A statistical perspective on distillation

Aditya Krishna Menon

Ankit Singh Rawat

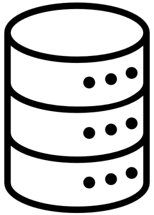
Seungyeon Kim

Sashank Reddi

Sanjiv Kumar

Google Research

Supervised learning: standard



Labelled data

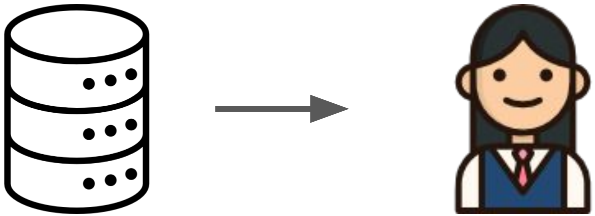
$$(x_1,\, y_1)$$

$$\vdots$$

$$(x_n,\, y_n)$$

Icon made by Freepik from www.flaticon.com

Supervised learning: standard



Labelled data

Learner

$$(x_1,\, y_1)$$
$$\vdots$$
$$(x_n,\, y_n)$$

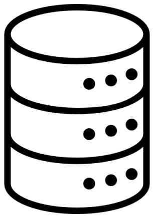
$$\min_f \frac{1}{N} \sum_{n \in [N]} e_{y_n}^\top \ell(f(x_n))$$

Loss vector

One-hot encoding of labels

$$[\ell(1, f(x_n)), \ldots, \ell(L, f(x_n))]$$

Supervised learning: standard



Labelled data

(x_1, y_1)

$\vdots$

(x_n, y_n)

Learner

$$\min_f \frac{1}{N} \sum_{n \in [N]} e_{y_n}^\top \ell(f(x_n))$$

Predictor

$f : \mathcal{X} \to \mathbb{R}^L$

Loss vector

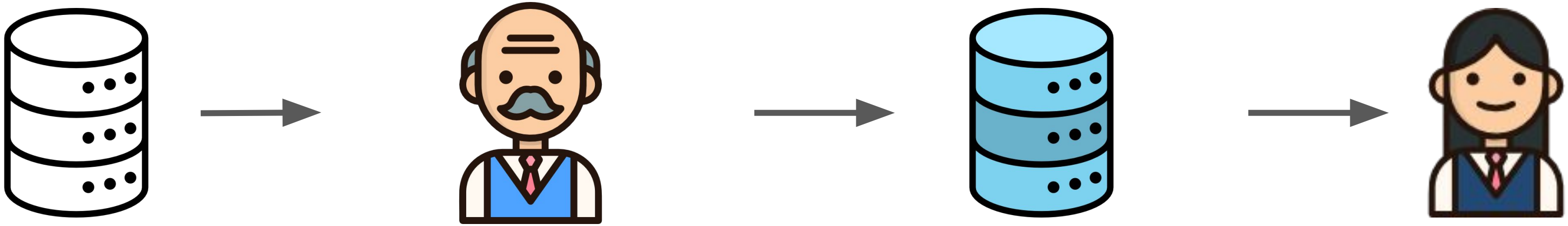
One-hot encoding of labels

$[\ell(1, f(x_n)), \ldots, \ell(L, f(x_n))]$

Google

Icon made by Freepik from www.flaticon.com

Supervised learning: distillation



Labelled data

$$(x_1, y_1)$$
$$\vdots$$
$$(x_n, y_n)$$

"Teacher" model

$$\min_f \frac{1}{N} \sum_{n \in [N]} e_{y_n}^\top \ell(p^t(x_n))$$

Distribution over labels

Teacher-labelled data

$$(x_1, p^t(x_1))$$
$$\vdots$$
$$(x_n, p^t(x_n))$$

"Student" model

$$\min_f \frac{1}{N} \sum_{n \in [N]} p^t(x_n)^\top \ell(f(x_n))$$

Replaces one-hot labels

Successes of distillation

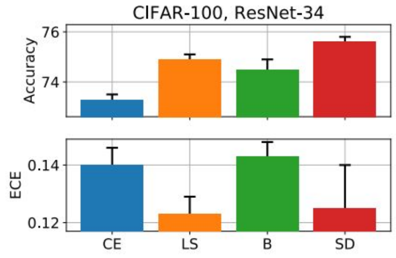
Empirically, distillation has demonstrated wide success:

System	Test Frame Accuracy
Baseline	58.9%
10xEnsemble	61.1%
Distilled Single model	60.8%

Hinton et al., 2015

Network	Teacher	BAN
DenseNet-112-33	18.25	**16.95**
DenseNet-90-60	17.69	**16.69**
DenseNet-80-80	17.16	**16.36**
DenseNet-80-120	16.87	**16.00**

Furlanello et al., 2018

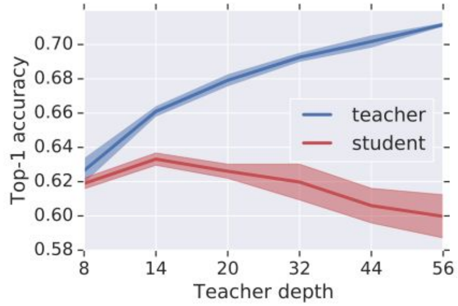


Zhang et al., 2020

But *why* does distillation help?

