SCARF: 使用隨機特徵損壞的自監督對比學習 - 論文解析

總體重點歸納

1. 創新貢獻：**SCARF** 提出了一種簡單但有效的表格數據自監督對比學習技術，通過損壞隨機特徵子集生成不同視圖。
2. 關鍵機制：
    - 從特徵的經驗邊際分佈中抽樣替換隨機選取的特徵
    - 使用對比學習框架最大化同一輸入不同視圖的相似性
    - 全面微調編碼器和分類頭，只使用任務監督
3. 技術優勢：
    - 對超參數設置（批量大小、損壞率、溫度）不敏感
    - 對特徵縮放方法選擇不敏感
    - 實現簡單，無需複雜工程
4. 廣泛驗證：
    - 在 **69** 個真實世界表格數據集上進行全面測試
    - 涵蓋完全監督、半監督和標籤噪聲三種關鍵場景
    - 與多種基線（自編碼器、**dropout**、**mixup** 等）進行比較
5. 優越性能：
    - 在完全監督設置中提高 **1-2%**相對精度
    - 在標籤噪聲環境下提高 **2-3%**
    - 在半監督環境下提高 **2-4%**
    - 可與現有方法互補，進一步提高性能
6. 與 **VIME** 的主要區別：
    - **SCARF** 使用對比損失而非自編碼器損失
    - 微調所有模型權重而非僅任務頭

    - 僅使用任務監督進行微調，不結合重構損失
7. **實用價值：**
    - 為表格數據提供有效的預訓練策略
    - 在標記數據稀缺場景中尤為有價值
    - 提供了應對標籤噪聲的穩健方法
8. **研究意義：**
    - 將自監督對比學習成功擴展到表格數據領域
    - 為表格數據表示學習提供了新視角
    - 填補了重要研究空白，促進自監督學習的普適性

SCARF 簡單而強大的設計使其成為處理表格數據挑戰的重要工具，尤其是在標記數據有限或不可靠的實際場景中。

Abstract (摘要)

翻譯： 自監督對比表示學習在視覺和自然語言領域已被證明非常成功，實現了最先進的性能，且所需標記數據量減少了數個數量級。然而，這類方法都是特定領域的，很少有工作能夠在真實世界的表格數據集上利用這種技術。我們提出 SCARF，一種簡單、廣泛適用的對比學習技術，其中視圖是通過損壞一個隨機特徵子集形成的。當應用於 OpenML-CC18 基準測試中的 69 個真實世界表格分類數據集進行深度神經網絡預訓練時，SCARF 不僅在完全監督設置中提高了分類精度，在存在標籤噪聲的情況下以及在只有一部分可用訓練數據被標記的半監督設置中也同樣提高了精度。我們展示 SCARF 能夠補充現有策略，並優於自編碼器等替代方法。我們進行了全面的消融研究，詳細說明了各種因素的重要性。

解析與重點：

- SCARF 是一種針對表格數據的自監督對比學習方法
- 核心機制：通過損壞隨機特徵子集來創建數據的不同視圖
- 在 69 個真實世界表格數據集上進行了測試

- 優勢：
    1. 提高完全監督學習下的分類精度
    2. 提高存在標籤噪聲情況下的精度
    3. 提高半監督學習情況下的精度
    4. 優於自編碼器等傳統方法
    5. 可與現有策略互補
- 進行了全面的消融研究以驗證各種設計選擇的重要性

1 INTRODUCTION (引言)

翻譯： 在許多機器學習任務中，未標記數據豐富但標記數據成本高昂，需要人工標記者。自監督學習的目標是利用大量未標記數據來學習對下游任務（如分類）有用的表示。自監督學習在計算機視覺（Grill 等，2020；Misra & Maaten，2020；He 等，2020；Tian 等，2019）和自然語言處理（Song 等，2020；Wang 等，2019；Raffel 等，2019）領域已被證明至關重要。一些最近的例子包括：Chen 等（2020）展示了在他們提出的 SimCLR 學習的表示上訓練線性分類器，顯著優於先前最先進的圖像分類器，且僅需 100 倍更少的標籤；Brown 等（2020）通過其 GPT-3 語言模型展示，在大型文本語料庫上預訓練後，只需少量標記示例即可針對各種任務進行特定微調。

