具有抗噪对比损失的部分视图对齐表示学习

Mouxing Yang1 , Yunfan Li1 , Zhenyu Huang1 , Zitao Liu2 , Peng Hu1 , Xi Peng1* 1 College of Computer Science, Sichuan University. 2 TAL Education Group, Beijing China.

{yangmouxing, yunfanli.gm, zyhuang.gm, zitao.jerry.liu, penghu.ml, pengx.gm}@gmail.com

摘要

在实际应用中，由于空间、时间或时空异步，通常只有一部分数据跨视图对齐，从而导致所谓的部分视图对齐问题 (PVP)。为了在没有标签帮助的情况下解决这种较少接触的问题，我们建议使用抗噪对比损失同时学习表示和对齐数据。简而言之，对于来自一个视图的每个样本，我们的方法旨在从其他视图中识别其类别内对应物，因此可以建立跨视图对应关系。由于对比学习需要数据对作为输入，我们使用已知对应关系构建正对，使用随机采样构建负对。为了减轻甚至消除由随机采样引起的假阴性的影响，我们提出了一种抗噪声对比损失，可以自适应地防止假阴性主导网络优化。据我们所知，这可能是使对比学习对嘈杂标签具有鲁棒性的第一次成功尝试。事实上，这项工作可能会通过嘈杂的标签显着丰富学习范式。更具体地说，传统的嘈杂标签被定义为对分类等监督任务的错误注释。相比之下，这项工作提出视图对应可能是错误的，这与广泛接受的噪声标签定义截然不同。大量实验表明，与聚类和分类任务中的 10 种最先进的多视图方法相比，我们的方法具有良好的性能。代码将在 https://pengxi.me 公开发布。

1.简介

多视图表示学习 (MvRL) [2, 16, 24, 37, 44] 旨在从多视图/模态数据中学习一致的表示，以促进下游任务，包括但不限于聚类、分类和检索。所有现有作品 [2, 37, 44] 的成功在很大程度上依赖于两个假设，即数据的完整性和视图的一致性。 具体来说，完整性假设要求实例在所有视图中呈现，一致性假设要求来自不同视图的数据必须严格对齐。当这两个假设之一不满足时，就不可能执行 MvRL。 然而，在实践中，这两个假设在数据收集或传输中很容易被违反，从而导致部分数据丢失问题（PDP）和部分视图对齐问题（PVP，见图 1（a））更具体地说，PDP 发生在某些视图中丢失某些数据时，从而导致数据不完整。 PVP 是指只对齐了一部分数据，从而导致数据不一致的情况。 最近，有几项工作在 PDP 上取得了显着进展 [15, 27, 40]，但只有少数研究用于解决 PVP。

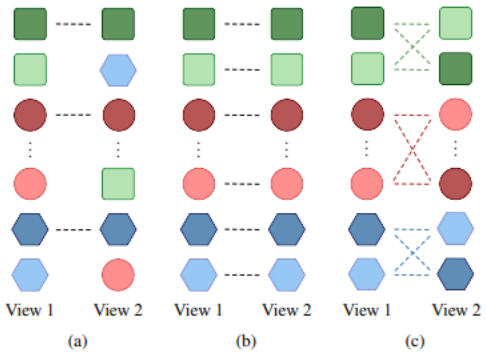

Figure 1. 本文的动机。 图中，不同的颜色表示不同的实例将出现在多个视图中，不同的形状表示不同的类别，虚线表示所需的对应关系。(a)部分视图对齐问题：由于数据收集和传输的复杂性，只有一部分数据与已知的对应关系；(b) 实例级对齐：旨在建立同一实例的两个跨视图样本之间的对应关系；(c) 类别级对齐：每对由属于同一类别的样本组成。 考虑到聚类和分类等下游任务，类别级对齐比实例级对齐更可取，因为它具有更高的可访问性和可扩展性。

在本文中，我们尝试在没有数据注释的帮助下解决 PVP。 我们的观察和动机如图 1 所示。理想情况下，高度期望数据在实例级别完全对齐，如图 1(b) 所示。为了实现这一目标，一个简单的解决方案是使用匈牙利算法作为预处理步骤来构建两个视图的对应关系，然后将对齐的数据传递到标准的多视图方法中来学习表示。然而，这种两阶段学习范式的性能是次优的，因为匈牙利算法 i) 不能应用于异构的多视图原始空间； ii) 不会利用数据中的已知对应关系。最近，部分视图对齐聚类 (PVC) [18] 提出了一种匈牙利算法的可微神经模块，因此数据对齐和表示学习可以通过一个阶段的方式实现。 然而，vanilla Hungarian 算法和 PVC 都旨在实现实例级对齐，这对于多视图聚类和分类可能过于充足。与检索[8, 17]等一对一映射任务不同，聚类和分类的本质是一对多映射。 因此，由于其更高的可访问性和可扩展性，类别级对齐比实例级对齐更适用于聚类和分类。直观地说，对于给定的跨视图实例，它在实例和类别级别正确对齐的随机概率为 1/N 和 1/K，其中 N 和 K 是实例和类别的数量，K ≪ N。换句话说，类别级别的对齐具有更高的可访问性。 另一方面，匈牙利算法等实例级对齐方法的计算复杂度为$O(N^3)$，这使其无法处理大规模数据集。

基于上述观察和动机，我们通过尝试实现类别而不是实例级对齐来解决 PVP，如图 1(c) 所示。 最后，我们提出了一种新颖的部分视图对齐的表示学习方法，称为具有噪声鲁棒损失的多视图对比学习（MvCLN）。我们的基本思想是将视图对齐问题重新表述为识别任务。具体来说，以双视图数据为展示，对于来自一个视图的每个样本，MvCLN 旨在从另一个视图识别属于同一类别的对应物。为了训练 MvCLN，我们使用可用的对齐数据构建正对，使用随机采样构建负对 (NP)。为了减轻甚至消除由随机采样引起的假负对 (FNP) 的影响，我们的 MvCLN 具有新颖的抗噪声对比损失。 这项工作的贡献可以总结如下：

