Deep Reference Priors:

What is the best way to pretrain a model?

Yansong Gao*, Rahul Ramesh* and Pratik Chaudhari

University of Pennsylvania, *equal contribution

Best way to pre-train models?

Exploiting the extra data is a powerful way to reduce the number of training samples required to learn a given task.

Best way to pre-train models?

Exploiting the extra data is a powerful way to reduce the number of training samples required to learn a given task.

We think of two types of the extra data:

- unlabelled data from the same task, e.g., semi-supervised learning;

Best way to pre-train models?

Exploiting the extra data is a powerful way to reduce the number of training samples required to learn a given task.

We think of two types of the extra data:

- unlabelled data from the same task, e.g., semi-supervised learning;
- labeled data from other related tasks, e.g., transfer, multi-task, and meta-learning.

Best way to pre-train models?

Exploiting the extra data is a powerful way to reduce the number of training samples required to learn a given task.

We think of two types of the extra data:

- unlabelled data from the same task, e.g., semi-supervised learning;
- labeled data from other related tasks, e.g., transfer, multi-task, and meta-learning.

  If we have a pool of data—be it labeled or unlabeled, from the same task, or from another related task—what is the optimal way to pre-train a model?

Best way to pre-train models?

Exploiting the extra data is a powerful way to reduce the number of training samples required to learn a given task.

We think of two types of the extra data:

- unlabelled data from the same task, e.g., semi-supervised learning;
- labeled data from other related tasks, e.g., transfer, multi-task, and meta-learning.

  If we have a pool of data—be it labeled or unlabeled, from the same task, or from another related task—what is the optimal way to pre-train a model?

We formalize this question using the theory of reference priors [1].

Notations

- $w \in \mathbb{R}^p$: the weights of a probabilistic model.

Notations

- $w \in \mathbb{R}^p$: the weights of a probabilistic model.
- A prior on weights: $w \sim \pi(w)$.

Notations

- $w \in \mathbb{R}^p$: the weights of a probabilistic model.
- A prior on weights: $w \sim \pi(w)$.
- Consider a dataset $z^n = (x^n, y^n)$;
  $x^n = (x_1, \ldots, x_n)$ denotes all inputs (e.g., images);
  $y^n = (y_1, \ldots, y_n)$ denotes labels.

Notations

- $w \in \mathbb{R}^p$: the weights of a probabilistic model.
- A prior on weights: $w \sim \pi(w)$.
- Consider a dataset $z^n = (x^n, y^n)$;
  $x^n = (x_1, \ldots, x_n)$ denotes all inputs (e.g., images);
  $y^n = (y_1, \ldots, y_n)$ denotes labels.
- Bayes rule

$$p(w \mid z^n) \propto p(z^n \mid w)\pi(w)$$

Notations

- $w \in \mathbb{R}^p$: the weights of a probabilistic model.
- A prior on weights: $w \sim \pi(w)$.
- Consider a dataset $z^n = (x^n, y^n)$;
  $x^n = (x_1, \ldots, x_n)$ denotes all inputs (e.g., images);
  $y^n = (y_1, \ldots, y_n)$ denotes labels.
- Bayes rule

$$p(w \mid z^n) \propto p(z^n \mid w)\pi(w)$$

To make the choice of priors less subjective Bernado [1] suggested that uninformative priors should maximize some divergence between posterior $p(w \mid z^n)$ and prior $\pi(w)$.

Reference Priors: Uninformative Bayesian priors

KL divergence measures the difference between $p(w \mid z^n)$ and $\pi(w)$ as

$$\mathrm{KL}(p(w \mid z^n) \mid\mid \pi(w)) = \int \mathrm{d}w \, p(w \mid z^n) \log \frac{p(w \mid z^n)}{\pi(w)}.$$

Reference Priors: Uninformative Bayesian priors

KL divergence measures the difference between $p(w \mid z^n)$ and $\pi(w)$ as

$$\mathrm{KL}(p(w \mid z^n) \,||\, \pi(w)) = \int \mathrm{d}w \; p(w \mid z^n) \log \frac{p(w \mid z^n)}{\pi(w)}.$$

Since we do not know the data *a priori* while picking the prior, we should maximize the average KL divergence over the data distribution $p(z^n)$.

$$\begin{aligned} \pi_n^* &= \arg\max_{\pi} \int \mathrm{d}z^n \; p(z^n) \cdot \mathrm{KL}(p(w \mid z^n) \,||\, \pi(w)) \\ &= \arg\max_{\pi} I_\pi(w; z^n) \end{aligned} \tag{1}$$

