Entropy and Cross Entropy and Kl divergence

熵 Entropy

我们用熵（Entropy）来描述一个系统或一个事件的不确定性。”熵“这一概念本身来自于热力学，香农的信息熵是本文的主要讨论内容。

> 回顾一下概率论，考虑一个随机试验 E ，举例为”抛掷一颗骰子，观察出现的点数“；随机试验 E 所有可能出现的基本事件（基本事件：结果单一不可再分解，每一个基本事件彼此不相容）组成的集合，称为样本空间 S ，对于本例就是，$S=\{1_{点朝上},2_{点朝上},...\}$；考虑事件域 $\mathscr{F}$ 中的一个元素/事件”骰子点数为偶数“，即$A=\{_{骰子点数为偶数}\}$，用随机变量 X 表示（随机变量），则 $X=\{2,4,6\}$；（随机变量是一个实值函数，将一个事件映射为一个值）
>
> https://www.zhihu.com/question/20642770
>
> （后续另辟一章笔记：概率论三要素）

考虑一事件A，其对应离散随机变量X的取值范围为$\{x_1,x_2,...,x_N\}$,则离散随机变量X的信息熵定义如下：

$$H(X)=-\sum_{i=1}^{N}p(x_i)\cdot\log p(x_i) \tag{1}$$

其中$p(\cdot)$表示其中基本事件发生的概率。

无损编码长度的最小长度

交叉熵 Cross Entropy

考虑两个概率分布p,q 交叉熵定义如下：

$$H(p,q)=-\sum_{x}p(x)\log q(x) \tag{2}$$

表示用q去编码p的冗余编码长度

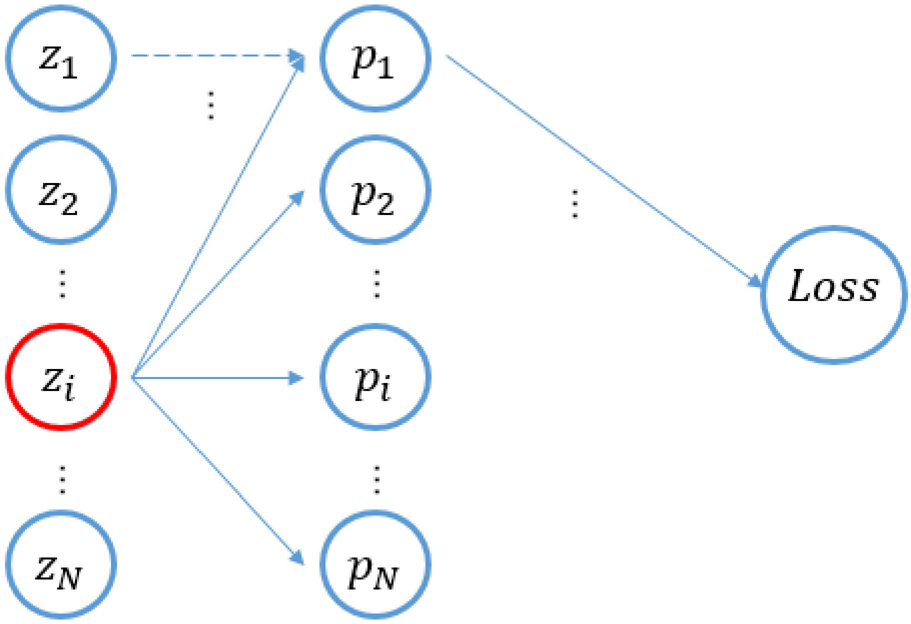
假定标签分布/真实分布为 y_i ，网络/模型输出 $logits$ 经过$softmax$处理后分布为 $p(x_i)$ ，则交叉熵计算如下，

