

DiffMG: Differentiable Meta Graph Search for Heterogeneous Graph Neural Networks

Yuhui Ding¹, Quanming Yao^{2,3}, Huan Zhao², Tong Zhang¹

¹Department of Computer Science and Engineering, Hong Kong University of Science and Technology

²4Paradigm Inc. ³Department of Electronic Engineering, Tsinghua University

yuhui.ding@connect.ust.hk, {yaoquanming, zhaohuan}@4paradigm.com, tongzhang@ust.hk

ABSTRACT

In this paper, we propose a novel framework to automatically utilize task-dependent semantic information which is encoded in heterogeneous information networks (HINs). Specifically, we search for a meta graph, which can capture more complex semantic relations than a meta path, to determine how graph neural networks (GNNs) propagate messages along different types of edges. We formalize the problem within the framework of neural architecture search (NAS) and then perform the search in a differentiable manner. We design an expressive search space in the form of a directed acyclic graph (DAG) to represent candidate meta graphs for a HIN, and we propose task-dependent type constraint to filter out those edge types along which message passing has no effect on the representations of nodes that are related to the downstream task. The size of the search space we define is huge, so we further propose a novel and efficient search algorithm to make the total search cost on a par with training a single GNN once. Compared with existing popular NAS algorithms, our proposed search algorithm improves the search efficiency. We conduct extensive experiments on different HINs and downstream tasks to evaluate our method, and experimental results show that our method can outperform state-of-the-art heterogeneous GNNs and also improves efficiency compared with those methods which can implicitly learn meta paths¹.

CCS CONCEPTS

- **Information systems** → **Data mining**: *Recommender systems*;
- **Computing methodologies** → **Neural networks**.

KEYWORDS

Heterogeneous information networks; Neural architecture search; Graph neural networks.

ACM Reference Format:

Yuhui Ding, Quanming Yao, Huan Zhao, Tong Zhang. 2021. DiffMG: Differentiable Meta Graph Search for Heterogeneous Graph Neural Networks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and*

Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3447548.3467447>

1 INTRODUCTION

Heterogeneous information networks (HINs) are ubiquitous in real-world applications, such as web mining [34], recommender systems [14] and biomedicine [41], to model complex relations among individual entities. Alongside the network topology, a HIN also encodes rich semantic information which consists of multiple node types, edge types and node features which come from different sources. For example, Microsoft Academic Graph [25] has multiple types of nodes: authors, papers, venues, etc., and multiple types of relations among them (see Figure 1(a)). Such additional semantic information increases the difficulty of representation learning on HINs, compared with homogeneous networks [7, 26, 32, 38].

Recently, graph neural networks (GNNs), which typically fall into a common message passing framework [9], have achieved remarkable progress in many graph-based machine learning tasks [9, 15, 36]. Within this framework, each node generates its representation by aggregating features of its neighbors. Despite the success of GNNs on homogeneous networks, it has been observed that, naive message passing on HINs which ignores node types and edge types often leads to sub-optimal performance [26, 32]. To utilize semantic information encoded in HINs, MAGNN [7] and HAN [26] are designed specifically for HINs. They employ meta paths [21] to define what neighbors to aggregate information from. However, it can be extremely difficult to design meta paths by hand with little or no domain knowledge, especially for complicated heterogeneous systems (e.g., various types of chemical atoms).

Other heterogeneous GNN variants, e.g., GTN [32] and HGT [12], implicitly learn meta paths by fusing different types based on attention mechanisms. However, such methods face two potential issues. First, due to the heterogeneity, not all semantic information is useful to the downstream task. For example, when we want to predict research areas for authors in an academic network (Figure 1(a)), institutions that authors are affiliated with are unrelated information and may bring noise to author representations. Second, mechanisms used to fuse different node types and edge types increase model complexity and harm the efficiency. For example, HGT [12] adapts the transformer architecture [23] and keeps distinct hidden weight matrices for each node type and each edge type at each layer to compute attention scores, which makes it much less efficient than homogeneous counterparts.

Moreover, above methods only consider meta paths, which have limited capacity to capture intricate semantic proximity compared with more expressive meta graphs [13, 39]. A more recent method GEMS [11] utilizes the evolutionary algorithm to search for meta

¹Code is available at: <https://github.com/AutoML-4Paradigm/DiffMG>. Q. Yao is the corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
KDD '21, August 14–18, 2021, Virtual Event, Singapore

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8332-5/21/08...\$15.00

<https://doi.org/10.1145/3447548.3467447>

Table 1: Comparison of heterogeneous GNNs. Semantics: what kind of semantic information a method utilizes. Construction: how a method finds the semantics.

Method	Semantics	Construction
HAN [26]	Meta path	Manual design
MAGNN [7]	Meta path	Manual design
GTN [32]	Meta path	Implicitly learn by attention
HGT [12]	Meta path	Implicitly learn by attention
GEMS [11]	Meta graph	Genetic search
DiffMG	Meta graph	Differentiable search

graphs between source node types and target node types for recommendation, but it is too slow to fully explore the search space within a reasonable time budget. Therefore, it has not been well addressed yet how GNNs can automatically utilize task-dependent semantic information in HINs and at the same time keep high efficiency (see Table 1 for a summary).

In this paper, we propose a novel method to address the above issue, namely DiffMG (Differentiable Meta Graph search). We are inspired by recent advancements in one-shot neural architecture search (NAS) [3, 17] which automatically discovers novel promising CNN architectures. Within the context of HIN, DiffMG searches for a meta graph in a differentiable fashion to guide how GNNs propagate messages along different types of edges. DiffMG is trained in a two stage manner. At the search stage, DiffMG defines a search space in the form of a directed acyclic graph (DAG) to represent a candidate meta graph, and then architecture parameters are introduced to weight candidate edge types for each link in the DAG search space. Moreover, we apply task-dependent type constraint to particular links in the DAG, in order to filter out those candidate edge types along which message passing has no effect on the representations of nodes which are related to the downstream task. The architecture parameters are then optimized end to end through bi-level optimization [6]. At the evaluation stage, we derive a meta graph by selecting the most promising edge type for each link in the search space, indicated by architecture parameters, and then DiffMG is trained with the derived meta graph. To improve the search efficiency, we propose an efficient search algorithm, so that only one edge type, instead of all candidates, for each link in the DAG is involved in the computation per iteration during the search. We evaluate DiffMG extensively on different heterogeneous datasets, and experimental results show that DiffMG outperforms state-of-the-art heterogeneous GNNs and also improves efficiency compared with methods that learn meta paths implicitly [12, 32]. To summarize, the main contributions of our work are:

- We address the problem of utilizing task-dependent semantic information in HINs by searching for a meta graph within the framework of NAS;
- We design an expressive DAG search space to represent candidate meta graphs and propose task-dependent type constraint to filter out unrelated edge types;
- We propose an efficient search algorithm to make the search cost on a par with training a single GNN once;

- We conduct extensive experiments to demonstrate the effectiveness and efficiency of our proposed method.

