ± 0.04
Gradient	57.15 ± 0.07	57.10 ± 0.03	57.75 ± 0.06	$\mathbf{54.95 \pm 0.08}$	54.86 ± 0.10	55.00 ± 0.07
Variance	57.15 ± 0.07	57.60 ± 0.02	57.15 ± 0.05	54.40 ± 0.03	55.20 ± 0.10	55.15 ± 0.04
Hard-mixing	57.25 ± 0.08	57.72 ± 0.03	57.75 ± 0.06	54.70 ± 0.07	$\mathbf{55.48 \pm 0.09}$	55.03 ± 0.11
Soft-mixing	$\mathbf{57.28 \pm 0.06}$	$\mathbf{57.92 \pm 0.06}$	$\mathbf{58.20 \pm 0.10}$	$\mathbf{54.95 \pm 0.08}$	55.42 ± 0.10	$\mathbf{55.28 \pm 0.03}$

Table 9: Testing accuracies of varying the target fine-tuning data size from *Yelp-2015* to *Yelp-2016* / *Yelp-2017* (1st set).

Percentage of source set	*Yelp2015* to *Yelp-2016*			*Yelp2015* to *Yelp-2017*		
	25%	50%	75%	25%	50%	75%
TgtOnly	50.23 ± 0.11	50.23 ± 0.11	50.23 ± 0.11	54.06 ± 0.11	54.06 ± 0.11	54.06 ± 0.11
SrcOnly	53.43 ± 0.09	54.60 ± 0.11	55.80 ± 0.07	51.78 ± 0.10	52.28 ± 0.11	54.45 ± 0.11
All	54.77 ± 0.08	56.05 ± 0.05	56.27 ± 0.11	52.26 ± 0.11	54.57 ± 0.09	54.80 ± 0.10
Uniform	54.58 ± 0.10	55.77 ± 0.07	56.26 ± 0.05	53.63 ± 0.05	55.00 ± 0.10	55.05 ± 0.06
Gradient	54.58 ± 0.11	56.30 ± 0.10	$\mathbf{57.56 \pm 0.06}$	54.38 ± 0.06	55.00 ± 0.10	55.15 ± 0.08
Variance	55.18 ± 0.09	55.83 ± 0.08	56.38 ± 0.08	54.65 ± 0.09	55.67 ± 0.11	55.48 ± 0.04
Hard-mixing	54.58 ± 0.11	55.97 ± 0.02	56.84 ± 0.10	54.73 ± 0.06	55.70 ± 0.09	55.43 ± 0.03
Soft-mixing	$\mathbf{55.23 \pm 0.09}$	$\mathbf{56.37 \pm 0.07}$	$\mathbf{57.56 \pm 0.06}$	$\mathbf{54.83 \pm 0.07}$	$\mathbf{55.73 \pm 0.10}$	$\mathbf{55.55 \pm 0.03}$

Table 10: Testing accuracies of varying the source pre-training data size from *Yelp-2015* to *Yelp-2016* / *Yelp-2017* (1st set).

Method		*Yelp-2015* to *Yelp-2016* (2nd set)	*Yelp-2015* to *Yelp-2016* (3rd set)
Baselines	TgtOnly	51.32 ± 0.09	51.90 ± 0.08
	SrcOnly	56.55 ± 0.10	56.50 ± 0.04
	All	56.35 ± 0.06	57.60 ± 0.08
Fine-tuned	Uniform	57.14 ± 0.07	58.28 ± 0.03
	Gradient	57.88 ± 0.05	58.28 ± 0.03
	Variance	57.24 ± 0.06	$\mathbf{58.92 \pm 0.03}$
	Hard-mixing	57.74 ± 0.08	58.40 ± 0.06
	Soft-mixing	$\mathbf{57.90 \pm 0.05}$	$\mathbf{58.92 \pm 0.03}$

Table 11: Testing accuracies of transfer learning from *Yelp-2015* to *Yelp-2016*.