$$\mathcal{L}=H(y,p)=-\sum_{i=1}^{N}y_i\log p(x_i) \tag{3}$$

其中 $softmax$ 处理操作如下，

$$p(x_i)=\frac{\exp(z_i)}{\sum_{j=1}^{N}\exp(z_j)} \tag{4}$$

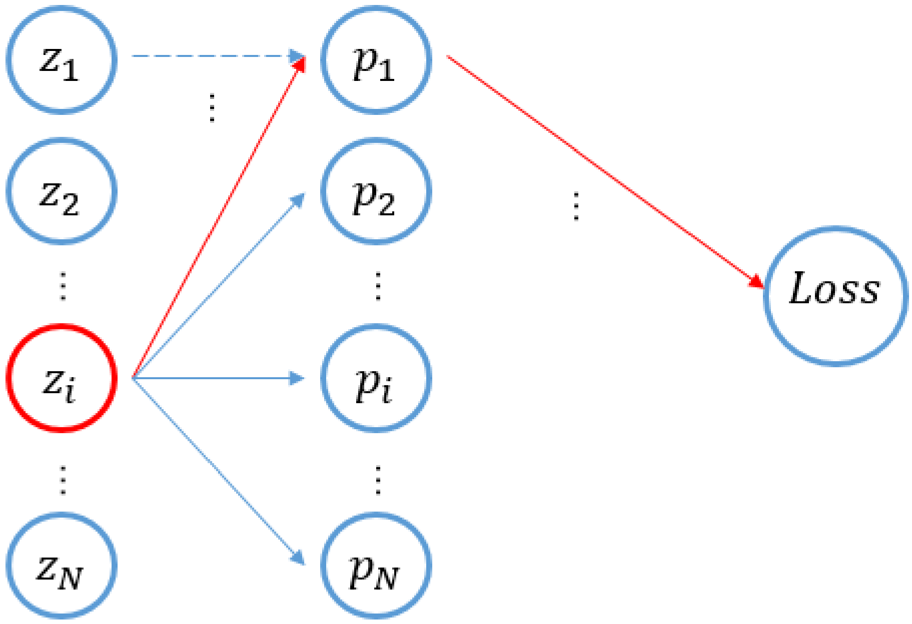
做gradient flow图如下，



接下来我们考虑交叉熵损失函数对$logits$ 的导数：

$$\frac{\partial \mathcal{L}}{\partial z_i} = \sum_{j=1}^{N} \left(\frac{\partial \mathcal{L}}{\partial p_j} \cdot \frac{\partial p_j}{\partial z_i} \right) \tag{5}$$

求和中的子项即计算图中的一条"路径"：



由于 $softmax$ 操作，所有 p_j 都是 z_i 的函数，所以需要求和操作；但由于$softmax$操作，i, j 不同/相同的情况下 p_j 对 z_i 的偏导不一致，当 $i \neq j$ 时有：

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial p_j} \cdot \frac{\partial p_j}{\partial z_i}\Big|_{i \neq j} &= \frac{\partial}{\partial p_j}\left[-\sum_{k=1}^{N} y_k \log p_k\right] \cdot \frac{\partial p_j}{\partial z_i} && (6) \\
&= -\frac{y_j}{p_j} \cdot \frac{\partial p_j}{\partial z_i} && (7) \\
&= -\frac{y_j}{p_j} \cdot \frac{\partial}{\partial z_i}\left[\frac{\exp(z_j)}{\sum_{k=1}^{N} \exp(z_k)}\right] && (8) \\
&= -\frac{y_j}{p_j} \cdot \frac{\partial}{\partial z_i}\left[\frac{\exp(z_j)}{\sum_{k \neq i}^{N} \exp(z_k) + \exp(z_i)}\right] && (9) \\
&= -\frac{y_j}{p_j} \cdot \left[\frac{-\exp(z_i) \cdot \exp(z_j)}{[\sum_{k=1}^{N} \exp(z_k)]^2}\right] && (10) \\
&= -\frac{y_j}{p_j} \cdot \left[- \cdot \frac{\exp(z_i)}{\sum_{k=1}^{N} \exp(z_k)} \cdot \frac{\exp(z_j)}{\sum_{k=1}^{N} \exp(z_k)}\right] && (11) \\
&= -\frac{y_j}{p_j} \cdot -p_i \cdot p_j && (12) \\
&= y_j \cdot p_i && (13)
\end{aligned}
$$