2 PRELIMINARIES AND NOTATIONS

Definition 2.1 (Heterogeneous Information Network (HIN) [29]). A heterogeneous information network is a directed graph which has the form $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathcal{T}, \mathcal{R}, f_{\mathcal{T}}, f_{\mathcal{R}}\}$, where \mathcal{V} denotes the set of nodes and \mathcal{E} denotes the set of edges. $f_{\mathcal{T}} : \mathcal{V} \rightarrow \mathcal{T}$ is a mapping function which maps each node $v \in \mathcal{V}$ to a node type $f_{\mathcal{T}}(v) \in \mathcal{T}$. Similarly, $f_{\mathcal{R}}$ maps each edge $e \in \mathcal{E}$ to an edge type $f_{\mathcal{R}}(e) \in \mathcal{R}$. We call $\mathcal{S} = \{\mathcal{T}, \mathcal{R}\}$ the network schema of a heterogeneous information network, with $|\mathcal{T}| + |\mathcal{R}| > 2$.

We use \mathbf{A}_r to denote the adjacency matrix formed by the edges of type $r \in \mathcal{R}$, and \mathcal{A} to denote the collection of all \mathbf{A}_r 's. An edge type is determined by its source node type and target node type. For example, in Figure 1(a), \mathbf{A}_{PA} represents directed edges from author nodes (A) to paper nodes (P), while \mathbf{A}_{AP} represents directed edges from P to A. If the source node type and the target node type of an edge type r are the same, then \mathbf{A}_r is symmetric.

Definition 2.2 (Meta Path [21]). A meta path P is a composite relation defined on \mathcal{S} which consists of multiple edge types, i.e., $P = t_1 \xrightarrow{r_1} t_2 \xrightarrow{r_2} \dots \xrightarrow{r_l} t_{l+1}$, where $t_1, \dots, t_{l+1} \in \mathcal{T}$ and $r_1, \dots, r_l \in \mathcal{R}$. One meta path corresponds to many meta path instances in the underlying HIN.

Definition 2.3 below offers a natural generalization of meta paths, which allows the in-degree of each node type (except the source node type) to be larger than 1. Intuitively, meta graphs can represent more intricate semantic relations than meta paths. For example, to determine whether two people are classmates or alumni [37], school information alone is not enough. Class information is also required.

Definition 2.3 (Meta Graph [39]). A meta graph is a directed acyclic graph (DAG) on \mathcal{S} , with a single source node type (i.e., with in-degree 0) and a single sink (target) node type (i.e., with out-degree 0).

We use N to denote the total number of nodes in the HIN and d to denote the hidden dimension. We use uppercase bold letters (e.g., \mathbf{X}) to denote matrices, and lowercase bold letters (e.g., \mathbf{z}) to denote column vectors.

3 RELATED WORK

3.1 Heterogeneous GNNs

Heterogeneous GNNs are designed for HINs to utilize semantic information which is ignored by homogeneous ones. One category of heterogeneous GNNs [7, 14, 26, 38] employs hand-designed meta paths (Definition 2.2) to define neighbors. HAN [26] extracts multiple homogeneous sub-networks based on different meta paths and uses semantic-level attention to combine node representations from different sub-networks. MAGNN [7] utilizes RotatE [22] to encode intermediate nodes along each meta path instance and combines multiple meta paths like HAN. NIREC [14] proposes an interaction model which is based on meta path guided neighborhood for recommendation. Compared with these methods, DiffMG does not require prior knowledge of domain specific rules.

Another category of heterogeneous GNNs [12, 32, 33] designs mechanisms to fuse information from different types of nodes, in order to eliminate the need for hand-designed meta paths. GTN [32] combines adjacency matrices of different edge types using attention scores and learns a new meta path based graph via matrix multiplication, but it suffers from huge memory cost. HetGNN [33] also mixes information of different node types by attention. However, HetGNN is designed for a scenario where each node is associated with multi-modal contents (e.g., text and image), which is different from our setting, and it employs bi-directional LSTM to encode different types of contents. HGT [12] keeps distinct hidden weight matrices for each type at each layer and computes mutual attention scores to weight messages coming along different types of relations. Different from these methods, DiffMG can select task-dependent semantic information rather than fuse all candidate node types and edge types together. In this way, DiffMG filters out potential noise which is unrelated to the downstream task and also improves model efficiency.

Moreover, none of the above methods takes meta graphs [4, 13, 39] into account, which show stronger capacity to express finer-grained semantics [37]. GEMS [11] utilizes the evolutionary algorithm to search for meta graphs between source nodes and target nodes for recommendation on HINs. However, new candidate meta graphs which are generated by evolution have to be retrained with the GNN module from scratch for evaluation, which makes it too inefficient. Compared with GEMS, DiffMG performs search in a differentiable fashion and the total search cost is on a par with training a single GNN once.

3.2 Neural Architecture Search (NAS)

Neural architecture search (NAS) [3] has become a promising way to automatically discover novel architectures which outperform hand-designed ones. Both the search space and the search algorithm are crucial to the effectiveness and efficiency of NAS.

