One-Class Learning for Text Causal Discovery through Hypergraph Neural Networks

Marcos Paulo Silva Gôlo[1] and Ricardo Marcondes Marcacini[1]

Institute of Mathematical and Computer Sciences, University of Sao Paulo
marcosgolo@usp.br and ricardo.marcacini@icmc.usp.br

Abstract. We explore the problem of causal discovery between text pairs. We propose a new method called eCOLGAT (**e**dge **C**lassification through **O**ne-c**L**ass **G**raph **AT**tention autoencoder) that exploits hypergraphs to better learn the representation of edges (causal relations), graph attention networks to perform edge classification in causal graphs, one-class learning to better model the problem and reduce the labeling effort, and interpretability to improve the understanding of the causal discovery learning process. eCOLGAT outperformed other one-class methods and large language models (state-of-the-art for causal discovery), proving to be a promising method for causal discovery in text pairs.

Keywords: Event Causal Discovery · One-Class Classification · Graph Neural Networkds · Text Pair Causal Discovery.

1 Introduction

Understanding an event's causal relations is a challenging task that directly impacts society because this information can be applied in government analyses through decision-makers to reduce harm to society [11]. Although the vast majority of research and applications in causal discovery focus on effect inference tasks, the growing prevalence of textual event reporting in society has introduced new challenges and opportunities. Thus, there is an increasing demand for causal discovery methods specifically tailored to textual data. In particular, we are interested in binary classification tasks that address the causal relationship between two textual statements. For example, does a "*weakening economic environment*" cause "*rising unemployment rates*"? This task is called causal discovery in text pairs in the machine learning literature [8, 13, 12].

Studies have performed textual analysis to discover causality between text pairs by exploiting the Bidirectional Encoder Representations from Transformers (BERT) model [8, 13, 12]. Furthermore, considering text-based models, the state-of-the-art (SOTA) models are Large Language Models (LLMs) [16]. Even obtaining SOTA results, LLMs do not observe the causal relation between sentences, since each pair of texts is analyzed individually. This fact can harm the performance of the models since modeling these relations can allow the models to explore more information. Graphs are an alternative since graphs model this

task naturally because each node in the graph is a text, and an edge is created between the nodes when there is a causal relation between the texts [28].

Graph Neural Networks (GNNs) have been widely used to discover cause-effect relations in text pairs modeled by graphs [20, 22, 23]. However, we highlight the limitation of GNNs in learning representations for edges due to biased message passing for node representation learning [10]. Furthermore, GNN works for causal discovery are typically based on binary learning, i.e., during training they learn from labeled instances of causal and non-causal classes. We highlight some gaps of binary learning, such as the need for a significant amount of labeled instances for the algorithm learning step. Furthermore, we highlight the large scope of non-causal relations, which makes labeling challenging [6].

We present a new method for edge classification through one-class learning (OCL) in graphs. The method is called eCOLGAT (**e**dge **C**lassification through **O**ne-c**L**Ass **G**raph **AT**tention autoencoder). eCOLGAT is based on hypergraphs that improve representation learning through GNNs for edges since they transform edges into nodes and thus GNNs can learn better representations [10]. Furthermore, we propose to model the problem of causal discovery in text pairs through OCL since, in OCL, the algorithm trains with only one class (causal relations) and can predict two (causal or non-causal relations). In this way, covering the entire scope of non-causal relations is unnecessary, and the user's labeling effort is reduced since it only labels causal instances [24, 4]. In this sense, we base eCOLGAT on the SOTA loss functions for GNNs and OCL [7, 29]. Finally, we explicitly learn three-dimensional representations to introduce interpretability in eCOLGAT naturally. In summary, our contributions are:

1. We model causal discovery between text pairs through **one-class learning**, providing labeling advantages and more natural modeling for the problem.
2. We model causal discovery between text pairs through **hypergraphs** to better exploit graph neural networks' representation learning for edges.
3. We introduce **interpretable** representation learning on hypergraphs through graph neural networks and one-class learning for causal discovery.

2 Related Work

Hassanzade et al. [8] is one of the pioneering studies in causal discovery between text pairs using language models based on deep neural networks. The authors propose an unsupervised method based on the pre-trained BERT language model called NLM-BERT. First, the authors generated embeddings for a corpus of 17 million causal sentences using Bidirectional Encoder Representations from Transformers (BERT). Second, the method generates embeddings for text pairs to answer whether the first text causes the second. Using a technique based on cosine similarity between the top k similar embeddings, NLM-BERT generates two scores and compares these values with a threshold to decide whether there is causality between the texts. The authors obtained an f1-score of 67%, outperforming four other methods in the Risk Models dataset.

