Cross-Domain Few-Shot Graph Classification with a Reinforced Task Coordinator

Qiannan Zhang[1]*, Shichao Pei[2]*, Qiang Yang[1], Chuxu Zhang[3]†, Nitesh Chawla[2], Xiangliang Zhang[2]†‡

[1]King Abdullah University of Science and Technology, Saudi Arabia
[2]University of Notre Dame, USA
[3]Brandeis University, USA
{qiannan.zhang, qiang.yang}@kaust.edu.sa, chuxuzhang@brandeis.edu,
{spei2, nchawla, xzhang33}@nd.edu

Abstract

Cross-domain graph few-shot learning attempts to address the prevalent data scarcity issue in graph mining problems. However, the utilization of cross-domain data induces another intractable domain shift issue which severely degrades the generalization ability of cross-domain graph few-shot learning models. The combat with the domain shift issue is hindered due to the coarse utilization of source domains and the ignorance of accessible prompts. To address these challenges, in this paper, we design a novel Cross-domain Task Coordinator to leverage a small set of labeled target domain data as prompt tasks, then model the association and discover the relevance between meta-tasks from the source domain and the prompt tasks. Based on the discovered relevance, our model achieves adaptive task selection and enables the optimization of a graph learner using the selected fine-grained meta-tasks. Extensive experiments conducted on molecular property prediction benchmarks validate the effectiveness of our proposed method by comparing it with state-of-the-art baselines.

Introduction

Canonical graph learning models have achieved remarkable progress in modeling and inference on graph-structured data. These models commonly require abundant annotated (labeled) data and inevitably suffer from the label scarcity issue in practice due to the expense and hardship of data annotation. This issue leads to the difficulty of inference on novel classes with scarce annotations and harms the generalization of graph learning models. To alleviate the impact of label scarcity, graph few-shot learning, mainly empowered by meta-learning (Hospedales et al. 2020), formulates graph learning into a few-shot framework, which allows for the fast adaptation to novel tasks with scarce labeled data after training with a sufficient number of related tasks, as applied in various node classification (Huang and Zitnik 2020; Wang et al. 2020; Yao et al. 2020) and graph classification (Guo et al. 2021; Chauhan, Nathani, and Kaul 2020) tasks.

Current graph few-shot learning models assume that the meta-training and meta-testing tasks are sampled from the same task distribution. Yet, the assumption cannot be well held in real-world scenarios. Collecting adequate training data from a single domain is non-trivial in some cases, while the data from distinct but related data-rich domains might contain implicitly correlated knowledge, which is transferable to facilitate the few-shot learning on the data-scarce domain. Learning from cross-domain data thus is a feasible solution to mitigate the label scarcity issue and a requisite for many real-world applications. However, meta-learning could be outperformed by pre-training and fine-tuning methods when there is a large shift between source and target domains (Chen et al. 2019). Hence, addressing domain shift issue and improving cross-domain generalizability for graph few-shot learning are of great interest.

The studies of cross-domain few-shot learning mainly fall in the computer vision field, and tackle the domain shift issue by learning generalizable statistics for feature augmentation (Tseng et al. 2020), feature masking (Das, Yun, and Porikli 2022), and batch normalization (Du et al. 2020). A recent attempt (Hassani 2022) on graphs leverages a bunch of source domains to pre-train a designed graph encoder for the target domain with scarce labeled data. However, existing works are far from perfect due to two major challenges.

The coarse utilization of source domains. Recent solutions (Jiang et al. 2019; Zhou et al. 2021; Shen et al. 2020a) to cross-domain knowledge transfer leverage whole information from one or multiple source domains to gain generalizability. Yet, source domains inevitably include target-relevant and target-irrelevant information that could cause the domain shift issue. Roughly fusing all source domains without distinction into the learning model would introduce noisy and redundant target-irrelevant knowledge, which might degrade the generalization on the target domain. A recent effort (Guo, Pasunuru, and Bansal 2020) on text classification selects closer domains under certain criteria from all source domains. However, the strategy is limited to the coarse choice at the level of domains and unavoidably faces

*Equal contribution.
†Corresponding authors.
‡Dr. Xiangliang Zhang is secondly affiliated with the King Abdullah University of Science and Technology, Saudi Arabia.

a similar problem. Directly adopting the inadequate strategies for cross-domain graph few-shot learning is not an ideal choice, and it is more desirable to adaptively select beneficial and fine-grained knowledge from source domains.

The ignorance of accessible prompts. A line of research proposes to ease the domain shift issue by involving a large quantity of unlabeled target domain data during model training (Yang et al. 2021; Huang, Xu, and Wang 2020). However, it is not effortless in many cases to collect the required unlabeled data, especially for fine-grained categories. Oppositely, a small amount of labeled target domain data is easy to obtain, yet has not been well leveraged by existing works in model training. Specifically, the small set of labeled target domain data can be regarded as the prompt to evaluate and steer the selection of beneficial knowledge from abundant source domain data, such that the learning model can gain better generalization on the target domain.

Motivated by the aforementioned challenges, we aim to promote graph few-shot learning by selectively leveraging beneficial and fine-grained knowledge from the source domain, with the assistance of limited labeled data from the target domain. In this paper, we propose a novel **C**ross-**D**omain **T**ask **C**oordinator (a.k.a. **CDTC**) to achieve the adaptive knowledge selection and thus mitigate the domain shift issue. Specifically, the goal of CDTC is to select suitable *meta-tasks* from the source domain for training a graph base learner with optimization-based meta-learning. In detail, in order to leverage the accessible labeled target domain data as the prompt for selecting beneficial knowledge, CDTC organizes the target domain data in the form of few-shot tasks as prompt tasks, then models the association and discovers the relevance between the meta-tasks from the source domain and the prompt tasks by a task bipartite graph, and refines the task representations using Graph Neural Networks. With the refined task representations, CDTC calculates the task sampling probabilities based on which it achieves adaptive task selection for choosing appropriate meta-tasks to optimize the graph base learner in a meta-learning way. Note that the prompt data only guides knowledge selection rather than being directly involved in training the graph base learner.