- 为了促进像聚类这样的一对多映射任务，我们建议通过建立类别而不是实例级对齐来解决 PVP。 如上述分析和以下实验所示，这种特定于任务的对齐具有更高的可访问性和可扩展性；
- 我们将对齐问题重新表述为在对比学习框架下进一步执行的视图识别任务。 据我们所知，这可能是通过对比学习实现类别级别对齐的首批作品之一；
- 为了使用对比学习建立视图对应，我们提出了一种新的抗噪声对比损失，可以减轻甚至消除在对构建过程中引入的噪声标签（即 FNP）的影响。 据我们所知，这可能是第一个具有处理噪声标签能力的对比学习方法。 **需要指出的是，传统的嘈杂标签被定义为对分类等监督任务的错误注释。 相比之下，这项工作提出视图对应可能是错误的，这与传统定义有很大不同**。 因此，我们的研究可能会用嘈杂的标签丰富学习范式。

2.相关工作

在本节中，我们简要回顾了与这项工作相关的一些最新进展。

2.1. 多视图表示学习

通常，大多数现有的 MvRL 方法高度依赖于多视图数据的完整性和一致性假设。 如第 1 节所述，大多数多视图表示学习方法无法处理部分数据丢失问题 (PDP) 和部分视图对齐问题 (PVP)。从这个角度来看，现有的多视图学习作品可以分为三类。 即，vanilla 多视图学习方法 [2, 4, 32, 36, 37, 41, 42, 44, 47] 旨在利用不同视图的同质和互补信息来学习表示；不完整的多视图学习方法 [15,26,27,40] 利用完整的视图来预测缺失的视图；部分视图对齐表示学习方法 [18,21,39] 建立未对齐数据的对应关系，几乎所有现有研究都实现了实例级别的对应关系。

在上述研究中，[18,21,39] 与这项工作最相关。 与它们不同的是，我们在类别而不是实例级别执行对齐。 具体来说，在我们的研究中，两个任意的交叉视图样本被定义为对齐。

它们属于同一类别。 考虑到包括聚类和分类在内的下游任务，这种类别级对齐方案比实例级对齐更可取，后者在可访问性和可扩展性上更昂贵。 此外，基于度量学习 [39] 的方法在监督学习场景下实现对齐，比我们的设置更具挑战性，因为类别级别的对齐可以自然地从注释中导出。

2.2. 对比学习

对比学习是最近提出的无监督学习范式 [5,6,10,12,23,31]，已经在各种任务中达到了最先进的水平。 它们的主要区别在于使用的数据增强策略和对比损失。 简而言之，大多数对比学习方法首先通过一系列数据增强在实例级别构建正负对。 之后，不同的对比损失，如 Triplet [33]、NCE [11] 和 NT-Xent [6]，可用于最大化正对之间的相似性，同时最小化负对之间的相似性。

下面给出了这项工作与现有方法 [5, 6, 10-12, 31, 33] 的区别。 首先，本研究旨在处理多视图数据而不是单视图数据。 换句话说，这些对比学习方法不能直接用于处理多视图数据，尤其是在 PVP 发生时。 其次，我们不使用数据增强来构建数据对。 相反，我们直接使用可用的对齐数据作为正样本，并对观察到的数据执行随机抽样以构建负样本，这会导致嘈杂的标签问题。 第三，我们的方法具有新颖的对比损失，它对噪声标签具有鲁棒性。 据我们所知，到目前为止还没有涉及到带有噪声标签的对比学习。

2.3. 用嘈杂的标签学习

最近，已经进行了一些研究 [13,14,28,34,38] 使神经网络能够抵抗噪声标签，这引起了社区的兴趣。 一般来说，这些现有的工作旨在处理分类等监督任务的错误注释。 与它们不同的是，这项工作提出了视图对应可能是假的，并努力解决这种特殊的嘈杂标签问题。

3.方法

在本节中，我们提出了一种部分视图对齐的表示学习方法来解决 PVP，称为具有噪声鲁棒损失的多视图对比学习 (MvCLN)。 本节组织如下。 首先，第 3.1 节介绍了如何将对齐问题重新表述为类别级别的识别任务，该任务通过对比学习进一步实现。 第 3.2 节详细阐述了所提出的噪声鲁棒对比损失，以减轻甚至消除噪声对的影响，这对于无监督对构建是不可避免的。3.3 节从理论和实验的角度介绍了我们两阶段优化的必要性。最后，第 3.4 节介绍了我们模型的实现细节。

3.1. 问题表述

设$\{\mathbf{X}^i\}_{i=1}^v = \{\mathbf{x}_1^i, \mathbf{x}_2^i, \dots, \mathbf{x}_N^i\}_{i=1}^v$是一个部分视图对齐的数据集，即$\{\mathbf{X}^i\}_{i=1}^v = \{\boldsymbol{A}^i, \boldsymbol{U}^i\}_{i=1}^v$，其中 v 是指视图的数量，对齐和未对齐的数据分别表示为$\{\mathbf{A}^i\}_{i=1}^v = \{\mathbf{a}_1^i, \mathbf{a}_2^i, \dots, \mathbf{a}_N^i\}_{i=1}^v$和$\{\mathbf{U}^i\}_{i=1}^v = \{\mathbf{u}_1^i, \mathbf{u}_2^i, \dots, \mathbf{u}_N^i\}_{i=1}^v$。我们的目标是利用$\{\mathbf{A}^i\}_{i=1}^v$对齐$\{\mathbf{U}^i\}_{i=1}^v$，同时学习整个数据集的通用表示。

