Transferable Calibration with Lower Bias and Variance in Domain Adaptation

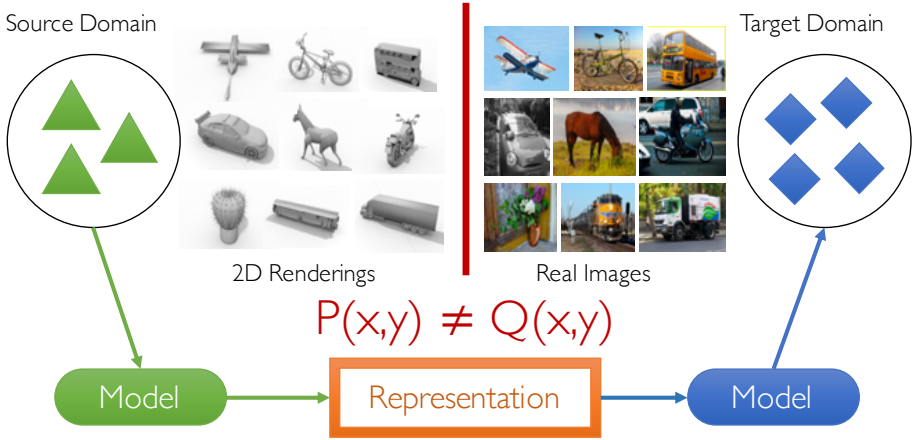
Ximei Wang, Mingsheng Long (✉), Jianmin Wang, and Michael I. Jordan[♯]

School of Software, Tsinghua University
National Engineering Laboratory for Big Data Software
[♯]University of California, Berkeley

wxm17@mails.tsinghua.edu.cn
https://wxm17.github.io/
Neural Information Processing Systems (NeurIPS), 2020

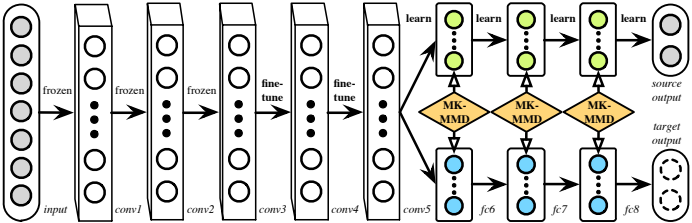
Domain Adaptation (DA)

Transfer from a labeled source domain to an unlabeled target one.

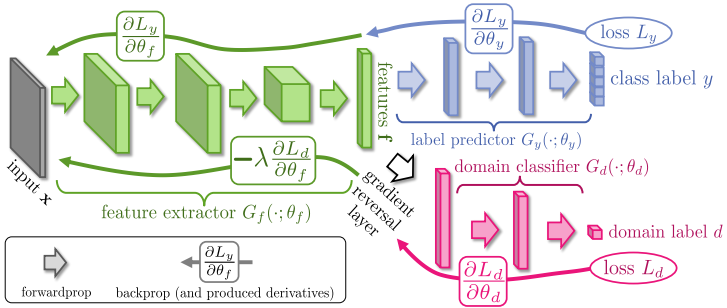


Domain Adaptation Approaches

- **Moment Matching**
- Deep Adaptation Netwotk etc.



- **Adversarial Training**
- Domain Adversarial Neural Network etc.



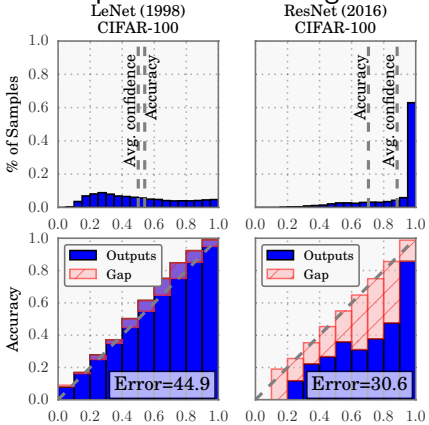
Confidence Calibration in Deep Learning[1]

