Stochastic Beams and Where to Find Them

Денис Золотухин

НИУ ВШЭ

26 сентября 2019

Какая решается задача?

Многие задачи машинного обучения могут быть представлены как факторизация распределений на последовательностях:

$$p_\theta(y_{t:1}) = p_\theta(y_t|y_{t-1:1}) \cdot p_\theta(y_{t-1:1}) = \prod_t p_\theta(y_t|y_{t-1:1})$$

где θ, обычно, - обучаемый параметр модели.

- Как получать примеры/ответы из такой модели?

Этот доклад: способ сэмплировать без возвращений

Beam Search

$$
\overset{\text{t-1}}{\begin{bmatrix} y_{t-1,1} \\ y_{t-1,2} \\ \vdots \\ y_{t-1,k} \end{bmatrix}}
\rightarrow
\begin{bmatrix}
\begin{bmatrix} y_{t,1,1} \\ \vdots \\ {\color{red} y_{t,1,k}} \end{bmatrix} \\
\begin{bmatrix} {\color{red} y_{t,2,1}} \\ \vdots \\ y_{t,2,k} \end{bmatrix} \\
\vdots \\
\begin{bmatrix} y_{t,k,1} \\ \vdots \\ {\color{red} y_{t,k,k}} \end{bmatrix}
\end{bmatrix}
\rightarrow
\overset{\text{t}}{\begin{bmatrix} {\color{red} y_{t,1}} \\ {\color{red} y_{t,2}} \\ \vdots \\ {\color{red} y_{t,k}} \end{bmatrix}}
$$

Категориальное распределение

Случайная величина I на конечном множестве $(N = \{1, \dots, n\})$ элементов имеет категориальное распределние. Его можно представить как

$$p(i) = \frac{\exp(\phi_i)}{\sum_{j \in N} \exp(\phi_j)}$$

$$I \sim \mathrm{Cat}\left(\frac{\exp(\phi_i)}{\sum_{j \in N} \exp(\phi_j)}\right)$$

ϕ_i называются лог-вероятностями.

Распределение Гумбеля

Если $U(x)$ - функция распределения U$[0,1]$, то

$$G(x) = \phi - \log(-\log(U(x)))$$

- функция распределения Гумбеля (Gumbel) со сдвигом ϕ.

- Чем она так хороша?

Распределение Гумбеля

Возьмём

$$G_i \sim \mathrm{Gumbel}(0)$$

и лог-вероятности ϕ_i. Положим $G_{\phi_i} = G_i + \phi_i$ и

$$\xi = \arg\max_i \{G_{\phi_i}\},$$

$$\eta = \max_i \{G_{\phi_i}\}$$

Тогда,

$$\xi \sim \mathrm{Cat}\left(\frac{\exp(\phi_i)}{\sum_{j \in N} \exp(\phi_j)}\right)$$

$$\eta \sim \mathrm{Gumbel}\left(\log \sum \exp(\phi_i)\right)$$

Распределение Гумбеля

Более того, если

$$\xi_1, \ldots, \xi_k = \arg \operatorname{top}\ k\{G_{\phi_i}\}$$

То $\xi_1, \ldots, \xi_k$ - выбор из $\mathrm{Cat}\left(\frac{\exp(\phi_i)}{\sum_{j \in N} \exp(\phi_j)}\right)$ **без возвращения**.

Последовательность как дерево

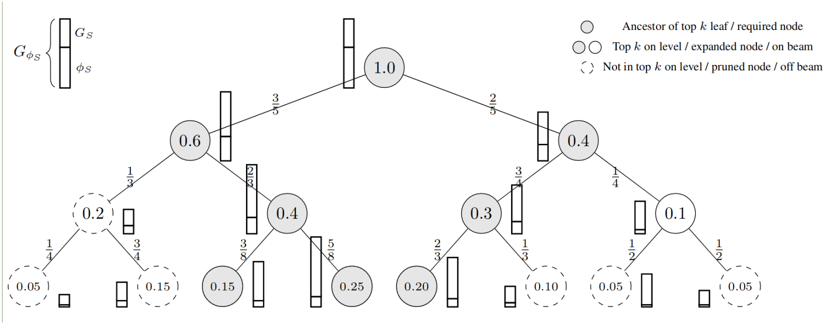


Рис.: Представление последовательности в виде дерева

Последовательность как дерево

- Каждый узел - неполная последовательность $y_{t:1}$ и одновременно все последовательности с таким началом.
- Листы y^i с лог-вероятностями $\phi_i = \log p_\theta(y^i)$.

Сэмпл из листов размера k без возвращения:

- Считаем все ϕ_i
- Сэмплируем G_{ϕ_i}
- $\arg\operatorname{top} k$
- Profit?

