Learning the Regularization Strength for Deep Fine-Tuning via a Data-Emphasized Variational Objective

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

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

Introduction

A number of popular transfer learning methods rely on grid search to select regularization hyperparameters that control over-fitting. This grid search requirement has several key disadvantages:

1) the search is computationally expensive,
2) requires carving out a validation set that reduces the size of available data for model training,
3) and requires practitioners to specify candidate values.

In this paper, we propose an alternative to grid search: directly learning regularization hyperparameters on the full training set via model selection techniques based on the *evidence lower bound* ("ELBo") objective from variational methods.

Background

Deep learning view.

$$L(\theta, \lambda^{-1}) := \frac{1}{N}\left(\sum_{i=1}^{N} \ell(y_i, f_\theta(x_i)) + \frac{\lambda^{-1}}{2}||\theta||_2^2\right)$$

$$\theta_{t+1} \leftarrow \theta_t - \eta \nabla_{\theta_t} L(\theta_t, \lambda^{-1})$$

Grid search learning rate η and regularization strength λ^{-1}.

Methods

Bayesian view.

$$L_{\text{DE-ELBo}}(\theta, \sigma, \lambda) := -\kappa \mathbb{E}_{q(\theta)}\left[\log \sum_{i=1}^{N} p(y_i|\theta)\right] + \mathbb{KL}(q(\theta)\|p(\theta|\lambda))$$

$$\begin{aligned}
\lambda_t^* &\leftarrow \frac{1}{D}\Big[\sigma_t^2 \operatorname{Tr}(\Sigma_p^{-1}) + (\mu_p - \theta_t)^T \Sigma_p^{-1}(\mu_p - \theta_t)\Big] & \\
\theta_{t+1} &\leftarrow \theta_t - \eta \nabla_{\theta_t} L_{\text{DE-ELBo}}(\theta_t, \sigma_t, \lambda_t^*) & \text{approx. w/ 1 sample} \\
\sigma_{t+1} &\leftarrow \sigma_t - \eta \nabla_{\sigma_t} L_{\text{DE-ELBo}}(\theta_t, \sigma_t, \lambda_t^*) & \text{approx. w/ 1 sample}
\end{aligned}$$

Grid search learning rate η.

Need for validation set and grid search. Selecting λ^{-1} to directly minimize $L(\theta, \lambda^{-1})$ on the training set alone is not a coherent way to guard against over-fitting.
Backbone prior mean/covariance. Several recent transfer learning approaches correspond to specific settings of the backbone mean and covariance μ_p, Σ_p.

Table 1: Mean and covariance of backbone weights w for several transfer learning approaches.

Method	$p(w)$	Init.
L2-zero	$\mathcal{N}(0, \lambda I)$	μ
L2-SP	$\mathcal{N}(\mu, \lambda I)$	μ
PTYL	$\mathcal{N}(\mu, \lambda \Sigma)$	μ

