Deep Fusion Clustering Network

Abstract

Deep clustering is a fundamental yet challenging task for data analysis. Recently we witness a strong tendency of combining autoencoder and graph neural networks to exploit structural information for clustering performance enhancement. However, we observe that existing literature 1) lacks a dynamic fusion mechanism to adaptively integrate the information of node attributes and graph structure for consensus and discriminative representation learning; 2) fails to extract information from the both sides for robust pseudo label construction. To tackle the above issues, we propose a **D**eep **F**usion **C**lustering **N**etwork (**DFCN**). Specifically, in our network, an interdependency learning-based Structure and Attribute Information Fusion (SAIF) module is proposed to explicitly merge the representations learned by an autoencoder and a graph autoencoder for discriminative representation learning. Also, a consensus pseudo label extraction method and a triplet self-supervision strategy, which facilitate cross-modality information exploitation, are designed for network training. Extensive experiments on six benchmark datasets have demonstrated that the proposed DFCN consistently outperforms the state-of-the-art deep clustering methods.

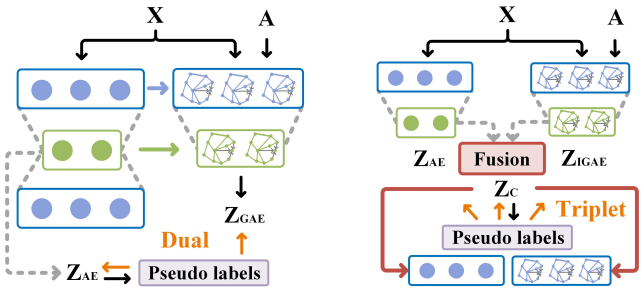

Figure 1: Network structure comparison. Different from the existing attribute and structure information fusion networks (such as SDCN), our proposed algorithm is enhanced with an information fusion module. With this module, 1) both the decoder of AE and GCN reconstruct the input with a learned consensus latent representation. 2) The pseudo label is constructed with sufficient negotiation between AE and GCN. 3) A self-supervised triplet learning mechanism is designed.

Introduction

Deep clustering, which aims to train a neural network for learning discriminative feature representations to divide data into several disjoint groups without intense manual guidance, is becoming an increasingly appealing direction to the machine learning researchers. Thanks to the strong representation learning capability of deep learning methods, researches in this field have achieved promising performance in many applications including anomaly detection (Markovitz et al. 2020), social network analysis (Hu, Chan, and He 2017), and face recognition (Wang et al. 2019b). Two important factors, i.e., the optimization objective and the fashion of feature extraction, significantly determine the performance of a deep clustering method. Specifically, in the unsupervised clustering scenario, without the guidance of labels, designing a subtle objective function and an elegant architecture to enable the network to collect more comprehensive and discriminative information for intrinsic structure revealing is extremely crucial and challenging.

According to the network optimization objective, existing deep clustering methods can be roughly grouped into five categories, i.e., subspace clustering-based methods (Zhou et al. 2019; Ji et al. 2017; Peng et al. 2017), generative adversarial network-based methods (Mukherjee et al. 2019; Ghasedi et al. 2019), spectral clustering-based methods (Yang et al. 2019b; Shaham et al. 2018), Gaussian Mixture Model-based methods (Yang et al. 2019a; Chen et al. 2019), and pseudo label-based self-supervision methods (Xie, Girshick, and Farhadi 2016; Guo et al. 2017). Our method falls into the last category. In the early state, the above deep clustering methods mainly concentrate on exploiting the attribute information in the original feature space of data and have achieved good performance in many circumstances. To further improve the clustering accuracy, recent literature shows a strong tendency in extracting geometrical structure information and then integrates it with attribute information for representation learning. Specifically, Yang et al. design a novel stochastic extension of graph embedding to add local data structures into probabilistic deep Gaussian mixture model (GMM) for clustering (Yang

et al. 2019a). Distribution Preserving Subspace Clustering (DPSC) first estimate the density distribution of the original data space and the latent feature space with kernel density estimation. Then it preserves the intrinsic cluster structure within data by minimizing the distribution inconsistency between the two spaces (Zhou et al. 2019). More recently, graph convolutional networks (GCN), which aggregates the neighborhood information for better sample representation have attracted the attention of many researchers. The work in deep attentional embedded graph clustering (DAEGC) exploits both graph structure and node attributes with a graph attention encoder. It reconstructs the adjacency matrix by a self-optimizing embedding algorithm (Wang et al. 2019a). Following the setting of DAEGC, Adversarially Regularized Graph Autoencoder (ARGA) further develop an adversarial regularizer to guide the learning of latent representations (Pan et al. 2020). After that, structural deep clustering network (SDCN) (Bo et al. 2020) integrates an autoencoder and a graph convolutional network (GCN) into a unified framework by designing a information passing delivery operator and a dual self-supervised learning mechanism.

Although the former efforts have achieve preferable performance enhancement by leveraging both kinds of information, we find that 1) the existing methods lack an cross-modality dynamic information fusion and processing mechanism. Information from two sources are simply aligned or concatenated, leading to insufficient information interaction and merging; 2) the construction of the pseudo label in the literature have seldom considered of using information from both sources, making the guidance of network training less comprehensive and accurate. As a consequence, the negotiate between two information sources is obstructed, resulting in unsatisfying clustering performance.

