SubTab: 用于自监督表示学习的表格数据子集化方法 - 详细解析与重点归纳

总体重点归纳

1. 核心创新：**SubTab** 将表格数据分割为多个可能重叠的特征子集，从子集重建完整数据，创造多视角学习机会。
2. 关键机制：
    - 从特征子集重建完整数据（而非从损坏版本重建）
    - 协作推理：测试时聚合各子集的表示
    - 三种损失函数组合：重建损失、对比损失和距离损失
3. 技术优势：
    - 无需预测噪声掩码（避免不平衡分类问题）
    - 能够处理缺失特征（训练和测试时）
    - 识别信息丰富的特征区域
    - 可使用更小模型减少过拟合
4. 实验证明：
    - 在表格格式 **MNIST** 上达到 **98.31%**准确率，与 **CNN** 模型相当
    - 在 **TCGA** 数据集上性能几乎是基线的两倍
    - 在 **Obesity** 数据集上超越所有监督和非监督方法
    - 证明了协作推理的有效性和对缺失特征的鲁棒性
5. 应用潜力：
    - 医疗健康数据（处理不完整特征）
    - 金融和法律数据分析
    - 跨数据集知识迁移
    - 分布式学习环境
6. 未来方向：

- 扩展到排列不变架构
- 开发分层版本识别个别重要特征
- 应用于其他数据类型（图像、音频、文本）
- 探索特征重要性分析

SubTab 的核心贡献是将自监督学习的成功从图像、音频和文本领域扩展到表格数据，通过子集化特征解决了表格数据缺乏固有结构的挑战，其灵活的框架为处理表格数据的实际问题（如缺失特征、高维度特征）提供了优雅解决方案。

Abstract（摘要）翻译与解析

翻译： 自监督学习已被证明在学习有用的表示方面非常有效，然而大部分成功是在图像、音频和文本等数据类型上取得的。这种成功主要通过利用数据中的空间、时间或语义结构进行增强来实现的。然而，这种结构可能在医疗保健等领域常用的表格数据集中不存在，这使得设计有效的增强方法变得困难，阻碍了在表格数据环境中取得类似的进展。在本文中，我们引入了一个新框架，表格数据的子集化（SubTab），通过将输入特征划分为多个子集，将从表格数据学习的任务转变为多视角表示学习问题。我们认为，在自编码器设置中，从特征子集而非其损坏版本重建数据可以更好地捕捉其潜在表示。在这个框架中，联合表示可以在测试时表示为子集潜在变量的聚合，我们称之为协作推理。我们的实验表明，SubTab 在表格设置的 MNIST 上达到了 98.31%的最先进（SOTA）性能，与基于 CNN 的 SOTA 模型相当，并在其他三个真实世界数据集上大幅超越现有基线。

解析与重点：

- 自监督学习在图像、音频和文本领域成功的关键在于这些数据固有的空间、时间或语义结构
- 表格数据缺乏这些结构，导致难以设计有效的增强方法
- SubTab 框架将表格数据分割成多个子集，创造多视角学习问题
- 核心思想：从特征子集重建完整数据，而非从损坏版本重建原始数据
- 协作推理：在测试时聚合各子集的潜在表示

- 性能亮点：在表格格式的 MNIST 上达到 98.31%准确率，与 CNN-SOTA 模型相当
- 在三个真实世界数据集上明显优于现有方法

1 Introduction（引言）翻译与解析

翻译： 近年来，自监督学习已成功用于在自然语言处理中学习数据的有意义表示。图像和音频领域也取得了类似的成功。这一进展主要通过利用数据通过数据增强的空间、语义或时间结构，生成前置任务和通过架构选择使用归纳偏置（例如用于图像的 CNN）来实现的。然而，在医疗保健、广告、金融和法律等许多领域常用的表格数据中缺乏此类结构和偏置时，这些方法可能不那么有效。一些增强方法，如裁剪、旋转、颜色变换等是特定领域的，不适用于表格设置。难以设计类似有效的针对表格数据的方法是为什么自监督学习在这个领域研究不足的原因之一。

表格数据中最常见的方法是通过添加噪声破坏数据。自编码器将数据的损坏示例映射到潜在空间，然后从那里映射回未损坏的数据。通过这个过程，它学习对输入中的噪声具有鲁棒性的表示。这种方法可能不太有效，因为它平等对待所有特征，就好像特征都同样具有信息量。然而，扰动无信息特征可能不会达到破坏的预期目标。最近的一项工作通过引入前置任务，在表格数据设置中利用自监督学习，其中带有连接到表示层的分类器的去噪自编码器在损坏的数据上进行训练。分类器的任务是预测损坏特征的位置。然而，这个框架仍然依赖于输入中的嘈杂数据。此外，对于高维数据，在不平衡的二元掩码上训练分类器可能不是学习有意义表示的理想方法。

