Derivazione della InfoNCE Loss

Introduzione

Questa relazione fornisce una derivazione dettagliata del gradiente della InfoNCE loss rispetto ai vettori di feature Z. La perdita InfoNCE è utilizzata nei metodi di *self-supervised learning* contrastivo per avvicinare coppie di esempi "positivi" e allontanare esempi "negativi" nello spazio delle rappresentazioni.

1 Definizione della perdita

Siano $N = 2B$ il numero totale di vettori di feature e $Z = \{Z_1, Z_2, \ldots, Z_N\}$ con $Z_i \in \mathbb{R}^d$ e $\|Z_i\|_2 = 1$. Si definiscono:

$$
\begin{aligned}
L_{ij} &= \frac{1}{\tau}\, Z_i \cdot Z_j, \\
P_{ij} &= \frac{\exp(L_{ij})}{\sum_{k=1}^{N} \exp(L_{ik})}, \\
\mathcal{L} &= -\frac{1}{N} \sum_{i=1}^{N} \log P_{i,p(i)},
\end{aligned}
$$

ove $p(i) = (i + B) \bmod N$ identifica l'indice del campione positivo per ogni riga i.

2 Derivata rispetto ai logits

Per ciascuna riga i si considera

$$
\ell_i = -\log P_{i,p(i)}
$$

La derivata di ℓ_i rispetto a L_{ij}, applicando la cross-entropy sulla softmax, è:

$$
\frac{\partial \ell_i}{\partial L_{ij}} = P_{ij} - \begin{cases} 1 & j = p(i), \\ 0 & j \neq p(i). \end{cases}
$$

Poiché $\mathcal{L} = \frac{1}{N} \sum_i \ell_i$, allora

$$
\frac{\partial \mathcal{L}}{\partial L_{ij}} = \frac{1}{N} \big(P_{ij} - \mathbb{1}_{j=p(i)} \big).
$$

3 Derivata dei logits rispetto a Z

Ricordando $L_{ij} = \frac{1}{\tau} Z_i \cdot Z_j$, si ha per ogni componente vettoriale Z_k:

$$
\frac{\partial L_{ij}}{\partial Z_k} = \frac{1}{\tau} (\delta_{ik} Z_j + \delta_{jk} Z_i),
$$

dove δ_{ik} è il delta di Kronecker.

4 Catena di derivazione finale

Applicando la regola della catena:

$$\frac{\partial \mathcal{L}}{\partial Z_k} = \sum_{i,j} \frac{\partial \mathcal{L}}{\partial L_{ij}} \frac{\partial L_{ij}}{\partial Z_k}.$$

Separando i contributi e definendo la matrice G con

$$G_{ij} = P_{ij} - \mathbb{1}_{j=p(i)},$$

si ottiene:

$$\frac{\partial \mathcal{L}}{\partial Z_k} = \frac{1}{N\tau}\Big(\sum_j G_{kj} Z_j + \sum_i G_{ik} Z_i\Big).$$

In forma compatta:

$$\nabla_Z \mathcal{L} = \frac{1}{N\tau}(G + G^T)Z.$$

Conclusioni

La derivazione mostra come il gradiente della InfoNCE loss sia ottenuto dalla combinazione delle probabilità softmax corrette dai marcatori dei positivi, moltiplicata per la matrice delle feature. Il termine simmetrico $(G + G^T)$ garantisce che ogni coppia di feature contribuisca in modo reciproco alla spinta verso i positivi e all'allontanamento dai negativi.