VIME: 將自監督與半監督學習的成功擴展到表格數據領域

重點總結

1. 問題背景：
    - 自監督和半監督學習在圖像和語言領域取得了顯著成功，但由於缺乏明確的結構特徵，這些方法難以直接應用於表格數據
    - 表格數據中，特徵間的相關結構不明顯且因數據集而異，無法像圖像和語言那樣使用固定的結構假設
2. **VIME** 框架的創新點：
    - 引入了兩個互補的前置任務：特徵向量估計和掩碼向量估計
    - 開發了專門的表格數據增強技術，通過掩碼和填充創建有意義的增強樣本
    - 提出了統一的自監督和半監督學習框架，從未標記數據中學習特徵相關性
3. 技術方法：
    - 自監督學習部分：通過掩碼和重建訓練編碼器學習特徵相關性
    - 半監督學習部分：利用訓練好的編碼器為每個樣本生成多個增強版本，並通過一致性正則化改進預測模型
    - 兩部分共同工作，充分利用標記和未標記數據
4. 實驗結果：
    - 在基因組學數據上，**VIME** 在僅使用一半標記數據的情況下能達到基準方法的性能
    - 在臨床數據上，**VIME** 是唯一顯著優於監督學習模型的自/半監督方法
    - 在公共表格數據集上，**VIME** 始終優於監督學習和其他自/半監督學習方法
    - 消融研究證明了框架每個組件都貢獻了性能提升

5. **應用價值：**
    - **適用於醫療、金融和基因組學等表格數據豐富但標記稀少的領域**
    - **提供了處理表格數據的新思路，填補了自監督和半監督學習在表格數據上的研究空白**
    - **提出的方法可以擴展到數據增強、特徵編碼和缺失值填充等相關任務**

VIME 成功地將自監督和半監督學習的理念擴展到表格數據領域，通過巧妙地設計掩碼估計和特徵恢復任務，克服了表格數據缺乏明顯結構的限制，實現了在少量標記數據情況下的有效學習，為實際應用中常見的表格數據處理提供了新的解決方案。

摘要 (Abstract)

自監督和半監督學習框架在圖像和語言領域使用有限標記數據訓練機器學習模型方面取得了顯著進展。這些方法嚴重依賴於領域數據集的獨特結構（如圖像中的空間關係或語言中的語義關係）。它們無法適應缺乏與圖像和語言數據相同顯性結構的一般表格數據。在本文中，作者填補了這一差距，提出了用於表格數據的新型自監督和半監督學習框架，統稱為 VIME（Value Imputation and Mask Estimation，值填充和掩碼估計）。除了用於自監督學習的重建前置任務外，作者還創建了一個從受損表格數據中估計掩碼向量的新穎前置任務。作者還為自監督和半監督學習框架引入了一種新穎的表格數據增強方法。在實驗中，作者在來自各種應用領域的多個表格數據集上評估了所提出的框架，如基因組學和臨床數據。與現有基線方法相比，VIME 取得了更優的性能。

1. 引言 (Introduction)

在各種應用中（如圖像分類、目標檢測和語言翻譯），通過在大型標記數據集（如 ImageNet）上進行監督學習，深度學習模型已經取得了巨大成功。不幸的是，在若干領域中，收集足夠大的標記數據集是昂貴的，甚至是不可能的（例如，針對特別罕見疾病的醫療數據集）。在這些情況下，通常有大量的未標記數據可用 - 數據集通常從大型人群中收集，但目標標籤僅適用於一小部分人群。例如，「10 萬基因組計劃」對來自約 85,000 名受罕見疾病（如癌症）影響的 NHS 患者的 10 萬個基因組進行了測序。按定義，罕見疾病發生在每 2000 人中

