

Second-Order Pooling for Graph Neural Networks

Zhengyang Wang and Shuiwang Ji, *Senior Member, IEEE*

Abstract—Graph neural networks have achieved great success in learning node representations for graph tasks such as node classification and link prediction. Graph representation learning requires graph pooling to obtain graph representations from node representations. It is challenging to develop graph pooling methods due to the variable sizes and isomorphic structures of graphs. In this work, we propose to use second-order pooling as graph pooling, which naturally solves the above challenges. In addition, compared to existing graph pooling methods, second-order pooling is able to use information from all nodes and collect second-order statistics, making it more powerful. We show that direct use of second-order pooling with graph neural networks leads to practical problems. To overcome these problems, we propose two novel global graph pooling methods based on second-order pooling; namely, bilinear mapping and attentional second-order pooling. In addition, we extend attentional second-order pooling to hierarchical graph pooling for more flexible use in GNNs. We perform thorough experiments on graph classification tasks to demonstrate the effectiveness and superiority of our proposed methods. Experimental results show that our methods improve the performance significantly and consistently.

Index Terms—Graph neural networks, graph pooling, second-order statistics.

1 INTRODUCTION

In recent years, deep learning has been widely explored on graph structured data, such as chemical compounds, protein structures, financial networks, and social networks [1], [2], [3]. Remarkable success has been achieved by generalizing deep neural networks from grid-like data to graphs [4], [5], [6], [7], resulting in the development of various graph neural networks (GNNs), like graph convolutional network (GCN) [8], GraphSAGE [9], graph attention network (GAT) [10], jumping knowledge network (JK) [11], and graph isomorphism networks (GINs) [12]. They are able to learn representations for each node in graphs and have set new performance records on tasks like node classification and link prediction [13]. In order to extend the success to graph representation learning, graph pooling is required, which takes node representations of a graph as inputs and outputs the corresponding graph representation.

While pooling is common in deep learning on grid-like data, it is challenging to develop graph pooling approaches due to the special properties of graphs. First, the number of nodes varies in different graphs, while the graph representations are usually required to have the same fixed size to fit into other machine learning models. Therefore, graph pooling should be capable of handling the variable number of node representations as inputs and producing fixed-sized graph representations. Second, unlike images and texts where we can order pixels and words according to the spatial structural information, there is no inherent ordering relationship among nodes in graphs. Indeed, we can set pseudo indices for nodes in a graph. However, an

isomorphism of the graph may change the order of the indices. As isomorphic graphs should have the same graph representation, it is required for graph pooling to create the same output by taking node representations in any order as inputs.

Some previous studies employ simple methods such as averaging and summation as graph pooling [12], [14], [15]. However, averaging and summation ignore the feature correlation information, hampering the overall model performance [16]. Other studies have proposed advanced graph pooling methods, including DIFFPOOL [17], SORTPOOL [16], TOPKPOOL [18], SAGPOOL [19], and EIGENPOOL [20]. DIFFPOOL maps nodes to a pre-defined number of clusters but is hard to train. EIGENPOOL involves the computation of eigenvectors, which is slow and expensive. SORTPOOL, SAGPOOL and TOPKPOOL rely on the top- K sorting to select a fixed number (K) of nodes and order them, during which the information from unselected nodes is discarded. It is worth noting that all the existing graph pooling methods only collect first-order statistics [21].

In this work, we propose to use second-order pooling as graph pooling. Compared to existing graph pooling methods, second-order pooling naturally solves the challenges of graph pooling and is more powerful with its ability of using information from all nodes and collecting second-order statistics. We analyze the practical problems in directly using second-order pooling with GNNs. To address the problems, we propose two novel and effective global graph pooling approaches based on second-order pooling; namely, bilinear mapping and attentional second-order pooling. In addition, we extend attentional second-order pooling to hierarchical graph pooling for more flexible use in GNNs. We perform thorough experiments on ten graph classification benchmark datasets. The experimental results show that our methods improve the performance significantly and

• Zhengyang Wang and Shuiwang Ji are with the Department of Computer Science and Engineering, Texas A&M University, College Station, TX, 77843.
E-mail: sji@tamu.edu

Manuscript received October, 2019.

consistently.

2 RELATED WORK

In this section, we review two categories of existing graph pooling methods in Section 2.1. Then in Section 2.2, we introduce what second-order statistics are, as well as their applications in both transitional machine learning and deep learning. In addition, we discuss the motivation of using second-order statistics in graph representation learning.

2.1 Graph Pooling: Global versus Hierarchical

Existing graph pooling methods can be divided into two categories according to their roles in graph neural networks (GNNs) for graph representation learning. One is global graph pooling, also known as graph readout operation [12], [19]. The other is hierarchical graph pooling, which is used to build hierarchical GNNs. We explain the details of the two categories and provide examples. In addition, we discuss advantages and disadvantages of the two categories.

Global graph pooling is typically used to connect embedded graphs outputted by GNN layers with classifiers for graph classification. Given a graph, GNN layers produce node representations, where each node is embedded as a vector. Global graph pooling is applied after GNN layers to process node representations into a single vector as the graph representation. A classifier takes the graph representation and performs graph classification. The “global” here refers to the fact that the output of global graph pooling encodes the entire graph. Global graph pooling is usually used only once in GNNs for graph representation learning. We call such GNNs as flat GNNs, in contrast to hierarchical GNNs. The most common global graph pooling methods include averaging and summation [12], [14], [15].

Hierarchical graph pooling is more similar to pooling in computer vision tasks [21]. The output of hierarchical graph pooling is a pseudo graph with fewer nodes than the input graph. It is used to build hierarchical GNNs, where hierarchical graph pooling is used several times between GNN layers to gradually decrease the number of nodes. The most representative hierarchical graph pooling methods are DIFFPOOL [17], SORTPOOL [16], TOPKPOOL [18], SAGPOOL [19], and EIGENPOOL [20]. A straightforward way to use hierarchical graph pooling for graph representation learning is to reduce the number of nodes to one. Then the resulted single vector is treated as the graph representation. Besides, there are two other ways to generate a single vector from the pseudo graph outputted by hierarchical graph pooling. One is introduced in SAGPOOL [19], where global and hierarchical graph pooling are combined. After each hierarchical graph pooling, global graph pooling with an independent classifier is employed. The final prediction is an average of all classifiers. On the other hand, SORTPOOL [16] directly applies convolutional neural networks (CNNs) to reduce the number of nodes to one. In particular, it takes advantage of a property of the pseudo graph outputted by hierarchical graph pooling. That is, the pseudo graph is a graph with a fixed number of nodes and there is an inherent ordering relationship among nodes determined by the trainable parameters in the hierarchical

graph pooling. Therefore, common deep learning methods like convolutions can be directly used. In fact, we can simply concatenate node presentations following the inherent order as the graph representation.

Given this property, most hierarchical graph pooling methods can be flexibly used as global graph pooling, with the three ways introduced above. For example, SORTPOOL [16] is used to build flat GNNs and applied only once after all GNN layers. While the idea of learning hierarchical graph representations makes sense, hierarchical GNNs do not consistently outperform flat GNNs [19]. In addition, with advanced techniques like jumping knowledge networks (JK-Net) [11] to address the over-smoothing problem of GNN layers [22], flat GNNs can go deeper and achieve better performance than hierarchical GNNs [12].

In this work, we first focus on global graph pooling as second-order pooling naturally fits this category. Later, we extend one of our proposed graph pooling methods to hierarchical graph pooling in Section 3.6.

2.2 Second-Order Statistics

In statistics, the k -order statistics refer to functions which use the k -th power of samples. Concretely, consider n samples (x_1, x_2, \dots, x_n) . The first and second moments, i.e., the mean $\mu = \frac{1}{n} \sum_i x_i$ and variance $\sigma^2 = \frac{1}{n} \sum_i (x_i - \mu)^2$, are examples of first and second-order statistics, respectively. If each sample is a vector, the covariance matrix is an example of second-order statistics. In terms of graph pooling, it is easy to see that existing methods are based on first-order statistics [21].

Second-order statistics have been widely explored in various computer vision tasks, such as face recognition, image segmentation, and object detection. In terms of traditional machine learning, the scale-invariant feature transform (SIFT) algorithm [23] utilizes second-order statistics of pixel values to describe local features in images and has become one of the most popular image descriptors. Tuzel et al. [24], [25] use covariance matrices of low-level features with boosting for detection and classification. The Fisher encoding [26] applies second-order statistics for recognition as well. Carreira et al. [27] employs second-order pooling for semantic segmentation. With the recent advances of deep learning, second-order pooling is also used in CNNs for fine-grained visual recognition [28] and visual question answering [29], [30], [31].

