Байесовская дистилляция моделей на базе трансформеров

Игорь Николаевич Игнашин

Московский физико-технический институт

Научный руководитель: к. ф.-м. н. А. В. Грабовой

18 мая 2024 г.

Слайд об исследованиях

Исследуется проблема снижения размерности пространства параметров аппроксимирующих моделей.

Цель исследования:

Адаптация методов построения выравнивающих преобразований структуры модели учителя в модель ученика для моделей трансформеров.

Решение:

Построение последовательности выравнивающих преобразований позволяющих выровнять структуры модели учителя в модель ученика на базе архитектуры трансформера.

Дистилляция Дж. Хинтона[1]

Заданы

1) признаки $\mathbf{x}_i \in \mathbb{R}^n$,
2) $y_i \in \mathbb{Y} = \{1, \cdots, K\}, \quad \mathbb{Y}' = \mathbb{R}^K$.

Параметрические семейства учителя и ученика:

$$\mathfrak{F}_{\text{cl}} = \left\{\mathbf{f} | \mathbf{f} = \text{softmax}\bigl(\mathbf{v}(\mathbf{x})/T\bigr), \quad \mathbf{v} : \mathbb{R}^n \to \mathbb{R}^K\right\},$$

$$\mathfrak{G}_{\text{cl}} = \left\{\mathbf{g} | \mathbf{g} = \text{softmax}\bigl(\mathbf{z}(\mathbf{x})/T\bigr), \quad \mathbf{z} : \mathbb{R}^n \to \mathbb{R}^K\right\},$$

где $\mathbf{z}, \mathbf{v}$ — дифференцируемые по параметрам функции заданной структуры, T — параметр температуры. Оптимальная модель учителя $\hat{\mathbf{f}} \in \mathfrak{F}_{\text{cl}}$.

Функция ошибки

$$\mathcal{L}(\mathbf{g}) = -\underbrace{\sum_{i=1}^{m}\sum_{k=1}^{K} y_i^k \log g_k(\mathbf{x}_i)\big|_{T=1}}_{\text{исходная функция потерь}} - \underbrace{\sum_{i=1}^{m}\sum_{k=1}^{K} \hat{f}_k(\mathbf{x}_i)\big|_{T=T_0} \log g_k(\mathbf{x}_i)\big|_{T=T_0}}_{\text{слагаемое дистилляции}},$$

где $\cdot\big|_{T=t}$ фиксирует температуру T.

Оптимальная модель выбирается из класса, $\hat{\mathbf{g}} = \arg\min_{\mathbf{g} \in \mathfrak{G}_{\text{cl}}} \mathcal{L}(\mathbf{g})$.

[1] *Hinton G., et al* Distilling the knowledge in a neural network // NIPS, 2015.

Байесовская постановка задачи дистилляции

Задана обучающая выборка $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^m \quad x_i \in \mathbb{R}^n, \quad y_i \in \mathbb{Y}$.
Модель учителя

$$\mathbf{f}(\mathbf{x}) = \sigma \circ \mathbf{U}_T \sigma \circ \mathbf{U}_{T-1} \sigma \circ \cdots \circ \mathbf{U}_1 \mathbf{x},$$

где $\mathbf{U}$ матрицы линейных отображений, σ монотонная вектор-функция.
Параметры учителя фиксированы

$$\mathbf{u} = \mathrm{vec}\big([\mathbf{U}_T, \mathbf{U}_{T-1}, \cdots \mathbf{U}_1]\big).$$

На основе выборки $\{\mathbf{x}_i, y_i\}_{i=1}^m$ и значений учителя $\mathbf{f}(\hat{\mathbf{u}}, \mathbf{x})$ требуется выбрать модель ученика:

$$\mathbf{g}(\mathbf{x}) = \sigma \circ \mathbf{W}_L \sigma \circ \cdots \circ \mathbf{W}_1 \mathbf{x}, \quad \mathbf{W}_l \in \mathbb{R}^{n_s \times n_{s-1}}, L \leq T,$$

