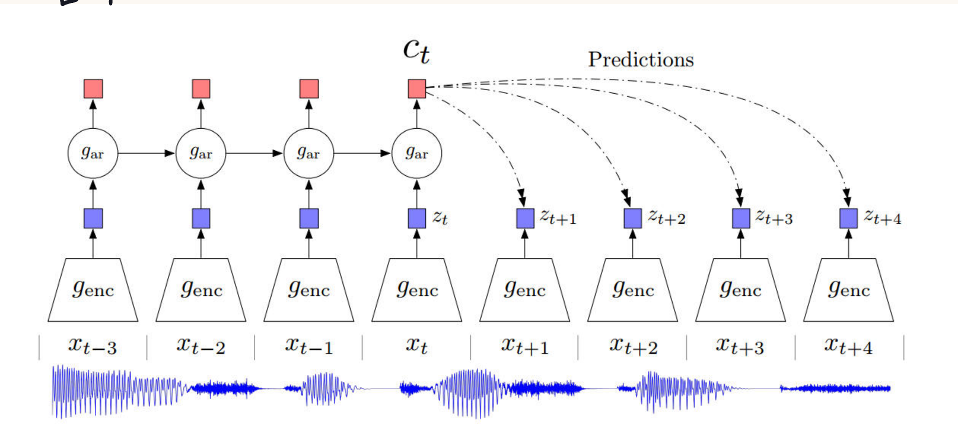
CPC (contrastive Predictive Coding) 表征学习

框架



原始数据 $x \xrightarrow[g_{enc}]{\text{非线性编码器}}$ 潜在表征序列 z_t $\xrightarrow[g_{ar}]{\text{自回归函数}}$ 历史信息表征集 c_t

$z_t = g_{enc}(x_t)$

$c_t = g_{ar}(z_{\le t})$

利用历史信息 c_t 来预测未来表征 z_{t+k}

我们把模型的预测能力建模成函数 $f_k(x_{t+k}, c_t)$

$$f_k(x_{t+k}, c_t) = \exp(\underbrace{z_{t+k}^T}_{\text{真实特征}} \underbrace{\boxed{W_k c_t}}_{\text{根据历史信息，经过权重}W_k\text{预测值}})$$

我们希望对于输入 x_{t+k}, c_t, f_k 的输出尽可能大

W_k 是可学习的参数

那么如何学习到 W_k 呢?

通过构建正、负样本对，使正样本对得分高于负样本对

一组 sample $X = \{x_1, x_2, \cdots, x_n\}$，其中：

正样本对 $\{c_t^{(i)}, x_i\}$，其余为负样本对

所以这里存在一个分类任务：

我们希望：

$$p(d=i \mid X, c_t) \gg p(d=j_{(j\neq i)} \mid X, c_t)$$

$$p(d=i \mid X, c_t) = \frac{p(d=i, X \mid c_t)}{\sum_{j=1}^{N} p(d=j, X \mid c_t)} = \frac{p(x_i \mid c_t) \cdot \prod_{l\neq i} p(x_l)}{\sum_{j=1}^{N}\left[p(x_j \mid c_t)\prod_{l\neq j} p(x_l)\right]}$$

$$= \frac{\frac{p(x_i \mid c_t)}{p(x_i)}}{\sum_{j=1}^{N} \frac{p(x_j \mid c_t)}{p(x_j)}}$$

$p(d=i \mid X, c_t)$ 是正确分类的条件概率

可以认为 正确预测 和 正确分类 两个任务是一定程度等价的

所以可以得出

$$f_k(x_{t+k}, c_t) \propto \frac{p(x_{t+k} \mid c_t)}{p(x_{t+k})}$$

涉及真实条件概率
和高维数据，
很难计算

损失函数设计：

$$\mathcal{L}_N = -\mathop{E}_{X}\left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}\right]$$

进一步证明损失函数的有效性

$$\mathcal{L}_N = -\mathop{E}_{X}\left[\log \frac{\frac{p(x_{t+k} \mid c_t)}{p(x_{t+k})}}{\frac{p(x_{t+k} \mid c_t)}{p(x_{t+k})} + \sum_{x_j \in X_{neg}} \frac{p(x_j \mid c_t)}{p(x_j)}}\right]$$

$$= \mathop{E}_{X} \log\left[1 + \frac{p(x_{t+k})}{p(x_{t+k}|c_t)} \sum_{x_j \in X_{neg}} \frac{p(x_j|c_t)}{p(x_j)}\right]$$

$$\approx \mathop{E}_{X} \log\left[1 + \frac{p(x_{t+k})}{p(x_{t+k}|c_t)} (N-1) \mathop{E}_{x_j} \frac{p(x_j|c_t)}{p(x_j)}\right]$$

$$= \mathop{E}_{X} \log\left[1 + \frac{p(x_{t+k})}{p(x_{t+k}|c_t)} (N-1)\right] \qquad \frac{p(x_{t+k})}{p(x_{t+k}|c_t)} < 1$$

$$\geq \mathop{E}_{X} \log\left[\frac{p(x_{t+k})}{p(x_{t+k}|c_t)} N\right]$$

$$= \mathop{E}_{X} \log \frac{p(x_{t+k})}{p(x_{t+k}|c_t)} + \log N$$

$$= -\sum_{x \in X} p(x, c) \log \frac{p(x_{t+k}|c_t)}{p(x_{t+k})} + \log N$$

$$= \underbrace{-I(x_{t+k}, c_t)}_{\text{互信息}} + \log N$$

$$\therefore I(x_{t+k}, c_t) \geq \log N - \mathcal{L}_N$$

$\mathcal{L}_N \downarrow$, $I \uparrow$