Информативные априорные предположения в задаче привилегированного обучения

Нейчев Радослав Георгиев

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель:
д.ф.-м.н, н.с. ВЦ РАН В.В. Стрижов

6 июня 2018

Цели исследования

Цель исследования: создать метод построения моделей оптимальной сложности для задач распознования и прогнозирования

Проблема: Неустойчивая сходимость моделей в зависимости от начальной инициализации параметров. Привилегированное обучение - способ уточнения структуры и параметров модели за счет привлечения дополнительной информации.

Задача: Сформулировать метод построения моделей, который позволит:

- соблюдать баланс между точностью и сложностью модели;
- использовать дополнительную (априорную) информацию на этапе обучения;
- использовать неполные признаковые описания объектов (т.е. работать и с объектами, априорная информация о которых отсутствует).

Основная литература

- Vladimir Vapnik and Rauf Izmailov.
  Learning using privileged information: Similarity control and knowledge transfer.
  JMLR, 16:2023–2049, 2015.
- Bernhard Schölkopf Vladimir Vapnik David Lopez-Paz, Léon Bottou.
  Unifying distillation and privileged information.
  ICLR, 2016.
- Ben Poole Eric Jang, Shixiang Gu.
  Categorical reparameterization with gumbel-softmax.
  ICLR, 2017.
- В.В. Стрижов А.А. Адуенко.
  Совместный выбор объектов и признаков в задачах многоклассовой классификации коллекции документов.
  Инфокоммуникационные технологии, 1:47–53, 2014.

Постановка задачи декодирования, прогнозирования и классификации

Матрица плана $\bar{\mathbf{X}}$, где $\mathbf{Y}$ содержит метки классов, распределения или целевые значения в зависимости от задачи.

$$\bar{\mathbf{X}} = \left[\begin{array}{c|ccc} \hat{\mathbf{y}}' & \mathbf{x}'_0 & \mathbf{x}''_0 & \cdots \\ \hline \mathbf{y}'_1 & \mathbf{x}'_1 & \mathbf{x}''_1 & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{y}'_m & \mathbf{x}'_m & \mathbf{x}''_m & \cdots \end{array}\right] = \left[\begin{array}{c|c} \underset{1\times r}{\hat{\mathbf{y}}} & \underset{1\times n}{\mathbf{x}_0} \\ \hline \underset{m\times r}{\mathbf{Y}} & \underset{m\times n}{\mathbf{X}} \end{array}\right].$$

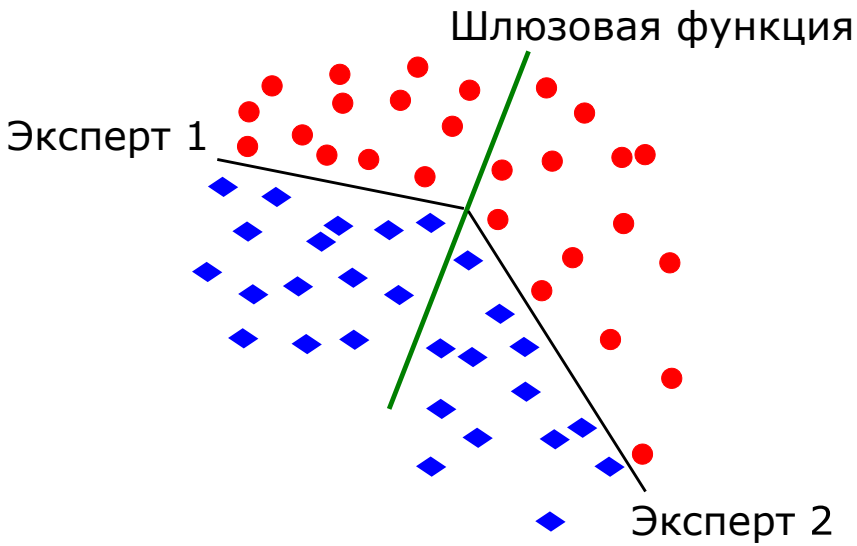
Оптимальная модель $\hat{\mathbf{f}} : \mathbf{X} \to \mathbf{Y}$ минимизирует заданную функцию ошибки $S(\mathbf{f}, \mathbf{X}, \mathbf{Y})$ при ограничении на сложность:

$$\hat{\mathbf{f}} = \operatorname{argmin}_{\mathbf{f}} S(\mathbf{f}, \mathbf{X}, \mathbf{Y}) \Big| |\hat{\mathbf{f}}|_c \leq M_c$$

Предлагается использовать *привилегированную и априорную информацию* при построении $\hat{\mathbf{f}}$.

