AUTOG: TOWARDS AUTOMATIC GRAPH CONSTRUCTION FROM TABULAR DATA

統整

Abstract (摘要):

這篇論文關注圖機器學習(GML)領域中一個重要但經常被忽視的問題 - 如何從表格數據中構建合適的圖結構。主要貢獻有兩個:

- 提出了一套用於評估圖構建方法的數據集
- 提出了一個基於大語言模型(LLM)的解決方案 AutoG,能夠自動生成高質量的圖架構

Introduction (引言):

引言部分說明了研究背景和動機:
GML 在各個領域都有廣泛應用,如生命科學、電商、社交網絡等
目前的研究主要集中在開發強大的模型,而忽視了如何從原始數據構建圖的問題
從表格數據構建圖存在兩個主要挑戰:

- 缺乏專門的數據集來評估圖構建方法的有效性
- 現有的自動構建方法只能應用於特定場景

Preliminaries (預備知識):

這部分介紹了一些基本概念:
表格數據的定義:包含表名、列名、數據類型等元信息
表結構描述(Schema):使用 YAML 格式存儲表的元信息
從表格數據到圖的轉換:如何將表格數據轉換為異構圖

Design Dataset for Evaluating Graph Construction (設計評估數據集):

提出了 4 個核心挑戰來設計評估數據集:

- C1: 識別非主外鍵關係的邊
- C2: 從單個表中擴充多個節點或邊的類型
- C3: 將表轉換為合適的節點或邊類型
- C4: 為不同下游任務生成合適的圖

Method (方法):

提出了 AutoG 框架,主要包含:

- LLM 作為生成器:生成候選圖結構
- 啟發式圖構建器:將表轉換為圖
- Oracle 作為判別器:評估生成圖的質量

Experimental Results (實驗結果):

在 8 個數據集上進行了全面評估
與多個基線方法比較,表明 AutoG 能生成接近人類專家水平的圖
進行了深入的消融實驗分析各個組件的作用

Related Work (相關工作):

主要介紹了三個相關領域:

- 圖機器學習在表格數據上的應用
- 使用 LLM 進行自動化數據科學
- 異構圖學習

Conclusion (結論):

總結了 AutoG 框架的主要貢獻
指出了兩個主要限制:

- 僅適用於已包含關係信息的數據集
- 過於依賴語義信息
  提出未來可以探索的方向
  這篇論文的主要創新點在於:
- 首次系統性地研究了從表格數據構建圖的問題
- 提出了評估框架和基準數據集
- 開發了基於 LLM 的自動化解決方案

分段翻譯

Abstract (摘要):

近年來,圖機器學習(GML)取得了重大進展,其應用範圍涵蓋眾多領域。然而,GML 的研究重點主要集中在開發強大的模型上,常常忽視了一個關鍵的初始步驟:從常見的數據格式(如表格數據)中構建合適的圖。
這個構建過程是應用基於圖的模型的基礎,但它仍然缺乏深入研究和形式化。我們的研究旨在通過將圖構建問題形式化並提出有效的解決方案來填補這一空白。
我們識別出實現這一目標的兩個關鍵挑戰:
缺乏專門用於形式化和評估圖構建方法有效性的數據集
現有的自動構建方法只能應用於某些特定情況,而生成高質量的圖需要繁瑣的人工工程
為了應對這些挑戰,我們提出了兩方面的貢獻:
引入一組數據集來形式化和評估圖構建方法
提出一個基於 LLM 的解決方案 AutoG,可以自動生成高質量的圖架構,無需人工干預
實驗結果表明,構建的圖的質量對下游任務的性能至關重要,而 AutoG 能夠生成可與人類專家產出媲美的高質量圖。

Introduction (引言):