Kayesh et al. [13] proposed fine-tuning the BERT model and its variations to detect causality. The authors fine-tuned with a semi-supervised dataset of 100,000 pairs of causal and non-causal sentences. The authors compared their results with the results of Hassanzade et al. [8]. The developed methods could not outperform NLM-BERT, obtaining an f1-score of 66% on the Risk Models dataset. In the same line of research, Kayesh et al. [12] extended the work of Kayesh et al. [13] by adding another training dataset of 197,000 sentence pairs, three new methods, and three combinations of these methods to detect causality, and used a causality graph through each of the training sets (197,000 and 100,000 sentence pairs). The proposal of Kayesh et al. [12] is based on a knowledge fusion that fuses representations generated in two stages: extraction of causal features (embedding generation through the graph) and extraction of contextual features (embedding generation through attention mechanisms). For the Risk Models dataset, the new model trained on the set of 100 thousand sentence pairs could not outperform NLM-BERT, obtaining 66% of f_1.

The state-of-the-art (SOTA) for causal discovery in textual data is through large language models (LLMs) [16]. LLMs are pre-trained models on a corpus with trillions of words capable of generating text from an input text. In this sense, LLMs can be queried in natural language whether one sentence causes another to work as causal discovery models. In this sense, the PyWhy-LLM library was developed[1]. Other methods that also obtain SOTA results for causal discovery are graph neural networks (GNNs) methods since graphs express the causal relations explicitly, and GNNs obtain SOTA results in graphs [20, 22, 23].

Minghim et al. [20] obtain contextualized embeddings for the words used to build a causal graph. The authors used a gated GNN and recurrent neural network decoder for graph neural network encoding. Finally, a fully connected neural network was used for Event Causality Identification binary classification. The model outperformed baselines, including the BERT model. Sakaji et al. [22] is another study that explores BERT and GNNs. This work explored Graph Attention Networks, pointing out improvements in relation to graph convolutional networks. Finally, the authors used two fully connected neural networks to classify sentences as cause and effect separately. The model outperformed baselines, including the BERT model. Finally, Sasaki et al. [23] perform edge classification in the causal graph through GNNs, in which the edges are the causality relations and the nodes are the sentences. A classifier was used in the final layer of the GNN. The authors do not compare with baselines but obtain 98% of f_1.

Text-based methods (BERT and LLM) do not explore information about the relations between sentences because they do not model the data through graphs. Even using graphs in the strategy, Kayesh et al. [12] do not use graphs as the main part of the method nor GNNs, obtaining inferior results than text-based methods. GNNs explore this relational information and can explore textual information because they need an initial representation for the texts. On the other hand, GNNs focus on node message passage, i.e., in the sentences and not in edges, in which the classification will be performed, thus making causal

[1] `https://github.com/py-why/pywhy-llm/tree/main`

discovery difficult. Furthermore, we highlight that GNNs are explored as binary supervised methods, i.e., they require labeling of causality and non-causality, making knowledge discovery difficult since labeling what is not causal is costly due to lack of scope. In this sense, in the next section, we present a method based on hypergraphs and one-class learning for causal discovery in text pairs.

3 eCOLGAT: edge Classification through One-cLAss Graph ATtention autoencoder

We propose a novel method called **e**dge **C**lassification through **O**ne-c**L**Ass **G**raph **AT**tention autoencoder (eCOLGAT). Our method for causal detection in text pairs is based on hypergraphs, one-class learning (OCL), and graph neural networks (GNNs). Our method presents several novelties for causal discovery between text pairs. First, eCOLGAT is a pioneering one-class method for causal discovery. Second, using hypergraphs with graph neural networks for causal discovery in text pairs is also novel. Third, exploiting three-dimensional representation learning to provide interpretability for causal discovery in text pairs is another novelty of our method.

Hypergraphs lead better with the edge-representation gap through GNNs. eCOLGAT explores a state-of-the-art one-class loss function to encapsulate causal relations closer to the center within the sphere. Our proposal learns a new three-dimensional latent space to provide interpretable learning, where causal relations are positioned inside a sphere and non-causal relations outside. We learn the new space through a graph attention autoencoder to explore the reconstruction loss as a constraint to the sphere loss function and the attention mechanism to learn better representations for the edges. Finally, our final loss function combines the sphere loss function with the reconstruction loss function. Figure 1 presents an eCOLGAT illustration.

