Notation

- z_i: i-th logit of the student model (pre-softmax)
- v_i: i-th logit of the teacher model
- T: temperature
- $p_i = \text{softmax}(z_i/T)$: softened student probability
- $q_i = \text{softmax}(v_i/T)$: softened teacher probability
- N: number of classes

KL-based Distillation Loss

$$L_{\text{KD}} = T^2 \,\text{KL}(q \,\|\, p) = T^2 \sum_{i=1}^{N} q_i \log \frac{q_i}{p_i} = -T^2 \sum_{i=1}^{N} q_i \log p_i + \text{const.} \tag{1}$$

Since q_i does not depend on z_i, we can ignore the constant term in the gradient.

—

Gradient w.r.t. Student Logits z_j

$$\frac{\partial L_{\text{KD}}}{\partial z_j} = -T^2 \sum_{i=1}^{N} q_i \frac{\partial}{\partial z_j} \log p_i \tag{2}$$

$$= -T^2 \sum_{i=1}^{N} q_i \frac{1}{p_i} \frac{\partial p_i}{\partial z_j} \tag{3}$$

—

Derivative of Softmax

The softmax derivative w.r.t. logits is:

$$\frac{\partial p_i}{\partial z_j} = \frac{1}{T} p_i (\delta_{ij} - p_j) \tag{4}$$

where δ_{ij} is the Kronecker delta:

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

—

Substitute Softmax Derivative

$$\frac{\partial L_{\text{KD}}}{\partial z_j} = -T^2 \sum_{i=1}^{N} q_i \frac{1}{p_i} \cdot \frac{1}{T} p_i(\delta_{ij} - p_j) \tag{5}$$

$$= -T \sum_{i=1}^{N} q_i(\delta_{ij} - p_j) \tag{6}$$

$$= -T \left(q_j - \sum_{i=1}^{N} q_i p_j \right) \tag{7}$$

Since $\sum_{i=1}^{N} q_i = 1$, this simplifies to:

$$\frac{\partial L_{\text{KD}}}{\partial z_j} = T(p_j - q_j) \tag{8}$$

—

Approximation for Zero-Mean Logits

If the logits are zero-meaned, i.e.,

$$\sum_i z_i = 0, \quad \sum_i v_i = 0,$$

and for small logits (linear approximation of softmax):

$$p_i \approx \frac{1}{N} + \frac{z_i}{N}, \quad q_i \approx \frac{1}{N} + \frac{v_i}{N},$$

then

$$\frac{\partial L_{\text{KD}}}{\partial z_i} \approx \frac{1}{N}(z_i - v_i) \tag{9}$$

which shows that the gradient is simply the **difference between student and teacher logits**, scaled by $1/N$.

—