Knowledge Distillation as Semiparametric Inference

Lester Mackey*

April 28, 2021

Collaborators: Tri Dao†, Govinda M. Kamath*, Vasilis Syrgkanis*, Ruishan Liu†, and Nicolo Fusi*

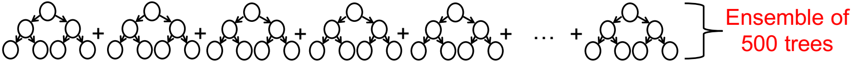
Microsoft Research*, Stanford University†

Knowledge Distillation in a Nutshell

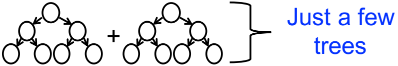
Knowledge Distillation (KD)

[Bucila, Caruana, and Niculescu-Mizil, 2006, Li, Zhao, Huang, and Gong, 2014, Hinton, Vinyals, and Dean, 2015]

1. Train your favorite accurate classifier (called the **teacher**)



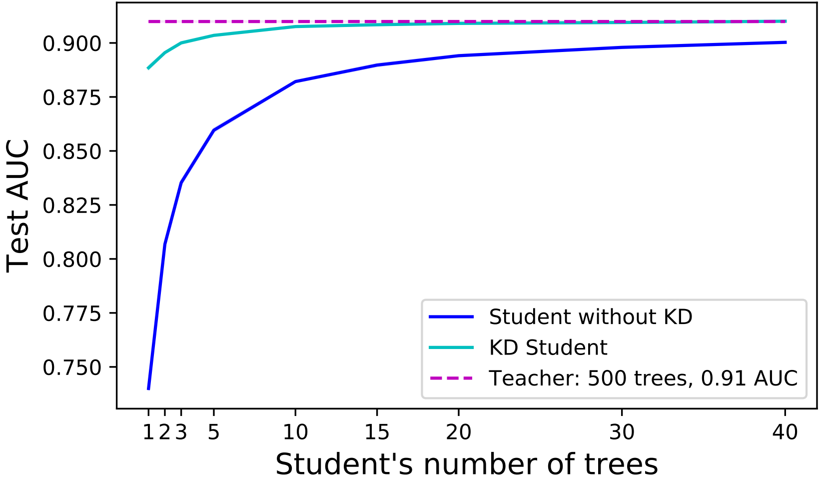
2. Train a simpler model (the **student**) to mimic the teacher's predicted class probabilities



3. That's it: there are only two steps!

Knowledge Distillation (KD) in Action

Task: Predict income level from census data



KD Student: 10 trees $\Rightarrow$ .91 AUC and simpler to deploy

$50\times$ less storage and computation

Knowledge Distillation in a Nutshell

Knowledge Distillation (KD)

[Bucila, Caruana, and Niculescu-Mizil, 2006, Li, Zhao, Huang, and Gong, 2014, Hinton, Vinyals, and Dean, 2015]

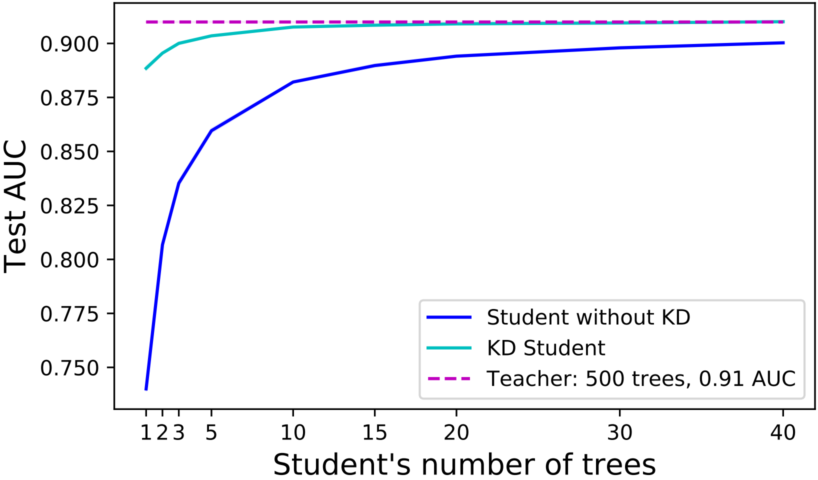
1. Train your favorite accurate classifier (called the **teacher**)
2. Train a simpler model (the **student**) to mimic the teacher's predicted class probabilities

Benefits

1. Simpler student often retains most of the teacher accuracy
   - Reduces test-time computation and storage costs; ideal for resource-constrained devices
2. KD often more accurate than training same student from scratch
3. Same strategy applies to any classifier (be it a random forest or a neural net) and any domain (be it tabular, image, or language)

Knowledge Distillation (KD) in Action

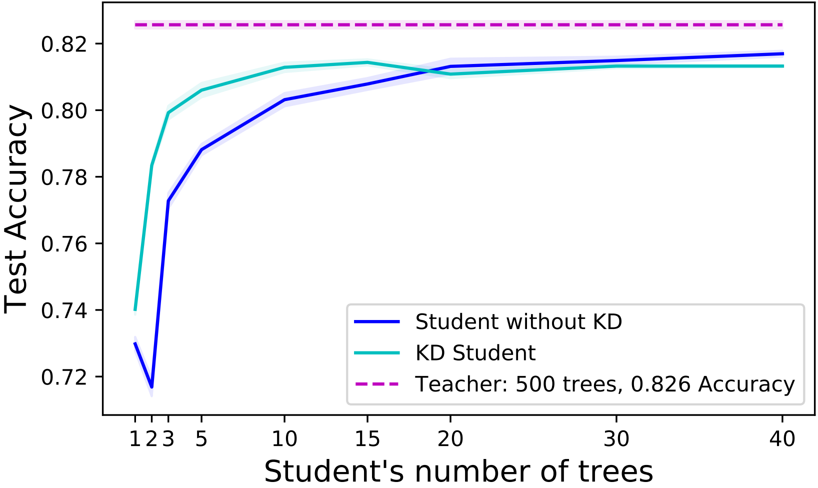
Task: Predict income level from census data



Warning: KD doesn't always work quite this well...

Knowledge Distillation (KD) in Action

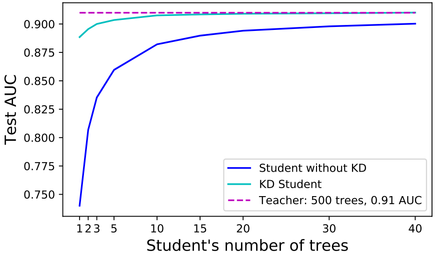
Task: Distinguish ephemeral and evergreen websites

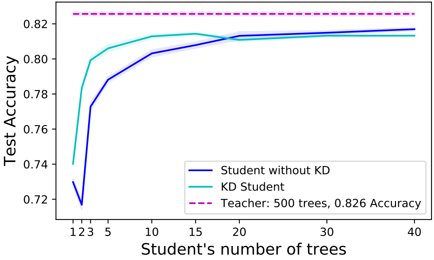


KD Student: 3 trees ⇒ .80 Acc. 40 trees ⇒ .81 Acc.

