Transfer Learning with Informative Priors: Simple Baselines Better than Previously Reported

Tufts UNIVERSITY

Ethan Harvey[1*] Mikhail Petrov[2*] Michael C. Hughes[1]

[1]Department of Computer Science, Tufts University [2]Department of Mechanical Engineering, Tufts University

Findings

- Standard transfer learning better than reported in Shwartz-Ziv et al. (2022).
- Relative gains of informed priors over standard transfer learning vary across datasets.
- Large variability in quality of alignment between training and test loss landscapes.

Background

Bayesian transfer learning. Recent work by Shwartz-Ziv et al. (2022) proposed Bayesian transfer learning, where a re-scaled posterior from the source task is used as the prior for the target task. This work is motivated by sequential Bayesian updating, where some source data $\mathcal{D}_S$ is acquired, a posterior $p(w|\mathcal{D}_S)$ over weights w is formed, and this posterior is used as an informed prior for a target dataset $\mathcal{D}_T$

$$p(w|\mathcal{D}_S) \propto p(\mathcal{D}_S|w)p(w)$$
$$p(w|\mathcal{D}_T, \mathcal{D}_S) \propto p(\mathcal{D}_T|w)p(w|\mathcal{D}_S).$$

Common framework for MAP transfer learning

Probabilistic model for target task.

$$p(w) = \mathcal{N}(w \mid \mathbf{m}, \lambda \mathbf{S}) \qquad \text{source-informed prior on backbone}$$
$$p(V) = \mathcal{N}(\text{vec}(V) \mid 0, \tau I), \qquad \text{prior on clf. head}$$
$$p(y_{1:n}|x_{1:n}, w, V) = \prod_{i=1}^{n} \text{Cat}\big(y_i|\text{softmax}(V f_w(x_i))\big) \qquad \text{likelihood}$$

Target task MAP estimation. We can fit the above model to the target dataset via a MAP point estimation strategy, finding values of weights w, V that minimize the objective

$$L(w, V) := -\frac{1}{n}\Big[\sum_{i=1}^{n} \log p(y_i|x_i, w, V) + \log p(w) + \log p(V)\Big].$$

Table 1: Possible methods for point estimation of neural network weights for a target task.

Method	also known as	Prior	Init.	Shwartz-Ziv et al. (2022)	Špendl & Pirc (2023)	ours
StdPrior fromScratch	SGD Non-Learned Prior in Shwartz-Ziv et al.	$\mathcal{N}(0, \lambda I)$	random	✓	✓	✗
StdPrior fromImgNet	SGD Transfer Init in Shwartz-Ziv et al.	$\mathcal{N}(0, \lambda I)$	μ	✓	✗	✓
LearnedPriorIso fromImgNet	MAP adaptation in Chelba & Acero	$\mathcal{N}(\mu, \lambda I)$	μ	✗	✗	✓
LearnedPriorLR fromImgNet	SGD Learned Prior in Shwartz-Ziv et al.	$\mathcal{N}(\mu, \lambda \Sigma)$	μ	✓	✓	✓

Experimental Procedures

Fixes to Shwartz-Ziv et al. (2022)'s code. Shwartz-Ziv et al. (2022)'s released code performs inconsistent scaling by λ of the low-rank and diagonal components of the covariance matrix of their informed prior.