Смесь экспертов в качестве мультимодели

Пусть для описания данных используются K моделей.
Шлюзовая функция (англ. *gating function*) — отображение $\pi_k(\mathbf{x}) : X \to [0, 1]$, определяющая правдоподобие k-й модели на векторе $\mathbf{x} \in \mathbf{X}$, где $\mathbf{X}$ есть признаковое пространство.



В качестве шлюзовой функции может быть использован softmax (σ):

$$\pi_k(\mathbf{x}, \mathbf{V}) = \frac{e^{\mathbf{v}_k^\top \mathbf{x}}}{\sum_{i=1}^{K} e^{\mathbf{v}_i^\top \mathbf{x}}}, \quad \mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_K], \quad \mathbf{f} = \sum_{i=1}^{K} \pi_k(\mathbf{x}, \mathbf{V}) f_k(\mathbf{x}, \mathbf{w}_k)$$

Прогноз экспертов $f_1, \ldots, f_K$ с учетом *гауссовского* шума:

$$\mathbf{y} = f_k(\mathbf{x}, \mathbf{w}) + \varepsilon, \quad \mathbf{y} \sim \mathcal{N}(f_k(\mathbf{x}, \mathbf{w}), \beta_k).$$

Обозначим вектор гиперпараметров за $\boldsymbol{\theta}$:

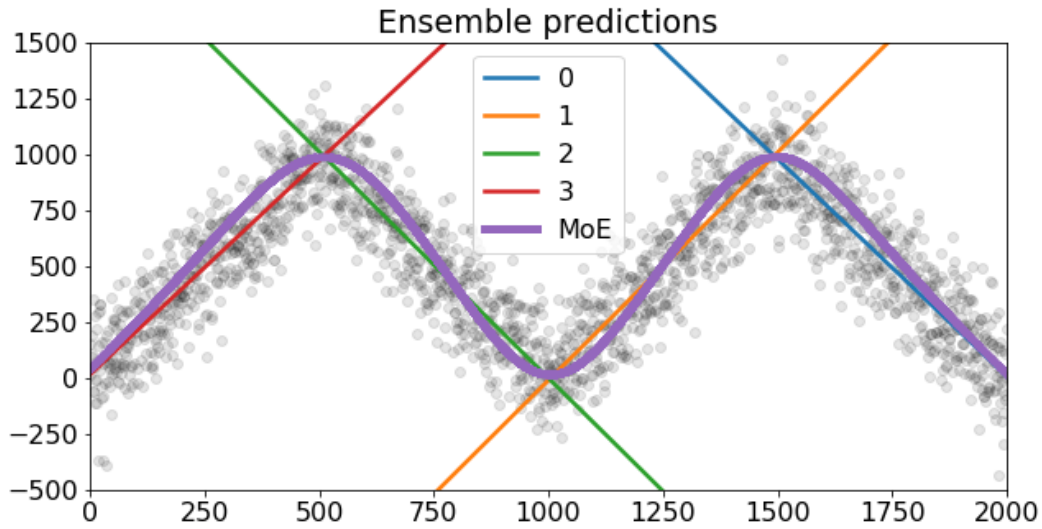
$$\boldsymbol{\theta} = [w_1, \ldots, w_K, \mathbf{V}, \boldsymbol{\beta}]$$

