Transferable Calibration with Lower Bias and Variance in Domain Adaptation

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

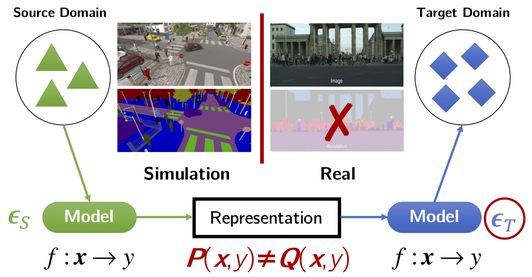
School of Software, KLiss, BNRist, Tsinghua University ♯University of California, Berkeley

Summary

- A Transferable Calibration (TransCal) method, achieving more accurate calibration with lower bias and variance in a unified hyperparameter-free optimization framework.
- 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.
- Extensive experiments on various DA methods, datasets, and calibration metrics, while the effectiveness of our method has been justified both theoretically and empirically.
- Code available @ `github.com/thuml/TransCal`

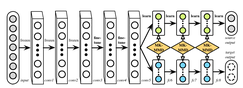
Domain Adaptation (DA)

- Deep learning across domains: $(P \neq Q)$
  - Non independent and identically distributed distributions (Non-IID)

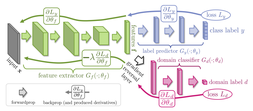


Mainstream Approaches to DA

- Numerous deep DA methods can be mainly grouped into two categories: moment matching and adversarial training.
- Most of DA methods focus on improving the accuracy in the target domain but fail to estimate the predictive uncertainty, falling short of a miscalibration problem.



(a) Moment Matching: DAN



(b) Adversarial Training: DANN

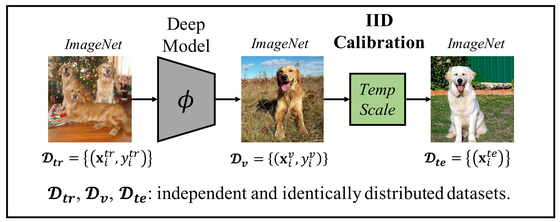
Confidence Calibration in Deep Learning

- A model should output a prediction probability reflecting the true frequency of an event:

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

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

- DNNs learn high accuracy at the cost of over-confidence.



Calibration Metric

- Expected Calibration Error (ECE)

$$\mathcal{L}_{\text{ECE}} = \sum_{m=1}^{B} \frac{|B_m|}{n} |\mathbb{A}(B_m) - \mathbb{C}(B_m)|$$

$$\mathbb{A}(B_m) = |B_m|^{-1} \sum_{i \in B_m} 1(\widehat{\mathsf{y}}_i = \mathsf{y}_i) \quad (\textit{Accuracy}) \tag{2}$$

$$\mathbb{C}(B_m) = |B_m|^{-1} \sum_{i \in B_m} \max_k p(\widehat{\mathsf{y}}_i^k | \mathsf{x}_i, \boldsymbol{\theta}) \quad (\textit{Confidence})$$

Temperature Scaling for IID Calibration

- Fix the neural model trained on the training set $\mathcal{D}_{tr}$
- Attain the optimal temperature T^* by minimizing

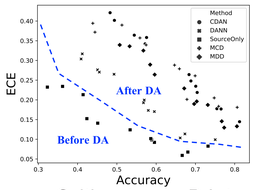
$$T^* = \arg\min_T \mathsf{E}_{(\mathsf{x}_v, \mathsf{y}_v) \in \mathcal{D}_v} \mathcal{L}_{\text{NLL}}(\sigma(\mathsf{z}_v / T), \mathsf{y}_v) \tag{3}$$

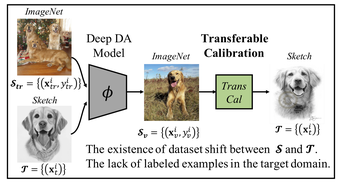
σ is the *softmax* function, $\mathcal{L}_{\text{NLL}}$ is Negative Log-Likelihood.