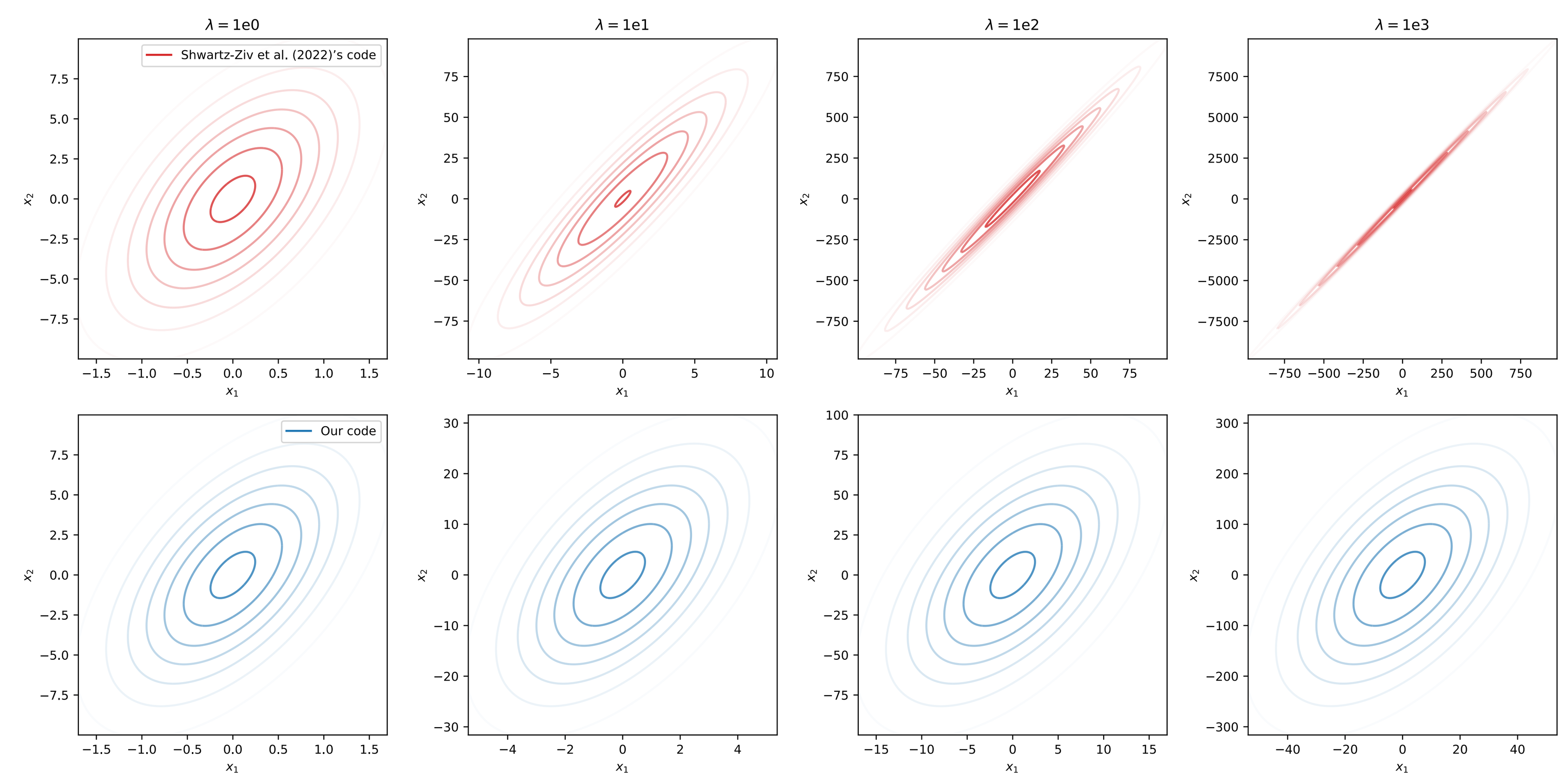

Figure 1: Example scaling the low-rank and diagonal components of a 2D covariance matrix with Shwartz-Ziv et al. (2022)'s code (top row) and our code (bottom row).

Results

Finding 1: Standard transfer learning better than reported in Shwartz-Ziv et al. (2022).