不失一般性，我们以 v=2 作为展示。 当x_k^1和x_k^2属于同一类别时，数据集在类别级别对齐，即，

$$C(\mathbf{x}_k^1) = C(\mathbf{x}_k^2), \forall k \in [1, N], \qquad (1)$$

其中 C(x)表示 x 的类别。 类别级对齐可以通过解决一个识别任务来实现，该任务旨在为x_k^1识别对应的x_k^2以满足上述目标。

为了完成识别任务，可以使用类别级对比学习 [12]，其目的是增加正对的相似性，同时最小化负对的相似性。然而，由于以下限制，不可能直接使用对比学习来执行识别任务。

一方面，我们的设置只包含正对$\{\mathbf{A}^i\}_{i=1}^{v}$，因此有必要从数据中构造负对。另一方面，如果没有标记数据的帮助，尽管使用了数据对构建方法，也不可避免地会得到一些嘈杂的负对。 因此，为了简单起见，我们建议使用随机抽样来生成负数。具体来说，我们从$\{\mathbf{A}^i\}_{i=1}^{v}$中随机选择两个样本$a_i^1$和$a_j^2$作为负对，其中 i≠j。直观地，当类别均匀分布时，构造的对有 1/K 的概率是有噪声的，其中 K 是类别号。 因此，我们的目标是使对比学习对嘈杂的标签（即假阴性）具有鲁棒性。

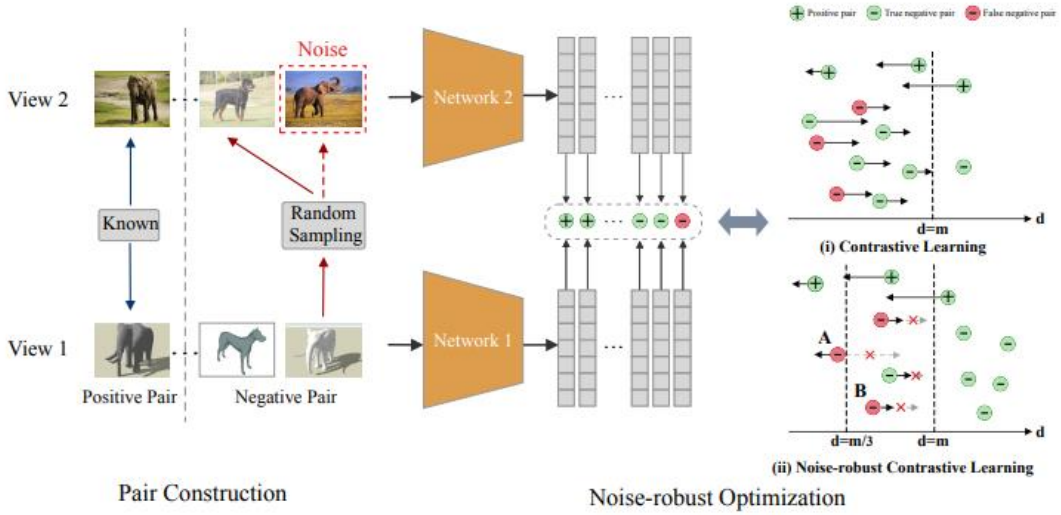

图 2. 为了从部分视图对齐的数据集中学习公共表示，建议的 MvCLN 包括对构建和噪声鲁棒优化。

为了获得数据对，MvCLN 将对齐的数据$\{\mathbf{A}^i\}_{i=1}^{2}$作为正对，将$\{\boldsymbol{A}^1\}$中的所有样本作为锚点。每个 anchor 在一起，MvCLN 从$\{\boldsymbol{A}^2\}$中随机选择 M 个样本形成 M 个负对。 在这样的随机抽样过程中，一些正对会被错误地视为负对，从而导致噪声标签问题。 基于我们的分析（见 3.3 节），MvCLN 通过采用两阶段优化策略解决了这个问题.。更具体地说，(i) 对比学习：它的目的是在数据自适应边际 m 上增加真负例的距离，以便最大限度地区分真假负例。 在(i)之后，大多数真负例的距离大于 m，一些负例的距离在(0, m/3)范围内，而其他负例的距离将落入(m/3, m)。 然后，我们切换到 (ii) 噪声鲁棒对比学习：它将通过降低梯度的幅度（见 B 点）甚至反转梯度的方向（见 A 点）来减轻假阴性的影响。 在（i）和（ii）中，箭头的方向和长度分别指的是损失梯度的方向和大小。

3.2. 抗噪对比损失

为了减轻甚至消除假阴性的影响，我们提出以下损失函数：

$$\mathcal{L} = \frac{1}{2N}\sum_{i=1}^{N}\left(P\mathcal{L}_i^{pos} + (1-P)\mathcal{L}_i^{neg}\right), \qquad (2)$$

其中 N 表示数据对的数量，P = 1/0 表示正/负对。 显然，L_i^{pos}将在对为正时起作用，而L_i^{neg}对负对起作用。

对于正交叉视图样本a_i^1和a_i^2，我们的目标是通过最小化它们在潜在空间中的距离

$$\mathcal{L}_i^{pos} = d(\mathbf{a}_i^1, \mathbf{a}_i^2), \qquad (3)$$

其中

$$d(\mathbf{a}_i^1, \mathbf{a}_i^2) = \left\| f_1(\mathbf{a}_i^1) - f_2(\mathbf{a}_i^2) \right\|_2^2, \qquad (4)$$

f_1和f_2分别表示两个参数化神经网络，将两个视图（数据对）投影到潜在空间中。

如果只最小化正对的距离，所有样本可能会崩溃到一个点。为了避免琐碎的解决方案，以下对比术语可能会有所帮助：

$$\mathcal{L}_i^{ctr} = \max(m - d(\mathbf{a}_i^1, \mathbf{a}_j^2), 0)^2, \qquad (5)$$