To tackle the above issues, we propose a Deep Fusion Clustering Network (DFCN). The main idea of our solution is to design an information dynamic fusion module to finely process the attribute information and structural information extracted from autoencoder and GCN for more comprehensive and accurate representation construction. Then, with the reliable representation, we further generate the pseudo label to provide more dependable guidance for network training. Specifically, inspired by the recent developments in self-attention learning, we propose a structure and attribute information fusion (SAIF) module for elaborate information processing. After that, by estimating the similarity between sample points and pre-calculated cluster centers in the latent feature space with Students' t-distribution, we acquire more precise pseudo labels. Finally, we design a triplet self-supervision mechanism which uses the pseudo labels to guide the training of autoencoder, GAE, and information fusion simultaneously. Moreover, we also develop an improved graph autoencoder (IGAE) with symmetric structure and reconstruct the adjacent matrix with both the latent representation and the feature representation reconstructed by the graph decoder. The key contributions of this paper are listed as follow:

- We propose a deep fusion clustering network (DFCN). In this network, a structure and attribute information fusion (SAIF) module is designed for better information interaction between AE and GCN. With this module, 1) since both the decoder of AE and GCN reconstruct the inputs of each other using a consensus latent representation, the generalization capacity of the latent features is boosted. 2) The discriminative capability of the generated pseudo label is enhanced by integrating the complementary information between AE and GCN. 3) The self-supervised triplet learning mechanism integrates the learning of AE, GAE and the fusion module in a unified and robust system, thus further improves the clustering performance.
- We develop a symmetric graph autoencoder, i.e., improved graph autoencoder (IGAE), to further improve the generalization capability of the proposed algorithm.
- Extensive experiment results on six public benchmark datasets have demonstrated that our method is highly competitive and consistently outperforms the state-of-the-art ones with a preferable margin.

Related work

Attributed Graph Clustering

Benefiting from the strong representation power of graph convolutional networks (GCNs) (Kipf and Welling 2017), GCN-based clustering methods that jointly learn graph structure and node attributes have been widely studied in the recent years (Fan et al. 2020; Cheng et al. 2020; Sun, Lin, and Zhu 2020). Specifically, graph autoencoder (GAE) and variational graph autoencoder (VGAE) propose to integrate graph structure into node attributes via iterative weighted linear aggregating neighborhood sample representations (Kipf and Welling 2016). After that, ARGA (Pan et al. 2020), AGAE (Tao et al. 2019), DAEGC (Wang et al. 2019a), and MinCutPool (Bianchi, Grattarola, and Alippi 2020) improves the performance of the early-stage algorithms with adversarial training, attention, and graph pooling mechanisms, respectively. Although the performance of the corresponding algorithms has been improved considerably, the over-smoothing phenomenon of the GCNs still limits the accuracy of these algorithms. More recently, SDCN (Bo et al. 2020) propose to integrate autoencoder and GCN for better representation learning. Through careful theoretical and experimental analysis, authors find that in their proposed network, autoencoder can help provide complementary structure information and help relieve the over-smoothing phenomenon of GCN, while GCN provides high-order structure information to autoencoder. Although SDCN proves that combining autoencoder and GCN can boost the clustering performance of both components, in this work, the GCN acts only as a regularizer of autoencoder, the learned feature of the GCN is insufficiently utilized and the representation learning of the framework lacks the negotiation between the two sub-networks. Differently, in our proposed algorithm, an information fusion module (i.e., SAIF module) is proposed to finely integrate the features learned by the autoencoder and IGAE. As a consequence, the complementary information from the two sub-networks are finely merged, and more discriminative representations are learned.

Pseudo Label Generation

Since reliable guidance is missing in network training, many deep clustering algorithms seek to generate pseudo labels for discriminative representation learning (Ren et al. 2019; Xu et al. 2019; Li et al. 2019). The early method (DEC) in this category first trains an encoder, and then with the pre-trained network, it further creates the pesudo labels with Student's *t*-distribution and fine-tunes the network with stronger guidance (Xie, Girshick, and Farhadi 2016). To increase the accuracy of pseudo labels, IDEC jointly optimizes cluster label assignment and learns features that are suitable for clustering with local structure preservation (Guo et al. 2017). After that, to better train the autoencoder and GAE integrated network, SDCN designs a dual self-supervised learning mechanism which conducts pseudo label refinement and sub-network training in a unified system (Bo et al. 2020). Despite their success, these methods generate pseudo labels with only the information of autoencoder or GCN, none of them considers combining the information from both sides and then come up with a more robust guidance. In contrast, in our method, as the information fusion module allows the information from the two sub-networks to adequately interacts with each other, the resultant pseudo label has the potential to be more discriminative than of the single-source counterparts.

The proposed method

Our proposed method is consist of four parts, i.e., an autoencoder, an improved graph autoencoder, a fusion module, and the optimization targets (please check Fig. 1 for the diagram of our network structure). The encoder part of both AE and IGAE are similar with that of the existing literature. In the following part, we will first introduce the basic notations and then introduce the decoder of both networks, the fusion module, and the optimization targets in detail.

Notations

Given an undirected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where N is the sample number, $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$ and E are the node set the edge set, respectively. The graph is characterized by its attribute matrix $\mathbf{X} \in \mathbb{R}^{N\times d}$ and adjacent matrix $\mathbf{A} = (a_{ij})_{N\times N} \in \mathbb{R}^{N\times N}$. Here, d is the attribute dimension and $a_{ij} = 1$ if $(v_i, v_j) \in \mathcal{E}$, otherwise $a_{ij} = 0$. The corresponding degree matrix is $\mathbf{D} = diag(d_1, d_2, ..., d_N) \in \mathbb{R}^{N\times N}$ and $d_i = \sum_{v_j\in V} a_{ij}$. With $\mathbf{D}$, the adjacent matrix can be further normalized through calculating $\mathbf{D}^{-\frac{1}{2}}\widetilde{\mathbf{A}}\mathbf{D}^{-\frac{1}{2}}$, where $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{I} \in \mathbb{R}^{N\times N}$ indicates that each node in $\mathcal{V}$ is linked with a self-loop structure.