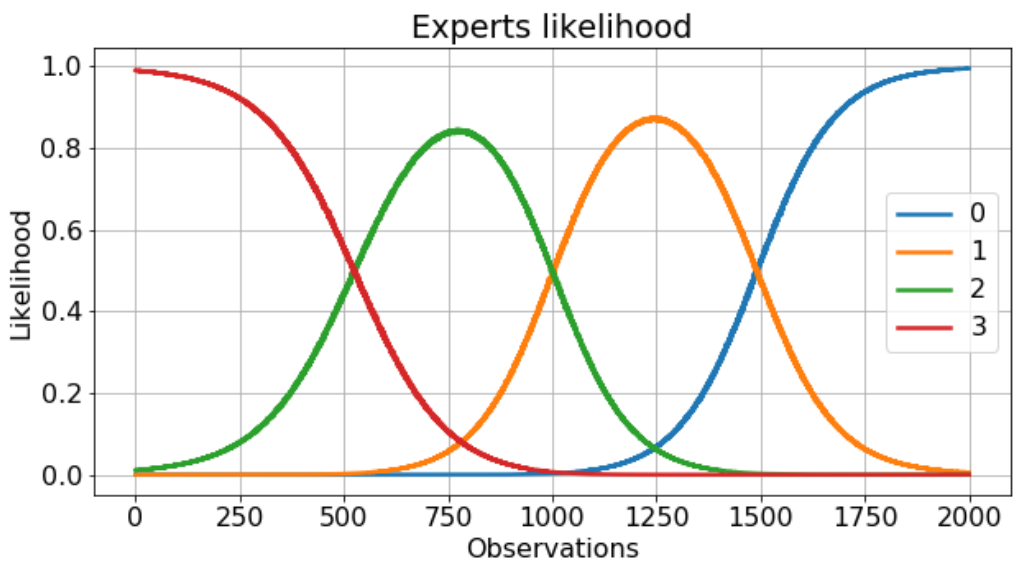
Правдоподобие f_k на паре $(\mathbf{x}, \mathbf{y})$ обозначим $p(k|\mathbf{x}, \mathbf{w})$.
Апостериорное распределение на $\mathbf{y}$:

$$p(y|\mathbf{x}, \boldsymbol{\theta}) = \sum_{k=1}^{K} p(y, k|\mathbf{x}, \boldsymbol{\theta}) = \sum_{k=1}^{K} p(k|\mathbf{x}, \boldsymbol{\theta}) p(y|k, \mathbf{x}, \boldsymbol{\theta}) =$$

$$= \sum_{k=1}^{K} \frac{\exp(\mathbf{v}_k^{\mathsf{T}}\mathbf{x})}{\sum_{k'=1}^{K} \exp(\mathbf{v}_{k'}^{\mathsf{T}}\mathbf{x})} \exp\left(-\frac{1}{2\beta_k}(\mathbf{y} - f_k(\mathbf{x}, bw))^2\right).$$

Модели $f_1, \ldots, f_K$ и шлюзовая функция $\pi_k(\mathbf{x}, \mathbf{V})$ обучаются с помощью EM-алгоритма.





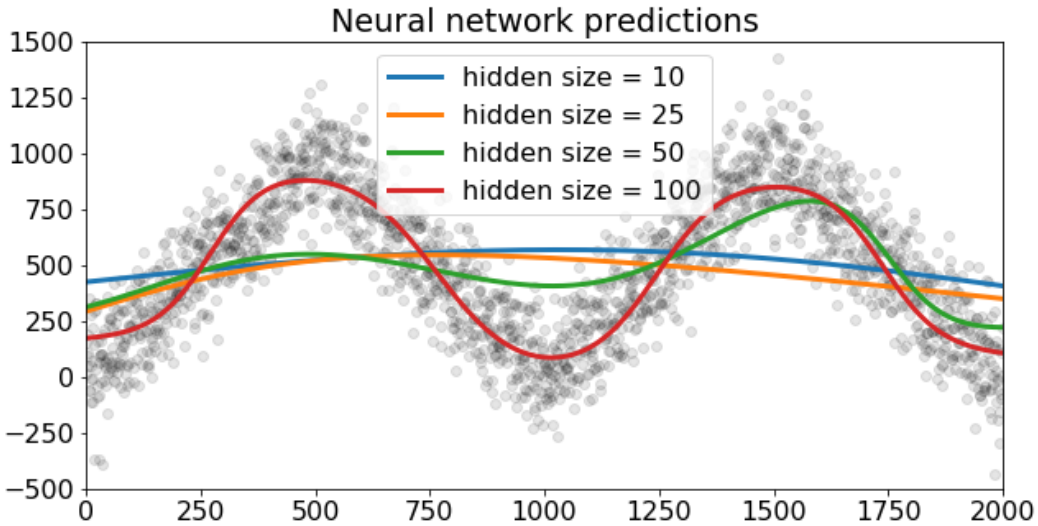
Синтетические данные.