Underperforms student without KD after 20 trees

Knowledge Distillation (KD) in Action





Questions

1. When should we expect KD to succeed or fail?
2. Can we enhance the performance of KD?

Knowledge Distillation (KD) in a Nutshell

Question: When should KD succeed or fail?

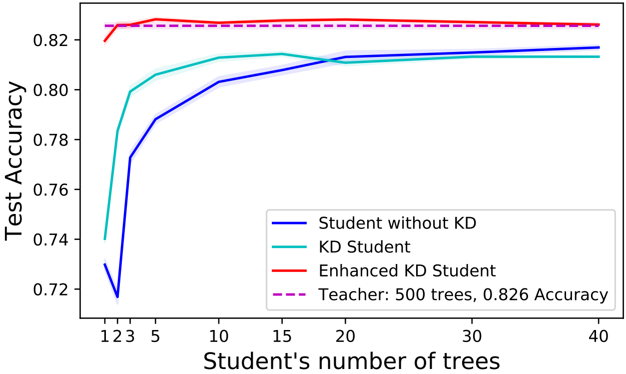
Hypotheses and partial answers

- Probabilities more informative than labels [Hinton, Vinyals, and Dean, 2015]
- Linear students exactly mimic linear teachers [Phuong and Lampert, 2019]
- Students can learn at a faster rate given knowledge of datapoint difficulty (LUPI) [Lopez-Paz, Bottou, Schölkopf, and Vapnik, 2015]
- Regularization for kernel ridge regression [Mobahi, Farajtabar, and Bartlett, 2020]
- Teacher class probabilities $\hat{p}(x)$ are proxies for the true **Bayes class probabilities** $p_0(x) = \mathbb{E}[Y \mid x]$ [Menon, Rawat, Reddi, Kim, and Kumar, 2020]

This talk: Cast KD as learning with nuisance

- **Goal:** fit an accurate, simple student model $\hat{f}$
  - Nuisance: true Bayes class probabilities p_0
  - Plug-in estimate: teacher's predicted class probabilities $\hat{p}$
- Analyze the success and failure modes of KD
- Develop two improvements for enhanced KD performance

Knowledge Distillation (KD) in a Nutshell



This talk: Cast KD as learning with nuisance

- **Goal:** fit an accurate, simple student model $\hat{f}$
  - Nuisance: true Bayes class probabilities p_0
  - Plug-in estimate: teacher's predicted class probabilities $\hat{p}$
- Analyze the success and failure modes of KD
- Develop two improvements for enhanced KD performance

Knowledge Distillation as Learning with Nuisance

Given: n datapoints $z_i = (x_i, y_i)$ drawn independently from $\mathbb{P}$

- Feature vector $x_i \in \mathcal{X}$ and label vector $y_i \in \{e_1, \dots, e_k\}$

Goal: Learn a simple, accurate **student scoring rule** $\hat{f} : \mathcal{X} \to \mathbb{R}^k$

- **Student function class:** $\hat{f} \in \mathcal{F}$
- **Loss function:** $\ell(f(x), p_0(x))$ depending on unknown Bayes class probabilities $p_0(x) = \mathbb{E}[Y \mid x]$ (the nuisance)

Example (Standard KD losses)

- **Squared error logit loss** [Ba and Caruana, 2014]

$$\ell_{\text{se}}(f(x), p(x)) \triangleq \sum_{j\in[k]} \frac{1}{2}(f_j(x) - \log(p_j(x)))^2$$

- **Annealed cross-entropy loss** [Hinton, Vinyals, and Dean, 2015]

$$\ell_\beta(f(x), p(x)) = -\sum_{j\in[k]} \frac{p_j(x)^\beta}{\sum_{l\in[k]} p_l(x)^\beta} \log\left(\frac{\exp(\beta f_j(x))}{\sum_{l\in[k]} \exp(\beta f_l(x))}\right)$$

with inverse temperature parameter $\beta \in (0, 1)$

Knowledge Distillation as Learning with Nuisance

Given: n datapoints $z_i = (x_i, y_i)$ drawn independently from $\mathbb{P}$

- Feature vector $x_i \in \mathcal{X}$ and label vector $y_i \in \{e_1, \ldots, e_k\}$

Goal: Learn a simple, accurate **student scoring rule** $\hat{f} : \mathcal{X} \to \mathbb{R}^k$

- **Student function class:** $\hat{f} \in \mathcal{F}$
- **Loss function:** $\ell(f(x), p_0(x))$ depending on unknown Bayes class probabilities $p_0(x) = \mathbb{E}[Y \mid x]$ (the nuisance)
- **Optimal student:** $f_0 = \operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{E}[\ell(f(X), p_0(X))]$ (the target)

Vanilla KD = Plug-in ERM

1. Form teacher estimate $\hat{p}$ of nuisance p_0 using $(x_i, y_i)_{i=1}^n$
2. Student minimizes plug-in empirical risk (using the same data!):

$$\hat{f} = \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \ell(f(x_i), \hat{p}(x_i))$$

When Does Knowledge Distillation Work?

Theorem (Fast Rates for Vanilla KD [Dao, Kamath, Syrgkanis, and Mackey, 2021])

With high probability, the Vanilla KD student $\hat{f}$ satisfies

$$\|\hat{f} - f_0\|_2^2 = O(\tfrac{1}{n} + \|\hat{p} - p_0\|_n^2 + \delta_n(\mathcal{F}, p_0)^2)$$

when $\mathcal{F}$ is convex, $\ell(f(x), p(x))$ is strongly convex in $f(x)$, and $\ell, \nabla_{f(x)}\ell$, and $\nabla_{f(x),p(x)}\ell$ are bounded.

Student error: $\|\hat{f} - f_0\|_2^2 \triangleq \mathbb{E}_{X\sim\mathbb{P}}\|\hat{f}(X) - f_0(X)\|_2^2$

- How well $\hat{f}$ matches the optimal student f_0 on test points

Teacher error: $\|\hat{p} - p_0\|_n^2 \triangleq \frac{1}{n}\sum_{i=1}^n \|\hat{p}(x_i) - p_0(x_i)\|_2^2$

- How well the teacher matches the nuisance p_0 on **training** points

Complexity of noiseless student regression: $\delta_n(\mathcal{F}, p_0)^2$

- Localized Rademacher critical radius of $\ell(\mathcal{F}, p_0) - \ell(f_0, p_0)$
- How well $\ell(f, p_0) - \ell(f_0, p_0)$ approximates random noise
- Tight bounds for many $\mathcal{F}$; $\tilde{O}(\frac{1}{n})$ for parametric, VC, & kernel $\mathcal{F}$

When Does Knowledge Distillation Work?

Theorem (Fast Rates for Vanilla KD [Dao, Kamath, Syrgkanis, and Mackey, 2021])

With high probability, the Vanilla KD student $\hat{f}$ satisfies

$$\|\hat{f} - f_0\|_2^2 = O(\tfrac{1}{n} + \|\hat{p} - p_0\|_n^2 + \delta_n(\mathcal{F}, p_0)^2)$$

when $\mathcal{F}$ is convex, $\ell(f(x), p(x))$ is strongly convex in $f(x)$, and $\ell, \nabla_{f(x)}\ell$, and $\nabla_{f(x),p(x)}\ell$ are bounded.