Fusion-based Autoencoders

Input of the Decoders. Most of the existing autoencoders, either classic autoencoder or graph autoencoder, reconstruct the inputs with only its own latent representation. However, in our proposed method, with the compressed representations of AE and GAE, we first integrate the information from both sources for a consensus latent representation. Then, with this feature as input, both the decoder of AE and GAE reconstruct the input of the network. This is very different from the existing methods that our proposed method fuses heterogeneous attribute and structural information with a carefully designed fusion module and then reconstruct the input of both networks with the consensus latent representation. Detailed information about the fusion module will be introduced in a specific section in the following part.

Improved Graph Autoencoder. In the existing literature, the classic autoencoders are usually symmetric, while graph convolutional networks are usually asymmetric (Kipf and Welling 2016; Wang et al. 2019a; Tao et al. 2019). They require only the latent representation to reconstruct the adjacent information and overlook that the attribute information can also be exploited for improving the generalization capability of the corresponding network. To better make use of both the adjacent information and the attribute information, we design a symmetric improved graph autoencoder (IGAE). This network requires to reconstruct both the adjacent matrix and the weighted feature matrix simultaneously. In the proposed network, a layer in the encoder and decoder is formulated as:

$$\mathbf{Z}^{(l)} = \sigma(\mathbf{D}^{-\frac{1}{2}}\widetilde{\mathbf{A}}\mathbf{D}^{-\frac{1}{2}}\mathbf{Z}^{(l-1)}\mathbf{W}^{(l)}), \tag{1}$$

$$\widetilde{\mathbf{Z}}^{(h)} = \sigma(\mathbf{D}^{-\frac{1}{2}}\widetilde{\mathbf{A}}\mathbf{D}^{-\frac{1}{2}}\widetilde{\mathbf{Z}}^{(h)}\widetilde{\mathbf{W}}^{(h)}), \tag{2}$$

where $\mathbf{W}^{(l)}$ and $\widetilde{\mathbf{W}}^{(h)}$ denotes the learnable parameters of the *l*-th encoder layer and *h*-th decoder layer. σ is a nonlinear activation function, such as ReLU or Tanh. To minimize both the reconstruction loss over the adjacent matrix and the weighted attributed matrix, our IGAE is designed to minimize a hybrid loss function:

$$L_{IGAE} = L_w + \gamma L_a, \tag{3}$$

In Eq.(3), γ = 0.1 is a hyper-parameter that balances the weight of the two reconstruction loss functions. Specially, L_w and L_a are defined as follows:

$$L_w = \frac{1}{2N}\|\mathbf{D}^{-\frac{1}{2}}\widetilde{\mathbf{A}}\mathbf{D}^{-\frac{1}{2}}\mathbf{X} - \widetilde{\mathbf{Z}}\|_F^2, \tag{4}$$

$$L_a = \frac{1}{2N}\|\mathbf{D}^{-\frac{1}{2}}\widetilde{\mathbf{A}}\mathbf{D}^{-\frac{1}{2}} - \widehat{\mathbf{A}}\|_F^2. \tag{5}$$

In Eq.(4), $\widetilde{\mathbf{Z}} \in \mathbb{R}^{N\times d}$ is the output of the graph decoder. In Eq.(5), $\widehat{\mathbf{A}} \in \mathbb{R}^{N\times N}$ is the reconstructed adjacent matrix generated by an inner product operation with the latent representation of the network. By minimizing both Eq.(4) and Eq.(5), the proposed IGAE is termed to minimize the reconstruction loss over the weighted feature matrix and adjacent matrix at the same time. Experimental results in the following parts validate the effectiveness of this setting.

Structure and Attribute Information Fusion

To sufficiently explore the node attributes and graph structure information extracted by the AE and IGAE, we propose our structure and attribute information fusion (SAIF) module. This module consists of two parts, i.e., a dynamic fusion mechanism and a triplet self-supervised learning strategy. The overall structure of SAIF is illustrated in Fig. 2

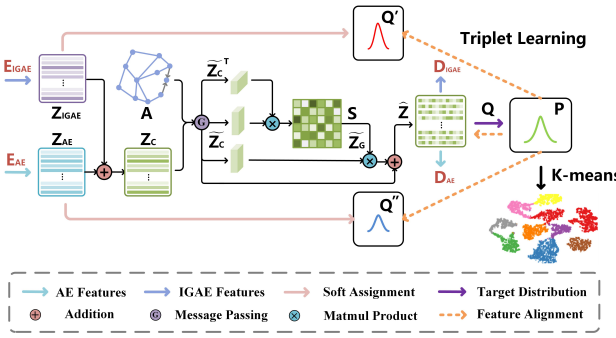

Figure 2: Illustration of the Structure and Attribute Information Fusion (SAIF) module.

Cross-modality dynamic fusion mechanism. The information integration within our fusion module includes four steps. First, we combine the latent representation of the AE ($\mathbf{Z}_{AE}^{(m)} \in \mathbb{R}^{N\times d}$) and IGAE ($\mathbf{Z}_{IGAE}^{(n)} \in \mathbb{R}^{N\times d}$) with a linear combination operation:

$$\mathbf{Z}_C = \alpha \mathbf{Z}_{AE}^{(m)} + (1-\alpha)\mathbf{Z}_{IGAE}^{(n)}, \tag{6}$$

where, m and n are the depth of the encoder of AE and IGAE, respectively. In Eq.(6), α is a learnable coefficient which can adaptively determine the importance of two information sources according to the property of the corresponding dataset. In our paper, it is initialized as 0.5 and then tuned automatically with a gradient decent algorithm.

