Байесовская дистилляция моделей на базе трансформеров

И. Н. Игнашин

Московский физико-технический институт

18 мая 2024 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 метод удаления аддитивного внимания в модели RNN.
Предложен метод удаления слоя энкодера/декодера в модели трансформера.
Предложен метод уменьшения числа голов в модуле multi-head-attention трансформера

Постановка задачи в случае машинного перевода

Заданы

1) Выборка $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$, где $x_i = [x_i^1, x_i^2, ..., x_i^{k_i}]$, $x_i^k \in \overline{0, 1, ..., v_{rus}}$, $y_j = [y_i^1, y_i^2, ..., y_i^{l_j}]$, $y_i^k \in \overline{0, 1, ..., v_{eng}}$
2) Модель учителя (RNN with attention/Transformer)
3) Модель ученика (RNN/Transformer)
4) Апостериорные распределния параметров учителя $p(u|D)$

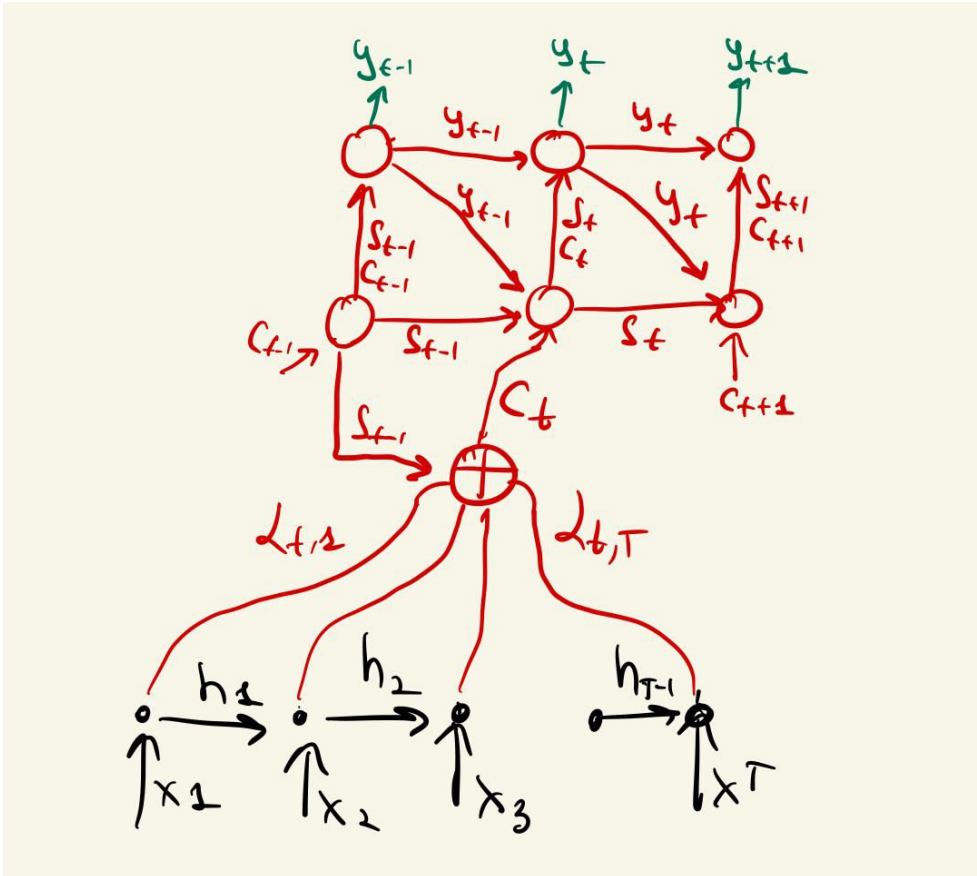
Требуется найти зависимость параметров априорного распределения модели ученика $p(w|A)$ от апостериорного распределения параметров модели учителя $p(u|\mathcal{D})$.

Модель RNN

В качестве модели учителя используется модель RNN с вниманием следующей структуры:

$$a_{t,j} = a(h_j, s_{t-1}) = U_0 \tanh(U_1 h_j + U_2 s_{t-1})$$

$$c_t = \sum_{j=1}^{T} soft \max_{t \in 1,2,\ldots,T} (a_{t,j}) h_j$$



Модель трансформера

В качестве модели учителя взята стандартная модель трансформера.
Энкодер представим в виде преобразований:

$$H'' = Enc(H) \quad H \in \mathbb{R}^{T \times d_{model}} \tag{1}$$

$$Enc = E_1 \circ E_2 \circ, ..., E_{N_{enc}} \tag{2}$$

$$E_i = I + norm \circ FFN \circ (I + norm \circ SA) \tag{3}$$

$$SA = dropout \circ MHA \tag{4}$$

$$FFN(H_t) = dropout \circ (\sigma \circ (H_t \cdot W_1 + b_1)W_2 + b_2) \tag{5}$$

$$head_i = soft\max(\{\frac{Q_i K_i^T}{\sqrt{d_k}}\})V_i \tag{6}$$

$$MHA(H) = concat(head_1, head_2, ..., head_{N_{heads}}) \cdot W_o, \tag{7}$$

где E_i – слои энкодера, H – последовательность T входных эмбеддингов.