- The search space is defined as the set of all possible architectures. It is task-specific and should be general enough to cover existing models. At the same time, it should be tractable to avoid high search cost. For example, by exploring the transferable blocks across different tasks, the search space of CNNs has evolved from a macro architecture [42] to a micro cell structure [17], which significantly reduces the search cost on large image datasets [43].
- The search algorithm aims to find the desired architecture in a given space. Pioneer works [20, 43] evaluate candidate architectures through stand alone training and require thousands of GPU days to obtain a good convolutional neural network. Later, one-shot NAS methods [17, 19, 27, 30] have improved the search efficiency by orders of magnitude via parameter sharing, i.e., candidate child neural networks share parameters of a single super neural network.

More recently, some works have extended NAS to homogeneous GNNs [8, 16, 31, 40]. GraphNAS [8] utilizes reinforcement learning to select proper GNN components (e.g., aggregation function, number of attention heads). GraphGym [31] proposes a GNN design space and a GNN task space to comprehensively evaluate model-task combinations. Policy-GNN [16] uses reinforcement learning to determine the number of aggregation layers for each

node. SANE [40] follows DARTS [17] to search for combinations of aggregation functions.

However, how to design a search space for heterogeneous GNNs which is aware of semantic information in HINs is challenging, and how to efficiently search the given space is also a non-trivial problem. Both problems have not been addressed by above works.

4 THE PROPOSED METHOD

In this section, we present the components of our proposed method in detail. In Section 4.1, we define a search space to represent candidate meta graphs, and describe how a meta graph in this search space determines message passing along different types of edges in the underlying HIN. Our search space has two key features:

- First, it is a DAG which allows more than one incoming link for each intermediate node (Section 4.1.1). This feature enhances its capacity to represent more complex semantics than meta paths, i.e., meta graphs.
- Second, it applies task-dependent type constraint in order to filter out those edge types along which message passing has no effect on the final representations of nodes which are related to the downstream task (Section 4.1.2).

Next, in Section 4.2 we define the search objective, which presents a continuous relaxation of the designed search space. Finally, in Section 4.3 we describe how DiffMG performs search in a differentiable way and propose an efficient search algorithm.

In the sequel, we will build our method based on graph convolutional network (GCN) [15], which is the most popular GNN instance. Other variants which also fall into the message passing framework, like GraphSAGE [10], GAT [24] and GIN [28], can be employed and extended in a similar way. We leave this for future work and focus on searching for meta graphs in this paper.

4.1 Search Space

Recall that a GCN layer for homogeneous networks has the form:

$$\mathbf{Z} = \sigma(\hat{\mathbf{A}}\mathbf{X}\Theta), \quad (1)$$

where $\mathbf{X} \in \mathbb{R}^{N \times d}$ and $\mathbf{Z} \in \mathbb{R}^{N \times d}$ denote the input and output node representations of this layer, respectively. Θ denotes a hidden weight matrix shared by all nodes, and σ denotes the activation function. $\hat{\mathbf{A}}$ denotes the normalized adjacency matrix which represents information propagation between homogeneous one-hop neighbors. However, for heterogeneous networks, (1) cannot differentiate different node types and thus loses semantic information.

To utilize the semantic information, it is important to differentiate messages which come along different edge types. Moreover, different combinations of messages from different edge types will generate different node representations. Therefore, we can construct a meta graph to describe the message passing process, by selecting edge types and forming appropriate connections between them. This motivates us to extend (1) to the following heterogeneous message passing layer:

$$\mathbf{Z} = \sigma(F_{\mathcal{A}}(\mathbf{X}'\Theta)), \quad (2)$$

where \mathbf{X}' denotes the projected features output by type-specific transformation [7, 26] which projects features of different node

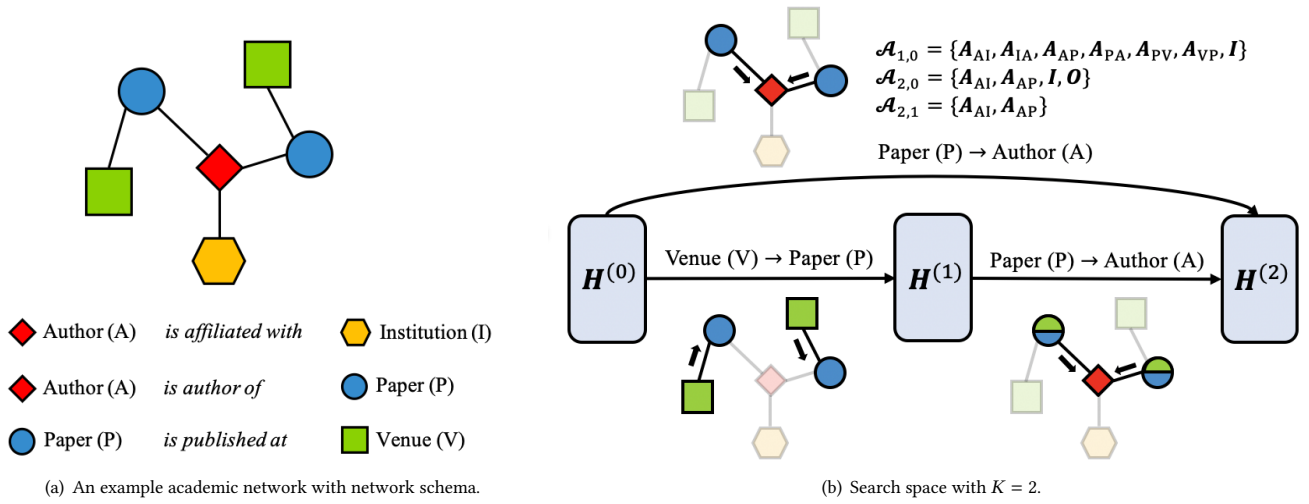


Figure 1: Illustration of $F_{\mathcal{A}}(\cdot)$ with $K = 2$ for an example academic network (best viewed in color). Hidden weight matrices and non-linearity are omitted for ease of illustration. Here author (A) is the target node type related to the evaluation task which predicts research areas, and the derived meta graph does not propagate information from I to A. Candidate edge types for each link in the search space are shown with task-dependent type constraint.

types into a common latent space. Here $F_{\mathcal{A}}(\cdot)$ represents the message passing process which is aware of the edge types, i.e., \mathcal{A} , and can be described by a meta graph.