Then, we process the combined information with a graph convolution-like operation. In the convolution, the adjacent matrix is the self-looped adjacent matrix. With this operation, we enhance $\mathbf{Z}_C$ by considering the local structure within data.

$$\widetilde{\mathbf{Z}_C} = \sigma(\mathbf{D}^{-\frac{1}{2}}\widetilde{\mathbf{A}}\mathbf{D}^{-\frac{1}{2}}\mathbf{Z}_C). \tag{7}$$

In Eq.(7), $\widetilde{\mathbf{Z}}_C$ denotes the local structure enhanced $\mathbf{Z}_C$.

After that, we further introduce a self-correlated learning mechanism to exploit the non-local relationship among samples. Specifically, we first calculate the normalized self-correlation matrix through Eq.(8):

$$\mathbf{S}_{ij} = \frac{e^{(\widetilde{\mathbf{Z}_C}\widetilde{\mathbf{Z}_C}^{\mathbf{T}})_{ij}}}{\sum_{k=1}^{N} e^{(\widetilde{\mathbf{Z}_C}\widetilde{\mathbf{Z}_C}^{\mathbf{T}})_{ik}}}, \tag{8}$$

With $\mathbf{S}$ as coefficients, we recombine $\mathbf{Z}_C$ by considering the global correlation among samples: $\widetilde{\mathbf{Z}_G} = \mathbf{S}\widetilde{\mathbf{Z}_C}$.

Finally, we adopt a skip connection to encourage information to pass smoothly within the fusion module.

$$\widehat{\mathbf{Z}} = \beta\widetilde{\mathbf{Z}_G} + \widetilde{\mathbf{Z}_C}, \tag{9}$$

where β is the weight of the original features. We follow the setting in (Fu et al. 2019) and initialize it as 0 and learn its weight while training the network. Technically, our cross-modality dynamic fusion mechanism can finely fuse the information from both AE and IGAE by considering both the local and global structure information among samples.

Triplet self-supervised strategy. To generate more discriminative guidance for clustering network training, we first adopt the more robust $\widehat{\mathbf{Z}}$ which has integrated the information from both AE and IGAE for pseudo label generation. As shown in Eq.(10) and Eq.(11), the generation process includes two steps.

$$\mathbf{q}_{ij} = \frac{(1+\|\hat{\mathbf{z}}_i - \mathbf{u}_j\|^2/v)^{-\frac{v+1}{2}}}{\sum_{j'}(1+\|\hat{\mathbf{z}}_i - \mathbf{u}_{j'}\|^2/v)^{-\frac{v+1}{2}}}, \tag{10}$$

$$\mathbf{p}_{ij} = \frac{\mathbf{q}_{ij}^2/\sum_i \mathbf{q}_{ij}}{\sum_{j'}\mathbf{q}_{ij'}^2/\sum_i \mathbf{q}_{ij'}}, \tag{11}$$

In the first step, we calculate the similarity of i-th sample ($\hat{\mathbf{z}}_i$) in the fused feature space between the j-th pre-calculated clustering center ($\mathbf{u}_j$) using Student's t-distribution as kernel. In Eq.(10), v is the degree of freedom for Student's t-distribution and $\mathbf{q}_{ij}$ indicates the probability of assigning i-th node to j-th centroid. The matrix $\mathbf{Q}$ reflects the distribution of all samples. In the second step, to increase the confidence of cluster assignment, we introduce Eq.(11) to drive the samples to get closer to the cluster centers. Specifically, $\mathbf{p}_{ij}$ is the elements of target distribution matrix (i.e., pseudo labels) $\mathbf{P}$ in our method, $0 \le \mathbf{p}_{ij} \le 1$. With the generated pseudo labels, we then calculate the soft assignment distribution of the AE and IGAE using Eq.(10) over the latent features of the two networks, respectively. We denote the soft assignment distribution of AE and IGAE as $\mathbf{Q}^{'}$ and $\mathbf{Q}^{''}$.

To train the network in a unified framework and improve the representative capability of each component, we design a triplet loss by adapting the KL-divergence in the following form:

$$L_{KL} = \sum_i \sum_j \mathbf{p}_{ij} log \frac{\mathbf{p}_{ij}}{(\mathbf{q}_{ij} + \mathbf{q}_{ij}^{'} + \mathbf{q}_{ij}^{''})/3}. \tag{12}$$

As in this formulation, the summation of label assignment distribution of AE, IGAE, and the fused representation are aligned with the robust pseudo label simultaneously. Also, since the pseudo label is generated without human guidance, we name the loss function triplet loss and the corresponding training mechanism as triplet self-supervised strategy.

Joint loss and Optimization

The overall learning objective is consist of two main parts, i.e., the reconstruction loss of AE and IGAE, and the triplet loss which is correlated with the pseudo label:

$$L = \underbrace{L_{AE} + L_{IGAE}}_{Reconstruction} + \underbrace{\lambda L_{KL}}_{Clustering}. \tag{13}$$

In Eq.(13), L_{AE} is the common mean square error (MSE) reconstruction loss of AE, λ is a predefined hyper-parameter which balances the importance of reconstruction and clustering.

Table 1: Dataset summary

Dataset	Type	Samples	Classes	Dimension
USPS	Image	9298	10	256
HHAR	Record	10299	6	561
REUT	Text	10000	4	2000
ACM	Graph	3025	3	1870
DBLP	Graph	4058	4	334
CITE	Graph	3327	6	3703

Experiments

Benchmark Datasets

We evaluate the proposed DFCN on six popular public datasets, including three graph datasets (ACM[1] , DBLP[2], and CITE[3]) and three non-graph datasets (USPS (LeCun et al. 1990), HHAR (Lewis et al. 2004), and REUT (Stisen et al. 2015)) . Table 1 summarizes the brief information of these datasets. For the construction details of the datasets please check Appendix A.

