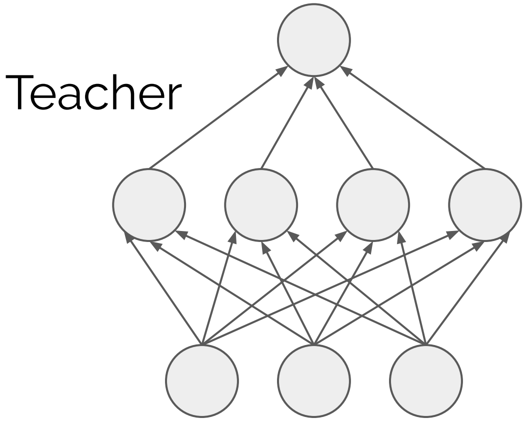
Towards Understanding Knowledge Distillation

Mary Phuong, Christoph H. Lampert

IST AUSTRIA
Institute of Science and Technology

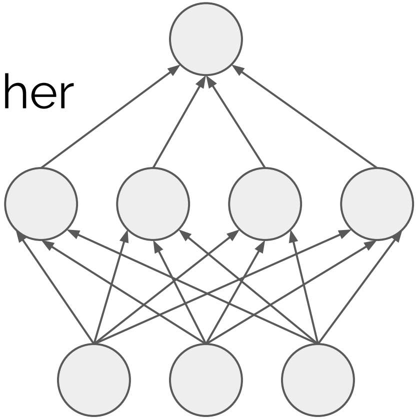
poster #254

Knowledge distillation [Hinton et al. 2015]

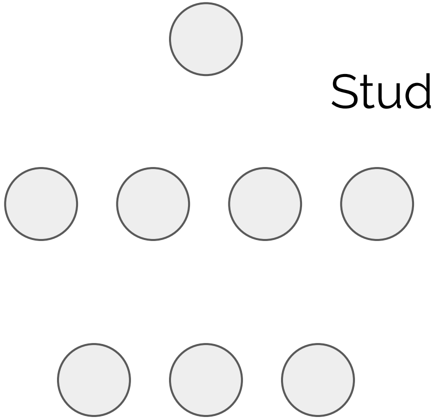


Knowledge distillation [Hinton et al. 2015]