Many studies motivates the use of second-order statistics as taking advantage of the Riemannian geometry of the space of symmetric positive definite matrices [25], [27], [32]. In these studies, certain regularizations are cast to guarantee that the applied second-order statistics are symmetric positive definite [33], [34]. Other work relates second-order statistics to orderless texture descriptors for images [26], [28].

In this work, we propose to incorporate second-order statistics in graph representation learning. Our motivations lie in three aspects. First, second-order pooling naturally fits the goal and requirements of graph pooling, as discussed in Sections 3.1 and 3.2. Second, second-order pooling is able to capture the correlations among features, as well as topology information in graph representation learning,

as demonstrated in Section 3.2. Third, our proposed graph pooling methods based on second-order pooling are related to covariance pooling [24], [25], [33], [34] and attentional pooling [35] used in computer vision tasks, as pointed out in Section 3.5. In addition, we show that both covariance pooling and attentional pooling have certain limitations when employed in graph representation learning, and our proposed methods appropriately address them.

3 SECOND-ORDER POOLING FOR GRAPHS

In this section, we introduce our proposed second-order pooling methods for graph representation learning. First, we formally define the aim and requirements of graph pooling in Section 3.1. Then we propose to use second-order pooling as graph pooling, analyze its advantages, and point out practical problems when directly using it with GNNs in Section 3.2. In order to address the problems, we propose two novel second-order pooling methods for graphs in Sections 3.3 and 3.4, respectively. Afterwards, we discuss why our proposed methods are more suitable as graph pooling compared to two similar pooling methods in image tasks in Section 3.5. Finally, while both methods focus on global graph pooling, we extend second-order pooling to hierarchical graph pooling in Section 3.6.

3.1 Properties of Graph Pooling

Consider a graph $G = (A, X)$ represented by its adjacency matrix $A \in \{0, 1\}^{n \times n}$ and node feature matrix $X \in \mathbb{R}^{n \times d}$, where n is the number of nodes in G and d is the dimension of node features. The node features may come from node labels or node degrees. Graph neural networks (GNNs) are known to be powerful in learning good node representation matrix H from A and X :

$$H = [h_1, h_2, \dots, h_n]^T = \text{GNN}(A, X) \in \mathbb{R}^{n \times f}, \quad (1)$$

where rows of H , $h_i \in \mathbb{R}^f, i = 1, 2, \dots, n$, are representations of n nodes, and f depends on the architecture of GNNs. The task that we focus on in this work is to obtain a graph representation vector h_G from H , which is then fed into a classifier to perform graph classification:

$$h_G = g([A], H) \in \mathbb{R}^c, \quad (2)$$

where $g(\cdot)$ is the graph pooling function and c is the dimension of h_G . Here, $[A]$ means that the information from A can be optionally used in graph pooling. For simplicity, we omit it in the following discussion.

Note that $g(\cdot)$ must satisfy two requirements to serve as graph pooling. First, $g(\cdot)$ should be able to take H with variable number of rows as the inputs and produce fixed-sized outputs. Specifically, different graphs may have different number of nodes, which means that n is a variable. On the other hand, c is supposed to be fixed to fit into the following classifier.

Second, $g(\cdot)$ should output the same h_G when the order of rows of H changes. This permutation invariance property is necessary to handle isomorphic graphs. To be concrete, if two graph $G_1 = (A_1, X_1)$ and $G_2 = (A_2, X_2)$ are isomorphic, GNNs will output the same multiset of node representations [12], [16]. That is, there exists a permutation matrix $P \in \{0, 1\}^{n \times n}$ such that $H_1 = PH_2$, for

$H_1 = \text{GNN}(A_1, X_1)$ and $H_2 = \text{GNN}(A_2, X_2)$. However, the graph representation computed by $g(\cdot)$ should be the same, i.e., $g(H_1) = g(H_2)$ if $H_1 = PH_2$.

3.2 Second-Order Pooling

In this work, we propose to employ second-order pooling [27], also known as bilinear pooling [28], as graph pooling. We show that second-order pooling naturally satisfies the two requirements above.

We start by introducing the definition of second-order pooling.

Definition. Given $H = [h_1, h_2, \dots, h_n]^T \in \mathbb{R}^{n \times f}$, second-order pooling (SOPool) is defined as

$$\text{SOPool}(H) = \sum_{i=1}^n h_i h_i^T = H^T H \in \mathbb{R}^{f \times f}. \quad (3)$$

In terms of graph pooling, we can view $\text{SOPool}(H)$ as an f^2 -dimensional graph representation vector by simply flattening the matrix. Another way to transform the matrix into a vector is discussed in Section 3.4. Note that, as long as SOPool meets the two requirements, the way to transform the matrix into a vector does not affect its eligibility as graph pooling.

Now let us check the two requirements.

Proposition 1. SOPool always outputs an $f \times f$ matrix for $H \in \mathbb{R}^{n \times f}$, regardless of the value of n .

Proof. The result is obvious since the dimension of $H^T H$ does not depend on n . \square

Proposition 2. SOPool is invariant to permutation so that it outputs the same matrix when the order of rows of H changes.

Proof. Consider $H_1 = PH_2$, where P is a permutation matrix. Note that we have $P^T P = I$ for any permutation matrix. Therefore, it is easy to derive

$$\begin{aligned} \text{SOPool}(H_1) &= H_1^T H_1 \\ &= (PH_2)^T (PH_2) \\ &= H_2^T P^T P H_2 \\ &= H_2^T H_2 = \text{SOPool}(H_2). \end{aligned} \quad (4)$$

This completes the proof. \square

In addition to satisfying the requirements of graph pooling, SOPool is capable of capturing second-order statistics, which are much more discriminative than first-order statistics computed by most other graph pooling methods [27], [28], [29]. In detail, the advantages can be seen from two aspects. On one hand, we can tell from $\text{SOPool}(H) = \sum_{i=1}^n h_i h_i^T$ that, for each node representation h_i , the features interact with each other, enabling the correlations among features to be captured. On the other hand, topology information is encoded as well. Specifically, we view $H \in \mathbb{R}^{n \times f}$ as $H = [l_1, l_2, \dots, l_f]$, where $l_j \in \mathbb{R}^n, j = 1, 2, \dots, f$. The vector l_j encodes the spatial distribution of the j -th feature in the graph. Based on this view, $\text{SOPool}(H) = H^T H$ is able to capture the topology information.

However, we point out that the direct application of second-order pooling in GNNs leads to practical problems.

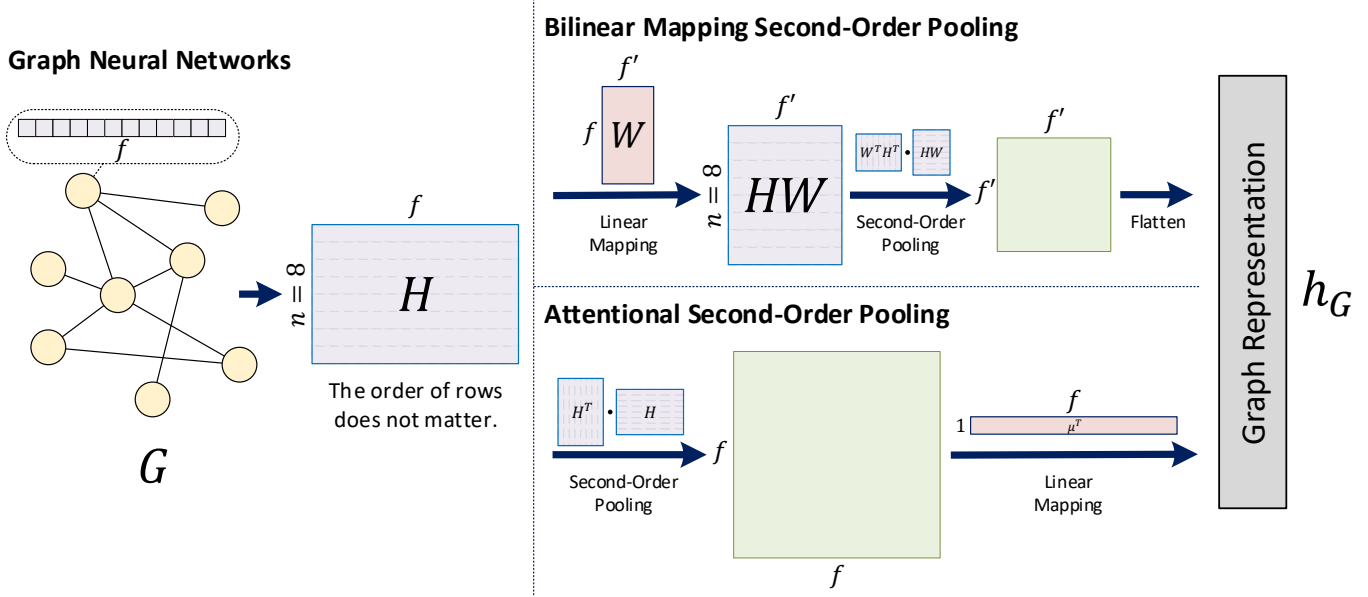


Fig. 1. Illustrations of our proposed graph pooling methods: bilinear mapping second-order pooling (SOPool_{bimap}) in Section 3.3 and attentional second-order pooling (SOPool_{attn}) in Section 3.4. This is an example for a graph G with $n = 8$ nodes. GNNs can learn representations for each node and graph pooling processes node representations into a graph representation vector h_G .