Experimental settings

The training of the proposed DFCN includes three steps. First, we pre-train the AE and IGAE independently by minimizing the reconstruction losses. Then, we train the AE and IGAE together without the triplet learning mechanism. With the learned latent representation, we generate the centroids of different clusters. Finally, with the learned centroids and under the guidance of the triplet learning strategy, we train the whole network until convergence. The clustering label is acquired by perform K-means algorithm over the consensus latent embedding $\widehat{\mathbf{Z}}$. For all the compared algorithms, to alleviate the adverse influence of randomness, we repeat each experiment for 10 times and report the average values. For the inputs of the compared algorithms, we adopt the same Principal Components Analysis (PCA) to extract more compact and discriminative features. All the algorithms are trained with the Adam optimizer for at least 200 iterations until we meet a plateau in the validation loss values. The learning rate is set to 10^{-3} for USPS, HHAR and 10^{-4} for REUT, ACM, DBLP, and CITE. More implementation details are presented in Appendix B. We employ four metrics to evaluate the clustering performance of all algorithms: Accuracy (ACC), Normalized Mutual Information (NMI), Average Rand Index (ARI), and macro F1-score (F1).

Comparison with the State-of-the-art Algorithms

In this part, we compare our proposed method with nine state-of-the-art deep clustering algorithms to illustrate its effectiveness. These algorithms include representative algorithms from four categories. Among them, K-means (Hartigan and Wong 1979) is the representative method of classic shallow clustering methods. AE (Hinton and Salakhutdinov 2006), DEC (Xie, Girshick, and Farhadi 2016), and IDEC (Guo et al. 2017) represent the autoencoder-based clustering methods which learn the representation for clustering through training an autoencoder. GAE/VGAE (Kipf and Welling 2016), ARGA (Pan et al. 2020), and DAEGC (Wang et al. 2019a) are typical methods of graph convolutional network-based methods. In these methods, the clustering representation is embedded with structural information with GCN. SDCN and SDCN$_Q$ (Bo et al. 2020) are representatives of hybrid methods which take advantage of both AE and GCN for clustering.

The clustering performance of our method and 10 baseline algorithms on six benchmark datasets are summarized in Table 2. Based on the results, we have the following observations:

1) DFCN shows superior performance against the compared algorithms in most of the circumstances. Specifically, K-means performs clustering on raw data, AE, DEC, and IDEC merely exploits flat-table representations for clustering. These methods seldom take structure information into account, leading to sub-optimal performance. In contrast, DFCN successfully leverages available data by adaptively integrate the information of node attributes and graph structure, which complements each other for consensus and discriminative representation learning and greatly improves clustering performance.

2) It is obvious that GCN-based methods such as GAE, VGAE, ARGA, and DAEGC are not comparable to ours, because these methods under-utilize abundant information from data itself and might be limited to the over-smoothing phenomenon. Differently, DFCN incorporates feature-based representations learned by AE into the whole clustering framework, and mutually explores node attributes and graph structure with a fusion mechanism for discriminative representations. As a result, the proposed DFCN contributes to significant improvements on clustering performance compared with the above methods.

3) DFCN achieves better clustering results than the strongest baseline SDCN$_Q$ and SDCN in the majority of cases, especially on HHAR, DBLP, and CITE datasets. On DBLP dataset for instance, our method represents a relative increase of 7.97%, 4.15%, 7.80%, and 8.03% with respect to ACC, NMI, ARI and F1 against SDCN. This is because DFCN not only achieves a dynamic interaction between node attributes and graph structure to reveal the intrinsic clustering structure, but also adopts a triplet learning mechanism to provide precise network training guidance.

Ablation Studies

Effectiveness of IGAE As shown in Fig. 3, we carry out ablation studies to verify the effectiveness of IGAE. GAE-L_w or GAE-L_a denotes the method optimized by the reconstruction loss of weighted feature matrix or adjacent matrix only. We can find out that GAE-L_w (i.e., red line) consistently performs better than GAE-L_a (i.e., green line) on six datasets. Besides, IGAE (i.e., blue line) clearly improves the clustering performance over the method leveraging sparse information from structure source only. Both observations illustrate that our proposal is able to exploit more attribute information for improving the generalization capability of

[1]http://dl.acm.org/

[2]https://dblp.uni-trier.de

[3]http://citeseerx.ist.psu.edu/index

Table 2: Clustering performance on six datasets. The red and blue values indicate the best and the runner-up results, respectively.