圖機器學習(GML)因其在生命科學、電子商務和社交網絡等多個領域的廣泛應用而備受關注。GML 通常涉及應用圖神經網絡(GNNs)等模型來利用給定任務的底層圖結構,例如使用好友網絡進行用戶推薦和識別新的藥物相互作用。
儘管 GML 受到廣泛關注並快速發展,但從工業表格數據等常見數據格式構建圖仍是一個未被充分探索的主題。這主要源於一個被廣泛接受的假設:認為下游任務已有適當的圖數據集,就像已建立的基準數據集一樣。然而,在許多實際的企業場景中,現成的圖數據集並不存在。
首先,對於以表格等常見存儲格式存在的輸入數據,可以在其上定義多種合理的圖架構和結構。圖架構的選擇會影響 GML 的下游性能。其次,將源數據轉換為圖格式需要專業的數據工程和處理。
儘管基於 GNN 的方法在 Kaggle 排行榜上表現出色,但將原始表格數據轉換為可供 GML 使用的圖需要繁瑣的預處理和專門的技能。
本研究的目標是通過建立真實世界的數據集來形式化圖構建中的挑戰,並實現從輸入表格數據自動構建圖。現有的表格圖數據集(如 Wang 等人(2024b)和 Fey 等

人(2024))假設已有格式良好的圖,包含明確的關係,例如完整的外鍵和主鍵對。
在這些情況下,可以使用諸如 Row2Node 這樣的啟發式方法輕鬆構建圖,方法是將每個表轉換為一個節點類型。然而,在現實場景中也廣泛存在隱式關係,如具有相似語義的列或具有分類類型的列,這些無法通過啟發式方法解決。
為評估圖構建而設計的數據集應該反映建模隱式關係的重要性。此外,可以基於同一數據集定義不同的任務。而且,不同的圖構建方式如何影響不同任務的性能這一問題尚未得到充分研究。
因此,理想的圖構建數據集還應該包含不同的下游任務以反映這個問題。從解決方案的角度來看,圖構建涉及在所有可能的圖結構中找到最佳候選者。但是,考慮到巨大的搜索空間,通過窮舉搜索來找到圖結構是不可行的。
因此,一個有效的自動圖構建方法應該能夠從眾多可能的圖結構/架構中高效地識別出高質量的候選者。
為了應對上述挑戰,我們提出了一套評估數據集和基於大語言模型(LLM)的圖構建解決方案。我們首先從 Kaggle、Codalab 和其他數據源提取原始表格數據集,以設計能反映真實世界圖構建挑戰的數據集。
這些數據集與之前的工作不同,它們未經過專家處理,現有的圖構建方法在這些數據集上表現較差。為了解決圖構建問題,我們提出了基於 LLM 的自動圖構建解決方案 AutoG,其靈感來自 LLM 在代理任務規劃和表格數據處理方面的推理能力。然而,我們觀察到 LLM 傾向於生成無效的圖或關係較少的圖。我們通過引導 LLM 進行封閉式函數調用來解決這個問題。具體來說,我們將圖結構的生成分解為四個應用於表格數據的基本轉換:
建立兩列之間的鍵關係
擴展特定列
基於列生成新表
操作主鍵
結合每個動作的思維鏈提示示例,AutoG 生成一系列動作來獲得增強的架構並構建圖。為了進一步提高生成質量,它將採用訓練好的 GML 模型的早期驗證性能作為 oracle 來高效選擇結果。
我們的主要貢獻可以總結如下:
a) 用一組數據集形式化圖構建問題:我們創建了一組涵蓋不同圖構建挑戰的數據集,包括來自學術、電子商務、醫療等領域的八個數據集。
b) 基於 LLM 的自動圖構建方法 AutoG:為了在無需手動數據工程的情況下解決圖構建問題,我們提出了一個基於 LLM 的基線方法,可以自動生成圖候選項並高

效地選擇最佳候選項。
c) 全面評估:我們在提出的數據集上將 AutoG 與不同的基線方法進行了比較。AutoG 展現出接近數據工程專家水平的良好性能。在 12 個測試任務中,它在 9 個下游任務上達到了人類專家設計提示性能的 98.5%。

Preliminaries (預備知識):