[1] Attention Is All You Need , 2017

Предложенный метод

Заданы

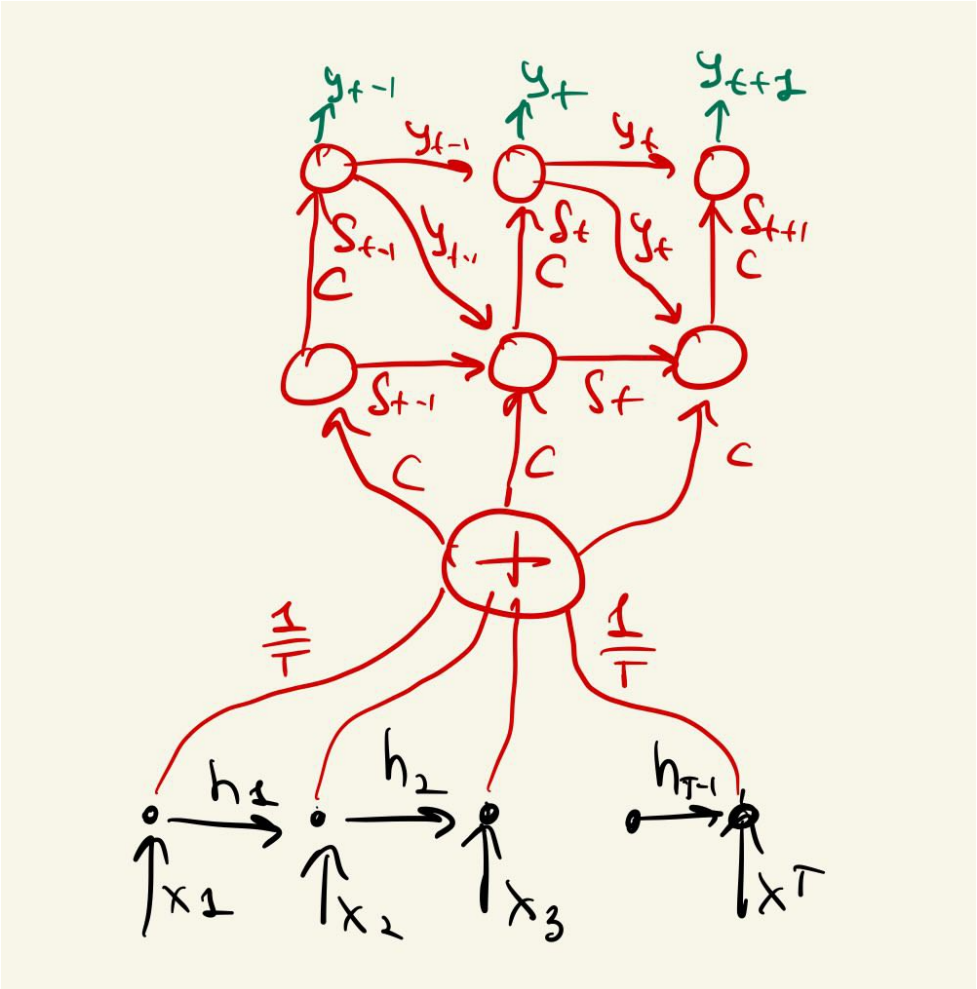
1) Апостериорное распределение модели учителя RNN : $p(u|\mathcal{D})$, по предположению нормальное.
2) Аддитивное внимание в модели учителя $a(h, h') = U_0^T th(U_1 h + U_2 h')$.

Для эквивалентности модели учителя и модели ученика достаточно занулить только вектор U_0 или только пару векторов U_1, U_2, либо все сразу.
Апостериорное распределение для модели учителя получается нормальным с параметрами, вычисляемыми аналитически различными способами:

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

Анализ предложенного метода

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



Предложенный метод для трансформеров

Theorem

Заданы

1) *Апостериорное распределение модели учителя :* $p(u|\mathcal{D}) = \mathcal{N}(\mu, \Theta)$.
2) *Модель трансформера имеет число голов в multi-head-attention большее одной или число слоев энкодера/декодера большее одного*
3) *Удаляемые и зануляемые параметры:* v_1 *и* v_2

Тогда апостериорное распределение параметров преобразованной модели учителя описывается нормальным распределением

$$m = \mu_z + \Theta_{z,v_2}\Theta_{v_2}^{-1}(0 - \mu_z)$$
$$R = \Theta_z - \Theta_{z,v_2}\Theta_{v_2}^{-1}\Theta_{v_2,z}$$
$$p(\nu|\mathcal{D}) = \mathcal{N}(m_\nu, R_{\nu,\nu}),$$

где $z = [\nu, v_1]$ *,*R *– ковариационная матрица между векторами* z, v_2

Анализ предложенного метода

Для эквивалентности модели учителя и модели ученика достаточно:

- Занулить параметры модуля feed-forward и удалить остальные параметры модулей в слое энкодера/декодера.
- Занулить часть параметров последнего линейного преобразования в multi-head-attention и удалить все параметры, соответствующие данной голове.

Выбранные удаляемые и зануляемые параметры подставляютя в условия теоремы для получения апостериорного распределения модели учителя, которое используется в качестве априорного распределения модели ученика.

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.