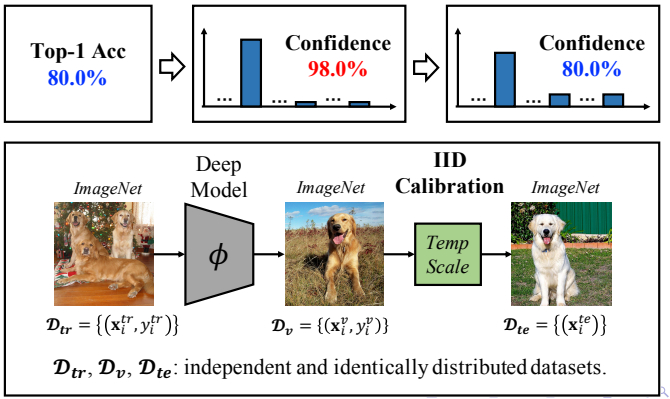
- A model should output a probability reflecting the true frequency:

$$\mathbb{P}(\widehat{Y} = Y | \widehat{P} = c) = c, \ \forall \ c \in [0, 1] \tag{1}$$

where $\widehat{Y}$ is the class prediction and $\widehat{P}$ is its associated confidence.

- Deep networks learn high accuracy at the expense of over-confidence.





Temperature Scaling for IID Calibration

- **Calibration Metric**: Expected Calibration Error (ECE)

$$\begin{aligned}
\mathcal{L}_{\mathrm{ECE}} &= \sum_{m=1}^{B} \frac{|B_m|}{n} |\mathbb{A}(B_m) - \mathbb{C}(B_m)| \\
\mathbb{A}(B_m) &= |B_m|^{-1} \sum_{i \in B_m} 1(\hat{\mathrm{y}}_i = \mathrm{y}_i) \quad \textbf{(Accuracy)} \\
\mathbb{C}(B_m) &= |B_m|^{-1} \sum_{i \in B_m} \max_k p(\hat{\mathrm{y}}_i^k | \mathrm{x}_i, \boldsymbol{\theta}) \quad \textbf{(Confidence)}
\end{aligned} \tag{2}$$

- **IID Calibration**: Temperature Scaling

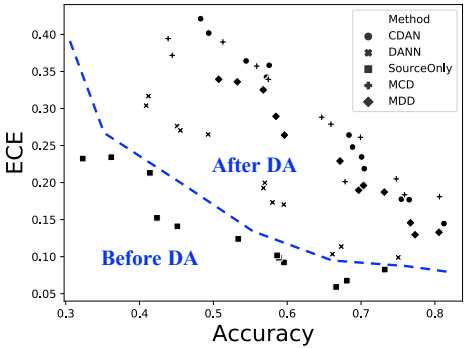
$$T^* = \arg\min_T \mathrm{E}_{(\mathrm{x}_v, \mathrm{y}_v) \in \mathcal{D}_v} \mathcal{L}_{\mathrm{NLL}} \left(\sigma(\mathrm{z}_v / T), \mathrm{y}_v \right) \tag{3}$$

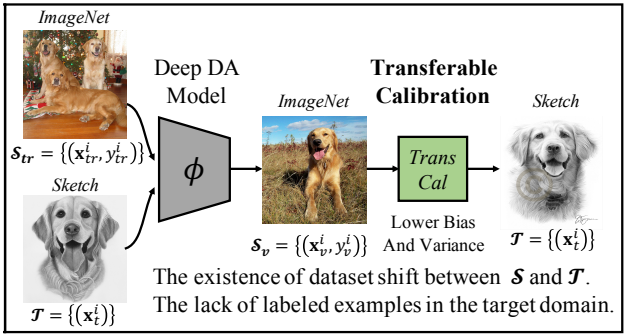
  σ is the *softmax* function, $\mathcal{L}_{\mathrm{NLL}}$ is Negative Log-Likelihood loss.

- Transform logits z_{te} into calibrated probabilities $p_{te} = \sigma(\mathrm{z}_{te} / T^*)$.

Dilemma of Accuracy vs Confidence in DA

- Transfer models yield high accuracy at the expense of over-confidence.





- Calibration in transfer learning is challenging due to the coexistence:
  - Domain shift — ECE should be unbiased to the target domain
  - Unlabeled target — ECE on the target domain is incomputable
- Bias-Variance-Shift Dilemma of confidence calibration in Transfer Learning

