Weak Supervision

Dr. Fayyaz ul Amir Afsar Minhas
PIEAS Biomedical Informatics Research Lab
Department of Computer and Information Sciences
Pakistan Institute of Engineering & Applied Sciences
PO Nilore, Islamabad, Pakistan
http://faculty.pieas.edu.pk/fayyaz/

Primary resources

- Learning using privileged information: Similarity control and knowledge transfer (Vapnik & Izmailov, JMLR: 2015)
- Unifying distillation and privileged information (Lopez-paz, Bottou, Scholkopf, Vapnik, ICR: 2016)
- Weak supervision and other non-standard classification problems: A taxonomy (Hernandez-Gonzalez and Lozano, Pattern Recognition Letters, 2015)

Traditional Supervised Learning

- One Instance, One Label
- Labeling is expensive or impractical
  - Impossibility of observing individual examples
  - Reliability of labeling
- Solutions to Data Hungry AI
  - Develop methods that can work with
    - Partial Labeling during training
    - Use partial labeling during testing
    - Can use variety of data from different sources

Instance Label Relationship

Four possible definitions of the target function H. An example is composed of a single (SI) or multiple (MI) instances. The categorization is composed of a single (SL) or multiple (ML) class labels.

		Categorization	
		SL	ML
Example	SI	$H : \mathcal{X} \to \mathcal{C}$	$H : \mathcal{X} \to 2^{\mathcal{C}}$
	MI	$H : 2^{\mathcal{X}} \to \mathcal{C}$	$H : 2^{\mathcal{X}} \to 2^{\mathcal{C}}$

Types of Supervision

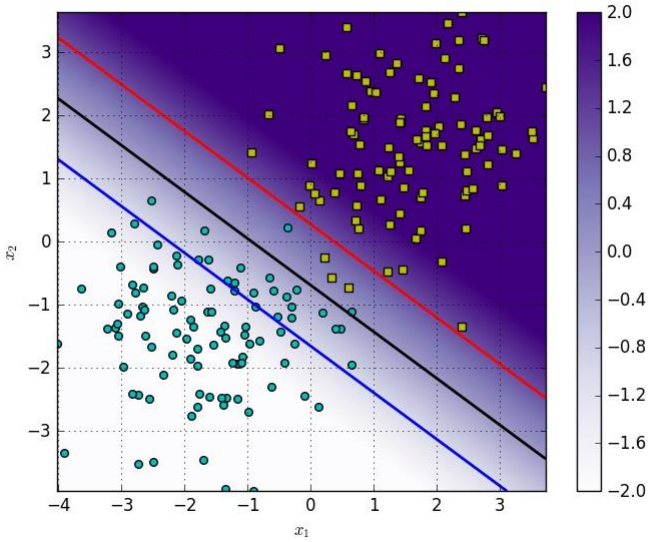
Model	References	Description
Full-supervision	[9,24,34,43]	For each example, complete class information is provided.
Unsupervision	[24]	No class information is provided with the examples.
Semi-supervision	[5]	Part of the examples are provided fully supervised. The rest are unsupervised.
Positive-unlabeled	[4,10,21,32]	Part of the examples are provided fully supervised, all of them with the same categorization. The rest are unsupervised.
Candidate labels	[7,13,16]	For each example, a set of class labels is provided. In this set, the class label(s) that compose the real categorization of the example are included.
Probabilistic labels	[18]	For each example, the probability of belonging to each class label is provided. This probability distribution is expected to assign high probability to the real label(s).
Incomplete	[3,33,42]	For each example, a subset of the labels that compose its real categorization is provided (SIML or MIML, Table 1).
Noisy labels	[2,44]	For each example, complete class information is provided, although its correctness is not guaranteed.
Crowd	[30,40]	For each example, many different non-expert annotators provide their (noisy) categorization.
Mutual label constraints	[19,20,31]	For each **group** of examples, an explicit relationship between their class labels is provided (e.g., all the examples have the same categorization).
Candidate labeling vectors	[22]	For each **group** of examples, a set of labeling vectors (including the real one) is provided. A labeling vector provides a class label for each examples of a group.
Label proportions	[15,25,28]	For each **group** of examples, the proportion of examples belonging to each class label is provided.