4.1.1 Structure of $F_{\mathcal{A}}(\cdot)$. Specifically, we define $F_{\mathcal{A}}(\cdot)$ as a DAG, in which ordered nodes denote intermediate representations in the message passing process. The input of $F_{\mathcal{A}}(\cdot)$ is denoted as $H^{(0)}$ and the output is $H^{(K)}$, with K the predefined number of intermediate states. In this DAG, $H^{(k)}$ ($1 \leq k \leq K$) is generated based on all its predecessors, i.e.,

$$H^{(k)} = \sum_{0 \leq i < k} \tilde{f}_{k,i}(H^{(i)}; \mathcal{A}_{k,i}), \quad (3)$$

where $\tilde{f}_{k,i}(\cdot)$ denotes one message passing step which aggregates $H^{(i)}$ along edges of a certain type which is selected from $\mathcal{A}_{k,i}$. Here we choose graph convolution [15] as our aggregation function.

In Figure 1(b), we illustrate a message passing process $F_{\mathcal{A}}(\cdot)$ with $K = 2$ for an example academic network (Figure 1(a)). $\tilde{f}_{1,0}(\cdot)$ aggregates information from venue nodes to paper nodes, and then $\tilde{f}_{2,1}(\cdot)$ passes the aggregated information to the target author node. At the same time, $\tilde{f}_{2,0}(\cdot)$ also passes information from paper nodes to the author node. Note that both $\tilde{f}_{2,1}(\cdot)$ and $\tilde{f}_{2,0}(\cdot)$ choose the edge type A_{AP} to propagate information, however, the messages they pass are different because paper representations at the state $H^{(1)}$ already aggregate information from venue nodes (mixing green and blue in Figure 1(b)). This example shows that the DAG structure can flexibly combine multiple information propagation paths.

4.1.2 Contents of $\mathcal{A}_{k,i}$. Here we describe what is contained in each $\mathcal{A}_{k,i}$. In addition to edge types of the original HIN, i.e., \mathcal{A} , we add the identity matrix I to $\mathcal{A}_{k,i}$, which allows the number of actual message passing steps in the searched meta graph to be flexible. Moreover, we add the empty matrix O to $\mathcal{A}_{k,i}$ ($i < k - 1$), so that if the empty matrix receives the highest weight after search,

we will drop the contribution of $H^{(i)}$ to $H^{(k)}$ ($i < k - 1$). This mechanism allows the number of incoming links for $H^{(k)}$ in the searched meta graph to be flexible in a range of $[1, k]$, and if all $H^{(k)}$'s have only one incoming link, then the searched meta graph becomes a meta path. These two additional matrices enable us to search for a flexible meta graph in the predefined search space.

Besides, we notice that in a HIN only node representations of particular types are related to the downstream task. For example, when predicting research areas for authors in Figure 1(a), only author representations are related to the evaluation. We denote the collection of those edge types whose target node type is related to the evaluation as $\overline{\mathcal{A}}$, and then candidate edge types for message passing steps that contribute to final representations $H^{(K)}$ should be in $\overline{\mathcal{A}}$, otherwise message passing will have no effect on the nodes related to the evaluation. To summarize, we define $\mathcal{A}_{k,i}$ as:

$$\mathcal{A}_{k,i} = \begin{cases} \mathcal{A} \cup \{I\} & k < K \text{ and } i = k - 1 \\ \mathcal{A} \cup \{I\} \cup \{O\} & k < K \text{ and } i < k - 1 \\ \overline{\mathcal{A}} & k = K \text{ and } i = K - 1 \\ \overline{\mathcal{A}} \cup \{I\} \cup \{O\} & k = K \text{ and } i < K - 1 \end{cases}, \quad (4)$$

Even though the task-dependent type constraint filters out certain candidate edge types, the search space is still very large. For Douban movie recommendation dataset with 11 edge types (see Appendix B), when $K = 4$, to generate user representations the total number of possible meta graphs is $(11+1)^3 \times (11+2)^3 \times 3 \times (3+2)^3 \approx 1.4 \times 10^9$, which makes it necessary to design an efficient search algorithm.

REMARK 1. To integrate Definition 2.3 into the message passing framework, we omit the constraint of a single source node type, and generalize a meta graph to a more flexible subgraph defined on the network schema which only requires a single target node type.

4.2 Search Objective

At the search stage, to differentiate the importance of candidate edge types for each link in the DAG, we introduce the architecture parameters $\lambda_{k,i}^m$ ($m = 1, \dots, |\mathcal{A}_{k,i}|$) to mix them in $\tilde{f}_{k,i}(\cdot)$:

$$\alpha_{k,i}^m = \exp(\lambda_{k,i}^m) / \sum_{m'=1}^{|\mathcal{A}_{k,i}|} \exp(\lambda_{k,i}^{m'}), \quad (5)$$

$$\tilde{f}_{k,i}(\mathbf{H}^{(i)}; \mathcal{A}_{k,i}) = \sum_{m=1}^{|\mathcal{A}_{k,i}|} \alpha_{k,i}^m \cdot f(\mathbf{H}^{(i)}; \mathcal{A}_{k,i}^m), \quad (6)$$

where $f(\mathbf{H}^{(i)}; \mathcal{A}_{k,i}^m)$ denotes to aggregate information $\mathbf{H}^{(i)}$ from neighbors defined by $\mathcal{A}_{k,i}^m$, the m -th edge type in $\mathcal{A}_{k,i}$.

Let ω denote the conventional parameters (hidden weight matrices and biases) of our model and λ be the collection of all architecture parameters. Following the NAS framework [17, 30], at the search stage, we aim to solve a bi-level optimization problem:

$$\min_{\lambda} \mathcal{L}_{\text{val}}(\omega^*(\lambda), \lambda), \text{ s.t. } \omega^*(\lambda) = \arg \min_{\omega} \mathcal{L}_{\text{tra}}(\omega, \lambda), \quad (7)$$

where \mathcal{L}_{val} and \mathcal{L}_{tra} represent the validation loss and the training loss, respectively. Next, we show two popular tasks over HINs which will be examined in experiments.