$$f_i = w_i \mathbf{x} + b_i$$

$\pi_k(\mathbf{x}, \mathbf{V})$ — нейросеть с 1 скрытым слоем из 50 нейронов.

$$\mathbf{f}_e = \sum_{i=1}^{K} \pi_k f_k$$

$$|\mathbf{f}_e|_C \sim 10^2$$



Синтетические данные.

$$|\mathbf{f}_{\text{nn}}|_C \sim 10^4$$

Сравнение нейросетей (два скрытых слоя) с различным числом нейронов. Лишь экземпляр $\mathbf{f}_{\text{nn}}$ с размером скрытого слоя 100 смог верно описать данные.

$$S(\mathbf{f}_{\text{e}}) \approx S(\mathbf{f}_{\text{nn}}), \quad |\mathbf{f}_{\text{e}}|_C \ll |\mathbf{f}_{\text{nn}}|_C$$

Проблема: Мультимодель $\mathbf{f}_{\text{e}}$ очень плохо сходится ($\sim$ в 10% запусков).

Мета-обучение (distillation)

Для простоты, рассмотрим задачу классификации, Δ^c — вектор вероятностей, $\sum_{i=1}^{c} \Delta_i^c = 1$.
Пусть для некоторых объектов $\mathbf{x} \in \mathbf{X}$ доступна *привелегированная* информация $\mathbf{x}^* \in \mathbf{X}^*$. Введем функции ученика $\mathbf{f}_\mathsf{s} \in \mathcal{F}_\mathsf{s}$ (student) и учителя $\mathbf{f}_\mathsf{t} \in \mathcal{F}_\mathsf{t}$ (teacher):

$$\mathbf{f}_\mathsf{s} : \mathbf{X} \longrightarrow \mathbf{Y}, \quad \mathbf{f}_\mathsf{t} : \mathbf{X} \oplus \mathbf{X}^* \longrightarrow \mathbf{Y}$$