где $\mathbf{W}$, σ вводятся как и отображения учителя. Задача выбора модели $\mathbf{g}$ состоит в оптимизации вектора $\mathbf{w}$. Решается вариационным выводом

$$\hat{\mathbf{w}}, \hat{\mu}, \hat{\Sigma} = \arg\min_{\mu, \Sigma, \mathbf{w}} \mathrm{D}_{\mathrm{KL}}\big(q(\mathbf{w}|\mu, \Sigma) || p(\mathbf{w}|\mathbf{A})\big) - \mathrm{E}_{\mathbf{w} \sim q} \sum_{i=1}^{m} \log p(y_i|\mathbf{x}_i, \mathbf{w}).$$

Априорное распределение $p(\mathbf{w}|\mathbf{A})$ задается как функция от апостериорного распределения параметров учителя $p(\mathbf{u}|\mathbf{X}, \mathbf{y})$. Оно задано

$$p(\mathbf{u}|\mathbf{X}, \mathbf{y}) = \mathcal{N}(\mathbf{m}, \Sigma).$$

Так как размерности $\mathbf{u}$ и $\mathbf{w}$ не совпадают, то применяется выравнивающее преобразование – приведение параметров моделей в одно общее пространство:

$$p(\mathbf{w}|\mathbf{A}) = p(\upsilon|\mathbf{X}, \mathbf{y}), \quad \upsilon = \psi(t, \mathbf{u}), \quad \psi(t) : \mathbb{R}^{\mathbf{p_{tr}}} \to \mathbb{R}^{\mathbf{p_{tr}} - n_t n_{t-1}}.$$

---

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

Отличие архитектуры трансформера от полносвязной сети

Задана выборка $\mathcal{D} = \{x_i, y_i\}_{i=1}^{n}$, где $x_i = [x_i^1, x_i^2, \dots, x_i^{k_i}]$, $x_i^k \in \overline{0, 1, \dots, v_o}$,
$y_j = [y_i^1, y_i^2, \dots, y_i^{l_j}]$, $y_i^k \in \overline{0, 1, \dots, v_d}$.
Модель учителя

$$\mathbf{f}\left(\mathbf{x}, \mathbf{y}\right) = \mathbf{G} \circ \mathbf{D}_T \circ \dots \circ \mathbf{D}_2 \circ \mathbf{D}_1 \circ \mathbf{E}[\mathbf{x}, \mathbf{y}].$$

Модель ученика

$$\mathbf{g}\left(\mathbf{x}, \mathbf{y}\right) = \mathbf{G} \circ \mathbf{D}_T \circ \dots \circ \mathbf{D}_{k+1} \circ \mathbf{D}_{k-1} \circ \dots \circ \mathbf{D}_2 \circ \mathbf{D}_1 \circ \mathbf{E}[\mathbf{x}, \mathbf{y}].$$

Энкодер

$$\mathbf{E} : \underbrace{\mathbb{R}^{L \times d_{model}}}_{\text{токены x}} \times \underbrace{\mathbb{R}^{M \times v_{eng}}}_{\text{токены y}} \to \underbrace{\mathbb{R}^{L \times d_{model}}}_{\text{память}} \times \underbrace{\mathbb{R}^{M \times d_{model}}}_{\text{токен эмбеддинги y}}.$$

Слой декодера

$$\mathbf{D}_k : \underbrace{\mathbb{R}^{L \times d_{model}}}_{\text{память}} \times \underbrace{\mathbb{R}^{M \times d_{model}}}_{\text{эмбеддинги декодера}} \to \underbrace{\mathbb{R}^{L \times d_{model}}}_{\text{память}} \times \underbrace{\mathbb{R}^{M \times d_{model}}}_{\text{новые эмбеддинги декодера}}.$$