輸入表格數據使用 RDB 語言表示為架構文件。隨後,我們介紹表架構以及如何使用它們來描述圖。我們首先介紹 RDB 語言的基本元素。
定義. 表格數據 D 包含 K 個表的數組 D := {Ti}K(i=1)。每個表 Ti 可以被視為一個集合 Ti = (Ci, Ri, Mi),其中:
Ci = (Ci,1, ..., Ci,li)是一個表示列名的字符串數組,li 表示表 Ti 中的列數。
Ri 是一個矩陣,其中每一行 Ri,j = (Ri,j,1, ..., Ri,j,li)包含表 Ti 的第 j 行的值。
Mi = (Mi,1, ..., Mi,li)是一個指定每列數據類型的數組。
在本文中,我們考慮以下數據類型{category(類別), numeric(數值), text(文本), primary_key(PK 主鍵), foreign_key(FK 外鍵), set(集合), timestamp(時間戳)}。例如,如果 Mi,1 = text,則同一列中的所有值 Ri,1,1, · · · , Ri,mi,1 都是文本類型(mi 指表 Ti 的行數)。每種數據類型的詳細描述可以在附錄 A.1 中找到。
上述定義關注單個表的屬性。對於多個表(K > 1),它們可以通過一組 n 個 PK-FK 對{xmPK, ymPK, xmFK, ymFK}相關聯,其中 m = 1, ..., M。x 和 y 分別表示表 D 中的索引和列的索引。在實際場景中,通常只有部分 PK-FK 關係是顯式的。必須手動識別其他隱式連接才能很好地支持下游任務。
表架構和圖架構描述。基於這種語言,我們通過以 YAML 等結構化格式存儲所有元信息來定義表架構。附錄 A.2 中展示了一個示例。表架構按照 RDB 語言以結構化方式定義表的元信息。
圖架構是表架構的一種特殊類型。與一般的表架構相比,圖架構呈現的表具有適當的列設計和 PK-FK 關係。這些特徵使得將圖架構(如 2.2 節所述)轉換為適合下游任務的理想圖結構變得簡單。

基於表格數據的定義,圖構建的目標是將關係型表格數據 D 轉換為圖 G。參照 Fey 等人(2024)和 Wang 等人(2024b)的工作,我們將 G 視為由節點集 V 和邊集 E 特徵化的異構圖。
節點和邊的組織方式為 V = ∪v∈V Vv 且 E = ∪e∈E Ee,其中 Vv 表示類型 v 的節點集,Ee 表示類型 e 的邊集。圖構建的主要挑戰在於從表格數據的架構中提取適當的節點類型和邊類型。

如果我們將每個表作為一個節點類型,將每個 PK-FK 關係作為一個邊類型,這個過程可能很直觀。然而,這種方法可能會為一般的表架構生成次優的圖。例如,當兩個實體放在同一個表中時,一個實體可能被視為另一個實體的特徵,導致生成的圖無法有效反映結構關係,從而影響下游任務的性能。

Design Dataset for Evaluating Graph Construction (設計評估數據集):

為了使圖構建問題具體化並提供一組用於比較不同方法的數據集,我們首先確定了在圖構建過程中需要解決的關鍵問題,這些可以被視為設計空間。基於這些問題,我們從不同領域精心選擇了 8 個多表數據集來構建用於圖構建的數據集。
我們提出在將表格數據轉換為圖時需要解決的四個核心挑戰。這些挑戰的示例在圖 1 中展示。
C1: 識別非 PK-FK 關係中的邊:像 Row2Node 這樣的傳統方法只將 PK-FK 關係轉換為邊,而這些關係通常是不完整的,這就需要自動連接發現或人工干預。
C2: 從一個表中增強多個節點或邊類型:多個節點類型和邊類型可能不恰當地放在一個表中。例如,圖 1 中的"Field"列可以誘導出有用的關係,因此應該添加一個增強表。
C3: 將表轉換為適當的節點或邊類型:如何將表轉換為適當的類型會影響下游任務性能和生成圖的有效性。例如,圖 1 中的"Ratings"表應該更好地建模為邊類型,因為它是關於預測用戶和電影類型之間的屬性。
C4: 為不同的下游任務生成適當的圖:考慮到可以基於同一個表格數據定義多個任務,單一的圖設計可能無法適用於所有任務。這個論點尚未得到充分研究,將在實驗中進行驗證。
這些挑戰的設計理念。這四個挑戰受到現有工作的啟發,但超出了它們的範圍。具體來說,C1 是數據湖和關係型數據庫中自動數據工程的常見問題。當構建圖是最終目標時,可連接列的檢測變得更加重要,因為找到關係至關重要。
C2 是通過比較 Kaggle 的原始架構和 Wang 等人(2024b)使用的圖架構得出的。人類專家引入了多個增強表,這些對 GML 模型的性能至關重要。這些增強表背後的機制尚未得到充分研究,我們首次在數據集中引入它們。
C3 源自諸如(Harper & Konstan, 2015)等真實世界的數據集,我們發現當無法從架構中推導出表的適當類型時,簡單的啟發式方法可能效果不佳。C4 自然源自表格數據上定義的多個任務。我們是首個研究圖對不同下游任務性能影響的工作。
與傳統數據庫分析的關係。數據庫分析(包括規範化)是與我們工作相關的概念。