The direct way to use second-order pooling as graph pooling is represented as

$$h_G = \text{FLATTEN}(\text{SOPool}(\text{GNN}(A, X))) \in \mathbb{R}^{f^2}. \quad (5)$$

That is, we apply SOPool on $H = \text{GNN}(A, X)$ and flatten the output matrix into an f^2 -dimensional graph representation vector. However, it causes an explosion in the number of training parameters in the following classifier when f is large, making the learning process harder to converge and easier to overfit. While each layer in a GNN usually has outputs with a small number of hidden units (e.g. 16, 32, 64), it has been pointed out that graph representation learning benefits from using information from outputs of all layers, obtaining better performance and generalization ability [11]. It is usually achieved by concatenating outputs across all layers in a GNN [12], [16]. In this case, H has a large final f , making direct use of second-order pooling infeasible. For example, if a GNN has 5 layers and each layer's outputs have 32 hidden units, f becomes $32 \times 5 = 160$. Suppose h_G is sent into a 1-layer fully-connected classifier for c graph categories in a graph classification task. It results in $160^2 c = 25,600c$ training parameters, which is excessive. We omit the bias term for simplicity.

3.3 Bilinear Mapping Second-Order Pooling

To address the above problem, a straightforward solution is to reduce f in H before $\text{SOPool}(H)$. Based on this, our first proposed graph pooling method, called bilinear mapping second-order pooling (SOPool_{bimap}), employs a linear mapping on H to perform dimensionality reduction. Specifically, it is defined as

$$\begin{aligned} \text{SOPool}_{bimap}(H) &= \text{SOPool}(HW) \\ &= W^T H^T H W \in \mathbb{R}^{f' \times f'}, \end{aligned} \quad (6)$$

where $f' < f$ and $W \in \mathbb{R}^{f \times f'}$ is a trainable matrix representing a linear mapping. Afterwards, we follow the same process to flatten the matrix and obtain an f'^2 -dimensional graph representation vector:

$$h_G = \text{FLATTEN}(\text{SOPool}_{bimap}(\text{GNN}(A, X))) \in \mathbb{R}^{f'^2}. \quad (7)$$

Figure 1 provides an illustration of the above process. By selecting an appropriate f' , the bilinear mapping second-order pooling does not suffer from the excessive number of training parameters. Taking the example above, if we set $f' = 32$, the total number of parameters in SOPool_{bimap} and a following 1-layer fully-connected classifier is $32 \times 160 + 32^2 c = 5,120 + 1,024c$, which is much smaller than $25,600c$.

3.4 Attentional Second-Order Pooling

Our second proposed graph pooling method tackles with the problem by exploring another way to transform the matrix computed by SOPool into the graph representation vector, instead of simply flattening. Similarly, we use a linear mapping to perform the transformation, defined as

$$h_G = \text{SOPool}(\text{GNN}(A, X)) \cdot \mu \in \mathbb{R}^f, \quad (8)$$

where $\mu \in \mathbb{R}^f$ is a trainable vector. It is interesting to note that $h_G = H^T H \mu$, which is similar to the sentence attention in [36]. To be concrete, consider a word embedding matrix $E = [e_1, e_2, \dots, e_l]^T \in \mathbb{R}^{l \times d_w}$ for a sentence, where l is the number of words and d_w is the dimension of word embeddings. The sentence attention is defined as

$$\alpha_i = \frac{\exp(e_i^T \mu_s)}{\sum_{j=1}^l \exp(e_j^T \mu_s)}, i = 1, 2, \dots, l \quad (9)$$

$$s = \sum_{i=1}^l \alpha_i e_i, \quad (10)$$

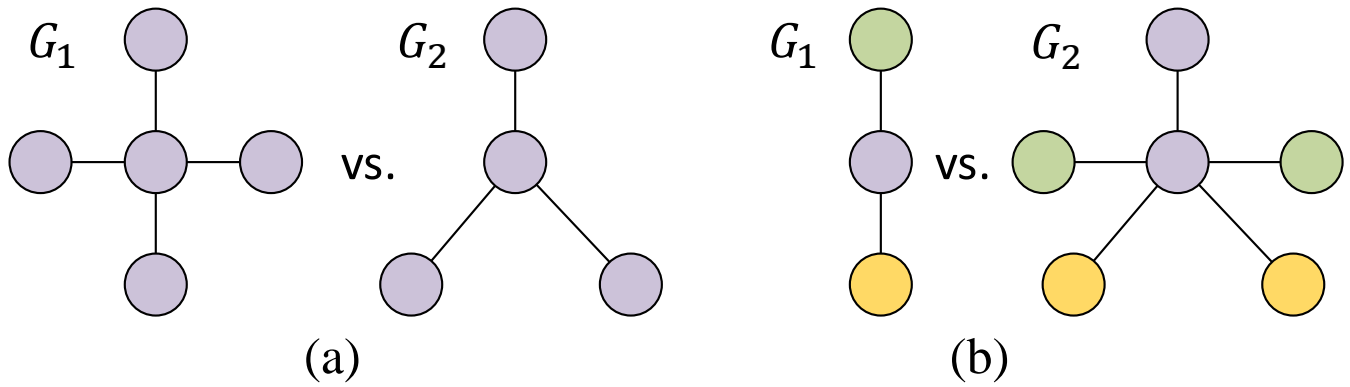


Fig. 2. Examples of graphs that pooling methods discussed in Section 3.5 fail to distinguish, *i.e.*, producing the same graph representation for different graphs G_1 and G_2 . The same color denotes the same node representation. (a) Covariance pooling (COVPOOL) and attentional pooling (ATTNPOOL) both fail. COVPOOL fails because subtracting the mean results in $h_{G_1} = h_{G_2} = 0$. ATTNPOOL computes the mean of node representations, leading to $h_{G_1} = h_{G_2}$ as well. (b) ATTNPOOL fails in this example with the same μ .

where $\mu_s \in \mathbb{R}^{d_w}$ is a trainable vector and s is the resulted sentence embedding. Note that Eqn. (9) is the SOFTMAX function and serves as a normalization function [37]. Rewriting the sentence attention into matrix form, we have $s = E^T \text{SOFTMAX}(E\mu_s)$. The only difference between the computation of h_G and that of s is the normalization function. Therefore, we name our second proposed graph pooling method as attentional second-order pooling (SOPOOL_{attn}), defined as

$$\text{SOPOOL}_{\text{attn}}(H) = H^T H \mu \in \mathbb{R}^f, \quad (11)$$

where $\mu \in \mathbb{R}^f$ is a trainable vector. It is illustrated in Figure 1. We take the same example above to show that SOPOOL_{attn} reduces the number of training parameters. The total number of parameters in SOPOOL_{attn} and a following 1-layer fully-connected classifier is just $160 + 160c$, significantly reducing the amount of parameters compared to 25,600c.

3.5 Relationships to Covariance Pooling and Attentional Pooling

The experimental results in Section 4 show that both our proposed graph pooling methods achieve better performance significantly and consistently than previous studies. However, we note that, there are pooling methods in image tasks that have similar computation processes to our proposed methods, although they have not been developed based on second-order pooling. In this section, we point out the key differences between these methods and ours and show why they matter in graph representation learning.

Note that images are usually processed by deep neural networks into feature maps $I \in \mathbb{R}^{h \times w \times c}$, where h , w , c are the height, width, and number of feature maps, respectively. Following [33], [34], [35], we reshape I into the matrix $H \in \mathbb{R}^{n \times f}$, where $n = hw$ and $f = c$ so that different pooling methods can be compared directly.