4.2.1 Node Classification. Node representations output by the heterogeneous message passing layer can be adapted for different downstream tasks. For the node classification task, we append a linear layer after the heterogeneous message passing layer, which reduces the hidden dimension to the number of classes, i.e.,

$$\bar{\mathbf{Y}} = \text{softmax}(\mathbf{Z}\mathbf{W}_o),$$

where $\mathbf{W}_o \in \mathbb{R}^{d \times C}$ is the output weight matrix and C is the number of classes. Then, we use cross-entropy loss over all labeled nodes as

$$\mathcal{L} = - \sum_{v \in \mathcal{V}_L} \sum_{c=1}^C \mathbf{y}_v[c] \log \tilde{\mathbf{y}}_v[c],$$

where \mathcal{V}_L denotes the set of labeled nodes, \mathbf{y}_v is a one-hot vector indicating the label of node v , and $\tilde{\mathbf{y}}_v$ is the predicted label for the corresponding node in $\bar{\mathbf{Y}}$.

4.2.2 Recommendation. For the recommendation task, we use the following loss:

$$\mathcal{L} = - \sum_{(u,v) \in \Omega^+} \log \sigma(\mathbf{z}_u^\top \mathbf{z}_v) - \sum_{(u',v') \in \Omega^-} \log \sigma(-\mathbf{z}_{u'}^\top \mathbf{z}_{v'}), \quad (8)$$

where Ω^+ and Ω^- denote the set of observed positive pairs and the set of negative pairs respectively, and σ denotes the sigmoid function. $\mathbf{z}_u, \mathbf{z}_v, \mathbf{z}_{u'}, \mathbf{z}_{v'}$ are node representations output by the heterogeneous message passing layer.

4.3 Search Algorithm

The architecture parameters enable us to mix candidate meta graphs as in (6). However, naive training with mixed operations, as existing one-shot NAS algorithms (e.g., DARTS [17]), has two potential issues. First, to evaluate training and validation losses we need to compute message passing once for each candidate edge type in the mixed operation (6) and for each link in the DAG search space, which is not efficient enough. Second, training with mixed operations is not well consistent with our final goal, i.e., to derive a single promising meta graph, because different candidate message passing steps in (6) may correlate with each other [1] and thus metrics evaluated using the mixed model at the search stage cannot

well indicate the performance of the derived model. Such issues motivate us to propose a new search algorithm to facilitate the search process.

4.3.1 First-order Approximation. While $\omega^*(\lambda)$ is coupled with λ in (7), such an issue can be empirically addressed by approximating $\omega^*(\lambda)$ with the ω dynamically maintained during training [17, 27, 30]. Then, the optimization is performed with alternate updates, where in the first phase of each iteration λ is fixed and ω is updated based on $\partial \mathcal{L}_{\text{tra}} / \partial \omega$, and in the second phase λ is updated based on $\partial \mathcal{L}_{\text{val}} / \partial \lambda$ while ω being fixed.

4.3.2 Differentiating Argmax Operation. To align the search procedure with the goal of deriving a good meta graph, in each iteration we consider transforming $\alpha_{k,i}^m$ in (6) into $\bar{\alpha}_{k,i}^m$ for loss evaluation:

$$\bar{\alpha}_{k,i}^m = \begin{cases} \alpha_{k,i}^m & m = m^* \\ 0 & \text{otherwise} \end{cases}, \quad (9)$$

where $m^* = \arg \max_m \alpha_{k,i}^m$. With (9), only the message passing step associated with the maximum weight in (6) needs to be computed to obtain $\partial \mathcal{L}_{\text{tra}} / \partial \omega$. However, when we are updating λ , the transformation from $\alpha_{k,i}$ to $\bar{\alpha}_{k,i}$ in (9) is not differentiable and prevents gradients from propagating backwards to $\alpha_{k,i}$ and $\lambda_{k,i}$. We address this issue with a softmax approximation:

$$\bar{\alpha}_{k,i}^m = \lim_{t \rightarrow 0^+} \alpha_{k,i}^m \cdot h(m; t), \quad (10)$$

where $h(m; t) = \exp(\alpha_{k,i}^m / t) / \sum_{p=1}^{|\mathcal{A}_{k,i}|} \exp(\alpha_{k,i}^p / t)$, and $t > 0$ is a temperature parameter. Then (10) enables us to compute $\partial \mathcal{L}_{\text{val}} / \partial \alpha_{k,i}$ approximately, and as shown by Proposition 1, when $t \rightarrow 0^+$, only the gradient with respect to the maximum weight $\alpha_{k,i}^{m^*}$ in (6) is non-zero and can further propagate backwards to $\lambda_{k,i}$ via (5). We provide the proof of Proposition 1 in Appendix A.

Proposition 1. Let $m^* = \arg \max_m \alpha_{k,i}^m$. Then, we have, i) $\lim_{t \rightarrow 0^+} \partial \mathcal{L}_{\text{val}} / \partial \alpha_{k,i}^m = 0$ for $m \neq m^*$; ii) $\lim_{t \rightarrow 0^+} \partial \mathcal{L}_{\text{val}} / \partial \alpha_{k,i}^{m^*} = \partial \mathcal{L}_{\text{val}} / \partial \bar{\alpha}_{k,i}^{m^*}$.