不到 1 人。像這樣的數據集為自監督和半監督學習算法提供了巨大的機會，這些算法可以利用未標記數據進一步提高預測模型的性能。

不幸的是，現有的自監督和半監督學習算法對表格數據無效，因為它們嚴重依賴圖像或語言數據的空間或語義結構。標準的自監督學習框架設計一個（或一組）前置任務，從原始輸入特徵中學習信息性表示。對於語言領域，BERT 引入了 4 個不同的前置任務（例如，從前面的詞預測未來的詞）來學習語言數據的表示。在圖像領域，旋轉、拼圖和上色可以作為前置任務來學習圖像的表示。標準的半監督學習方法也遭受相同的問題，因為它們使用的預測模型正則化器基於這些數據結構的一些先驗知識。例如，一致性正則化器鼓勵預測模型在樣本及其增強變體上具有相同的輸出分佈，如圖像及其旋轉變體，或兩幅圖像及其凸組合。旋轉的概念在表格數據中根本不存在。此外，在許多情況下，變量通常是分類的，不允許有意義的凸組合。即使在所有變量都是連續的情況下，也不保證數據流形是凸的，因此採取凸組合要麼會產生分佈外樣本（從而降低模型性能），要麼僅限於生成與實際樣本非常接近的樣本（限制了數據增強的有效性）。

貢獻：在本文中，作者提出了用於表格數據的新型自監督和半監督學習框架。對於自監督學習，除了特徵向量估計外，作者還引入了一個新穎的前置任務，即掩碼向量估計。為了解決這些前置任務，編碼器函數從未標記數據中的原始特徵學習構建信息性表示。對於半監督學習，作者引入了一種新穎的表格數據增強方案。作者使用訓練過的編碼器為每個數據點生成多個增強樣本，方法是使用多個不同的掩碼對每個點進行掩碼，然後填充每個掩碼數據點的損壞值。最後，作者提出了用於表格數據的系統性自監督和半監督學習框架 VIME（Value Imputation and Mask Estimation），結合這些思想，在具有少量標記樣本的各種領域的多個表格數據集上產生最先進的性能。

2. 相關工作 (Related Works)

自監督學習 (Self-supervised learning)框架是使用未標記數據的表示學習方法。它可分為兩類：使用前置任務和對比學習。大多數現有的前置任務工作僅適用於圖像或自然語言：(i) 代理類預測（縮放和平移），(ii) 旋轉角度預測，(iii) 著色，(iv) 區塊相對位置估計，(v) 拼圖求解，(vi) 圖像去噪，(vii) 部分到部分配准，以及(viii) 下一個詞和前一個詞預測。大多數現有的對比學習工作也僅適用於圖像或自然語言，因為它們的數據增強方案以及定義相似性的時間和空間關係：(i) 對比預測編碼，(ii) 對比多視角編碼，(iii) SimCLR，(iv) 動量對

比。

有一些現有的自監督學習工作可以應用於表格數據。在去噪自編碼器中，前置任務是從損壞樣本中恢復原始樣本。在上下文編碼器中，前置任務是從損壞樣本和掩碼向量中重建原始樣本。TabNet 和 TaBERT 的自監督學習前置任務也是恢復損壞的表格數據。

在本文中，作者提出了一個新的前置任務：恢復掩碼向量，並且還使用了一種新穎的損壞樣本生成方案來恢復原始樣本。此外，作者提出了一種新穎的表格數據增強方案，可以與各種對比學習框架結合，將自監督學習擴展到表格領域。

半監督學習 (Semi-supervised learning)框架可分為兩類：熵最小化和一致性正則化。熵最小化鼓勵分類器在未標記數據上輸出低熵預測。例如，從未標記數據上的高置信度預測構建硬標籤，並以監督方式使用這些偽標籤和標記數據一起訓練網絡。一致性正則化鼓勵樣本和其隨機變異版本之間的某種一致性。Π-model 使用 L2 損失來鼓勵預測的一致性。Mean teacher 使用 L2 損失來鼓勵中間表示的一致性。虛擬對抗訓練(VAT)通過最小化樣本與多個增強版本之間預測的最大差異來鼓勵預測一致性。MixMatch 和 ReMixMatch 使用 MixUp 作為數據增強方法，在一個統一框架中結合熵最小化和一致性正則化。還有一系列有趣的基於圖的半監督學習工作，考慮了樣本通過給定邊緣連接的網絡數據的特殊情況，例如引文網絡，其中文章與其引用相連。在這裡，作者引入了一種適用於一般表格數據的新穎數據增強方法，可以與各種半監督學習框架結合，以半監督方式訓練預測模型。