Our contributions in this work are summarized as follows: (1) We study the cross-domain few-shot graph classification problem, and propose a novel model called CDTC to tackle the domain shift issue; (2) We explore an innovative and effective way to utilize a small set of labeled target domain data as the prompt to achieve adaptive knowledge selection; (3) The proposed CDTC can be well integrated with the optimization-based meta-learning process and trained with reinforcement learning in an end-to-end manner; (4) Experimental results on molecular property prediction benchmarks demonstrate the effectiveness of CDTC in comparison to the state-of-the-art approaches for multiple cross-domain few-shot graph classification tasks.

Related Work

Graph Few-shot Learning Meta-learning has been the dominant paradigm for few-shot learning and developed into two main branches: 1) optimization-based methods, which solve the few-shot learning problem as an optimization problem (Finn, Abbeel, and Levine 2017; Li et al. 2017) and 2) metric-based methods, which learn a generalized metric space for distance-based predictions (Snell, Swersky, and Zemel 2017). Recently, graph few-shot learning techniques have surged by combining few-shot learning with graph neural networks (GNNs) (Wu et al. 2020), such as meta-learning a GNN (Kipf and Welling 2017; Wang et al. 2020) with MAML (Finn, Abbeel, and Levine 2017), learning transferable graph signatures (Bose et al. 2019), collecting meta-gradients from local subgraphs (Huang and Zitnik 2020), calculating weighted class prototypes (Liu et al. 2019; Ding et al. 2020), learning with heterogeneity (Zhang et al. 2022a,b), etc. Particularly, molecular graphs are widely utilized to evaluate graph few-shot learning approaches, especially for graph classification tasks. Meta-MGNN (Guo et al. 2021) utilizes a pre-trained GNN (Hu et al. 2020) to encode molecular information and incorporates auxiliary self-supervised tasks to facilitate few-shot molecular property prediction tasks. PAR (Wang et al. 2021) leverages contextual information within mini-batches to augment graph embeddings for improving few-shot learning. Apart from the above works solving few-shot problems in one single domain, the study in (Abbasi et al. 2019) aims to leverage knowledge from other domains. However, it requires access to large amounts of unlabeled data in the target domain at the training phase. A recent attempt (Hassani 2022) proposes to pre-train a multi-view graph encoder with molecular data in multiple domains, while they focus on the design of the encoder rather than the crucial domain shift issue. How to address the domain shift issue to enable cross-domain graph few-shot learning remains an open problem.

Domain Adaptation To address the domain shift issue caused by learning from distinct but relevant domains, a family of domain adaptation techniques has been proposed, by generally assuming the availability of a large amount of unlabeled data from the target domain during training (Wang and Deng 2018). These approaches are divided into the multi-source domain adaptation (Zhao et al. 2020; Venkat et al. 2020) adopting multiple source domains for knowledge transfer and the prevalent single-source domain adaptation leveraging only one source domain. Recent studies include exploring domain-invariant feature spaces (Zhao et al. 2019), domain mapping (Kirchmeyer et al. 2022), classifier ensemble (Venkat et al. 2020), etc. Existing works generally leverage the whole available data in the source domain, thus probably introducing noisy and redundant information into the learning model. In light of this, a study in (Guo, Pasunuru, and Bansal 2020) proposes to select beneficial domains from all the source domains for text classification, yet they focus on the coarse choice at the level of domains. Therefore, it is unsuitable to directly adopt inadequate strategies to select the beneficial and fine-grained knowledge within one or more source domains.

Cross-domain Graph Learning Recently, increasing attention has been drawn to leveraging cross-domain information to promote graph learning in the target domain. Research works combine domain adaptation with

graph-structured data by domain adversarial learning (Shen et al. 2020a), domain-invariant feature learning (Shen et al. 2020b), etc. For instance, DANE (Zhang et al. 2019) associates GNNs with adversarial learning to align graph distributions in different domains. MuSDAC (Yang et al. 2021) utilizes a multi-channel GCN for channel-wise alignment between the source and target domain. However, these approaches may not target few-shot learning problems, which require fast adaptation ability. Besides, they rely on common nodes in distinct domains or assume the same label space across domains. Also, they may require to access unlabeled target domain data during the model training phase, while such data can be strenuous to obtain in practice.

Preliminaries

Problem Formulation Let $G = (V, E)$ denote a graph where V represents a set of nodes and E denotes a set of edges. Generally, we define a domain $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}, \mathcal{P}_{\mathcal{X},\mathcal{Y}}\}$ as a joint distribution $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$ over the feature space $\mathcal{X}$ and the label space $\mathcal{Y}$. And we focus, in this paper, on the cross-domain few-shot graph classification problem where a source domain $\mathcal{D}^S$ and a target domain $\mathcal{D}^T$ exist. The source domain $\mathcal{D}^S$ consists of a number of graphs $\mathcal{G}^S = \{G_1^S, \ldots, G_d^S\}$ and a set of distinct labels $\mathcal{Y}^S = \{Y_1^S, \ldots, Y_c^S\}$. Similarly, the target domain $\mathcal{D}^T$ contains a few graphs $\mathcal{G}^T = \{G_1^T, \ldots, G_m^T\}$ and a set of different labels $\mathcal{Y}^T = \{Y_1^T, \ldots, Y_z^T\}$. d and m denote the number of graphs, while c and z show the number of labels in $\mathcal{D}^S$ and $\mathcal{D}^T$ respectively. It is assumed that $\mathcal{D}^S$ and $\mathcal{D}^T$ have different marginal distributions $\mathcal{P}_{\mathcal{G}^S}$ and $\mathcal{P}_{\mathcal{G}^T}$, and the label space $\mathcal{Y}^S$ and $\mathcal{Y}^T$ are disjoint. Cross-domain few-shot graph classification aims to learn a model with good generalization ability that can predict the label of graphs in the target domain $\mathcal{D}^T$ given few-shot annotated examples from the target domain and the source domain data $\mathcal{D}^S$.