Student error: $\|\hat{f} - f_0\|_2^2 \triangleq \mathbb{E}_{X\sim\mathbb{P}}\|\hat{f}(X) - f_0(X)\|_2^2$

Teacher error: $\|\hat{p} - p_0\|_n^2 \triangleq \frac{1}{n}\sum_{i=1}^{n}\|\hat{p}(x_i) - p_0(x_i)\|_2^2$

Complexity of noiseless student regression: $\delta_n(\mathcal{F}, p_0)^2$

Takeaway: Vanilla KD "works" when teacher approximates p_0 well on training set and noiseless student regression is relatively simple

- Result applies to standard KD losses with bounded f and $\log p$

When Does Knowledge Distillation Fail?

Theorem (Fast Rates for Vanilla KD [Dao, Kamath, Syrgkanis, and Mackey, 2021])

With high probability, the Vanilla KD student $\hat{f}$ satisfies

$$\|\hat{f} - f_0\|_2^2 = O\left(\frac{1}{n} + \|\hat{p} - p_0\|_n^2 + \delta_n(\mathcal{F}, p_0)^2\right)$$

when $\mathcal{F}$ is convex, $\ell(f(x), p(x))$ is strongly convex in $f(x)$, and $\ell, \nabla_{f(x)}\ell$, and $\nabla_{f(x),p(x)}\ell$ are bounded.

Guess: KD fails when teacher approximates p_0 poorly on training set

1. Teacher **underfitting** from model misspecification, an overly restrictive teacher function class, or insufficient training
2. Teacher **overfitting**: $\hat{p}$ approximates p_0 well on test data but overconfident or miscalibrated on training set

Next: Simple lower-bounding examples showing KD suffers from both teacher underfitting and teacher overfitting

Impact of Teacher Underfitting on KD

Example (Impact of Teacher Underfitting [Dao, Kamath, Syrgkanis, and Mackey, 2021]**)**

There exists a classification problem in which, with high probability:

- p_0 and $f_0 = \log(p_0)$ are **constant** (independent of x)
- **Ridge regression teacher** $\hat{p} = \frac{1}{n(1+\lambda)} \sum_{i=1}^{n} y_i$ for $\lambda = \frac{1}{n^{1/4}}$
- **SEL loss** $\ell_{\text{se}}(f(x), p(x)) \triangleq \sum_{j \in [k]} \frac{1}{2}(f_j(x) - \log(p_j(x)))^2$
- **Vanilla KD** with constant $\hat{f}$ satisfies

  $$\|\hat{f} - f_0\|_2^2 = \Omega(\|\hat{p} - p_0\|_n^2) = \Omega(\tfrac{1}{\sqrt{n}})$$

  matching upper bound up to a constant
- Enhanced KD with loss correction satisfies $\|\hat{f} - f_0\|_2^2 = O(\frac{1}{n})$

Takeaway: Vanilla KD is not robust to teacher underfitting

Impact of Teacher Overfitting on KD

Example (Impact of Teacher Overfitting [Dao, Kamath, Syrgkanis, and Mackey, 2021])

There exists a classification problem in which, with high probability:

- $f_0 = \mathbb{E}[\log(p_0(X))]$ is **constant** (independent of x)
- **Teacher interpolates** $\|\hat{p} - p_0\|_n^2 = \Omega(1)$ but still **generalizes** $\mathbb{E}\|\hat{p} - p_0\|_2^2 = O(n^{-\frac{4}{4+d}})$ [Belkin, Rakhlin, and Tsybakov, 2019]
- **SEL loss** $\ell_{\text{se}}(f(x), p(x)) \triangleq \sum_{j\in[k]} \frac{1}{2}(f_j(x) - \log(p_j(x)))^2$
- **Vanilla KD** with constant $\hat{f}$ is **inconsistent** with

$$\|\hat{f} - f_0\|_2^2 = \Omega(\|\hat{p} - p_0\|_n^2) = \Omega(1)$$

  matching upper bound up to a constant
- Enhanced KD with cross-fitting satisfies $\|\hat{f} - f_0\|_2^2 = O(n^{-\frac{4}{4+d}})$

Takeaway: Vanilla KD is not robust to teacher overfitting

Enhancing Knowledge Distillation

Failure Modes of KD

1. Teacher underfitting
2. Teacher overfitting

KD Enhancements

1. Loss correction
2. Cross-fitting

Fighting Overfitting with Cross-fitting

Problem

- Student only observes teacher's **training set** predictions
- Training predictions are susceptible to **overfitting**

Idea: Sample splitting

- Hold out a fraction of the data for training the student
- Downside: Student accuracy suffers from reduced training data

Better idea: Cross-fitting

[Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins, 2018]

1. Split data into B batches $S_1, \ldots, S_B$
2. For $t \in \{1, \ldots, B\}$, fit teacher estimate $\hat{p}^{(t)}$ of p_0 **excluding** S_t
3. Student minimizes the cross-fitted risk:

$$\hat{f} = \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{B} \sum_{i \in S_t} \ell(f(X_i), \hat{p}^{(t)}(X_i))$$

- Each teacher $\hat{p}^{(t)}$ queried only on held-out points S_t
- Student trained on all n datapoints

Fighting Overfitting with Cross-fitting

Theorem (Fast Rates for Cross-fit KD [Dao, Kamath, Syrgkanis, and Mackey, 2021])

With high probability, the Cross-fit KD student $\hat{f}$ satisfies

$$\|\hat{f} - f_0\|_2^2 = O\left(\frac{1}{n} + \frac{1}{B}\sum_{t=1}^{B} \|\hat{p}^{(t)} - p_0\|_2^2 + \delta_{n/B}(\mathcal{F}, \hat{p}^{(t)})^2\right)$$

when $\mathcal{F}$ is convex, $\ell(f(x), p(x))$ is strongly convex in $f(x)$, and $\ell, \nabla_{f(x)}\ell$, and $\nabla_{f(x),p(x)}\ell$ are bounded.