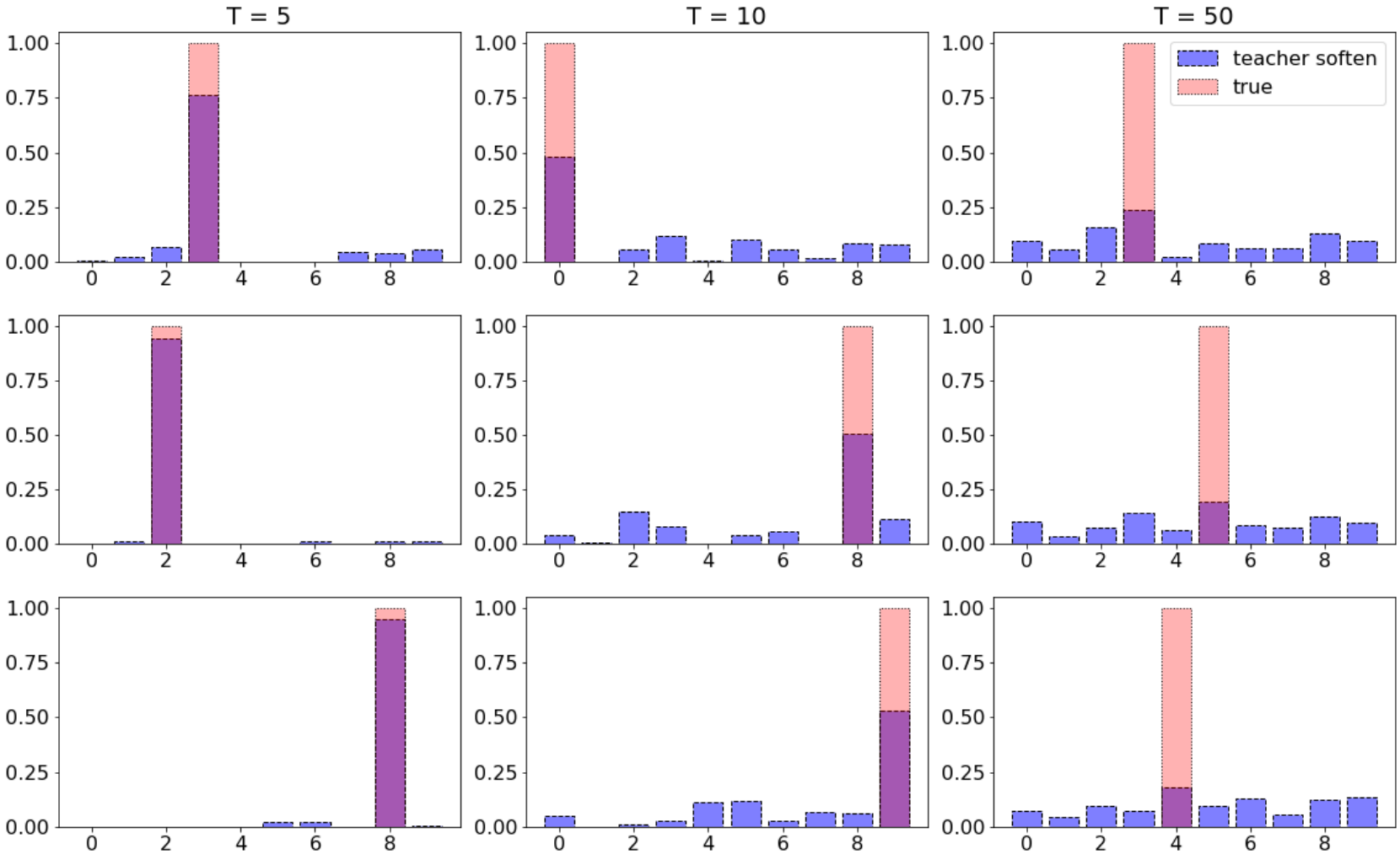
$$\mathbf{f}_\mathsf{s} = \operatorname{argmin}_{\mathbf{f} \in \mathcal{F}_s} \frac{1}{n} \sum_{i=1}^{n} \Big[(1-\lambda)\ell(\mathbf{y}_i, \sigma(\mathbf{f}(\mathbf{x}_i))) + \lambda \ell(\mathbf{s}_i, \sigma(\mathbf{f}(x_i))) \Big],$$