Transferable Calibration Framework

- $\mathrm{E}_{\mathrm{x}\sim p}\left[w(\mathrm{x})\mathcal{L}_{(\cdot)}(\phi(\mathrm{x}),\ \mathrm{y})\right]$ is an **unbiased** estimator of the target calibration error $\mathbb{E}_q$

$$\begin{aligned}\mathrm{E}_{\mathrm{x}\sim q}\left[\mathcal{L}_{(\cdot)}(\phi(\mathrm{x}),\ \mathrm{y})\right] &= \int_q \mathcal{L}_{(\cdot)}(\phi(\mathrm{x}),\ \mathrm{y})q(\mathrm{x})\mathrm{d}\mathrm{x} \\ &= \int_p \frac{q(\mathrm{x})}{p(\mathrm{x})}\mathcal{L}_{(\cdot)}(\phi(\mathrm{x}),\ \mathrm{y})p(\mathrm{x})\mathrm{d}\mathrm{x} = \mathrm{E}_{\mathrm{x}\sim p}\left[w(\mathrm{x})\mathcal{L}_{(\cdot)}(\phi(\mathrm{x}),\ \mathrm{y})\right]\end{aligned} \tag{4}$$

- Discriminative density ratio estimation method: **LogReg**
- Use Bayesian formula to derive $\widehat{w}(\mathrm{x})$ from a logistic regression classifier

$$\widehat{w}(\mathrm{x}) = \frac{q(\mathrm{x})}{p(\mathrm{x})} = \frac{v(\mathrm{x}|d=0)}{v(\mathrm{x}|d=1)} = \frac{P(d=1)}{P(d=0)}\frac{P(d=0|\mathrm{x})}{P(d=1|\mathrm{x})} \tag{5}$$

Transferable Calibration: Bias Reduction

- Importance-weighting for an unbiased estimate of target ECE if $\widehat{w}(\mathrm{x}) = w(\mathrm{x})$
- The **bias** between the estimated ECE and the ground-truth ECE

$$\begin{aligned} &\left|\mathrm{E}_{\mathrm{x}\sim q}\left[\mathcal{L}_{\mathrm{ECE}}^{\widehat{w}(\mathrm{x})}\right] - \mathrm{E}_{\mathrm{x}\sim q}\left[\mathcal{L}_{\mathrm{ECE}}^{w(\mathrm{x})}\right]\right| \\ =&\left|\mathrm{E}_{\mathrm{x}\sim p}\left[\widehat{w}(\mathrm{x})\mathcal{L}_{\mathrm{ECE}}(\phi(\mathrm{x}),\ \mathrm{y})\right] - \mathrm{E}_{\mathrm{x}\sim p}\left[w(\mathrm{x})\mathcal{L}_{\mathrm{ECE}}(\phi(\mathrm{x}),\ \mathrm{y})\right]\right| \\ =&\left|\mathrm{E}_{\mathrm{x}\sim p}\left[(w(\mathrm{x}) - \widehat{w}(\mathrm{x}))\mathcal{L}_{\mathrm{ECE}}(\phi(\mathrm{x}),\ \mathrm{y})\right]\right|. \end{aligned} \tag{6}$$

- The **bias** of them can be further bounded by

$$\begin{aligned} &\left|\mathrm{E}_{\mathrm{x}\sim p}\left[(w(\mathrm{x}) - \widehat{w}(\mathrm{x}))\,\mathcal{L}_{\mathrm{ECE}}(\phi(\mathrm{x}),\ \mathrm{y})\right]\right| \\ \leq&\sqrt{\mathrm{E}_{\mathrm{x}\sim p}\left[(w(\mathrm{x}) - \widehat{w}(\mathrm{x}))^2\right]\mathrm{E}_{\mathrm{x}\sim p}\left[(\mathcal{L}_{\mathrm{ECE}}(\phi(\mathrm{x}),\ \mathrm{y}))^2\right]} \qquad (\mathrm{Cachy-Schwarz\ Ineqaulity}) \\ \leq&\frac{1}{2}\left(\mathrm{E}_{\mathrm{x}\sim p}\left[(w(\mathrm{x}) - \widehat{w}(\mathrm{x}))^2\right] + \mathrm{E}_{\mathrm{x}\sim p}\left[(\mathcal{L}_{\mathrm{ECE}}(\phi(\mathrm{x}),\ \mathrm{y}))^2\right]\right) \quad (\mathrm{AM/GM\ Inequality}) \end{aligned} \tag{7}$$

Transferable Calibration: Bias Reduction

- For any x s.t. $P(d=1|\mathrm{x}) \neq 0$, the following inequality holds:

$$\frac{1}{M+1} \leq P(d=1|\mathrm{x}) \leq 1, \quad \text{since } w(\mathrm{x}) = \frac{P(d=0|\mathrm{x})}{P(d=1|\mathrm{x})} = \frac{1-P(d=1|\mathrm{x})}{P(d=1|\mathrm{x})} = \frac{1}{P(d=1|\mathrm{x})} - 1. \tag{8}$$