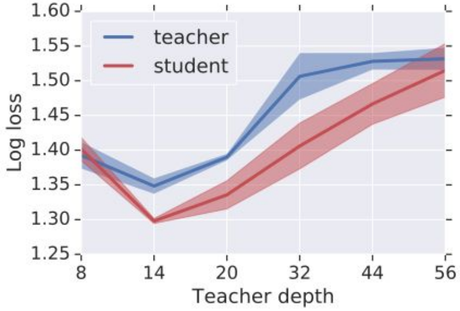
Summary of our work

Q: Why does distillation help?

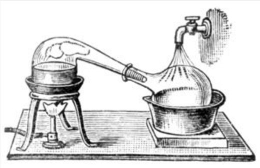


A: We provide a **statistical perspective**:

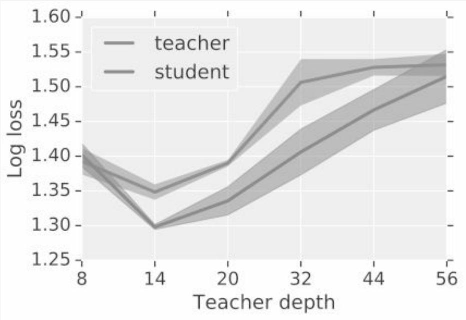
- Teacher approximates Bayes probabilities
- Exact Bayes probabilities ⟶ reduce variance of objective
- Approximate Bayes probabilities ⟶ bias-variance tradeoff



Distilling from a Bayes teacher



Distilling from an imperfect teacher



Applications

$$\widetilde{\mathrm{PD}}(f) \propto \sum_{i\in S, j\in S-\{i\}} p^{\mathrm{t}}(x_i)\cdot(1-p^{\mathrm{t}}(x_j))\cdot[\![f(x_i) < f(x_j)]\!].$$

$$z(x_n) \doteq \log\left(\sum_{y'\in[L]} \alpha_{y'}(x_n)\cdot e^{f_{y'}(x_n)}\right)$$

$$\alpha_{y'}(x_n) \doteq 1 - [\![y' \neq y_n]\!]\cdot p^{\mathrm{t}}(y' \mid x_n).$$

Google

Statistical learning setup

Suppose our training samples $(x, y) \sim \mathbb{P}$

Underlying "Bayes" distribution

Student goal: minimise the population risk, i.e., expected loss:

$$R(f) = \mathop{\mathbb{E}}_{x} \mathop{\mathbb{E}}_{y|x} [\ell(y, f(x))]$$

Statistical learning setup

Suppose our training samples $(x, y) \sim \mathbb{P}$

Underlying "Bayes" distribution

Student goal: minimise the population risk, i.e., expected loss:

$$R(f) = \mathbb{E}_{x} \mathbb{E}_{y|x} [\ell(y, f(x))]$$

$$= \mathbb{E}_{x} [p^*(x)^\top \ell(f(x))]$$

($\mathbf{P}^*(y = 1 \mid x), \ldots, \mathbf{P}^*(y = L \mid x)$)

Inherently smooths loss by Bayes-probabilities!

Google

“Bayes teacher” distillation

“Bayes-distilled” training loss:

$$\hat{R}_*(f) = \frac{1}{N} \sum_{n \in [N]} p^*(x_n)^\top \ell(f(x_n))$$

Predictions from a “Bayes teacher”

Like standard empirical loss, $\mathbb{E}[\hat{R}_*(f)] = R(f)$

But has an important advantage...

Google

Why does "Bayes distillation" help?

Bayes-distilled loss lowers variance over empirical loss:

$$\mathop{\mathbb{V}}_{S\sim\mathbb{P}^N}[\hat{R}_*(f)] \leq \mathop{\mathbb{V}}_{S\sim\mathbb{P}^N}[\hat{R}(f)]$$

Variance over draws of training set

Lower variance → better generalisation bound:

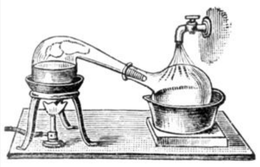
$$R(\mathbf{f}) \leq \hat{R}_*(\mathbf{f}; S) + \mathcal{O}\left(\sqrt{\mathbb{V}_N^*(\mathbf{f})/N} \cdot \sqrt{\log\left(\mathcal{M}_N^*/\delta\right)} + \log\left(\mathcal{M}_N^*/\delta\right)/N\right),$$

"Bayes distillation" can improve generalisation!