Reference Priors: Uninformative Bayesian priors

KL divergence measures the difference between $p(w \mid z^n)$ and $\pi(w)$ as

$$\mathrm{KL}(p(w \mid z^n) \,||\, \pi(w)) = \int \mathrm{d}w \; p(w \mid z^n) \log \frac{p(w \mid z^n)}{\pi(w)}.$$

Since we do not know the data *a priori* while picking the prior, we should maximize the average KL divergence over the data distribution $p(z^n)$.

$$\begin{aligned} \pi_n^* &= \arg\max_{\pi} \int \mathrm{d}z^n \; p(z^n) \cdot \mathrm{KL}(p(w \mid z^n) \,||\, \pi(w)) \\ &= \arg\max_{\pi} I_\pi(w; z^n) \end{aligned} \tag{1}$$

π_n^* is the *n*-reference prior proposed by Bernardo in 1979.

Reference Priors

π_n^* gives the unseen ground truth labels the maximum capacity to dominate the posterior [1].

Reference Priors

π_n^* gives the unseen ground truth labels the maximum capacity to dominate the posterior [1].

Reference priors are supported on discrete sets [3].

Reference Priors

π_n^* gives the unseen ground truth labels the maximum capacity to dominate the posterior [1].

Reference priors are supported on discrete sets [3].

Most importantly, reference priors select diverse parts of the hypothesis space [2].

Reference priors select diverse parts of the hypothesis space

Consider the estimation of the bias of a coin $w \in [0, 1]$ using n trials. z^n denotes the number of heads we observe. The atoms of π_n^* have diverse inference on the n samples.

Reference priors select diverse parts of the hypothesis space

Consider the estimation of the bias of a coin $w \in [0, 1]$ using n trials. z^n denotes the number of heads we observe. The atoms of π_n^* have diverse inference on the n samples.

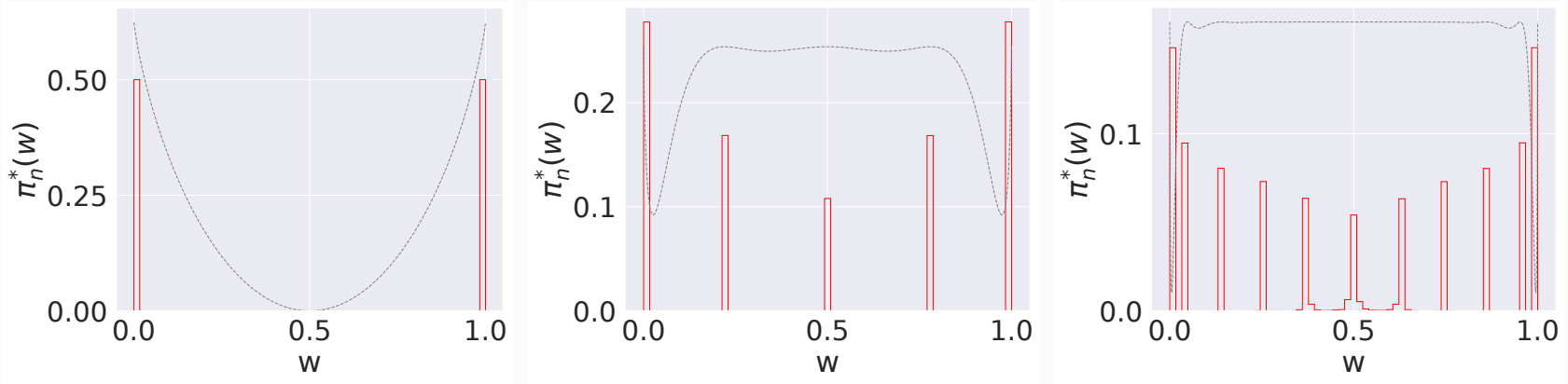

Figure 1: π_n^* for the coin-tossing model for $n = 1, 10, 50$ (from left to right). π_n^* is discrete for $n < \infty$. Atoms of the prior are maximally different from each other, e.g., for $n = 1$, they are on opposite corners of the parameter space.

Reference priors select diverse parts of the hypothesis space

Consider the estimation of the bias of a coin $w \in [0, 1]$ using n trials. z^n denotes the number of heads we observe. The atoms of π_n^* have diverse inference on the n samples.

