Байесовская дистилляция моделей на базе трансформеров

И. Н. Игнашин

Московский физико-технический институт

16 декабря 2023 г.

Слайд об исследованиях

Исследуется проблема дистилляции моделей. То есть понижения сложности аппроксимирующих моделей.

Цель исследования —

Предложить метод дистилляции модели трансформера, а также дистилляции RNN модели с Attention.

Требуется предложить

1) метод удаления слоя Attention в модели RNN,
2) метод удаления слоя Attention в модели трансформера

Решение

Предложен способ удаления аддитивного внимания в модели RNN.

Постановка задачи в случае полносвязной нейросети

Заданы

1) Выборка $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^m \quad x_i \in \mathbb{R}^n \quad y_i \in \mathbb{Y}$,
2) Модель учителя $f = \sigma \circ U_T\sigma \circ ...U_2\sigma \circ U_1 \quad u = vec([U_T, U_{T-1}, ..U_1])$
3) Модель ученика
   $f = \sigma \circ U_T\sigma \circ ...U_k\sigma \circ U_{k-2}...U_2\sigma \circ U_1 \quad w = vec([U_T, U_{T-1}, ..U_1])$
4) Апостериорные распределния параметров учителя $p(u|D)$

Требуется найти зависимость параметров априорного распределения модели ученика $p(w|A)$ от апостериорного распределения параметров модели учителя $p(u|\mathcal{D})$

[1] *Грабовой А.В.* Априорное распределение параметров в задачах выбора моделей глубокого обучения , 2022

Предложенный способ

Заданы

1) Апостериорное распределение модели учителя RNN : $p(u|\mathcal{D})$, по предположению нормальное.
2) Аддитивное внимание в модели учителя $a(h, h') = w^T th(Uh + Vh')$.

Удаление аддитивного внимания посредством зануления некоторых параметров, приводящее к обычной RNN без внимания.

Можно занулять параметры различными способами:

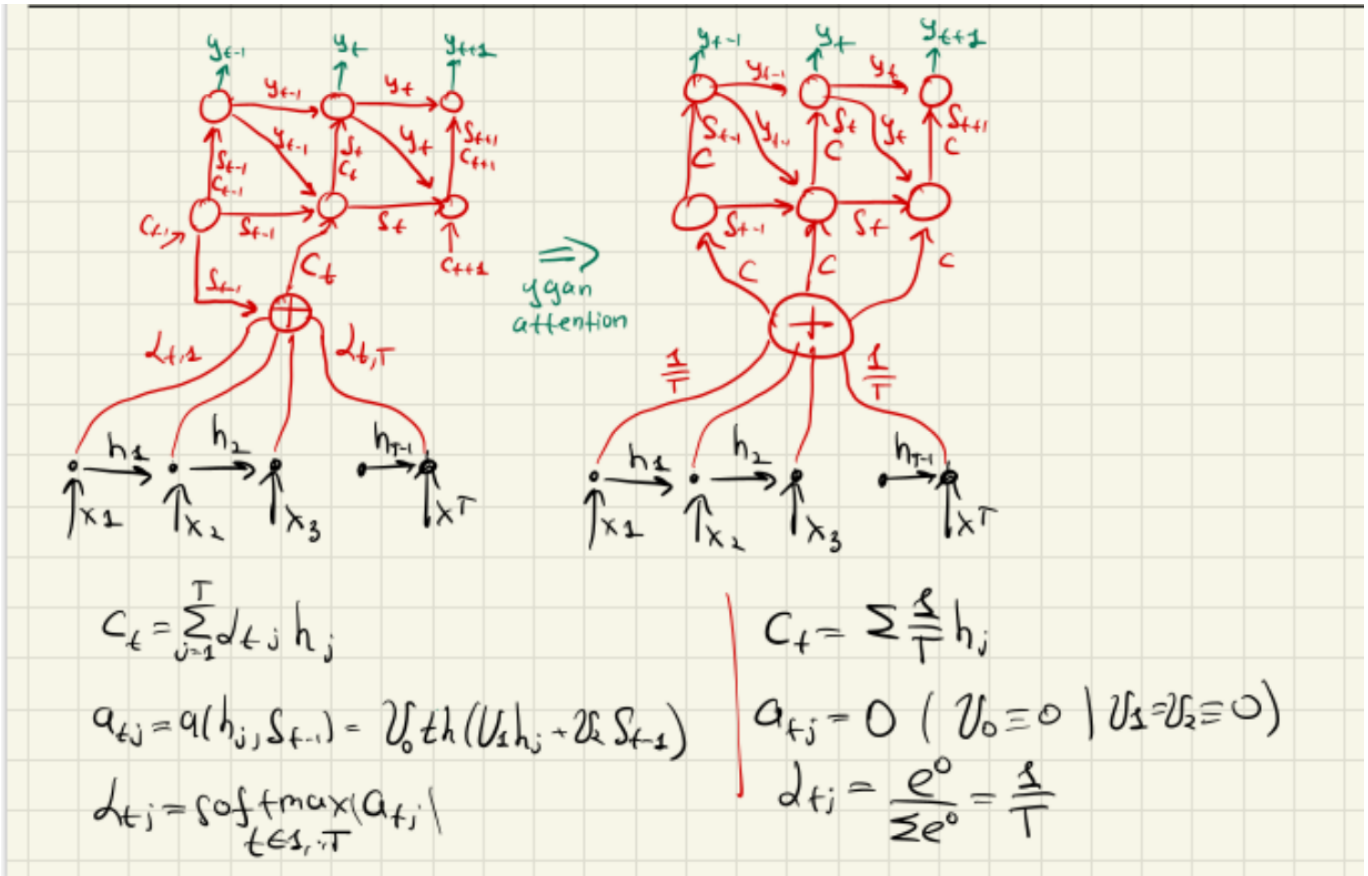
1. $w = 0$
2. $U = 0$ и $V = 0$
3. Оставшиеся способы с занулением w, U, V , приводящие к $a(h, h') \equiv 0$

Априорное распределение для модели учителя получается нормальным с параметрами, вычисляемыми аналитически различными способами:

- $p(U^*|\mathcal{D}) = \int\limits_{U_1, U_2} p([U^*, U_!, U_2]|\mathcal{D}) dU_1 dU_2$
- $p(U^*|\mathcal{D}) = p(U|\mathcal{D}, U_0, U_1, U_2 = 0)$
  $\sim \mathcal{N}(m_* + R_{U,U'} R_{U'}^{-1}(0 - m_{U'}), R_U - R_{U,U'} R_{U'}^{-1} R_{U',U})$

Анализ предложенного способа

Зануление параметров приводит к $a(h, h') = 0$.
Это в свою очередь приводит к $softmax(a_{t,j}) = \frac{1}{T}$.
Следовательно получается структура RNN без внимания.



Выводы

1. Предложен способ дистилляции модели RNN.
2. Ожидается большее качество у дистиллированной модели, чем у произвольно инициализированной, той же структуры.
3. В случае трансформеров происходит поиск способа дистилляции за счет удаления attention слоя.