

Classifier Guidance для текстовой диффузии

1 Постановка задачи

Дано:

- y – префикс (условие)
- x – продолжение, которое хотим сгенерировать
- Цель: сэмплировать из $p(x | y)$

Используем диффузионную модель в обратном процессе.

2 Разложение градиента

Начнём с формулы Байеса для условного распределения маргинала x_t :

$$p(x_t | y) = \frac{p(y | x_t) \cdot p(x_t)}{p(y)}$$

Формула Байеса, где:

- $p(x_t | y)$ – апостериорная вероятность (что нам нужно)
- $p(y | x_t)$ – вероятность условия при данных x_t (предсказывает классификатор)
- $p(x_t)$ – априорная вероятность (предсказывает диффузионная модель)
- $p(y)$ – вероятность условия (константа при фиксированном y)

Логарифмируем обе части уравнения:

$$\log p(x_t | y) = \log p(y | x_t) + \log p(x_t) - \log p(y)$$

Теперь берём градиент по x_t :

$$\nabla_{x_t} \log p(x_t | y) = \nabla_{x_t} \log p(y | x_t) + \nabla_{x_t} \log p(x_t) - \nabla_{x_t} \log p(y)$$

Замечаем, что $p(y)$ не зависит от x_t , поэтому $\nabla_{x_t} \log p(y) = 0$. Получаем:

$$\boxed{\nabla_{x_t} \log p(x_t | y) = \nabla_{x_t} \log p(x_t) + \nabla_{x_t} \log p(y | x_t)} \quad (1)$$

Это ключевое уравнение: градиент логарифма условного распределения равен сумме градиента логарифма безусловного распределения (диффузия) и градиента логарифма правдоподобия (классификатор).

3 Первое слагаемое: диффузионный score

Диффузионная модель обучается предсказывать шум $\epsilon_\theta(x_t, t)$. В параметризации DDPM прямой процесс:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

Score (градиент логарифма плотности) аппроксимируется как:

$$\nabla_{x_t} \log p(x_t) = -\frac{\epsilon_\theta(x_t, t)}{\sqrt{1 - \bar{\alpha}_t}} \quad (2)$$

Тогда функция потерь для диффузии

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{x_0, \epsilon, t} [\|\epsilon_\theta(x_t, t) - \epsilon\|^2]$$

4 Второе слагаемое: классификатор

Обучаем классификатор (например, BERT) на зашумлённых состояниях x_t . Модель выдаёт логит $f(x_t, t, y) \in \mathbb{R}$, тогда:

$$p(y | x_t) = \sigma(f(x_t, t, y)), \quad \sigma(z) = \frac{1}{1 + e^{-z}}$$

4.1 Подробный вывод градиента классификатора

Рассмотрим $\log p(y | x_t) = \log \sigma(f)$.

Для сигмоидной функции $\sigma(f) = \frac{1}{1+e^{-f}}$:

$$\begin{aligned} \frac{d}{df} \log \sigma(f) &= \frac{1}{\sigma(f)} \cdot \frac{d\sigma(f)}{df} \\ \frac{d\sigma(f)}{df} &= \frac{e^{-f}}{(1 + e^{-f})^2} = \sigma(f) \cdot (1 - \sigma(f)) \\ \frac{d}{df} \log \sigma(f) &= \frac{1}{\sigma(f)} \cdot \sigma(f)(1 - \sigma(f)) = 1 - \sigma(f) \end{aligned}$$

Используя цепное правило, получаем:

$$\nabla_{x_t} \log p(y | x_t) = (1 - \sigma(f(x_t, t, y))) \cdot \nabla_{x_t} f(x_t, t, y) \quad (3)$$

4.2 Функция потерь для классификатора (BCE)

$$\mathcal{L}_{\text{cls}} = -\mathbb{E}_t [\log \sigma(f(x_t^+, t, y)) + \log(1 - \sigma(f(x_t^-, t, y)))]$$

где x_t^+ – зашумлённое правильное продолжение, x_t^- – зашумлённый случайный текст.