這些進展的共同主題是學習對同一輸入的不同視圖或變形具有魯棒性的表示；這通常通過最大化同一輸入的視圖之間的相似性，並最小化不同輸入的視圖之間的相似性，通過對比損失實現。然而，生成視圖或損壞的技術迄今為止大多是特定領域的（例如，視覺中的顏色變形（Zhang 等，2016）和裁剪（Chen 等，2020），以及 NLP 中的令牌遮罩（Song 等，2020））。儘管自監督學習的重要性，但在尋找適用於各領域，特別是適用於表格數據的方法方面，工作驚人地少。

在本文中，我們提出 SCARF，一種簡單且多功能的對比預訓練程序。我們通過選擇輸入特徵的一個隨機子集，並用從各自特徵的經驗邊際分佈中隨機抽取的值替換它們，為給定輸入生成一個視圖。在實驗中，我們在 OpenML-CC18 基準測試（Vanschoren 等，2013；Bischl 等，2017；Feurer 等，2019）上測試 SCARF，這是一個包含 72 個真實分類數據集的集合。我們展示，SCARF 預訓

練不僅在完全監督設置中提高了分類精度，在存在標籤噪聲的情況下以及在只有一部分可用訓練數據被標記的半監督設置中也同樣提高了精度。此外，我們展示將 SCARF 預訓練與這些問題的其他解決方案相結合進一步改進了它們，展示了 SCARF 的多功能性及其學習有效任務無關表示的能力。然後，我們進行廣泛的消融研究，展示各種設計選擇的效果和對超參數的穩定性。我們的消融研究表明，SCARF 構建視圖的方式比替代方法更有效。我們展示 SCARF 對特徵縮放不敏感，並且對批量大小、損壞率和 softmax 溫度等各種超參數穩定。

解析與重點：

- 背景問題：標記數據成本高昂，而未標記數據豐富
- 自監督學習在視覺和 NLP 領域已取得顯著成功
    - SimCLR：使用 100 倍少的標籤超越先前最佳性能
    - GPT-3：僅需少量標記示例即可完成特定任務微調
- 對比學習的核心：

1. 最大化同一輸入不同視圖間的相似性
2. 最小化不同輸入視圖間的相似性

- 現存方法局限性：多為領域特定（如圖像裁剪、顏色變形，NLP 的標記遮罩）
- SCARF 創新點：

1. 簡單且多功能的對比預訓練方法
2. 通過替換隨機特徵子集生成數據視圖
3. 適用於表格數據，填補了研究空白

- 實驗證明：

1. 完全監督下提高分類精度
2. 提高標籤噪聲情況下的性能
3. 提高半監督學習的性能
4. 可與現有方法互補

5. 對特徵縮放不敏感，參數穩健性強

2 RELATED WORKS (相關工作)

翻譯： 計算機視覺領域已提出多種自監督學習技術（Zhai 等，2019；Tung 等，2017；Jing & Tian，2020）。一種框架涉及通過各種方法學習基於生成圖像的特徵，包括使用 GAN（Goodfellow 等，2014；Donahue 等，2016；Radford 等，2015；Chen 等，2018），預測像素（Larsson 等，2016），預測著色（Zhang 等，2016；Larsson 等，2017），確保局部和全局一致性（Iizuka 等，2017），以及學習合成偽像（Jenni & Favaro，2018）。與我們方法最相關的是對比學習方法（Tian 等，2019；Hassani & Khasahmadi，2020；Oord 等，2018；Henaff，2020；Li 等，2016；He 等，2020；Bojanowski & Joulin，2017；Wang & Gupta，2015；Gidaris 等，2018）。特別是，我們的框架類似於 SimCLR（Chen 等，2020），其中涉及通過基於圖像的損壞如隨機裁剪、顏色變形和模糊生成單個圖像的視圖；然而，我們生成適用於表格數據的視圖。