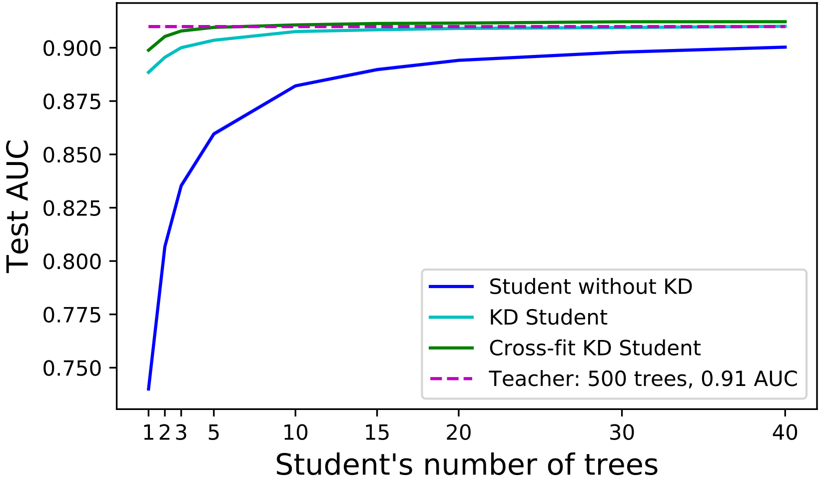
Teacher error: $\|\hat{p}^{(t)} - p_0\|_2^2 \triangleq \mathbb{E}_{X\sim\mathbb{P}}\|\hat{p}^{(t)}(X) - p_0(X)\|_2^2$

- How well the teacher matches the nuisance p_0 on **test** points

Takeaway: Cross-fit KD is robust to teacher overfitting

Cross-fit KD in Action

Task: Predict income level from census data [Dheeru and Karra Taniskidou, 2017]

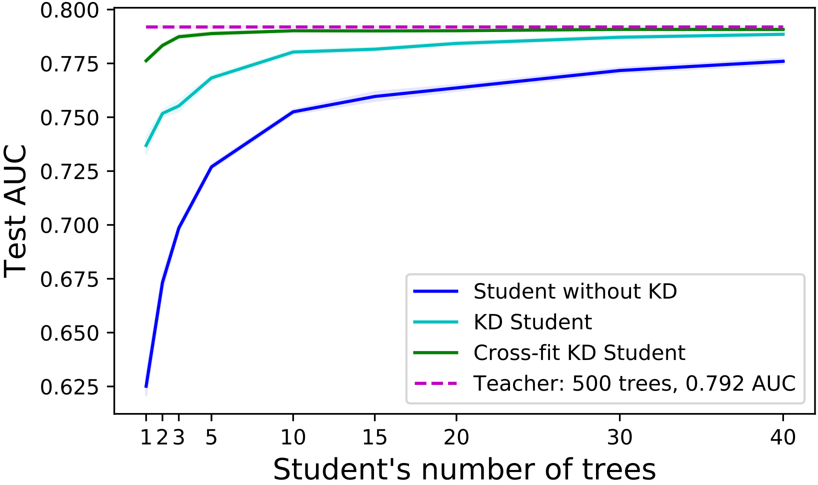


Without KD: Not great **Vanilla:** 10 trees, .91 AUC

Cross-fit: 3 trees, .91 AUC

Cross-fit KD in Action

Task: Predict loan repayment [FIC]

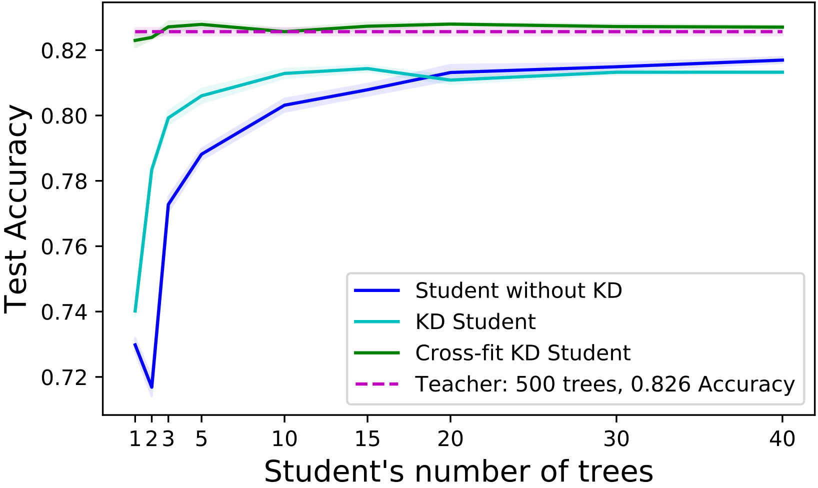


Without KD: Not great **Vanilla:** 40 trees, .789 AUC

Cross-fit: 5 trees, .789 AUC

Cross-fit KD in Action

Task: Distinguish ephemeral and evergreen websites [Eve]

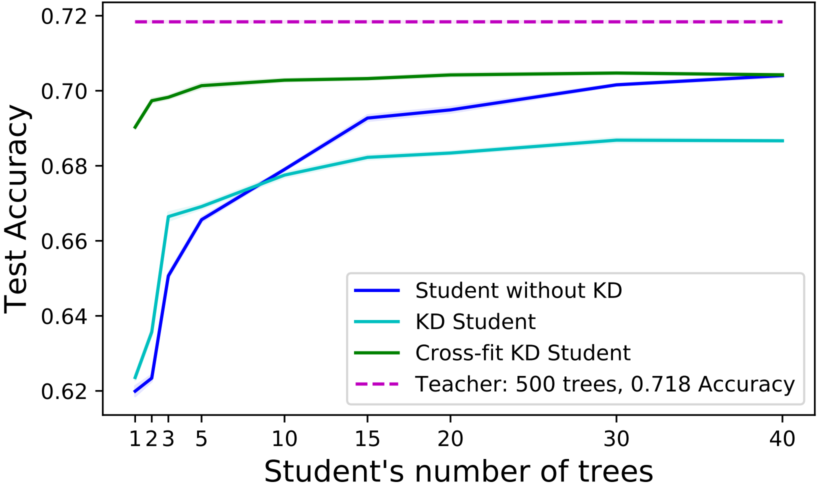


Without KD: 40 trees, .817 Acc. **Vanilla:** 15 trees, .814 Acc.

Cross-fit: 3 trees, .827 Acc.

Cross-fit KD in Action

Task: Detect Higgs boson production [Dheeru and Karra Taniskidou, 2017]



Without KD: 40 trees, .70 Acc. **Vanilla:** 40 trees, .69 Acc.

Cross-fit: 5 trees, .70 Acc.

Fighting Underfitting with Loss Correction

Problem

- KD relies wholly on the accuracy of the teacher
- Suffers when **0th-order** approximation $\ell(f, \hat{p})$ of $\ell(f, p_0)$ is poor

First-order correction: $\ell(f, \hat{p}) + \langle p_0 - \hat{p}, \nabla_{\hat{p}} \ell(f, \hat{p}) \rangle$

- Issue: We don't know p_0!

Unbiased estimate: $\ell(f, \hat{p}) + \langle y - \hat{p}, \nabla_{\hat{p}} \ell(f, \hat{p}) \rangle$

- *Neyman-orthogonal* loss [Foster and Syrgkanis, 2019]: robust to errors in $\hat{p}$
- SEL loss: $\frac{1}{2}(f(x) - \log \hat{p}(x))^2 + \langle y - \hat{p}(x), \mathrm{diag}(\frac{1}{\hat{p}(x)}) f(x) \rangle$
- Issue: Variance explodes if $\hat{p}(x)$ is small!