從關係數據到圖的圖構建的目標是找出哪種關係信息對下游任務有利。
例如,挑戰 2 的目標是考慮由分類值引起的關係是否有益。這個決定需要考慮此列與相應下游任務之間的語義關係,這是無法通過規範化來解決的。
相比之下,分析的目標是最小化數據冗餘並提高數據完整性。儘管有重疊,但數據分析方法無法完全解決圖構建任務。
基於從關係型表格數據進行圖構建的設計空間,我們從各個領域收集了 8 個數據集來評估圖構建方法。我們從以下來源收集這些數據集:
現有表格圖數據集的源頭,如 Diginetica
從現有表格圖數據集增強而來,如 Stackexchange
為圖構建改編的傳統表格數據集,包括 IEEE-CIS 和 Movielens
評估。為了評估生成圖的質量,我們採用定量評估方法來評估下游任務性能,即使用固定的 GML 模型(RGCN, RGAT, HGT, R-PNA)來比較不同圖構建方法的影響。更好的下游任務性能表明更高的圖質量。我們同時考慮了單個模型的性能和不同模型的平均性能。

Method (方法):

本節介紹一個自動圖構建解決方案,用於解決 3.1 節中的五個挑戰。如 2.2 節所討論的,我們將圖構建視為從具有隱式關係的原始表架構到具有顯式關係的最終圖架構的轉換。我們採用 LLM 作為決策者來自動生成轉換。
4.1 AUTOG: 基於 LLM 的圖構建框架
受經典的生成器-判別器結構啟發,我們首先設計一個生成器來產生合理的候選項,然後通過判別器評估生成的結果。在之前的工作中,人類數據科學家通常扮演生成器的角色,基於他們的專家知識生成輸出。
與人類類似,LLM 也展示了基於先驗知識生成分子結構或代碼格式增強的能力。因此,我們採用 LLM 作為生成器,並為其提供輸入表格數據以生成轉換。如圖 2 所示,我們提出的 AutoG 框架由以下模塊組成。
輸入模塊。AutoG 的輸入由兩部分組成:

輸入表架構,表示與數據相關的元數據
提示指令
我們遵循 Wang 等人(2024b)的工作,使用 2.1 節介紹的表架構格式來表示輸入數據。
一個示例可以在附錄 A.2 中找到。輸入架構文件可以從表格存儲(例如 Pandas DataFrames)中輕易生成,列數據類型可以是用戶定義的,也可以使用 LLM 從採樣

的列值中推斷得出。對於提示指令,我們包括:

圖構建任務的一般描述
相應下游任務的一句話描述
數據補充信息,包括數據集統計信息和示例列值

LLM 作為生成器。基於輸入模塊,我們進一步利用 LLM 生成轉換後的架構。一個直接的方法是讓 LLM 直接生成結構化輸出,如 YAML 格式的代碼。然而,我們發現開放式生成通常會產生無效的圖結構。為了解決這個問題,受函數調用思想的啟發,我們基於 5 個圖構建挑戰設計了基本的增強動作,然後通過增強鏈提示引導輸出,這在 4.2 節中詳細說明。

基於啟發式的圖構建器。一旦生成候選表架構,我們就使用啟發式算法將表轉換為圖。例如,如果我們選擇 Row2Node/Edge 啟發式算法,我們會將至少有兩個 FK 列且沒有 PK 的表,以及剩餘的 PK-FK 關係轉換為異構圖的邊,同時將其他表轉換為節點。

Oracle 作為判別器。在生成圖之後,我們設計一個 oracle 作為判別器來生成反饋。LLM 基於表的語義信息和統計信息生成候選結果。這些信息可以作為有價值的先驗,但無法評估生成的圖與特定下游任務的有效性和兼容性。

因此,我們採用圖構建的結果(無論是否成功)或執行 GML 模型訓練模塊來獲取生成圖的(估計)性能。這種反饋將進一步作為歷史信息附加到提示指令中。我們在 4.3 節中詳細說明 oracle 的設計。