Covariance pooling. Covariance pooling (COVPOOL) [24], [25], [33], [34] has been widely explored in image tasks, such as image categorization, facial

expression recognition, and texture classification. Recently, it has also been explored in GNNs [38]. The definition is

$$\text{COVPOOL}(H) = (H - \mathbf{1}\bar{H})^T (H - \mathbf{1}\bar{H}) \in \mathbb{R}^{f \times f}, \quad (12)$$

where $\mathbf{1}$ is the n -dimensional all-one vector and $\bar{H} \in \mathbb{R}^{1 \times f}$ is the mean of rows of H . It differs from SOPOOL defined in Eqn. (3) only in whether to subtract the mean. However, subtracting the mean makes COVPOOL less powerful in terms of distinguishing graphs with repeating node embeddings [12], which may cause the performance loss. Figure 2(a) gives an example of this problem.

Attentional pooling. Attentional pooling (ATTNPOOL) [35] has been used in action recognition. As shown in Section 3.4, it is also used in text classification [36], defined as

$$\text{ATTNPOOL}(H) = H^T \text{SOFTMAX}(H\mu) \in \mathbb{R}^f, \quad (13)$$

where $\mu \in \mathbb{R}^f$ is a trainable vector. It differs from SOPOOL_{attn} only in the SOFTMAX function. We show that the SOFTMAX function leads to similar problems as other normalization functions, such as mean and max-pooling [12]. Figure 2 provides examples in which ATTNPOOL does not work.

To conclude, our methods derived from second-order pooling are more suitable as graph pooling. We compare these pooling methods through experiments in Section 4.3. The results show that COVPOOL and ATTNPOOL suffer from significant performance loss on some datasets.

3.6 Multi-Head Attentional Second-Order Pooling

The proposed SOPOOL_{bimap} and SOPOOL_{attn} belong to the global graph pooling category. As discussed in Section 2.1, they are used in flat GNNs after all GNN layers and output the graph representation for the classifier. While flat GNNs outperform hierarchical GNNs in most benchmark datasets [12], developing hierarchical graph pooling is still desired, especially for large graphs [17], [19], [20]. Therefore, we explore a hierarchical graph pooling method based on second-order pooling.

Unlike global graph pooling, hierarchical graph pooling outputs multiple vectors corresponding to node representations in the pooled graph. In addition, hierarchical graph pooling has to update the adjacency matrix to indicate how nodes are connected in the pooled graph. **To be specific, given the adjacency matrix $A \in \mathbb{R}^{n \times n}$ and node representation matrix $H \in \mathbb{R}^{n \times f}$, a hierarchical graph pooling function $g_h(\cdot)$ can be written as**

$$A', H' = g_h(A, H), \quad (14)$$

where $A' \in \mathbb{R}^{k \times k}$ and $H' \in \mathbb{R}^{k \times f}$. Here, k is a hyperparameter determining the number of nodes in the pooled graph. Note that Eqn. (14) does not conflict with Eqn. (2), as we can always transform H' into a vector h_G , as discussed in Section 2.1.

We note that the proposed $\text{SOPool}_{\text{attn}}$ in Section 3.4 is closely related to the attention mechanism and can be easily extended to a hierarchical graph pooling method based on the multi-head technique in the attention mechanism [10], [37]. The multi-head technique means that multiple independent attentions are performed on the same inputs. Then the outputs of multiple attentions are then concatenated together. Based on this insight, we propose multi-head attentional second-order pooling ($\text{SOPool}_{\text{m_attn}}$), defined as

$$H' = \text{SOPool}_{\text{m_attn}}(H) = UH^T H \in \mathbb{R}^{k \times f}, \quad (15)$$

where $U \in \mathbb{R}^{k \times f}$ is a trainable matrix. To illustrate its relationship to the multi-head technique, we can equivalently write it as

$$\text{SOPool}_{\text{m_attn}}(H) = [H^T H \mu_1, \dots, H^T H \mu_k]^T, \quad (16)$$

where we decompose U in Eqn. (15) as $U = [\mu_1, \mu_2, \dots, \mu_k]^T$. The relationship can be easily seen by comparing Eqn. (16) with Eqn. (11).

The multi-head technique enables $\text{SOPool}_{\text{m_attn}}$ to output the node representation matrix for the pooled graph. We now describe how to update the adjacency matrix. **In particular, we employ a contribution matrix C in updating the adjacency matrix. The contribution matrix is a $k \times n$ matrix, whose entries indicate how nodes in the input graph contribute to nodes in the pooled graph. In $\text{SOPool}_{\text{m_attn}}$, we can simply let $C = UH^T \in \mathbb{R}^{k \times n}$. With the contribution matrix C , the corresponding adjacency matrix A' of the pooled graph can be computed as**

$$A' = CAC^T \in \mathbb{R}^{k \times k}. \quad (17)$$

The proposed $\text{SOPool}_{\text{m_attn}}$ is closely related to DiffPool [17]. The contribution matrix C corresponds to the assignment matrix in DiffPool . However, DiffPool applied GNN layers with normalization on H to obtain C , preventing the explicit use of second-order statistics. In the experiments, we evaluate $\text{SOPool}_{\text{m_attn}}$ as both global and hierarchical graph pooling methods, in flat and hierarchical GNNs, respectively.

4 EXPERIMENTS

We conduct thorough experiments on graph classification tasks to show the effectiveness of our proposed

graph pooling methods, namely bilinear mapping second-order pooling ($\text{SOPool}_{\text{bimap}}$), attentional second-order pooling ($\text{SOPool}_{\text{attn}}$), and multi-head attentional second-order pooling ($\text{SOPool}_{\text{m_attn}}$). Section 4.1 introduces the datasets, baselines, and experimental setups for reproduction. The following sections aim at evaluating our proposed methods in different aspects, by answering the questions below:

- Can GNNs with our proposed methods achieve improved performance in graph classification tasks? Section 4.2 provides the comparison results between our methods and existing methods in graph classification tasks.
- Do our proposed methods outperform existing global graph pooling methods with the same flat GNN architecture? The ablation studies in Section 4.3 compare different graph pooling methods with the same GNN, eliminating the influences of different GNNs. In particular, we use hierarchical graph pooling methods as global graph pooling methods in this experiment, including $\text{SOPool}_{\text{m_attn}}$.
- Is the improvement brought by our proposed method consistent with various GNN architectures? Section 4.4 shows the performance of the proposed $\text{SOPool}_{\text{bimap}}$ and $\text{SOPool}_{\text{attn}}$ with different GNNs.
- Is $\text{SOPool}_{\text{m_attn}}$ effective as hierarchical graph pooling methods? We compare $\text{SOPool}_{\text{m_attn}}$ with other hierarchical graph pooling methods in the same hierarchical GNN architecture in Section 4.5.

4.1 Experimental Setup

Reproducibility. The code used in our experiments is available at <https://github.com/divelab/sopool>. Details of datasets and parameter settings are described below.

Datasets. We use ten graph classification datasets from [1], including five bioinformatics datasets (MUTAG, PTC, PROTEINS, NCI1, DD) and five social network datasets (COLLAB, IMDB-BINARY, IMDB-MULTI, REDDIT-BINARY, REDDIT-MULTI5K). Note that only bioinformatics datasets come with node labels. Below are the detailed descriptions of datasets:

- MUTAG is a bioinformatics dataset of 188 graphs representing nitro compounds. Each node is associated with one of 7 discrete node labels. The task is to classify each graph by determining whether the compound is mutagenic aromatic or heteroaromatic [39].
- PTC [40] is a bioinformatics dataset of 344 graphs representing chemical compounds. Each node comes with one of 19 discrete node labels. The task is to predict the rodent carcinogenicity for each graph.
- PROTEINS [41] is a bioinformatics dataset of 1,113 graph structures of proteins. Nodes in the graphs refer to secondary structure elements (SSEs) and have discrete node labels indicating whether they represent a helix, sheet or turn. And edges mean that two nodes are neighbors along the amino-acid sequence or in space. The task is to predict the protein function for each graph.