Episodic Training To mimic the few-shot circumstances at the testing phase, an episodic training paradigm has been proposed by composing small support and query sets in one meta-task (Finn, Abbeel, and Levine 2017; Li et al. 2019). Under N-way K-shot setting, the optimization-based meta-learning that our work employs, seeks a generalizable initialization for the parameters of a base model by learning with a series of related N-way K-shot meta-tasks at the training phase. Specifically, each meta-task τ can be generated by randomly choosing N different classes $C_\tau, C_\tau \subset \mathcal{Y}, |C_\tau| = N$, where $\mathcal{Y}$ is the label space of a domain. Upon that, K and J labeled examples are sampled to form the support set Ω^s and query set Ω^q of meta-task τ. Episodic training is achieved by minimizing the loss on the query set Ω^q (meta-update in the outer loop), based on the parameters obtained by minimizing the loss on its support set Ω^s (inner-update in the inner loop).

Graph Base Learner To encode given graphs, we employ GNNs as the graph base learner for episodic training. Here, we adopt the optimization-based meta-learning strategy to meta-train the graph base learner with a batch of meta-tasks at each iteration. Considering our focus on adaptive task selection to combat domain shift, we leverage the simple yet effective Graph Isomorphism Network (GIN) (Xu et al. 2019) as the graph base learner. The (l)-th layer of a GIN updates the node embedding $\mathbf{h}_v$ of node v as: $\mathbf{h}_v^{(l)} = \mathrm{MLP}_{\mathbf{W}^l}((1+\epsilon^l)\cdot \mathbf{h}_v^{(l-1)} + \sum_{u\in N_v} \mathbf{h}_u^{(l-1)})$, where $\mathrm{MLP}_{\mathbf{W}^l}$ denotes the multilayer perception at layer l, ϵ^l is learnable, and N_v represents the neighbors of node v. With $\mathbf{h}^{(0)}$ initialized as the input node features, the graph-level representation $\mathbf{h}_G$ is obtained as follows:

$$\mathbf{h}_G = \mathrm{READOUT}(\{\mathbf{h}_v^L | v \in V\}), \tag{1}$$

where $\mathrm{READOUT}(\cdot)$ function produces the embedding of a graph by aggregating embeddings of all nodes at the last L-th layer (Mesquita, Souza, and Kaski 2020). We simply adopt mean pooling to derive the graph-level representations. Besides, we add a classification head $f(\cdot)$ to predict graph label $y = f(\mathbf{h}_G)$ for graph classification.

Overall Framework

The overall framework of the proposed model is illustrated in Figure 1. To conduct graph classification, a graph base learner is first employed to extract information from graph examples. Then a cross-domain task coordinator (CDTC) is proposed to adaptively coordinate and select the beneficial meta-tasks at each iteration to circumvent the noisy and redundant meta-tasks and mitigate the domain shift issue. The graph base encoder is thus meta-trained with the selected meta-tasks to gain cross-domain generalization on the target domain. Lastly, we reinforce CDTC to improve its selection based on feedback from the target domain. Next, we elaborate on the model details and the optimization procedure.

Cross-domain Task Coordinator

To combat domain shift, we endeavor to adaptively select meta-tasks beneficial for generalization to the target domain in episodic training by the well-designed cross-domain task coordinator (CDTC). Specifically, CDTC consists of two components. One is the *task association modeling* for discovering the relevance between meta-tasks sampled from the source domain and prompt data from the target domain. Another is *task selection* for choosing appropriate meta-tasks to optimize the graph base learner in a meta-learning paradigm. We next elaborate on the components in detail.

Task Association Modeling

With limited labeled data from the target domain as the prompt and a group of meta-tasks $\{\tau_i\}_{i=1}^{M_s} \subseteq \tau_S$ from the source domain $\mathcal{D}^S$ as the candidates at each iteration, we can explore and model their associations to learn accurate task representations for later task selection. Note that the prompt data aims to only assist task selection rather than being involved in the training of the graph base learner.

Task Representation Learning We base the selection of beneficial and fine-grained knowledge on the level of tasks, i.e., to select B beneficial meta-tasks from M_s candidate

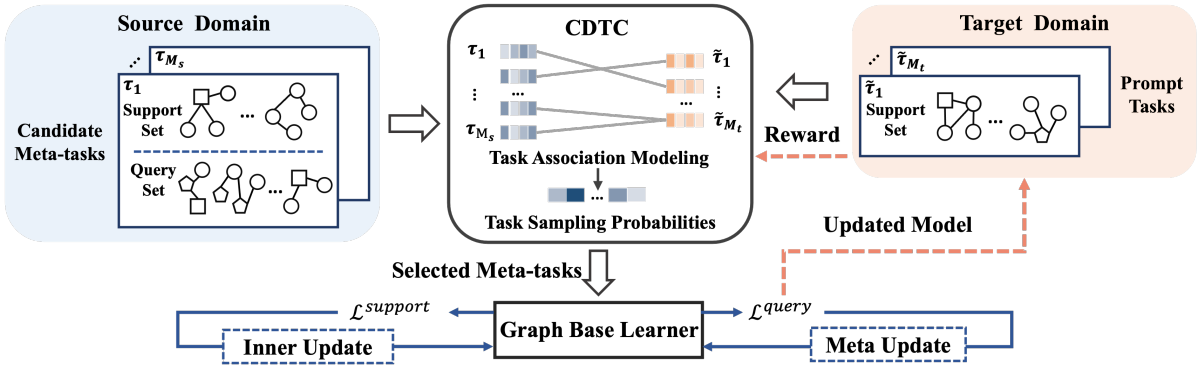


Figure 1: Illustration of cross-domain few-shot graph classification with the proposed CDTC. (1) CDTC models the relationships between candidate meta-tasks from the source domain and prompt tasks from the target domain to generate task sampling probabilities for selecting beneficial meta-tasks (hollow arrow); (2) Graph base learner is meta-updated with the selected meta-tasks (blue arrow); (3) CDTC receives the reward to update itself with reinforcement learning, where the reward is the performance improvement evaluated on prompt tasks before and after the update of the graph base learner (orange arrow).