- The discrepancy between $\widehat{w}(\mathrm{x})$ and $w(\mathrm{x})$ can be bounded by

$$\begin{aligned} \mathbb{E}_{\mathrm{x}\sim p}\left[\left(w(\mathrm{x}) - \widehat{w}(\mathrm{x})\right)^2\right] &= \mathrm{E}_{\mathrm{x}\sim p}\left[\left(\frac{P(d=1|\mathrm{x}) - \widehat{P}(d=1|\mathrm{x})}{P(d=1|\mathrm{x})\widehat{P}(d=1|\mathrm{x})}\right)^2\right] \\ &\leq (M+1)^4 \mathrm{E}_{\mathrm{x}\sim p}\left[\left(P(d=1|\mathrm{x}) - \widehat{P}(d=1|\mathrm{x})\right)^2\right]. \end{aligned} \tag{9}$$

- Use λ $(0 \leq \lambda \leq 1)$ to control the bound M of the importance weights

$$T^*, \lambda^* = \arg\min_{T,\lambda} \mathrm{E}_{\mathrm{x}_v\sim p}\left[\widetilde{w}(\mathrm{x}_v)\mathcal{L}_{\mathrm{ECE}}(\sigma(\phi(\mathrm{x}_v)/T),\ \mathrm{y})\right], \quad \widetilde{w}(\mathrm{x}_v^i) = \left[\widehat{w}(\mathrm{x}_v^i)\right]^\lambda. \tag{10}$$

Control Variate Method

- (a) Feature adaptation reduces distribution discrepancy $d_{\alpha+1}(q||p)$

$$\begin{aligned}\mathrm{Var}_{x\sim p}[\mathcal{L}^w_{\mathrm{ECE}}] &= \mathsf{E}_{x\sim p}\left[(\mathcal{L}^w_{\mathrm{ECE}})^2\right] - (\mathsf{E}_{x\sim p}[\mathcal{L}^w_{\mathrm{ECE}}])^2 \\ &\leq d_{\alpha+1}(q||p)(\mathsf{E}_{x\sim p}\mathcal{L}^w_{\mathrm{ECE}})^{1-\frac{1}{\alpha}} - (\mathsf{E}_{x\sim p}\mathcal{L}^w_{\mathrm{ECE}})^2, \quad \forall\alpha > 0.\end{aligned} \tag{11}$$

- (b) Control variate explicitly reduces the variance [2]
  1. Given two unbiased estimators: $\mathbb{E}[z] = \zeta, \mathbb{E}[t] = \tau$
  2. Construct a new estimator: $z^\star = z + \eta(t - \tau)$
  3. $z^\star$ is still unbiased: $\mathbb{E}[z^\star] = \mathbb{E}[z] + \eta\mathbb{E}[t - \tau] = \zeta + \eta(\mathbb{E}[t] - \mathbb{E}[\tau]) = \zeta$
  4. $\mathrm{Var}[z^\star] = \mathrm{Var}[z + \eta(t - \tau)] = \eta^2\mathrm{Var}[t] + 2\eta\mathrm{Cov}(z, t) + \mathrm{Var}[z]$
  5. $\min \mathrm{Var}[z^\star] = (1 - \rho^2_{z,t})(\mathrm{Var}[z],$ when $\hat{\eta} = -\frac{\mathrm{Cov}(z,t)}{\mathrm{Var}[t]}$
  6. Since $0 \leq \rho^2_{z,t} \leq 1$, $\mathrm{Var}[z^\star] \leq \mathrm{Var}[z]$, the variance is reduced.

[2] Lemieux. Control variates. In Wiley StatsRef: Statistics Reference Online, American Cancer Society, 2017.