Fixed	*Yelp-2015* to *Yelp-2016* (2nd set)			*Yelp-2015* to *Yelp-2016* (3rd set)		
parameters	encoder	embedding	None	encoder	embedding	None
Uniform	56.66 ± 0.02	56.98 ± 0.06	57.14 ± 0.07	57.73 ± 0.05	58.24 ± 0.07	58.28 ± 0.03
Gradient	56.68 ± 0.02	57.72 ± 0.08	57.88 ± 0.05	57.73 ± 0.05	58.24 ± 0.07	58.28 ± 0.03
Variance	56.92 ± 0.02	57.14 ± 0.05	57.24 ± 0.06	$\mathbf{57.88 \pm 0.03}$	$\mathbf{58.72 \pm 0.08}$	$\mathbf{58.92 \pm 0.03}$
Hard-mixing	56.84 ± 0.02	57.48 ± 0.03	57.74 ± 0.08	57.73 ± 0.05	58.30 ± 0.04	58.40 ± 0.06
Soft-mixing	$\mathbf{56.92 \pm 0.03}$	$\mathbf{57.92 \pm 0.08}$	$\mathbf{57.90 \pm 0.05}$	$\mathbf{57.88 \pm 0.03}$	$\mathbf{58.72 \pm 0.08}$	$\mathbf{58.92 \pm 0.03}$

Table 12: Testing accuracies of varying fine-tuning parameters from *Yelp-2015* to *Yelp-2016*.

Percentage	*Yelp-2015* to *Yelp-2016* (2nd set)			*Yelp-2015* to *Yelp-2016* (3rd set)		
of target set	25%	50%	75%	25%	50%	75%
TgtOnly	44.74 ± 0.10	47.94 ± 0.11	50.02 ± 0.08	38.23 ± 0.09	46.90 ± 0.08	50.00 ± 0.06
SrcOnly	56.55 ± 0.10	56.55 ± 0.10	56.55 ± 0.10	56.50 ± 0.04	56.50 ± 0.04	56.50 ± 0.04
All	56.12 ± 0.09	55.90 ± 0.11	56.70 ± 0.09	56.37 ± 0.05	56.73 ± 0.08	57.00 ± 0.08
Uniform	57.05 ± 0.05	56.92 ± 0.07	57.05 ± 0.10	57.78 ± 0.01	58.00 ± 0.05	58.22 ± 0.03
Gradient	57.63 ± 0.06	57.48 ± 0.10	$\mathbf{57.68 \pm 0.06}$	57.78 ± 0.01	58.00 ± 0.05	58.22 ± 0.03
Variance	57.05 ± 0.05	57.36 ± 0.07	57.13 ± 0.09	$\mathbf{58.78 \pm 0.03}$	58.50 ± 0.06	58.96 ± 0.05
Hard-mixing	57.55 ± 0.05	57.52 ± 0.09	57.43 ± 0.11	58.60 ± 0.06	58.27 ± 0.03	58.74 ± 0.02
Soft-mixing	$\mathbf{57.58 \pm 0.05}$	$\mathbf{57.76 \pm 0.11}$	$\mathbf{57.68 \pm 0.06}$	$\mathbf{58.78 \pm 0.03}$	$\mathbf{58.53 \pm 0.04}$	$\mathbf{58.90 \pm 0.05}$

Table 13: Testing accuracies of varying target fine-tuning data size from *Yelp-2015* to *Yelp-2016*.