γ-**Loss correction:** $\ell(f(x), \hat{p}(x)) + \langle y - \hat{p}(x), \gamma(x) f(x) \rangle$

- Select correction matrix $\gamma(x)$ to trade off bias and variance

Enhanced KD: Cross-fitting + loss correction with $\gamma^{(t)}$ fit per batch

Fighting Underfitting with Loss Correction

Theorem (Fast Rates for Enhanced KD [Dao, Kamath, Syrgkanis, and Mackey, 2021])

With high probability, the Enhanced KD student satisfies

$$\|\hat{f}-f_0\|_2^2 = O\Big(\tfrac{1}{n} + \tfrac{1}{B}\textstyle\sum_{t=1}^B \|\hat{p}^{(t)} - p_0\|_4^4 + \delta_{n/B}(\mathcal{F},\hat{p}^{(t)})^2$$
$$+ \tfrac{1}{B}\textstyle\sum_{t=1}^B \|(\mathrm{diag}(\tfrac{1}{\hat{p}^{(t)}}) - \hat{\gamma}^{(t)})(\hat{p}^{(t)} - p_0)\|_2^2$$
$$+ \tfrac{1}{B}\textstyle\sum_{t=1}^B \delta_{n/B}(\mathcal{F},\hat{p}^{(t)})^2\sqrt{\mathbb{E}[\|\hat{\gamma}^{(t)}(X)(Y-\hat{p}^{(t)}(X))\|_2^4]}\Big)$$

with SEL loss, convex $\mathcal{F}$, and $\ell, \nabla_{f(x)}\ell$, and $\nabla_{f(x),p(x)}\ell$ bounded.

Teacher error: $\|\hat{p}^{(t)} - p_0\|_4^4$ **= reduced impact**

- Small even when teacher converges slowly

γ **bias:** $\|(\mathrm{diag}(\frac{1}{\hat{p}^{(t)}}) - \hat{\gamma}^{(t)})(\hat{p}^{(t)} - p_0)\|_2^2$

- Exactly 0 when $\hat{\gamma}^{(t)} = \mathrm{diag}(\frac{1}{\hat{p}^{(t)}})$; product of $\hat{\gamma}$ and $\hat{p}$ errors

γ **variance:** $\sqrt{\mathbb{E}[\|\hat{\gamma}^{(t)}(X)(Y-\hat{p}^{(t)}(X))\|_2^4]}$

- Exactly 0 when $\hat{\gamma}^{(t)} = 0$; often explodes when $\hat{\gamma}^{(t)} = \mathrm{diag}(\frac{1}{\hat{p}^{(t)}})$

Fighting Underfitting with Loss Correction

Theorem (Fast Rates for Enhanced KD [Dao, Kamath, Syrgkanis, and Mackey, 2021]**)**

With high probability, the Enhanced KD student satisfies

$$\begin{aligned}\|\hat{f}-f_0\|_2^2 &= O\Big(\tfrac{1}{n} + \tfrac{1}{B}\textstyle\sum_{t=1}^{B} \|\hat{p}^{(t)} - p_0\|_4^4 + \delta_{n/B}(\mathcal{F}, \hat{p}^{(t)})^2 \\ &\quad + \tfrac{1}{B}\textstyle\sum_{t=1}^{B} \|(\mathrm{diag}(\tfrac{1}{\hat{p}^{(t)}}) - \hat{\gamma}^{(t)})(\hat{p}^{(t)} - p_0)\|_2^2 \\ &\quad + \tfrac{1}{B}\textstyle\sum_{t=1}^{B} \delta_{n/B}(\mathcal{F}, \hat{p}^{(t)})^2 \sqrt{\mathbb{E}[\|\hat{\gamma}^{(t)}(X)(Y - \hat{p}^{(t)}(X))\|_2^4]}\Big)\end{aligned}$$

with SEL loss, convex $\mathcal{F}$, and $\ell, \nabla_{f(x)}\ell$, and $\nabla_{f(x),p(x)}\ell$ bounded.

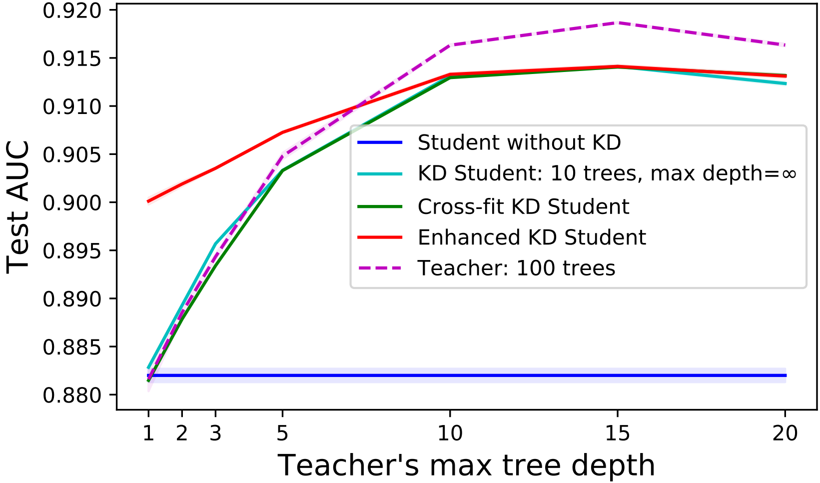
Takeaway: Enhanced KD avoids teacher overfitting and mitigates teacher underfitting when γ chosen to balance bias and variance

Example: Minimize pointwise estimate of bias-variance sum

$$\hat{\gamma}^{(t)}(x) = \mathrm{argmin}_{\gamma} \|\gamma\,(y - \hat{p}^{(t)}(x))\|_2^2 + \alpha\|\,\mathrm{diag}(\tfrac{1}{\hat{p}^{(t)}(x)}) - \gamma\|_2^2$$

Enhanced KD in Action

Task: Predict income level from census data [Dheeru and Karra Taniskidou, 2017]

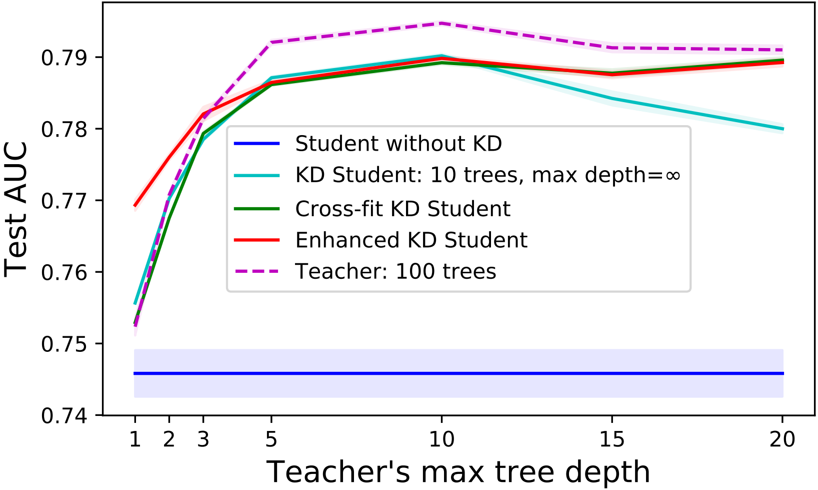


Teacher: Underfits for low depths

High depths: KD $\gg$ no KD **Low depths:** Enhanced $\gg$ Teacher!