Data	Metric	K-means	AE	DEC	IDEC	GAE	VGAE	ARGA	DAEGC	SDCN$_Q$	SDCN	DFCN
USPS	ACC	66.82	71.04	73.31	76.22	63.10	56.19	66.75	73.55	77.09	78.08	79.51
	NMI	62.63	67.53	70.58	75.56	60.69	51.08	61.60	71.12	77.71	79.51	82.83
	ARI	54.55	58.83	63.70	67.86	50.30	40.96	51.11	63.33	70.18	71.84	75.34
	F1	64.78	69.74	71.82	67.86	61.84	53.63	66.06	72.45	75.88	76.98	78.25
HHAR	ACC	59.98	68.69	69.39	71.05	62.33	71.30	63.33	76.51	83.46	84.26	87.06
	NMI	58.86	71.42	72.91	74.19	55.06	62.95	57.06	69.10	78.82	79.90	82.24
	ARI	46.09	60.36	61.25	62.83	42.63	51.47	44.68	60.38	71.75	72.84	76.35
	F1	58.33	66.36	67.29	68.64	62.64	71.55	61.05	76.89	81.45	82.50	87.33
REUT	ACC	54.04	74.90	73.58	75.43	54.40	60.85	56.16	65.60	79.30	77.15	77.68
	NMI	41.54	49.69	47.50	50.28	25.92	25.51	28.67	30.55	56.89	50.82	59.92
	ARI	27.95	49.55	48.44	51.26	19.61	26.18	24.51	31.12	59.58	55.36	59.77
	F1	41.28	60.96	64.25	63.21	43.53	57.14	51.10	61.82	66.15	65.48	69.62
ACM	ACC	67.31	81.83	84.33	85.12	84.52	84.13	86.08	86.94	86.95	90.45	90.84
	NMI	32.44	49.30	54.54	56.61	55.38	53.20	55.72	56.18	58.90	68.31	69.39
	ARI	30.60	54.64	60.64	62.16	59.46	57.72	62.85	59.35	65.25	73.91	74.93
	F1	67.57	82.01	84.51	85.11	84.65	84.17	86.11	87.07	86.84	90.42	90.78
DBLP	ACC	38.65	51.43	58.16	60.31	61.21	58.59	61.55	62.05	65.74	68.05	76.02
	NMI	11.45	25.40	29.5	31.17	30.80	26.92	26.81	32.49	35.11	39.50	43.65
	ARI	6.97	12.21	23.92	25.37	22.02	17.92	22.66	21.03	34.00	39.15	46.95
	F1	31.92	52.53	59.38	61.33	61.41	58.69	61.79	61.75	65.78	67.71	75.74
CITE	ACC	39.32	57.08	55.89	60.49	61.35	60.97	56.87	64.54	61.67	65.96	69.54
	NMI	16.94	27.64	28.34	27.17	34.63	32.69	34.45	36.41	34.39	38.71	43.93
	ARI	13.43	29.31	28.12	25.70	33.55	33.13	33.38	37.78	35.50	40.17	45.45
	F1	36.08	53.80	52.62	61.62	57.36	57.70	54.82	62.20	37.82	63.62	64.27

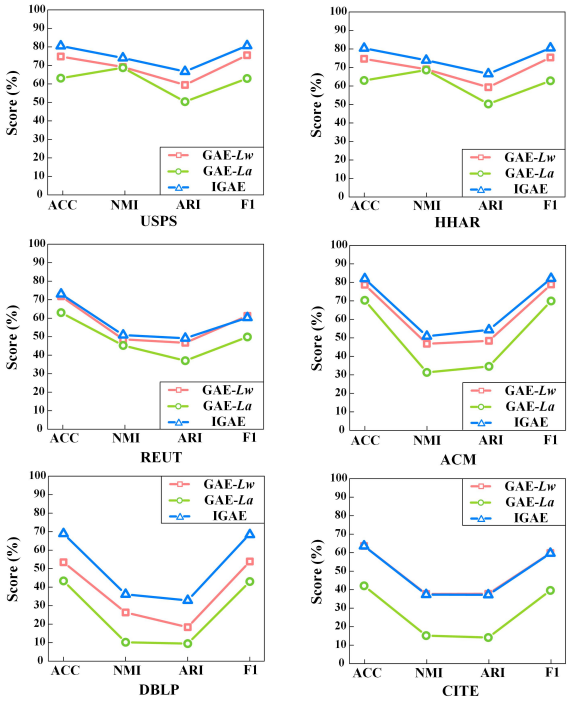

Figure 3: Clustering results of IGAE with different reconstruction strategy.

the deep clustering network. By this means, the latent embedding inherits more properties from the feature space of the original graph, preserving representative meaningful features that generate better clustering decisions.

Analysis of SAIF module In this part, we conduct several experiments to measure the performance of SAIF module. Here the baseline refers to a naive united framework consisting of AE and IGAE. C, S, and T indicate that the method utilizes the cross-modality dynamic fusion mechanism, single or triplet self-supervised strategy, respectively. As summarized in Fig. 4, we observe that 1) Compared with the baseline, Baseline-C method has about 0.5% to 5.0% performance improvements, indicating that exploring both graph structure and node attributes in a dynamic fusion manner is helpful to generate more accurate pseudo labels for better clustering; 2) The performance of Baseline-C-T method is consistently better than Baseline-C-S method on all datasets, which clearly demonstrate the superiority of our self-supervised strategy. According to these observations, SAIF module significantly contributes to improve the clustering performance.

Influence of complementary learning We compare our method with two variants to validate the effectiveness of two-modality complementary learning for deep clustering. As reported in Table 3, w/o-AE or w/o-IGAE refers to the DFCN without AE or IGAE part, respectively. On the one hand, the performance of w/o-AE is better than w/o-IGAE and vice versa in different cases, which proves that both are equally essential for deep clustering. On the other hand, DFCN encoding both DNN-based and GCN-based representations consistently achieves a stable improvement than the case leveraging information from single source. This indicates that DFCN can facilitate the complementary and optimization of two-modality information, which ensures the

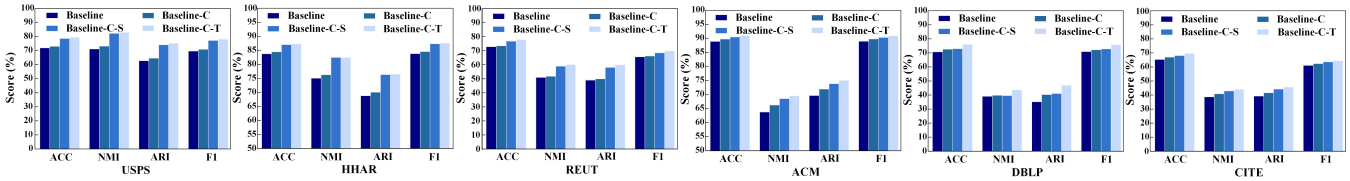

Figure 4: Ablation comparison of cross-modality dynamic fusion mechanism and triplet self-supervised strategy in SAIF.