Percentage	*Yelp-2015* to *Yelp-2016* (2nd set)			*Yelp-2015* to *Yelp-2016* (3rd set)		
of source set	25%	50%	75%	25%	50%	75%
TgtOnly	51.32 ± 0.09	51.32 ± 0.09	51.32 ± 0.09	51.90 ± 0.08	51.90 ± 0.08	51.90 ± 0.08
SrcOnly	54.10 ± 0.11	54.52 ± 0.09	55.10 ± 0.08	52.17 ± 0.09	53.20 ± 0.06	54.97 ± 0.09
All	53.70 ± 0.09	55.70 ± 0.09	56.25 ± 0.05	53.53 ± 0.08	56.10 ± 0.05	56.53 ± 0.06
Uniform	54.78 ± 0.10	56.24 ± 0.05	56.80 ± 0.08	56.13 ± 0.09	56.30 ± 0.04	56.83 ± 0.07
Gradient	55.38 ± 0.08	$\mathbf{56.94 \pm 0.06}$	56.80 ± 0.08	56.13 ± 0.09	56.50 ± 0.01	56.83 ± 0.07
Variance	54.78 ± 0.10	56.24 ± 0.05	$\mathbf{57.37 \pm 0.07}$	$\mathbf{57.08 \pm 0.05}$	56.40 ± 0.07	$\mathbf{57.17 \pm 0.06}$
Hard-mixing	54.78 ± 0.10	56.70 ± 0.08	57.10 ± 0.06	56.83 ± 0.04	56.48 ± 0.08	56.97 ± 0.06
Soft-mixing	$\mathbf{55.43 \pm 0.07}$	$\mathbf{56.94 \pm 0.06}$	$\mathbf{57.37 \pm 0.07}$	$\mathbf{57.08 \pm 0.05}$	$\mathbf{56.82 \pm 0.08}$	$\mathbf{57.17 \pm 0.06}$

Table 14: Testing accuracies of varying pre-training source data size from *Yelp-2015* to *Yelp-2016*.

Method		*Yelp-2015* to *Yelp-2017* (2nd set)	*Yelp-2015* to *Yelp-2017* (3rd set)
Baselines	TgtOnly	50.27 ± 0.09	51.14 ± 0.09
	SrcOnly	53.60 ± 0.10	54.45 ± 0.10
	All	54.37 ± 0.09	54.80 ± 0.07
Fine-tuned	Uniform	56.07 ± 0.07	56.52 ± 0.05
	Gradient	56.07 ± 0.07	56.72 ± 0.04
	Variance	$\mathbf{56.77 \pm 0.09}$	57.62 ± 0.04
	Hard-mixing	56.63 ± 0.08	57.62 ± 0.04
	Soft-mixing	$\mathbf{56.77 \pm 0.09}$	$\mathbf{57.74 \pm 0.03}$

Table 15: Testing accuracies of transfer learning from *Yelp-2015* to *Yelp-2017*.

Fixed	*Yelp-2015* to *Yelp-2017* (2nd set)			*Yelp-2015* to *Yelp-2017* (3rd set)		
parameters	encoder	embedding	None	encoder	embedding	None
Uniform	53.02 ± 0.03	55.88 ± 0.06	56.07 ± 0.07	55.86 ± 0.03	56.38 ± 0.06	56.52 ± 0.05
Gradient	53.02 ± 0.03	55.88 ± 0.06	56.07 ± 0.07	55.86 ± 0.03	56.92 ± 0.04	56.72 ± 0.04
Variance	53.90 ± 0.04	$\mathbf{56.40 \pm 0.03}$	$\mathbf{56.77 \pm 0.09}$	$\mathbf{56.50 \pm 0.03}$	56.60 ± 0.04	57.62 ± 0.04
Hard-mixing	53.92 ± 0.03	56.28 ± 0.07	56.63 ± 0.08	56.50 ± 0.05	56.64 ± 0.03	57.62 ± 0.04
Soft-mixing	$\mathbf{53.94 \pm 0.03}$	$\mathbf{56.40 \pm 0.03}$	$\mathbf{56.77 \pm 0.09}$	$\mathbf{56.50 \pm 0.03}$	$\mathbf{57.04 \pm 0.02}$	$\mathbf{56.74 \pm 0.03}$

Table 16: Testing accuracies of varying fine-tuning parameters from *Yelp-2015* to *Yelp-2017*.