自監督學習在語言建模方面產生了特別大的影響（Qiu 等，2020）。一種流行的方法是遮罩語言建模，其中模型被訓練預測被故意遮罩的輸入標記（Devlin 等，2018；Raffel 等，2019；Song 等，2019）以及對此方法的增強（Liu 等，2019；Dong 等，2019；Bao 等，2020；Lample & Conneau，2019；Joshi 等，2020）和涉及置換標記的變體（Yang 等，2019；Song 等，2020）。去噪自編碼器被用於訓練，通過從損壞版本（例如通過標記遮罩、刪除和填充產生）重建輸入（Lewis 等，2019；Wang 等，2019；Freitag & Roy，2018）。對比方法包括隨機替換單詞並區分真實和假短語（Collobert 等，2011；Mnih & Kavukcuoglu，2013），隨機標記替換（Mikolov 等，2013；Clark 等，2020），以及相鄰句子（Joshi 等，2020；Lan 等，2019；de Vries 等，2019）。

在對比學習框架中，損失函數的選擇很重要。InfoNCE（Gutmann & Hyvärinen，2010；Oord 等，2018），可以解釋為表示熵的非參數估計（Wang & Isola，2020），是一個受歡迎的選擇。此後，已有多種提案（Zbontar 等，2021；Grill 等，2020；Hjelm 等，2018）；然而，我們展示 InfoNCE 對我們的框架有效。

最近，Yao 等（2020）以類似我們方法的方式將對比框架適用於大規模推薦系統。主要區別在於方法生成多個視圖的方式。Yao 等（2020）提出以相關方式屏蔽隨機特徵，並對分類特徵應用 dropout，而我們的方法涉及基於各自特徵

的經驗邊際分佈（以非相關方式）隨機化隨機特徵。為任務生成此類視圖是一個困難問題：已有大量工作在理解和設計視圖（Wu 等，2018；Purushwalkam & Gupta，2020；Gontijo-Lopes 等，2020；Lopes 等，2019；Perez & Wang，2017；Park 等，2019）以及學習視圖（Ratner 等，2017；Cubuk 等，2020；Ho 等，2019；Lim 等，2019；Zhang 等，2019b；Tran 等，2017；Tamkin 等，2020）方面。

最後，與我們工作類似的是 VIME（Yoon 等，2020），它為表格數據提出了與我們相同的損壞技術。他們在未標記數據上預訓練編碼器網絡，通過在編碼器狀態之上附加"遮罩估計器"和"特徵估計器"頭，並教導模型從損壞輸入中恢復二進制遮罩和原始未損壞輸入。預訓練的編碼器網絡隨後通過附加特定任務頭和最小化監督損失以及自編碼器重構損失用於半監督學習。VIME 已被證明在基因組學和臨床數據集上達到了最先進的結果。與我們工作的主要區別是我們使用對比損失預訓練，這被證明比部分構成 VIME 的去噪自編碼器損失更有效。此外，預訓練後我們微調所有模型權重，包括編碼器（與 VIME 不同，VIME 只微調任務頭），且我們僅使用任務監督進行微調。

解析與重點：

- **計算機視覺自監督方法**：
    - 基於生成模型（GAN、像素預測、著色預測等）
    - 對比學習（SimCLR 等）-特別相關
    - 視覺對比學習通常使用裁剪、顏色變形等
- **自然語言處理自監督方法**：
    - 遮罩語言建模（如 BERT）
    - 去噪自編碼器
    - 對比方法（詞替換、相鄰句子預測等）
- **對比損失函數**：
    - InfoNCE 是受歡迎的選擇
    - 最近有許多變體被提出
- **最相關工作**：

1. **Yao 等（2020）**：針對推薦系統的對比學習，但使用相關遮罩
2. **VIME（Yoon 等，2020）**：與 SCARF 使用相同的損壞技術，但：
    - VIME 使用自編碼器損失，SCARF 使用對比損失
    - VIME 只微調任務頭，SCARF 微調所有模型權重
    - VIME 結合自編碼器重構損失，SCARF 僅使用任務監督