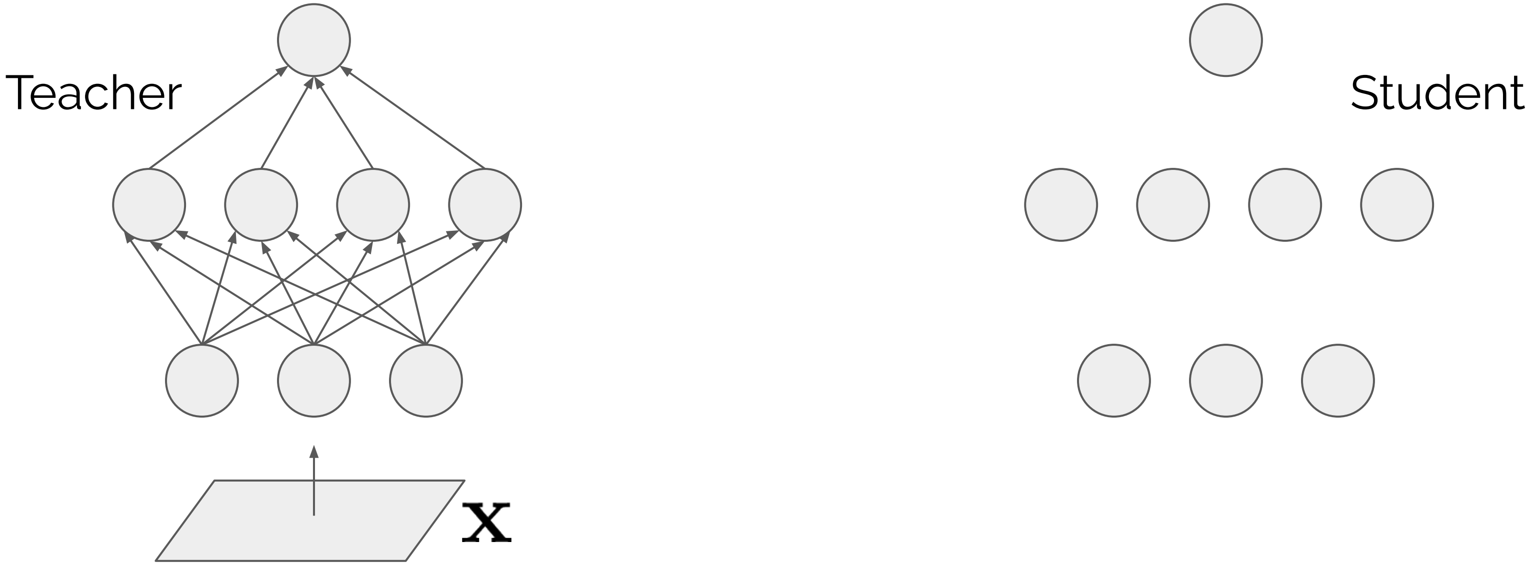
Teacher



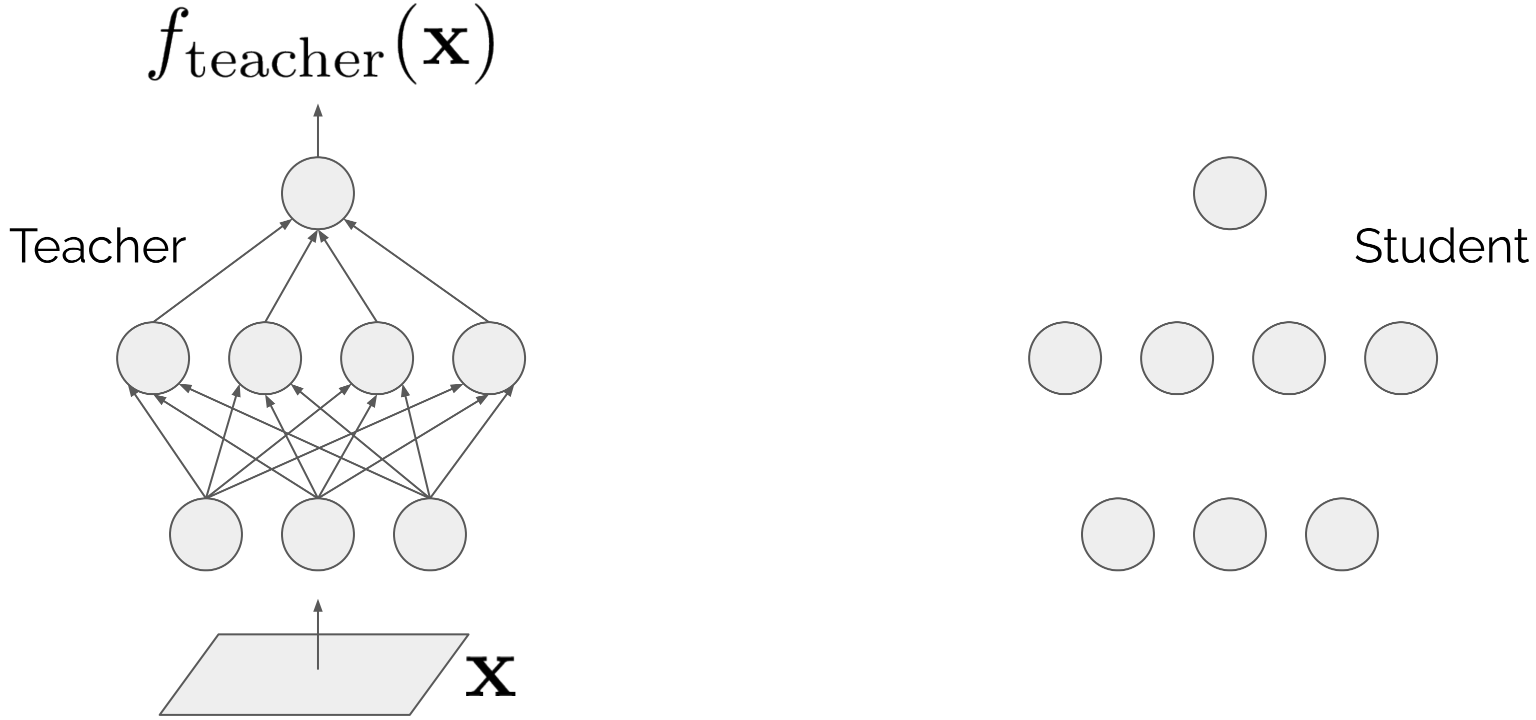