Генератор

$$\mathbf{G} : \underbrace{\mathbb{R}^{L \times d_{model}}}_{\text{память}} \times \underbrace{\mathbb{R}^{M \times d_{model}}}_{\text{эмбеддинги декодера}} \to \underbrace{\mathbb{R}^{M \times d_{model} \times v_d}}_{\text{логиты}}.$$

Байесовская дистилляция трансформера

Задана модель учителя, суперпозиция

$$\log p(\mathbf{y}_{1:t+1}) = \mathbf{f}\big(\mathbf{x}, \mathbf{y}_{0:t}\big) = \mathbf{G}(\mathbf{U}_{T+1}) \circ \mathbf{D}_T(\mathbf{U}_T) \circ \mathbf{D}_{T-1}(\mathbf{U}_{T-1}) \circ \cdots \circ \mathbf{D}_1(\mathbf{U}_1) \circ \mathbf{E}(\mathbf{U}_0)[\mathbf{x}, \mathbf{y}_{0:t}],$$

где $\mathbf{D}$ слои декодера, $\mathbf{E}$ энкодер, $\mathbf{G}$ генератор, возвращающий логиты следующих токенов. Параметры учителя фиксированы

$$\mathbf{u} = \mathrm{vec}\big([\mathbf{U}_{T+1}, \mathbf{U}_T, \mathbf{U}_{T-1}, \cdots \mathbf{U}_1, \mathbf{U}_0]\big),$$

где $\mathbf{U}_k$ векторизированные параметры соответствующих модулей.
На основе выборки $\{\mathbf{x}_i, \mathbf{y_i}\}_{i=1}^m$ и значений учителя $\mathbf{f}(\hat{\mathbf{u}}, \mathbf{x}, \mathbf{y})$ требуется выбрать модель ученика:

$$\mathbf{g}\big(\mathbf{x}, \mathbf{y}_{0:t}\big) = \mathbf{G}(\mathbf{W}_{L+1}) \circ \mathbf{D}_L(\mathbf{W}_L) \circ \cdots \circ \mathbf{D}_1(\mathbf{W}_1) \circ \mathbf{E}(\mathbf{W}_0)[\mathbf{x}, \mathbf{y}_{0:t}], \quad L \leq T,$$

где $\mathbf{G}, \mathbf{D}, \mathbf{E}$ вводятся как и отображения учителя. Задача выбора модели $\mathbf{g}$ состоит в оптимизации вектора $\mathbf{w} = \mathrm{vec}\big([\mathbf{W}_{L+1}, \mathbf{W}_L, \mathbf{W}_{L-1}, \cdots, \mathbf{W}_1, \mathbf{W}_0]\big)$. Решается вариационным выводом

$$\hat{\mathbf{w}}, \hat{\mu}, \hat{\Sigma} = \arg\min_{\mu, \Sigma, \mathbf{w}} \mathrm{D_{KL}}\big(q(\mathbf{w}|\mu, \Sigma) || p(\mathbf{w}|\mathbf{A})\big) - \mathrm{E}_{\mathbf{w}\sim q} \sum_{i=1}^{m} \log p\big(\mathbf{y}_i|\mathbf{x}_i, \mathbf{w}\big).$$

Априорное распределение $p\big(\mathbf{w}|\mathbf{A}\big)$ задается как функция от апостериорного распределения параметров учителя $p\big(\mathbf{u}|\mathbf{X}, \mathbf{Y}\big)$. Оно задано:

$$p\big(\mathbf{u}|\mathbf{X}, \mathbf{Y}\big) = \mathcal{N}\big(\mathbf{m}, \Sigma\big).$$