meta-tasks at each iteration. As each meta-task can be characterized by the task examples, we derive the task representation for each meta-task with its included graph examples. Specifically, considering a N-way K-shot meta-task τ_i sampled from the source domain $\mathcal{D}^S$, it can be depicted by N class prototypes $\{\mathbf{p}_n\}_{n=1}^N$, each of which serves as the center of the graph examples belonging to class n. To calculate the class prototype, we average the graph embeddings in the support set learned by the graph base learner. Task representation $\mathbf{p}_{\tau_i}$ of meta-task τ_i can be obtained as follows:

$$\mathbf{p}_{\tau_i} = \mathrm{MLP}_{\mathbf{W}_a}(\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N), \tag{2}$$

where $\{\mathbf{p}_n\}_{n=1}^N$ can be combined via concatenation or in an order-invariant manner, e.g., mean-pooling. And $\mathbf{W}_a$ is the trainable parameters. Without loss of generality, we conduct binary classification in experiments and calculate task representations by concatenating the prototypes: $\mathbf{p}_{\tau_i} = \mathrm{MLP}_{\mathbf{W}_a}(\mathbf{p}_{pos} \| \mathbf{p}_{neg})$, where $\|$ denotes the concatenation operation, and $\mathbf{p}_{pos}$ (resp. $\mathbf{p}_{neg}$) is the positive (resp. negative) class prototype. When the class number (N) grows, the order-invariant manner for combining class prototypes, e.g., mean-pooling, has reduced computational cost.

Furthermore, we resort to a limited set of labeled target domain data to select meta-tasks in the source domain. Hence, we organize them into few-shot tasks, which are expected to work as the prompt to model the relationship between the source and target domain. Specifically, a N-way K-shot prompt task is composed by randomly sampling K graph examples from N novel classes in the target domain. We denote the number of prompt tasks $\{\tilde{\tau}_j\}_{j=1}^{M_t}$ built from the target domain as M_t, and note that each prompt task only consists of a support set and remains unchangeable during meta-training. Likewise, the task representation of a prompt task $\tilde{\tau}_j$, denoted as $\mathbf{p}_{\tilde{\tau}_j}$, is obtained by leveraging its class prototypes as described in Eq. (2).

Task Bipartite Graph To better explore the association between the candidate meta-tasks and the prompt tasks, we model them in a task bipartite graph in which each vertice denotes a task. Specifically, the task bipartite graph $\mathcal{G}_{\mathcal{B}} = (\mathcal{U}, \mathcal{V}, \mathcal{E})$ consists of two disjoint parts. Here, $\mathcal{U}$ denotes the set of vertices representing the candidate meta-tasks and $\mathcal{V}$ contains vertices referring to the prompt tasks. The vertices in both sets are featured with the obtained task representations. We assume that one possible edge $e_{ij} \in \mathcal{E}$ can exist only between a candidate meta-task $\tau_i \in \mathcal{U}$ and a prompt task $\tilde{\tau}_j \in \mathcal{V}$. With this constraint, candidate meta-tasks can directly interact with the prompt tasks. Yet, two candidate meta-tasks may correspond with each other only through an intermediate prompt task, such that the task bipartite graph could explore the knowledge related to the prompt tasks from the target domain.

Therefore, we build the task bipartite graph by considering the relevance between candidate meta-tasks and the prompt tasks under the aforementioned constraint, and correspondingly initialize the adjacency matrix $\mathbf{A} \in \mathbb{R}^{(M_s+M_t)\times(M_s+M_t)}$ of the task bipartite graph $\mathcal{G}_{\mathcal{B}}$. Its element $A(\tau_i, \tilde{\tau}_j)$, w.r.t. task τ_i and $\tilde{\tau}_j$, is calculated as follows:

$$A(\tau_i, \tilde{\tau}_j) = \begin{cases} 1 & \text{if } ||\mathbf{p}_{\tau_i} - \mathbf{p}_{\tilde{\tau}_j}||_2 \geq \delta; \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $||\mathbf{p}_{\tau_i} - \mathbf{p}_{\tilde{\tau}_j}||_2$ measures the similarity between a candidate meta-task τ_i and a prompt task $\tilde{\tau}_j$ computed by Euclidean distance. Here δ is a pre-defined threshold for connecting task pairs with high relevance.

Task Representation Refinement Since the adoption of separately calculated task representations and predefined similarity measurement (i.e., Euclidean distance) in the initialization of $\mathcal{G}_{\mathcal{B}}$ may not accurately depict the relevance between tasks. We aim to refine task representations by information propagation over the task bipartite graph. Though several research works (Ma et al. 2020; Wang et al. 2021) in few-shot learning incorporate the relationships among examples in meta-tasks for improving model performance, we conduct task-level interaction to better explore their associations for task selection rather than the example-level relationship. It is expected that similar or related tasks would be close in the embedding space, and the task embedding space could well reflect the relevance between the candidate meta-tasks and prompt tasks.

With this objective, we employ a two-layer GCN on the task bipartite graph. Specifically, let $\mathbf{P} \in \mathbb{R}^{(M_s+M_t)\times d}$ denote the feature matrix for vertexes in $\mathcal{G}_{\mathcal{B}}$, where d is feature dimension. Task representations at the (l)-th layer are updated as follows:

$$\mathbf{P}^{(l)} = \sigma(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{P}^{(l-1)}\mathbf{W}_g^{(l)}), \tag{4}$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathrm{I}$, $\tilde{\mathbf{D}} = \sum_i \tilde{A}_{i,j}$. $\mathbf{W}_g^{(l)}$ denotes the trainable weight at the l-th layer, and σ is the activation function. Here, $\mathbf{P}^{(0)}$ is initialized with $\{\mathbf{p}_{\tau_i}\}_{i=1}^{M_s}$ and $\{\mathbf{p}_{\tilde{\tau}_j}\}_{j=1}^{M_t}$. With a slight abuse of notation, we denote the refined task representations as $\{\mathbf{p}_{\tau_i}\}_{i=1}^{M_s}$ and $\{\mathbf{p}_{\tilde{\tau}_j}\}_{j=1}^{M_t}$ respectively. The refined task representations are expected to preserve more accurate information regarding the target domain.