Examples

- Sample Weighting to reflect confidence in labeling or confidence in features

Examples

- Semi-Supervised Classification
- Transductive Classification
  - How can semi-supervised learning be viewed as Learning using privileged information?



$$\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{L}\xi_i + C^*\sum_{i=L+1}^{L+U}\xi_i$$

$$y_i f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \geq 1-\xi_i, \quad i=1,\ldots,L$$

$$|f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)| \geq 1-\xi_i, \quad i=L+1,\ldots,L+U$$

$$\frac{1}{U}\sum_{i=L+1}^{L+U} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = \frac{1}{L}\sum_{i=1}^{L} y_i .$$

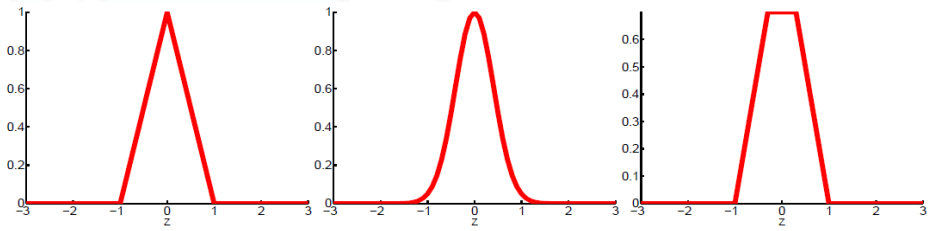


Figure 1: Three loss functions for unlabeled examples, from left to right (i) the Symmetric Hinge $H_1(|t|) = \max(0, 1-|t|)$, (ii) Symmetric Sigmoid $S(t) = \exp(-3t^2)$; and (iii) Symmetric Ramp loss, $R_s(|t|) = \min(1+s, \max(0, 1-|t|))$. The last loss function has a plateau of width $2|s|$ where $s \in (-1, 0]$ is a tunable parameter, in this case $s = -0.3$.

Large Scale Transductive SVMs by Collobert et al. JMLR (2006)

Presentations

- On classification with bags, groups and sets (Sadaf)
- Learning from candidate label sets (Amina)
- Learning from partial labels (Dawood)

Learning using privileged information

- LUPI

The LUPI paradigm describes a more complex model: given a set of iid triplets

$$(x_1, x_1^*, y_1), ..., (x_\ell, x_\ell^*, y_\ell), \quad x_i \in X, \quad x_i^* \in X^*, \quad y_i \in \{-1, +1\},$$

LUPI: Examples

Example 1. Suppose that our goal is to find a rule that predicts the outcome y of a surgery in three weeks after it, based on information x available before the surgery. In order to find the rule in the classical paradigm, we use pairs (x_i, y_i) from previous patients.

However, for previous patients, there is also additional information x^* about procedures and complications during surgery, development of symptoms in one or two weeks after surgery, and so on. Although this information is not available *before* surgery, it does exist in historical data and thus can be used as privileged information in order to construct a rule that is better than the one obtained without using that information. The issue is how large an improvement can be achieved.

Example 2. Let our goal be to find a rule $y = f(x)$ to classify biopsy images x into two categories y: cancer ($y = +1$) and non-cancer ($y = -1$). Here images are in a pixel space X, and the classification rule has to be in the same space. However, the standard diagnostic procedure also includes a pathologist's report x^* that describes his/her impression about the image in a high-level holistic language X^* (for example, "aggressive proliferation of cells of type A among cells of type B" etc.).

The problem is to use the pathologist's reports x^* as privileged information (along with images x) in order to make a better classification rule for images x just in pixel space X. (Classification by a pathologist is a time-consuming procedure, so fast decisions during surgery should be made without consulting him or her).

Example 3. Let our goal be to predict the direction of the exchange rate of a currency at the moment t. In this problem, we have observations about the exchange rates before t, and we would like to predict if the rate will go up or down at the moment $t + \Delta$. However, in the historical market data we also have observations about exchange rates *after* moment t. Can this future-in-the-past privileged information be used for construction of a better prediction rule?