4.3 Общий случай с булевой меткой Бернулли: подробный вывод

Для булевой метки $\varphi \in \{0, 1\}$ (например, $\varphi = 1$ означает, что x_t является правильным продолжением префикса y):

Распределение Бернулли: $p(\varphi | x_t) = \sigma(f)^\varphi \cdot (1 - \sigma(f))^{1-\varphi}$

Пояснение:

- Если $\varphi = 1$: $p(1 | x_t) = \sigma(f)^1 \cdot (1 - \sigma(f))^0 = \sigma(f)$
- Если $\varphi = 0$: $p(0 | x_t) = \sigma(f)^0 \cdot (1 - \sigma(f))^1 = 1 - \sigma(f)$

Логарифмируем:

$$\log p(\varphi | x_t) = \varphi \log \sigma(f) + (1 - \varphi) \log(1 - \sigma(f))$$

Теперь вычислим градиент по x_t . Используем цепное правило и полученные ранее производные:

$$\begin{aligned} \nabla_{x_t} \log p(\varphi | x_t) &= \nabla_{x_t} [\varphi \log \sigma(f) + (1 - \varphi) \log(1 - \sigma(f))] \\ &= \varphi \cdot \nabla_{x_t} \log \sigma(f) + (1 - \varphi) \cdot \nabla_{x_t} \log(1 - \sigma(f)) \end{aligned}$$

Из предыдущих вычислений:

- $\nabla_{x_t} \log \sigma(f) = (1 - \sigma(f)) \cdot \nabla_{x_t} f$
- $\nabla_{x_t} \log(1 - \sigma(f)) = -\sigma(f) \cdot \nabla_{x_t} f$

Подставляем:

$$\begin{aligned} \nabla_{x_t} \log p(\varphi | x_t) &= \varphi \cdot (1 - \sigma(f)) \cdot \nabla_{x_t} f + (1 - \varphi) \cdot (-\sigma(f)) \cdot \nabla_{x_t} f \\ &= [\varphi(1 - \sigma(f)) - (1 - \varphi)\sigma(f)] \cdot \nabla_{x_t} f \end{aligned}$$

Упрощаем выражение в квадратных скобках:

$$\begin{aligned} \varphi(1 - \sigma(f)) - (1 - \varphi)\sigma(f) &= \varphi - \varphi\sigma(f) - \sigma(f) + \varphi\sigma(f) \\ &= \varphi - \sigma(f) \end{aligned}$$

Таким образом:

$$\nabla_{x_t} \log p(\varphi | x_t) = (\varphi - \sigma(f(x_t, t, y))) \cdot \nabla_{x_t} f(x_t, t, y)$$

При $\varphi = 1$ (хотим, чтобы продолжение было правильным) получаем:

$$\nabla_{x_t} \log p(\varphi = 1 | x_t) = (1 - \sigma(f(x_t, t, y))) \cdot \nabla_{x_t} f(x_t, t, y)$$

что совпадает с формулой (3).

5 Полный score

Подставляем (2) и (3) в (1):

$$\nabla_{x_t} \log p(x_t | y) = -\frac{\epsilon_\theta(x_t, t)}{\sqrt{1 - \bar{\alpha}_t}} + (1 - \sigma(f(x_t, t, y))) \cdot \nabla_{x_t} f(x_t, t, y)$$

6 Шаг сэмплирования

В DDPM обратный шаг (без условия):

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right)$$

С classifier guidance модифицируем среднее, добавляя градиент классификатора:

$$\tilde{\mu}(x_t) = \mu_\theta(x_t, t) + \lambda \sigma_t^2 \nabla_{x_t} \log p(y | x_t)$$

где λ – коэффициент guidance, σ_t^2 – дисперсия обратного шага.