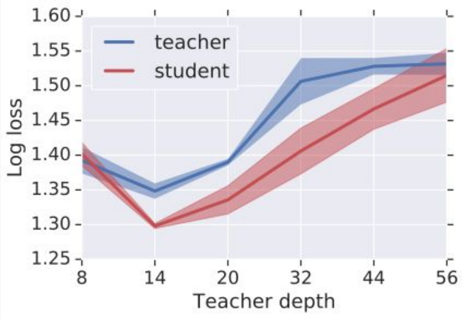
See paper for more!

Google

Distilling from a Bayes teacher



Distilling from an imperfect teacher



Applications

$$\widetilde{\mathrm{PD}}(f) \propto \sum_{i\in S, j\in S-\{i\}} p^{\mathrm{t}}(x_i)\cdot(1-p^{\mathrm{t}}(x_j))\cdot[\![f(x_i)<f(x_j)]\!].$$

$$z(x_n) \doteq \log\left(\sum_{y'\in[L]} \alpha_{y'}(x_n)\cdot e^{f_{y'}(x_n)}\right)$$

$$\alpha_{y'}(x_n) \doteq 1 - [\![y'\neq y_n]\!]\cdot p^{\mathrm{t}}(y'\mid x_n).$$

Distilling from imperfect teacher

"Bayes teacher" helps; but what about other teachers?

Better approximation of *p** ⟶ better generalisation:

$$\mathbb{E}\left[(\tilde{R}(\mathbf{f};S) - R(\mathbf{f}))^2\right] \leq \frac{1}{N} \cdot \mathbb{V}\left[\mathbf{p}^{\mathrm{t}}(x)^\top \ell(\mathbf{f}(x))\right]$$
$$+\mathcal{O}(\mathbb{E}\|\mathbf{p}^{\mathrm{t}}(x) - \mathbf{p}^*(x)\|_2^2)$$

Distilling from imperfect teacher

"Bayes teacher" helps; but what about other teachers?

Better approximation of *p** ⟶ better generalisation:

$$\mathbb{E}\left[(\tilde{R}(\mathbf{f}; S) - R(\mathbf{f}))^2\right] \leq \frac{1}{N} \cdot \mathbb{V}\left[\mathbf{p}^{\mathrm{t}}(x)^\top \ell(\mathbf{f}(x))\right]$$
$$+ \mathcal{O}\left(\|\mathbb{E}[\mathbf{p}^{\mathrm{t}}(x)] - \mathbf{p}^*(x)\|_2^2 + \mathbb{V}\left[\mathbf{p}^{\mathrm{t}}(x)\right] \right).$$