- **SCARF 的區別**：
    - 使用特徵的經驗邊際分佈進行非相關隨機化
    - 透過對比學習進行預訓練
    - 全面微調所有權重

3 SCARF (方法)

翻譯： 我們現在描述我們提出的方法（算法 1），也在圖 1 中描述。對於未標記訓練數據的每個小批量樣本，我們為每個樣本 x^(i)生成一個損壞版本 x̃^(i)，方法如下。我們均勻隨機抽樣一些比例的特徵，並用從該特徵的經驗邊際分佈中隨機抽取的值替換每個特徵，該經驗邊際分佈定義為在整個訓練數據集中該特徵取值的均勻分佈。然後，我們將 x^(i)和 x̃^(i)都通過編碼器網絡 f 處理，其輸出我們通過預訓練頭網絡 g 處理，得到 z^(i)和 z̃^(i)。注意，預訓練頭網絡對輸出進行 ℓ2-歸一化，使它們位於單位超球面上 - 這在實踐中被證明是至關重要的（Chen 等，2020；Wang & Isola，2020）。我們在 InfoNCE 對比損失上進行訓練，鼓勵所有 i 的 z^(i)和 z̃^(i)接近，而 i≠j 的 z^(i)和 z̃^(j)遠離，並通過 SGD 優化 f 和 g 的參數。

然後，為了通過微調訓練任務分類器，我們採用編碼器網絡 f，並附加一個分類頭 h，該頭以 f 的輸出作為輸入，預測樣本的標籤。我們優化交叉熵分類損失，並調整 f 和 h 的參數。

雖然預訓練可以像正常監督訓練一樣運行預定數量的輪次，但需要多少很大程度上取決於模型和數據集。為此，我們建議在驗證 InfoNCE 損失上使用早停。給定未標記驗證數據，我們循環若干輪次，運行我們提出的方法來生成(x^(i), x̃^(i))對。一旦建立，該靜態集合的損失在預訓練期間被跟踪。典型的損失曲線如附錄所示。

解析與重點：

- **SCARF 核心流程**：
    1. **視圖生成**：
        - 對於每個輸入 x^(i)，選擇一個隨機特徵子集
        - 將選定特徵替換為從相應特徵經驗邊際分佈中抽樣的值，創建損壞視圖 x̃^(i)
    2. **表示學習**：
        - 將原始輸入 x^(i)和損壞輸入 x̃^(i)都通過共享編碼器 f 處理
        - 將編碼器輸出通過預訓練頭網絡 g 處理，得到 z^(i)和 z̃^(i)
        - 這些輸出被 ℓ2-歸一化到單位超球面上
    3. **對比學習**：
        - 使用 InfoNCE 損失：鼓勵同一輸入的不同視圖（正對）接近
        - 同時使不同輸入的視圖（負對）遠離
        - 通過 SGD 優化 f 和 g 的參數
    4. **微調階段**：
        - 取預訓練的編碼器 f
        - 附加一個分類頭 h
        - 使用有標籤數據優化交叉熵分類損失
        - 同時調整 f 和 h 的參數
- **早停策略**：
    - 在驗證數據上監控 InfoNCE 損失
    - 建立靜態(x^(i), x̃^(i))對集合用於驗證
    - 當驗證損失不再下降時停止預訓練
- **算法結構**：

- 預訓練階段：學習強大的特徵表示
- 微調階段：利用這些表示進行特定任務訓練
- 與典型的自監督學習流程一致，但專為表格數據設計

4 EXPERIMENTS (實驗)

翻譯： 我們評估 SCARF 預訓練在三種不同設置中監督微調後對測試準確率的影響：在完整數據集上，在完整數據集上但只有 25%的樣本有標籤而其餘 75%沒有，以及在完整數據集上但 30%的樣本經歷標籤噪聲。