LUPI Formulation

Our goal is to we minimize the functional

$$\mathcal{T}(w, w^*, b, b^*) = \frac{1}{2}[(w, w) + \gamma(w^*, w^*)] + C\sum_{i=1}^{\ell}[y_i((w^*, z_i^*) + b^*)]_+$$

subject to the constraints

$$y_i[(w, z_i) + b] \geq 1 - [y_i((w^*, z_i^*) - b^*)]_+.$$

To find the solution of this optimization problem, we approximate this non-linear optimization problem with the following quadratic optimization problem: minimize the functional

$$\mathcal{T}(w, w^*, b, b^*) = \frac{1}{2}[(w, w) + \gamma(w^*, w^*)] + C\sum_{i=1}^{\ell}[y_i((w^*, z_i^*) + b^*) + \zeta_i] + \Delta C\sum_{i=1}^{\ell}\zeta_i \quad (8)$$

(here $\Delta > 0$ is the parameter of approximation[6]) subject to the constraints

$$y_i((w, z_i) + b) \geq 1 - y_i((w^*, z^*) + b^*) - \zeta_i, \quad i = 1, ..., \ell, \quad (9)$$

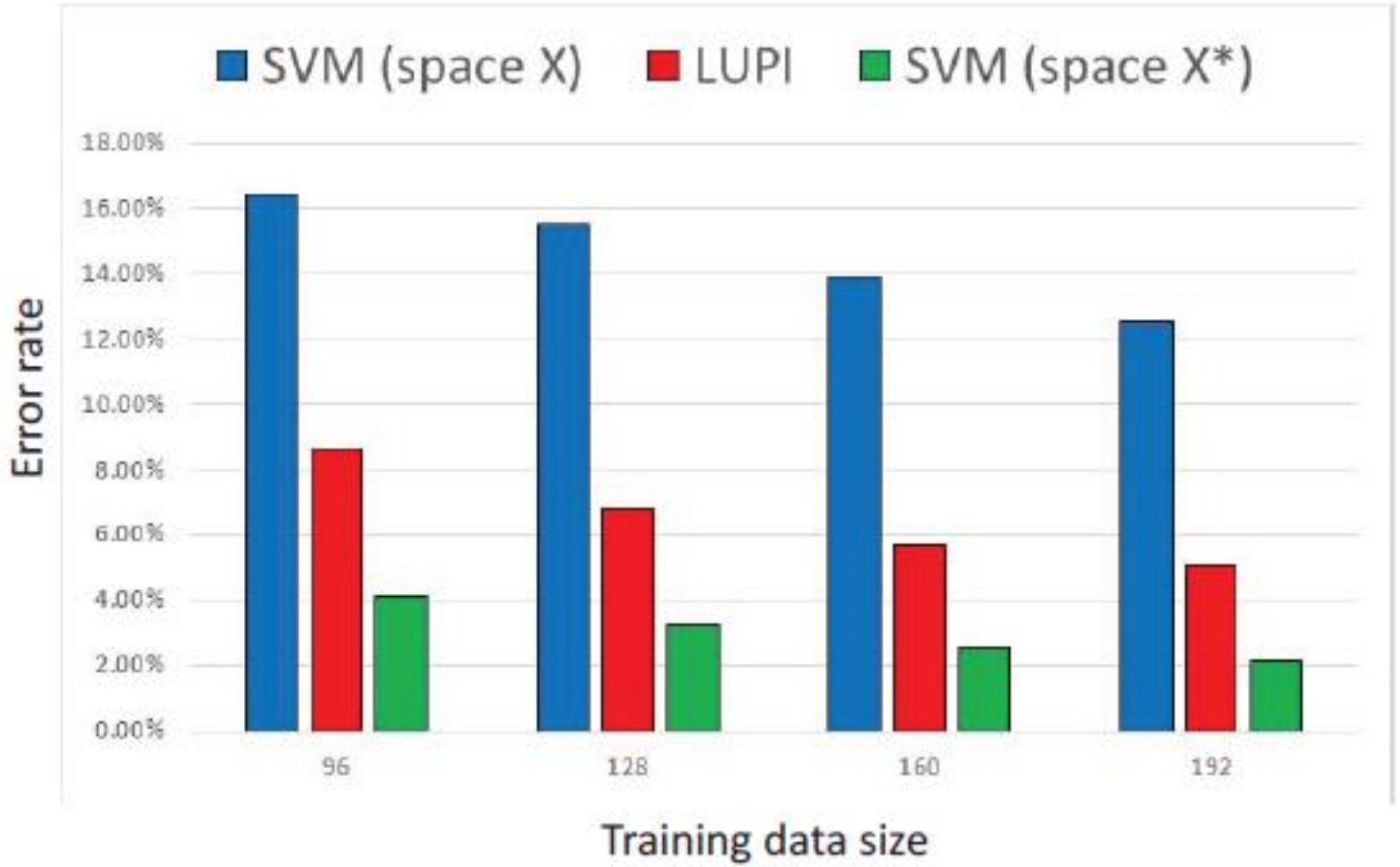
the constraints

$$y_i((w^*, z_i^*) + b^*) + \zeta_i \geq 0, \quad \forall i = 1, ..., \ell, \quad (10)$$

and the constraints

$$\zeta_i > 0, \quad \forall i = 1, ..., \ell. \quad (11)$$

LUPI Results



LUPI Knowledge Transfer

- LUPI can be viewed as a mechanism for knowledge transfer from one space to another
  - LUPI allows a student to learn from a teacher
    - Teacher controls the boundary of the student from what the student would draw otherwise
    - The teach controls the concept of similarity between examples that the student learns (Kernel Space)

One can give the following general mathematical justification for our model of knowledge transfer. Teacher knows that the goal of Student is to construct a good rule in space X with one of the functions from the set $f(x, \alpha),\ x \in X,\ \alpha \in \Lambda$ with capacity VC_X. Teacher also knows that there exists a rule of the same quality in space X^* – a rule that belongs to the set $f^*(x^*, \alpha^*),\ x^* \in X^*,\ \alpha^* \in \Lambda^*$ and that has a much smaller capacity VC_{X^*}. This knowledge can be defined by the ratio of the capacities

$$\kappa = \frac{VC_X}{VC_{X^*}}.$$