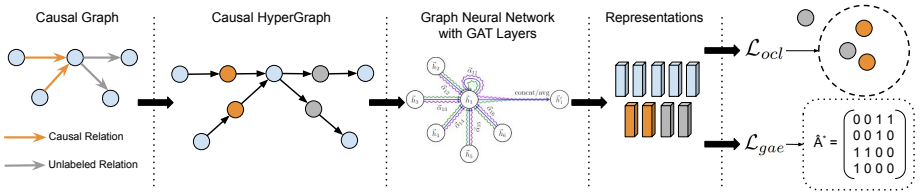

Fig. 1. Our proposed method eCOLGAT. We show all the steps from eCOLGAT: hypergraph generation, representation learning through GAT, one-class sphere loss ($\mathcal{L}_{ocl}$), and GAE loss ($\mathcal{L}_{gae}$).

Causal discovery between text pairs can be defined as a binary classifier with two inputs that output a causal or non-causal label. We define $\{s_1, s_2, s_3, ..., s_m\} \in \mathcal{S}$ as a set of m natural language sentences and $\{causal, non-causal\} \in \mathcal{C}$ as the

set of classes. This classifier can be defined as a function $f : \mathcal{S} \rightarrow \mathcal{C}$ that maps two text pairs $(s_i, s_j) \in \mathcal{S}$ to the causal or non-causal label. In this case, $\boldsymbol{S} \in \mathbb{R}^d$ represents the sentence feature space. We define one-class learning for causal discovery on text pairs as the function f^* from a training set with only causal labels $\{((s_1, s_2); causal), ((s_5, s_3); causal), ..., ((s_i, s_j); causal)\}$ that approximates the unknown mapping function f.

In the context of this work, $\boldsymbol{S} \in \mathbb{R}^d$ will be a representation generated by Bidirectional Encoder Representations from Transformers (BERT) for each of our sentences, where d has value 384 [21]. Given the limitations of text-based models, this study models causality pairs through a directed graph. A directed graph is formally defined as $\mathcal{G} = (\mathcal{V}, \boldsymbol{A})$, where each node $v_i \in \mathcal{V}$ and $\boldsymbol{A}$ is the adjacency matrix containing relations between nodes. Thus, $\mathcal{V} \equiv \mathcal{S}$, i.e., the sentences are the nodes, and $\boldsymbol{A}$ contain the causal and non-causal relations [17]. Our method also exploits GNNs, state-of-the-art methods for different tasks, including node classification [9]. Due to the limitation of GNNs in learning edge representation and classification, this study explores the use of hypergraphs to model causality in graphs [10].

We transform each edge into a node to transform our directed graph into a directed hypergraph. Thus, when a node v_i was connected to node v_j, we create a node v_o (which we will call here node-edge) and connect v_i with v_o and v_o with v_j [10]. In this way, GNNs can be explored without the limitation of message passing for edges because the edges are now nodes. We define a hypergraph for causal discovery as $\mathcal{G}^* = (\mathcal{V}^*, \boldsymbol{A}^*)$, where $\mathcal{V}^*$ is the nodes set $\mathcal{V}$ plus the edge set of $\mathcal{G}$ and $\boldsymbol{A}^*$ is the adjacency matrix of the hypergraph with the new generated relations [10]. Another point is the need for GNNs in each node to have an initial representation. In the case of our proposal, part of the set $\mathcal{V}^*$ (the $\mathcal{V}$ set) has the initial representation of BERT, but the nodes-edge does not. Thus, to use GNNs, we add the average of the adjacent nodes-edge as an initial representation of the nodes. In the above example, $\boldsymbol{v_o} = avg(\boldsymbol{v_i}, \boldsymbol{v_j})$.

We exploit GNNs to learn representations in our causality hypergraph. The GNNs consider the structured representation of each node $\boldsymbol{v}_i \in \boldsymbol{V}^*$ and the adjacency matrix $\boldsymbol{A}^*$ as input for the representation learning process. Therefore, $g(\boldsymbol{V}^*, \boldsymbol{A}^*; \mathbf{W})$ represents a GNN with trainable weights $\mathbf{W} = \{\boldsymbol{W}^{(1)}, \cdots, \boldsymbol{W}^{(L)}\}$ in L hidden layers. Formally, for the l-th layer, the GNN propagation can be summarized as follows [26]:

$$\boldsymbol{H}^{(l+1)} = g(\boldsymbol{H}^{(l)}, \boldsymbol{A}^*; \boldsymbol{W}^{(l)}), \tag{1}$$

in which $\boldsymbol{H}^{(l)}$ is the input to the l-th GNN layer, and $\boldsymbol{H}^{l+1}$ is the output of this layer. The representations $\boldsymbol{V}^*$ are the inputs for the first layer, i.e., $\boldsymbol{V}^* \equiv \boldsymbol{H}^{(0)}$. In this sense, $\boldsymbol{H}^{(L)}$ are the learned embeddings for each node. GNNs learn representations $\boldsymbol{H}^{(L)}$ by aggregating information from neighbors.