在这项工作中，我们通过将特征划分为子集，类似于图像领域中的裁剪或集成学习中的特征装袋，将从单一视角学习表示的问题转变为从多个视角学习的问题，以生成数据的不同视角。每个子集可以被视为不同的视角。我们表明，从特征子集重建数据迫使编码器学习比通过现有方法（如添加噪声）学习到的表示更好的表示。我们在自监督设置中训练我们的模型，并在下游任务（如分类和聚类）上评估它。我们使用五个不同的数据集：表格格式的 MNIST、癌症基因组图谱（TCGA）、肥胖队列的人类肠道元基因组样本（Obesity）、UCI 成人收入（Income）和 UCI BlogFeedback（Blog）。

SubTab 可以：i）通过使用子集表示的聚合构建更好的表示，这个过程我们称为协作推理 ii）通过测量每个子集的预测能力发现信息丰富的特征区域，这在高

维数据中特别有用 iii）通过忽略相应的子集在训练和推理过程中处理缺失特征 iv）通过减少输入维度使用更小的模型，使其不太容易过拟合。

解析与重点：

- 自监督学习在 NLP、图像和音频领域成功的关键：数据的固有结构特性和增强技术
- 表格数据的挑战：缺乏空间、时间或语义结构，导致传统增强方法不适用
- 现有方法的局限性：
    1. 通过添加噪声来损坏数据 - 但平等对待所有特征不合理
    2. 使用前置任务（如 VIME-self）- 但仍依赖嘈杂输入和在不平衡掩码上训练分类器
- SubTab 的核心思想：将数据分割为多个子集，从而创建多视角学习问题
- 作者使用了 5 个数据集进行验证：MNIST（表格格式）、TCGA、Obesity、Income 和 Blog
- SubTab 的四大优势：
    1. 通过协作推理构建更好的表示
    2. 发现信息丰富的特征区域
    3. 处理缺失特征
    4. 使用更小的模型减少过拟合

2 Method（方法）翻译与解析

翻译： 图像领域中常用的增强方法包括添加噪声、旋转、裁剪等。其中，裁剪被证明是最有效的技术。受此启发，我们提出了表格数据特征子集化。

图 1 展示了 SubTab 框架，其中我们有一个编码器（E）、一个解码器（D）和一个可选的投影（G）。在本文中，我们将 h 称为潜在表示，z 称为投影，$\tilde{x}$ 和 $\tilde{X}$ 分别指子集和整个数据的重建。小写字母与子集相关，而大写字母与整个特征集相关。此外，在本文中，当我们说表示"好"时，我们指的是它在使用线性模型的分类任务中的表现。

在 SubTab 框架中，我们将表格数据划分为多个子集。相邻的子集可以有重叠区域，定义为子集维度的百分比。每个子集都被输入到相同的编码器（即参数共享）以获得对应的潜在表示。共享解码器用于重建输入到编码器的子集，或完整的表格数据（即从特征子集重建所有特征）。在我们的实验中，我们选择了后者，因为它在学习好的表示方面更有效。我们还应该注意到，在后一种情况下，自编码器不能学习恒等映射，消除了对瓶颈（即表示）维度的约束。我们为每个子集计算一个重建损失项。

此外，我们可以通过使用所有子集投影的所有组合对选择性地向我们的目标添加对比损失。例如，如图 1 所示，给定三个子集，有三种两两组合：$n/k = 3/2 = 3!/(2!(1)!) = 3$。对于四个子集，将有 6 对组合，依此类推。我们可以添加另一个损失项，称为距离损失，通过使用均方误差（MSE）等损失函数来减少子集投影对之间的距离。所有三个损失项都对正样本施加拉力，而对比损失还对正负样本之间施加推力，如图 2a 所示。

一旦数据在数据准备步骤中被划分为子集，这个过程类似于集成学习中的特征装袋，它们的位置是固定的。因此，我们在训练过程中不改变子集中特征的相对顺序，因为标准神经网络架构不是排列不变的。这是为了确保相同的特征被馈送到神经网络的相同输入单元。然而，我们的方法可以扩展到排列不变设置，作为下一步。

2.1 添加噪声的策略

我们的框架与表格数据设置中使用的其他增强技术互补。因此，我们通过使用三种类型的噪声在每个子集中随机选择的条目上添加噪声进行了实验：i）添加高斯噪声，$N(0, \sigma^2)$，ii）用从同一列随机采样的另一个值覆盖所选条目的值，称为交换噪声，iii）将随机选择的条目归零，称为置零噪声。