The larger is κ, the more knowledge Teacher can transfer to Student; also the larger is κ, the fewer examples will Student need to select a good classification rule.

Distillation

- Learning one classifier from another

We focus on c-class classification, although the same ideas apply to regression. Consider the data

$$\{(x_i, y_i)\}_{i=1}^n \sim P^n(x, y),\ x_i \in \mathbb{R}^d,\ y_i \in \Delta^c. \tag{2}$$

Here, Δ^c is the set of c-dimensional probability vectors. Using (2), we are interested in learning the representation

$$f_t = \arg\min_{f \in \mathcal{F}_t} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \sigma(f(x_i))) + \Omega(\|f\|), \tag{3}$$

where $\mathcal{F}_t$ is a class of functions from $\mathbb{R}^d$ to $\mathbb{R}^c$, the function $\sigma : \mathbb{R}^c \to \Delta^c$ is the softmax operation

$$\sigma(z)_k = \frac{e^{z_k}}{\sum_{j=1}^{c} e^{z_j}},$$

for all $1 \leq k \leq c$, the function $\ell : \Delta^c \times \Delta^c \to \mathbb{R}_+$ is the cross-entropy loss

$$\ell(y, \hat{y}) = -\sum_{k=1}^{c} y_k \log \hat{y}_k,$$

and $\Omega : \mathbb{R} \to \mathbb{R}$ is an increasing function which serves as a regularizer.

Unifying Distillation and LUPI

$$f_s = \arg\min_{f \in \mathcal{F}_s} \frac{1}{n} \sum_{i=1}^{n} \Big[(1-\lambda)\ell(y_i, \sigma(f(x_i))) + \lambda \ell(s_i, \sigma(f(x_i))) \Big], \tag{4}$$

where

$$s_i = \sigma(f_t(x_i)/T) \in \Delta^c \tag{5}$$

are the *soft predictions* from f_t about the training data, and $\mathcal{F}_s$ is a function class simpler than $\mathcal{F}_t$. The temperature parameter $T > 0$ controls how much do we want to soften or smooth the class-probability predictions from f_t, and the imitation parameter $\lambda \in [0, 1]$ balances the importance between imitating the soft predictions s_i and predicting the true hard labels y_i. Higher temperatures lead to softer class-probability predictions s_i. In turn, softer class-probability predictions reveal label dependencies which would be otherwise hidden as extremely large or small numbers. After distillation, we can use the simpler $f_s \in \mathcal{F}_s$ for faster prediction at test time.

