Trompt 論文詳細解釋

我將針對這篇「Trompt: Towards a Better Deep Neural Network for Tabular Data」論文的主要部分進行詳細解釋，包括導言、相關工作、Trompt 模型架構、討論和結論。

1. Introduction（導言）

這部分介紹了表格數據的重要性和深度學習在表格數據上的挑戰。

表格數據的重要性

表格數據在現實世界中扮演著至關重要的角色，應用於多種實際領域：

- 金融領域：銀行用財務報表評估公司信用度
- 醫療領域：醫生利用診斷報告確定病因
- 電子商務：平台利用客戶記錄發現潛在興趣

表格數據一般由異質特徵（不同類型的特徵）組成，具有廣泛的實用價值。

深度學習的成就與挑戰

深度學習在多個領域取得了巨大成功，包括：

- 計算機視覺
- 自然語言處理
- 機器人技術

深度學習的端到端優化優勢包括：

1. 使用流數據進行在線學習
2. 多模態整合（結合圖像和文本等不同類型輸入）
3. 表示學習（實現半監督學習和生成建模）

然而，在表格數據領域，深度神經網絡仍然落後於基於樹的模型。研究人員嘗試通過以下方式改進表格數據的深度學習：

1. 使用 Transformer 架構
2. 研究歸納偏置（幫助模型更好地學習表格數據結構）

Trompt 的提出

論文提出了一種新的架構 Trompt（Tabular Prompt），靈感來自語言模型的提示學習。Trompt 將表格數據的學習策略分為兩部分：

1. 表格的內在信息（列的固有性質）
2. 樣本間變化的信息（不同樣本的特徵重要性）

Trompt 在 Grinsztajn45 基準測試上評估，結果表明它優於最先進的深度神經網絡，並與基於樹的模型相當。

2. Related Work（相關工作）

相關工作部分主要涵蓋了三個關鍵領域：提示學習、表格神經網絡和 Trompt 的獨特性。

2.1 提示學習（Prompt Learning）

提示學習的目的是將下游任務的輸入和輸出轉換為構建預訓練模型時使用的原始任務格式。與微調不同，提示學習不需要更改模型權重，而是通過提示來引導模型。

提示學習的特點：

- 使用少量數據甚至零樣本學習就能達到良好效果
- 大大提高了預訓練模型的應用多樣性
- 提示可以是離散的（由自然語言詞彙組成）或軟的（學習的表示）

例如，在自然語言處理中，可以在句子前插入特定任務的提示，引導模型調整其響應以適應不同任務。

2.2 表格神經網絡（Tabular Neural Network）

這一小節介紹了兩種表格神經網絡的研究方向：

Transformer 架構

自從 2017 年自注意力機制革新了自然語言處理後，它很快被其他領域採用。在表格數據領域：

- TabTransformer：首個基於 Transformer 的表格神經網絡，但只將分類特

徵輸入到 Transformer 塊

- FT-Transformer：通過同時輸入分類和數值特徵到 Transformer 塊解決了上述問題
- SAINT：進一步改進，不僅在特徵維度上應用注意力，還在樣本維度上應用

歸納偏置研究

歸納偏置指模型對特定數據結構的先驗假設，有助於學習：

- CNN 在圖像上表現良好，因為卷積核設計用於捕獲局部模式
- RNN 在語言理解中廣泛使用，因為它能很好地封裝詞之間的因果關係

然而，表格數據的歸納偏置尚未被充分發現。鑒於基於樹的模型一直是表格數據的最佳選擇，兩個模型提出了假設：

- Net-DNF：理論證明決策樹等同於某些析取範式，提出了析取神經範式
- TabNet：利用順序注意力和稀疏正則化模擬樹模型的學習策略

2.3 Trompt 的獨特性

作者認為表格數據的列重要性對所有樣本並非不變，可以分組為多種模態。Trompt 的獨特性在於：

- 首個受提示啟發的表格神經網絡
- 與基於 Transformer 的模型相比，學習分離的列重要性而非專注於列之間的交互
- 與 TabNet 和 Net-DNF 相比，通過模擬提示學習而非決策樹的分支分割來處理多模態性