3. 問題表述 (Problem Formulation)

在本節中，作者介紹了自監督和半監督學習的一般表述。假設我們有一個小型標記數據集 Dl = {xi, yi}^(Nl)*(i=1)和一個大型未標記數據集 Du = {xi}^(Nl+Nu)*(i=Nl+1)，其中 Nu >> Nl，xi ∈ X ⊆ R^d 且 yi ∈ Y。在單任務學習中標籤 yi 是標量，而在多任務學習中可以是多維向量。我們假設 Dl 和 Du 中的每個輸入特徵 xi 都是從特徵分佈 pX 獨立同分佈採樣的，而 Dl 中的標記數據對(xi, yi)則來自聯合分佈 pX,Y。當從 pX,Y 可獲得的標記樣本有限時，僅通過監督學習訓練的預測模型 f：X → Y 很可能會過擬合訓練樣本，因為我們最小化的經驗監督損失∑^(Nl)*(i=1) l(f(xi), yi)與期望監督損失 E*(x,y)~pX,Y[l(f(x), y)]之間存在顯著偏差，其中 l(·,·)是某種標準監督損失函數（例如交叉熵）。

3.1 自監督學習

自監督學習旨在從未標記數據中學習信息性表示。在本小節中，作者專注於使用各種自監督/前置任務的自監督學習，讓前置模型來解決。這些任務被設計成具有挑戰性但與我們試圖解決的下游任務高度相關。理想情況下，前置模型在解決前置任務的過程中將從原始數據中提取一些有用信息。然後，下游任務的預測模型 f 可以利用提取的信息。一般來說，自監督學習構建一個編碼器函數 e：X → Z，該函數接受樣本 x ∈ X 並返回信息性表示 z = e(x) ∈ Z。表示 z 經過優化以解決用偽標籤 ys ∈ Ys 和自監督損失函數 lss 定義的前置任務。例如，前置任務可以是預測原始數據集中某些旋轉圖像的旋轉角度，其中 ys 是真實旋轉角度，而 lss 是預測的旋轉角度與 ys 之間的平方差。作者將前置預測模型定義為 h：Z → Ys，該模型與編碼器函數 e 一起通過最小化期望自監督損失函數 lss 進行聯合訓練：

min_(e,h) E_(xs,ys)~pXs,Ys[lss(ys, (h∘e)(xs))]

其中 pXs,Ys 是一個前置分佈，用於生成偽標記樣本(xs, ys)以訓練編碼器 e 和前置預測模型 h。請注意，我們有足夠的樣本來近似上述目標函數，因為對於 Du 中的每個輸入樣本，我們可以免費生成前置樣本(xs, ys)，例如，旋轉圖像 xi 創建 xs，並將旋轉角度作為標籤 ys。訓練後，編碼器函數 e 可用於從原始數據中提取更好的數據表示，以解決各種下游任務。請注意，在下游任務（及其損失）事先已知的情況下，編碼器可以與下游任務的模型聯合訓練。

3.2 半監督學習

半監督學習通過聯合最小化監督損失函數與在輸出空間 Y 上定義的某種非監督損失函數來優化預測模型 f。形式上，半監督學習表述為以下優化問題：

min_f E_(x,y)~~pXY[l(y, f(x))] + β · E_x~~pX,x'~p̃X(x'|x)[lu(f(x), f(x'))]