Generalized Distillation and LUPI

We now have all the necessary background to describe *generalized distillation*. To this end, consider the data $\{(x_i, x_i^\star, y_i)\}_{i=1}^n$. Then, the process of generalized distillation is as follows:

1. Learn teacher $f_t \in \mathcal{F}_t$ using the input-output pairs $\{(x_i^\star, y_i)\}_{i=1}^n$ and Eq. 3.
2. Compute teacher soft labels $\{\sigma(f_t(x_i^\star)/T)\}_{i=1}^n$, using temperature parameter $T > 0$.
3. Learn student $f_s \in \mathcal{F}_s$ using the input-output pairs $\{(x_i, y_i)\}_{i=1}^n$, $\{(x_i, s_i)\}_{i=1}^n$, Eq. 4, and imitation parameter $\lambda \in [0, 1]$.[2]

We say that generalized distillation reduces to *Hinton's distillation* if $x_i^\star = x_i$ for all $1 \leq i \leq n$ and $|\mathcal{F}_s|_C \ll |\mathcal{F}_t|_C$, where $|\cdot|_C$ is an appropriate function class capacity measure. Conversely, we say that generalized distillation reduces to *Vapnik's learning using privileged information* if $x_i^\star$ is a privileged description of x_i, and $|\mathcal{F}_s|_C \gg |\mathcal{F}_t|_C$.

Generalized Distillation

This comparison reveals a subtle difference between Hinton's distillation and Vapnik's privileged information. In Hinton's distillation, $\mathcal{F}_t$ is *flexible*, for the teacher to exploit her *general purpose* representation $x_i^\star = x_i$ to learn intricate patterns from *large* amounts of labeled data. In Vapnik's privileged information, $\mathcal{F}_t$ is *simple*, for the teacher to exploit her *rich* representation $x_i^\star \neq x_i$ to learn intricate patterns from *small* amounts of labeled data. The space of privileged information is thus a specialized space, one of "metaphoric language". In our running example of biopsy images, the space of medical reports is much more specialized than the space of pixels, since the space of pixels can also describe buildings, animals, and other unrelated concepts. In any case, the teacher must develop a language that effectively communicates information to help the student come up with better representations. The teacher may do so by incorporating invariances, or biasing them towards being robust with respect to the kind of distribution shifts that the teacher may expect at test time. In general, having a teacher is one opportunity to learn characteristics about the decision boundary which are not contained in the training sample, in analogy to a good Bayesian prior.

Extensions

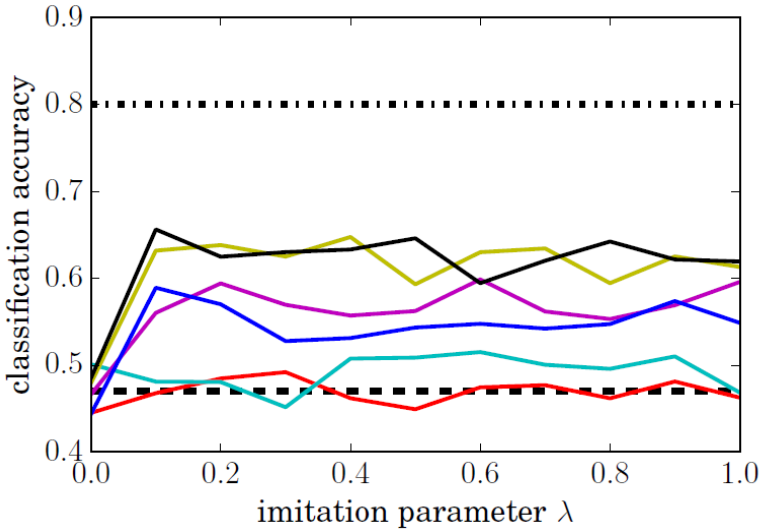
- LUPI
  - Difference: In Vapnik's LUPI, both the decision function in the input space and the correcting slack function in the privileged space are learned simultaneously
  - In generalized distillation, the learning of the privileged space classifier is separate from the input space classifier
- Semi-Supervised Learning
- Learning the universum
- Multiple Task Learning
- Curriculum and reinforcement learning