где

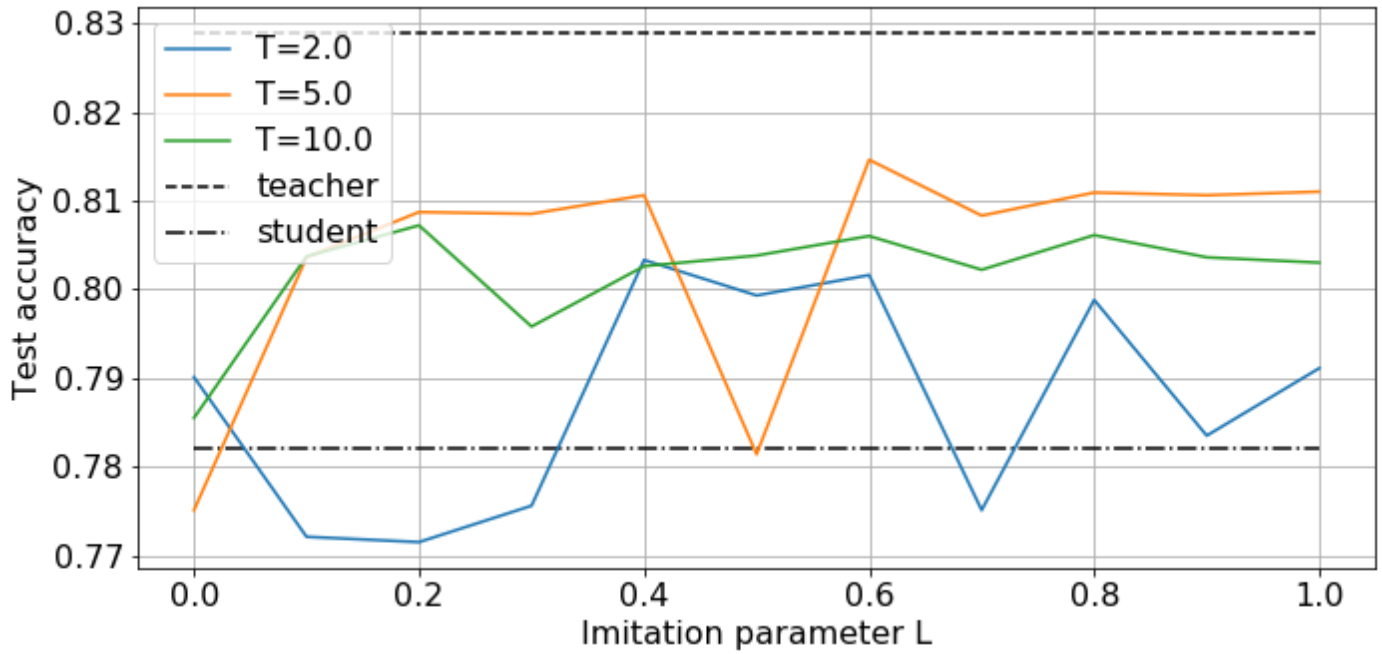
$$\mathbf{s}_i = \sigma(\mathbf{f}_\mathsf{t}(x_i)/T) \in \Delta^c, \quad \ell(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{k=1}^{c} \mathbf{y}_k \log \hat{\mathbf{y}}_k,$$

В случае $|\mathcal{F}_\mathsf{t}|_C \gg |\mathcal{F}_\mathsf{c}|_C$, $\mathbf{X}^* = \emptyset$ — дистилляция (Хинтон).
В случае $|\mathcal{F}_\mathsf{t}|_C \ll |\mathcal{F}_\mathsf{c}|_C$, $\mathbf{X}^* \neq \emptyset$ — привилегированное обучение (Вапник).

Иллюстрация сглаженных предсказаний учителя $\mathbf{s}_i$ в зависимости от значения параметра T на примере классификации датасета MNIST.



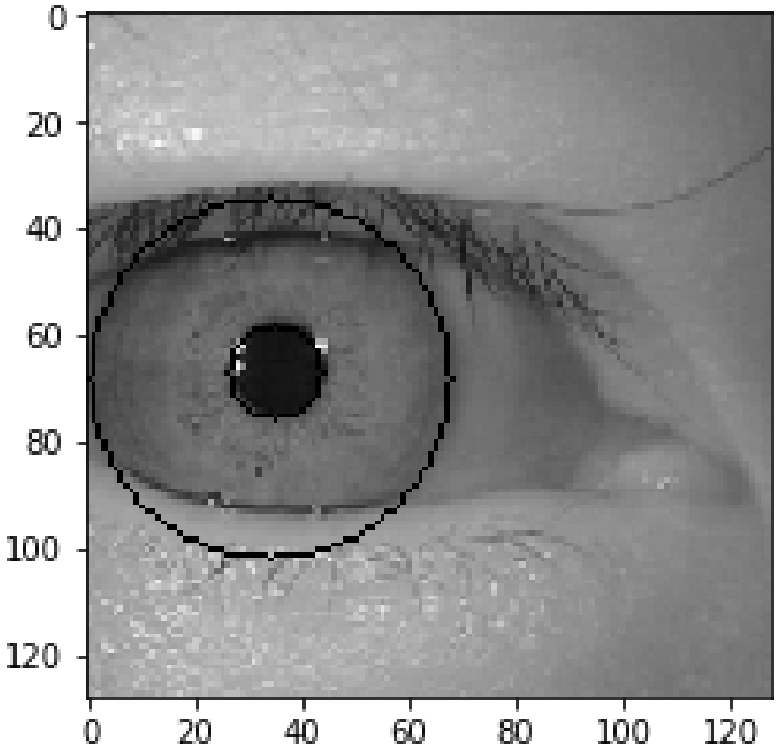
Качество классификации ученика, обученного методом дистилляции в зависимости от параметров T и L.