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

Dilemma of Accuracy vs Confidence in DA

- DA models yield high acc at the cost of poorly-calibration.





- Calibration in DA is challenging due to the existence of domain shift and the lack of target label

Transferable Calibration Framework

- Estimate the target ECE by importance weighting

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

- Estimate density ratio from a logistic regression classifier

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

Transferable Calibration: Bias Reduction

- Bias between the estimated ECE and the ground-truth one

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

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

$$\mathbb{E}_{\mathsf{x}\sim p}\left[(w(\mathsf{x}) - \widehat{w}(\mathsf{x}))^2\right] \leq (M+1)^4 \mathsf{E}_{\mathsf{x}\sim p}\left[\left(P(d=1|\mathsf{x}) - \widehat{P}(d=1|\mathsf{x})\right)^2\right]. \tag{7}$$

- Use λ $(0 \leq \lambda \leq 1)$ to control the bound M of $\widehat{w}(\mathsf{x})$

$$T^* = \arg\min_{T,\lambda} \mathsf{E}_{\mathsf{x}_v\sim p}\left[\widetilde{w}(\mathsf{x}_v) \mathcal{L}_{\text{ECE}}(\sigma(\phi(\mathsf{x}_v)/T),\ \mathsf{y})\right], \quad \widetilde{w}(\mathsf{x}_v^i) = \left[\widehat{w}(\mathsf{x}_v^i)\right]^\lambda. \tag{8}$$

Transferable Calibration: Variance Reduction

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

$$\begin{aligned}u^* &= u + \eta_1(t_1 - \tau_1) \\ u^{**} &= u^* + \eta_2(t_2 - \tau_2)\end{aligned} \tag{9}$$

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

$$\mathbb{E}_q^*(\widehat{\mathsf{y}}, \mathsf{y}) = \widetilde{\mathbb{E}}_q(\widehat{\mathsf{y}}, \mathsf{y}) - \frac{1}{n_s} \frac{\mathrm{Cov}(\mathcal{L}_{\text{ECE}}^{\widetilde{w}}, \widetilde{w}(\mathsf{x}))}{\mathrm{Var}[\widetilde{w}(\mathsf{x})]} \sum_{i=1}^{n_s} [\widetilde{w}(\mathsf{x}_s^i) - 1]. \tag{10}$$

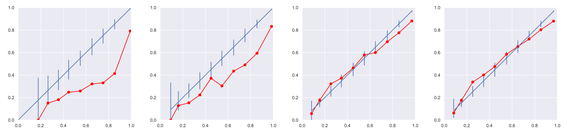
- Second, use the prediction correctness $r(\mathsf{x}_s)$ as another one

$$\mathbb{E}_q^{**}(\widehat{\mathsf{y}}, \mathsf{y}) = \mathbb{E}_q^*(\widehat{\mathsf{y}}, \mathsf{y}) - \frac{1}{n_s} \frac{\mathrm{Cov}(\mathcal{L}_{\text{ECE}}^{\widetilde{w}*}, r(\mathsf{x}))}{\mathrm{Var}[r(\mathsf{x})]} \sum_{i=1}^{n_s} [r(\mathsf{x}_s^i) - c], \tag{11}$$

Experiments and Results

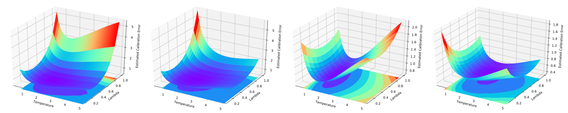
Table 2: ECE (%) vs. Acc (%) via various calibration methods on *Office-Home* with CDAN

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



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

Figure 2: Reliability diagrams from *Clipart* to *Product* with CDAN [30] before and after calibration.



(a) *Art* → *Clipart* (b) *Art* → *Product* (c) *Art* → *Real-World* (d) *Clipart* → *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.