數據集：我們使用來自公共 OpenML-CC18 基準測試的 69 個數據集，該基準測試具有 CC-BY 許可證。它由 72 個為有效基準測試手動策劃的真實世界分類數據集組成。由於我們關注的是表格數據集，我們移除了 MNIST、FashionMNIST 和 CIFAR10。對於每個 OpenML 數據集，我們形成 70%/10%/20%的訓練/驗證/測試分割，每次試驗生成不同的分割，且所有方法使用相同的分割。用於驗證和測試的百分比從未改變，只有訓練標籤在標籤噪聲實驗中被損壞。

數據集預處理：我們通過一熱編碼表示分類特徵，大多數在消融中探索的損壞方法都是在這些數據的一熱編碼表示上（除了 SCARF，其中邊際採樣在一熱編碼之前完成）。我們預處理缺失數據如下：如果特徵列總是缺失，我們將其丟棄。否則，如果特徵是分類的，我們用在整個數據集上計算的眾數或最頻繁類別填充缺失條目。對於數值特徵，我們用均值替代。我們通過 z-score、min-max 和 mean 規範化探索數值特徵的縮放。我們發現對於普通網絡（即控制組），z-score 規範化在除了三個數據集（OpenML 數據集 id 4134、28 和 1468）外表現最佳，這三個數據集以不進行縮放為最佳。因此，我們不對這三個數據集進行縮放，而對所有其他數據集進行 z-score 規範化。

模型架構與訓練：除非另有指定，我們在所有實驗中使用以下設置。如前所述，我們將神經網絡分解為編碼器 f、預訓練頭 g 和分類頭 h，其中 g 和 h 的輸入是 f 的輸出。我們選擇所有三個部件模型為隱藏維度為 256 的 ReLU 網絡。f 由 4 層組成，而 g 和 h 都有 2 層。SCARF 和自編碼器基線都使用 g（用於預訓練和共同訓練，後面描述），但對於自編碼器，輸出維度是輸入特徵維度，並應用均方誤差重構損失。我們使用 Adam 優化器訓練所有模型及其組件，使用默認學習率 0.001。對於預訓練和微調，我們使用 128 的批量大小。無監督預訓練方法都在驗證損失上使用耐心為 3 的早停，除非另有說明。監督微調使用相

同的標準（和驗證分割），但分類錯誤用作驗證早停指標，因為它表現稍好。我們設置微調最大輪次為 200，預訓練輪次為 1000，我們使用 10 輪次建立靜態驗證集。除非另有說明，我們對基於 SCARF 的方法使用損壞率 c 為 0.6 和溫度 τ 為 1。所有運行重複 30 次，使用不同的訓練/驗證/測試分割。實驗在 CPU 雲集群上運行，總共使用了約一百萬 CPU 核心小時。

評估方法：我們使用以下方法有效地傳達所有數據集的結果。

勝率矩陣：給定 M 個方法，我們計算大小為 M × M 的"勝率"矩陣 W，其中(i, j) 項定義為： $W_{i,j} = \frac{\sum_{d=1}^{69} 1[\text{方法 i 在數據集 d 上勝過 j}]}{\sum_{d=1}^{69} 1[\text{方法 i 在數據集 d 上勝過 j}] + 1[\text{方法 i 在數據集 d 上輸給 j}]}$

"勝過"和"輸給"僅在均值不是統計平手時定義（使用 Welch 的 t 檢驗，假設方差不等，p 值為 0.05）。勝率為 0/1 意味著在 69 次（成對）比較中，只有一次是顯著的，並且是一次失敗。因為 0/1 和 0/69 有相同的值，但後者更確信 i 比 j 差，我們以分數形式呈現值並使用熱圖。我們在矩陣中添加一列，代表每行的最小勝率。

箱形圖：勝率矩陣有效地傳達一個方法比另一個頻繁勝出的程度，但不能捕捉程度差異。為此，對於每種方法，我們計算相對於某個參考方法在每個數據集上的相對百分比改進。然後我們建立箱形圖，描繪跨數據集的相對改進分佈，將觀察結果繪製為小點。我們只考慮方法和參考的均值差異達到 p 值 0.20 的數據集。我們在此處使用比計算勝率時更大的 p 值，因為否則某些方法將沒有足夠的點來使箱形圖有意義。