We chose Graph Attention Networks (GAT) [3] as GNNs, given their improved performance in node classification tasks, including causality discovery [22]. The GAT learns the most important edges through the attention mechanism [3], i.e., the GAT has attention to the main relations in the graph, improving

information aggregation. The GAT aggregation step can be defined by Equation 2 [3]:

$$\boldsymbol{h}_{v_i}^{l} = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{v_j \in N(v_i)} \alpha_{ij}^{k}\boldsymbol{W}^{k}\boldsymbol{h}_{v_j}^{l-1}\right), \tag{2}$$

in which $\boldsymbol{h}_{v_i}^{l}$ is the aggregation result of the v_i neighbors defined as N_{v_i}, and $\boldsymbol{h}_{v_j}^{l-1}$ is the feature vector of the node v_j at the $l-1^{th}$ layer. $\boldsymbol{W}^{k}$ are the GAT weights associated with the head k. K is the number of heads in the GAT, and α_{ij}^{k} is the attention computed bu the $k-th$ attention head defined by the Equation 3.

$$\alpha_{ij} = \frac{exp(\boldsymbol{a}^{\intercal} LeakyReLU(\boldsymbol{W}\boldsymbol{h}_{v_i}^{l-1} || \boldsymbol{W}\boldsymbol{h}_{v_j}^{l-1}))}{\sum_{v_u \in N(v_i)} exp(\boldsymbol{a}^{\intercal} LeakyReLU(\boldsymbol{W}\boldsymbol{h}_{v_i}^{l-1} || \boldsymbol{W}\boldsymbol{h}_{v_u}^{l-1}))}, \tag{3}$$

in which, $\boldsymbol{a}$ is the shared attention mechanism, $(\cdot)^{\intercal}$ represents transposition and $||$ is the concatenation operation.

GNNs with sphere loss functions are state-of-the-art for one-class graph neural networks [27, 7]. These methods learn $\boldsymbol{h}_{v_i}^{L}$ encapsulating the nodes of interest. To detect causality through our hypergraph, we explore this strategy [7]. We use the sphere loss function $\mathcal{L}_{ocl}$ defined in Equation 4 [7].

$$\mathcal{L}_{ocl}(\boldsymbol{W}) = \frac{1}{|\mathcal{V}^{\text{in}}|}\sum_{i=1}^{|\mathcal{V}^{\text{in}}|}\begin{cases} o_i + 1, \text{ if } o_i > 0 \\ \exp(o_i), \text{otherwise} \end{cases}, \tag{4}$$

$$o_i = \|\boldsymbol{h}_{v_i}^{(L)} - \boldsymbol{c}\|^2 - r^2. \tag{5}$$

in which equation 5 represents the value indicating whether the interest instance v_i is within the hypersphere with radius r and center $\boldsymbol{c}$ and $\mathcal{V}^{\text{in}}$ are the set of interest nodes-edges in our hypergraph (with the causal class).

By using only $\mathcal{L}_{ocl}$, all nodes will converge to the center. In this sense, following the work of [7], we use our GAT layer in a graph autoencoder (GAE) since GAEs have an unsupervised loss function that is a constraint to mitigate the collapse of the sphere. Thus, we combine the sphere loss function with the loss function of GAEs [7], which obtained superior and state-of-the-art results compared to other methods [27]. A GAE uses GNN layers as an encoder and an inner product of the latent representation as a decoder to learn node representations. Equation 6 describes a GAE [14]:

$$GAE = \begin{cases} Encoder : \boldsymbol{H}^{(L)} = g(\mathbf{V}^{*}, \boldsymbol{A}^{*}; \mathbf{W}) \\ Decoder : \hat{\boldsymbol{A}}^{*} = \sigma(\boldsymbol{H}^{(L)} \cdot \boldsymbol{H}^{(L)^{\intercal}}) \end{cases}, \tag{6}$$

in which $\sigma(.)$ is a logistic sigmoid function. The GAE loss function $\mathcal{L}_{gae}$ is defined in Equation 7 (binary cross entropy loss applied in the adjacency matrix). Therefore, our final loss function is defined by: $\mathcal{L} = \mathcal{L}_{ocl} + \mathcal{L}_{gae}$.