Enhanced KD in Action

Task: Predict loan repayment [FIC]

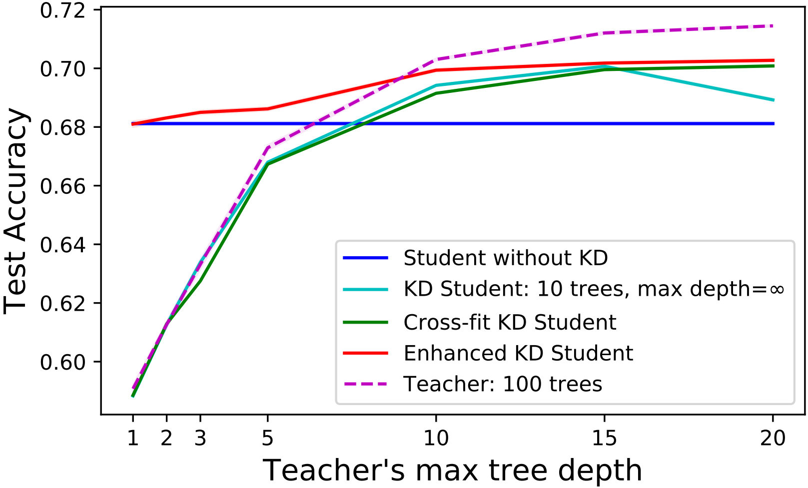


Low depths: Enhanced $\gg$ Teacher!

Mid depths: KD $\gg$ no KD **High depths:** Enhanced $\gg$ Vanilla

Enhanced KD in Action

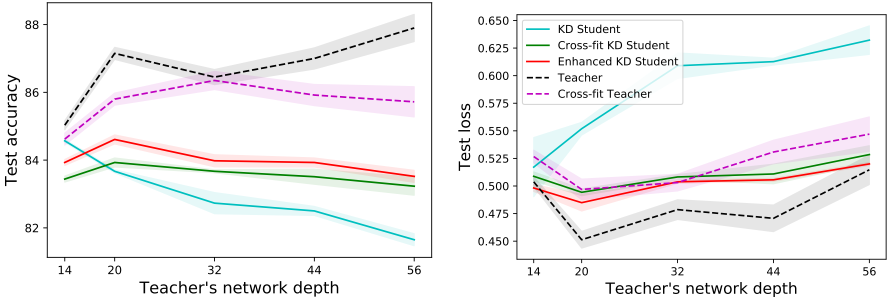
Task: Detect Higgs boson production [Dheeru and Karra Taniskidou, 2017]



Low depths: Enhanced $\gg$ no KD $\gg$ Teacher!

Mid depths: KD $\gg$ no KD **High depths:** Enhanced $\gg$ Vanilla

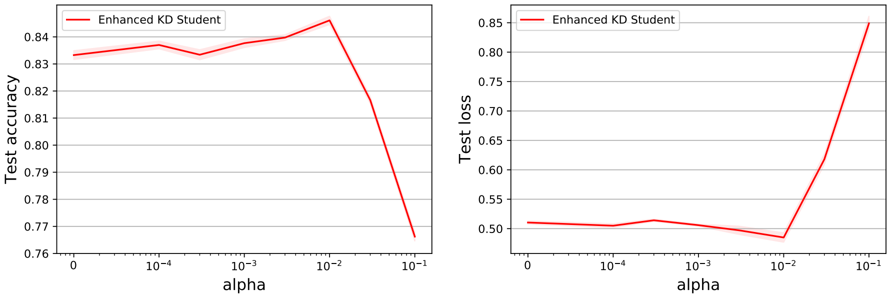
Image Classification with ResNets



Task: CIFAR-10 image classification [Krizhevsky and Hinton, 2009]

- Student = ResNet-10 [He, Zhang, Ren, and Sun, 2016]
- Teacher = ResNet with depth in $\{14, 20, 32, 44, 56\}$
- Vanilla suffers from teacher overfitting
- Cross-fitting corrects for overfitting
- Enhanced benefits from loss-correction

Effect of the Bias-Variance Tradeoff Parameter α



Task: CIFAR-10 image classification [Krizhevsky and Hinton, 2009]

Recall: $\hat{\gamma}^{(t)}(x) = \operatorname{argmin}_{\gamma} \|\gamma\,(y - \hat{p}^{(t)}(x))\|_2^2 + \alpha\|\operatorname{diag}(\frac{1}{\hat{p}^{(t)}(x)}) - \gamma\|_2^2$

- α trades off bias and variance in loss correction
- $\alpha = \infty \Rightarrow$ high-variance Neyman-orthogonal loss
- $\alpha = 0 \Rightarrow$ no loss correction

Summary

What have we accomplished?

- Framed knowledge distillation as learning with nuisance
- Proved that KD succeeds when the teacher's training set probabilities are accurate and noiseless regression is simple
- Identified two KD failure modes: teacher over- and underfitting
- Developed two KD enhancements to mitigate these failures: cross-fitting and loss correction

Paper: Knowledge Distillation as Semiparametric Inference

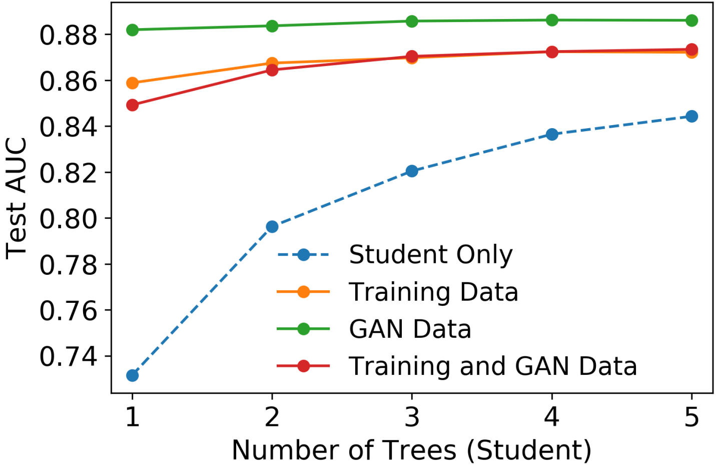
Code: github.com/microsoft/semiparametric-distillation

Future Directions

Many opportunities for future development

1. Can other tools from semiparametric inference improve KD?
   - Example: Targeted Maximum Likelihood [Van Der Laan and Rubin, 2006]
2. Self-distilled students often **outperform** their teachers! [Furlanello, Lipton, Tschannen, Itti, and Anandkumar, 2018]
   - What explains their surprising success?
3. Synthetic data augmentation often improves KD, even when it harms the original supervised learning task
   - Teacher-Student Compression with Generative Adversarial Networks [Liu, Fusi, and Mackey, 2018], MUNGE [Bucila, Caruana, and Niculescu-Mizil, 2006]
   - What characterizes a good generative model for KD?