基線：我們使用以下基線：

- **標籤平滑**（Szegedy 等，2016），被證明對準確率（Müller 等，2019）和標籤噪聲（Lukasik 等，2020）有效。我們在平滑項上使用 0.1 的權重。
- **Dropout**。我們使用標準 dropout（Srivastava 等，2014），在所有層上使用 0.04 的比率。Dropout 已被證明能改善性能並對標籤噪聲具有魯棒性（Rusiecki，2020）。
- **Mixup**（Zhang 等，2017），使用 α = 0.2。
- **自編碼器**（Rumelhart 等，1985）。我們用它作為我們的主要消融預訓練基線。我們使用經典自編碼器（"無噪聲 AE"），使用高斯加性噪聲的去

噪自編碼器（Vincent 等，2008；2010）（"加噪聲 AE"），以及 SCARF 的損壞方法（"SCARF AE"）。我們使用 MSE 作為重構損失。我們嘗試了預訓練和與監督任務共同訓練，當共同訓練時，我們將 0.1 倍的自編碼器損失添加到監督目標中。我們在附錄中討論共同訓練，因為它不如預訓練有效。

- **SCARF 數據增強**。為了隔離我們提出的特徵損壞技術的效果，我們跳過預訓練，而是在監督微調期間對損壞的輸入進行訓練。我們在附錄中討論這個基線的結果，因為它不如其他基線有效。
- **判別性 SCARF**。這裡，我們的預訓練目標是區分原始輸入特徵和使用我們提出的技術損壞的特徵。為此，我們更新預訓練頭網絡以包括最終線性投影，並用二元邏輯損失替換 InfoNCE。我們使用分類錯誤而非邏輯損失作為早停的驗證指標，因為我們發現它表現稍好。
- **自蒸餾**（Hinton 等，2015；Zhang 等，2019a）。我們首先在標記數據上訓練模型，然後在標記和未標記數據上訓練最終模型，使用第一個模型的預測作為兩個數據集的軟標籤。
- **Deep k-NN**（Bahri 等，2020），一種最近提出的標籤噪聲方法。我們設置 k = 50。
- **Bi-tempered 損失**（Amid 等，2019），一種最近提出的標籤噪聲方法。我們使用 5 次迭代，t1 = 0.8，t2 = 1.2。
- **自訓練**（Yarowsky，1995；McClosky 等，2006）。一種經典的半監督方法 - 每次迭代，我們在偽標記數據上訓練（初始化為原始標記數據集），並將高置信度預測添加到訓練集中，使用預測作為標籤。然後我們在最終數據集上訓練我們的最終模型。我們使用 0.75 的 softmax 預測閾值，運行 10 次迭代。
- **三重訓練**（Zhou & Li，2005）。類似於自訓練，但使用三個通過引導採樣具有不同初始標記數據的模型。每次迭代，每個模型的訓練集通過只添加其他兩個模型預測一致的未標記點更新。它在現代半監督 NLP 任務中被證明具有競爭力（Ruder & Plank，2018）。我們使用與自訓練相同的超參數。

4.1 SCARF 預訓練提高預測性能

翻譯： 圖 2 顯示了我們的結果。從第一個勝率矩陣圖中，我們看到所有五種考慮的預訓練技術都比無預訓練（控制組）有所改進，並且 SCARF 優於其他技術，有更多統計顯著的勝利。第二個勝率矩陣顯示，SCARF 預訓練提升了 mixup、標籤平滑、蒸餾和 dropout 的性能，且比替代方法做得更好。換句話說，預訓練補充了現有解決方案，表明這裡有一個不同的機制在起作用。箱形圖擴展了第二個勝率矩陣，顯示每種預訓練策略相對於基線的相對改進。