其中 lu：Y × Y → R 是一個非監督損失函數，引入超參數 β ≥ 0 來控制監督和非監督損失之間的權衡。x'是 x 的擾動版本，假設從條件分佈 p̃X(x'|x)中抽取。第一項使用小型標記數據集 Dl 進行估計，而第二項使用 Du 中的所有輸入特徵進行估計。非監督損失函數(lu)通常受到下游任務的一些先驗知識的啟發。例如，一致性正則化鼓勵模型 f 在其輸入被擾動(x')時產生相同的輸出分佈。

4. 提出的模型：VIME

在本節中，作者描述了 VIME，這是他們為表格數據設計的自監督和半監督學習

系統性方法。作者首先在自監督學習中提出兩個前置任務，然後使用通過自監督學習從前置任務中學習到的編碼器開發半監督學習中的非監督損失函數。

4.1 表格數據的自監督學習

作者引入了兩個前置任務：特徵向量估計和掩碼向量估計。目標是優化前置模型，從其受損變體中恢復輸入樣本（特徵向量），同時估計應用於樣本的掩碼向量。

在該框架中，兩個前置任務共享單一前置分佈 pXs,Ys。首先，掩碼向量生成器輸出一個二元掩碼向量 m = [m1, ..., md]^T ∈ {0, 1}^d，其中每個 mj 從概率為 pm 的伯努利分佈中隨機抽樣（即 pm = ∏^d_(j=1) Bern(mj|pm)）。然後，前置生成器 gm：X × {0, 1}^d → X 將來自 Du 的樣本 x 和掩碼向量 m 作為輸入，並生成一個掩碼樣本 x̃。x̃ 的生成過程如下：

x̃ = gm(x, m) = m ⊙ x̄ + (1 − m) ⊙ x

其中 x̄ 的第 j 個特徵從經驗分佈 p̂Xj = (1/Nu) ∑^(Nl+Nu)_(i=Nl+1) δ(xj = xi,j)中抽樣，其中 xi,j 是 Du 中第 i 個樣本的第 j 個特徵（即每個特徵的經驗邊際分佈）。這個生成過程確保了受損樣本 x̃ 不僅是表格的，而且與 Du 中的樣本相似。與標準的樣本損壞方法相比，如對缺失特徵添加高斯噪聲或用零替換，作者的方法生成的 x̃ 更難與 x 區分。這種難度對於自監督學習至關重要。

在作者的前置分佈 pXs,Ys 中有兩方面的隨機性。顯式地，m 是從伯努利分佈中隨機抽樣的向量。隱式地，前置生成器 gm 也是一個隨機函數，其隨機性來自 x̄。這種隨機性增加了從 x̃ 中重建 x 的難度。通過更改超參數 pm（Bern(·|pm)中的概率），可以調整難度水平，該超參數控制將被掩碼和損壞的特徵比例。

按照自監督學習的慣例，編碼器 e 首先將掩碼和受損的樣本 x̃ 轉換為表示 z，然後引入前置預測模型來從 z 中恢復原始樣本 x。這比現有的前置任務更具挑戰性，比如糾正圖像的旋轉或為灰度圖像上色。旋轉或灰度圖像仍然包含有關原始特徵的一些信息。相比之下，掩碼完全移除了 x 中的一些特徵，並用可能來自 Du 中不同隨機樣本的每個特徵的噪聲樣本 x̄ 替換它們。結果樣本 x̃ 可能不包含有關缺失特徵的任何信息，甚至難以識別哪些特徵缺失。為了解決這樣一個具有挑戰性的任務，作者首先將其分為兩個子任務（前置任務）：

(1) 掩碼向量估計：預測哪些特徵已被掩碼； (2) 特徵向量估計：預測已被損壞的特徵的值。

作者為每個前置任務引入了單獨的前置預測模型。這兩個模型都基於由編碼器 e 給出的表示 z 運作，並試圖協作估計 m 和 x。這兩個模型及其功能是：

- 掩碼向量估計器，sm：Z → [0, 1]^d，接受 z 作為輸入並輸出一個向量 m̂ 來預測 x̃ 的哪些特徵已被有噪聲的對應物替換（即 m）；
- 特徵向量估計器，sr：Z → X，接受 z 作為輸入並返回 x̂，原始樣本 x 的估計。