其中 m 是将底片的距离强制为适度大的余量，(a_i^1, a_j^2)表示负对。这就是著名的 SIAMESE 网络的损失[12]。

由于上述损失没有明确包含对噪声标签的鲁棒性，它会混淆真假阴性对，从而获得如图 3(c) 和 3(d) 所示的性能下降。 因此，为了享受对假阴性的鲁棒性，我们提出以下噪声鲁棒损失：

$$\mathcal{L}_i^{neg} = \frac{1}{m} \max(m d^{\frac{1}{2}}(\mathbf{a}_i^1, \mathbf{a}_j^2) - d^{\frac{3}{2}}(\mathbf{a}_i^1, \mathbf{a}_j^2), 0)^2 \qquad (6)$$

其中 m 是在初始状态下通过只计算一次：

$$m = \frac{1}{N_p} \sum d(\mathbf{a}_i^1, \mathbf{a}_i^2) + \frac{1}{N_n} \sum d(\mathbf{a}_i^1, \mathbf{a}_j^2), \qquad (7)$$

其中N_p和N_n分别是正数和负数。

由于公式的制定。 如图 6 所示，MvCLN 可以防止网络拟合假阴性甚至纠正错误的优化方向，如图 3(c) 和 3(d) 所示。下一节将进行详细分析，以从数学和实验的角度解释为什么我们的损失可以享受上述理想属性。

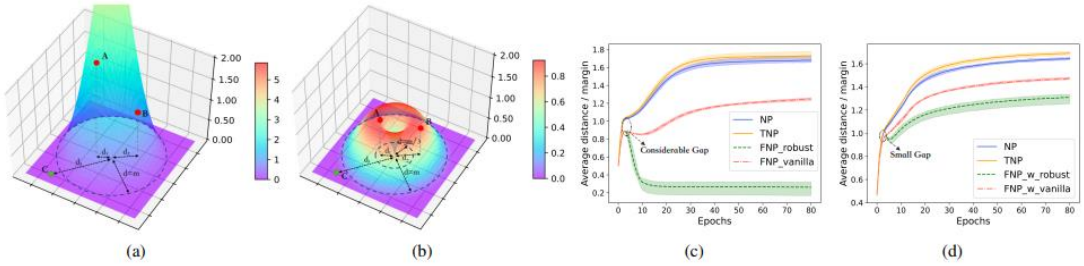

Figure 3.我们的抗噪对比损失的数学和实验分析。(a-b) 等式的损失值。5 和方程。6 w.r.t. 数据对的距离。为了更好地说明为什么我们的损失可以对嘈杂的标签具有鲁棒性，我们考虑了性能表面上的所有三种可能情况。 即 A、B、C 指的是距离 d1 < m/3 的假阴性对(FNP)，m/3 < d2 < m 的假阴性对(FNP)，距离 d3 >的真阴性对(TNP) 米。(a) 表明香草对比损失（方程 5）会增加包括 A、B 和 C 在内的所有底片的距离，无法处理嘈杂的标签。 相比之下，（b）表明我们的损失可以减少 A 的距离并缓慢增加 B 的距离，从而享受对嘈杂标签的鲁棒性。 (cd) 在 NoisyMNIST 和 Reuters 数据集上，随着 epoch 的增加，平均距离与边距的比率，其中 NP、FNP 稳健和 FNP vanilla 表示负对、由我们的损失优化的假负对（方程 6）和假负 通过 vanilla loss 优化的对（方程 5）。 彩色区域表示五个网络初始化的差异。可以观察到，随着更多的训练时期，我们的损失将显着扩大 TNP 和 FNP 之间的距离差距。 事实上，在 NoisyMNIST 上，我们的损失甚至可以纠正嘈杂标签的梯度方向，即 FNP 可以根据需要被视为真阳性。

3.3. 拟议损失分析

在本节中，我们进行数学和实验分析，以说明为什么提出的损失函数对噪声标签具有鲁棒性，以及为什么我们的模型采用两阶段优化策略。

让负数的距离 d 的L^{neg}的梯度为零，我们只需要考虑 d≤m，即

$$\frac{\partial \mathcal{L}^{neg}}{\partial d} = \frac{\partial(\frac{1}{m}d^3 - 2d^2 + md)}{\partial d}$$
$$= \frac{3}{m}d^2 - 4d + m, \qquad (8)$$

然后 d = m/3 或 d = m。 因此，性能表面将被划分为两个区域，即 0 < d < m/3 和 m/3 < d < m。

为了直观地说明上述理论结果，我们展示了损失面 w.r.t. 图中给定负对的距离。 3(a) - 3(b)。 从结果可以看出，与 vanilla loss（方程 5）相比，我们的噪声鲁棒损失（方程 6）的优化不会单调增加负对的距离，因此具有以下两个特点：

逆向优化（0 < d < m/3）：对于定位在空洞区域中的负对（例如参见 A），我们的损失梯度将被反转，因此负对的距离将减小。

慢优化（m/3 < d < m）：对于位于 m/3 < d < m 区域的对（例如参见 B），我们损失的优化速度会比 vanilla 慢 loss，因为梯度总是为负值，并且前者的梯度大于后者的梯度。 在数学上，

$$\Delta = \frac{\partial \mathcal{L}^{neg}}{\partial d} - \frac{\partial \mathcal{L}^{ctr}}{\partial d}$$
$$= \frac{\partial(\frac{1}{m}d^3 - 2d^2 + md)}{\partial d} - \frac{\partial(d^2 - 2dm + m^2)}{\partial d}$$
$$= \frac{\partial(d-m)^2}{\partial d} \geq 0. \qquad (9)$$

显然，如果假阴性被限制为 0 < d < m/3，则第一个特征可用于消除假阴性的影响。 或者，第二个可以用于通过将假阴性限制为 m/3 < d < m 来减轻假阴性的影响。 然而，如何区分假阴性和真阴性的问题在实践中是一项艰巨的任务。

