Stanford University

CS 329P : Practical Machine Learning (2021 Fall)

12.2 Knowledge Distillation

Qingqing Huang, Mu Li, Alex Smola

https://c.d2l.ai/stanford-cs329p

Knowledge Distillation

- Use large models (teachers) to guide the training of small models (students), e.g.
  - Random forest → decision tree
  - ResNet-152 → ResNet-34
  - BERT-Base → BERT-mini
- Better than training students directly as teachers
  - Tell what they learned that are easier to train than the original data
  - Augment data with pseudo labels

Function Approximation to Distillation

- Teacher f learned by empirical risk minimization (ERM) on data $D_n = \{(x_i, y_i)\}_{i=1}^n$ sampled from p
- Learn student g close to f such as it generalizes better than just learn from D_n
- Given distance function d we can learn g^* by ERM again

$$\mathcal{F}(f, g, D_n) = \frac{1}{n} \sum_{i=1}^{n} d(f(x_i), g(x_i))$$

  - We pay twice for the statistical error due to sampling D_n from p, one for learning f, the other for distilling g^*
  - If we reduce the later, then g can be as good as f

Surrogate Approximation

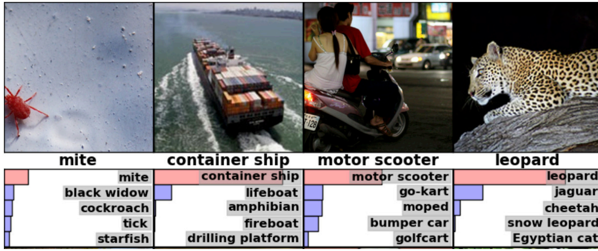
- Sample D'_m from surrogate distribution q such that $m \gg n$
- Under assumptions about search space and distance function, there exists constant V such that with probability at least $1 - \delta$

$$\underbrace{\mathscr{F}(f, g^*, p)}_{\text{Generalized error}} \leq \underbrace{\mathscr{F}(f, g^*, D'_m)}_{\text{Training error}} + \underbrace{\sqrt{(V - \log \delta)/m}}_{\text{As more data as possible}} + \underbrace{\|p - q\|_1}_{q \text{ should be close to } p}$$

Data Augmentation

- Sampling from q: perform augmentation, obtain labels from teacher, to
- Tabular (FAST-DAD):
  - For each feature/column i, estimate $p(x^i \mid x^{-i})$
  - Iterative sample features by Gibbs sampling
- Image: we learned how to do various image augmentations
- Text:
  - Use pre-trained BERT to fill randomly masked tokens
  - Other common ways such as back-translation, mixup

Distillation with Soft Targets



- Softmax outputs of negative classes contain information that are not available in labels
- $S_T(\mathbf{x})$ is the softmax output with temperature T: $\exp(x_i/T) \,/\, \sum_j \exp(x_j/T)$
- Match student's softmax outputs with teacher's and also label

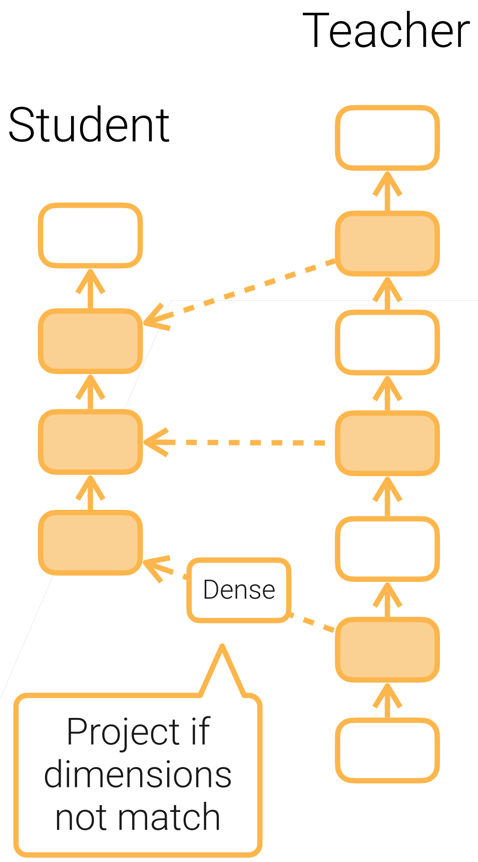
$$\mathrm{CE}(S_T(g(\mathbf{x})), S_T(f(\mathbf{x})) + \lambda\, \mathrm{CE}(S_1(g(\mathbf{x})), y)$$