編碼器 e 和前置預測模型（在我們的情況下，兩個估計器 sm 和 sr）在以下優化問題中聯合訓練：

min_(e,sm,sr) E_x~~pX,m~~pm,x̃~gm(x,m)[lm(m, m̂) + α · lr(x, x̂)]

其中 m̂ = (sm ∘ e)(x̃)和 x̂ = (sr ∘ e)(x̃)。第一個損失函數 lm 是掩碼向量每個維度的二元交叉熵損失之和：

lm(m, m̂) = -(1/d) ∑^d_(j=1) [mj log((sm ∘ e)j(x̃)) + (1 - mj) log(1 - (sm ∘ e)j(x̃))]

第二個損失函數 lr 是重建損失：

lr(x, x̂) = (1/d) ∑^d_(j=1) (xj - (sr ∘ e)j(x̃))^2

α 調整這兩個損失之間的權衡。對於分類變量，作者將方程式 6 修改為交叉熵損失。

編碼器學到了什麼？ 這兩個損失函數共享編碼器 e。它是我們在下游任務中將利用的唯一部分。為了理解編碼器如何有利於這些下游任務，我們考慮編碼器必須能夠做什麼來解決我們的前置任務。作者做出以下直觀觀察：對於 e 來說，捕捉 x 的特徵之間的相關性並輸出一些能夠恢復 x 的潛在表示 z 是很重要的。在這種情況下，sm 可以從特徵值之間的不一致性識別掩碼特徵，而 sr 可以通過從相關的非掩碼特徵學習來填充掩碼特徵。例如，如果一個特徵的值與其相關特徵相差很大，這個特徵可能被掩碼和損壞了。作者注意到，在其他自監督學習框架中也學習了相關性，例如旋轉圖像中的空間相關性和未來與過去詞之間的自相關性。作者的框架在學習表格數據相關性方面是新穎的，這些相關性結構比圖像或語言中的相關性結構不那麼明顯。無論對象類型如何（例如，語言、圖像或表格數據），捕捉對象不同部分之間相關性的學習表示，對各種下游任務來說都是有信息的輸入。

4.2 表格數據的半監督學習

現在作者展示了如何在半監督學習中使用上一小節中的編碼器函數 e。他們的半監督學習框架遵循第 3 節所給的結構。令 fe = f ∘ e 且 ŷ = fe(x)。作者通過最小化目標函數來訓練預測模型 f：

L_final = L_s + β · L_u

監督損失 L_s 由以下公式給出：

L_s = E_(x,y)~pXY[ls(y, fe(x))]

其中 ls 是標準監督損失函數，例如回歸的均方誤差或分類的分類交叉熵。非監督（一致性）損失 L_u 定義在原始樣本(x)與從受損和掩碼樣本(x̃)重建的樣本之間：

L_u = E_x~~pX,m~~pm,x̃~gm(x,m)[(fe(x̃) - fe(x))^2]

作者的一致性損失受到一致性正則化器思想的啟發：鼓勵預測模型 f 在其輸入被擾動時返回相似的輸出分佈。然而，在作者的框架中，擾動是通過自監督框架學習的，而在之前的工作中，擾動是來自手動選擇的分佈，如旋轉。

對於固定樣本 x，方程(9)中的內部期望是相對於 pm 和 gm(x, m)取的，可以解釋為受損和掩碼樣本預測的方差。β 是另一個超參數，用於調整監督損失 L_s 和一致性損失 L_u。在訓練的每次迭代中，對於批次中的每個樣本 x ∈ Du，作者通過重複方程(3)中的操作 K 次創建 K 個增強樣本 x̃1, ..., x̃K。每次使用 Du 中的樣本 x 於批次時，作者都會重新創建這些增強樣本。L_u 的隨機近似由以下給出：

L̂_u = (1/(N_b K)) ∑^(N_b)*(i=1)* ∑^*K*(k=1) [(fe(x̃i,k) - fe(xi))^2] = (1/(N_b K)) ∑^(N_b)*(i=1)* ∑^*K*(k=1) [(f(zi,k) - f(zi))^2]

