Машинное обучение (Machine Learning)

Distillation and Batch Normalization

Уткин Л.В.

Санкт-Петербургский политехнический университет Петра Великого

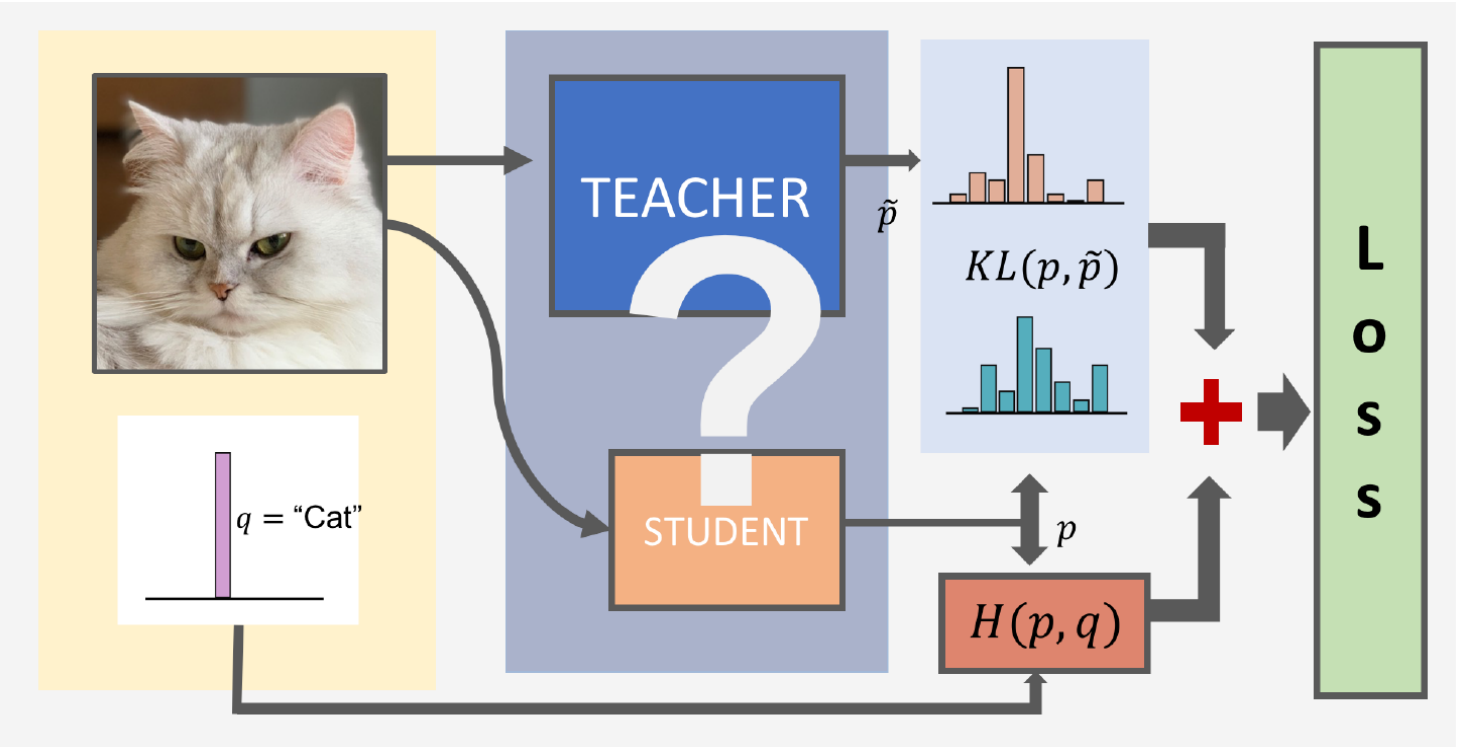
Distillation (idea)

- G. Hinton, O. Vinyals, J. Dean. Distilling the Knowledge in a Neural Network, arXiv:1503.02531.
- Мотивация: использовать результаты обучения "сложной" модели для обучения "простой"
- Отличие transfer learning и distillation: в последнем перенос обобщения (transfer of generalization)
- Понятия: сети учитель и студент, понятие температуры в softmax, "темные" знания (dark knowledge) и мягкие вероятности

Distillation (модель учитель-студент)