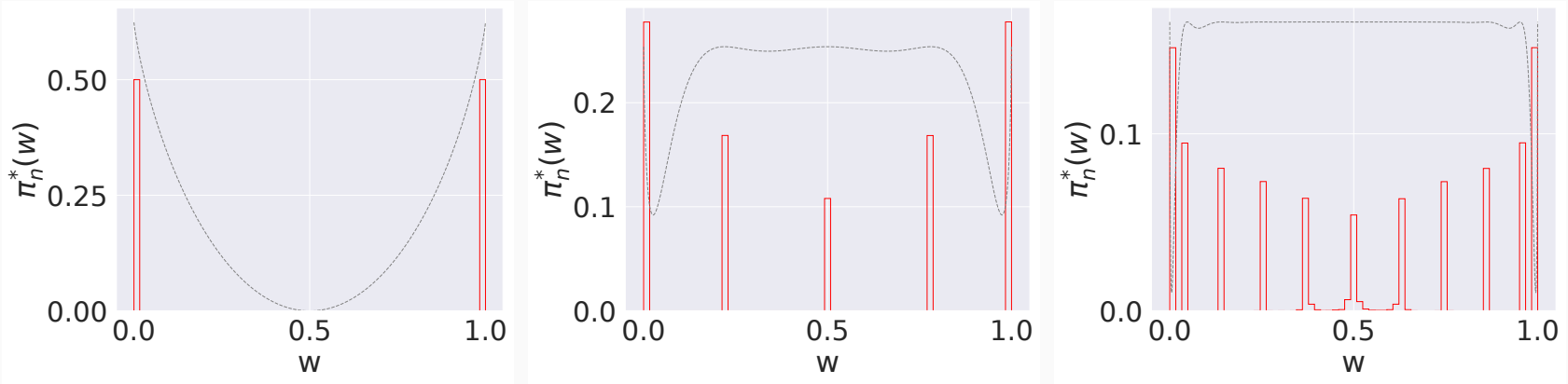

Figure 1: π_n^* for the coin-tossing model for $n = 1, 10, 50$ (from left to right). π_n^* is discrete for $n < \infty$. Atoms of the prior are maximally different from each other, e.g., for $n = 1$, they are on opposite corners of the parameter space.

This ability of the prior to select a small set of representative models is extremely useful for training deep networks with few data and it was our primary motivation.

Application: pre-training models with unlabeled data

Given a few labeled data and a pool of unlabeled data, we combine reference prior and the likelihood function into a single objective

Application: pre-training models with unlabeled data

Given a few labeled data and a pool of unlabeled data, we combine reference prior and the likelihood function into a single objective

$$\max_{\pi} \gamma I_{\pi}(w; y^u, x^u) + \mathbb{E}_{w \sim \pi} \left[\log p(y^n \mid x^n, w) \right], \tag{2}$$

where γ is a hyper parameter, x^n, y^n are labeled samples, x^u is an unlabeled sample.

Experiments on semi-supervised learning

Accuracy (%) of semi-supervised learning methods on CIFAR-10

Method	#Samples				
	50	100	250	500	1000
PiModel	-	-	46.58	58.18	68.47
PseudoLab	-	-	50.02	59.45	69.09
Mixup	-	-	52.57	63.86	74.28
VAT	-	-	63.97	73.89	81.32
Mean Teacher	-	-	52.68	57.99	82.68
MixMatch	64.21	80.29	88.91	90.35	92.25
FixMatch	86.19	90.1	94.9	94.0	94.3
SelfMatch	93.19 (40)	-	95.13	-	-
FlexMatch	95.0	-	95.2	-	-
Deep Reference Prior	85.5	88.5	92.1	93.1	93.5

Discussion

Reference priors have a unique ability to select diverse parts of the model space.

Discussion

Reference priors have a unique ability to select diverse parts of the model space.

Reference priors provide an information-theoretically optimal way to pre-train models using unlabeled data.

Discussion

Reference priors have a unique ability to select diverse parts of the model space.

Reference priors provide an information-theoretically optimal way to pre-train models using unlabeled data.

This is the first implementation of reference priors for deep networks that maintains their characteristic feature, namely that they are supported on a discrete set.

Discussion

Reference priors have a unique ability to select diverse parts of the model space.

Reference priors provide an information-theoretically optimal way to pre-train models using unlabeled data.

This is the first implementation of reference priors for deep networks that maintains their characteristic feature, namely that they are supported on a discrete set.

A two-stage reference prior can be used for transfer learning.

Github: github.com/grasp-lyrl/deep_reference_priors

Arxiv: arxiv.org/abs/2202.00187

Visit our poster (#711) at Hall E - Wednesday 6:30pm to 8:30pm

References

[1] Jose M Bernardo. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):113–128, 1979.

[2] Henry H Mattingly, Mark K Transtrum, Michael C Abbott, and Benjamin B Machta. Maximizing the information learned from finite data selects a simple model. *Proceedings of the National Academy of Sciences*, 115(8):1760–1765, 2018.

[3] Zhongxin Zhang. *Discrete noninformative priors*. PhD thesis, Yale University, 1994.