$$\mathcal{L}_{gae}(\boldsymbol{W}) = -\frac{1}{|V|}\sum_{i=0}^{|V|}\sum_{j=0}^{|V|}(\boldsymbol{A}_{ij}^{*} \cdot log\hat{\boldsymbol{A}}^{*}{}_{ij} + (1 - \boldsymbol{A}_{ij}^{*}) \cdot log(1 - \hat{\boldsymbol{A}}^{*}{}_{ij})). \tag{7}$$

We show the causal graph, causal hypergraph, the GAT step for representation learned in the hypergraph, the sphere loss, and GAE loss. Even though it is possible to understand the model's decision by observing a sphere and instances inside the sphere (causal nodes-edge) and outside the circle/sphere (non-causal nodes-edge), it is challenging to interpret the learning that generated this decision, since we cannot visualize the representations generated during learning if the dimension is greater than three. On the other hand, with representations in three dimensions, we can observe and interpret the representations generated during learning. In this sense, we bias eCOLGAT learning so that our method learns representations in size three to provide interpretability for representation learning in the scenario of one-class learning for causal discovery in text pairs.

4 Experimental Evaluation

This section presents the experimental evaluation of this article. We present the used dataset, experimental settings, results, and discussion. Our research goal is to demonstrate that our eCOLGAT proposal outperforms other SOTA methods for causal discovery in text pairs. Another goal is to demonstrate that our method learns low-dimensional representations, providing interpretability for the causal discovery scenario. The experimental evaluation codes are publicly available[2].

4.1 Dataset

There are some benchmark datasets for detecting causality between text pairs, such as the three explored by Hassanzadeh et al. [8]. We used the largest of them to perform the empirical evaluation. The other two are very small, as they have only 160 and 59 causal pairs, making them unfeasible to train our one-class learning model. In this sense, we explored the Risk Models dataset [8].

Hassanzadeh et al. [8] explored models built by expert analysts to configure a decision support system [25] as a source of causal knowledge by human experts. The models are graphs in which the nodes are texts represented by descriptions of conditions or events, and the edges show causal relations. The models are based on enterprise risk management, expert knowledge, literature study, and reports. The authors created the cause-effect pairs dataset by transforming each edge in the graph into a pair of texts with the causal label. Finally, the dataset has 368 causal pairs with 223 unique cause/effect sentences. For the non-interest class, the authors randomly chose 368 pairs to be non-causal.

4.2 Experimental Setting

We focus only on unsupervised and one-class methods due to the difficulty of creating a large enough training set with reasonable quality and coverage. We compare the eCOLGAT with the state-of-the-art text-based methods for causal

[2] `https://github.com/GoloMarcos/eCOLGAT`

discovery [16]. We compare our methods with five large language models using the strategy of Pywhy-LLM library. Different LLMs have been proposed in the last three years, and they have differences that generate advantages and disadvantages for each model [30]. We use the 5-fold cross-validation for our experiments. We use four folds of the causal class to train, the remaining fold to test, and one fold of the non-causal class to test. Finally, we use the f_1-macro to compare all models.

We explore in our methodology four families of LLMs open-source: the LLM from meta (LLaMa) [18], from Microsoft (Phi) [19], from Google (Gemma) [5], and from Alibaba Cloud (Qwen) [2]. Each LLM has a number of parameters: Llama 3 (8 and 70 billion of parameters), Phi 3 (14 billion of parameters), Gemma 2 (27 billion of parameters), and Qwen 2 (7 billion of parameters). We also compare eCOLGAT with one-class methods since we have an initial representation for each sentence (BERT embedding) and can generate an initial representation for causal relations (average of causal and effect sentences). In this sense, we explore two one-class methods: One-Class Support Vector Machines (OCSVM) [1] and Isolation Forest (IsoForest) [15]. We use the following parameters for the methods:

- **LLMs**: parameter-free;
- **OCSVM**: kernel = {rbf, poly, sigmoid, linear}, $\nu = \{0.05 * b\}, b \in [1..19]$, and $\gamma =$ { scale, auto };
- **IsoForest**: nº of estimators = $\{1, 2, 5, 10, 50, 100, 200, 500\}$, maximum samples and maximum features = $\{0.1 * b\}, b \in [1..10]$;
- **eCOLGAT**: radius - $\{0.35, 0.45, 0.5\}$, epochs = $\{700, 1000, 1500\}$, heads for GAT = $\{1, 2, 3\}$, learning rate = $\{0.001, 0.0001, 0.0005\}$.