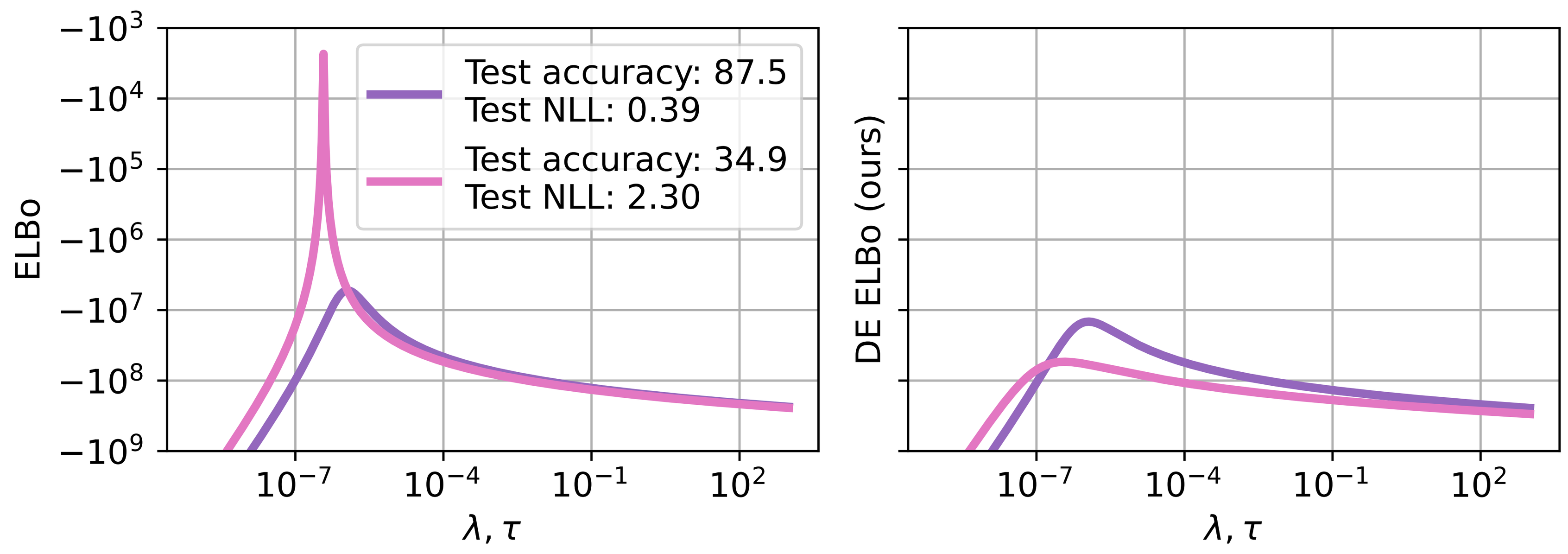

Figure 1: Model selection comparison between the ELBo (left) and our *data-emphasized ELBo* (DE ELBo) for two ResNet-50s trained on CIFAR-10 $N = 1000$. For both models, we fix the estimated posterior q and vary λ, τ. **Takeaway: Without enough training data or with too many model parameters, the ELBo has a preference for simpler models.**

Table 2: Computational time comparison between methods for transfer learning with informative priors using grid search and using our *data-emphasized ELBo* (DE ELBo) for CIFAR-10 at $N = 50000$. Runtime measured on one NVIDIA A100 40GB PCIe GPU.

Model	Method	Avg. SGD runtime	lr search space	λ, τ search space	Total GS time
L2-SP	MAP + GS	39 mins. 11 secs.	4	6	16 hrs. 15 mins.
	DE ELBo	39 mins. 0 secs.	4	n/a	2 hrs. 36 mins.
PTYL	MAP + GS	37 mins. 15 secs.	4	60	149 hrs. 36 mins.
	DE ELBo	40 mins. 0 secs.	4	n/a	2 hrs. 39 mins.

Table 3: Accuracy on CIFAR-10 test set for different probabilistic models and methods. We report mean (min-max) over 3 separately-sampled training sets.

Model	Method	$N =$ 100 (10/cl.)	1000 (100/cl.)	10000 (1k/cl.)	50000 (5k/cl.)
L2-zero	MAP + GS	67.7 (66.0-68.6)	87.8 (87.5-88.4)	95.0 (94.4-95.5)	97.2 (97.1-97.2)
	DE ELBo	60.9 (58.9-63.1)	87.2 (87.0-87.4)	91.2 (90.7-92.0)	93.2 (93.0-93.3)
L2-SP	MAP + GS	68.1 (66.7-68.9)	87.3 (87.2-87.3)	95.3 (95.1-95.7)	97.1 (97.0-97.1)
	DE ELBo	70.6 (68.7-72.7)	87.2 (86.8-87.4)	95.0 (94.8-95.2)	96.8 (96.7-96.9)
PTYL	MAP + GS	67.5 (65.7-68.4)	87.9 (86.9-89.2)	95.2 (95.0-95.4)	97.3 (97.3-97.3)
	DE ELBo	70.6 (68.7-72.6)	87.2 (86.9-87.6)	95.1 (94.9-95.4)	96.9 (96.8-96.9)

Outlook. Our proposed approach saves practitioners time by learning an optimal regularization strength without need for expensive grid search. We hope our data-emphasized ELBo for efficient hyperparameter tuning may eventually prove useful across a wide array of classifier tasks beyond transfer learning, such as semi-supervised learning, few-shot learning, continual learning, and beyond.