Финальный сэмплинг:

$$x_{t-1} = \tilde{\mu}(x_t) + \sigma_t \xi, \quad \xi \sim \mathcal{N}(0, I)$$

В развернутом виде:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right) + \lambda \sigma_t^2 (1 - \sigma(f)) \nabla_{x_t} f + \sigma_t \xi$$

7 Псевокод

7.1 Обучение диффузионной модели

Algorithm 1 Обучение диффузионной модели

```
1: Инициализировать параметры диффузионной модели  $\theta$ 
2: for эпоха = 1 ... N do
3:   for батч  $x_0$  из датасета do
4:     Выбрать случайный шаг  $t \sim \text{Uniform}\{1, \dots, T\}$ 
5:     Сгенерировать шум  $\epsilon \sim \mathcal{N}(0, I)$ 
6:     Вычислить зашумлённое состояние  $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$ 
7:     Получить предсказание шума  $\epsilon_\theta = \text{DiffusionModel}(x_t, t)$ 
8:     Вычислить лосс:  $\mathcal{L}_{\text{diff}} = \|\epsilon_\theta - \epsilon\|^2$ 
9:     Обновить  $\theta$ :  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{diff}}$ 
10:   end for
11: end for
```

7.2 Обучение классификатора

7.3 Инференс

Algorithm 2 Обучение классификатора

- 1: Инициализировать параметры классификатора ϕ
- 2: **for** эпоха = 1 . . . N **do**
- 3: **for** батч (x_0, y) из датасета **do**
- 4: Выбрать случайный шаг $t \sim \text{Uniform}\{1, \dots, T\}$
- 5: Сгенерировать шум $\epsilon \sim \mathcal{N}(0, I)$
- 6: Вычислить $x_t^+ = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$ (правильное продолжение)
- 7: Выбрать случайный текст x_0^- из датасета
- 8: Сгенерировать шум $\epsilon' \sim \mathcal{N}(0, I)$
- 9: Вычислить $x_t^- = \sqrt{\bar{\alpha}_t}x_0^- + \sqrt{1 - \bar{\alpha}_t}\epsilon'$ (неправильное продолжение)
- 10: Получить логиты: $f^+ = \text{Classifier}(x_t^+, t, y)$, $f^- = \text{Classifier}(x_t^-, t, y)$
- 11: Вычислить лосс: $\mathcal{L}_{\text{cls}} = -\log \sigma(f^+) - \log(1 - \sigma(f^-))$
- 12: Обновить ϕ : $\phi \leftarrow \phi - \eta \nabla_{\phi} \mathcal{L}_{\text{cls}}$
- 13: **end for**
- 14: **end for**

Algorithm 3 Генерация текста с classifier guidance

- 1: **Вход:** префикс y , коэффициент guidance λ , число шагов T
- 2: Инициализировать $x_T \sim \mathcal{N}(0, I)$
- 3: **for** $t = T, T - 1, \dots, 1$ **do**
- 4: **Шаг 1: Получить диффузионный шум**
 $\epsilon_{\theta} = \text{DiffusionModel}(x_t, t)$
- 6: **Шаг 2: Вычислить градиент классификатора**
Включить режим вычисления градиентов для x_t
 $f = \text{Classifier}(x_t, t, y)$
 $\sigma_f = \sigma(f)$ ▷ Вероятность правильного продолжения
- 10: Вычислить градиент: $g = \nabla_{x_t} \log p(y|x_t) = (1 - \sigma_f) \cdot \nabla_{x_t} f$
- 11: **Шаг 3: Модифицировать среднее обратного шага**
Вычислить $\mu_{\theta} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_{\theta} \right)$
- 13: $\tilde{\mu} = \mu_{\theta} + \lambda \sigma_t^2 g$
- 14: **Шаг 4: Сэмплировать следующее состояние**
Сгенерировать шум $\xi \sim \mathcal{N}(0, I)$
 $x_{t-1} = \tilde{\mu} + \sigma_t \xi$
- 17: Выключить режим вычисления градиентов для x_t
- 18: **end for**
- 19: **Выход:** x_0 ▷ Сгенерированное продолжение префикса y