Так как размерности $\mathbf{u}$ и $\mathbf{w}$ не совпадают, то применяется выравнивающее преобразование – приведение параметров моделей в одно общее пространство:

$$p\big(\mathbf{w}|\mathbf{A}\big) = p\big(\upsilon|\mathbf{X}, \mathbf{y}\big), \quad \upsilon = \psi\big(\mathbf{u}\big), \quad \psi : \mathbb{R}^{\mathbf{p_{tr}}} \to \mathbb{R}^{\mathbf{p_{st}}}.$$

Выравнивание модели трансформера

Построим выравнивающее преобразование ψ из пространства параметров модели учителя в пространство параметров модели ученика:

$$\psi : \mathbb{R}^{\mathbf{p_{tr}}} \to \mathbb{R}^{\mathbf{p_{tr}} - 4\cdot d_{model} \times d_k - 2d_v \times d_{model} - 2\cdot N_{heads} \cdot d_v \times d_{model} - 2d_{model} \times d_{ff} - d_{ff} - d_{model}}.$$

Модель учителя:

$$\mathbf{f}(\mathbf{x}, \mathbf{y}) = \mathbf{G}(\mathbf{U}_{T+1}) \circ \mathbf{D}_T(\mathbf{U}_T) \circ \cdots \circ \mathbf{D}_2(\mathbf{U}_2) \circ \mathbf{D}_1(\mathbf{U}_1) \circ \mathbf{E}(\mathbf{U}_0)[\mathbf{x}, \mathbf{y}]$$

$$\mathbf{D}_k(\mathbf{x}, \mathbf{y}) \equiv [\mathbf{x}, \mathbf{D}_k^0(\mathbf{x}, \mathbf{y})]$$

$$\mathbf{D}_k^0(\mathbf{x}, \mathbf{y}) = \underbrace{(\mathbf{I} + \mathbf{FFN})}_{\text{feed-forward}} \circ \underbrace{(\mathbf{I} + \mathbf{MHA})[\mathbf{x}]}_{\text{multi-head attention}} \circ \underbrace{(\mathbf{I} + \mathbf{SA})}_{\text{self attention}} (\mathbf{y})$$

$$\mathbf{FFN}(\mathbf{y}) = (\mathbf{b}_2 + \sigma \circ \mathbf{U}_2^F \circ (\mathbf{b}_1 + \mathbf{y} \cdot \mathbf{U}_1^F))$$

$$\mathbf{MHA}(\mathbf{x}, \mathbf{y}) = \text{heads}[\mathbf{U}_M^h](\mathbf{x}, \mathbf{y}) \cdot \mathbf{U}^M \qquad \mathbf{SA}(\mathbf{y}) = \text{heads}[\mathbf{U}_S^h](\mathbf{y}) \cdot \mathbf{U}^S,$$

где σ функция активации, параметры модуля U_k:
$\mathbf{U}_1^F \in \mathbb{R}^{d_{model} \times d_{ff}}, \mathbf{U}_2^F \in \mathbb{R}^{d_{ff} \times d_{model}}, \mathbf{b}_1 \in \mathbb{R}^{d_{ff}}, \mathbf{b}_2 \in \mathbb{R}^{d_{model}}$,
$\mathbf{U}^M, \mathbf{U}^S \in \mathbb{R}^{N_{heads} \cdot d_v \times d_{model}}, \mathbf{U}_M^h, \mathbf{U}_S^h \in \mathbb{R}^{2\cdot d_k \times d_{model} + \cdot d_v \times d_{model}}$.
Модель ученика:

$$\mathbf{g}(\mathbf{x}, \mathbf{y}) = \mathbf{G}(\mathbf{W}_{T+1}) \circ \cdots \circ \mathbf{D}_{k+1}(\mathbf{W}_{k+1}) \circ \mathbf{D}_{k-1}(\mathbf{W}_{k-1}) \circ \cdots \circ \mathbf{D}_1(\mathbf{W}_1) \circ \mathbf{E}(\mathbf{W}_0)[\mathbf{x}, \mathbf{y}].$$