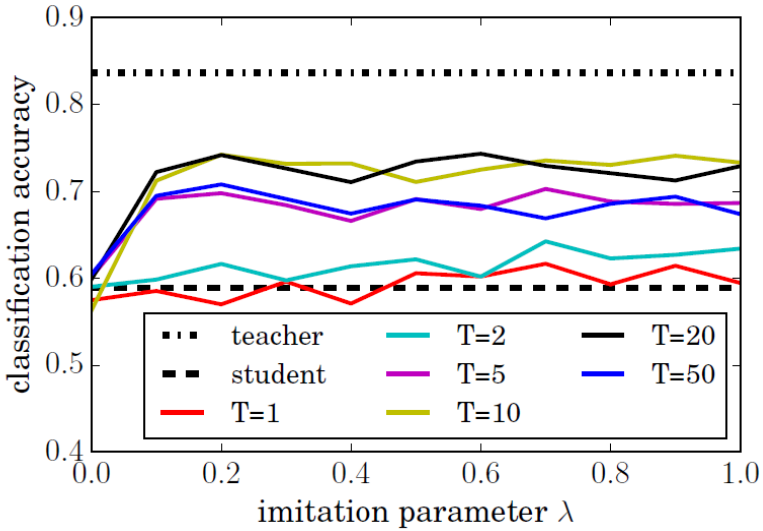
Toy Examples

- Clean labels as Priveleged Information
  - LUPI is useful: P = 96%, I = 88%, L = 95%
- Clean Features as privileged information
  - LUPI is not useful because the noise is completely random: P = 90%, I = 68%, L = 70%
- Relevant Features as privileged information
  - LUPI is useful: P = 98%, I = 89%, L = 97%

Practical Examples

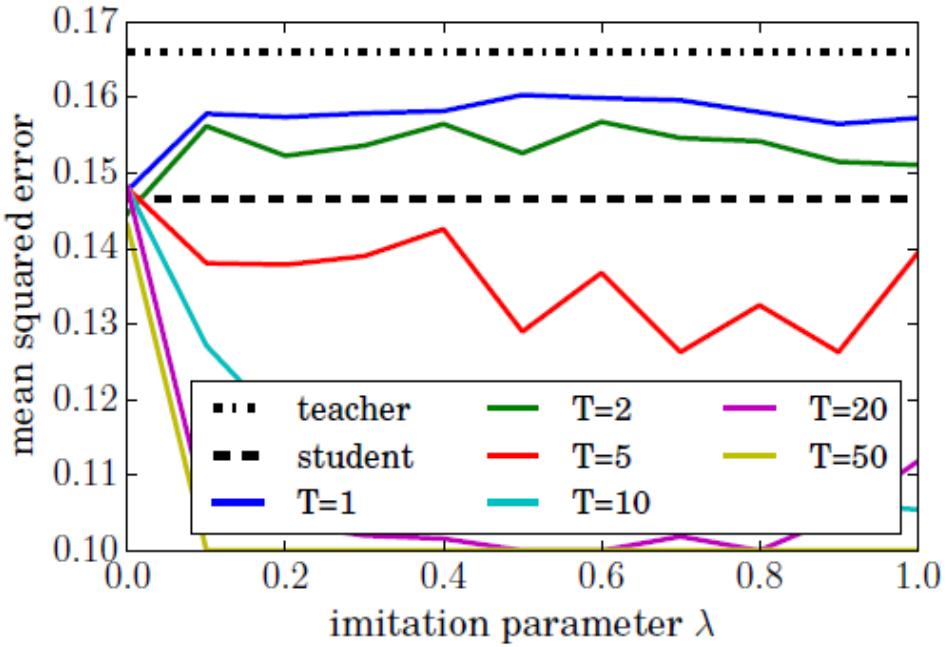
- MNIST
  - Use 28x28 images as privileged information and down sampled 7x7 images as input space information





Practical Examples

- CIFAR 10 Semi-Supervised Classification
- 300 labeled images
- 50K unlabeled
- Privileged Features: 32x32 images
- Input Features: 32x32 images with additive noise
- Train teacher on 300 labeled images (MSE: ~16.5%)
- Classify the unlabeled images using the teacher
- Train a student using labeled data (300) and unlabeled data using predictions from the teacher: (MSE: ~0.10)
- Compare it to a classifier trained on all 50K labeled images (MSE: ~0.15)



End of Lecture

better than a thousand days of diligent study is one day with a great teacher.

- Danny Hillis