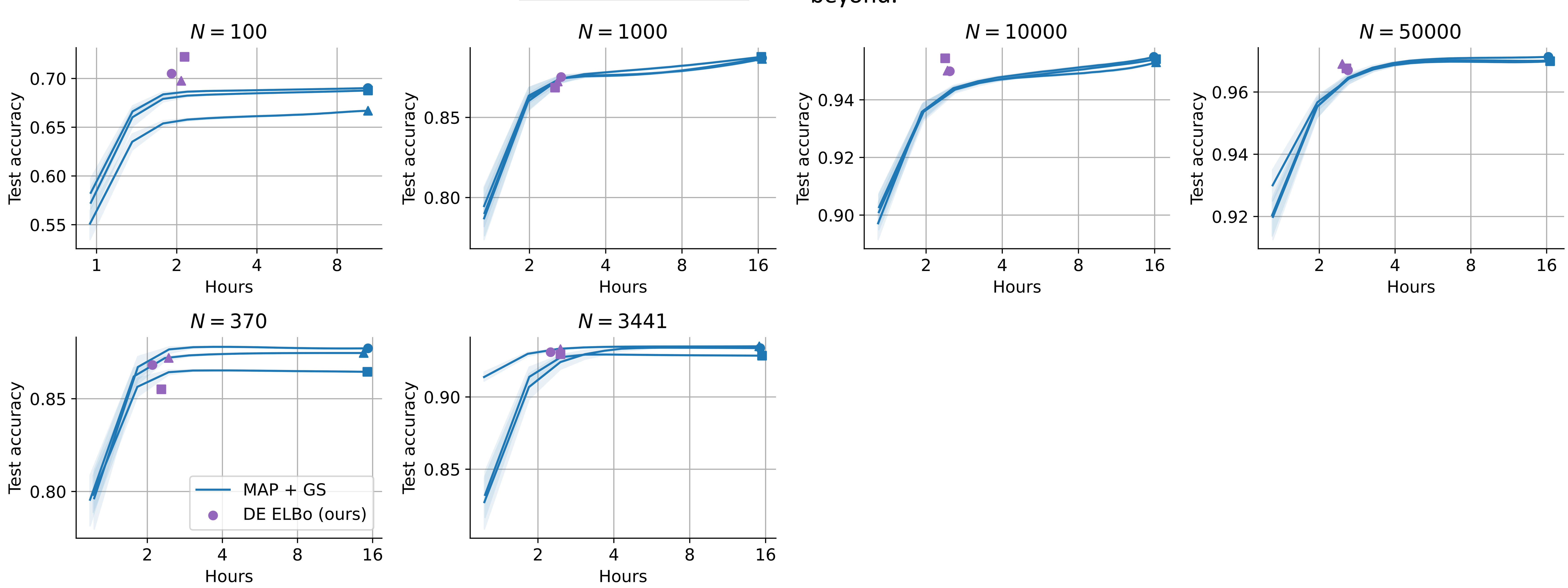

Figure 2: Test-set accuracy on CIFAR-10 (top row) and Oxford-IIIT Pet (bottom row) over training time for L2-SP with *MAP + grid search (GS)* and our *data-emphasized ELBo* (DE ELBo). We run each method on 3 separate training sets of size N (3 different marker styles)**. Takeaway: Our DE ELBo achieves as good or better performance at small dataset sizes and similar performance at large dataset sizes with far less compute time.** To make the blue curves, we did the full grid search once (markers). Then, at each given shorter compute time, we subsampled a fraction of all hyperparameter configurations with that runtime and chose the best via validation NLL. Averaging this over 500 subsamples at each runtime created each blue line.

*Both authors made lead author level contributions

Open-source code: **github.com/tufts-ml/data-emphasized-ELBo**

Ethan.Harvey@tufts.edu