其中 N_b 是批次大小。在訓練期間，預測模型 f 被正則化以對 zi 和 zi,k, k = 1, ..., K 做出相似的預測。訓練 f 後，新測試樣本 x_t 的輸出由 ŷ = fe(x_t)給出。

5. 實驗

在本節中，作者進行了一系列實驗，以展示其框架(VIME)在來自不同應用領域的多個表格數據集上的有效性，包括基因組學和臨床數據。作者使用 Min-max scaler 將數據標準化到 0 和 1 之間。對於自監督學習，作者將 VIME 與兩個基準進行比較，分別是去噪自編碼器(DAE)和上下文編碼器(Context Encoder)。對於半監督學習，作者使用數據增強方法 MixUp 作為主要基準。作者排除了僅適

用於圖像或語言數據的自監督和半監督學習基準。作為基線，作者還包括了監督學習基準。在實驗中，自監督和半監督學習方法同時使用標記數據和未標記數據，而監督學習方法僅使用標記數據。

5.1 基因組學數據：全基因組多基因評分

在本小節中，作者評估了方法在來自 UK Biobank 的大型基因組學數據集上的表現，該數據集包含約 400,000 個人的基因組信息(SNPs)和 6 個相應的血細胞特徵：(1)平均網狀紅細胞體積(MRV)，(2)平均血小板體積(MPV)，(3)平均紅細胞血紅蛋白(MCH)，(4)紅細胞中網狀紅細胞比例(RET)，(5)血小板壓積(PCT)，以及(6)白細胞中單核細胞百分比(MONO)。數據集的特徵包括約 700 個 SNPs（經過標準 p 值過濾過程後），每個 SNP 取值為{0, 1, 2}，被視為分類變量（具有三個類別）。這裡，作者有 6 個不同的血細胞特徵需要預測，作者將每個特徵視為獨立的預測任務（不同血細胞特徵選擇的 SNPs 是不同的）。

為了測試自監督和半監督學習在小標記數據設置中的有效性，VIME 和基準方法被要求預測 6 個血細胞特徵，同時作者逐漸增加標記數據點的數量從 1,000 到 100,000 個樣本，而將剩餘數據作為未標記數據（超過 300,000 個樣本）。由於線性模型（Elastic Net）在基因組學數據集上相比其他非線性模型（如多層感知器和隨機森林）表現更優，作者使用其作為預測模型。

圖 3 顯示了 MSE 性能（y 軸）與標記數據點數量（x 軸，對數尺度）的關係，數量從 1,000 增加到 10,000。提出的模型(VIME)優於所有基準，包括純監督方法 ElasticNet、自監督方法 Context Encoder 和半監督方法 MixUp。事實上，在許多情況下，即使 VIME 只能訪問一半的標記數據點（與基準相比），它仍然表現相似。

5.2 臨床數據：患者治療預測

在本小節中，作者評估了方法在臨床數據上的表現，使用英國和美國前列腺癌數據集（分別來自 Prostate Cancer UK 和 SEER 數據集）。特徵包括患者的臨床信息（例如，年齡、等級、階段、Gleason 評分）- 共 28 個特徵。作者預測英國前列腺癌患者的 2 種可能治療：(1)激素治療（患者是否接受激素治療），(2)根治性治療（患者是否接受根治性治療）。這兩項任務都是二元分類。在英國前列腺癌數據集中，作者只有約 10,000 個標記患者樣本。美國前列腺癌數據集包含超過 200,000 個未標記患者樣本，是標記英國數據集的二十倍。作者使用英國數據集的 50%（作為標記數據）和整個美國數據集（作為未標記數據）進行