4.3.3 Complete Algorithm. The search algorithm for DiffMG based on the above discussion is described in Algorithm 1. Note that the softmax approximation in (10) is only employed for intermediate analysis, and we do not need it in the final algorithm by driving t to 0^+ . In the i -th iteration at the search stage, we sample one candidate message passing step, indicated by $\alpha_{k,i}^*$, for each mixed $\tilde{f}_{k,i}(\cdot)$:

$$\alpha_{k,i}^* = \begin{cases} \alpha_{k,i}^{m^*} & \text{with probability } 1 - \epsilon_i \\ \text{rand}(\{\alpha_{k,i}^m\}_{m=1}^{|\mathcal{A}_{k,i}|}) & \text{with probability } \epsilon_i \end{cases}, \quad (11)$$

where $\text{rand}(\cdot)$ means sampling one element from the given set randomly and uniformly. $\epsilon_i \in (0, 1)$ is a small parameter that encourages exploring different message passing options at the beginning and decreases to 0 as i increases (using a decay factor of 0.9 in our experiments). Forward and backward propagations in the neural network for updating both ω and λ only involve one sampled message passing step for each link in the DAG, which accelerates the training. After search, to derive the meta graph for evaluation, we pick the edge type $\mathcal{A}_{k,i}^{m^*}$, which is associated with the maximum weight $\alpha_{k,i}^{m^*}$ in (6), for each $\tilde{f}_{k,i}(\cdot)$ in the DAG search space. Then,

the message passing step determined by the derived meta graph is $\tilde{f}_{k,i}(H^{(i)}; \mathcal{A}_{k,i}) = f(H^{(i)}; \mathcal{A}_{k,i}^{m*})$.

Algorithm 1 Search algorithm for DiffMG

Require: step sizes $\eta_\omega, \eta_\lambda$, and edge types \mathcal{A} .

- 1: Initialize parameters ω and architecture parameters λ ;
 - 2: **for** the i -th iteration **do**
 - 3: Compute weights $\alpha_{k,i}$ as in (5) for each $\tilde{f}_{k,i}(\cdot)$;
 - 4: Sample one candidate message passing step for each $\tilde{f}_{k,i}(\cdot)$ as in (11). The collection of sampled $\alpha_{k,i}^*$'s is denoted as α^* ;
 - 5: $\omega \leftarrow \omega - \eta_\omega \frac{\partial \mathcal{L}_{\text{tra}}(\omega, \alpha^*)}{\partial \omega}$;
 - 6: $\lambda \leftarrow \lambda - \eta_\lambda \frac{\partial \mathcal{L}_{\text{val}}(\omega, \alpha^*)}{\partial \alpha^*} \frac{d\alpha^*}{d\lambda}$ (by Proposition 1 and (5));
 - 7: **end for**
 - 8: Derive a meta graph by selecting $\mathcal{A}_{k,i}^{m*}$ for each $\tilde{f}_{k,i}(\cdot)$;
 - 9: **return** the derived meta graph.
-

4.3.4 Comparison with Other One-shot NAS Algorithms. Table 2 summarizes the difference between our search algorithm and some popular one-shot NAS algorithms [17, 27, 30]. In Algorithm 1, evaluation of derivatives with respect to both parameters ω and architecture parameters λ only involves one candidate message passing step for each mixed operation, and thus speeds up the search process.

Table 2: Comparison of our search algorithm with existing one-shot NAS algorithms. ✓ and ✗ indicate whether update needs to compute all candidates of a mixed operation.

Method	Updating ω	Updating λ
DARTS [17]	✓	✓
SNAS [27]	✓	✓
NASP [30]	✗	✓
Ours	✗	✗

The most related one-shot NAS method to ours is NASP [30] which also transforms $\alpha_{k,i}$ into one-hot $\tilde{\alpha}_{k,i}$ as in (9). However, NASP performs proximal gradient descent which needs to evaluate $\partial \mathcal{L}_{\text{val}} / \partial \tilde{\alpha}_{k,i}$. Note that $\partial \mathcal{L}_{\text{val}} / \partial \tilde{\alpha}_{k,i}^m$ ($m \neq m^*$) is not zero even though the value of $\tilde{\alpha}_{k,i}^m$ ($m \neq m^*$) itself is zero. Therefore, NASP still needs to compute the candidate operation associated with $\tilde{\alpha}_{k,i}^m$ ($m \neq m^*$) to obtain $\partial \mathcal{L}_{\text{val}} / \partial \tilde{\alpha}_{k,i}^m$ ($m \neq m^*$). Different from NASP, we differentiate the transformation of (9) through a softmax approximation, in order to compute $\partial \mathcal{L}_{\text{val}} / \partial \alpha_{k,i}$ directly. The Gumbel-softmax trick [27] also uses softmax with a temperature parameter t to approximate arg max, however, in this paper we point out by Proposition 1 that when $t \rightarrow 0^+$ only the gradient with respect to $\alpha_{k,i}^{m^*}$ is non-zero, and thus avoid computing the operation associated with $\alpha_{k,i}^m$ ($m \neq m^*$). Besides, we do not need to design an annealing schedule for t like [27].

5 EXPERIMENTS

In this section, we demonstrate the effectiveness and efficiency of DiffMG through extensive experiments on different kinds of heterogeneous datasets. We want to address the following questions:

- How does DiffMG perform in the node classification task and the recommendation task compared with state-of-the-art baselines?
- How efficient is our proposed search algorithm compared against genetic search and popular one-shot NAS algorithms?
- How efficient is DiffMG compared against methods which can learn meta paths implicitly?
- What architectures are discovered by DiffMG over different real-world heterogeneous datasets?

5.1 Experimental Setup

5.1.1 Datasets. We evaluate DiffMG on two popular tasks [29]: node classification and recommendation. In the node classification task, we want to predict labels for nodes of a particular type, based on the network structure and node features. We use three real-world datasets: DBLP, ACM and IMDB. DBLP contains three types of nodes: papers (P), authors (A) and conferences (C), and authors are labeled by their research areas. ACM contains three types of nodes: papers (P), authors (A) and subjects (S), and papers are labeled by research areas. IMDB contains three types of nodes: movies (M), actors (A) and directors (D), and labels are genres of movies. Nodes in these datasets have bag-of-words representations as input features. We follow the splits provided by GTN [32]². Statistics of these datasets are summarized in Appendix B.