同理，当 $j = i$ 时，求导如下：

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial p_j} \cdot \frac{\partial p_j}{\partial z_i}\Big|_{i = j} &= \frac{\partial}{\partial p_i}\left[-\sum_{k=1}^{N} y_i \log p_k\right] \cdot \frac{\partial p_i}{\partial z_i} && (14) \\
&= -\frac{y_i}{p_i} \cdot \frac{\partial p_i}{\partial z_i} && (15) \\
&= -\frac{y_i}{p_i} \cdot \frac{\partial}{\partial z_i}\left[\frac{\exp(z_i)}{\sum_{k=1}^{N} \exp(z_k)}\right] && (16) \\
&= -\frac{y_i}{p_i} \cdot \frac{\partial}{\partial z_i}\left[\frac{\exp(z_i)}{\sum_{k \neq i}^{N} \exp(z_k) + \exp(z_i)}\right] && (17) \\
&= -\frac{y_i}{p_i} \cdot \left[\frac{\exp(z_i) \cdot \sum_{k=1}^{N} \exp(z_k) - \exp(z_i) \cdot \exp(z_i)}{[\sum_{k=1}^{N} \exp(z_k)]^2}\right] && (18) \\
&= -\frac{y_i}{p_i} \cdot \left[\cdot \frac{\exp(z_i)}{\sum_{k=1}^{N} \exp(z_k)} \cdot \frac{\left(\sum_{k=1}^{N} \exp(z_k)\right) - \exp(z_i)}{\sum_{k=1}^{N} \exp(z_k)}\right] && (19) \\
&= -\frac{y_i}{p_i}\left[\cdot \frac{\exp(z_i)}{\sum_{k=1}^{N} \exp(z_k)} \cdot \left(1 - \frac{\exp(z_i)}{\sum_{k=1}^{N} \exp(z_k)}\right)\right] && (20) \\
&= -\frac{y_i}{p_i} \cdot p_i \cdot (1 - p_i) && (21) \\
&= y_i \cdot (p_i - 1) && (22)
\end{aligned}
$$

考虑公式(5)，则求导改写为：

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial p_j} \cdot \frac{\partial p_j}{\partial z_i} &= \sum_{j \neq i}^{N} y_j \cdot p_i + y_i \cdot (p_i - 1) && (23) \\
&= p_i \sum_{j \neq i}^{N} y_j + (p_i \cdot y_i - y_i) && (24) \\
&= p_i \sum_{j}^{N} y_j - y_i && (25) \\
&= p_i - y_i && (26)
\end{aligned}
$$

> 那为什么label smooth能work？（从求导角度的一种解释）

label smooth 区别于传统 one-hot的label：

$$y_i = \begin{cases} 1-\epsilon & ,\text{i is true label} \\ \dfrac{\epsilon}{K-1} & ,\text{otherwise} \end{cases} \tag{27}$$

由于$p_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$，对于one-hot的label来说要求$\exp(z_i) = \sum_{j\neq i} \exp(z_j) + \exp(z_i)$，也就是要求$\sum_{j\neq i} \exp(z_j) = 0$ 要求所有不是label对应维度的logits趋近于负无穷，而对于label smooth来说要求$\exp(z_i) = (1-\epsilon)\left[\sum_{j\neq i} \exp(z_j) + \exp(z_i)\right]$, 移项化简有：$\exp(z_i) = \frac{(1-\epsilon)}{\epsilon} \cdot \sum_{j\neq i} \exp(z_j)$；有助于优化（待续）

KL散度 Kullback-Leibler Divergence