- NCI [42] is a bioinformatics dataset of 4,110 graphs representing chemical compounds. It contains data published by the National Cancer Institute (NCI). Each node is assigned with one of 37 discrete node labels. The graph classification label is decided by NCI anti-cancer screens for ability to suppress or inhibit the growth of a panel of human tumor cell lines.
- COLLAB is a scientific collaboration dataset of 5,000 graphs corresponding to ego-networks generated using the method in [43]. The dataset is derived from 3 public collaboration datasets [44]. Each ego-network contains different researchers from each field and is labeled by the corresponding field. The three fields are High Energy Physics, Condensed Matter Physics, and Astro Physics.
- IMDB-BINARY is a movie collaboration dataset of 1,000 graphs representing ego-networks for actors/actresses. The dataset is derived from collaboration graphs on Action and Romance genres. In each graph, nodes represent actors/actresses and edges simply mean they collaborate the same movie. The graphs are labeled by the corresponding genre and the task is to identify the genre for each graph.
- IMDB-MULTI is multi-class version of IMDB-BINARY. It contains 1,500 ego-networks and has three extra genres, namely, Comedy, Romance and Sci-Fi.
- REDDIT-BINARY is a dataset of 2,000 graphs where each graph represents an online discussion thread. Nodes in a graph correspond to users appearing in the corresponding discussion thread and an edge means that one user responded to another. Datasets are crawled from top submissions under four popular subreddits, namely, IAmA, AskReddit, TrollXChromosomes, atheism. Among them, AmA and AskReddit are question/answer-based subreddits while TrollXChromosomes and atheism are discussion-based subreddits, forming two classes to be classified.
- REDDIT-MULTI5K is a similar dataset as REDDIT-BINARY, which contains 5,000 graphs. The difference lies in that REDDIT-MULTI5K crawled data from five different subreddits, namely, worldnews, videos, AdviceAnimals, aww and mildlyinteresting. And the task is to identify the subreddit of each graph instead of determining the type of subreddits.
- DD [45] is a bioinformatics dataset of 1,178 graph structures of proteins. Nodes in the graphs represent amino acids. And edges connect nodes that are less than 6 Ångströms apart. The task is a two-way classification task between enzymes and non-enzymes. DD is only used in Section 4.5. The average number of nodes in DD is 284.3.

More statistics of these datasets are provided in the “datasets” section of Table 1. The input node features are different for different datasets. For bioinformatics datasets, the nodes have categorical labels as input features. **For social network datasets, we create node features. To be specific, we set all node feature vectors to be the same for REDDIT-**

BINARY and REDDIT-MULTI5K [12]. And for the other social network datasets, we use one-hot encoding of node degrees as features.

Configurations. In Sections 4.2, 4.3 and 4.4, the flat GNNs we use with our proposed graph pooling methods are graph isomorphism networks (GINs) [12]. The original GINs employ averaging or summation (SUM/AVG) as the graph pooling function; specifically, summation on bioinformatics datasets and averaging on social datasets. We replace averaging or summation with our proposed graph pooling methods and keep other parts the same. There are seven variants of GINs, two of which are equivalent to graph convolutional network (GCN) [8] and GraphSAGE [9], respectively. In Sections 4.2 and 4.3, we use GIN-0 with our methods. In Section 4.4, we examine our methods with all variants of GINs. Details of all variants can be found in Section 4.4.

The hierarchical GNNs used in Section 4.5 follow the hierarchical architecture in [19], allowing direct comparisons. To be specific, each block is composed of one GNN layer followed by a hierarchical graph pooling. After each hierarchical pooling, a classifier is used. The final prediction is the combination of all classifiers.

Training & Evaluation. Following [1], [46], model performance is evaluated using 10-fold cross-validation and reported as the average and standard deviation of validation accuracies across the 10 folds. For the flat GNNs, we follow the same training process in [12]. All GINs have 5 layers. Each multi-layer perceptron (MLP) has 2 layers with batch normalization [47]. For the hierarchical GNNs, we follow the the same training process in [19]. **There are three blocks in total. Dropout [48] is applied in the classifiers. The Adam optimizer [49] is used with the learning rate initialized as 0.01 and decayed by 0.5 every 50 epochs. The number of total epochs is selected according to the best cross-validation accuracy. We tune the number of hidden units (16, 32, 64) and the batch size (32, 128) using grid search.**

Baselines. We compare our methods with various graph classification models as baselines, including both kernel-based and GNN-based methods. The kernel-based methods are graphlet kernel (GK) [50], random walk kernel (RW) [51], Weisfeiler-Lehman subtree kernel (WL) [52], deep graphlet kernel (DGK) [1], and anonymous walk embeddings (AWE) [53]. Among them, DGK and AWE use deep learning methods as well. The GNN-based methods are diffusion-convolutional neural network (DCNN) [54], PATCHSCAN [46], ECC [55], deep graph CNN (DGCNN) [16], differentiable pooling (DIFFPOOL) [17], graph capsule CNN (GCAPS-CNN) [38], self-attention graph pooling (SAGPOOL) [19], GIN [12], and eigenvector-based pooling (EigenGCN) [20]. We report the performance of these baselines provided in [12], [16], [17], [19], [20], [38].

4.2 Comparison with Baselines

The comparison results between our methods and baselines are reported in Table 1. GIN-0 equipped with our proposed graph pooling methods, “GIN-0 + SOPOOL_{attn}”, “GIN-0 + SOPOOL_{bimap}”, and “GIN-0 + SOPOOL_{m_attn}”, outperform all the baselines significantly on seven out of

TABLE 1

Comparison results between our proposed methods and baselines described in Section 4.1. We report the accuracies of these baselines provided in [12], [16], [17], [20], [38]. The best models are highlighted with boldface. If a kernel-based baseline performs the best than all GNN-based models, we highlight the best GNN-based model with boldface and the best kernel-based baseline with boldface and asterisk.

Datasets		MUTAG	PTC	PROTEINS	NCI1	COLLAB	IMDB-B	IMDB-M	RDT-B	RDT-M5K
Datasets	# graphs	188	344	1113	4110	5000	1000	1500	2000	5000
	# classes	2	2	2	2	3	2	3	2	5
	# nodes (max)	28	109	620	111	492	136	89	3783	3783
	# nodes (avg.)	18.0	25.6	39.1	29.9	74.5	19.8	13.0	429.6	508.5
Kernel	GK [2009]	81.4±1.7	57.3±1.4	71.7±0.6	62.3±0.3	72.8±0.3	65.9±1.0	43.9±0.4	77.3±0.2	41.0±0.2
	RW [2010]	79.2±2.1	57.9±1.3	74.2±0.4	>1 day	-	-	-	-	-
	WL [2011]	90.4±5.7	59.9±4.3	75.0±3.1	86.0±1.8*	78.9±1.9	73.8±3.9	50.9±3.8	81.0±3.1	52.5±2.1
	DGK [2015]	-	60.1±2.6	75.7±0.5	80.3±0.5	73.1±0.3	67.0±0.6	44.6±0.5	78.0±0.4	41.3±0.2
	AWE [2018]	87.9±9.8	-	-	-	73.9±1.9	74.5±5.9	51.5±3.6	87.9±2.5	54.7±2.9
GNN	DCNN [2016]	67.0	56.6	61.3	56.6	52.1	49.1	33.5	-	-
	PATCHSCAN [2016]	92.6±4.2	60.0±4.8	75.9±2.8	78.6±1.9	72.6±2.2	71.0±2.2	45.2±2.8	86.3±1.6	49.1±0.7
	ECC [2017]	-	-	72.7	76.8	67.8	-	-	-	-
	DGCNN [2018]	85.8±1.7	58.6±2.5	75.5±1.0	74.4±0.5	73.8±0.5	70.0±0.9	47.8±0.9	76.0±1.7	48.7±4.5
	DIFFPOOL [2018]	80.6	-	76.3	76.0	75.5	-	-	-	-
	GCAPS-CNN [2018]	-	66.0±5.9	76.4±4.2	82.7±2.4	77.7±2.5	71.7±3.4	48.5±4.1	87.6±2.5	50.1±1.7
	GIN-0 + SUM/AVG [2018]	89.4±5.6	64.6±7.0	76.2±2.8	82.7±1.7	80.2±1.9	75.1±5.1	52.3±2.8	92.4±2.5	57.5±1.5
	EigenGCN [2019]	79.5	-	76.6	77.0	-	-	-	-	-
Ours	GIN-0 + SOPOOL _{attn}	93.6±4.1	72.9±6.2	79.4±3.2	82.8±1.4	81.1±1.8	78.1±4.0	54.3±2.6	91.7±2.7	58.3±1.4
	GIN-0 + SOPOOL _{bimap}	95.3±4.4	75.0±4.3	80.1±2.7	83.6±1.4	79.9±1.9	78.4±4.7	54.6±3.6	89.6±3.3	58.4±1.6
	GIN-0 + SOPOOL _{m_attn}	95.2±5.4	74.4±5.5	79.5±3.1	84.5±1.3	77.6±1.9	78.5±2.8	54.3±2.1	90.0±0.8	55.8±2.2

TABLE 2

Comparison results between our proposed methods and other graph pooling methods by fixing the GNN before graph pooling to GIN-0, as described in Section 4.3. The best models are highlighted with boldface.