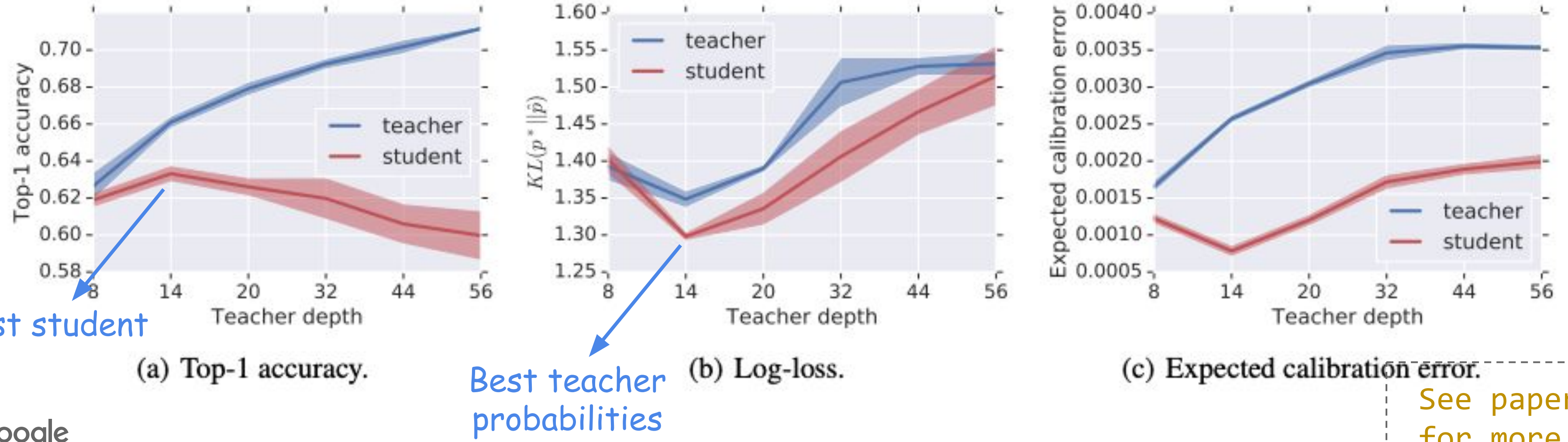
Bias-variance tradeoff
for modelling p*

Implications: bias-variance bound

Bound is **not** on teacher accuracy

$$\mathbb{E}[(\tilde{R}(\mathbf{f};S) - R(\mathbf{f}))^2] \le \frac{1}{N} \cdot \mathbb{V}\left[\mathbf{p}^{\mathrm{t}}(x)^\top \ell(\mathbf{f}(x))\right] + \mathcal{O}\left(\|\mathbb{E}[\mathbf{p}^{\mathrm{t}}(x)] - \mathbf{p}^*(x)\|_2^2 + \mathbb{V}\left[\mathbf{p}^{\mathrm{t}}(x)\right]\right).$$

- Teacher can be accurate but poorly calibrated
- cf. finding that accurate teachers may distill poorer [Muller et al., 2019]

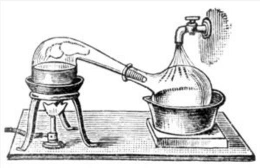


(a) Top-1 accuracy. (b) Log-loss. (c) Expected calibration error.

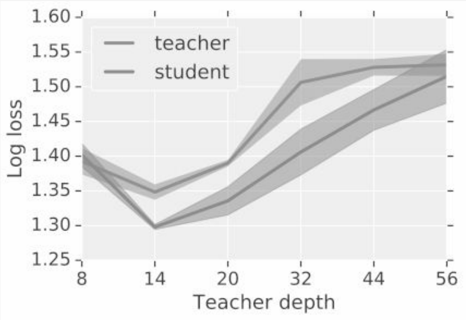
See paper for more!

Google

Distilling from a Bayes teacher



Distilling from an imperfect teacher



Applications

$$\widetilde{\mathrm{PD}}(f) \propto \sum_{i\in S, j\in S-\{i\}} p^{\mathrm{t}}(x_i)\cdot(1-p^{\mathrm{t}}(x_j))\cdot[\![f(x_i) < f(x_j)]\!].$$

$$z(x_n) \doteq \log\left(\sum\nolimits_{y'\in[L]} \alpha_{y'}(x_n)\cdot e^{f_{y'}(x_n)}\right)$$
$$\alpha_{y'}(x_n) \doteq 1 - [\![y' \neq y_n]\!]\cdot p^{\mathrm{t}}(y' \mid x_n).$$

Applications of statistical view

Use teacher's estimates in place of Bayes probabilities $p^*(x)$

Example: bipartite ranking, where goal is to minimise

$$\mathrm{PD}(f) = \mathop{\mathbb{P}}_{x|y=+1} \mathop{\mathbb{P}}_{x|y=-1} (f(x) < f(x'))$$

Distilled bipartite risk:

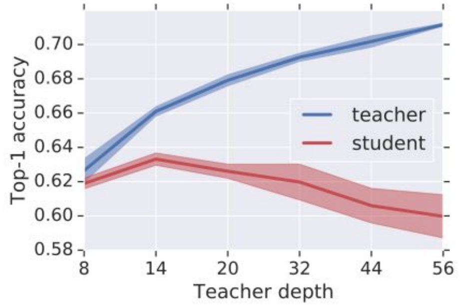
$$\widetilde{\mathrm{PD}}(f) \propto \sum_{i \in S, j \in S - \{i\}} p^{\mathrm{t}}(x_i) \cdot (1 - p^{\mathrm{t}}(x_j)) \cdot [\![f(x_i) < f(x_j)]\!].$$

Additional weighting on "negatives"

See paper for more!

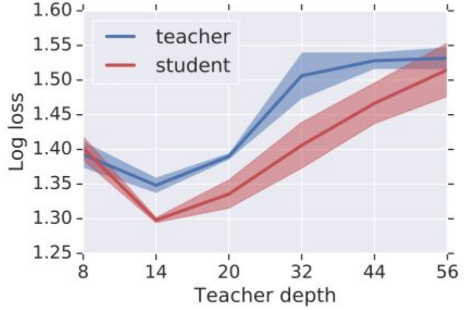
Summary of our work

Q: **Why does distillation help?**



A: We provide a **statistical perspective**:

- Teacher approximates Bayes probabilities
- Exact Bayes probabilities ⟶ reduce variance of objective
- Approximate Bayes probabilities ⟶ bias-variance tradeoff



See paper for more!