Task: Distinguish ephemeral and evergreen websites [Eve]

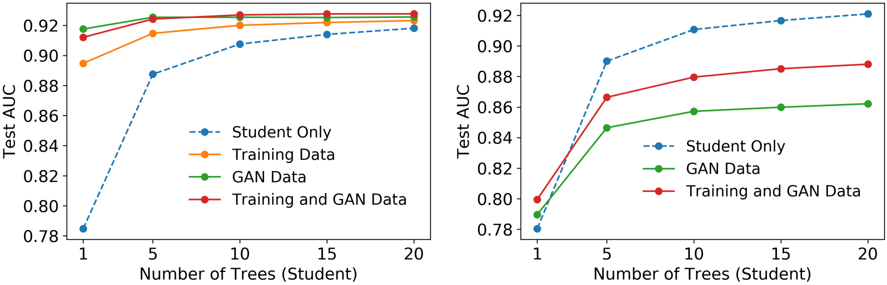


Teacher: 500 trees, .889 AUC

Augmented Student: 1 tree, .882 AUC

Teacher-Student Compression with GANs (GAN-TSC)

[Liu, Fusi, and Mackey, 2018]



(a) GAN augmentation improves KD student performance

(b) Same GAN augmentation impairs student without KD

Task: Distinguish gamma telescope signals

[Dheeru and Karra Taniskidou, 2017]

What's a GAN?

Generative Adversarial Networks (GANs)

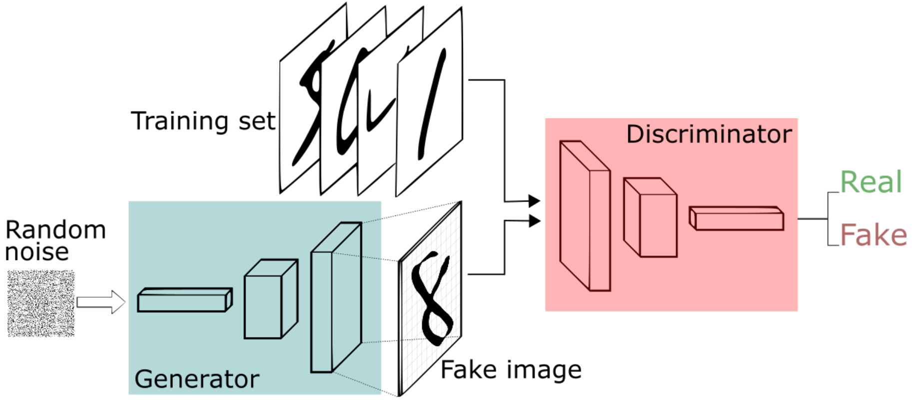


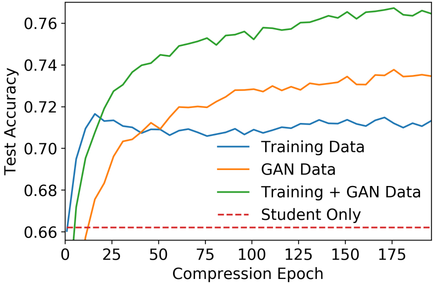
Image credit: Thalles Silva

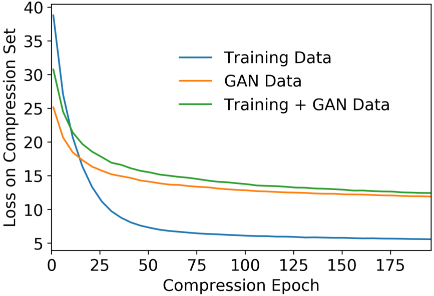
- We train Auxiliary Classifier GANs (AC-GANs) [Odena, Olah, and Shlens, 2017]

Augmenting KD with GAN Data

Teacher-Student Compression with GANs (GAN-TSC)

[Liu, Fusi, and Mackey, 2018]





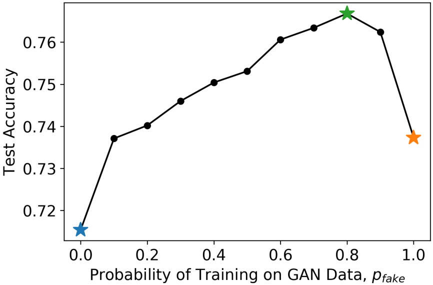
Task: CIFAR-10 image classification [Krizhevsky and Hinton, 2009]

- Teacher: **78.1%** accuracy, NIN [Lin, Chen, and Yan, 2014]
- Without KD: 66% accuracy, LeNet [LeCun, Bottou, Bengio, and Haffner, 1998]
- Vanilla KD: 71% accuracy
- **GAN-TSC**: **76%** accuracy

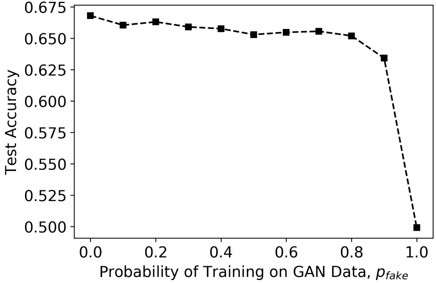
Augmenting KD with GAN Data

Teacher-Student Compression with GANs (GAN-TSC)

[Liu, Fusi, and Mackey, 2018]



(a) GAN augmentation improves KD student performance



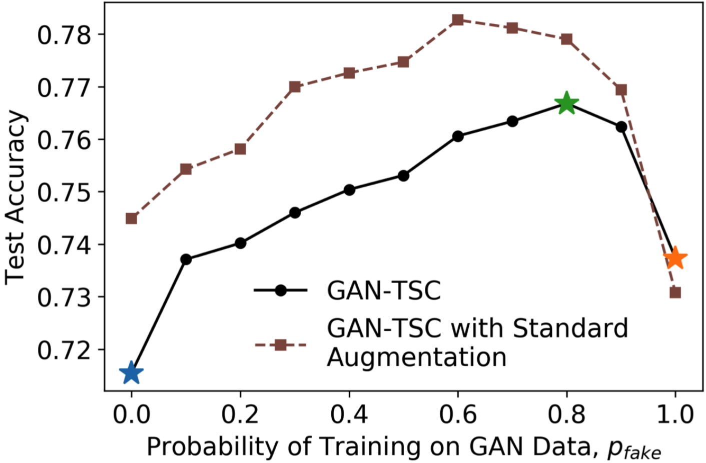
(b) Same GAN augmentation impairs student without KD

Task: CIFAR-10 image classification

[Krizhevsky and Hinton, 2009]

Teacher-Student Compression with GANs (GAN-TSC)

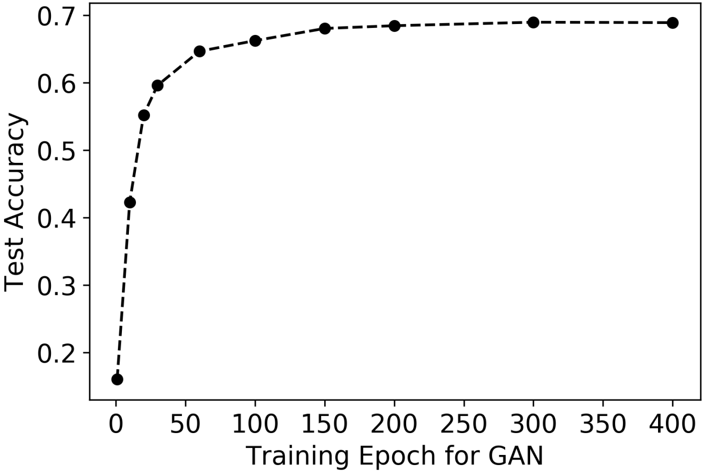
[Liu, Fusi, and Mackey, 2018]