Models	MUTAG	PTC	PROTEINS	NCI1	COLLAB	IMDB-B	IMDB-M	RDT-B	RDT-M5K
GIN-0 + SUM/AVG	89.4±5.6	64.6±7.0	76.2±2.8	82.7±1.7	80.2±1.9	75.1±5.1	52.3±2.8	92.4±2.5	57.5±1.5
GIN-0 + DIFFPOOL	94.8±4.8	66.1±7.7	78.8±3.1	76.6±1.3	75.3±2.2	74.4±4.0	50.1±3.2	-	-
GIN-0 + SORTPOOL	95.2±3.9	69.5±6.3	79.2±3.0	78.9±2.7	78.2±1.6	77.5±2.7	53.1±2.9	81.6±4.6	48.4±4.8
GIN-0 + TOPKPOOL	94.7±3.5	68.4±6.4	79.1±2.2	79.6±1.7	79.6±2.1	77.8±5.1	53.7±2.8	-	-
GIN-0 + SAGPOOL	93.9±3.3	69.0±6.6	78.4±3.1	79.0±2.8	78.9±1.7	77.8±2.9	53.1±2.8	-	-
GIN-0 + ATTNPOOL	93.2±5.8	71.2±8.0	77.5±3.3	80.6±2.1	81.8±2.2	77.1±4.4	53.8±2.5	92.5±2.3	57.9±1.7
GIN-0 + SOPOOL _{attn}	93.6±4.1	72.9±6.2	79.4±3.2	82.8±1.4	81.1±1.8	78.1±4.0	54.3±2.6	91.7±2.7	58.3±1.4
GIN-0 + COVPOOL	95.3±3.7	73.3±5.1	80.1±2.2	83.5±1.9	79.3±1.8	72.1±5.1	47.8±2.7	90.3±3.6	58.4±1.7
GIN-0 + SOPOOL _{bimap}	95.3±4.4	75.0±4.3	80.1±2.7	83.6±1.4	79.9±1.9	78.4±4.7	54.6±3.6	89.6±3.3	58.4±1.6
GIN-0 + SOPOOL _{m_attn}	95.2±5.4	74.4±5.5	79.5±3.1	84.5±1.3	77.6±1.9	78.5±2.8	54.3±2.1	90.0±0.8	55.8±2.2

nine datasets. On NCI1, WL has better performance than all GNN-based models. However, “GIN-0 + SOPOOL_{m_attn}” is the second best model and has improved performance over other GNN-based models. On REDDIT-BINARY, our methods achieve comparable performance to the best one.

It is worth noting that the baseline “GIN-0 + SUM/AVG” is the previous state-of-the-art model [12]. Our methods differ from it only in the graph pooling functions. The significant improvement demonstrates the effectiveness of our proposed graph pooling methods. In the next section, we compare our methods with other graph pooling methods by fixing the GNN before graph pooling to GIN-0, in order to eliminate the influences of different GNNs.

4.3 Ablation Studies in Flat Graph Neural Networks

We perform ablation studies to show that our proposed methods are superior to other global graph pooling methods under a fair setting. Starting from the baseline “GIN-0 + SUM/AVG”, we replace SUM/AVG with different graph pooling methods and keep all other configurations

unchanged. The graph pooling methods we include are DIFFPOOL [17], SORTPOOL from DGCNN [16], TOPKPOOL from Graph U-Net [18], SAGPOOL [19], and COVPOOL and ATTNPOOL described in Section 3.5. DIFFPOOL, TOPKPOOL, and SAGPOOL are used as hierarchical graph pooling methods in their works, but they achieve good performance as global pooling methods as well [17], [19]. EIGENPOOL from EigenGCN suffers from significant performance loss as a global pooling method [20] so that we do not include it in the ablation studies. COVPOOL and ATTNPOOL use the same settings as our proposed methods.

Table 2 provides the comparison results. Our proposed SOPOOL_{bimap} and SOPOOL_{attn} achieve better performance than DIFFPOOL, SORTPOOL, TOPKPOOL, and SAGPOOL on all datasets, demonstrating the effectiveness of our graph pooling methods with second-order statistics.

To support our discussion in Section 3.5, we analyze the performance of COVPOOL and ATTNPOOL. Note that the same bilinear mapping technique used in SOPOOL_{bimap} is applied on COVPOOL, in order to avoid the excessive number of parameters. COVPOOL achieves comparable per-

TABLE 3

Results of our proposed methods with different GNNs before graph pooling, as described in Section 4.4. The architectures of different GNNs come from variants of GINs in [12], whose details can be found in the supplementary material. The best models are highlighted with boldface.

GNNs	POOLS	MUTAG	PTC	PROTEINS	NCI1	COLLAB	IMDB-B	IMDB-M
SUM-MLP (GIN-0)	SUM/AVG	89.4±5.6	64.6±7.0	76.2±2.8	82.7±1.7	80.2±1.9	75.1±5.1	52.3±2.8
	SOPOOL _{attn}	93.6±4.1	72.9±6.2	79.4±3.2	82.8±1.4	81.1±1.8	78.1±4.0	54.3±2.6
	SOPOOL _{bimap}	95.3±4.4	75.0±4.3	80.1±2.7	83.6±1.4	79.9±1.9	78.4±4.7	54.6±3.6
SUM-MLP (GIN- ϵ)	SUM/AVG	89.0±6.0	63.7±8.2	75.9±3.8	82.7±1.6	80.1±1.9	74.3±5.1	52.1±3.6
	SOPOOL _{attn}	92.6±5.4	73.6±5.5	79.2±1.9	83.1±1.8	80.6±1.6	78.1±4.3	55.4±3.7
	SOPOOL _{bimap}	93.7±5.3	73.5±7.0	79.3±1.8	83.6±1.4	80.4±2.4	77.5±4.5	54.5±3.5
SUM-1-LAYER	SUM/AVG	90.0±8.8	63.1±5.7	76.2±2.6	82.0±1.5	80.6±1.9	74.1±5.0	52.2±2.4
	SOPOOL _{attn}	94.2±4.4	73.6±6.5	79.0±2.9	81.2±1.5	81.2±1.6	78.6±4.1	54.5±3.0
	SOPOOL _{bimap}	95.8±4.2	71.8±6.1	80.1±2.5	82.4±1.3	80.5±2.0	78.2±3.6	54.1±3.4
MEAN-MLP	SUM/AVG	83.5±6.3	66.6±6.9	75.5±3.4	80.9±1.8	79.2±2.3	73.7±3.7	52.3±3.1
	SOPOOL _{attn}	92.6±4.5	74.9±6.6	79.4±2.8	80.6±1.1	80.0±2.0	77.5±3.9	55.2±3.3
	SOPOOL _{bimap}	90.4±6.2	72.7±4.0	79.3±2.4	81.1±1.6	80.4±1.7	77.9±4.7	55.0±3.7
MEAN-1-LAYER (GCN)	SUM/AVG	85.6±5.8	64.2±4.3	76.0±3.2	80.2±2.0	79.0±1.8	74.0±3.4	51.9±3.8
	SOPOOL _{attn}	90.0±5.1	76.7±5.6	78.5±2.8	78.0±1.8	80.2±1.6	78.9±4.2	54.8±3.1
	SOPOOL _{bimap}	90.9±5.7	70.9±4.1	78.7±3.1	78.8±1.1	80.4±2.1	77.7±4.5	54.5±4.0
MAX-MLP	SUM/AVG	84.0±6.1	64.6±10.2	76.0±3.2	77.8±1.3	-	73.2±5.8	51.1±3.6
	SOPOOL _{attn}	90.0±7.3	72.4±4.7	78.3±3.1	78.6±1.9	-	78.1±4.1	54.1±3.4
	SOPOOL _{bimap}	88.8±7.0	73.3±5.5	78.4±3.0	78.0±1.9	-	78.2±4.7	54.6±3.5
MAX-1-LAYER (GraphSAGE)	SUM/AVG	85.1±7.6	63.9±7.7	75.9±3.2	77.7±1.5	-	72.3±5.3	50.9±2.2
	SOPOOL _{attn}	90.0±6.8	72.1±5.9	79.0±2.9	77.4±1.8	-	77.4±5.1	54.1±3.1
	SOPOOL _{bimap}	89.9±5.8	73.6±5.1	78.9±2.8	77.0±2.0	-	78.6±4.7	54.2±3.9

formance to SOPOOL_{bimap} on most datasets. However, huge performance loss is observed on PTC, IMDB-BINARY, and IMDB-MULTI, indicating that subtracting the mean is harmful in graph pooling.