幸运的是，Bengio 等人。 [3] 凭经验发现神经网络倾向于首先拟合简单的模式，这为我们提供了动力。 具体而言，我们建议将 TNP 视为简单模式，将 FNP 视为复杂模式。 因此，可以合理地推测具有普通对比损失的神经网络将比 FNP 更快地拟合 TNP，如图 3(c) 和 3(d) 所示。 更具体地说，这些数字表明，由于 TNP 的拟合速度更快，因此在早期训练阶段，TNP 和 FNP vanilla 之间存在差距。

由于上述观察，我们建议采用两阶段优化策略来区分 FNP 和 TNP。 简而言之，第一阶段将使用普通对比损失（方程 5）来优化我们的模型，直到所有负对的平均距离大于 m。 因此，由于 TNP 的拟合速度较快，大多数 TNP 和 FNP 将分别位于 d > m 和 d < m 的区域。 然后，我们的模型将切换到具有噪声鲁棒对比损失的第二个优化阶段，即等式。 6. 在此阶段，由于大多数 FNP 将位于 m/3 < d < m 或 0 < d < m/3，因此 FNP 的距离将缓慢增加（参见图 3（d）中的 FNP 稳健）或 减少（参见图 3（c）中的 FNP 稳健性），从而减轻甚至消除噪声标签的影响。 同时，它对真负对的影响可以忽略不计，因为到目前为止它们的大部分距离都大于 m。

3.4. 实验细节

在本节中，我们首先详细说明所提出的 MvCLN 的实现细节，然后展示如何在学习不同视图的通用表示的同时进行 MvCLN 进行类别级对齐。

如图 2 所示，MvCLN 首先将可用的对齐数据$\{\mathbf{A}^i\}_{i=1}^2$简单地视为正对并执行随机采样以

获得负对来构建数据对。 更具体地说，MvCLN 以$\{\mathbf{A}^1\}$中的每个样本为锚点，从$\{\mathbf{A}^2\}$中随机抽取 M 个样本，形成 M 个负对。 换句话说，正负比为 1/M。

获得数据对后，MvCLN 会将它们传递到两个维度为 D-1024-1024-1024-10 的神经网络（f_1和f_2）中，其中 D 是输入的维度。 所有层都紧密连接，然后是批量归一化 [20]、ReLU [29] 和 Dropout 层 [35]。

如第 3.3 节所述，我们的模型采用两阶段优化策略。 简而言之，MvCLN 使用带有 SGD 的 vanilla 对比损失（方程 5）进行优化，直到所有 NP 的平均距离达到边界 m。 之后，MvCLN 使用抗噪对比损失（Eq. 6）不断优化。

一旦我们的模型收敛，可以通过以下两个步骤实现类别级别的对齐：

- 步骤 1（距离计算）：获得表示$f_1(X^1)$和$f_2(X^2)$并计算它们的距离矩阵$D \in R^{N*N}$。在我们的实现中，我们简单地采用欧几里德距离。
- 步骤 2（对齐）：对于一个视图中的每个样本x_i^1，它在另一个视图中的对应关系是最小的D_{ij}。

在建立跨视图的对应关系后，我们为下游任务连接对齐数据的表示。

4. 实验

我们在四个广泛使用的多视图数据集上进行实验，并通过聚类和分类任务评估学习到的表示。 由于篇幅限制，我们在补充材料中展示了分类结果和更多细节。 为了验证我们的 MvCLN 在聚类中的有效性，我们使用 10 种最先进的多视图聚类方法作为基线，并使用 ACC、NMI 和 ARI 作为性能指标

4.1. 实验配置

我们在 PyTorch 1.5.0 中实施 MvCLN，并在配备 NVIDIA 2080Ti GPU 的标准 Ubuntu-16.04 操作系统上进行所有评估。为了优化 MvCLN，采用了初始学习率为 0.001 的 Adam 优化器 [22]，并且没有使用调度器或权重衰减。所有数据集的批量大小固定为 1024。

在实验中，使用了四个多视图数据集，即 Scene-15 [7, 9] 和 Caltech-101 [25, 45]，其中提取了两个图像特征作为视图，路透社 [1, 19] 使用前两种语言 （英语和法语）作为两个视图，NoisyMNIST [37] 随机选择了 30,000 个样本，因为基线无法处理原始的大规模数据集。 更多细节在补充材料中描述。除非另有说明，对于每个数据集$\{\mathbf{X}^v\}_{v=1}^2$，我们将其随机分成大小相等的两个分区，即$\{\mathbf{A}^v\}_{v=1}^2$和$\{\mathbf{U}^v\}_{v=1}^2$.在训练中，只有$\{\mathbf{A}^v\}_{v=1}^2$用作正对，负对通过执行第 3.4 节中详述的随机采样获得。

4.2. 与最先进技术的比较

在本节中，我们将提出的 MvCLN 与 10 种多视图聚类方法进行比较，包括 CCA [36]、KCCA [4]、DCCA [2]、DCCAE [37]、LMSC [43]、MvCDMF [46]、SwMC [30]、BMVC [45]、AE2-Nets [44] 和 PVC [18]。 对于所有基线，我们按照原始论文中的建议调整参数以实现最佳性能。对于我们的 MvCLN，我们将所有数据集的负/正比 M 固定为 30。 为了实现聚类，除了 Mvc-DMF、SwMC 和 BMVC 之外，对所有测试方法学习的表示进行了 k-means。

由于只有 PVC 和我们的 MvCLN 可以解决 PVP，为了公平比较，其他测试方法的结果在以下两种设置下报告：

- 设置 1（部分视图对齐数据）：我们首先使用 PCA 将原始数据投影到一个与 MvCLN 维度相同的潜在空间中，以便可以应用匈牙利算法来建立部分视图对齐数据的对应关系。 之后，我们对重新调整的数据进行这些基线。 对于 PVC 和 MvCLN，我们只需在部分视图对齐的数据上运行它们。
- 设置 2（Full View-aligned Data）：我们在完全对齐的原始数据上直接运行除 PVC 和 MvC IN 之外的所有方法