Эквивалентность моделей:

$$\mathbf{f} \mid_{\mathbf{U}_2^F, \mathbf{b}_2, \mathbf{U}^M, \mathbf{U}^S = 0} \equiv \mathbf{g}.$$

Параметры $\mathbf{u}$ модели $\mathbf{f}$ делятся на удаляемые $\xi_2 = vec([b_1, U_1^F, U_M^h, U_S^h])$, зануляемые $\xi_1 = \text{vec}([b_2, U_2^F, U^S, U^M])$, оставшиеся $\upsilon = \text{vec}\big([\mathbf{U}_{T+1}, \cdots, \mathbf{U}_{k+1}, \mathbf{U}_{k-1}, \cdots, \mathbf{U}_0]\big)$.

Решение задачи выравнивания структур моделей

Параметры $\mathbf{u}$ модели $\mathbf{f}$:

1. удаляемые $\xi_2 = vec([b_1, U_1^F, U_M^h, U_S^h])$,
2. зануляемые $\xi_1 = \text{vec}([b_2, U_2^F, U^S, U^M])$,
3. оставшиеся $\upsilon = \text{vec}\big([\mathbf{U}_{T+1}, \cdots, \mathbf{U}_{k+1}, \mathbf{U}_{k-1}, \cdots, \mathbf{U}_0]\big)$.

Апостериорное распределение параметров $\upsilon = \psi_{transformer}(\mathbf{u}, k)$ модели $\mathbf{f}$:

$$p(\upsilon|\mathfrak{D}) = \int\limits_{\xi_2} p(\bar{\xi_1}|\mathfrak{D}, \xi_1 = \mathbf{0}) d\xi_2,$$

где $\bar{\xi_1} = [\upsilon, \xi_2]$.

Из свойства распределения $p(\bar{\xi}_1|\mathfrak{D}, \xi_1 = \mathbf{0}) = \mathcal{N}(\mu, \Xi)$, с параметрами μ, Ξ:

$$\mu = \mathbf{m}_{\bar{\xi}_1} + \mathbf{\Sigma}_{\bar{\xi}_1, \xi_1} \mathbf{\Sigma}_{\xi_1, \xi_1}^{-1} \left(\mathbf{0} - \mathbf{m}_{\xi_1}\right),$$

$$\Xi = \mathbf{\Sigma}_{\bar{\xi}_1, \bar{\xi}_1} - \mathbf{\Sigma}_{\bar{\xi}_1, \xi_1} \mathbf{\Sigma}_{\xi_1, \xi_1}^{-1} \mathbf{\Sigma}_{\xi_1, \bar{\xi}_1}.$$

Маргинализация нормального распределения $p(\upsilon|\mathfrak{D}) = \mathcal{N}(\mu_\upsilon, \Xi_{\upsilon, \upsilon})$.

Theorem (Игнашин, 2024)

Пусть апостериорное распределение параметров модели учителя $\mathbf{u}$ *имеют распределение* $\mathcal{N}(\mathbf{m}, \Sigma)$. *Модель ученика имеет схожую структуру с моделью учителя, но без одного слоя декодера* $\mathbf{D}_k(\mathbf{U}_k)$. *Тогда апостериорное распределение имеет вид:*

$$p(\psi_{transformer}(\mathbf{u}, k)|\mathfrak{D}) = \mathcal{N}(\mathbf{m}_\upsilon + \mathbf{\Sigma}_{\upsilon, \xi_1} \mathbf{\Sigma}_{\xi_1, \xi_1}^{-1} \left(\mathbf{0} - \mathbf{m}_{\xi_1}\right), \mathbf{\Sigma}_{\upsilon, \upsilon} - \mathbf{\Sigma}_{\upsilon, \xi_1} \mathbf{\Sigma}_{\xi_1, \xi_1}^{-1} \mathbf{\Sigma}_{\xi_1, \upsilon}),$$