Percentage	*Yelp-2015* to *Yelp-2017* (2nd set)			*Yelp-2015* to *Yelp-2017* (3rd set)		
of target set	25%	50%	75%	25%	50%	75%
TgtOnly	39.70 ± 0.10	45.10 ± 0.11	48.77 ± 0.11	40.55 ± 0.11	46.74 ± 0.09	49.62 ± 0.09
SrcOnly	53.60 ± 0.10	53.60 ± 0.10	53.60 ± 0.10	54.45 ± 0.10	54.45 ± 0.10	54.45 ± 0.10
All	52.90 ± 0.09	53.93 ± 0.08	54.43 ± 0.09	52.77 ± 0.07	53.93 ± 0.08	54.73 ± 0.08
Uniform	54.70 ± 0.06	55.58 ± 0.03	55.13 ± 0.04	55.57 ± 0.06	56.36 ± 0.03	56.37 ± 0.04
Gradient	54.70 ± 0.06	55.58 ± 0.03	55.13 ± 0.04	55.63 ± 0.04	56.74 ± 0.08	57.27 ± 0.02
Variance	$\mathbf{55.33 \pm 0.02}$	$\mathbf{56.34 \pm 0.05}$	$\mathbf{55.87 \pm 0.04}$	56.17 ± 0.06	56.58 ± 0.02	56.60 ± 0.07
Hard-mixing	55.20 ± 0.03	56.28 ± 0.08	55.80 ± 0.03	$\mathbf{56.43 \pm 0.08}$	56.64 ± 0.06	56.70 ± 0.03
Soft-mixing	$\mathbf{55.33 \pm 0.02}$	$\mathbf{56.34 \pm 0.05}$	$\mathbf{55.87 \pm 0.04}$	56.17 ± 0.06	$\mathbf{57.10 \pm 0.04}$	$\mathbf{57.27 \pm 0.02}$

Table 17: Testing accuracies of varying target fine-tuning data size from *Yelp-2015* to *Yelp-2017*.

Percentage	*Yelp-2015* to *Yelp-2017* (2nd set)			*Yelp-2015* to *Yelp-2017* (3rd set)		
of source set	25%	50%	75%	25%	50%	75%
TgtOnly	50.27 ± 0.09	50.27 ± 0.09	50.27 ± 0.09	51.14 ± 0.09	51.14 ± 0.09	51.14 ± 0.09
SrcOnly	50.17 ± 0.08	51.87 ± 0.10	52.30 ± 0.08	46.13 ± 0.09	49.13 ± 0.10	50.77 ± 0.09
All	51.10 ± 0.09	52.20 ± 0.11	52.77 ± 0.09	51.17 ± 0.08	52.83 ± 0.07	54.70 ± 0.08
Uniform	55.27 ± 0.06	53.47 ± 0.10	53.84 ± 0.05	52.98 ± 0.03	54.93 ± 0.07	55.62 ± 0.04
Gradient	55.27 ± 0.06	53.47 ± 0.10	54.42 ± 0.06	53.20 ± 0.09	55.18 ± 0.06	55.52 ± 0.09
Variance	$\mathbf{56.10 \pm 0.06}$	53.83 ± 0.07	54.24 ± 0.09	53.58 ± 0.04	55.65 ± 0.08	56.00 ± 0.09
Hard-mixing	55.97 ± 0.06	53.73 ± 0.09	54.12 ± 0.09	$\mathbf{53.68 \pm 0.06}$	55.38 ± 0.09	56.00 ± 0.09
Soft-mixing	$\mathbf{56.10 \pm 0.06}$	$\mathbf{54.10 \pm 0.09}$	$\mathbf{54.48 \pm 0.05}$	53.66 ± 0.03	$\mathbf{55.68 \pm 0.08}$	$\mathbf{56.16 \pm 0.05}$

Table 18: Testing accuracies of varying source pre-training data size from *Yelp-2015* to *Yelp-2017*.