为了避免由于随机性导致的性能变化，我们将所有方法运行五次，并根据三个性能指标（即 ACC、NMI 和 ARI）报告平均性能。 请注意，由于硬件和软件环境的差异，我们实验中某些基线的结果与 [18] 中报告的结果略有不同。 从表 1 可以看出：

Aligned	Methods	Scene-15			Caltech-101			Reuters			NoisyMNIST		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Fully	CCA (NeurIPS'03)	36.37	36.91	19.82	20.25	45.41	16.34	44.31	20.34	14.52	71.31	52.60	48.46
	KCCA (JMLR'02)	37.93	37.42	21.38	21.45	45.58	17.62	**50.87**	22.34	**20.61**	**96.85**	**92.10**	**93.23**
	DCCA (ICML'13)	36.61	39.20	21.03	27.60	47.84	**30.86**	47.95	**26.57**	12.71	89.64	88.33	83.95
	DCCAE (ICML'15)	34.58	39.01	19.65	19.84	45.05	14.57	41.98	20.30	8.51	78.00	81.24	68.15
	LMSC (CVPR'17)	38.46	35.50	20.54	26.87	48.80	18.06	38.56	20.12	15.48	–	–	–
	MvC-DMF (AAAI'17)	30.99	31.35	15.68	24.35	44.98	14.82	33.83	14.89	12.59	74.39	63.22	49.79
	SwMC (IJCAI'17)	33.89	32.98	11.78	**30.74**	36.07	7.75	33.65	16.02	5.90	–	–	–
	BMVC (TPAMI'18)	**40.74**	**41.67**	**24.19**	27.59	46.43	21.28	42.39	21.86	15.14	88.31	77.01	76.58
	AE^2-Nets (CVPR'19)	37.17	40.47	22.24	20.79	45.01	15.89	42.39	19.76	14.87	42.11	43.38	30.42
Partially	CCA (NeurIPS'03)	32.73	34.24	18.80	20.06	41.56	**16.62**	40.87	**15.82**	12.68	34.46	29.83	17.89
	KCCA (JMLR'02)	33.09	31.43	16.35	12.57	31.36	7.65	40.08	11.80	11.27	26.57	18.19	10.55
	DCCA (ICML'13)	34.27	36.55	18.83	12.52	32.13	7.63	39.71	13.83	**14.38**	29.22	20.24	11.08
	DCCAE (ICML'15)	33.62	36.56	18.54	11.75	30.54	6.60	**41.42**	12.82	13.61	27.61	19.45	10.00
	LMSC (CVPR'17)	26.27	20.45	10.93	**21.54**	**40.26**	15.51	32.17	11.34	7.19	–	–	–
	MvC-DMF (AAAI'17)	28.49	24.31	11.22	9.54	23.41	3.84	32.58	12.36	11.08	27.34	22.96	6.85
	SwMC (IJCAI'17)	31.03	30.39	12.94	19.03	22.75	3.73	31.92	11.03	5.40	–	–	–
	BMVC (TPAMI'18)	**36.81**	**36.55**	**20.20**	12.13	31.33	7.11	38.15	11.57	12.07	28.47	24.69	14.19
	AE^2-Nets (CVPR'19)	28.56	26.58	12.96	10.45	29.51	7.90	35.49	10.61	8.07	**38.25**	**34.32**	**22.02**
Partially	PVC (NeurIPS'20)	37.88	39.12	20.63	22.11	**47.82**	17.98	42.07	20.43	16.95	81.84	82.29	82.03
	MvCLN (Mean)	**38.53**	**39.90**	**24.26**	**30.09**	43.07	**38.34**	**50.16**	**30.65**	**24.90**	**91.05**	**84.15**	**83.56**
	MvCLN (Best)	39.87	40.47	24.83	35.72	45.25	51.44	56.62	33.62	27.37	94.51	86.77	88.42

Table 1. 四个广泛使用的多视图数据集的聚类比较，其中每个设置的最佳结果以粗体显示，“-”表示该方法由于时间或内存成本过高而无法获得结果

- 在第一个设置中，我们的 MvCLN 显着优于所有测试方法。 特别是，与最佳基线相比，MvCLN 在 Caltech-101 和路透社上分别实现了 113.2% 和 46.9% 的 ARI 改进。 这验证了我们的主张和动机，即类别级别的对齐比实例级别的对齐更可取；
- 在第二种设置中，尽管 MvCLN 是在部分视图对齐的数据上进行的，而基线是在完全视图对齐的数据上进行的，但它仍然取得了有竞争力的结果。

4.3. 消融研究和参数分析

在本节中，我们对 NoisyMNIST 进行以下实验分析，即消融研究、正负比的影响、对齐比例的影响以及两个优化阶段之间切换时间的影响。

除了使用的聚类性能指标外，我们还引入了一个称为类别级别对齐率 (CAR) 的指标来衡量类别级别对齐的比率。 数学上，

$$CAR = \frac{\sum_{i=1}^{N} \delta(C(\hat{\mathbf{x}}_i^1), C(\hat{\mathbf{x}}_i^2))}{N}, \qquad (10)$$

其中$(\hat{\mathbf{x}}_i^1, \hat{\mathbf{x}}_i^2)$表示重新排列的对，$\delta$ 是狄利克雷函数，N 是数据对的数量。

噪声鲁棒对比损失的有效性：为了证明所提出的噪声鲁棒对比损失的有效性，我们将其替换为普通对比损失，即等式 5. 如表 2 所示，虽然 vanilla contrastive loss 也可以取得一些有希望的结果，但明显比 MvCNL 差。

$\mathcal{L}_{neg}$	ACC	NMI	ARI	CAR
✗	88.2	75.73	75.89	84.48
✓	**92.17**	**85.57**	**84.51**	**88.24**