- Учитель - это “сложная” глубокая нейронная сеть, которая была обучена на большом количестве данных (или любая другая модель - ансамбль) с хорошим обобщением
- Студент - это “простой” сеть, цель - выучить большинство обобщений учителя и оставаться “простой”

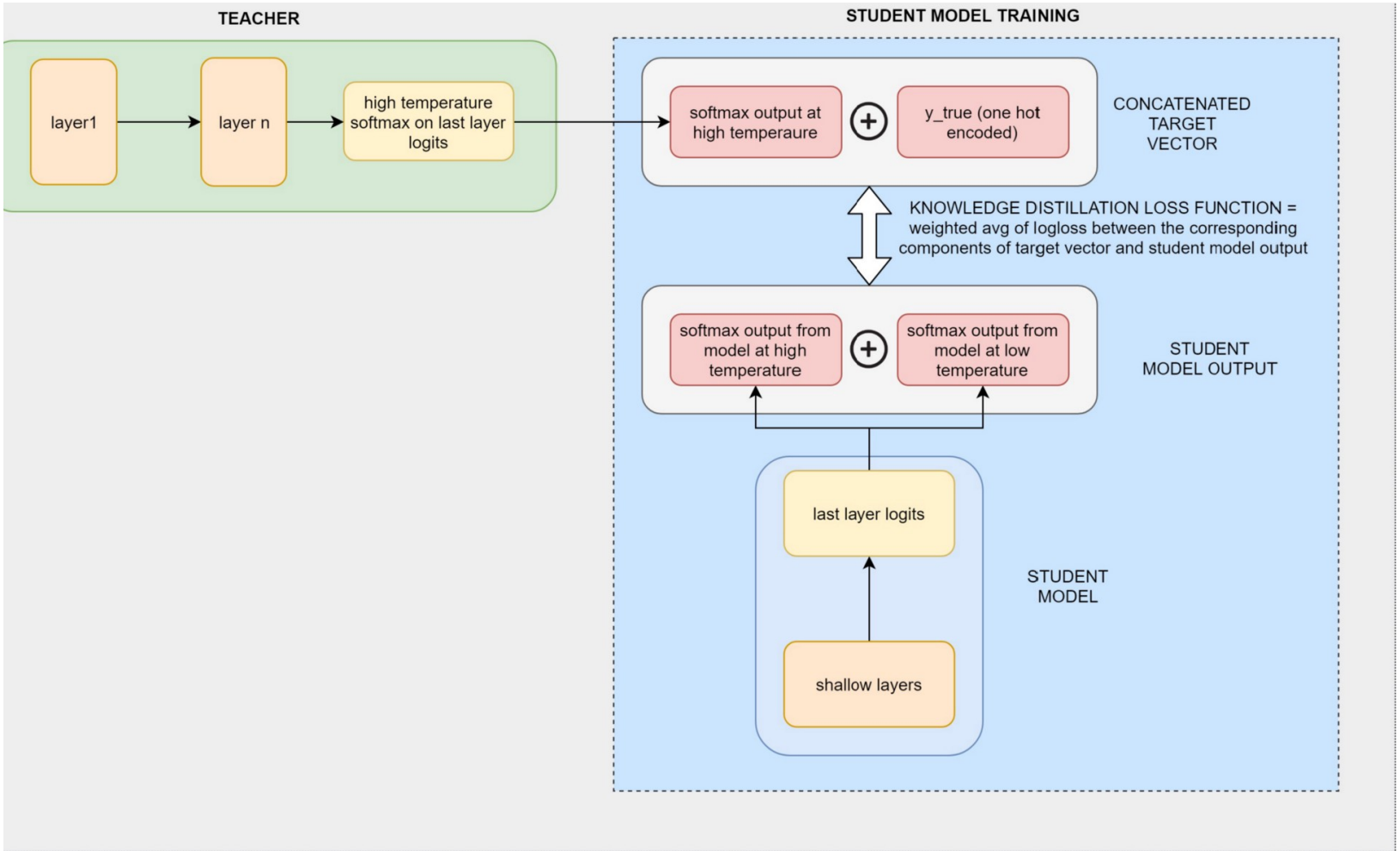
Distillation (модель учитель-студент)



J.H. Cho and B. Hariharan. On the Efficacy of Knowledge Distillation.