此外，我们在选择要添加噪声的特征时使用三种不同的策略，如图 2b 所示：i）随机块的相邻列（NC），ii）随机列（RC）iii）每个样本的随机特征（RF）。为了添加噪声，我们创建一个二项掩码 m 和一个与子集形状相同的噪声矩阵 ε，其中掩码的条目以概率 p 被分配为 1，否则为 0。子集 x_1的损坏版本 x_1c 生成如下：

$$x_1c = (1 - m) \odot x_1 + m \odot \varepsilon \quad (1)$$

2.2 训练

我们的目标函数是：

$Lt = Lr + Lc + Ld$, (2)

其中 Lt、Lr、Lc 和 Ld 分别是总损失、重建损失、对比损失和距离损失。

i) 重建损失：给定一个子集，表示为 xk，我们可以重建相同的子集 x̃k 或整个特征空间 X̃k。然后，我们可以通过计算使用(xk, x̃k)或(X, X̃k)对的均方误差来计算第 k 个子集的重建损失，如图 1 所示。我们选择后者，因为它更有效。总体重建损失：

$Lr = (1/K) \sum_{k=1}^{K} sk$, 其中 $sk = (1/N) \sum_{i=1}^{N} ||X(i) - \tilde{X}k(i)||^2$ (3)

其中 K 是子集的总数，N 是批次的大小，sk 是第 k 个子集的重建损失，Lr 是所有子集上重建损失的平均值。

ii) 对比损失：如果数据集在类别数量上丰富，使得采样负样本的可能性很高，我们可以使用投影网络（G）获取表示 h 的投影 z。两个子集中相同行的样本 z1 和 z2 可以被视为正对，而子集中的其余行可以被视为对这些样本的负样本。这允许我们使用归一化的温度尺度交叉熵损失（NT-Xent）等损失函数计算每对投影的对比损失。对于三个子集{x1, x2, x3}，我们可以为集合 S = {{z1, z2}, {z1, z3}, {z2, z3}}中总共三对中的每一对{za, zb}计算这样的损失。总体对比损失是：

$Lc = (1/J) \sum_{\{za,zb\} \in S} p(za, zb)$, 其中 $p(za, zb) = (1/2N) \sum_{i=1}^{N} [l(za(i), zb(i)) + l(zb(i), za(i))]$ (4)

$l(za(i), zb(i)) = -\log(\exp(sim(za(i), zb(i))/\tau) / \sum_{k=1}^{N} 1[k \neq i] \exp(sim(za(i), zb(k))/\tau))$ (5)

其中 J 是集合 S 中的总对数，p(za, zb)是投影对{za, zb}的总对比损失，l(za(i), zb(i))是子集{za, zb}中相应正样本对(za(i), zb(i))的损失函数，Lc 是所有对上对比损失的平均值。

iii) 距离损失：我们还可以为子集的投影对添加均方误差（MSE）损失，因为子集中对应的样本应彼此接近。因此，我们可以计算整体 MSE 损失为：

$Ld = (1/J) \sum_{\{za,zb\} \in S} p(za, zb)$, 其中 $p(za, zb) = (1/N) \sum_{i=1}^{N} ||za(i) - zb(i)||^2$ (6)

算法的伪代码可以在附录的算法 1 中找到。我们应该注意，方程(2)中的 Lc 和 Ld 是可选的，我们只在一些实验中使用它们。

2.3 测试时间

在测试时，我们将测试集的子集输入到编码器，并获取所有可用子集表示的聚合，如图 2c 所示。请注意，我们可以使用均值、和、最小值、最大值或任何其他聚合方法来获取联合表示，这类似于计算机视觉中的池化，或图卷积网络中相邻节点的聚合。我们在所有实验中使用了均值聚合，但在附录 F.4 中比较了不同的聚合方法。我们的实验表明，我们可以只使用一个或几个子集的表示，仍然在测试时获得良好的性能。例如，在图 2c 中，我们可以只使用 h1，或 h1 和 h2 的聚合，而不是聚合所有子集（h1, h2, h3）。这允许模型即使在存在缺失特征的情况下也能从数据中推断，在这种情况下，我们可以忽略具有缺失特征的子集。我们还可以设计一个聚合函数，计算子集表示的加权平均值，因为某些子集可能比其他子集更具信息量：

$$h = (1/Z) \sum_{k=1 \text{ to } K} \eta_k * h_k, \text{ 和 } Z = \sum_{k=1 \text{ to } K} \eta_k, \quad (7)$$