Student



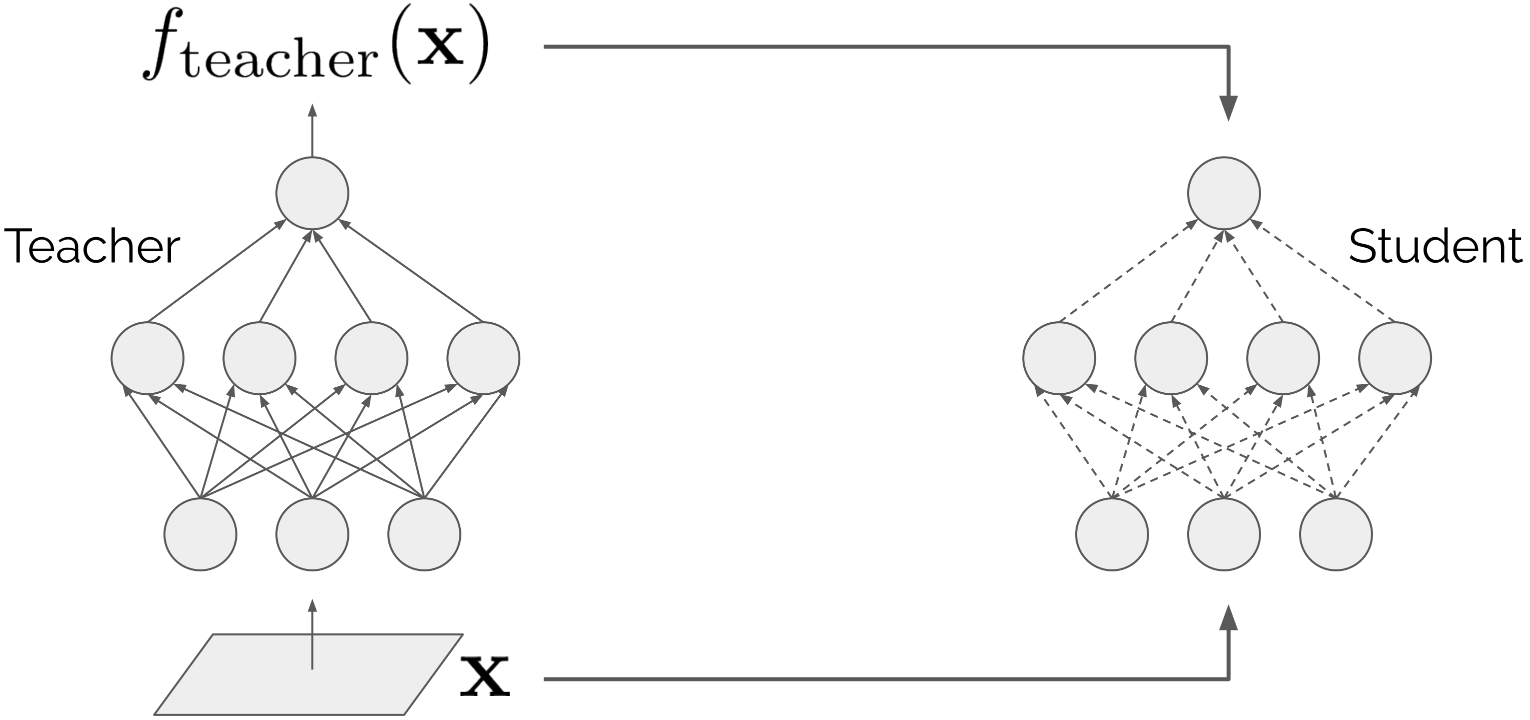
Knowledge distillation [Hinton et al. 2015]