arXiv:1910.01348v1

Distillation (модель учитель-студент)



Distillation (температура)

- Softmax возвращает вероятности каждого класса от 0 до 1, и их сумма =1, целевой класс имеет высокую вероятность

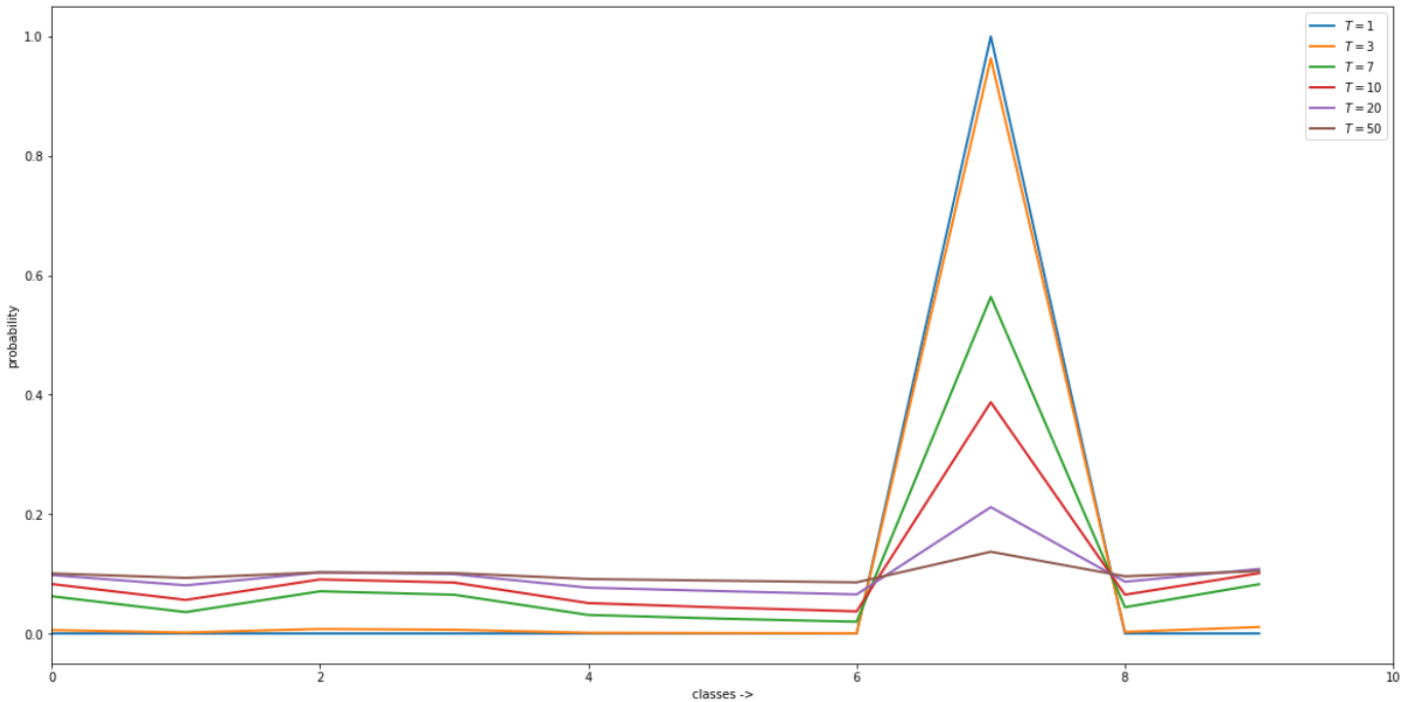
$$p_t(z_i) = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$$

- Softmax с температурой

$$p_t^*(z_i, T) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

- Больше температура - более размыты вероятности классов

Distillation (температура и MNIST)



Вероятности цифр MNIST, классифицируется цифра 7

Distillation (dark knowledge)

один семь или один?