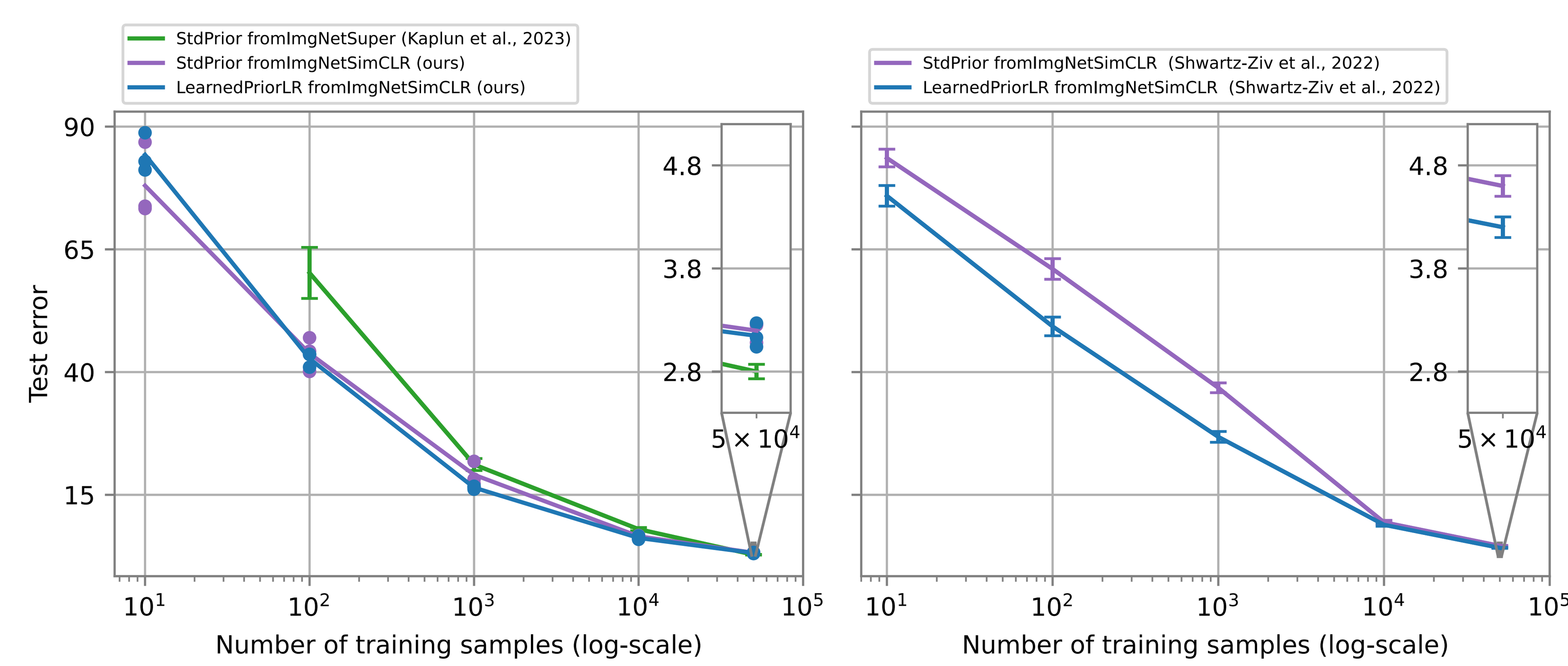

Figure 2: Error rate (lower is better) vs. target train set size on CIFAR-10, for various MAP estimation methods for transfer learning from ImageNet. Left: Our results. Right: Results copied from Shwartz-Ziv et al. (2022) (their Tab. 10). **Takeaway: In our experiments, standard transfer learning (*StdPrior*) does better than previously reported.**

Finding 2: Relative gains of informed priors over standard transfer learning vary across datasets.

Table 2: CIFAR-10 heldout accuracy (higher is better) as target train set size n increases.