Knowledge distillation [Hinton et al. 2015]



Knowledge distillation [Hinton et al. 2015]



How effective is distillation?

How effective is distillation?

- n_{data} examples $\sim P_{\mathbf{x}}$ for distillation

How effective is distillation?

- n_{data} examples $\sim P_{\mathbf{x}}$ for distillation

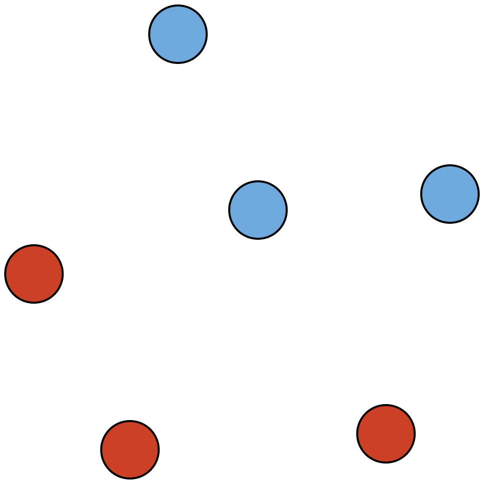
$$\underset{\mathbf{x}}{\mathbb{P}}\left[\operatorname{sign} f_{\text{student}}(\mathbf{x}) \neq \operatorname{sign} f_{\text{teacher}}(\mathbf{x})\right] = ?$$

How effective is distillation?

- n_{data} examples $\sim P_{\mathbf{x}}$ for distillation

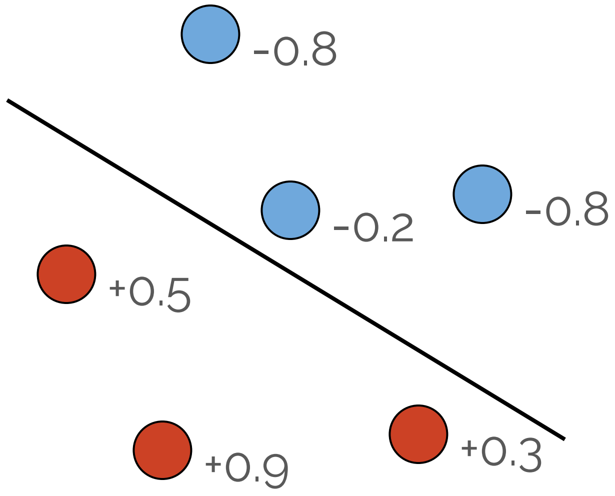
$$\underbrace{\mathbb{P}_{\mathbf{x}}\left[\operatorname{sign} f_{\text{student}}(\mathbf{x}) \neq \operatorname{sign} f_{\text{teacher}}(\mathbf{x})\right]}_{\text{“transfer risk”}} = ?$$

Setting: Linear distillation



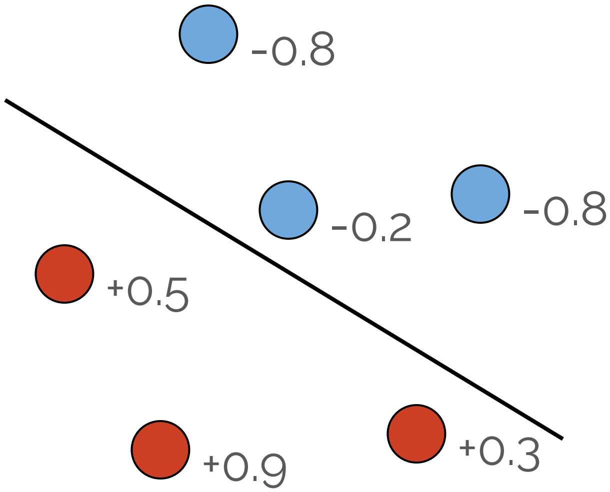
- Binary classification in $\mathbb{R}^d$