其中 K 是子集的数量，ηk 是第 k 个子集的权重。η 可以是半监督或监督设置中的可学习参数，通过使用注意力机制。我们也可以在方程(7)中使用 1D 卷积，通过在训练期间将子集的表示视为单独的通道。我们将这些想法作为未来的工作，并在我们的实验中使用了均值聚合（即 ηk = 1），除非明确说明。不同聚合方法的比较可以在附录中的表 A3 中找到。

解析与重点：

- SubTab 的灵感来源：图像领域中最有效的增强技术——裁剪
- 框架组成：编码器(E)、解码器(D)和可选的投影(G)
- 关键机制：将表格数据分割为可能重叠的子集
- 训练过程：
    1. 将每个子集输入到共享编码器获取潜在表示
    2. 使用共享解码器从子集表示重建完整数据（而非仅重建子集）
    3. 可选：通过所有子集投影对计算对比损失
    4. 可选：添加距离损失以减少子集投影间的距离

- 噪声策略：三种噪声类型（高斯、交换、置零）和三种特征选择策略
- 损失函数：总损失 = 重建损失 + 对比损失 + 距离损失
- 测试时的协作推理：聚合所有子集的表示（默认使用均值聚合）
- 主要优势：能够处理缺失特征、识别信息丰富的特征子集

3 Experiments（实验）翻译与解析

翻译： 我们在各种表格数据集上进行了各种实验，包括表格格式的 MNIST、癌症基因组图谱（TCGA）、肥胖队列的人类肠道元基因组样本（Obesity）、UCI 成人收入（Income）和 UCI BlogFeedback（Blog），以展示 SubTab 框架的有效性。我们将我们的方法与带和不带 dropout 的自编码器基线、其他自监督方法如 VIME-self、去噪自编码器（DAE）和上下文编码器（CAE）以及完全监督模型如逻辑回归、随机森林和 XGBoost 进行比较。对于每个数据集，一旦我们决定了特定的自编码器架构，我们就将其用于所有比较的模型（即 VIME-self、DAE、CAE 和我们的模型）。我们尝试了 ReLU 和 leakyReLU 作为所有模型的激活函数，两者表现同样好。SubTab 的代码已提供。模型架构和超参数的摘要在附录的表 A1 中。我们应该注意到，我们在附录 G 和 H 中分别使用 i）合成数据集和 ii）OpenML-CC18 数据集进行了更多实验。

3.1 数据

MNIST：我们将 28x28 的图像展平，并通过除以 255 进行缩放，就像在[46]中所做的那样。在搜索超参数时，我们将训练集分为训练和验证集（90-10%分割），然后使用所有训练集训练最终模型。测试集仅用于最终评估。

癌症基因组图谱（TCGA）：TCGA 是一个公共癌症基因组数据集，包含超过 20,000 个原发癌症和匹配的正常样本，持有超过 38 个队列的信息。任务是从反向相蛋白阵列（RPPA）数据集中分类癌症队列。它包括 6671 个样本，具有 122 个特征，我们将其分为 80-10-10%的训练-验证-测试集。一旦找到超参数，我们在训练和验证集的组合上训练模型。

Obesity：该数据集由公开可用的肥胖队列人类肠道元基因组样本组成。它源自全基因组散弹枪元基因组学研究。数据集包括 164 名肥胖患者和 89 名非肥胖对照，具有 425 个特征。我们使用最小-最大缩放来缩放数据集。由于它是一个小数据集，我们使用 10 个随机抽取的训练-测试（90-10%）分割评估模型，对于每个分割，我们使用 10 折交叉验证。

UCI 成人收入：这是一个著名的公共数据集，从 1994 年人口普查数据库中提取。它包括教育水平和人口统计等详细信息，用于预测一个人的收入是否超过 $50K/年。数据包括六个连续和八个分类特征。在对分类特征进行一热编码后，总共有 101 个特征。预处理步骤可以在附录的 B.1 节中找到。

UCI BlogFeedback：数据源自博客文章，最初用于回归任务，预测未来 24 小时内的评论数量。与 Yoon 等人[46]类似，我们将其转换为二元分类任务，预测博客文章是否收到评论。有 280 个整数和实值特征，并提供了单独的训练和测试数据集。更多信息可以在附录的 B.2 节中找到。

3.2 评估

对于自监督模型，一旦模型训练完成，我们通过在训练集的潜在表示上训练逻辑回归模型，并在测试集的潜在表示上测试它来评估它们。对于 SubTab，通过使用训练和测试集的子集嵌入的均值聚合获得联合潜在表示。我们使用分类任务上的性能作为表示质量的度量，这在自监督学习中通常这样做。MNIST 有 10 个类别，TCGA 有 38 个类别，其余（即 Obesity、Income 和 Blog）各有 2 个类别。

