$x \in R^{n\times 1}$ $W \in R^{m\times n}$ $E \in R^{m\times d}$

$s = Wx \in R^{m\times 1}$

$\hat{s} = \text{argtopk}(s) \Rightarrow s$ 中topk元素的index $\in R^{k\times 1}$，其每个元素 $\in [0, \cdots, m-1]$

hard: $res = \text{one\_hot}(\hat{s})E = E[\hat{s}] \in R^{k\times d}$

$loss = (res**2).sum()$

loss.backward() ⇒ 对E有梯度(仅从topk对应的entry)，对W没梯度

soft: $res = \text{one\_hot}(\hat{s})E \odot \sigma(s)$

$\sigma(s)$ → $\begin{cases} 0 & s_i < \gamma \\ s_i & s_i \geq \gamma \end{cases}$

$loss = (res**2).sum()$

loss.backward() ⇒ 对E有梯度(同上)，对W有梯度 (仅从 $s_i \geq \gamma$ 的entry)