表 1 總結了箱形圖通過平均相對增益。我們觀察到 SCARF 一般優於替代方案，並在整體上增加了 1-2%的相對增益。

4.2 SCARF 預訓練提高了標籤噪聲存在時的性能

翻譯： 為了展示預訓練如何在數據標籤不可靠時提高模型魯棒性，我們進行如下操作。首先，標籤噪聲魯棒性通常在兩種不同的設置中研究-(1)當訓練數據的某個子集保證未損壞且該集合預先已知，(2)當整個數據集都不可信任。為簡單起見，我們在實驗中考慮設置 2。我們損壞標籤如下：保持驗證和測試分割不變，我們選擇 30%的訓練數據進行隨機損壞，並對每個數據點，我們將其標籤替換為某個類別，均勻分布於所有類別（包括數據點的真實類別）。結果如圖 3 和表 1 所示。再次，我們看到 SCARF 表現優於其他方法，並通過 2-3%提高所有基線性能。

4.3 SCARF 預訓練在標記數據有限時提高性能

翻譯： 為了展示當未標記數據比標記數據更多時預訓練如何幫助，我們移除訓練分割中的標籤，使得只有原始分割的 25%保持標記。自編碼器、SCARF、自訓練和三重訓練都利用剩餘的未標記部分。結果如圖 4（附錄）和表 1 所示。SCARF 優於其他方法，為所有基線增加了非常可觀的 2-4%。

4.4 消融實驗

翻譯： 我們現在詳細說明 SCARF 中每個因素的重要性。我們在這裡只展示部分結果；其餘部分在附錄中。

其他損壞策略效果較差且對特徵縮放更敏感。 在這裡，我們消融了我們提出的邊際採樣損壞技術，用以下其他有希望的策略替換它，同時保持其他條件不變。

1. **無損壞**。我們不應用任何損壞 - 即 $\tilde{x}^{(i)} = x^{(i)}$，在算法 1 中。在這種情況下，正對之間的餘弦相似度總是為 1，模型學習使負對盡可能正交。

根據最近的觀點（Wang & Isola, 2020），對比損失包含兩個項 - 一個鼓勵同一個例子的視圖對齊 - 另一個鼓勵超球面嵌入均勻分布 - 我們看到沒有損壞時，預訓練可能只是學習將輸入例子均勻嵌入到超球面上。

2. **均值損壞**。確定要損壞的特徵後，我們用經驗邊際分布的均值替換它們的條目。
3. **加性高斯噪聲**。我們將 i.i.d $N(0, 0.5^2)$噪聲添加到特徵中。
4. **聯合採樣**。與其通過從其邊際隨機抽樣來替換特徵形成 $\tilde{x}^{(i)}$，我們隨機從訓練數據 X 中抽取 $\hat{x}^{(i)}$ - 即從經驗（聯合）數據分布中抽樣 - 然後設置 $\tilde{x}^{(i)}_j = \hat{x}^{(i)}_j$，$\forall j \in I_i$。
5. **缺失特徵損壞**。我們將選定的特徵標記為"缺失"，並為我們的模型添加每個特徵維度一個可學習的值。當特徵缺失時，它被相應的可學習缺失值填充。
6. **特徵 dropout**。我們將選定的特徵置零。

我們也在以下方式縮放輸入特徵的情況下檢驗損壞策略。

1. **Z-score 縮放**。這裡，x_j = [x_j - mean(X_j)]/std(X_j)。
2. **Min-max 縮放**。x_j = [x_j - min(X_j)]/[max(X_j) - min(X_j)]。
3. **均值縮放**。x_j = [x - mean(X_j)]/[max(X_j) - min(X_j)]。

圖 6（附錄）顯示了 z-score 和 min-max 縮放的結果。SCARF 邊際採樣一般優於不同類型特徵縮放的其他損壞策略。邊際採樣的優點是除了沒有超參數外，它還對縮放不變並保留了每個特徵的"單位"。相比之下，即使是簡單的乘法縮放也需要以相同方式縮放加性噪聲。