Task Selection

Next, we measure task relevance and conduct task selection with the refined task representations. Different from the existing domain adaption work that mainly leverages predefined measurement of closeness (Kirchmeyer et al. 2022), we propose to employ the prompt data in the target domain via reinforcement learning for adaptive task selection. Specifically, to reflect the characteristics of the target domain, we average task representations of prompt tasks as:

$$\mathbf{p}_T = \frac{1}{M_t}\sum_{j=1}^{M_t}\mathbf{p}_{\tilde{\tau}_j}, \tag{5}$$

where $\mathbf{p}_T$ denotes the representation for the target domain. Accordingly, the relevance score r_i between each candidate meta-task τ_i and target domain representation $\mathbf{p}_T$ is calculated as follows:

$$r_i = \mathrm{MLP}_{\mathbf{W}_c}(\mathbf{p}_{\tau_i} \odot \mathbf{p}_T), \tag{6}$$

where $\odot$ denotes the element-wise multiplication, $\mathrm{MLP}_{\mathbf{W}_c}$ is a multilayer perceptron parameterized with $\mathbf{W}_c$ and activated by the Sigmoid function.

Besides the relevance score, we assign an additional score, $o_i = \rho(w_i)$ where w_i is the trainable parameter and ρ denotes the Sigmoid activation, to each candidate meta-task τ_i as its task signature. The goal is to directly depict the task's importance for generalization to the target domain. Therefore, the overall score t_i for candidate meta-task τ_i is derived by combining the relevance score and task signature as $t_i = r_i + o_i$. The probability $\tilde{t}_i$ of selecting meta-task τ_i is thus computed as follows:

$$\tilde{t}_i = t_i / \sum_{m=1}^{M_s} t_m. \tag{7}$$

With the obtained sampling probabilities, we could select B meta-tasks from the candidate meta-tasks and conduct meta-optimization of graph base learner at iteration k as follows:

$$\theta^{(k+1)} = \theta^{(k)} - \alpha\nabla_{\theta^{(k)}}\frac{1}{B}\sum_{i=1}^{B}\mathcal{L}(\Omega_i^q; \theta_i^{(k)}). \tag{8}$$

Here, $\theta_i^{(k)}$ is the task-specific parameters obtained with the support set Ω_i^s of meta-task τ_i, computed as $\theta_i^{(k)} = \theta^{(k)} - \beta\nabla_{\theta^{(k)}}\mathcal{L}(\Omega_i^s; \theta^{(k)})$ where β is the inner-update rate. Besides, α denotes the meta-update rate, and $\mathcal{L}$ is the cross-entropy loss for graph classification.

Optimization and Training Procedure

To motivate CDTC towards the selection of suitable meta-tasks, we provide CDTC with feedback signals that reflect the appropriateness of selected meta-tasks. Specifically, we leverage graph classification performance on the target domain for evaluation. Let $\{\tau_i\}_{i=1}^{B}$ denote B meta-tasks selected by the current CDTC. The graph base learner (parameterized with θ_{OLD}) receives the selected meta-tasks and conducts a temporary update following Eq. (8) to obtain the parameters θ_{NEW}. Then, the reward R is defined as follows:

$$R = \tanh(\frac{1}{M_t}\sum_{j=1}^{M_t}\mathcal{L}_c(\tilde{\tau}_j, \theta_{\mathrm{NEW}}) - \mathcal{L}_c(\tilde{\tau}_j; \theta_{\mathrm{OLD}})), \tag{9}$$

where $\mathcal{L}_c$ is the cross-entropy loss measured on the prompt tasks. Performance improvement with θ_{NEW} against θ_{OLD} provides positive feedback to CDTC and reinforces CDTC to choose corresponding meta-tasks. Denote all the parameters of CDTC as $\psi = \{\mathbf{W_a}, \mathbf{W_g}, \mathbf{W_c}, \mathbf{w}\}$, including task representation learning, task representation refinement, and task selection modules. Due to the non-differentiability of the task sampling process, we employ the policy gradient algorithm REINFORCE (Williams 1992) to train CDTC as:

$$\psi^{(k+1)} = \psi^{(k)} - \gamma\nabla_{\psi^{(k)}}\log\pi(\psi^{(k)})(R - b), \tag{10}$$

where γ denotes the learning rate of CDTC and b is the baseline function, e.g., the moving average of the rewards, to reduce computational variance. And π refers to CDTC whose parameters ψ will be updated towards the direction of obtaining more rewards.

To jointly learn graph base learner (represented by θ) and CDTC (parameterized by ψ) at the meta-training phase, we adopt an alternative optimization strategy. At each iteration, a group of candidate meta-tasks is randomly sampled from the source domain. Then CDTC assigns the sampling probabilities to the candidate meta-tasks, in order to select a batch of meta-tasks. Thus the selected meta-tasks are utilized to execute a temporary update of graph base learner, and the reward is calculated to optimize CDTC. In practice, the temporary update can be conducted with a larger learning rate γ, such that CDTC collects feedback signals from an extensive range. Besides, CDTC is warmed up in the first iterations, after that the graph base learner will be updated along with the optimization of CDTC.

Experimental Evaluation

Experiment Setup

Datasets We conduct few-shot graph classification experiments with three datasets including **Tox21**, **SIDER**, and **MUV** from the molecular property prediction benchmarks. Each dataset is associated with a number of binary classification tasks. We explore cross-domain knowledge transfer between two molecular property prediction datasets, specifically including the transfer: (i) from Tox21 to SIDER, and vice versa; (ii) from MUV to Tox21; (iii) from MUV to SIDER. Domain shift exists in the above transfer process due to the distinct property prediction tasks and the diversely-structured molecular graphs in different datasets.