3.3 结果

MNIST：我们使用了一个简单的三层编码器架构，维度为[512, 256, 128]，称为基础模型，其中最后一层是线性层。在训练基础模型期间，我们使用了重建和对比损失。此外，我们在三种条件下训练我们的模型：i）输入数据没有任何噪声，ii）输入数据有噪声，iii）与 ii 相同，但我们还添加了为投影对{zi, zj, ...}计算的距离损失。

对于 SubTab，我们多次训练我们的基础模型，不在输入上添加噪声。对于每次训练，我们使用不同数量的子集，相邻子集之间有不同程度的重叠（图 3a）。对于少量子集（例如 2 或 3），当我们增加子集之间的重叠时，性能单调下降。但是，对于更高数量的子集，当我们增加相邻子集之间共享特征的数量时，性能通常会提高。总的来说，我们的结果表明，在 MNIST 数据集中，K = 4 且 75%重叠，以及 K = 7 且 50%重叠表现最佳，其中 K 指的是子集的数量。图 3 还显示了 K= 4 且 75%重叠的训练和测试集的 t-SNE 图，证明了聚类的高质量，而表 1 总结了测试集上所有模型的分类准确率。我们不添加噪声的基础模型优于自编码器基线和具有相同架构的其他自监督模型。我们对所有自监督模型实验了三种噪声类型，并观察到在输入上添加交换噪声会提高性能。对于

SubTab，添加距离损失并将最后一层的维度从 128 增加到 512 有助于进一步提高性能。此外，我们进行了三个额外的实验（详细信息在附录的 C.3 节中）：

在第一个实验中，对于 K = 4 且 75%重叠的最优情况，我们通过使用从不同数量的子集获得的联合表示训练和测试线性模型的准确率。从数据的单个子集开始，我们绘制了模型的训练和测试准确率（图 4a）。线性模型能够使用单个子集的表示达到 87.5%的测试准确率。当我们开始添加剩余子集的潜在表示时，训练和测试准确率都不断提高，最终在使用所有子集时达到最高准确率。与图 4a 对应的聚类演变可以在附录的图 A7 中看到。该实验表明，当我们无法访问其他特征的数据时，我们可以使用特征的小子集实现良好的性能。

在第二个实验中，我们评估了在训练期间特征缺失的条件下的 SubTab（图 5a）。为此，我们首先将 MNIST 的非随机特征分为七个不重叠的子集（对应于图 3a 中零重叠处的图例"7"的情况）。每个子集对应于 28x28 图像中的四行，从顶部四行（子集 1）到底部四行（子集 7）。然后，我们使用不同的子集组合训练基础模型；{4}，{4, 5}，{3, 4, 5}，{2, 3, 4, 5, 6}和{1, 2, 3, 4, 5, 6, 7}，得到五个不同的训练好的 SubTab 模型。请注意，我们选择的集合是从图像最具信息量的中间区域（即子集 4）扩展到最不具信息量的顶部和底部区域。

为了比较五个模型的性能，我们对每个训练好的模型执行以下步骤：1）首先获取训练和测试集所有七个子集的嵌入；2）然后使用以下七个集合中每一个的联合嵌入训练和评估逻辑回归模型：{1}，{1, 2}，{1, 2, 3}，...，{1, 2, 3, 4, 5, 6, 7}，即从第一个子集开始，我们不断添加新子集以增加集合中的信息内容。例如，对于集合{1, 2, 3}，我们首先通过使用训练集中子集 1、2 和 3 的联合嵌入训练逻辑回归模型，并通过使用测试集中相同子集的联合嵌入评估它。一个集合的联合嵌入是通过使用该集合中子集的嵌入的均值聚合获得的。除了五个模型外，我们还初始化了第六个 SubTab 模型，但保持它未经训练，并按照之前描述的相同程序使用它作为基线。结果如图 5a 所示。

在这个实验中，我们观察到即使当模型仅在单个子集（子集 4，或图 5a 中的蓝线）上训练时，聚合所有七个子集的表示（包括训练中未使用的子集）确实会提高结果。这是因为编码器能够将不同类别的样本映射到潜在空间中的不同点，即使它没有在这些样本上训练。由于我们使用相同类别的不同视角（即子集）的均值聚合，我们仍然可以使数据中的每个类别在潜在空间中与其他类别区分开来。我们还注意到，当模型在越来越多的子集上训练时，其性能不断提高。作为基线，我们还使用未训练的模型（图中的红线）进行了相同的测试，