Table 3: Ablation comparisons on complementary learning.

Dataset	Model	ACC	NMI	ARI	F1
USPS	w/o-AE	78.28	81.25	73.60	76.84
	w/o-IGAE	76.91	77.05	68.76	74.81
	DFCN	**79.51**	**82.83**	**75.34**	**78.25**
HHAR	w/o-AE	75.23	**82.76**	71.73	72.63
	w/o-IGAE	82.80	79.63	72.33	83.38
	DFCN	**87.06**	82.24	**76.35**	**87.33**
REUT	w/o-AE	69.30	48.48	44.56	58.34
	w/o-IGAE	71.39	52.53	48.95	61.46
	DFCN	**77.68**	**59.92**	**59.77**	**69.62**
ACM	w/o-AE	90.20	67.54	73.24	90.23
	w/o-IGAE	89.64	65.59	71.81	89.62
	DFCN	**90.84**	**69.39**	**74.93**	**90.78**
DBLP	w/o-AE	64.21	30.19	29.36	64.60
	w/o-IGAE	67.49	34.23	31.50	67.61
	DFCN	**76.02**	**43.65**	**46.95**	**75.74**
CITE	w/o-AE	69.30	42.86	44.67	**64.40**
	w/o-IGAE	67.90	41.78	43.03	63.74
	DFCN	**69.54**	**43.93**	**45.45**	64.27

pseudo labels to be more accurate for better clustering.

Analysis of Hyper-parameter λ

As can be seen in Eq.(13), DFCN introduces a hyper-parameter λ to make a trade off between the reconstruction and clustering. We conduct experiments to show the effect of this parameter on all datasets. Fig.5 presents four metrics of DFCN by varying λ from 0.01 to 100. From these figures, we observe that 1) hyper-parameters λ is effective in improving the clustering performance; 2) four metrics increase to a higher value and generally maintains it up to slight variation with the increasing value of λ; 3) the performance of DFCN is quite stable when λ is set to 10 across to all datasets.

Visualization of Clustering Results

To intuitively verify the effectiveness of DFCN, we visualize the distribution of the learned latent embedding in two-dimensional space by employing t-SNE algorithm (Maaten and Hinton 2008). As illustrated in Fig. 6, the visual results of DFCN show that there are fewer overlapping areas and samples belong to the same category clearly gather together.

Appendix C contains additional experimental results, including the effect of over-smoothing problem by varying the depth of IGAE, the number of nearest neighbors K-sensitivity analysis in the construction of KNN graph, and the convergence visualization of the proposed method.

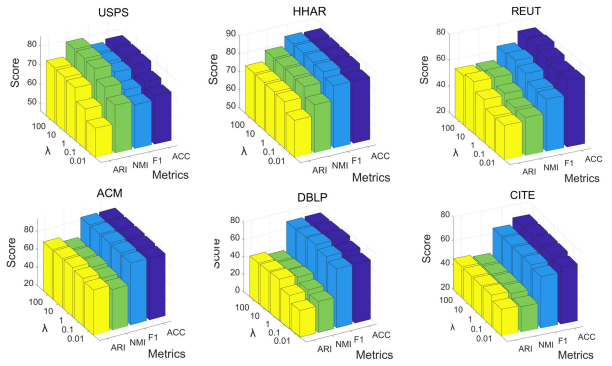

Figure 5: The sensitivity of DFCN with the variation of λ on six datasets.

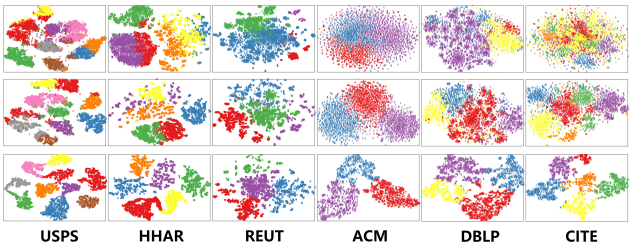

Figure 6: 2D visualization on six datasets. The first, second, and last row correspond to the distribution of raw data, baseline and DFCN, respectively.

Conclusion

In this paper, we propose a novel Deep Fusion Clustering Network (DFCN) for deep clustering. The core component SAIF module leverages both node attributes and graph structure via a cross-modality dynamic fusion mechanism and a triplet self-supervised strategy. In this way, more consensus and discriminative information from both sides is encoded to construct robust pseudo labels, which effectively provide precise network training guidance. Moreover, the proposed IGAE is able to assist in improving the generalization capability of the proposed algorithm. Experiments on six benchmarks show that DFCN consistently outperforms state-of-the-art baseline methods. In the future work, we plan to improve the cross-modality interplay strategy to make our method more robust and effective for some interesting clustering directions e.g., multi-view graph clustering and incomplete multi-view graph clustering.

References

Bianchi, F. M.; Grattarola, D.; and Alippi, C. 2020. Spectral Clustering with Graph Neural Networks for Graph Pooling. In *ICML*, 2729–2738.

Bo, D.; Wang, X.; Shi, C.; Zhu, M.; Lu, E.; and Cui, P. 2020. Structural Deep Clustering Network. In *WWW*, 1400–1410.

Chen, J.; Milot, L.; Cheung, H. M. C.; and Martel, A. L. 2019. Unsupervised Clustering of Quantitative Imaging Phenotypes Using Autoencoder and Gaussian Mixture Model. In *MICCAI*, 575–582.