Случай некорелллированных параметров

Рассмотрим частный случай, когда ковариационные матрицы нулевые $\Sigma_{\bar{\xi}_1,\xi_1} = 0$, то есть зануляемыые параметры ξ_1 и остальные параметры $\bar{\xi}_1$ – некорелллированы. Тогда распределение $p(\bar{\xi}_1|\mathfrak{D}, \xi_1 = \mathbf{0}) = \mathcal{N}(\mu, \Xi)$ имеет параметры μ, Ξ:

$$\mu = \mathbf{m}_{\bar{\xi}_1},$$
$$\Xi = \Sigma_{\bar{\xi}_1,\bar{\xi}_1}.$$

Следовательно апостериорное распределение параметров модели учителя υ:

$$p(\upsilon|\mathfrak{D}) = \mathcal{N}(m_\upsilon, \Sigma_{\upsilon,\upsilon}).$$

Априорное распределение модели ученика :

$$p(w|\mathbf{A}) = p(\upsilon|\mathfrak{D}).$$

Описание эксперимента

Обучающий датасет: 40 тысяч руско-английских переводов предложений.
В качестве аппроксимации вариационного вывода применяется обучение кроссэнтропийного лосса.
Инициализация модели ученика:

$$\mathbf{w}_0 = \mathbf{m}_\upsilon.$$

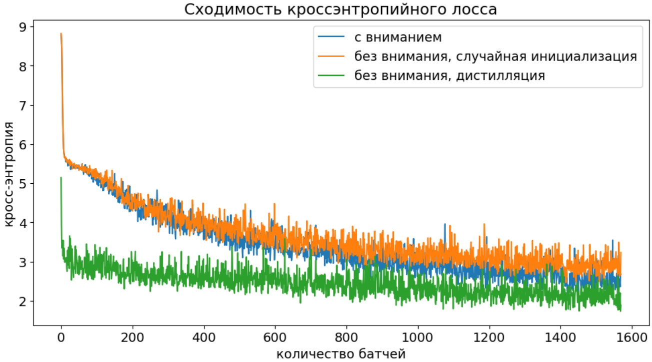
Аппроксимация $\mathbf{m}$ апостериорного распределения параметров модели учителя $p(\mathbf{u}|\mathbf{X}, \mathbf{Y}) = \mathcal{N}(\mathbf{m}, \mathbf{\Sigma})$:

$$\mathbf{m} = \mathbf{u}_K,$$

где $\mathbf{u}_K$ – параметры модели учителя на K-ой итерации обучения.
Выбраные рекуррентные модели:

1. Модель учителя,
2. Модель ученика $\mathbf{w}_0 \sim \mathit{Uniform}[-0.08, 0.08]$,
3. Модель ученика $\mathbf{w}_0 = \mathbf{m}_\upsilon$.

Результат вычислительного эксперимента



	модель учителя	**модель ученика**	**модель ученика (дистиллированная)**
BLEU	23.70	13.35	17.73

Дистиллированная модель сходится лучше, а также показывает лучшую метрику BLEU на отложенной выборке, чем модель той же структуры, но с произвольной инициализацией.

Выносится на защиту

Получены следующие результаты :

1. Предложен метод удаления слоя декодера трансформера.
2. Предложены методы выравнивания структур моделей глубокого обучения с механизмом внимания.
3. Проведен теоретический анализ предложенных методов
4. Доказана теорема о выравниванивании модели трансформера.
5. Проведен вычислительный эксперимент, показывающий состоятельность метода.

Публикации и выступления на конференциях:

1. О задаче поиска равновесного распределения потоков. «66-я Всероссийская научная конференция МФТИ», 2024.
2. Игнашин И.Н., Ярмошик Д.В. Модификации алгоритма Frank –Wolfe в задаче поиска равновесного распределения транспортных потоков. // Компьютерные исследования и моделирование, 2024 Т. 16 № 1 С. 53–68.