Methods	Tox21 → SIDER		SIDER → Tox21		MUV → Tox21		MUV → SIDER	
	1-shot	**10-shot**	**1-shot**	**10-shot**	**1-shot**	**10-shot**	**1-shot**	**10-shot**
GCN	54.46 ± 0.23	55.01 ± 1.89	61.61 ± 1.32	62.37 ± 1.52	63.82 ± 0.43	65.76 ± 0.95	55.19 ± 0.43	55.38 ± 0.73
GIN	54.73 ± 0.76	55.42 ± 2.32	62.41 ± 4.08	64.22 ± 2.19	62.17 ± 0.62	63.84 ± 0.83	54.79 ± 0.52	55.52 ± 0.89
Pre-GNN	54.82 ± 0.89	55.21 ± 1.01	63.61 ± 0.57	63.97 ± 1.23	66.57 ± 0.38	68.32 ± 0.43	53.77 ± 0.45	55.44 ± 0.56
MAML-GCN	54.65 ± 0.53	55.71 ± 0.26	59.37 ± 0.86	62.07 ± 1.23	63.18 ± 0.72	65.28 ± 1.53	54.25 ± 0.52	55.93 ± 0.75
Proto-GCN	54.12 ± 0.25	54.51 ± 0.53	59.24 ± 0.78	61.40 ± 0.62	61.32 ± 0.24	63.24 ± 0.43	53.36 ± 0.26	54.17 ± 0.42
MAML-GIN	54.93 ± 0.23	56.29 ± 0.72	61.63 ± 0.64	64.12 ± 1.15	58.29 ± 0.94	61.94 ± 1.43	55.53 ± 0.46	56.16 ± 0.47
Proto-GIN	54.01 ± 0.64	55.23 ± 0.56	60.92 ± 0.83	61.29 ± 0.45	58.36 ± 0.51	63.27 ± 0.21	54.19 ± 0.34	54.64 ± 0.26
PAR	54.96 ± 0.79	54.82 ± 0.11	63.72 ± 0.89	65.40 ± 1.60	66.49 ± 0.16	69.13 ± 0.13	54.47 ± 0.25	55.13 ± 0.12
CDTC	$\mathbf{55.87 \pm 0.23}$	$\mathbf{57.59 \pm 1.26}$	$\mathbf{67.48 \pm 0.74}$	$\mathbf{68.18 \pm 0.80}$	$\mathbf{70.16 \pm 0.17}$	$\mathbf{70.87 \pm 0.29}$	$\mathbf{56.18 \pm 0.09}$	$\mathbf{58.25 \pm 0.48}$

Table 1: Overall performance (ROC-AUC scores) of cross-domain few-shot graph classification.

Evaluation Metrics The performance of graph classification is evaluated by calculating the ROC-AUC score on the query set of each meta-testing task (Hu et al. 2020; Wang et al. 2021). Under the cross-domain settings, all the property prediction tasks in the target domain are employed for meta-testing. As the employed datasets basically contain binary classification tasks, we execute experiments under the 2-way K-shot settings.

Baselines We compare CDTC with baseline methods in two groups. First, we evaluate GNNs including **GCN** (Kipf and Welling 2017) and **GIN** (Xu et al. 2019), which are trained with source domain data and fine-tuned with few-shot annotations from the target domain for evaluation. We also compare to **Pre-GNN** (Hu et al. 2020) which pre-trains molecular graph encoder leveraging abundant molecules by various node-level and graph-level pre-training strategies. Second, we associate GNNs as the graph base learner with prevailing few-shot learning techniques including MAML (Finn, Abbeel, and Levine 2017) and ProtoNet (Snell, Swersky, and Zemel 2017) as the baselines. Specifically, we create **MAML-GCN**, **Proto-GCN**, **MAML-GIN**, **Proto-GIN** for comparison. Besides, we evaluate **PAR** (Wang et al. 2021) which uses target-related contextual information and pre-trained parameters (Hu et al. 2020) for few-shot graph classification.

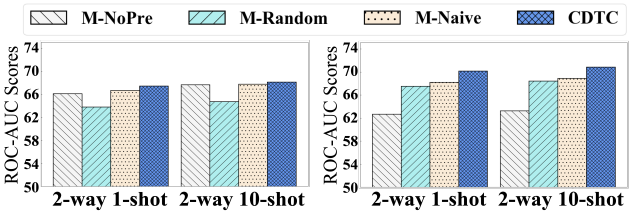
Experimental Settings We follow the general workflow of few-shot learning under the cross-domain setting and utilize the average performance of all meta-testing tasks in the target domain for evaluation. We conduct 5 random runs for each 2-way K-shot problem and present the average score over the repeated runs. Also, the standard deviation of ROC-AUC scores is reported. Regarding the randomness due to the small number of the support set, i.e., K=1 or K=10, model performance for each run is averaged based on 200 repeated meta-testing tasks. More experimental settings can be found in the reproducibility supplement of the Appendix.

Experimental Results

Overal Results The performance of CDTC and baseline models are reported in Table **??**, where the best results are highlighted in **bold** and the best baseline scores are underlined. It is observed that CDTC achieves the best performance under all the settings. Other observations are as follows. First, for the GNNs baselines, Pre-GNN shows better performance than GCN and GIN in most cases since GCN and GIN are trained from scratch. Yet in a few cases, Pre-GNN does not achieve significant improvements. We infer the cause to be that molecular graph datasets leveraged in Pre-GNN may preserve a large domain shift with the target domain, which thus limits the model generalization to the target domain. Second, GNNs perform competitively with a number of graph few-shot learning approaches, e.g. MAML-GCN. It implies that few-shot approaches with no special design for addressing domain shift issue might suffer from performance degradation, as domain shift violates the assumption of few-shot learning (Chen et al. 2019). Third, PAR outperforms other baselines in most cases probably due to the utilization of contextual information and pre-trained parameters. Lastly, our proposed CDTC achieves the best performance in all the cases. Compared to graph few-shot learning baselines without *solving the domain shift issue*, CDTC selects beneficial tasks during the meta-training phase to reduce the impact of domain shift.

Ablation Study To gain deeper insights into the effectiveness of different components in CDTC, we conduct ablation studies aiming for the following analysis: (RQ1) Whether the pre-trained parameters improve cross-domain knowledge transfer? (RQ2) Does CDTC provide beneficial meta-tasks to assist meta-training? (RQ3) Does task association modeling enhance task selection? Accordingly, we create three variants of CDTC, including: (a) **M-NoPre**, which initializes the GIN layers without pre-trained parameters; (b) **M-Random**, which conducts optimization-based meta-training with randomly sampled meta-tasks from the source domain; (c) **M-Naive**, which utilizes initial task representations for task selection bypassing task representation refinement. Performances of the variants are shown in Fig. 2, from which we find answers for:

- **RQ1**: M-NoPre shows inferior performances than CDTC, which demonstrates the advantage of using pre-trained parameters for knowledge transfer to the target domain.
- **RQ2**: M-Random is outperformed by CDTC, indicating the effectiveness of the proposed model on selecting beneficial meta-tasks and mitigating domain shift.
- **RQ3**: M-Naive has worse performances compared to CDTC, implying task association modeling could refine task representations for better task selection.



