下面给出一组 **SimCLR 及其损失函数 (InfoNCE Loss)** 相关的**计算题**与**应用题**。这些题目既包括**数值计算**，也包含**方法应用**，可用于更深入地理解 SimCLR 的工作机制与公式含义。

一、计算题

题目 1：批次相似度矩阵计算

假设我们有一个 **batch size = 3**，每张图像做了 2 种不同的数据增强，因此**总 embeddings 数为 $2N=6$**。我们记这 6 个嵌入向量为 ${z_1, z_2, z_3, z_4, z_5, z_6}$，其中：

- **正样本对**: (z_1, z_2), (z_3, z_4), (z_5, z_6)
- 不同对 (z_i, z_j with $i\neq j$) 视为负样本

现给定一个 **相似度矩阵** S，其中 $S_{ij} = \text{sim}(z_i, z_j)$（余弦相似度），如下：

$S = \begin{pmatrix} - & 0.95 & 0.10 & 0.05 & 0.02 & 0.00 \ 0.95 & - & 0.12 & 0.07 & 0.03 & 0.01 \ 0.10 & 0.12 & - & 0.90 & 0.06 & 0.04 \ 0.05 & 0.07 & 0.90 & - & 0.08 & 0.02 \ 0.02 & 0.03 & 0.06 & 0.08 & - & 0.93 \ 0.00 & 0.01 & 0.04 & 0.02 & 0.93 & - \end{pmatrix}$

其中 “-” 表示与自身的相似度无需计算。假设温度 $\tau = 0.1$。

1. **请计算** 对于正样本对 (z_1, z_2)，在 InfoNCE 损失公式下：

   L1,2=−logexp(sim(z1,z2)/τ)∑k=161k≠1exp(sim(z1,zk)/τ)L*{1,2} = -\log \frac{\exp(\text{sim}(z_1, z_2) / \tau)} {\sum*{k=1}^{6} \mathbf{1}_{k \neq 1} \exp(\text{sim}(z_1, z_k)/\tau)}

   - 写出分子与分母的具体数值，并求出最后的 $-\log(\dots)$ 值。
2. 同理，**计算** 正样本对 (z_3, z_4) 的损失 $L_{3,4}$。
3. **请问** 若我们要计算整个 batch 的总损失 L，应如何组合这 3 对正样本的损失？

> **提示**：对每个正样本对 (i,j) 都要单独计算一个损失 $L_{i,j}$，然后再取平均。

题目 2：温度参数的影响

1. 若将上题中的温度 τ 从 0.1 调整为 0.5，你预计分母中的指数项会发生怎样的变化？对最终损失值会有什么直观影响？
2. 在实际训练中，**调高温度** 通常会使相似度差异被“拉近”，请结合公式简要说明原因。

题目 3：小批次对损失的影响

1. 假设我们只能使用 **batch size = 2** (即 $N=1$)，那么**负样本**的数量变为 0（因为同一个 batch 里只有 1 对正样本，没有其他图像可作负样本）。
   - 试问，在 InfoNCE 公式中，此时分母会如何？损失还能否正常计算？
   - 实际训练会遇到什么问题？
2. 简述为什么 SimCLR 通常需要很大 batch size (如 256~2048) 才能取得较好效果。

二、应用题

题目 4：SimCLR 在小数据集上的应用

你有一个小型数据集（仅 1 万张图像），想用 **SimCLR** 进行自监督预训练，然后再做线性探针 (linear probe) 进行分类。请回答：

1. 如何在数据增强 (data augmentation) 上做设计，以确保对比学习能有效？请给出至少 3 种增强手段，并说明它们的意义。
2. 如果 GPU 资源有限，无法使用很大的 batch size，你可以采用哪些策略来近似地实现“丰富的负样本”？（例如使用队列、记忆库 (memory bank) 等思想，或借鉴 MoCo 的做法）

题目 5：SimCLR 表征的可视化与解释

在训练完 SimCLR 后，你希望**可视化**它学到的特征，以验证其对不同图像的区分效果：

1. 你可以如何操作来查看**同一张图像不同增强**在表示空间中的距离？
2. 如果发现某些不同图像的表示非常接近，可能是什么原因导致？
3. 若你想在特征空间中做 **t-SNE** 或 **UMAP** 降维，可观察到什么现象来验证 SimCLR 确实把相似图像拉近？

题目 6：SimCLR 与监督学习的对比

1. 假设你在 ImageNet 上分别训练一个 **有监督 ResNet** 与一个 **SimCLR ResNet (自监督)**，然后在某个下游任务（如分类、检测）中只用少量标注数据做微调。
   - 你认为两者在下游任务的表现可能有何差异？
   - 若标注数据极少（如 1% ImageNet labels），哪种方法更有优势？为什么？
2. SimCLR 训练好的表示能否**直接**用来做分类，而不经过微调？若可以，怎么做？它的准确率会接近有监督模型吗？

总结

- 以上 **计算题** 与 **应用题** 涉及了 **SimCLR 的损失公式 (InfoNCE)**、**温度参数**、**batch size**、**对比学习在小数据集或资源受限场景的应对**，以及 **如何可视化/验证自监督学到的特征。**
- 通过这些问题，你可以更深入理解 **SimCLR** 的数学原理与实际应用挑战，并与有监督学习进行比较，体会到大批量负样本、数据增强在对比学习中的关键作用。