SCARF 對批量大小不敏感。 像 SimCLR（Chen 等，2020）這樣的對比方法在增加批量大小 N 時顯示持續改進。批量大小和對比學習任務難度之間存在緊密耦合，因為在我們的情況下，每個例子 i 的損失項涉及 1 個正對和 N-1 個負對。對大型（例如 5000）批量大小的需求已經激發了支持它們的工程解決方案（Gao & Zhang, 2021），並被視為採用其他損失函數的理由（Zbontar 等, 2021）。圖 5（附錄）比較了一系列批量大小。我們看到將批量大小增加到 128 以上沒有帶來顯著改進。這裡的一個合理假設是，更高容量的模型和更難的任務從負例中受益更多。

SCARF 對損壞率和溫度相當不敏感。 我們在圖 5（附錄）中研究了損壞率 c 的影響。我們看到當比率在 50%-80%範圍內時，性能是穩定的。因此，我們建議默認設置為 60%。我們看到溫度超參數也有類似的穩定性（見附錄）。我們建議使用默認溫度為 1。

對損壞的微調效果不比原方法好。 附錄顯示了四種不同損壞方法微調的效果。我們沒有看到使用它們的任何理由。

InfoNCE 的替代方案效果不比原方法好。 我們調查了 InfoNCE 損失函數選擇的重要性，並查看了用最近提出的替代方案 Alignment and Uniformity（Wang & Isola, 2020）和 Barlow Twins（Zbontar 等, 2021）替換它的效果。我們發現在我們的設置中，這些替代損失幾乎匹配或比原始且流行的 InfoNCE 表現更差。詳細信息見附錄。

解析與重點：

- **多類環境下的實驗結果**：
    1. **SCARF 優於所有對比方法：**
        - 比控制組和其他預訓練方法有顯著提升
        - 平均提高了 1-2%的相對精度
    2. **標籤噪聲環境中的表現**：
        - 30%的隨機標籤損壞模型下，SCARF 提供 2-3%的改進
        - 優於專門針對標籤噪聲設計的方法
    3. **半監督情境下的優勢**：
        - 只有 25%數據有標籤時，SCARF 提供 2-4%的提升
        - 超越自訓練、三重訓練等傳統半監督方法
    4. **全面消融研究**：
        - **損壞策略對比**：邊際採樣優於其他方法（無損壞、均值損壞、高斯噪聲等）
        - **特徵縮放不敏感性**：在各種縮放方法下都表現穩定
        - **批量大小穩定性**：128 以上批量大小不再有顯著提升

        - **損壞率穩健性**：在 50%-80%範圍內性能穩定，推薦 60%
        - **InfoNCE 損失優勢**：優於 Alignment and Uniformity 和 Barlow Twins 等替代方案
    5. **互補性**：
        - SCARF 可有效補充現有方法如 mixup、dropout 等
        - 組合使用能獲得累積效果

- **核心優勢總結**：
    1. 簡單且廣泛適用
    2. 對超參數設置不敏感（批量大小、損壞率、溫度等）
    3. 對特徵縮放方法的選擇魯棒
    4. 在三種關鍵場景（全監督、半監督、標籤噪聲）下都表現優異
    5. 能與其他方法互補，提供額外提升

5 CONCLUSION (結論)

翻譯： 自監督學習在包括計算機視覺和自然語言處理在內的重要領域取得了深遠的成功，但在一般表格設置中的進展很少。我們提出了一種簡單且多功能的自監督學習方法，能夠學習在下游分類任務中有效的表示，即使在有限的標記數據或存在標籤噪聲的情況下也是如此。該方法可能的負面副作用包括學習強化輸入數據中出現的偏見的表示。在訓練過程中找到減輕這一點的方法是未來研究的潛在方向。

解析與重點：

- **研究貢獻**：
    1. 提出了一種專為表格數據設計的簡單自監督學習方法
    2. 證明了該方法在各種挑戰場景中的有效性
    3. 填補了表格數據自監督學習的研究空白
- **實際意義**：
    1. 提供了一種在標籤資源有限情況下提高模型性能的策略

2. 增強了模型對標籤噪聲的魯棒性
3. 與現有方法互補，可用於進一步改進模型

- **限制與未來方向**：
  - 可能強化數據中存在的偏見
  - 未來研究方向：尋找在訓練過程中減輕潛在偏見的方法