4.3 Results and Discussion

Table 1 presents the results of our study. We present the f_1-macro for all folds. We also present the average f_1 for the seven models explored and eCOLGAT. Higher values are in bold (best models). The second-highest values are underlined (second-best models). eCOLGAT outperforms the other two OCL methods since it obtained the higher f_1. OCSVM obtained the second-best results, followed by Phi 3 and LLaMa 3 (70b). LLaMa 3 (8b) obtains the worst f_1, followed by IsoForest and Qwen 2. Compared to LLM models, we highlight OCL models with state-of-the-art results and outperform these models.

Our method obtained the better f_1-macro in all analyzed folds compared to other methods. In addition, we obtained a 10% gain from the second-best model (OCSVM) and 25% from LLM models (Phi 3). It is worth mentioning that the second-best model also had a significant gain of 6% from LLM models, justifying the use of one-class learning for causal discovery in text pairs. Another point of attention in the results is Isolation Forest, which obtained a performance comparable with the other LLM algorithms. This shows that one-class learning is promising, but the choice of algorithm is important.

Table 1. f_1-macro for each method in the five folds and the average. The best results are in bold, and the second best are underlined.

Models	Folds					Average
	1	2	3	4	5	
LLaMa 3 (8b)	0.338	0.349	0.365	0.374	0.337	0.353
Phi 3 (14b)	<u>0.652</u>	0.612	0.580	0.585	0.657	0.618
Qwen 2 (7b)	0.497	0.467	0.516	0.492	0.524	0.499
Gemma 2 (27b)	0.451	0.521	0.502	0.502	0.556	0.508
LLaMa 3 (70b)	0.511	0.596	0.506	0.568	0.553	0.548
BERT + OCSVM	0.634	<u>0.659</u>	<u>0.659</u>	<u>0.688</u>	<u>0.718</u>	<u>0.672</u>
BERT + IsoForest	0.480	0.492	0.477	0.463	0.539	0.493
eCOLGAT	**0.764**	**0.811**	**0.707**	**0.766**	**0.804**	**0.771**

Typically, the number of LLM parameters influences the performance gain, as shown in other task performances considering LLMs. In the task explored in this study, models with fewer parameters such as Phi 3 obtained better performance than models with more parameters such as LLaMA 3 with 70 billion parameters or Gemma with 27 billion. Therefore, choosing LLMs for causal discovery in text pairs is also not trivial.

Figure 2 presents the confusion matrices for eCOLGAT and all baseline methods. The rows of the confusion matrix represent the true labels, while the columns represent the predicted labels. The cells of the main diagonal are the true positives (TP) and true negatives (TN), and the cells of the secondary diagonal are the false negatives (FN) and false positives (FP), where positive means our causal class and negative means the non-causal class. Thus, the higher the VP and VN values, the higher the accuracy, and the higher the FN and FP values, the higher the errors.

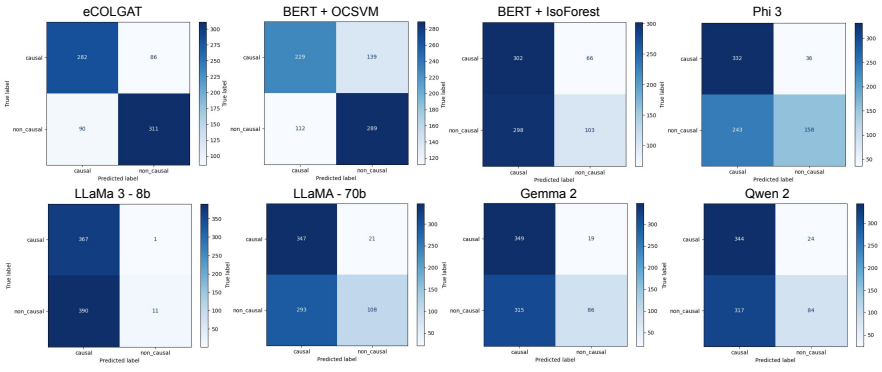

Fig. 2. Confusion Matrix for each method with the sum of the five folds. Higher values are in dark blue, and smaller values are in light blue.

eCOLGAT obtained the highest VN value and the lowest FP value, i.e., it correctly predicted more non-causal relations. LLaMA 3 (8b) obtained the highest VP value and the lowest FN value, i.e., it correctly predicted more causal relations. On the other hand, these correct predictions are due to the model's bias for this class. Note that it obtained the highest FP value and the lowest VN value, i.e., it predicted all instances for the causal class. This behavior was also observed in other models, such as IsoForest, Gemma 2, and Qwen 2. Even though it did not obtain the best VP and FN values, we emphasize that our model balanced its values, resulting in better classification performances.

