Speculative Decoding

正田 備也

masada@rikkyo.ac.jp

Decodingの効率化

- autoregressive LLMは、次の1トークンの生成を繰り返すことで、長いテキストを生成する。
- しかし、そのための計算コストは大きい。
- そこで、小規模なLMを補助的に使って、トークンをより高速に生成する。具体的には・・・
- speculative decodingを使うと、元のLLMからトークン列を生成するときに使われるのと全く同じ確率分布に基づきつつも、より効率的にトークン列を生成できるようになる。

Speculative decoding [arXiv:2211.17192][arXiv:2302.01318]

- 二つのカテゴリカル分布を使ってトークンを生成する。
  - Cat($\boldsymbol{\theta}$): 小規模 LM のある position でのカテゴリカル分布
  - Cat($\boldsymbol{\phi}$): LLM の同じ positionde でのカテゴリカル分布

1. Cat($\boldsymbol{\theta}$) からサンプリングし、得られたアイテムを x とする。
2. $\theta_x \leq \phi_x$ ならば x をそのまま採用する。
3. $\theta_x > \phi_x$ ならば、表の出る確率が $\frac{\phi_x}{\theta_x}$ のコインを投げる。
4. 表が出たら x を採用する。
5. 裏が出たら Cat$\big((\boldsymbol{\phi} - \boldsymbol{\theta})_+\big)$ からサンプリングし直す。

サンプリングし直すときに使うカテゴリカル分布

- $\mathrm{Cat}\big((\boldsymbol{\phi}-\boldsymbol{\theta})_+\big)$ は、アイテム x の出現確率が、以下のように定義されるカテゴリカル分布である。

$$(\phi_x-\theta_x)_+ \equiv \frac{\max(0,\phi_x-\theta_x)}{\sum_{x'}\max(0,\phi_{x'}-\theta_{x'})} \tag{1}$$

- つまり、$\phi_x > \theta_x$ を満たすアイテム x だけが非ゼロの確率を持つように規格化されたカテゴリカル分布である。

Speculative decodingにおけるアイテムの出現確率

- 上のrejection samplingで結果的に選ばれるアイテムを確率変数Xで表すと、$X = x$となる確率は以下のように書ける。

$$
\begin{aligned}
&P(X = x) \\
&= P_{\boldsymbol{\theta}}(X = x) P(x \text{が採用される} \,|X = x) \\
&\quad + \Big(P_{\phi}(X = x | \text{Cat}(\boldsymbol{\theta}) \text{からのサンプルが採用されない}) \\
&\qquad \times P(\text{Cat}(\boldsymbol{\theta}) \text{からのサンプルが採用されない}) \Big) \qquad (2)
\end{aligned}
$$

- この式を、以下、簡単な形に書き換える。

まず、

$$P(x \text{ が採用される } | X = x) = \begin{cases} 1 & \text{when } \theta_x \leq \phi_x \\ \frac{\phi_x}{\theta_x} & \text{when } \theta_x > \phi_x \end{cases} \tag{3}$$

この確率は、以下のように書き換えることができる。

$$P(x \text{ が採用される } | X = x) = \min\left(1, \frac{\phi_x}{\theta_x}\right) \tag{4}$$

$P_{\boldsymbol{\theta}}(X = x) \equiv \theta_x$ であるから、式 (2) の一つ目の項は、以下のように書き換えられる。

$$P_{\boldsymbol{\theta}}(X = x) P(x \text{ が採用される } | X = x) = \theta_x \times \min\left(1, \frac{\phi_x}{\theta_x}\right) = \min(\theta_x, \phi_x) \tag{5}$$

次に、式 (5) より

$$P_{\boldsymbol{\theta}}(X'=x')P(x' \text{ が採用されない } |X'=x') = \theta_{x'} \times \left(1 - \min\left(1, \frac{\phi_{x'}}{\theta_{x'}}\right)\right) \tag{6}$$

ところで、$\theta_x \leq \phi_x$ のとき $1 - \min\left(1, \frac{\phi_{x'}}{\theta_{x'}}\right) = 1 - 1 = 0$ であり、また、$\theta_x > \phi_x$ のとき $1 - \min\left(1, \frac{\phi_{x'}}{\theta_{x'}}\right) = 1 - \frac{\phi_{x'}}{\theta_{x'}}$ であるので、両方をまとめると、

$$1 - \min\left(1, \frac{\phi_{x'}}{\theta_{x'}}\right) = \max\left(0, 1 - \frac{\phi_{x'}}{\theta_{x'}}\right) \tag{7}$$

と書ける。したがって、

$$\begin{aligned} P_{\boldsymbol{\theta}}(X'=x')P(x' \text{ が採用されない } |X'=x') &= \theta_{x'} \times \left(1 - \min\left(1, \frac{\phi_{x'}}{\theta_{x'}}\right)\right) \\ &= \theta_{x'} \times \max\left(0, 1 - \frac{\phi_{x'}}{\theta_{x'}}\right) \\ &= \max(0, \theta_{x'} - \phi_{x'}) \end{aligned} \tag{8}$$

式 (8) を使うと、Cat($\boldsymbol{\theta}$) からのサンプルが採用されない確率は、

$$
\begin{aligned}
P(\text{Cat}(\boldsymbol{\theta}) \text{ からのサンプルが採用されない}) &= \sum_{x'} P(X' = x', x' \text{ が採用されない}) \\
&= \sum_{x'} P_{\boldsymbol{\theta}}(X' = x') P(x' \text{ が採用されない } | X' = x') \\
&= \sum_{x'} \max(0, \theta_{x'} - \phi_{x'}) \qquad (9)
\end{aligned}
$$

と書けることが分かる。
ここで、Cat($\boldsymbol{\theta}$) からのサンプルが採用されないときには、アイテム x の出現確率が

$$
(\phi_x - \theta_x)_+ \equiv \frac{\max(0, \phi_x - \theta_x)}{\sum_{x'} \max(0, \phi_{x'} - \theta_{x'})} \qquad (10)
$$

と定義されるカテゴリカル分布からサンプリングするのだったことを、思い出す。

式 (10) の分母は、式 (9) に一致している。よって、式 (2) の二つ目の項は以下のようになる。

$$
\begin{aligned}
& P_{\boldsymbol{\phi}}(X = x | \mathsf{Cat}(\boldsymbol{\theta}) \text{ からのサンプルが採用されない}) \\
& \quad \times P(\mathsf{Cat}(\boldsymbol{\theta}) \text{ からのサンプルが採用されない}) \\
&= (\phi_x - \theta_x)_+ \times \sum_{x'} \max(0, \theta_{x'} - \phi_{x'}) \\
&= \frac{\max(0, \phi_x - \theta_x)}{\sum_{x'} \max(0, \phi_{x'} - \theta_{x'})} \times \sum_{x'} \max(0, \theta_{x'} - \phi_{x'}) \\
&= \max(0, \phi_x - \theta_x) \qquad (11)
\end{aligned}
$$

以上より、式 (2) は、以下のように書き換えられる。

$$P(X = x) = \min(\theta_x, \phi_x) + \max(0, \phi_x - \theta_x) \qquad (12)$$

θ_x と ϕ_x の大小関係で場合分けして計算すると、

$$P(X = x) = \phi_x \qquad (13)$$