In the recommendation task, we want to predict links between source nodes (e.g., users) and target nodes (e.g., items). We consider three commonly used heterogeneous recommendation datasets³: Yelp, Douban movie (denoted as “Douban” in the sequel) and Amazon. Yelp is a platform where users review businesses. Douban is a social media community where users share reviews about movies. Amazon is a large e-commerce platform which contains users’ ratings for items. All the three datasets contain rich semantic relations, and their statistics are summarized in Appendix B. We convert ratings into binary class labels. Specifically, we randomly pick %50 of ratings which are higher than 3 as positive pairs, i.e., with label “1”, and all ratings which are lower than 4 serve as negative pairs, i.e., with label “0”. Positive pairs are then randomly split into a training set, a validation set and a test set according to a ratio of 3:1:1. We also randomly split negative pairs so that in each set the number of positive pairs and the number of negative pairs are the same. If the total number of negative pairs is less than the total number of positive pairs, we treat a source node and a target node which are unconnected in the original network as an additional negative pair, in order to balance two classes. To avoid the label leakage issue [35], both positive pairs and negative pairs are disconnected in the original network. Nodes in these three datasets are not associated with attributes, so we use one-hot IDs as input features.

5.1.2 Baselines. We compare against a random walk based network embedding method metapath2vec [2]⁴, two homogeneous GNNs: GCN [15] and GAT [24]⁵, two heterogeneous GNNs which

²https://github.com/seongjunyun/Graph_Transformer_Networks

³<https://github.com/librahu/HIN-Datasets-for-Recommendation-and-Network-Embedding>

⁴<https://ericdongyx.github.io/metapath2vec/m2v.html>

⁵We use implementations from PyTorch Geometric (PyG) [5] for GCN and GAT.

Table 3: Macro F1 scores (%) on the node classification task.

	metapath2vec	GCN	GAT	HAN	MAGNN	GTN	HGT	DiffMG
DBLP	89.93±0.45	90.46±0.41	93.92±0.28	92.13±0.26	92.81±0.30	93.98±0.32	93.67±0.22	94.45±0.15
ACM	67.13±0.50	92.56±0.20	92.50±0.23	91.20±0.25	91.15±0.19	92.62±0.17	91.83±0.23	92.65±0.15
IMDB	40.82±1.48	55.19±0.99	53.37±1.27	55.09±0.67	56.44±0.63	59.68±0.72	59.35±0.79	61.04±0.56

Table 4: AUC (%) on the recommendation task.

	metapath2vec	GCN	GAT	HAN	MAGNN	GEMS	GTN	HGT	DiffMG
Amazon	58.17±0.14	66.64±1.00	55.70±1.13	67.35±0.11	68.26±0.09	70.66±0.14	71.82±0.18	74.75±0.08	75.28±0.08
Yelp	51.98±0.14	58.98±0.52	56.55±0.05	64.28±0.20	64.73±0.24	65.12±0.27	66.27±0.31	68.07±0.35	68.77±0.13
Douban	51.60±0.07	77.95±0.05	77.58±0.33	82.65±0.08	82.44±0.17	83.00±0.05	83.26±0.10	83.38±0.06	83.78±0.09

require manually designed meta paths: HAN [26]⁶ and MAGNN [7]⁷, and two heterogeneous GNNs which use attention to implicitly learn meta paths: GTN [32]⁸ and HGT [12]⁹. In the recommendation task, we also compare against GEMS [11]¹⁰ which utilizes the evolutionary algorithm to search for meta graphs between source node type and target node type. All heterogeneous GNN baselines and our own model are also summarized in Table 1. Except metapath2vec, all methods are implemented using PyTorch [18].

5.1.3 Evaluation Metrics. In the node classification task, we use macro F1 score as the evaluation metric; in the recommendation task, we use AUC (area under the ROC curve) as the evaluation metric.

5.1.4 Hyper-parameters. Due to randomness of initialization and sampling as in (11), for each dataset we run the search algorithm three times with different random seeds and then derive the meta graph for final evaluation from the run which achieves the best validation performance. In the node classification task, we run the search algorithm for 50 epochs, while in the recommendation task, we run it for 100 epochs. At the search stage, we train λ using Adam with a learning rate of $3e-4$, and we train ω using Adam with a learning rate of 0.005 and weight decay of 0.001. We set $K = 4$ for all datasets, which is the same as the length of meta paths learned by GTN [32]. Since meta paths and meta graphs, either manually designed [7, 26] or learned [11, 32], enlarge GNN’s receptive field, we use 4 layers for GCN, GAT and HGT to keep a fair comparison. The hidden dimension is set to 64 for all baselines and our model. The number of attention heads is 8 for GAT, HAN, MAGNN and HGT, so the hidden dimension for each head is 8. The dimension of semantic-level attention vector for HAN [26] and MAGNN [7] is 128. All GNN methods are trained using full batch, for 100 epochs in the node classification task with early stopping, and for 200 epochs in the recommendation task. The optimizer is Adam and other hyper-parameters (learning rate, weight decay and input dropout) are tuned according to the validation performance. For metapath2vec, we follow its default hyper-parameter setting. For GEMS, we run the evolutionary algorithm for 100 generations

with a population size of 20. Experiments are conducted on a single RTX 2080 Ti GPU with 11GB memory, except that we parallelize GEMS on more than one GPU to speed up the genetic search.

5.2 Node Classification

Table 3 shows the macro F1 scores of different methods on the node classification task. For each method, we report the average score and standard deviation of 10 runs with different random seeds. First, GAT yields highly competitive performance on DBLP and ACM compared with heterogeneous GNN baselines. We conjecture this is because GAT can exploit semantic information by learning to assign different importance to different node types. Second, HAN and MAGNN which rely on hand-designed meta paths do not obtain desirable performance compared with GTN, HGT and DiffMG, and they can even perform worse than homogeneous GNNs. This indicates that hand-designed rules are limited in mining task-dependent semantic information and may even cause adverse effects. Finally, DiffMG consistently achieves the best performance on all three datasets, which demonstrates the importance of automatically utilizing task-dependent semantic information in HINs. Specifically, the performance improvement over GTN further validates that meta graphs are more powerful to capture semantic information than meta paths.

5.3 Recommendation

In Table 4, we report the AUC of DiffMG and baselines. The results are the average of 10 runs with different random seeds. The HINs used for the recommendation task are larger and have richer semantic relations than those used for the node classification task, and we observe that the performance of homogeneous GNNs is poor compared with heterogeneous GNNs which are tailored to utilize semantic information. HAN and MAGNN show performance improvement over homogeneous GNNs by leveraging meta paths. GEMS outperforms HAN and MAGNN by searching for a meta graph to capture semantic proximity between source nodes and target nodes, but it is too inefficient to fully explore the search space. Moreover, GEMS ignores representations of intermediate nodes along the meta graph, which limits its capacity. Again, our method is the best on all three datasets and exhibits significant improvement over the most competitive baseline HGT, which demonstrates that a task-dependent meta graph is superior to a much more complex architecture in representation learning on HINs.