并观察到类似的行为，即当我们在构建联合潜在表示时使用更多子集时，测试准确率通常会提高。此外，我们测量了各个子集的测试准确率，以了解每个子集有多信息量（图 5b）。结果符合预期，因为我们在这个实验中保持了特征未被打乱，并且知道对应于图像中间区域的子集（即子集 3、4 和 5）应该比对应于顶部和底部区域的子集（即子集 1 和 7）更具信息量。我们使用 28 个子集重复了相同的实验以获得更高分辨率，并在附录图 A8 中添加了结果。从这个实验中；i）我们看到当我们在训练和/或测试时包含更多子集（即子视图）时，联合表示会改善，ii）我们可以使用 SubTab 框架识别信息丰富的特征子集。

在第三个实验中，我们评估了 SubTab 在测试时处理缺失特征的能力。具体来说，我们使用在所有子集上训练的模型，并将其与未训练的模型（即我们的基线）进行比较。对于每个模型，我们通过使用所有七个子集嵌入的均值聚合获得训练集的联合嵌入，然后训练线性模型。线性模型的测试准确率通过使用；i）仅子集 4，ii）最具信息量的子集{3,4,5}的聚合，iii）排除最不具信息量的子集{2,3,4,5,6}的聚合，以及 iv）测试集的所有七个子集来测量（图 5c）。

结果表明 SubTab 可以在测试时适应缺失特征，并且仍然表现良好。这也可能表明使用子集可以在测试时缺失特征时更好地处理不确定性。随着模型以更多特征的形式收集更多信息，其预测改善（见图 5c）。我们也可以在训练期间有缺失子集的情况下训练模型，它仍然表现良好（例如，见图 5a 中对应于只在子集 4 上训练的模型的图例"4"）。我们的实验模拟了一个实际场景。例如，在医疗保健中，我们可能在一家医院无法获取某些特征，而在另一家医院可能有这些特征。因此，我们的方法在这类情况下将是有益的。

总的来说，从我们的实验中可以得出以下观察结果：i）信息量较少的子集可以为整体表示增加价值，或者至少不会损害性能（见图 5c 中{3,4,5}与"All"的聚合对比），ii）未训练的模型可用于分析哪些子集可能更具信息量，iii）一旦模型在子集上训练，单个子集的性能不会因为它是与其他子集一起训练还是单独训练而改变（例如，比较图 5b 中所有模型中子集 3、4 和 5 的性能），iv）我们框架背后的一般思想甚至适用于未训练的模型，以及 v）我们可能不需要在我们的框架中进行数据插补，因为我们可以简单地将它们视为缺失的子集，这很好，因为插补通常会扭曲数据和结果。

TCGA：我们使用了一个具有三层[1024, 784, 784]的编码器架构，其中第三层是线性的。对于 VIME-self、DAE、CAE 和我们的模型，我们尝试了三种噪声类型（高斯、交换和置零噪声）在不同百分比的掩码比率 p 下。我们观察到 p =

[0.15, 0.3]范围对所有模型都效果良好。对于高斯噪声，我们使用了均值为零，不同标准差（σ）的分布。在所有三种噪声类型中，标准差 σ = 0.1 的高斯噪声对所有模型效果最好。请注意，VIME-self 在其原始实现中使用交换噪声，但交换噪声在此数据集上效果不佳。对于 SubTab，与 MNIST 类似，我们使用了四个子集，75%重叠。SubTab 比其他自监督模型表现更好，差距显著，并且几乎使训练在原始数据上的逻辑回归模型的性能翻倍，如表 1 所示。

Obesity：我们使用了一个两层编码器，维度为[1024, 1024]。第二层是线性层。高斯噪声 N(0, 0.3)和掩码率 p = 0.2 对所有模型均效果良好。六个子集（K = 6）且 0%重叠对 SubTab 表现最佳。我们注意到，这个数据集有 164 名肥胖患者，共 253 名患者。因此，基线准确率为 164/253 = 64.82%。基于这个事实，我们可以说，除了我们的模型，所有模型在这个数据集上都表现不佳。添加高斯噪声的我们的模型准确率为 71.13 ± 4.08%，远高于所有模型，包括监督模型。这意味着我们的模型能够从数据中学习有用的表示。我们还应该注意到，我们模型的性能比 Oh 和 Zhang[36]报告的（66 ± 3.2%）要好得多，即使他们在相同的数据上训练了 DAE，并使用随机森林（一个非线性模型）而非线性模型报告了他们在学习表示上的结果。

UCI 成人收入和 BlogFeedback：对于这两个数据集，我们使用了与 Obesity 相同的架构。对于 Income 数据集，使用 5 个子集，25%重叠获得了最佳性能，而对于 Blog 数据集，我们使用了 7 个子集，75%重叠。对于基础模型，我们只使用了重建损失。向输入添加高斯噪声和向目标添加距离损失改善了两个数据集的性能。SubTab 在两个数据集上都优于其他自监督模型。所有实验的超参数选择和其他详细信息可以在附录 C.1 的表 A1 中找到。

