Abstract

Methods for selected problems in unsupervised graph representation learning

Machine learning (ML) methods have been studied in a variety of applications and data types. Since most downstream ML models expect a vector from a continuous space as input, representation learning methods have been developed to automatically create representation vectors (embeddings) for the input data. While many embedding methods exist for traditional data types, such as BERT for text or ResNet for images, this task is much more difficult for graph-structured data. Graphs can be used to describe objects (using nodes) and their relationships (using edges). For simple graphs, the main objective of representation learning methods is to capture the graph's structure. Nowadays, graphs contain multiple information sources, i.e., besides the structure (nodes and edges), one can assign attributes to nodes, edges, and even whole graphs (attributed graphs). Deriving appropriate graph representations is important for ML tasks to enable them to perform well, e.g., node classification or link prediction.

In recent years, various graph representation learning methods have been proposed. Despite the success of these methods, the following issues and research gaps are still not covered. First of all, most proposed approaches are inherently transductive (they optimize a fixed-size embedding matrix). Such a setting does not allow obtaining embeddings for previously unseen examples. Moreover, for real-world, large-scale graphs with millions or even billions of nodes, this solution requires an infeasible amount of memory and is impractical as graphs tend to evolve over time. The next major issue is that the (semi-)supervised setting requires labeled data. Obtaining such labels is an expensive and time-consuming task. Unsupervised (and self-supervised) learning has shown that graph representations can also be derived solely from the network structure and attributes. Yet, many proposed approaches were introduced in the semi-supervised setting, i.e., the model is jointly optimized in a particular downstream task. Hence, their expressive power is also limited by the choice of the downstream task and its connected loss function. In the unsupervised setting, the focus is on the only available structural and attribute information. Finally, the last issue concerns the problem of defining negative samples in contrastive loss functions utilized in unsupervised graph representation learning methods. The quality of the embeddings and, consequently, the performance of downstream tasks highly depend on how the negative samples are defined. Choosing appropriate negative samples is particularly difficult in graphs compared to other data types. Hence, the research community was exploring negative-sample-free methods in computer vision or natural language processing but were not explored for graphs.

The objectives of the doctoral dissertation were: 1) verifying whether the usage of the cross-correlation measure applied to augmented views of an attributed graph allows training a graph neural network in a self-supervised setting, such that this GNN computes better node representation vectors than existing self-supervised approaches, measured by the performance in downstream tasks and training time complexity, 2) developing a deep neural network model for calculating edge representation vectors that is trained using a combination of a contrastive learning objective with a feature reconstruction loss, such that this method's embedding vectors are better in downstream tasks than those obtained as aggregations of source and target node representations, 3) developing an incremental learning method for node representation vectors in dynamic graphs, which utilizes any kind of static node embeddings from consecutive graph snapshots and exhibits a lower time and memory complexity than contemporary methods for representation learning on dynamic graphs while providing competitive quality measure gains, and 4) developing a method for fusing together node attribute information with a given precomputed structural node embedding, resulting in a single low-dimensional embedding that performs better in downstream tasks than: the structural embedding, the node attributes or other attributed node representation learning methods.

The dissertation is a collection of thematically related works in the form of five scientific publications centered around the topic of unsupervised representation learning methods for graphs.

The first publication (*Graph Barlow Twins: A self-supervised representation learning framework for graphs*) proposes a novel framework for self-supervised representation learning of nodes in attributed graphs, which utilizes the empirical cross-correlation matrix between embeddings of two augmented views of a graph. It provides a simple yet powerful symmetrical neural network architecture and does not require any negative samples in the training process. Experimental evaluation shows that the node representations computed by the proposed framework achieved an analogous performance compared to state-of-the-art methods while requiring substantially fewer hyperparameters and converging in order of magnitude training steps earlier, leading to an overall speedup of up to 42 times compared to state-of-the-art methods.

In the second article (*AttrE2vec: Unsupervised attributed edge representation learning*), a novel method for learning edge representation vectors was introduced. It explores edge neighborhoods via random walks and applies an aggregation function to obtain neighborhood summaries, which are passed along with edge attributes to a deep neural network encoder module. The model is optimized using a compound loss function consisting of a contrastive learning one (to capture the graph structure) and a feature reconstruction loss (to ensure the edge attributes are encoded in the representation vector). Experiments showed that the proposed method build more powerful edge vector representations that achieved better performance in edge classification and edge clustering tasks than other state-of-the-art approaches.

The third publication (*FILDNE: A Framework for Incremental Learning of Dynamic Networks Embeddings*) proposes a novel framework for learning node representation vectors in dynamic graphs using an incremental learning approach. The model utilizes any provided static node representation learning method and applies it to the dynamic graph (modeled as a sequence of graph snapshots). Next, the framework aggregates the embedding vectors from consecutive snapshots using a linear convex combination function, whose parameters are estimated using a Dirichlet-Multinomial model based on an unsupervised link prediction task. The experimental evaluation showed that the proposed framework reduces memory and computational costs while providing competitive quality measure gains with respect to contemporary methods.

The fourth (*Retrofitting Structural Graph Embeddings with Node Attribute Information*) and fifth article (*A deeper look at Graph Embedding RetroFitting*) tackled the problem of updating (retrofitting) precomputed structural node representation vectors with node attribute information. First, a method based on a threefold objective function, consisting of an invariance loss (to preserve the structural embedding information), a graph neighbor loss (to increase the similarity of vectors of nodes connected by edges), and an attribute neighbor loss (to increase the similarity of vectors of nodes with similar attributes), was introduced. The experimental evaluation showed that the proposed method achieved better results in a node classification task than other attributed node representation learning methods. However, the method's hyperparameters required a manual adjustment, and the objective function contained redundant terms. Hence, the fifth article proposed an extension of the introduced method – it simplified the objective function and proposed an algorithm for automatic hyperparameter estimation. An extended experimental evaluation showed that the new method computed better representation vectors than existing approaches for attributed graphs.

To sum up, the results of studies conducted within this doctoral dissertation showed that there exist unsupervised representation learning methods for selected graph entities (nodes, edges) that compute better representation vectors than state-of-the-art unsupervised methods measured by means of downstream task evaluation.