A larger T make it equals to $\mathrm{MSE}(g(\mathbf{x}), f(\mathbf{x}))$, though $T = 1$ often works well

Normal classification objective

Distillation with Intermediate Representations

- Neural network hidden outputs have richer information than the output layer
  - Match student layers to teacher layers
  - Add a dense layer if layer output dimensions do not match
  - Loss can be MSE, L2, or even be learned

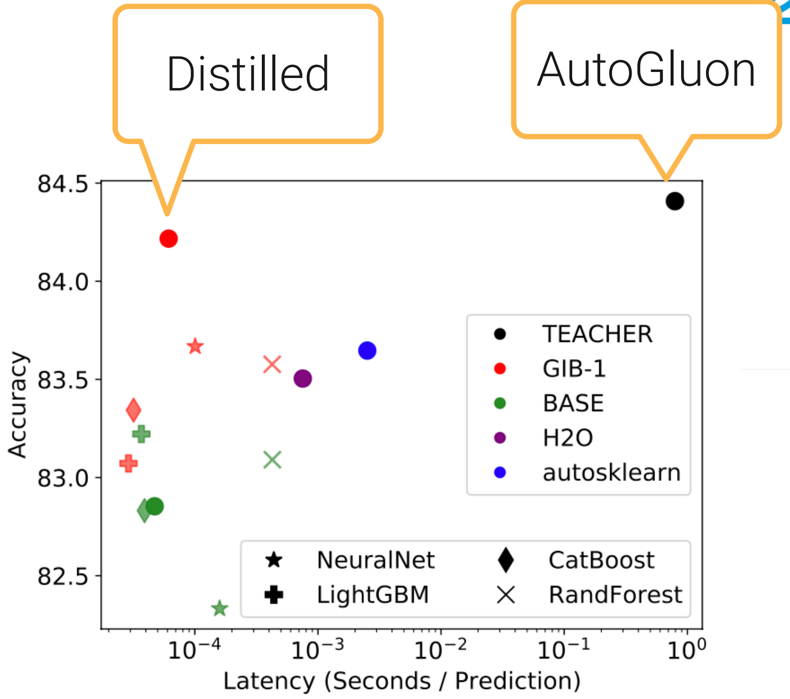


Results on Tabular

```
from autogluon.tabular import TabularPredictor

predictor = TabularPredictor(label=label).fit(...)
distilled = predictor.distill()
```

Sample from the data used to train predictor

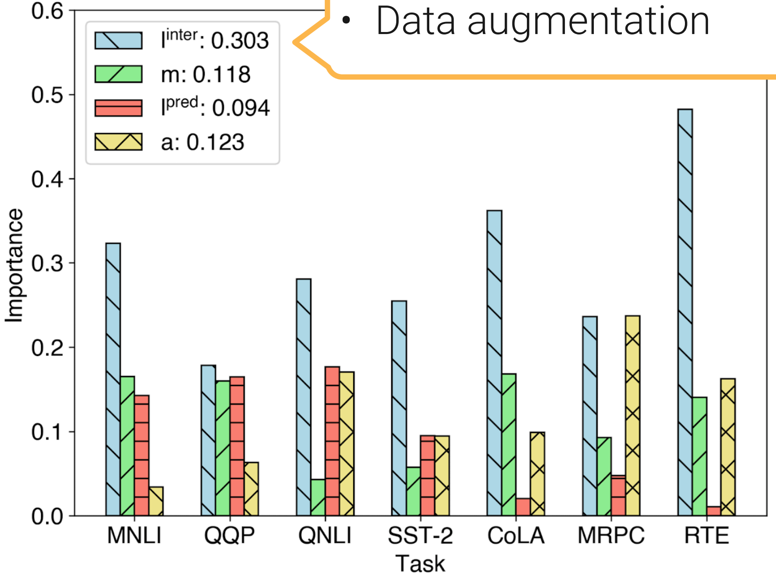


Averaged results on 30 datasets
(Fakoor et.al, NeurIPS'20)

Results on Text

- Distill Bert-base (110M parameters) to TinyBert (14M)
- The averaged accuracy on GLUE (9 tasks)
  - Bert-base: 79.6, TinyBERT 75.1

- Intermediate representations loss
- Matching layers strategies
- Output layer loss
- Data augmentation



Importance of each component (He et.al. 2021)

Summary

- Distill knowledges from teach models to students models
  - Teach models are big neural networks or model combinations
  - Students are smaller but have a similar generalization error with teachers