Transferable Calibration: Variance Reduction

- **Serial Control Variate:** $\mathrm{Var}[u^{**}] \leq \mathrm{Var}[u^{*}] \leq \mathrm{Var}[u]$

$$\begin{aligned} u^{*} &= u + \eta_1(t_1 - \tau_1) \\ u^{**} &= u^{*} + \eta_2(t_2 - \tau_2) \end{aligned} \tag{12}$$

- First, use importance weight $\widetilde{w}(\mathrm{x}_s)$ as a control covariate

$$\mathbb{E}_q^{*}(\widehat{\mathrm{y}}, \mathrm{y}) = \widetilde{\mathbb{E}}_q(\widehat{\mathrm{y}}, \mathrm{y}) - \frac{1}{n_s} \frac{\mathrm{Cov}(\mathcal{L}_{\mathrm{ECE}}^{\widetilde{w}}, \widetilde{w}(\mathrm{x}))}{\mathrm{Var}[\widetilde{w}(\mathrm{x})]} \sum_{i=1}^{n_s} [\widetilde{w}(\mathrm{x}_s^i) - 1]. \tag{13}$$

- Second, use the prediction correctness $r(\mathrm{x}_s)$ as another control variate

$$\mathbb{E}_q^{**}(\widehat{\mathrm{y}}, \mathrm{y}) = \mathbb{E}_q^{*}(\widehat{\mathrm{y}}, \mathrm{y}) - \frac{1}{n_s} \frac{\mathrm{Cov}(\mathcal{L}_{\mathrm{ECE}}^{\widetilde{w}*}, r(\mathrm{x}))}{\mathrm{Var}[r(\mathrm{x})]} \sum_{i=1}^{n_s} [r(\mathrm{x}_s^i) - c], \tag{14}$$

- Reduce bias, variance, and shift all-in-one for Transferable Calibration

TransCal Algorithm

Algorithm 1 Transferable Calibration in Domain Adaptation

Input: Labeled source dataset $\mathcal{S} = \left\{\left(\mathbf{x}_s^i, \mathbf{y}_s^i\right)\right\}_{i=1}^{n_s}$ and unlabled target dataset $\mathcal{T} = \left\{\left(\mathbf{x}_t^i\right)\right\}_{i=1}^{n_t}$

Parameter: Temperature T and learnable meta parameter λ

Partition $\mathcal{S}$ into $\mathcal{S}_{tr} = \left\{\left(\mathbf{x}_{tr}^i, \mathbf{y}_{tr}^i\right)\right\}_{i=1}^{n_{tr}}$ and $\mathcal{S}_v = \left\{\left(\mathbf{x}_v^i, \mathbf{y}_v^i\right)\right\}_{i=1}^{n_v}$

Train a DA model $\phi(\mathbf{x}) = G(F(\mathbf{x}))$ on $\mathcal{S}_{tr}$ and $\mathcal{T}$ via any DA method until convergy

Randomly upsample the source or the target dataset to make $n_{tr} = n_t$

Fix the DA model and compute features $\mathcal{F}_{tr} = \left\{\boldsymbol{f}_{tr}^i\right\}_{i=1}^{n_{tr}}$, $\mathcal{F}_v = \left\{\boldsymbol{f}_v^i\right\}_{i=1}^{n_v}$, $\mathcal{F}_t = \left\{\boldsymbol{f}_t^i\right\}_{i=1}^{n_t}$

Train a logistic regression model H to discriminate the features $\mathcal{F}_{tr}$ and $\mathcal{F}_t$ until converge

Compute $\widehat{w}(\mathbf{x}_v^i) = \left[1 - H(\boldsymbol{f}_v^i)\right] / H(\boldsymbol{f}_v^i)$ and $\widetilde{w}(\mathbf{x}_v^i) = \left[\widehat{w}(\mathbf{x}_v^i)\right]^{\lambda}$

Compute $\mathsf{E}_{\mathbf{x}\sim p}\mathcal{L}_{\mathrm{ECE}}^{\widetilde{w}}$, $\mathbb{E}_q^*(\widehat{\mathbf{y}}, \mathbf{y})$ and $\mathbb{E}_q^{**}(\widehat{\mathbf{y}}, \mathbf{y})$ as in Eq. 9, Eq. 11 and Eq. 13 respectively

Jointly optimize the transferable calibration objective as $T^* = \underset{T,\lambda}{\arg\min}\, \mathbb{E}_q^{**}\left(\sigma(\phi(\mathbf{x}_v)/T),\ \mathbf{y}_v\right)$