We present the representations generated by the eCOLGAT learning to demonstrate the interpretability of our method. In this sense, Figure 3 presents the eCOLGAT representations for the node-edges in our hypergraph focusing on the learning process. In the real world, we can show the video of the learning process since we can plot all learning epochs without processing the representation (our learned representations have three dimensions). The epochs are 30, 70, 110, 120, 140, 240, 280, and 360. Purple points represent the causal edges, and yellow points represent the non-causal edges.

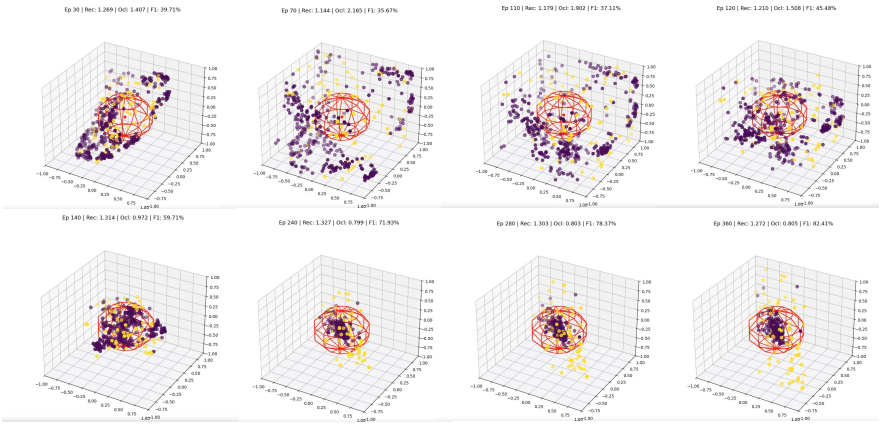

Fig. 3. Interpretability plot considering the three-dimensional last layer learned representations of eCOLGAT in the second fold. The colors indicate the causal (purple) and the non-causal (yellow) classes.

We observe the learning process of our proposal through eCOLGAT interpretability, as shown in Figure 3. In the initial steps, we observe eCOLGAT focusing on graph reconstruction (loss $\mathcal{L}_{gae}$). In the next steps, we observe learning through the two losses since the causal node-edge gets closer to the sphere while non-causal node-edge instances are outside the sphere. In the final steps, we observed that the one-class loss $\mathcal{L}_{ocl}$ encouraged the instances to continue coming to the sphere center while the $\mathcal{L}_{gae}$ maintained the non-causal node-edges outside the sphere.

5 Conclusion and Future Work

We propose eCOLGAT, a one-class GNN for causal discovery through edge classification. eCOLGAT explores a hypergraph to improve the edge representation learning and explores GAT as the GNN layer. Our eCOLGAT is also based on the SOTA sphere loss function for OCL and the reconstruction loss function. eCOLGAT explores three-dimensional representations during classification, providing interpretability for learning causal discorevy on text pairs.

Our experiments show that LLMs still have limitations in classifying causal relations, making it important to consider alternative representations. The proposed method incorporates both textual features from pre-trained models and relations between different cause-effect pairs, outperforming other LLM and OCL methods. In future work, we intend to explore a heterogeneous version of eCOLGAT on more causal text pair datasets, as well as incorporate classic effect inference tasks beyond textual data by mapping their attributes through natural language descriptions and modeling them in graphs.

Acknowledgments: This work was supported by FAPESP (grant numbers 2023/10100-4 and 2019/07665-4), CAPES (grant number 88887.671481/2022-00), CNPQ (grant number 316507/2023-7), Center for Artificial Intelligence (C4AI-USP), IBM Corporation and LatAm Google Ph.D. Fellowship.

