FPS-loss relation to mutual information

Kirill Tyshchuk

March 2021

Рассмотрим первую строку в матрице, к которой мы применяем FPS-loss. Пусть переменная z соответствует строке, и ей соответствует один позитивный пример x из столбцов и $N-1$ негативных примеров x_i^{neg}.

Пусть $f(x,z)$ обозначает выход нашей модели (включая экспоненту из Softmax):

$$f(x,z) = \exp(\langle embed(x), embed(z)\rangle)$$

Распишем матожидание вклада этой строки в минус FPS-loss, то есть кросс-энтропии с таргетом $(1,0,...,0)$:

$$-\mathbb{E}\mathcal{L}_{row} = -\mathbb{E}\left[\log \frac{f(x,z)}{f(x,z)+\sum\limits_{i=1}^{N-1} f(x_i^{negative},z)}\right] = \tag{1}$$

$$= \mathbb{E}\log f(x,z) - \mathbb{E}\log(f(x,z) + \sum_{i=1}^{N-1} f(x_i^{negative},z)) \leq \tag{2}$$

$$\leq \mathbb{E}\log f(x,z) - \mathbb{E}\log(\sum_{i=1}^{N-1} f(x_i^{negative},z)) \approx \tag{3}$$

$$\approx \mathbb{E}\log f(x,z) - \log(\mathbb{E}f(x^{neg},z)) - \log(N-1) = \tag{4}$$

$$= \mathbb{E}log\frac{f(x,z)}{\mathbb{E}f(x^{neg},z)} - log(N-1) \tag{5}$$

Если у нас нет каких-то сложных негативов, то здесь фигурируют два распределения: пара с позитивным примером (x,z) берётся из совместного распределения $\mathbb{P}_{XZ}$, а пары с негативным (x^{neg},z) берутся из произведения маргинальных распределений $\mathbb{P}_X \otimes \mathbb{P}_Z$ (взяли z и случайный x, не обязательно относящийся к делу). KL-дивергенция между этими рапспределениями равна взаимной информации между X и Z:

$$I(X,Z) = D_{KL}(\mathbb{P}_{XZ}||\mathbb{P}_X \otimes \mathbb{P}_Z) = \mathop{\mathbb{E}}_{\mathbb{P}_{XZ}} \frac{d\mathbb{P}_{\mathbb{XZ}}}{d\mathbb{P}_X \otimes \mathbb{P}_Z}$$

Введём третье распределение $\mathbb{G}$:

$$d\mathbb{G}(x,z) \propto f(x,z)p(x)p(z) = f(x,z)d\mathbb{P}_X \otimes \mathbb{P}_Z$$

Нормировочная константа как раз равна

$$C_f = \iint f(x,z)p(x)p(z)dxdz = \mathop{\mathbb{E}}_{\mathbb{P}_X \otimes \mathbb{P}_Z} f(x,z) = \mathbb{E}f(x^{neg}, z)$$

Распишем выражение из (5) в терминах взаимной информации и распределения $\mathbb{G}$:

$$\mathbb{E}log\frac{f(x,z)}{\mathbb{E}f(x^{neg},z)} = \mathbb{E}log\frac{d\mathbb{G}}{d\mathbb{P}_X \otimes \mathbb{P}_Z} = \tag{6}$$

$$= \mathbb{E}log\frac{d\mathbb{P}_{XZ}}{d\mathbb{P}_X \otimes \mathbb{P}_Z} - \mathbb{E}log\frac{d\mathbb{P}_{XZ}}{d\mathbb{G}} = \tag{7}$$

$$= D_{KL}(\mathbb{P}_{XZ}||\mathbb{P}_X \otimes \mathbb{P}_Z) - D_{KL}(\mathbb{P}_{XZ}||\mathbb{G}) = \tag{8}$$

$$= I(X,Z) - D_{KL}(\mathbb{P}_{XZ}||\mathbb{G}) \leq I(X,Z) \tag{9}$$

Собирая всё в кучу, получаем, что минус FPS-loss по всем N строкам имеет матожидание

$$-\mathbb{E}\mathcal{L}_{row} \approx N(I(X,Z) - D_{KL}(\mathbb{P}_{XZ}||\mathbb{G}) - \log(N-1))$$

Кроме того, оптимум достигается при

$$d\mathbb{G} = d\mathbb{P}_{XZ} \iff f(x,z) \propto \frac{p(x,z)}{p(x)p(z)} = \exp(PMI(x,z))$$

Выкладки получены адаптацией рассуждений из статьи об InfoNCE[2] и аппендикса из статьи о MINE[1]. В аппендиксе статьи об InfoNCE связь со взаимной информацией получена немного по-другому.

Список литературы

[1] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. MINE: Mutual Information Neural Estimation. *arXiv e-prints*, page arXiv:1801.04062, January 2018.

[2] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv e-prints*, page arXiv:1807.03748, July 2018.