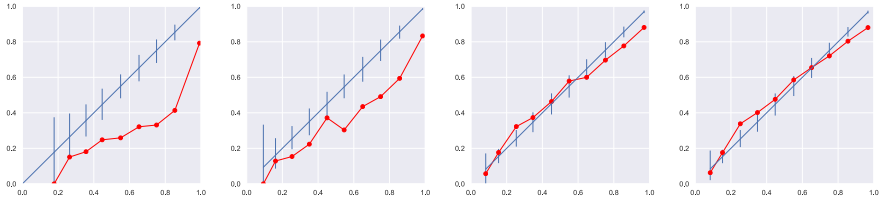
Calibrate the logit vectors on the target domain by $\widehat{\mathbf{y}}_t = \sigma(\phi(\mathbf{x}_t)/T^*)$

Experiments and Results

Table 2: ECE (%) vs. Acc (%) via various calibration methods on *Office-Home* with CDAN

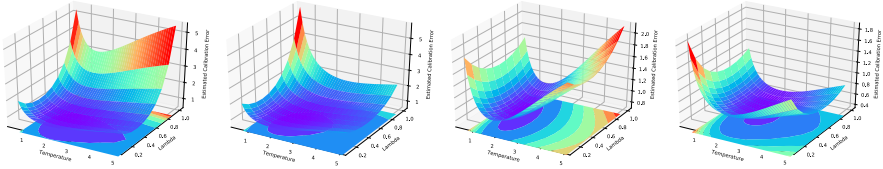
Metric	Cal. Method	A→C	A→P	A→R	C→A	C→P	C→R	R→A	R→C	R→P	Avg
Acc	Before Cal.	49.4	68.4	75.5	57.6	70.1	70.4	68.9	54.4	81.2	**68.3**
	MC-dropout [12]	47.2	66.2	71.4	57.1	65.7	70.6	68.3	53.6	80.7	66.7
	TransCal (ours)	49.4	68.4	75.5	57.6	70.1	70.4	68.9	54.4	81.2	**68.3**
ECE	Before Cal.	40.2	26.4	17.8	35.8	23.5	21.9	24.8	36.4	14.5	26.8
	MC-dropout [12]	33.1	21.3	15.0	24.2	20.5	13.2	25.6	14.2	22.4	19.6
	Matrix Scaling	44.7	28.8	19.7	36.1	25.4	24.1	38.1	15.7	29.5	29.1
	Vector Scaling	34.7	18.0	11.3	23.4	15.4	11.5	27.3	8.5	20.0	18.9
	Temp. Scaling	28.3	17.6	10.1	**21.2**	13.2	8.2	26.0	8.8	18.1	16.8
	CPCS [38]	35.0	29.4	8.3	21.3	29.0	**5.6**	**19.9**	9.1	20.3	19.8
	TransCal (w/o Bias)	**21.7**	10.8	5.8	27.6	**9.2**	6.0	27.4	5.2	16.9	14.5
	TransCal (w/o Variance)	31.2	16.4	6.5	31.1	14.7	16.1	27.5	**4.1**	20.0	18.6
	TransCal (ours)	22.9	**9.3**	**5.1**	21.7	14.0	6.4	21.6	4.5	**15.6**	**13.5**
	Oracle	5.8	8.1	4.8	10.0	7.7	4.2	5.5	3.9	6.2	6.2

Experiments and Results



(a) Before Calibration (b) IID Calibration (c) TransCal (d) Oracle

Figure 2: Reliability diagrams from *Clipart* to *Product* with CDAN [25] before and after calibration.



(a) *Art* $\rightarrow$ *Clipart* (b) *Art* $\rightarrow$ *Product* (c) *Art* $\rightarrow$ *Real-World* (d) *Clipart* $\rightarrow$ *Art*

Figure 3: The estimated calibration error with respect to different values of temperature T and meta parameter λ (both are *learnable*), showing that different models achieve optimal values at different λ.

Summary

- A dilemma in the open problem of Calibration in DA: existing domain adaptation models learn higher classification accuracy *at the expense of* well-calibrated probabilities.
- A Transferable Calibration (TransCal) method, achieving more accurate calibration with lower bias and variance in a unified hyperparameter-free optimization framework.
- Extensive experiments on various DA methods, datasets, and calibration metrics, while the effectiveness of our method has been justified both theoretically and empirically.
- Code will be available @ `github.com/thuml/TransCal`

Future Work

- **1. Design DA methods based on our ECE-Accuracy-Dilemma observation**
- **2. TransCal may still fall short under the following circumstances**:
- The domain gap is extremely large even after applying domain adaptation methods
- The source or the target dataset is too small to estimate importance weights
- TransCal is based on the covariate shift assumption and it remains unclear whether it can still perform well under label shift, especially when we meet with a long-tailed distribution.