4.2 基於增強鏈的引導生成

讓 LLM 生成架構最直接的方式是直接生成 YAML 格式的結構化輸出。然而,這種開放式生成存在以下陷阱:

LLM 生成的架構和增強代碼存在語法錯誤,導致流程無法自動進行。
LLM 傾向於遺漏那些需要多步增強的節點類型和關係。

以 Diginetica 數據集為例,可以通過首先將集合屬性列轉換為適當的增強列,然後從增強列中識別非 PK-FK 關係來找到關係。簡單地以單步方式生成架構無法提取這些關係。

為了緩解這些問題,我們提出了基於增強鏈的引導生成。首先,基於 3.1 節提出的四個挑戰,我們確定了以下基本增強動作。

1. CONNECT_TWO_COLUMNS(連接兩列):
   在兩列之間建立 PK-FK 關係,首先會確保它們滿足 PK 約束
   這個動作旨在解決挑戰 1

與可連接表發現(JTD)相比,這個動作更簡單,因為它直接基於 LLM 決策生成潛在的列對
JTD 也可以在需要更高準確性的場景中作為替代方案,但代價是更多的運行時間

2. GENERATE_NEW_TABLE(生成新表):
   通過移動列而不改變任何值,從原始表中導出新表
   這可以視為從原始表中識別多個節點或關係類型
   這個動作旨在解決挑戰 2

3. REMOVE(ADD)_PRIMARY_KEY(移除/添加主鍵):
   結合適當的啟發式方法,這個動作可以改變生成圖中表的類型(作為節點或邊類型)
   這個動作旨在解決挑戰 3

然後我們在提示中提供兩種類型的補充信息來幫助 LLM 決定採取什麼動作。
列的統計信息:

任務的文本描述和每列的統計信息被附加到提示指令中,引導 LLM 的決策
LLM 將根據增強表在語義上是否對任務有貢獻來決定像
GENERATE_NEW_TABLE 這樣的動作的有用性
例如:

如果任務是識別論文之間的引用關係,則"共同作者"關係高度相關,LLM 會傾向於生成表示這種關係的表
相反,"同年"關係信息量較少,LLM 生成它的可能性較小
此外,如果一個分類列只有兩個不同的值,生成的表會在圖中成為一個超級節點,這對模型訓練不理想,因此 LLM 會傾向於不生成這樣的表
思維鏈示例:

為每個動作提供示範，展示其用法。作者發現思維鏈(CoT)提示對動作生成至關重要。例如，在沒有 CoT 的情況下，LLM 傾向於僅找到具有相同名稱的列來建立非 PK-FK 關係。只有在引入 CoT 示範後，LLM 才能利用列的統計信息找到更一般的、具有不同列名的非 PK-FK 關係。

為確定終止步驟，作者將 null 動作添加到動作空間，並設置硬閾值 T 限制最大動作數，通常為 10。

4.3 設計 Oracle 生成反饋

生成模式候選後，需要 Oracle 評估其有效性並選擇最佳模式。儘管 LLM 能夠基於先驗知識生成模式，但它們無法定量預測不同模式如何影響下游任務性能。

實現定量 Oracle 的主要挑戰是高效獲取模型的近似性能。基於生成的模式使用啟發式構建圖後，AutoG 將自動執行 GML 模型擬合過程，並採用驗證性能作為最終指標。作者進一步探索加速此過程的潛力：

壓縮圖：通過在較小的圖上訓練和測試來提高評估效率。
採用早期階段訓練指標：例如驗證集性能。
簡化或無需訓練的模型：採用為異質圖設計的簡化模型，如線性 GNN。

作者比較了這些方法的有效性和效率，發現只有早期階段驗證性能能夠很好地估計下游任務性能。

4.4 候選和結果生成

在描述 LLM 的動作空間和 Oracle 後，AutoG 的最後一部分是候選生成策略。作者使用一種更簡單的策略，即一次生成一個動作來創建新候選，而不是使用複雜的基於樹的搜索策略（如 MCTS）。作者發現基於樹的搜索無法提高生成的候選質量，且許多候選是重複的。當生成無效動作時，AutoG 會回溯到最後一個有效狀態，並在連續錯誤後終止。為產生多樣化的模式，運行算法多次，選擇具有最佳 Oracle 分數的候選作為最終選擇。

