InfoNCE Loss Derivation

Introduction

This document provides a detailed derivation of the InfoNCE loss gradient with respect to the feature vectors Z. The InfoNCE loss is used in contrastive *self-supervised learning* methods to bring "positive" example pairs closer together and push "negative" examples apart in the representation space.

1 Loss definition

Let $N = 2B$ be the total number of feature vectors and $Z = \{Z_1, Z_2, \ldots, Z_N\}$ with $Z_i \in \mathbb{R}^d$ and $\|Z_i\|_2 = 1$. We define:

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

where $p(i) = (i + B) \bmod N$ identifies the index of the positive sample for each row i.

2 Derivative with respect to logits

For each row i we consider

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

The derivative of ℓ_i with respect to L_{ij}, applying cross-entropy on softmax, is:

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

Since $\mathcal{L} = \frac{1}{N} \sum_i \ell_i$, then

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

3 Derivative of logits with respect to Z

Remembering $L_{ij} = \frac{1}{\tau} Z_i \cdot Z_j$, for each vector component Z_k we have:

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

where δ_{ik} is the Kronecker delta.

4 Final derivation chain

Applying the chain rule:

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

Separating the contributions and defining matrix G with

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

we obtain:

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

In compact form:

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

Conclusions

The derivation shows how the InfoNCE loss gradient is obtained from the combination of softmax probabilities corrected by positive markers, multiplied by the feature matrix. The symmetric term $(G+G^T)$ ensures that each feature pair contributes reciprocally to pushing toward positives and away from negatives.