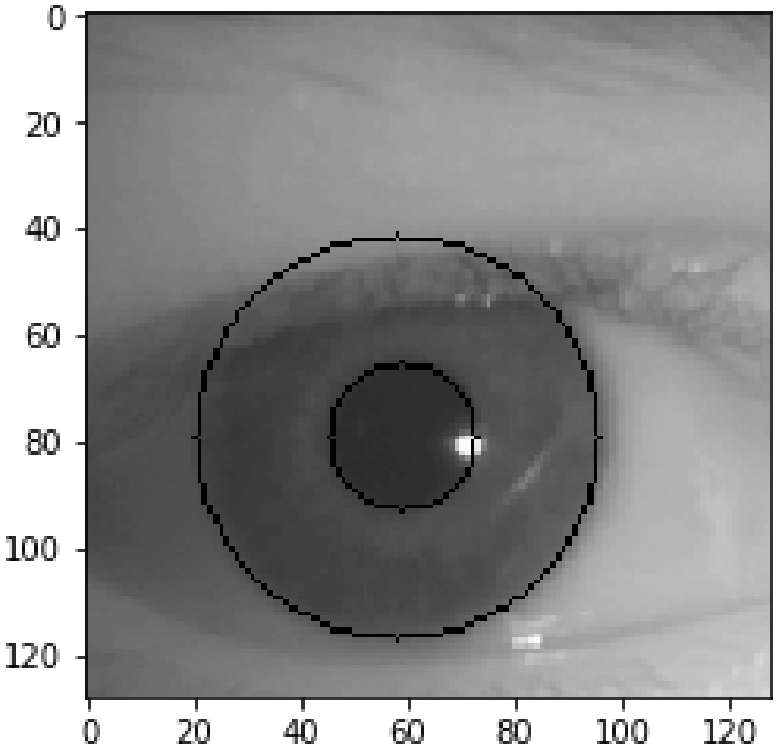


Обучающая выборка — 500 изображений из датасета MNIST, $\mathbf{x}^*$ — исходные изображения, $\mathbf{x}$ — изображения с разрешением в 4 раза меньше, $\mathbf{f}_t$ и $\mathbf{f}_s$ — нейросети с двумя скрытыми слоями из 50 нейронов и ReLU-активациями. Число параметров ученика значительно меньше, чем учителя:

$$|\mathbf{f}_s|_C = 1.5 \cdot 10^3 \ll 1.5 \cdot 10^4 = |\mathbf{f}_t|_C$$

Определение границы радужки





Coming soon!

Публикации по теме

- Выбор оптимального набора признаков из мультикоррелирующего множества в задаче прогнозирования. *Нейчев Р.Г., Катруца А.М., Стрижов В.В.* / Заводская лаборатория. №3 2016. Том 2.
- Heterogeneous model selection for multiscale time series forecasting. *Radoslav Neychev, Anastasia Motrenko, Eric Gaussier and Vadim Strijov* / Рассматривается редколлегией Journal of Applied Mathematics and Computation.

Сопутствующие результаты

Разработана и запущена в эксплуатацию автоматическая система прогнозирования энергопотребления ДЦ компании Яндекс.

Заключение

- Предложен метод построения модели меньшей сложности с использованием привилегированной информации.
- Предложенный метод позволяет использовать частично доступные данные.
- *Предложен метод автоматического подбора параметров дистилляции на основе анализа энтропии функции-учителя.
- *Создан фреймворк, позволяющий применять дистилляцию для обучения сторонних моделей.

Backup

Определения

Априорные знания — информация о предметной области/ограничениях на решение, не представленная в обучающей выборке в явном виде.
Привилегированная информация — дополнительная информация об объектах обучающей выборки, доступная только на этапе обучения.
Сложность модели $|f|_C.|.|_C$ используется оценка числа простейших арифметических операций на единичном входе