3. Trompt（模型架構）

Trompt 部分詳細介紹了模型的架構設計。這是論文的核心內容。

整體架構

如圖 2 所示，Trompt 由多個 Trompt Cell 和一個共享的 Trompt Downstream 組成：

- 每個 Trompt Cell 負責特徵提取和提供多樣化表示
- Trompt Downstream 用於預測

3.1 Trompt Cell

Trompt Cell 的架構如圖 3 所示，可分為三部分：

3.1.1 導出特徵重要性

第一部分基於列嵌入(Ecolumn)、前一個單元的輸出(Oprev)和提示嵌入(Eprompt)導出特徵重要性(Mimportance)。

關鍵步驟包括：

1. 提示嵌入與前一個單元的輸出融合（使用堆疊、連接和相加操作）
2. 列嵌入擴展（堆疊操作）
3. 透過矩陣乘法和 softmax 導出特徵重要性

這裡的設計使得特徵重要性能夠對不同的樣本進行調整，體現了提示學習的思想。

3.1.2 構建和擴展特徵嵌入

第二部分將分類特徵通過嵌入層處理，將數值特徵通過密集層處理，構建特徵嵌入(Efeature)。

第三部分通過密集層將特徵嵌入擴展為能夠容納 P 個提示的形式(Êfeature)。這樣設計使得每個提示可以有不同的特徵嵌入表示。

3.1.3 生成輸出

Trompt Cell 的輸出是 Êfeature 和 Mimportance 的元素相乘後按列求和的結果。這種設計使得模型能夠根據不同的重要性權重加權不同的特徵。

3.2 Trompt Downstream

Trompt Downstream 基於 Trompt Cell 的輸出進行預測：

1. 首先通過密集層和 softmax 激活函數導出每個提示的權重
2. 計算加權和

3. 通過兩個密集層進行最終預測

在訓練過程中，每個預測的損失單獨計算並相加以更新模型權重。在推理過程中，所有單元的預測簡單平均作為最終預測。

3.3 Trompt 的提示學習

這部分解釋了 Trompt 架構設計的基本原理及其與自然語言處理中提示學習的類比。

表格數據的特點：

- 結構化數據，每一列代表特定的數據屬性
- 基於樹的模型成功的關鍵是對個別樣本分配特徵重要性

Trompt 結合了列的內在屬性和樣本特定的特徵重要性，使用提示學習啟發的架構：

- 列嵌入表示每列的內在屬性
- 提示嵌入用於提示列嵌入，為給定提示生成特徵重要性
- 在提示列嵌入之前，提示嵌入與前一個 Trompt Cell 的輸出融合，使得輸入相關的表示能夠流動

表 1 概念性地類比了 Trompt 的提示學習與自然語言處理中的提示學習：

- 樣本不變的內在屬性（Ecolumn）類似於固定的大型語言模型
- 樣本特定的特徵重要性（Mimportance）類似於任務特定的預測

4. 討論（Discussion）

討論部分進一步探討了 Trompt 的"提示"機制以及 Trompt 與基於樹的模型的差異。

5.1 Trompt 中"提示"機制的進一步探討

Trompt 的"提示"機制通過式(4)實現，涉及擴展的提示嵌入和擴展的列嵌入的轉置的矩陣乘法。這種計算方式能計算基於餘弦的距離，偏好樣本特定表示和樣本不變內在屬性之間的高相似度。

與自注意力不同：

- 自注意力計算查詢和鍵之間的距離，導出詞元到詞元的相似度
- Trompt 計算 SEˆprompt 和 SEcolumn 之間的距離，導出樣本對內在屬性的相似度

這種計算的基本思想是捕捉每個樣本與表格數據集內在屬性之間的距離，假設將內在屬性納入表格神經網絡的建模有助於做出良好的預測。

5.2 Trompt 與基於樹的模型的差異

雖然 Trompt 和基於樹的模型都能實現樣本依賴的特徵重要性，但它們之間存在兩個主要差異：

1. 列信息的整合方式：
    - Trompt 使用列嵌入在樣本間共享列信息
    - 基於樹的模型在節點分割的本質中學習列信息
2. 學習特徵重要性的技術：
    - Trompt 通過提示學習顯式導出特徵重要性
    - 基於樹的模型在根到葉的路徑中隱式改變特徵重要性

6. 結論（Conclusion）

結論部分總結了 Trompt 的貢獻和未來研究方向。

主要貢獻：

- 引入 Trompt，一種用於表格數據分析的新型網絡架構
- Trompt 利用提示學習確定個別樣本中不同的特徵重要性
- 評估表明 Trompt 優於最先進的深度神經網絡（SAINT 和 FT-Transformer），縮小了深度神經網絡與基於樹的模型之間的性能差距

論文指出深度學習中提示學習的出現是有希望的。雖然 Trompt 的設計可能不直觀或不完全符合語言模型提示，但它展示了在表格數據分析中利用提示的潛力。

這項工作為深度神經網絡挑戰基於樹的模型提供了一種新策略，未來研究可以探索更多受提示啟發的架構。

通過這種創新架構，Trompt 在表格數據分析領域邁出了重要一步，為結合提示學習和表格數據處理開闢了新的研究方向。