3.4 消融研究

我们使用 MNIST 进行了全面的消融研究。表 2 总结了我们的实验。首先要注意的是，仅使用重建损失，我们的基础模型的性能已经很好。因此，我们可以认为从特征的子集重建原始特征空间是学习表示的一种非常有效的方式。通过向输入数据添加噪声，我们可以提高性能。对于 MNIST，交换噪声非常有效。此外，通过添加额外的损失（如对比和距离损失）以及将表示层的维度从 128 增加到 512，我们可以进一步改善结果。此外，我们打乱了 MNIST 的特征，以确保我们不会从相邻特征之间无意的空间相关性中获益。我们保持所有参数和随机种子相同以进行比较。如表所示，我们模型的性能变化不大。我们还尝试了连接子集的潜在变量，而不是在测试性能时聚合它们。比较表中的最后两行，

聚合被证明效果更好。请注意，我们在附录 F.4 中比较了不同的聚合函数，表明均值聚合效果最好。

最后，我们比较了 SubTab 在浅层和深层架构选择上的性能。我们训练并测试了 SubTab 的非常浅层架构（称为浅层 SubTab），并将它们与表 1 中使用的相对较深的 SubTab 模型（称为深层 SubTab）进行比较。对于 MNIST，我们使用了具有 784 维度的单层编码器和解码器，而对于其他数据集，我们使用了 1024 维度。在与较深的模型相同的条件下训练和评估浅层 SubTab。如表 3 所示，浅层 SubTab 在 MNIST 和 TCGA 上显著改善了结果，使我们的模型性能与基于 CNN 的 SOTA 模型相当，如图 4b 所示。Obesity 是唯一利用更深架构的数据集。

解析与重点：

- 研究团队使用了五个数据集进行实验：MNIST、TCGA、Obesity、Income 和 Blog
- 比较对象：自编码器基线、其他自监督方法和完全监督模型
- 主要实验发现：
    1. SubTab 在所有测试数据集上表现优于其他自监督方法
    2. 在表格格式的 MNIST 上达到 98.31%准确率，与基于 CNN 的 SOTA 模型相当
    3. 在 TCGA 数据集上性能几乎是原始逻辑回归的两倍
    4. 在 Obesity 数据集上超越所有现有方法，包括监督模型
- 进行了三个关键实验：
    1. 单子集到多子集的渐进测试，证明协作推理的有效性
    2. 训练时特征缺失的情况：证明即使只在部分子集上训练也有良好表现
    3. 测试时处理缺失特征的能力：证明框架的鲁棒性
- 消融研究重要发现：
    1. 仅使用重建损失已能达到良好效果
    2. 交换噪声在 MNIST 上最有效

3. 添加对比和距离损失提升性能
4. 均值聚合优于连接表示
5. 浅层 SubTab 在某些数据集上优于深层模型

- 实验表明 SubTab 适用于各种数据分布和结构的表格数据集

4 Related works（相关工作）翻译与解析

翻译： 我们请读者参考引言部分，其中列出了自监督学习中一些值得注意的近期工作。由于我们的工作专注于表格数据，我们将回顾一些在自监督框架中对表格数据进行的最近工作。最近的工作主要基于解决前置任务。例如，Yoon 等人[46]使用一个带有连接到其表示层的分类器的去噪自编码器。随机生成一个二元掩码来掩盖和覆写表格数据中的一部分条目，并将损坏的数据作为输入给到编码器。分类器用于预测掩码，而解码器用于重建未损坏的原始输入，类似于去噪自编码器[43]。尽管所提出的方法在实验中表现良好，但这种方法有几个缺点。首先，这种方法在非常高维、小而嘈杂的数据集中可能不会表现良好，因为模型可能很容易变得过参数化并容易过拟合到数据。其次，在这种设置下训练分类器可能具有挑战性，因为它需要预测非常高维、稀疏和不平衡的二元掩码，类似于在不平衡的二元数据集上训练模型时观察到的问题。

类似地，TabNet[1]和 TaBERT[45]也尝试从损坏的数据中恢复原始数据。

解析与重点：

- 作者先引导读者回顾引言部分中提到的自监督学习相关工作
- 表格数据自监督学习的主要现有方法基于前置任务（pretext task）设计
- VIME-self (Yoon 等人 2020)是最具代表性的工作：
    - 使用去噪自编码器架构
    - 添加分类器预测哪些特征被掩盖
    - 主要缺点：
        1. 在高维、小样本和嘈杂数据上可能效果不佳
        2. 训练分类器面临高维、稀疏和不平衡掩码预测挑战
- 其他相关工作包括 TabNet 和 TaBERT，核心思想也是从损坏数据恢复原