(a) SIDER → Tox21 (b) MUV → Tox21

Figure 2: Results of ablation study on few-shot graph classification tasks for variants of CDTC under cross-domain settings: (a) SIDER → Tox21 and (b) MUV → Tox21.

Methods	Tox21 → SIDER		MUV → Tox21	
	1-shot	10-shot	1-shot	10-shot
MEAN	55.36 ± 0.15	56.44 ± 0.64	70.89 ± 0.12	71.29 ± 0.52
CONCAT	55.87 ± 0.23	57.59 ± 1.26	70.16 ± 0.17	70.87 ± 0.29

Table 2: Few-shot graph classification performance (ROC-AUC) with different methods of task representation learning

Impact of Task Representation Learning Strategies In this work, we obtain task representations for 2-way K-shot tasks based on class prototype concatenation. When it comes to a larger N, the permutation-invariant function can be applied to reduce computational cost. We compare the performance of model variants adopting the mean-pooling operation (MEAN) and concatenation operation (CONCAT). As shown in Table 2, the variants of CDTC associated with the two different methods achieve competitive results.

Impact of Different Numbers of Prompt Tasks Here, we analyze the impact of the number of prompt tasks (M_t) employed in CDTC. We evaluate model performance with experiments on MUV → Tox21 by varying the numbers of prompt tasks. Considering that Tox21 contains 12 molecular property prediction problems, we sample one prompt task for each property prediction problem and thus determine the largest number of prompt tasks to be 12. We reduce the values of M_t by sampling one prompt task from a part of randomly chosen properties. The performance with different M_t is reported, averaged over 3 random runs. As shown in Fig. 3, CDTC obtains improved performance when the value of M_t increases. The reason can be two folds: 1) More prompt tasks facilitate task association modeling for more accurate task representations; 2) The reward received from more prompt tasks can better reflect the suitableness of the candidate meta-tasks over the target domain.

Analysis of Task Relevance Learning task relevance is one of the key features of CDTC. Therefore, we verify the relationship between task correlation and task selection with experiments on MUV → Tox21 under a 2-way 10-shot setting. Specifically, we randomly select a set of candidate meta-tasks and prompt tasks during meta-training after sufficient iterations and calculate their correlation by Pearson correlation coefficient with the task representations (in Fig. 4 (a)). It is expected that meta-tasks having higher correlation scores with the prompt tasks would be more beneficial

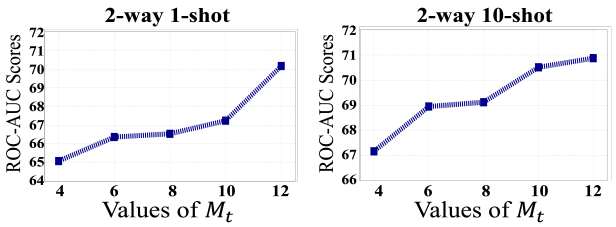


Figure 3: Few-shot graph classification performance (ROC-AUC) leveraging varied M_t values on MUV → Tox21.

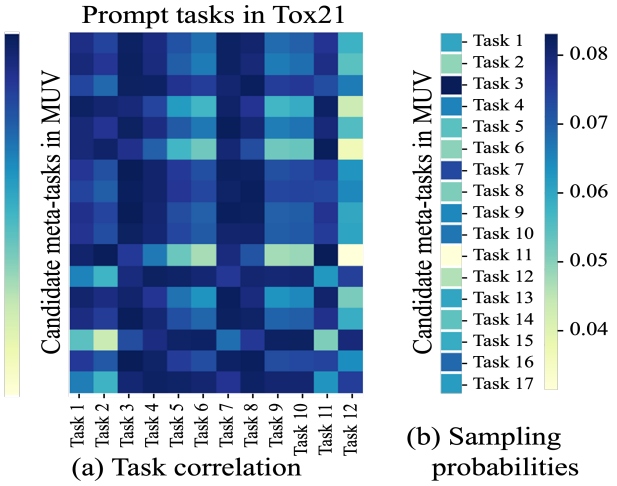


Figure 4: Analysis of task correlation and task sampling probabilities on MUV → Tox21.

for the generalization to the target domain. Meanwhile, we present the task sampling probabilities that CDTC produces for the set of candidate meta-tasks (in Fig. 4 (b)). It is observed that candidate meta-tasks having higher correlation scores regarding the prompt tasks in Fig. 4 (a) are assigned with higher sampling probabilities in Fig. 4 (b), demonstrating that CDTC is capable of selecting beneficial meta-tasks to improve model generalizability.

Conclusion

In this work, we target the problem of cross-domain few-shot graph classification and propose a novel model named CDTC to address the problem. CDTC adaptively selects appropriate meta-tasks for training the graph base learner with optimization-based meta-learning. To explore the relevance between tasks, CDTC develops task association modeling to obtain refined task representations from the task bipartite graph and task selection to choose beneficial meta-tasks. To better make use of small clues in the target domain, CDTC leverages the rewards evaluated on the target domain to reinforce its selection. The proposed CDTC can be integrated well with optimization-based meta-learning and trained in an end-to-end manner. Extensive experimental results on three molecular property prediction datasets verify the effectiveness of CDTC for few-shot graph classification for various cross-domain settings.

Acknowledgements

The research reported in this paper was partially supported by funding from the King Abdullah University of Science and Technology (KAUST).

References

Abbasi, K.; Poso, A.; Ghasemi, J.; Amanlou, M.; and Masoudi-Nejad, A. 2019. Deep transferable compound representation across domains and tasks for low data drug discovery. *Journal of chemical information and modeling*, 59(11): 4528–4539.

Bose, A. J.; Jain, A.; Molino, P.; and Hamilton, W. L. 2019. Meta-graph: Few shot link prediction via meta learning. *arXiv preprint arXiv:1912.09867*.