References

1. Alam, S., Sonbhadra, S.K., Agarwal, S., Nagabhushan, P.: One-class support vector classifiers: A survey. Knowledge-Based Systems **196**, 105754 (2020)
2. Alibaba: Qwen technical report (2023), `https://arxiv.org/abs/2309.16609`
3. Brody, S., Alon, U., Yahav, E.: How attentive are graph attention networks? In: International Conference on Learning Representations (2022)
4. Emmert-Streib, F., Dehmer, M.: Taxonomy of machine learning paradigms: A data-centric perspective. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **12**(5), e1470 (2022)
5. Gemma: Gemma 2: Improving open language models at a practical size (2024), `https://arxiv.org/abs/2408.00118`
6. Gôlo, M.P.S., Marcacini, R.M.: Text representation through multimodal variational autoencoder for one-class learning. In: Anais do XXXVI Concurso de Teses e Dissertações. pp. 148–157. SBC (2023)
7. Gôlo, M.P.S., Junior, J.G.B.M., Silva, D.F., Marcacini, R.M.: Olga: One-class graph autoencoder. arXiv (2024)
8. Hassanzadeh, O., Bhattacharjya, D., Feblowitz, M., Srinivas, K., Perrone, M., Sohrabi, S., Katz, M.: Answering binary causal questions through large-scale text mining an evaluation using cause-effect pairs from human experts. In: IJCAI (2019)
9. Jin, M., Koh, H.Y., Wen, Q., Zambon, D., Alippi, C., Webb, G.I., King, I., Pan, S.: A survey on graph neural networks for time series: Forecasting, classification, imputation, and anomaly detection. Transactions on Pattern Analysis and Machine Intelligence (2024)
10. Jo, J., Baek, J., Lee, S., Kim, D., Kang, M., Hwang, S.J.: Edge representation learning with hypergraphs. In: NeurIPS. pp. 7534–7546. NeurIPS foundation (2021)

11. Kayesh, H., Islam, M.S., Wang, J.: Event causality detection in tweets by context word extension and neural networks. In: Int. Conf. on parallel and distributed computing, applications and technologies. pp. 352–357. IEEE (2019)
12. Kayesh, H., Islam, M.S., Wang, J.: Answering binary causal questions using role-oriented concept embedding. Transactions on Artificial Intelligence (2022)
13. Kayesh, H., Islam, M.S., Wang, J., Anirban, S., Kayes, A., Watters, P.: Answering binary causal questions: A transfer learning based approach. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–9. IEEE (2020)
14. Kipf, T.N., Welling, M.: Variational graph auto-encoders. stat **1050**, 21 (2016)
15. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: 2008 eighth ieee international conference on data mining. pp. 413–422. IEEE (2008)
16. Liu, X., Xu, P., Wu, J., Yuan, J., Yang, Y., Zhou, Y., Liu, F., Guan, T., Wang, H., Yu, T., et al.: Large language models and causal inference in collaboration: A comprehensive survey. arXiv preprint arXiv:2403.09606 (2024)
17. Ma, J., Wan, M., Yang, L., Li, J., Hecht, B., Teevan, J.: Learning causal effects on hypergraphs. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 1202–1212 (2022)
18. Meta: The llama 3 herd of models (2024), `https://arxiv.org/abs/2407.21783`
19. Microsoft: Phi-3 technical report: A highly capable language model locally on your phone (2024), `https://arxiv.org/abs/2404.14219`
20. Minghim, R., Milos, E., Provia, K., et al.: Ggnn@ causal news corpus 2022: Gated graph neural networks for causal event classification from social-political news articles. In: Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE). pp. 85–90 (2022)
21. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics (2019)
22. Sakaji, H., Izumi, K.: Financial causality extraction based on universal dependencies and clue expressions. New Generation Computing **41**(4), 839–857 (2023)
23. Sasaki, H., Fujii, M., Sakaji, H., Masuyama, S.: Enhancing risk analysis with gnn: Edge classification in risk causality from securities reports. International Journal of Information Management Data Insights **4**(1), 100217 (2024)
24. Seliya, N., Abdollah Zadeh, A., Khoshgoftaar, T.M.: A literature review on one-class classification and its potential applications in big data. Journal of Big Data **8**, 1–31 (2021)
25. Sohrabi, S., Katz, M., Hassanzadeh, O., Udrea, O., Feblowitz, M.D., Riabov, A.: Ibm scenario planning advisor: Plan recognition as ai planning in practice. Ai Communications **32**(1), 1–13 (2019)
26. Tang, J., Liao, R.: Graph neural networks for node classification. Graph Neural Networks: Foundations, Frontiers, and Applications pp. 41–61 (2022)
27. Wang, X., Jin, B., Du, Y., Cui, P., Tan, Y., Yang, Y.: One-class graph neural networks for anomaly detection in attributed networks. Neural computing and applications **33**, 12073–12085 (2021)
28. Zanga, A., Ozkirimli, E., Stella, F.: A survey on causal discovery: theory and practice. International Journal of Approximate Reasoning **151**, 101–129 (2022)
29. Zhang, F., Fan, H., Wang, R., Li, Z., Liang, T.: Deep dual support vector data description for anomaly detection on attributed networks. International Journal of Intelligent Systems **37**(2), 1509–1528 (2022)
30. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al.: A survey of large language models. arXiv (2023)