Method	$n=$ 10 (1/cl.)	100 (10/cl.)	1000 (100/cl.)	10000 (1k/cl.)	50000 (5k/cl.)
StdPrior fromImgNet	22.0 (13.2-26.7)	56.2 (53.0-59.9)	80.9 (78.2-82.8)	93.4 (93.3-93.7)	96.8 (96.7-96.9)
LearnedPriorIso fromImgNet	20.7 (17.7-23.4)	56.7 (54.8-59.8)	81.9 (81.6-82.1)	94.3 (94.1-94.5)	97.3 (97.2-97.4)
LearnedPriorLR fromImgNet	15.7 (11.3-18.8)	58.7 (56.4-60.6)	83.5 (83.3-83.9)	93.8 (93.4-94.1)	96.9 (96.7-97.0)

Table 3: Oxford Flowers heldout accuracy (higher is better) as target train set size n increases.

Method	$n=$ 102 (1/cl.)	510 (5/cl.)	1020 (10/cl.)
StdPrior fromImgNet	31.1 (12.6-40.6)	78.9 (78.4-79.3)	87.9 (87.4-88.1)
LearnedPriorIso fromImgNet	19.2 (9.0-34.8)	79.1 (78.5-79.4)	88.6 (88.2-89.1)
LearnedPriorLR fromImgNet	28.8 (10.6-39.2)	77.9 (74.8-79.4)	88.4 (88.2-88.7)

Table 4: Oxford-IIIT Pets heldout accuracy (higher is better) as target train set size n increases.

Method	$n=$ 37 (1/cl.)	370 (10/cl.)	3441(93/cl.)
StdPrior fromImgNet	17.3 (14.5-20.0)	54.8 (53.2-57.5)	86.4 (86.0-86.6)
LearnedPriorIso fromImgNet	7.0 (5.4- 8.5)	55.5 (53.3-57.7)	86.4 (85.4-87.0)
LearnedPriorLR fromImgNet	6.6 (6.2- 7.3)	57.4 (56.2-58.2)	86.7 (85.0-87.8)

Table 5: FGVC-Aircraft heldout accuracy (higher is better) as target train set size n increases.

Method	$n=$ 100 (1/cl.)	500 (5/cl.)	1000 (10/cl.)	5000 (50/cl.)
StdPrior fromImgNet	2.6 (1.3-4.6)	22.5 (21.4-24.4)	40.6 (39.8-41.8)	85.7 (85.4-86.0)
LearnedPriorIso fromImgNet	3.5 (2.8-4.5)	25.9 (24.0-27.3)	51.3 (50.0-52.6)	85.9 (85.8-86.0)
LearnedPriorLR fromImgNet	3.8 (3.5-4.2)	24.5 (23.7-25.3)	50.9 (50.4-51.9)	84.2 (83.5-85.3)

Table 6: HAM10000 heldout AUROC (higher is better, macro-averaged across classes) as target train set size n increases.

Method	$n=$ 100	1000
StdPrior fromImgNet	78.1 (75.0-82.8)	85.6 (85.0-86.3)
LearnedPriorIso fromImgNet	78.0 (75.0-82.8)	86.5 (85.5-87.4)
LearnedPriorLR fromImgNet	78.7 (74.9-82.7)	86.6 (85.0-87.5)

Discussion and Conclusion

What transfer learning method is recommended? Based off our experiments, when point estimating neural network weights we recommend both *StdPrior* and *LearnedPriorIso* for their simplicity and performance.

References

Gal Kaplun, Andrey Gurevich, Tal Swisa, Mazor David, Shai Shalev-Shwartz, and Eran Malach. SubTuning: Efficient Finetuning for Multi-Task Learning. *arXiv preprint arXiv:2302.06354*, 2023.

Ravid Shwartz-Ziv, Micah Goldblum, Hossein Souri, Sanyam Kapoor, Chen Zhu, Yann LeCun, and Andrew G. Wilson. Pre-Train Your Loss: Easy Bayesian Transfer Learning with Informative Priors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[*]Equal contribution

[1]Code: **github.com/tufts-ml/bdl-transfer-learning**

ML Reproducibility Challenge
Conference on Neural Information Processing Systems

Ethan.Harvey@tufts.edu