Chauhan, J.; Nathani, D.; and Kaul, M. 2020. Few-Shot Learning on Graphs via Super-Classes based on Graph Spectral Measures. In *ICLR*.

Chen, W.-Y.; Liu, Y.-C.; Kira, Z.; Wang, Y.-C. F.; and Huang, J.-B. 2019. A Closer Look at Few-shot Classification. In *ICLR*.

Das, D.; Yun, S.; and Porikli, F. 2022. ConFeSS: A Framework for Single Source Cross-Domain Few-Shot Learning. In *ICLR*.

Ding, K.; Wang, J.; Li, J.; Shu, K.; Liu, C.; and Liu, H. 2020. Graph prototypical networks for few-shot learning on attributed networks. In *CIKM*.

Du, Y.; Zhen, X.; Shao, L.; and Snoek, C. G. 2020. Metanorm: Learning to normalize few-shot batches across domains. In *ICLR*.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*.

Guo, H.; Pasunuru, R.; and Bansal, M. 2020. Multi-source domain adaptation for text classification via distancenet-bandits. In *AAAI*.

Guo, Z.; Zhang, C.; Yu, W.; Herr, J.; Wiest, O.; Jiang, M.; and Chawla, N. V. 2021. Few-Shot Graph Learning for Molecular Property Prediction. In *WWW*.

Hassani, K. 2022. Cross-Domain Few-Shot Graph Classification. In *AAAI*.

Hospedales, T.; Antoniou, A.; Micaelli, P.; and Storkey, A. 2020. Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439*.

Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; and Leskovec, J. 2020. Strategies For Pre-training Graph Neural Networks. In *ICLR*.

Huang, K.; and Zitnik, M. 2020. Graph Meta Learning via Local Subgraphs. In *NeurIPS*.

Huang, T.; Xu, K.; and Wang, D. 2020. DA-HGT: Domain Adaptive Heterogeneous Graph Transformer. *arXiv preprint arXiv:2012.05688*.

Jiang, Z.; Wang, J.; Zhao, L.; Sun, C.; Lu, Y.; and Liu, X. 2019. Cross-domain Aspect Category Transfer and Detection via Traceable Heterogeneous Graph Representation Learning. In *CIKM*.

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

Kirchmeyer, M.; Rakotomamonjy, A.; de Bezenac, E.; and Gallinari, P. 2022. Mapping conditional distributions for domain adaptation under generalized target shift. In *ICLR*.

Li, D.; Zhang, J.; Yang, Y.; Liu, C.; Song, Y.-Z.; and Hospedales, T. M. 2019. Episodic training for domain generalization. In *ICCV*.

Li, Z.; Zhou, F.; Chen, F.; and Li, H. 2017. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*.

Liu, L.; Zhou, T.; Long, G.; Jiang, J.; and Zhang, C. 2019. Learning to propagate for graph meta-learning. In *NeurIPS*.

Ma, Y.; Bai, S.; An, S.; Liu, W.; Liu, A.; Zhen, X.; and Liu, X. 2020. Transductive Relation-Propagation Network for Few-shot Learning. In *IJCAI*.

Mesquita, D.; Souza, A.; and Kaski, S. 2020. Rethinking pooling in graph neural networks. In *NeuIPS*.

Shen, X.; Dai, Q.; Chung, F.-l.; Lu, W.; and Choi, K.-S. 2020a. Adversarial deep network embedding for cross-network node classification. In *AAAI*.

Shen, X.; Dai, Q.; Mao, S.; Chung, F.-l.; and Choi, K.-S. 2020b. Network together: Node classification via cross-network deep network embedding. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5): 1935–1948.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. In *NeurIPS*.

Tseng, H.-Y.; Lee, H.-Y.; Huang, J.-B.; and Yang, M.-H. 2020. Cross-Domain Few-Shot Classification via Learned Feature-Wise Transformation. In *ICLR*.

Venkat, N.; Kundu, J. N.; Singh, D.; Revanur, A.; et al. 2020. Your classifier can secretly suffice multi-source domain adaptation. In *NeuIPS*.

Wang, M.; and Deng, W. 2018. Deep visual domain adaptation: A survey. *Neurocomputing*, 312: 135–153.

Wang, N.; Luo, M.; Ding, K.; Zhang, L.; Li, J.; and Zheng, Q. 2020. Graph Few-shot Learning with Attribute Matching. In *CIKM*.

Wang, Y.; Abuduweili, A.; Yao, Q.; and Dou, D. 2021. Property-aware relation networks for few-shot molecular property prediction. In *NeuIPS*.

Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Reinforcement learning*, 5–32.

Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; and Philip, S. Y. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1): 4–24.

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2019. How powerful are graph neural networks? *ICLR*.

Yang, S.; Song, G.; Jin, Y.; and Du, L. 2021. Domain adaptive classification on heterogeneous information networks. In *IJCAI*.

Yao, H.; Zhang, C.; Wei, Y.; Jiang, M.; Wang, S.; Huang, J.; Chawla, N.; and Li, Z. 2020. Graph few-shot learning via knowledge transfer. In *AAAI*.

Zhang, Q.; Wu, X.; Yang, Q.; Zhang, C.; and Zhang, X. 2022a. Few-shot Heterogeneous Graph Learning via Cross-domain Knowledge Transfer. In *KDD*.

Zhang, Q.; Wu, X.; Yang, Q.; Zhang, C.; and Zhang, X. 2022b. HG-Meta: Graph Meta-learning over Heterogeneous Graphs. In *SDM*.

Zhang, Y.; Song, G.; Du, L.; Yang, S.; and Jin, Y. 2019. DANE: domain adaptive network embedding. In *IJCAI*.

Zhao, H.; Des Combes, R. T.; Zhang, K.; and Gordon, G. 2019. On learning invariant representations for domain adaptation. In *ICML*.

Zhao, S.; Li, B.; Xu, P.; and Keutzer, K. 2020. Multi-source domain adaptation in the deep learning era: A systematic survey. *arXiv preprint arXiv:2002.12169*.

Zhou, Y.; Zhu, F.; Song, P.; Han, J.; Guo, T.; and Hu, S. 2021. An Adaptive Hybrid Framework for Cross-domain Aspect-based Sentiment Analysis. In *AAAI*.