Последовательность как дерево

Пусть $S \subset N$ - все листы поддерева такого-то узла. Тогда

$$\phi_S = \log p_\theta(y^S) = \log \sum_{i \in S} \exp(\phi_i)$$

Пусть теперь $G_{\phi_S} = \max_{i \in S}(G_{\phi_i})$. Вспоминаем, что

$$G_{\phi_S} \sim \mathrm{Gumbel}\left(\log \sum_{i \in S} \exp(\phi_i)\right) \sim \mathrm{Gumbel}(\phi_S)$$

Сэмплирование снизу вверх

Мы можем посчитать G_{ϕ_S} снизу вверх рекурсивно:

$$G_{\phi_S} = \max_{S' \in \text{Children}(S)} G_{\phi_{S'}}$$

- Всё ещё нужно считать G_{ϕ_S} для всех узлов(

Сэмплирование сверху вниз

Заметим, что $\phi_N = 0$. Тогда $G_{\phi_N} \sim \text{Gumbel}(0)$. Для узла S мы бы хотели сэмплировать его детей G_{ϕ_S}, сохраняя свойство, что родитель - максимиу детей.

- Как?

А всё очень просто: возьмём независимо $G_{\phi_{S'}} \sim \text{Gumbel}(\phi_{S'})$ и $Z = \max_{S'} G_{\phi_{S'}}$. Тогда

$$\tilde{G}_{\phi_{S'}} = -\log(\exp(-G_{\phi_S}) - \exp(-Z) + \exp(-G_{\phi_{S'}}))$$

- это как раз то, что нам нужно: множество из распределения Гумбеля с заданным максимумом.

Стохастический Beam Search

- Чтобы насемплировать k последовательностей без возвращения нам нужно найти топ-k из ϕ_i.

Для этого достаточно при сэмплировании сверху вниз смотреть на поддеревья только k узлов с максимальными ϕ_S.

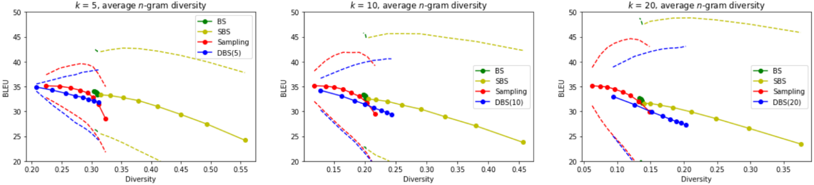
- Это то же, что и Beam Search, только с другими рейтингами узлов, которые определяюся стохастической процедурой.

Наивный подход?

А почему бы не запустить Beam Search и просто сэмплировать из приходящих последовательностей?

- Ничем не обосновано, ничего не значит, а в нашем алгоритме мы точно понимаем, что делаем.
- Маловероятные частичные последовательности скорее всего быстро исчезнут из пучка, чего не происходит в нашем алгоритме.

Эксперименты



(a) $k = 5$ (b) $k = 10$ (c) $k = 20$

Figure 2. Minimum, mean and maximum BLEU score vs. diversity for different sample sizes k. Points indicate different temperatures/diversity strengths, from 0.1 (low diversity, left in graph) to 0.8 (high diversity, right in graph).

Рис.: Тесты на разнообразие

Эксперименты

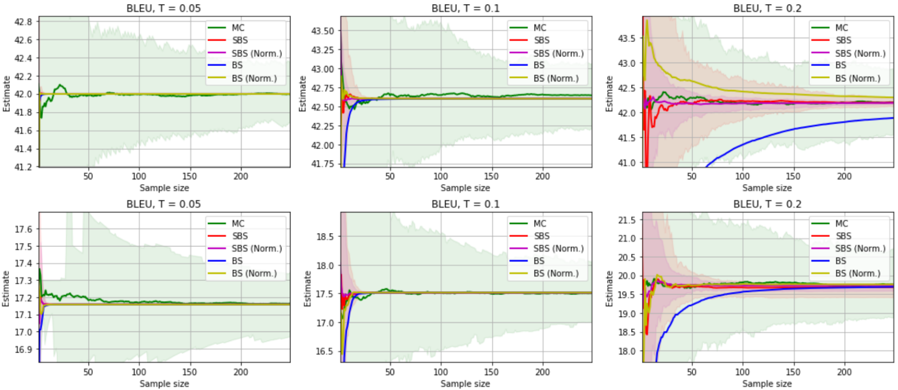


Рис.: Предсказание BLEU

Выводы

- Stochastic Beam Search - алгоритм сэмплирования последовательностей
- Основан на распределении Гумбеля
- Результат - сэмпл без возвращений
- Хорошо показывает себя во многих задачах

Ссылки

[1]. https://arxiv.org/abs/1903.06059