Compared to SOPOOL_{attn}, ATTNPOOL suffers from performance loss on all datasets except COLLAB and REDDIT-BINARY. The loss is especially significant on bioinformatics datasets (PTC, PROTEINS, NCI1). However, ATTNPOOL achieves the best performance on COLLAB and REDDIT-BINARY among all graph pooling methods, although the added SOFTMAX function results in less discriminative power. The reason might be capturing the distributional information is more important than the exact structure in these datasets. It is similar to GINs, where using averaging as graph pooling achieves better performance on social network datasets than summation [12].

4.4 Results with Different Graph Neural Networks

We've already demonstrated the superiority of our proposed SOPOOL_{bimap} and SOPOOL_{attn} over previous pooling methods. Next, we show that their effectiveness is robust to different GNNs. In this experiment, we change GIN-0 into other six variants of GINs. Note that these variants cover Graph Convolutional Networks (GCN) [8] and GraphSAGE [9], thus including a wide range of different kinds of GNNs.

We first give details of different variants of graph isomorphism networks (GINs) [12]. Basically, GINs iteratively update the representation of each node in a graph by aggregating representations of its neighbors, where the iteration is achieved by stacking several layers. Therefore, it suffice to describe the k -th layer of GINs based on one node.

Recall that we represent a graph $G = (A, X)$ by its adjacency matrix $A \in \{0, 1\}^{n \times n}$ and node feature matrix

$X \in \mathbb{R}^{n \times d}$, where n is the number of nodes in G and d is the dimension of node features. The adjacency matrix tells the neighboring information of each node. We introduce GINs by defining node representation matrices $H^{(k-1)} \in \mathbb{R}^{n \times f^{(k-1)}}$ and $H^{(k)} \in \mathbb{R}^{n \times f^{(k)}}$ as inputs and outputs to the k -th layer, respectively. We have $H^{(0)} = X$. Note that the first dimension n does not change during the computation, as GINs learn representations for each node.

Specifically, consider a node ν has corresponding representations $h_\nu^{(k-1)} \in \mathbb{R}^{f^{(k-1)}}$ and $h_\nu^{(k)} \in \mathbb{R}^{f^{(k)}}$, which are rows of $H^{(k-1)}$ and $H^{(k)}$, respectively. The set of neighboring nodes of ν is given by $\mathcal{N}(\nu)$. We describe the k -layer of the following variants:

- **SUM-MLP (GIN-0):**

$$h_\nu^{(k)} = \text{MLP}^{(k)}(h_\nu^{(k-1)} + \sum_{\mu \in \mathcal{N}(\nu)} h_\mu^{(k-1)})$$

- **SUM-MLP (GIN- ϵ):**

$$h_\nu^{(k)} = \text{MLP}^{(k)}((1 + \epsilon^{(k)})h_\nu^{(k-1)} + \sum_{\mu \in \mathcal{N}(\nu)} h_\mu^{(k-1)})$$

- **SUM-1-LAYER:**

$$h_\nu^{(k)} = \text{ReLU}(W^{(k)}(h_\nu^{(k-1)} + \sum_{\mu \in \mathcal{N}(\nu)} h_\mu^{(k-1)}))$$

- **MEAN-MLP:**

$$h_\nu^{(k)} = \text{MLP}^{(k)}(\text{MEAN}\{h_\mu^{(k-1)}, \forall \mu \in \nu \cup \mathcal{N}(\nu)\})$$

- **MEAN-1-LAYER (GCN):**

$$h_\nu^{(k)} = \text{ReLU}(W^{(k)}(\text{MEAN}\{h_\mu^{(k-1)}, \forall \mu \in \nu \cup \mathcal{N}(\nu)\}))$$

- **MAX-MLP:**

$$h_\nu^{(k)} = \text{MLP}^{(k)}(\text{MAX}\{h_\mu^{(k-1)}, \forall \mu \in \nu \cup \mathcal{N}(\nu)\})$$

TABLE 4

Comparison results between different hierarchical graph pooling methods. The hierarchical GNN architecture follows the one in [19]. We report the accuracies of the baselines provided in [19]. The best models are highlighted with boldface.

Models	DD	PROTEINS
DIFFPOOL	67.0±2.4	68.2±2.0
TOPKPOOL	75.0±0.9	71.1±0.9
SAGPOOL	76.5±1.0	71.9±1.0
SOPool _{m-attn}	76.8±1.9	77.1±3.8

TABLE 5

Comparison results of SOPool_{m-attn} with different hierarchical GNNs. The hierarchical GNN architecture follows the one in [19], where we change the number of blocks from one to three. The best models are highlighted with boldface.

Models	DD	PROTEINS
1 block	73.3±2.4	77.4±4.3
2 blocks	77.2±2.7	78.1±4.3
3 blocks	76.8±1.9	77.1±3.8

- **MAX-1-LAYER (GraphSAGE):**

$$h_{\nu}^{(k)} = \text{ReLU}(W^{(k)}(\text{MAX}\{h_{\mu}^{(k-1)}, \forall \mu \in \nu \cup \mathcal{N}(\nu)\}))$$

Here, the multi-layer perceptron (MLP) has two layers with ReLU activation functions. Note that MEAN-1-LAYER and MAX-1-LAYER correspond to GCN [8] and GraphSAGE [9], respectively, up to minor architecture modifications.

The results of these different GNNs with our graph pooling methods are reported in Table 3. Our proposed SOPool_{bimap} and SOPool_{attn} achieve satisfying performance consistently. In particular, on social network datasets, the performance does not decline when the GNNs before graph pooling become less powerful, showing the highly discriminative ability of second-order pooling.

4.5 Ablation Studies in Hierarchical Graph Neural Networks

SOPool_{m-attn} has shown its effectiveness as global graph pooling through the experiments in Sections 4.2 and 4.3. In this section, we evaluate it as hierarchical graph pooling in hierarchical GNNs. The hierarchical GNN architecture follows the one in [19], which contains three blocks of a GNN layer followed by graph pooling, as introduced in Section 4.1. The experiments are performed on DD and PROTEINS datasets, where hierarchical GNNs tend to achieve good performance [17], [19].

First, we compare SOPool_{m-attn} with different hierarchical graph pooling methods under the same hierarchical GNN architecture. Specifically, we include DIFFPOOL, TOPKPOOL, and SAGPOOL, which have been used as hierarchical graph pooling methods in their works. The comparison results are provided in Table 4. Our proposed SOPool_{m-attn} outperforms all the baselines on both datasets, indicating the effectiveness of SOPool_{m-attn} as a hierarchical graph pooling method.

In addition, we conduct experiments to evaluate SOPool_{m-attn} in different hierarchical GNNs by varying

the number of blocks. The results are shown in Table 5. On both datasets, SOPool_{m-attn} achieves the best performance when the number of blocks is two. The results indicate current datasets on graph classification are not large enough yet. And without techniques like jumping knowledge networks (JK-Net) [11], hierarchical GNNs tend to suffer from over-fitting, leading to worse performance than flat GNNs.

5 CONCLUSIONS

In this work, we propose to perform graph representation learning with second-order pooling, by pointing out that second-order pooling can naturally solve the challenges of graph pooling. Second-order pooling is more powerful than existing graph pooling methods, since it is capable of using all node information and collecting second-order statistics that encode feature correlations and topology information. To take advantage of second-order pooling in graph representation learning, we propose two global graph pooling approaches based on second-order pooling; namely, bilinear mapping and attentional second-order pooling. Our proposed methods solve the practical problems incurred by directly using second-order pooling with GNNs. We theoretically show that our proposed methods are more suitable to graph representation learning by comparing with two related pooling methods from computer vision tasks. In addition, we extend one of the proposed method to a hierarchical graph pooling method, which has more flexibility. To demonstrate the effectiveness of our methods, we conduct thorough experiments on graph classification tasks. Our proposed methods have achieved the new state-of-the-art performance on eight out of nine benchmark datasets. Ablation studies are performed to show that our methods outperform existing graph pooling methods significantly and achieve good performance consistently with different GNNs.

ACKNOWLEDGMENTS

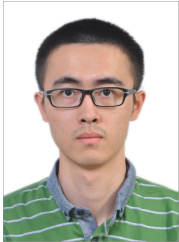
This work was supported in part by National Science Foundation grant IIS-1908166, and Defense Advanced Research Projects Agency grant N66001-17-2-4031.