Experimental Results (實驗結果):

5.1 實驗設置

為了研究不同圖構建方法的影響，作者固定 GML 模型以檢查下游任務性能。具體來說，選擇了兩個常用於異質圖的基線，RGCN 和 RGAT，並呈現了 RGCN 的結果（RGAT 的結果在附錄 E.1 中）。對於構建的圖，基於驗證集上的模型性能選擇最佳超參數。作者選擇 Claude's Sonnet-3.5 作為 LLM 的骨幹，並在 5.3 節中研究了不同 LLM 的影響。

考慮了以下基線方法：

XGBoost 和 DeepFM：直接應用於合併表格的兩個廣泛採用的表格數據基線。
TabGNN：基於每個分類值創建邊類型，並基於每個邊類型構建多重圖。
Row2Node 和 Row2Node/Edge：使用啟發式將表轉換為圖。
JTD with Row2Node/Edge：可聯接表發現（JTD）旨在找到表間可聯接的列，可

與啟發式結合生成具有更複雜關係的圖。
人類專家設計的圖模式。

5.2 定量評估
表 3 展示了不同圖構建方法的性能。評估遵循以下步驟：

使用相應的圖構建方法生成異質圖；
然後，使用構建的圖訓練 GML 模型以進行下游任務。模型性能用於確定圖的質量。

從實驗結果中，作者得出以下觀察結果：

AutoG 生成高質量圖：所提出的 AutoG 方法能夠超越其他自動圖構建方法，並達到接近人類專家的水平。
AutoG 相對於基於啟發式的方法的優勢：基於啟發式的自動發現方法只能應用於某些特殊情況。AutoG 在解決挑戰 2 方面具有獨特優勢。與挑戰 1 不同，挑戰 2 完全基於專家經驗解決。以 IEEE-CIS 為例，它有許多分類列。如果將所有分類列轉換為關係，會導致性能不佳（TabGNN）。相比之下，基於 LLM 的 AutoG 可以分析列之間的語義關係，例如將所有與卡相關的元信息分組到一個表中，從而獲得良好結果。
相同的圖可能對不同的下游任務效果不同：在 MAG 數據集上，作者觀察到專家設計的圖對年份預測任務不是最優的，比原始模式差得多。這表明需要根據任務自適應地生成圖，並說明自動圖構建的重要性。
5.3 深入分析
為了更好地理解 AutoG 的有效性，作者進一步研究了其組件的效果：
5.3.1 AutoG 變體研究
作者考慮了 AutoG 的兩個變體：AutoG-S（無預定義動作）和 AutoG-A（從 AutoG 中移除 Oracle）。結果顯示：

封閉式生成對於有效的模式生成是必要的。
Oracle 通常是不必要的，意味著 LLM 僅基於先驗知識就能生成好的候選。然而，在一些特定任務中，AutoG-A 表現不佳，這些任務可能存在潛在的噪聲關係。未來工作可能是在運行 AutoG 之前確定是否需要 Oracle，以提高整體效率。

5.3.2 LLM 的影響
作者評估了不同 LLM 的影響：Claude Sonnet 3.5、Mistral Large 和 Claude

Sonnet 3。結果顯示：

更強大的 LLM 能夠生成更好的模式，無效動作更少，這可能與指令遵循能力有關。
CoT 示範對 Mistral Large 效果不佳，這可能是由於不同 LLM 的預訓練策略不同。一般來說，對於能力超過 Sonnet3 的 LLM 模型，AutoG 能夠生成有希望的結果，並超越基於啟發式的對手。

5.3.3 AutoG 的工作機制
儘管 AutoG 表現出色，但 LLM 作為生成器由複雜的提示設計組成，使理解每個組件的角色以及它們如何應用於更一般類型的表格數據（例如具有匿名列的數據）變得具有挑戰性。作者進一步研究了不同提示組件的影響：