Cheng, J.; Wang, Q.; Tao, Z.; Xie, D.; and Gao, Q. 2020. Multi-View Attribute Graph Convolution Networks for Clustering. In *IJCAI*, 2973–2979.

Fan, S.; Wang, X.; Shi, C.; Lu, E.; Lin, K.; and Wang, B. 2020. One2Multi Graph Autoencoder for Multi-view Graph Clustering. In *WWW*, 3070–3076.

Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; and Lu, H. 2019. Dual Attention Network for Scene Segmentation. In *CVPR*, 3146–3154.

Ghasedi, K.; Wang, X.; Deng, C.; and Huang, H. 2019. Balanced Self-Paced Learning for Generative Adversarial Clustering Network. In *CVPR*, 4391–4400.

Guo, X.; Gao, L.; Liu, X.; and Yin, J. 2017. Improved Deep Embedded Clustering with Local Structure Preservation. In *IJCAI*, 1753–1759.

Hartigan, J. A.; and Wong, M. A. 1979. A K-Means Clustering Algorithm. *Applied Stats* 28(1): 100–108.

Hinton, G.; and Salakhutdinov, R. R. 2006. Reducing the dimensionality of data with neural networks. *Science* 313: 504–507.

Hu, P.; Chan, K. C. C.; and He, T. 2017. Deep Graph Clustering in Social Network. In *WWW*, 1425–1426.

Ji, P.; Zhang, T.; Li, H.; Salzmann, M.; and Reid, I. D. 2017. Deep Subspace Clustering Networks. In *NIPS*, 24–33.

Kipf, T. N.; and Welling, M. 2016. Variational Graph Auto-Encoders. *ArXiv abs/1611.07308* .

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*, 14.

LeCun, Y.; Matan, O.; Boser, B. E.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W. E.; Jacket, L. D.; and Baird, H. S. 1990. Handwritten zip code recognition with multilayer networks. In *ICPR*, 36–40.

Lewis, D. D.; Yang, Y.; Rose, T. G.; and Li, F. 2004. RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research* 5(2): 361–397.

Li, Z.; Wang, Q.; Tao, Z.; Gao, Q.; and Yang, Z. 2019. Deep Adversarial Multi-view Clustering Network. In *IJCAI*, 2952–2958.

Maaten, L. V. D.; and Hinton, G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9(2605): 2579–2605.

Markovitz, A.; Sharir, G.; Friedman, I.; Zelnik-Manor, L.; and Avidan, S. 2020. Graph Embedded Pose Clustering for Anomaly Detection. In *CVPR*, 10536–10544.

Mukherjee, S.; Asnani, H.; Lin, E.; and Kannan, S. 2019. ClusterGAN: Latent Space Clustering in Generative Adversarial Networks. In *AAAI*, 1965–1972.

Pan, S.; Hu, R.; Fung, S.-F.; Long, G.; Jiang, J.; and Zhang, C. 2020. Learning Graph Embedding with Adversarial Training Methods. *IEEE Transactions on Cybernetics* 50(6): 2475–2487.

Peng, X.; Feng, J.; Lu, J.; Yau, W.-Y.; and Yi, Z. 2017. Cascade Subspace Clusterings. In *AAAI*, 2478–2484.

Ren, Y.; Hu, K.; Dai, X.; Pan, L.; Hoi, S. C. H.; and Xu, Z. 2019. Semi-supervised deep embedded clustering. *Neurocomputing* 325(1): 121–130.

Shaham, U.; Stanton, K. P.; Li, H.; Basri, R.; Nadler, B.; and Kluger, Y. 2018. SpectralNet: Spectral Clustering using Deep Neural Networks. In *ICLR*.

Stisen, A.; Blunck, H.; Bhattacharya, S.; Prentow, T. S.; Kjærgaard, M. B.; Dey, A.; Sonne, T.; and Jensen, M. M. 2015. Smart devices are different: Assessing and mitigating mobile sensing heterogeneities for activity recognition. In *SENSYS*, 127–140.

Sun, K.; Lin, Z.; and Zhu, Z. 2020. Multi-Stage Self-Supervised Learning for Graph Convolutional Networks on Graphs with Few Labeled Nodes. In *AAAI*, 5892–5899.

Tao, Z.; Liu, H.; Li, J.; Wang, Z.; and Fu, Y. 2019. Adversarial Graph Embedding for Ensemble Clustering. In *IJCAI*, 3562–3568.

Wang, C.; Pan, S.; Hu, R.; Long, G.; Jiang, J.; and Zhang, C. 2019a. Attributed Graph Clustering: A Deep Attentional Embedding Approach. In *IJCAI*, 3670–3676.

Wang, Z.; Zheng, L.; Li, Y.; and Wang, S. 2019b. Linkage Based Face Clustering via Graph Convolution Network. In *CVPR*, 1117–1125.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised Deep Embedding for Clustering Analysis. In *ICML*, 478–487.

Xu, C.; Guan, Z.; Zhao, W.; Wu, H.; Niu, Y.; and Ling, B. 2019. Adversarial Incomplete Multi-view Clustering. In *IJCAI*, 3933–3939.

Yang, L.; Cheung, N.-M.; Li, J.; and Fang, J. 2019a. Deep Clustering by Gaussian Mixture Variational Autoencoders with Graph Embedding. In *ICCV*, 6440–6449.

Yang, X.; Deng, C.; Zheng, F.; Yan, J.; and Liu, W. 2019b. Deep Spectral Clustering Using Dual Autoencoder Network. In *CVPR*, 4066–4075.

Zhou, L.; Bai, X.; Wang, D.; Liu, X.; Zhou, J.; and Hancock, E. 2019. Latent Distribution Preserving Deep Subspace Clustering. In *IJCAI*, 4440–4446.