Distillation (dark knowledge)

- Модель дает более высокую вероятность для 1 одновременно прогнозируя 7 при высокой T
- Человек не может количественно определить, насколько 7 выглядит ближе к 1, а “высокотемпературная” модель делает это
- Т.о. “высокотемпературная” модель обладает “темными” знаниями - в дополнение к предсказанию числа 7, она также хранит информацию о том, насколько это число 7 напоминает число 1
- “Низкотемпературная” модель (обычная модель) хороша для точных прогнозов, но теряем эти “темные” знания
- Основная идея distillation - передача “темных” знаний от обученного учителя к простой модели студента

Distillation (обучение студента)

- Модель студента обучается при той же высокой температуре, что и учитель
- Функция потерь для студента

$$L = \alpha L_{\text{cross entopy}} + (1 - \alpha) L_{\text{knowledge distil.}}$$

$$L_{\text{knowledge distil.}} = -\tau \sum_i p_t^*(z_i, T) \ln p_s^*(z_i, T)$$

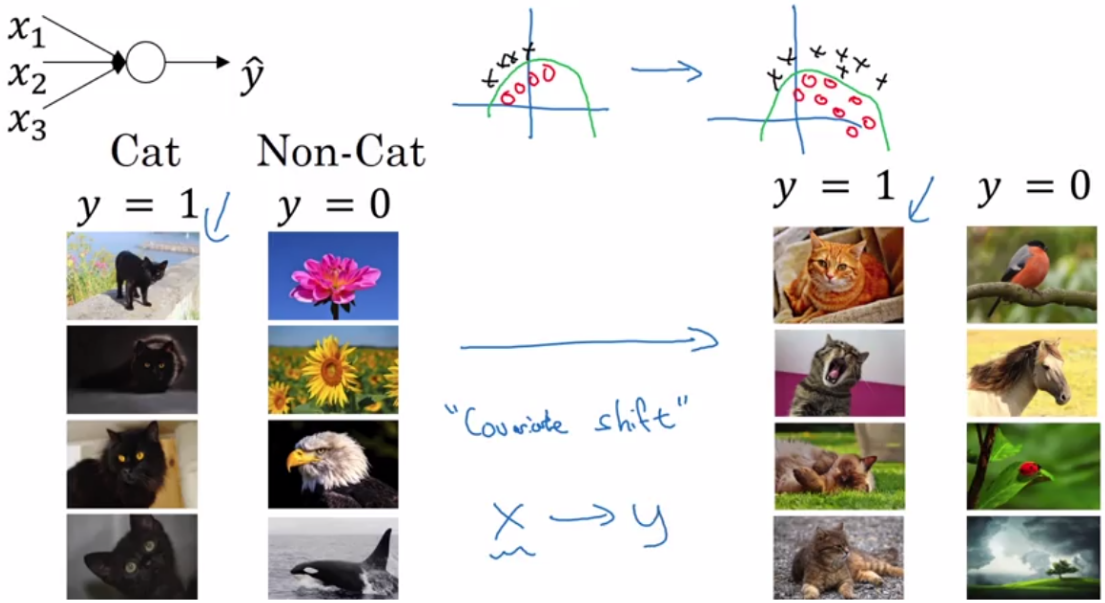
- Модель студента тестируется с обычной активацией softmax (т.е. без температуры).

Batch normalization (пакетная нормализация, BN) - зачем

- Есть функции активации от 0 до 1, а есть от 1 до 1000
- Если нормализуем входной слой, почему не сделать это для всех или части слоев
  - Это добавляет некторый шум к активациям аналогично dropout (регуляризация)
  - Уменьшает смещение
  - Делает слои сети более независимыми от других слоев
  - Высокая скорость обучения, т.к. нет очень больших или малых активации

BN - уменьшает смещение

Сеть по классификации кошек: обучаем только на черных кошках. Если применить сеть к цветными кошками, то будут ошибки. Обучающий и тестовый датасеты немного различаются. Batch normalization уменьшает смещение



Deeplearning.ai: Why Does Batch Norm Work? (C2W3L06)

BN - как

BN добавляет два обучаемых параметра к каждому слою: γ и β, что позволяет SGD выполнять денормализацию, изменяя только эти два веса для каждой активации, вместо потери стабильности сети путем изменения всех весов

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β
Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma^2_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma^2_{\mathcal{B}} + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

S. Ioffe, C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift // arXiv:1502.03167v3

BN - зачем gamma и beta

- Если использовать BN в предобученной сети, то это изменит обученные веса (плохо)
- Поэтому нужно определить γ и β, чтобы отменить изменение выходных данных

Вопросы

?