訓練，英國數據的其餘部分用作測試集。作者還測試了三種流行的監督學習模型：邏輯回歸、2 層多層感知器和 XGBoost。

表 1 顯示，VIME 取得了最佳預測性能，優於基準。更重要的是，VIME 是唯一顯著優於監督學習模型的自監督或半監督學習框架。這些結果揭示了使用 VIME 利用大型未標記表格數據集（如美國數據集）增強模型預測能力的獨特優勢。作者還展示了 VIME 在英國標記數據和美國未標記數據之間存在分佈偏移的情況下仍能良好表現（詳見補充材料第 2 節）。

5.3 公共表格數據

為了進一步驗證結果的普遍性和可重現性，作者使用三個公共表格數據集比較了 VIME 與基準方法：MNIST（解釋為具有 784 個特徵的表格數據）、UCI Income 和 UCI Blog。作者使用 10%的數據作為標記數據，剩餘 90%的數據作為未標記數據。在所有三個數據集上，使用單獨測試集上的預測準確率作為指標。如表 2 所示（類型 - 監督模型、自監督模型、半監督模型和 VIME），VIME 無論應用領域如何都達到了最佳準確率。這些結果進一步證實了 VIME 在多種表格數據集上的優越性。

5.4 消融研究

在本節中，作者進行了消融研究，分析 VIME 在 5.3 節介紹的表格數據集上各組件的性能增益。作者定義了 VIME 的三個變體：

- 僅監督（Supervised only）：排除自監督和半監督學習部分（即 2 層感知器）
- 僅半監督（Semi-SL only）：排除自監督學習部分（即移除圖 2 中的編碼器）
- 僅自監督（Self-SL only）：排除半監督學習部分（即 $\beta = 0$）。具體來說，首先通過自監督學習訓練編碼器，然後使用損失函數（在方程（7）中設 $\beta = 0$（僅利用標記數據））訓練預測模型。

表 2（類型 - VIME 變體和 VIME）顯示，與僅監督相比，僅自監督和僅半監督都顯示出性能增益，而 VIME 始終優於其變體。VIME 中的每個組件都可以提高預測模型的性能，當它們在統一框架中協同工作時，可以達到最佳性能。作者注意到，僅自監督導致的性能下降大於僅半監督，因為前者的預測模型僅在沒有非監督損失函數 Lu 的小型標記數據集上訓練，而後者的預測模型通過最小化

兩個損失但沒有編碼器進行訓練。

6. 討論：為什麼表格數據需要提出的模型(VIME)？

圖像和表格數據有很大的不同。圖像中像素之間的空間相關性或文本數據中詞之間的序列相關性是眾所周知的，且在不同數據集之間保持一致。相比之下，表格數據中特徵之間的相關結構未知且在不同數據集之間各不相同。換句話說，表格數據中沒有"共同"的相關結構（與圖像和文本數據不同）。這使得表格數據的自監督和半監督學習更具挑戰性。請注意，圖像領域的有效方法並不保證在表格領域也能取得良好的結果（反之亦然）。此外，圖像數據中使用的大多數增強和前置任務不適用於表格數據；因為它們直接利用圖像的空間關係進行增強（如旋轉）和前置任務（如拼圖和上色）。為了將自監督和半監督學習的成功從圖像領域轉移到表格領域，提出適用且恰當的表格數據前置任務和增強（作者的主要創新點）至關重要。請注意，更好的增強和前置任務可以顯著提高自監督和半監督學習的性能。

廣泛影響

表格數據是現實世界中最常見的數據類型。大多數數據庫包含表格數據，如醫療和金融數據集中的人口統計信息以及基因組數據集中的 SNPs。然而，深度學習（尤其是在圖像和語言領域）的巨大成功尚未完全擴展到表格領域。在表格領域，決策樹集成仍然達到最先進的性能。如果我們能有效地將圖像和語言的成功深度學習方法擴展到表格數據，機器學習在現實世界中的應用將大大擴展。本文在自監督和半監督學習框架方面邁出了一步，這些框架最近在圖像和語言領域取得了顯著成功。此外，提出的表格數據增強和表示學習方法可用於多個領域，如表格數據編碼、平衡表格數據的標籤以及缺失數據填充。