始数据

- 所有这些方法与 SubTab 的根本区别：它们依赖于添加噪声破坏数据，而 SubTab 使用特征子集

5 Conclusion（结论）翻译与解析

翻译： 在这项工作中，我们表明，一个简单的基于 MLP 的自编码器，在表格格式的 MNIST 上训练，可以在无监督/自监督框架中与在 MNIST 图像上训练的基于 CNN 的 SOTA 模型表现相当。SubTab 在表格设置的 MNIST 数据集上实现了 SOTA。我们还在其他常用的表格数据集上测试了我们的方法，并证明了它的好处。在 SubTab 中，主要的性能提升来自模型的两个部分：i）从特征子集重建所有特征，ii）通过聚合子集的嵌入来学习联合表示。

使用特征子集可能消除了训练期间对数据插补的需求，并允许在测试时使用特征子集进行推理。这可能为高维数据的分布式训练打开大门，因为模型可以同时在不同的特征子集上训练。我们还可以潜在地利用具有共同特征的不同数据集，通过将这些特征分配给相同的子集（即迁移学习）。我们应该注意，在我们的实验中，子集共享相同的自编码器，尽管如果某些特征与其他特征截然不同，我们可以为不同的子集使用单独的自编码器。

SubTab 在训练期间只使用重建损失时在计算上是可扩展的。然而，使用对比和/或距离损失需要投影的组合，这使得训练期间的计算复杂度变为二次，并限制了我们可以用来划分数据的子集数量。在这种情况下，测试时的计算复杂度仍然是线性的，因为我们只需要计算子集表示的聚合。此外，当我们将特征划分为子集时，我们在训练和测试期间保持每个子集中特征的位置相同，因为神经网络不是排列不变的。作为一个可能的解决方案，我们可以通过将特征集合视为一个集合，将我们的工作扩展到排列不变架构。我们还表明，SubTab 框架可以用于发现具有有限分辨率的最信息丰富的特征子集。SubTab 的分层版本可能用于识别个别重要特征，但我们将其作为未来的工作。

最后，尽管本工作的主要重点是表格数据设置，但 SubTab 可以扩展到其他领域，如图像、音频、文本等。我们将 SubTab 的扩展和应用作为未来的工作。

解析与重点：

- 主要成就：
    1. 表格格式 MNIST 上达到与 CNN-SOTA 模型相当的性能

2. 在多个表格数据集上证明了框架的优越性

- 性能提升的两大关键因素：
    1. 从特征子集重建所有特征
    2. 通过聚合子集嵌入学习联合表示
- 框架优势：
    1. 避免了数据插补需求
    2. 支持在测试时处理缺失特征
    3. 可用于分布式训练
    4. 支持跨数据集迁移学习
- 计算效率考量：
    1. 仅使用重建损失时计算可扩展
    2. 添加对比/距离损失导致训练时计算复杂度变为二次
    3. 测试时仍保持线性复杂度
- 局限性：
    1. 需要固定子集中特征位置（非排列不变）
    2. 可能存在表示坍塌风险（但在实践中较低）
- 未来工作方向：
    1. 扩展到排列不变架构
    2. 开发分层版本以识别个别重要特征
    3. 将方法扩展到其他数据类型（图像、音频、文本）

6 Broader Impact（更广泛影响）翻译与解析

翻译： 表格数据是医疗保健、金融、法律和许多其他领域常用的格式。尽管其广泛使用，但深度学习研究，特别是关于无监督表示学习的研究，主要集中在图像、文本和音频等其他数据类型上。我们的论文试图通过引入一个新框架来学习表格数据的良好表示，从而在无监督/自监督设置中缩小这一差距。这一研

究方向的进展将为表格数据在其他领域的广泛应用打开大门，如迁移学习、分布式学习和多视角学习，其中我们可以结合来自表格数据的知识（如人口统计和基因组学）与图像、文本和音频中的知识。然而，我们应该注意此类数据整合在偏差和隐私问题方面可能引入的缺点。

解析与重点：

- 社会意义：
    1. 弥补表格数据在深度学习研究中的不足
    2. 为医疗保健、金融、法律等领域提供更好的自监督学习方法
- 潜在应用：
    1. 迁移学习：跨数据集使用共同特征
    2. 分布式学习：在不同特征子集上并行训练
    3. 多视角学习：结合表格数据与其他模态数据
- 伦理考量：
    1. 数据整合可能引入新的偏见
    2. 隐私问题需要关注
    3. 需要考虑表格数据特有的敏感性（如医疗、金融信息）
- 研究价值：
    1. 填补理论空白：为表格数据设计专属自监督学习方法
    2. 实用价值：可直接应用于众多依赖表格数据的行业