REFERENCES

- [1] P. Yanardag and S. Vishwanathan, "Deep graph kernels," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 1365–1374.
- [2] S. Zhang, H. Tong, J. Xu, and R. Maciejewski, "Graph convolutional networks: Algorithms, applications and open challenges," in *International Conference on Computational Social Networks*. Springer, 2018, pp. 79–91.
- [3] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *arXiv preprint arXiv:1901.00596*, 2019.
- [4] W. Fan, Y. Ma, Q. Li, Y. He, E. Zhao, J. Tang, and D. Yin, "Graph neural networks for social recommendation," in *The World Wide Web Conference*. ACM, 2019, pp. 417–426.
- [5] H. Gao, J. Pei, and H. Huang, "Conditional random field enhanced graph convolutional neural networks," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '19. New York, NY, USA: ACM, 2019, pp. 276–284. [Online]. Available: <http://doi.acm.org/10.1145/3292500.3330888>
- [6] J. Ma, P. Cui, K. Kuang, X. Wang, and W. Zhu, "Disentangled graph convolutional networks," in *International Conference on Machine Learning*, 2019, pp. 4212–4221.

- [7] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, and P. S. Yu, "Heterogeneous graph attention network," in *The World Wide Web Conference*, ser. WWW '19. New York, NY, USA: ACM, 2019, pp. 2022–2032. [Online]. Available: <http://doi.acm.org/10.1145/3308558.3313562>
- [8] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2017.
- [9] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in Neural Information Processing Systems*, 2017, pp. 1024–1034.
- [10] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018.
- [11] K. Xu, C. Li, Y. Tian, T. Sonobe, K.-i. Kawarabayashi, and S. Jegelka, "Representation learning on graphs with jumping knowledge networks," in *International Conference on Machine Learning*, 2018, pp. 5449–5458.
- [12] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" in *International Conference on Learning Representations*, 2019.
- [13] K. Schütt, P.-J. Kindermans, H. E. S. Felix, S. Chmiela, A. Tkatchenko, and K.-R. Müller, "SchNet: A continuous-filter convolutional neural network for modeling quantum interactions," in *Advances in Neural Information Processing Systems*, 2017, pp. 991–1001.
- [14] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *Advances in Neural Information Processing Systems*, 2015, pp. 2224–2232.
- [15] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Advances in Neural Information Processing Systems*, 2016, pp. 3844–3852.
- [16] M. Zhang, Z. Cui, M. Neumann, and Y. Chen, "An end-to-end deep learning architecture for graph classification," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [17] Z. Ying, J. You, C. Morris, X. Ren, W. Hamilton, and J. Leskovec, "Hierarchical graph representation learning with differentiable pooling," in *Advances in Neural Information Processing Systems*, 2018, pp. 4800–4810.
- [18] H. Gao and S. Ji, "Graph U-Net," in *International Conference on Machine Learning*, 2019.
- [19] J. Lee, I. Lee, and J. Kang, "Self-attention graph pooling," in *International Conference on Machine Learning*, 2019, pp. 3734–3743.
- [20] Y. Ma, S. Wang, C. C. Aggarwal, and J. Tang, "Graph convolutional networks with eigenpooling," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '19. New York, NY, USA: ACM, 2019, pp. 723–731. [Online]. Available: <http://doi.acm.org/10.1145/3292500.3330982>
- [21] Y. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in vision algorithms," in *International Conference on Machine Learning*, vol. 345, 2010.
- [22] D. Chen, Y. Lin, W. Li, P. Li, J. Zhou, and X. Sun, "Measuring and relieving the over-smoothing problem for graph neural networks from the topological view," in *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.
- [23] D. G. Lowe, "Object recognition from local scale-invariant features," in *International Conference on Computational Vision*, vol. 99, no. 2, 1999, pp. 1150–1157.
- [24] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," in *European conference on computer vision*. Springer, 2006, pp. 589–600.
- [25] —, "Pedestrian detection via classification on riemannian manifolds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 10, pp. 1713–1727, 2008.
- [26] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *European Conference on Computer Vision*. Springer, 2010, pp. 143–156.
- [27] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu, "Semantic segmentation with second-order pooling," in *European Conference on Computer Vision*. Springer, 2012, pp. 430–443.
- [28] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1449–1457.
- [29] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 317–326.
- [30] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 457–468.
- [31] Z. Wang and S. Ji, "Learning convolutional text representations for visual question answering," in *Proceedings of the 2018 SIAM International Conference on Data Mining*. SIAM, 2018, pp. 594–602.
- [32] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Geometric means in a novel vector space structure on symmetric positive-definite matrices," *SIAM Journal on Matrix Analysis and Applications*, vol. 29, no. 1, pp. 328–347, 2007.
- [33] D. Acharya, Z. Huang, D. Pani Paudel, and L. Van Gool, "Covariance pooling for facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 367–374.
- [34] Q. Wang, J. Xie, W. Zuo, L. Zhang, and P. Li, "Deep CNNs meet global covariance pooling: Better representation and generalization," *arXiv preprint arXiv:1904.06836*, 2019.
- [35] R. Girdhar and D. Ramanan, "Attentional pooling for action recognition," in *Advances in Neural Information Processing Systems*, 2017, pp. 34–45.
- [36] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1480–1489.
- [37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.
- [38] S. Verma and Z.-L. Zhang, "Graph capsule convolutional neural networks," *arXiv preprint arXiv:1805.08090*, 2018.
- [39] A. K. Debnath, R. L. Lopez de Compadre, G. Debnath, A. J. Shusterman, and C. Hansch, "Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity," *Journal of Medicinal Chemistry*, vol. 34, no. 2, pp. 786–797, 1991.
- [40] H. Toivonen, A. Srinivasan, R. D. King, S. Kramer, and C. Helma, "Statistical evaluation of the predictive toxicology challenge 2000–2001," *Bioinformatics*, vol. 19, no. 10, pp. 1183–1193, 2003.
- [41] K. M. Borgwardt, C. S. Ong, S. Schöner, S. Vishwanathan, A. J. Smola, and H.-P. Kriegel, "Protein function prediction via graph kernels," *Bioinformatics*, vol. 21, no. suppl_1, pp. i47–i56, 2005.
- [42] N. Wale, I. A. Watson, and G. Karypis, "Comparison of descriptor spaces for chemical compound retrieval and classification," *Knowledge and Information Systems*, vol. 14, no. 3, pp. 347–375, 2008.
- [43] A. Shrivastava and P. Li, "A new space for comparing graphs," in *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE Press, 2014, pp. 62–71.
- [44] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," in *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*. ACM, 2005, pp. 177–187.
- [45] P. D. Dobson and A. J. Doig, "Distinguishing enzyme structures from non-enzymes without alignments," *Journal of Molecular Biology*, vol. 330, no. 4, pp. 771–783, 2003.
- [46] M. Niepert, M. Ahmed, and K. Kutzkov, "Learning convolutional neural networks for graphs," in *International Conference on Machine Learning*, 2016, pp. 2014–2023.
- [47] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.
- [48] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.
- [50] N. Shervashidze, S. Vishwanathan, T. Petri, K. Mehlhorn, and K. Borgwardt, "Efficient graphlet kernels for large graph comparison," in *Artificial Intelligence and Statistics*, 2009, pp. 488–495.

- [51] S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. M. Borgwardt, "Graph kernels," *Journal of Machine Learning Research*, vol. 11, no. Apr, pp. 1201–1242, 2010.
- [52] N. Shervashidze, P. Schweitzer, E. J. v. Leeuwen, K. Mehlhorn, and K. M. Borgwardt, "Weisfeiler-Lehman graph kernels," *Journal of Machine Learning Research*, vol. 12, no. Sep, pp. 2539–2561, 2011.
- [53] S. Ivanov and E. Burnaev, "Anonymous walk embeddings," in *International Conference on Machine Learning*, 2018, pp. 2191–2200.
- [54] J. Atwood and D. Towsley, "Diffusion-convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 1993–2001.
- [55] M. Simonovsky and N. Komodakis, "Dynamic edge-conditioned filters in convolutional neural networks on graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3693–3702.



Zhengyang Wang received the MSc degree in mathematics and computer science from New York University, New York City, New York, in 2015. Currently, he is working toward the PhD degree in the Department of Computer Science and Engineering, Texas A&M University, College Station, Texas. His research interests include machine learning, deep learning, and data mining.



Shuiwang Ji received the PhD degree in computer science from Arizona State University, Tempe, Arizona, in 2010. Currently, he is an Associate Professor in the Department of Computer Science and Engineering, Texas A&M University, College Station, Texas. His research interests include machine learning, data mining, and computational biology. He received the National Science Foundation CAREER Award in 2014. He is currently a Section Editor for BMC Bioinformatics, an Action Editor for Data Mining and Knowledge Discovery, and an Associate Editor for ACM Transactions on Knowledge Discovery from Data. He is a senior member of IEEE.