Setting: Linear distillation



- Binary classification in $\mathbb{R}^d$
- Linear teacher

Setting: Linear distillation



- Binary classification in $\mathbb{R}^d$
- Linear teacher
- Student is a deep linear network

$$f_{\text{student}}(\mathbf{x}) = \mathbf{W}_N \mathbf{W}_{N-1} \cdots \mathbf{W}_1 \mathbf{x}$$

- Distillation by gradient descent

Result 1: Closed-form student

$$f_{\text{teacher}}(\mathbf{x}) = \mathbf{w}^\intercal \mathbf{x}$$

Result 1: Closed-form student

$$f_{\text{teacher}}(\mathbf{x}) = \mathbf{w}^\intercal \mathbf{x}$$

$$\hat{f}_{\text{student}}(\mathbf{x}) = \begin{cases} \mathbf{w}^\intercal \mathbf{x}, & n_{\text{data}} \geq d, \\ & \end{cases}$$

Result 1: Closed-form student

$$f_{\text{teacher}}(\mathbf{x}) = \mathbf{w}^\intercal \mathbf{x}$$

$$\hat{f}_{\text{student}}(\mathbf{x}) = \begin{cases} \mathbf{w}^\intercal \mathbf{x}, & n_{\text{data}} \geq d, \\ \mathbf{w}^\intercal \mathbf{Proj}\{\mathbf{X}_{\text{trn}}\}\, \mathbf{x}, & n_{\text{data}} < d. \end{cases}$$

Result 2: Bound on transfer risk

- If $n_{\text{data}} \geq d$, $\mathbb{E}_{\mathbf{X}_{\text{trn}}}[\text{transfer risk}] = 0$

Result 2: Bound on transfer risk

- If $n_{\mathrm{data}} \geq d$, $\mathop{\mathbb{E}}_{\mathbf{X}_{\mathrm{trn}}}[\text{transfer risk}] = 0$
- If $n_{\mathrm{data}} < d$, $\mathop{\mathbb{E}}_{\mathbf{X}_{\mathrm{trn}}}[\text{transfer risk}] \lesssim \left(\frac{\log n_{\mathrm{data}}}{n_{\mathrm{data}}}\right)^{\kappa}$

Result 2: Bound on transfer risk

- If $n_{\text{data}} \geq d$, $\mathbb{E}_{\mathbf{X}_{\text{trn}}}[\text{transfer risk}] = 0$
- If $n_{\text{data}} < d$, $\mathbb{E}_{\mathbf{X}_{\text{trn}}}[\text{transfer risk}] \lesssim \left(\frac{\log n_{\text{data}}}{n_{\text{data}}}\right)^{\kappa}$
- $\kappa \in [0, \infty)$, "easiness" of data distribution $P_{\mathbf{x}}$

Result 2: Bound on transfer risk

#254

- If $n_{\text{data}} \geq d$, $\mathbb{E}_{\mathbf{X}_{\text{trn}}}[\text{transfer risk}] = 0$
- If $n_{\text{data}} < d$, $\mathbb{E}_{\mathbf{X}_{\text{trn}}}[\text{transfer risk}] \lesssim \left(\frac{\log n_{\text{data}}}{n_{\text{data}}}\right)^{\kappa}$
- $\kappa \in [0, \infty)$, "easiness" of data distribution $P_{\mathbf{x}}$

Paper link:

http://proceedings.mlr.press/v97/phuong19a.html