Table 2. NoisyMNIST 的消融研究。 “√”表示带组件的 Mv CLN，“×”表示不带组件的 MvCLN

正/负对比率的影响：我们的方法使用对齐的数据作为正数，随机选择的数据作为负数。

换句话说，很容易得到底片。 然而，过高的负/正对比率 (M) 会导致数据分布不平衡。 因此，直观地说，确定 M 的值是至关重要的。 在图 4(a) 中，我们通过以 5 的间隔将 M 从 1 增加到 50 来研究 MvCLN 的性能。根据结果，可以有以下观察结果。 一方面，增加 M 会在相当大的值范围内提高 MvCLN 的性能。 另一方面，当 M 范围为 [20, 40] 时，MvCLN 表现稳定，这表明其对参数的鲁棒性。

不同对齐比例的影响：为了研究 MvCLN 在不同对齐比例的数据上的性能，我们将对齐比例从 10% 增加到 100%，间隔为 10%。 从结果中，可以观察到：i) 有更多可用的对齐数据，MvCLN 取得了更好的结果； ii) 当对齐比例从 70% 增加到 100% 时，MvCLN 的性能略有提高。 可能的原因是 70% 的对齐数据足以让 MvCLN 学习对齐模式。

两个优化阶段之间切换时间的影响：我们的方法由两个优化阶段组成，它们以数据驱动的方式自动切换，如第 3.4 节所述。 在本节中，我们通过实验研究了以下七个切换标准的影响，即当负对的平均距离达到 0.0m、0.2m、0.4m、0.6m、0.8m、1.0m 和 1.2 时切换到 Stage 2 m，其中边距 m 是根据数据自动确定的。 如图 4(c) 所示，MvCLN 在 [0.2m, 1.0m] 范围内将获得稳定的结果。 如果没有第 1 阶段（即 0.0m），MvCLN 将获得较差的结果，这与图 3（b）中的分析一致。 当切换时间太晚（1.2m）时，大多数假阴性对的距离可能接近甚至超过 m，从而导致假阴性对和真阴性对混合。

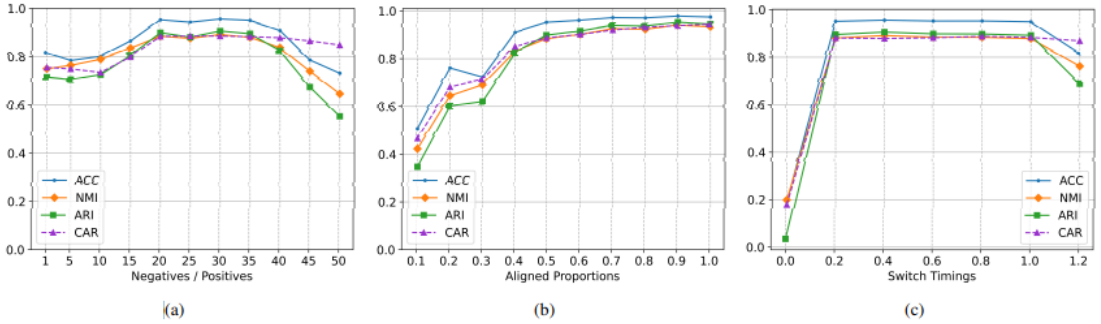

Figure 4. NoisyMNIST 数据集的性能分析。 (a) 负/正比 (M) 的表现； (b) 具有不同对齐比例的性能；(c) 两个优化阶段不同切换时间的性能

5. 结论

在本文中，我们提出了 MvCLN，它通过赋予对比学习对噪声标签的鲁棒性来处理部分视图对齐问题。与现有解决方案不同，我们的 MvCLN 旨在实现类别而不是实例级对齐。大量实验验证了我们学习范式的有效性和效率。此外，我们从理论上和实验上展示了为什么我们的模型可以对嘈杂的标签具有鲁棒性。据我们所知，所提出的方法可以被视为第一个使对比学习对噪声标签（即假阴性对应对）具有鲁棒性的研究。更重要的是，这项工作可能会通过将视图对应视为特殊的噪声标签问题来显着丰富噪声标签的学习范式。未来，我们计划为正负对都被噪声污染的情况探索一种更通用的解决方案。这种解决方案对各种应用都很有价值，包括但不限于 ReID、对象跟踪、人脸识别、图形匹配、图像翻译和恢复。

6. 致谢

这项工作得到了中国国家重点研发计划项目 2020YFB1406702 和 2020AAA0104500 的部分支持； 部分由 YJ201949 中央高校基本科研业务费资助； 部分由 NFSC 根据 Grant U19A2081、61625204、61836006、U19A2078 提供； 部分由四川大学明日生活基金资助；部分来自北京市科学技术委员会的北京新星计划 (Z201100006820068)。