GAN-TSC complements standard image augmentation

Teacher-Student Compression with GANs (GAN-TSC)

[Liu, Fusi, and Mackey, 2018]



Task: CIFAR-10 image classification [Krizhevsky and Hinton, 2009]

Evaluating GANs with Distillation

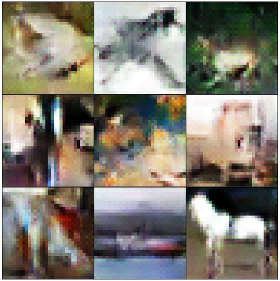
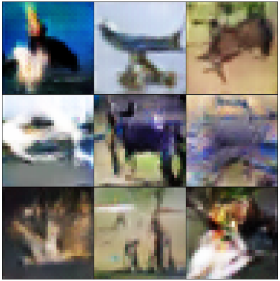
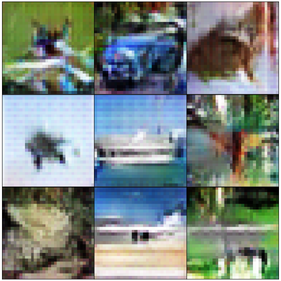
Teacher-Student Compression (TSC) Score [Liu, Fusi, and Mackey, 2018]

- Measures test accuracy of student distilled with synthetic data
  - Higher test accuracy indicates higher quality data
- Train student for single pass through data for rapid evaluation

Inception Score [Salimans, Goodfellow, Zaremba, Cheung, Radford, and Chen, 2016]

- Uses classifier confidence to quantify class affinity
- Does not account for within class diversity
- Easily misled by high-confidence unrealistic images

Evaluating GANs with Distillation

Real Data	Well-trained GAN	Inferior GAN
		
Inception: 11.2 ± 0.1 TSC: 0.994 ± 0.003	Inception: 5.80 ± 0.06 TSC: 0.778 ± 0.002	Inception: 5.93 ± 0.06 TSC: 0.702 ± 0.002

Timing: Inception (1436.6s), TSC Score (350.1s)
Code: `https://github.com/RuishanLiu/GAN-TSC-Score`
Paper: Teacher-Student Compression with Generative Adversarial Networks

Future Directions

Many opportunities for future development

1. Can other tools from semiparametric inference improve KD?
   - Example: Targeted Maximum Likelihood [Van Der Laan and Rubin, 2006]
2. Self-distilled students often **outperform** their teachers! [Furlanello, Lipton, Tschannen, Itti, and Anandkumar, 2018]
   - What explains their surprising success?
3. Synthetic data augmentation often improves KD, even when it harms the original supervised learning task
   - Teacher-Student Compression with Generative Adversarial Networks [Liu, Fusi, and Mackey, 2018], MUNGE [Bucila, Caruana, and Niculescu-Mizil, 2006]
   - What characterizes a good generative model for KD?

References I

Stumbleupon evergreen dataset. https://www.kaggle.com/c/stumbleupon.

FICO: Explanable machine learning challenge. https://community.fico.com/s/explainable-machine-learning-challenge.

J. Ba and R. Caruana. Do deep nets really need to be deep? In *Advances in neural information processing systems*, pages 2654–2662, 2014.

M. Belkin, A. Rakhlin, and A. B. Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR, 2019.

C. Bucila, R. Caruana, and A. Niculescu-Mizil. Model compression. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006*, pages 535–541, 2006. doi: 10.1145/1150402.1150464.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ectj.12097.

T. Dao, G. M. Kamath, V. Syrgkanis, and L. Mackey. Knowledge distillation as semiparametric inference, 2021. URL https://arxiv.org/abs/2104.09732.

D. Dheeru and E. Karra Taniskidou. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

D. J. Foster and V. Syrgkanis. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019.

T. Furlanello, Z. C. Lipton, M. Tschannen, L. Itti, and A. Anandkumar. Born again neural networks. *arXiv preprint arXiv:1805.04770*, 2018.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

J. Li, R. Zhao, J.-T. Huang, and Y. Gong. Learning small-size dnn with output-distribution-based criteria. In *Fifteenth annual conference of the international speech communication association*, 2014.

M. Lin, Q. Chen, and S. Yan. Network in network. In *ICLR*, 2014.

References II

R. Liu, N. Fusi, and L. Mackey. Teacher-student compression with generative adversarial networks. *arXiv preprint arXiv:1812.02271*, 2018. URL `https://arxiv.org/abs/1812.02271`.

D. Lopez-Paz, L. Bottou, B. Schölkopf, and V. Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

A. K. Menon, A. S. Rawat, S. J. Reddi, S. Kim, and S. Kumar. Why distillation helps: a statistical perspective. *arXiv preprint arXiv:2005.10419*, 2020.

H. Mobahi, M. Farajtabar, and P. L. Bartlett. Self-distillation amplifies regularization in hilbert space. *arXiv preprint arXiv:2002.05715*, 2020.

E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier GANs. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2642–2651, 2017.

M. Phuong and C. Lampert. Towards understanding knowledge distillation. In *International Conference on Machine Learning*, pages 5142–5151, 2019.

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

M. J. Van Der Laan and D. Rubin. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

Localized Rademacher Complexity

$$\mathcal{G} \triangleq \{z \to r\,(\ell(f(x), p_0(x)) - \ell(f_0(x), p_0(x))) : f \in \mathcal{F},\ r \in [0,1]\}$$

Definition (Critical radius δ_n [Wainwright, 2019, 14.1.1])

Satisfies $\mathcal{R}(\delta_n; \mathcal{G}) \leq \delta_n^2$ for the *localized Rademacher complexity*

$$\mathcal{R}(\delta; \mathcal{G}) = \mathbb{E}_{X_{1:n}, \epsilon_{1:n}}[\sup_{g \in \mathcal{G}: \|g\|_2 \leq \delta} \frac{1}{n} \sum_{i=1}^n \epsilon_i g(X_i)]$$

where ϵ_i are i.i.d. random variables uniform on $\{-1, 1\}$.

Nadaraya-Watson Kernel Smoothing

Definition (Nadaraya-Watson kernel smoothing estimator [Nadaraya, 1964, Watson, 1964])

$$\tilde{p}(x) \triangleq \begin{cases} y_i & \text{if} \quad x = x_i \\ \sum_{i=1}^n y_i K((x - x_i)/h) / \sum_{i=1}^n K((x - x_i)/h) & \text{otherwise} \end{cases}$$

with kernel $K(x) = \|x\|_2^{-a} \mathbb{I}[\|x\|_2 \leq 1]$, $a \in (0, d/2)$, and $h = n^{-1/(4+d)}$.