Байесовская дистилляция моделей на базе трансформеров

И. Н. Игнашин

Московский физико-технический институт

16 декабря 2023 г.

Слайд об исследованиях

Исследуется проблема дистилляции 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

Для удаления Attention в модели RNN .

Постановка задачи . . .

Заданы

1) признаки . . . ,
2) целевая переменная . . . ,
3) . . . .

. . .

Требуется выбрать модель . . . из множества

$$\mathfrak{G} = \{\mathbf{g} | \mathbf{g} : \mathbb{R}^n \to \mathbb{Y}'\}.$$

Оптимизационная задача . . . :

$$\mathbf{g} = \arg\min_{\mathbf{g} \in \mathfrak{G}} \mathcal{L}\big(\ldots\big),$$

где $\mathcal{L}$ — функция ошибки.

---

[1] *Lopez-Paz D., Bottou L., Scholkopf B., Vapnik V.* Unifying distillation and privileged information // ICLR, 2016.

[2] *Hinton G., Vinyals O., Dean J.* Distilling the knowledge in a neural network // NIPS, 2015.

Предложенный метод ...

Заданы

1) ...,
2) ....

Параметрические семейства:

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

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

где ....

Функция ошибки

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

где ....

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

Анализ предложенного метода . . .

На графике показана зависимость значения параметровw_i в зависимости от параметра l_1-регуляризации C.

С увеличением параметра регуляризации C число ненулевых параметров w_i уменьшается.

Выводы

1. Предложен ....
2. Доказаны теоремы ...,
   - ...,
   - ....
3. Предложен метод ...
   - ...,
   - ....
4. Предложены методы ...
   - ...,
   - ....
5. Предложена вероятностная интерпретации ....