参考文献

[1] Massih-Reza Amini, Nicolas Usunier, and Cyril Goutte. Learning from multiple partially observed views - an appli cation to multilingual text categorization. In NeruIPS, pages 28–36, 2009. 6
[2] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In ICML, pages 1247–1255, 2013. 1, 2, 6
[3] Devansh Arpit, Stanislaw K Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron C Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In ICML, pages 233–242, 2017. 5
[4] Francis R Bach and Michael I Jordan. Kernel independent component analysis. Journal of Machine Learning Research, 3(Jul):1–48, 2002. 2, 6
[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. arXiv:2006.09882, 2020. 3
[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. arXiv:2002.05709, 2020. 3
[7] Dengxin Dai and Luc Van Gool. Ensemble projection for semi-supervised image classification. In ICCV, pages 2072– 2079, 2013. 6
[8] Cheng Deng, Zhaojia Chen, Xianglong Liu, Xinbo Gao, and Dacheng Tao. Triplet-based deep hashing network for crossmodal retrieval. IEEE Transactions on Image Processing, 27(8):3893–3903, 2018. 2
[9] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In CVPR, volume 2, pages 524–531, 2005. 6
[10] Jean-Bastien Grill, Florian Strub, Florent Altche, Corentin ´ Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. arXiv:2006.07733, 2020. 3
[11] Michael Gutmann and Aapo Hyvarinen. Noise-contrastive ¨ estimation: A new estimation principle for unnormalized statistical models. In AISTATS, pages 297–304, 2010. 3
[12] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In CVPR, volume 2, pages 1735–1742, 2006. 3, 4
[13] Bo Han, Jiangchao Yao, Gang Niu, Mingyuan Zhou, Ivor Tsang, Ya Zhang, and Masashi Sugiyama. Masking: A new perspective of noisy supervision. arXiv:1805.08193, 2018. 3
[14] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Coteaching: Robust training of deep neural networks with extremely noisy labels. arXiv:1804.06872, 2018. 3
[15] Menglei Hu and Songcan Chen. Doubly aligned incomplete multi-view clustering. arXiv:1903.02785, 2019. 2
[16] Peng Hu, Dezhong Peng, Yongsheng Sang, and Yong Xiang. Multi-view linear discriminant analysis network. IEEE Transactions on Image Processing, 28(11):5352–5365, 2019. 1
[17] Peng Hu, Xi Peng, Hongyuan Zhu, Jie Lin, Liangli Zhen, and Dezhong Peng. Joint versus independent multiview hashing for cross-view retrieval. IEEE Transactions on Cybernetics, 2020. 2
[18] Zhenyu Huang, Peng Hu, Joey Tianyi Zhou, Jiancheng Lv, and Xi Peng. Partially view-aligned clustering. NeurIPS, 33, 2020. 2, 3, 6, 7
[19] Zhenyu Huang, Joey Tianyi Zhou, Xi Peng, Changqing Zhang, Hongyuan Zhu, and Jiancheng Lv.

Multi-view spectral clustering network. In IJCAI, pages 2563–2569, 2019. 6
[20] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167, 2015. 6
[21] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In CVPR, pages 3128–3137, 2015. 2, 3
[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv:1412.6980, 2014. 6
[23] Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In AAAI, 2021. 3
[24] Yingming Li, Ming Yang, and Zhongfei Zhang. A survey of multi-view representation learning. IEEE Transactions on Knowledge and Data Engineering, 31(10):1863–1883, 2018. 1
[25] F Li Fei-Fei, M Andreetto, MA Ranzato, and P Perona. Caltech101. Computational Vision Group, California Institute of Technology, 2003. 6
[26] Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, and Xi Peng. Completer: Incomplete multi-view clustering via contrastive prediction. In CVPR, 2021. 2
[27] Xinwang Liu, Miaomiao Li, Chang Tang, Jingyuan Xia, Jian Xiong, Li Liu, Marius Kloft, and En Zhu. Efficient and effective regularized incomplete multi-view clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020. 2
[28] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In ICML, pages 6543– 6553, 2020. 3
[29] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In ICML, pages 807– 814, 2010. 6
[30] Feiping Nie, Jing Li, Xuelong Li, et al. Self-weighted multi view clustering with multiple graphs. In IJCAI, pages 2564– 2570, 2017. 6
[31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. arXiv:1807.03748, 2018. 3
[32] Xi Peng, Zhenyu Huang, Jiancheng Lv, Hongyuan Zhu, and Joey Tianyi Zhou. Comic: Multi-view clustering without parameter selection. In ICML, pages 5092–5101, 2019. 2
[33] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In CVPR, pages 815–823, 2015. 3
[34] Hwanjun Song, Minseok Kim, Dongmin Park, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. arXiv preprint arXiv:2007.08199, 2020. 3
[35] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1):1929–1958, 2014. 6
[36] Alexei Vinokourov, Nello Cristianini, and John ShaweTaylor. Inferring a semantic representation of text via crosslanguage correlation analysis. In NeurIPS, pages 1497– 1504, 2003. 2, 6
[37] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In ICML, pages 1083–1092, 2015. 1, 2, 6
[38] Xiaobo Wang, Shuo Wang, Jun Wang, Hailin Shi, and Tao Mei. Co-mining: Deep face recognition with noisy labels. In CVPR, pages 9358–9367, 2019. 3
[39] Jiwei Wei, Xing Xu, Yang Yang, Yanli Ji, Zheng Wang, and Heng Tao Shen. Universal weighting metric learning for cross-modal matching. In CVPR, pages 13005–13014, 2020. 2, 3

[40] Chang Xu, Dacheng Tao, and Chao Xu. Multi-view learning with incomplete views. IEEE Transactions on Image Processing, 24(12):5812–5825, 2015. 2
[41] Ming Yin, Weitian Huang, and Junbin Gao. Shared generative latent representation learning for multi-view clustering. In AAAI, pages 6688–6695, 2020. 2
[42] Ming Yin, Wei Liu, Mingsuo Li, Taisong Jin, and Rongrong Ji. Cauchy loss induced block diagonal representation for robust multi-view subspace clustering. Neurocomputing, 427:84–95, 2021. 2
[43] Changqing Zhang, Qinghua Hu, Huazhu Fu, Pengfei Zhu, and Xiaochun Cao. Latent multi-view subspace clustering. In CVPR, pages 4279–4287, 2017. 6
[44] Changqing Zhang, Yeqing Liu, and Huazhu Fu. Ae2-nets: Autoencoder in autoencoder networks. In CVPR, pages 2577–2585, 2019. 1, 2, 6
[45] Zheng Zhang, Li Liu, Fumin Shen, Heng Tao Shen, and Ling Shao. Binary multi-view clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(7):1774– 1782, 2018. 6
[46] Handong Zhao and Zhengming Ding. Multi-view clustering via deep matrix factorization. In AAAI, pages 2921–2927, 2017. 6
[47] Tao Zhou, Changqing Zhang, Xi Peng, Harish Bhaskar, and Jie Yang. Dual shared-specific multiview subspace clustering. IEEE Transactions on Cybernetics, 50(8):3517–3530, 2019. 2