列的語義信息（列名）；
列的統計元信息；
提示中給出的例子；
每個動作的思維鏈示範。

作者建立了一個基於 MAG 的合成數據集，包括第 3.1 節中提出的挑戰 1-4，並確保測試數據不包含在 LLM 的預訓練集中。結果顯示：

示範對於 AutoG 生成有效動作是必要的。
CoT 和統計都對圖模式生成至關重要。特別是，LLM 只會找到瑣碎的增強（例如，具有相同列名的非 PK-FK 關係），這意味著 CoT 是 LLM 進行深度推理和充分利用統計數據的關鍵。
列名的語義信息對 AutoG 的性能至關重要，這是 AutoG 的一個限制。列名擴展可能被採用來增強 AutoG 在匿名數據上的有效性。

Related Work (相關工作):

近年來，GML 被廣泛採用來捕捉表格數據中的結構關係。關鍵挑戰之一是從表格數據中識別能夠使下游任務受益的圖結構。
早期在數據庫管理中使用基於規則的方法挖掘數據庫間的關係，但這些方法在大規模表格上的可擴展性有限。機器學習的興起導致了兩種基於 ML 的方法：基於啟發式和基於學習的方法。
基於啟發式的方法根據某些規則將表格數據轉換為圖：

Guo 等人(2021)基於表中具有分類值的列生成邊關係，通過多個列生成多重

圖。
Wu 等人(2021)和 You 等人(2020)基於每行表示樣本、每列表示特徵創建二部圖，其中 You 等人(2020)通過將數值存儲為邊屬性進一步支持數值。
Du 等人(2022)通過將每行視為超邊生成超圖。
Row2Node(Fey 等人，2024)和 Row2Node/Edge(Wang 等人，2024b)為具有顯式關鍵關係的多表提出。
Bai 等人(2021)設計了一個針對 RDB 預測任務的端到端模型。

這些方法仍限於滿足 RDB 規範的表格。
基於學習的方法旨在基於特徵間的相關性自動學習邊關係：

Chen 等人(2020)和 Franceschi 等人(2019)利用圖結構學習來學習樣本間的誘導邊關係。
Koutras 等人(2020)利用知識圖譜建立列間的關係圖，提取潛在的結構關係。
Dong 等人(2023)利用語言模型嵌入檢測表中的相似列，從而提取相關列。

為研究不同 GML 方法對表格數據的有效性，已開發多個基準，但它們的範圍僅限於模型評估或特徵評估，使圖構建評估成為一個未充分探索的領域。
基於 LLM 的自動數據科學：作者的工作也與應用 LLM 到自動數據科學相關。這些工作的核心原則是採用 LLM 的代碼生成能力，自動生成用於數據管理、數據增強或作為多樣化數據操作的通用接口的代碼。Zhang 等人(2024b)提出了一個評估 LLM 在各種數據科學場景中能力的基準。相比這些方法，AutoG 通過函數調用採用封閉式生成，確保生成的正確性。
基於異質圖的學習：具有多種節點和邊類型的異質圖自然抽象了關係數據庫數據。在這些圖上學習表示通常依賴於元路徑，將異質關係轉換為同質集合。早期方法專注於基於元路徑的相似度度量。隨著圖神經網絡(GNN)的出現，如 HAN(Wang 等人，2019)基於元路徑提取多個同質圖用於單獨編碼，MAGNN(Fu 等人，2020)進一步考慮元路徑中間節點的角色，而 RGCN(Schlichtkrull 等人，2018)和 G2S(Beck 等人，2018)則強調關係圖，其中邊攜帶豐富的語義信息。

Conclusion (結論):

在這篇論文中，作者通過建立基準和提出基於 LLM 的自動構建解決方案，正式化了圖構建問題。大量實驗結果表明，圖構建是一個重要步驟，可能顯著影響下游任務性能。所提出的 AutoG 能夠有效地處理這個重要任務，特別是當列呈現語義信息時。

然而，作者的方法仍有兩個限制：

數據集方面：所使用的數據集已經包含一些關係信息，可以通過啟發式方法轉換為圖結構（儘管此圖結構可能不夠有效）。因此，作者專注於相對簡單的場景，而下一個挑戰是更複雜的轉換，從原始非結構化文本文件開始。
方法方面：觀察到 LLM 嚴重依賴語義信息來做出有效決策，這在現實世界場景中是一個限制。將 AutoG 與命名擴展模塊結合可能是未來的一個潛在方向。

總的來說，這篇論文在圖構建領域做出了重要貢獻，提出了一個基準和一個基於 LLM 的自動解決方案，並通過實驗證明了其有效性和潛力。