⁶We implement with PyG following the authors’ TensorFlow code at <https://github.com/Jhy1993/HAN>.

⁷<https://github.com/cynricfu/MAGNN>

⁸https://github.com/seongjunyun/Graph_Transformer_Networks

⁹<https://github.com/acbull/pyHGT>

¹⁰<https://github.com/00showero0/GEMS>

5.4 Efficiency Analysis

5.4.1 Search Stage. In Table 5, we compare the search cost of our search algorithm (Algorithm 1), measured in GPU minutes, against genetic search and other popular one-shot NAS algorithms (Table 2), on the recommendation datasets. For one-shot search algorithms, we report the total search cost of three runs. We also report the evaluation cost which requires training the searched model from scratch once. The genetic search algorithm evaluates each candidate model individually, so it is extremely time consuming. On the contrary, one-shot search algorithms combine candidate models in a super model and reduce the search cost by dozens of times. Compared with existing one-shot NAS algorithms [17, 27, 30], our search algorithm significantly improves the search efficiency, so that the total search cost is on a par with the evaluation cost. The efficient search algorithm enables our framework DiffMG to be applied to large scale heterogeneous datasets.

Table 5: Search cost compared against evaluation cost, measured in GPU minutes. For one-shot NAS methods, we also report in parentheses the time required to update λ .

		Amazon	Yelp	Douban
Genetic	GEMS [11]	800	1500	2000
One-shot	DARTS [17]	8.9 (6.9)	15.8 (12.2)	19.4 (15.0)
	SNAS [27]	3.9 (1.9)	7.0 (3.4)	8.6 (4.2)
	NASP [30]	2.3 (1.9)	4.0 (3.4)	5.0 (4.2)
	Ours	0.7 (0.3)	1.1 (0.5)	1.6 (0.8)
Evaluation		0.5	0.7	1.0

5.4.2 Evaluation Stage. We study the efficiency of DiffMG at the evaluation stage, where the derived model is retrained from scratch, on two large heterogeneous datasets: Yelp and Douban. We compare against HGT [12] which can implicitly learn meta paths by attention and outperforms the other baselines in Table 4. Another strong baseline GTN [32] cannot fit into a single GPU on these datasets and requires much more time to train on CPU (nearly 20 minutes per epoch on Yelp), so we do not compare against it here. We also compare against the simplest architecture GCN to show the trade off between performance and efficiency.

In Figure 2, we plot the validation AUC with respect to GPU training time (measured in seconds) for these methods. They are trained for 200 epochs three times. We observe that DiffMG reaches its highest validation performance much faster than HGT. Compared with GCN, DiffMG gains large performance improvement while taking only a mild amount of additional time to finish training. Furthermore, the search cost of DiffMG is on a par with training the derived model once (see Section 5.4.1), which demonstrates that DiffMG is well applicable to large scale HINs.

5.5 Visualization

We visualize the architectures discovered by DiffMG on DBLP and Douban in Figure 3. In Figure 3(a), we observe that both conference nodes (C) and paper nodes (P) are involved in generating representations for authors (A), which is similar to hand-designed meta paths for academic networks [2]. Moreover, we observe that

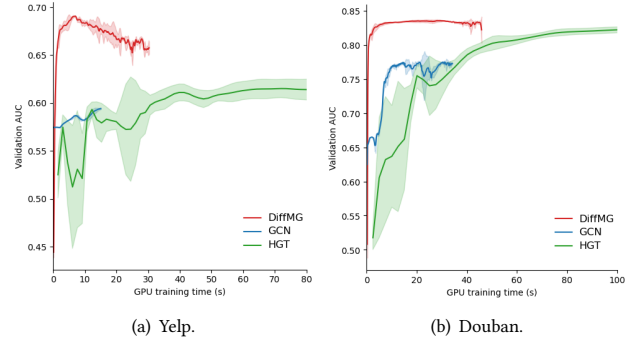


Figure 2: Efficiency of DiffMG at the evaluation stage. We do not compare against GTN because of OOM on a single GPU.

$H^{(3)}$ and $H^{(4)}$ have more than one incoming link, which combines multiple propagation paths together. This cannot be done by GTN [32] which only learns meta paths via matrix multiplication. In Figure 3(b), we show both the meta graph (above) whose target node type is user (U) and the meta graph (below) whose target node type is movie (M). DiffMG is able to obtain different meta graphs to utilize semantic information for different target node types. In our framework, each edge type in the derived meta graph corresponds to message passing along edges of this type in the underlying HIN, which enables DiffMG to encode information of intermediate nodes along a meta graph. On the contrary, GEMS [11] only propagates information between source nodes and target nodes, and ignores intermediate nodes along the meta graph.

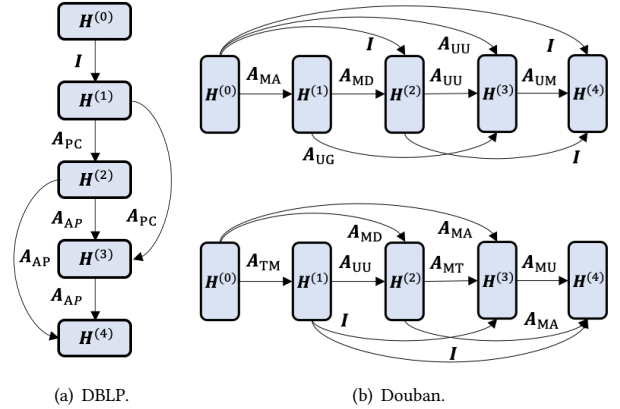


Figure 3: Architectures discovered by DiffMG.

5.6 Ablation Study

In Table 6, we study the influence of ϵ_0 on the final performance. Search and evaluation procedures for Table 6 are the same as Section 5.3. We observe that large ϵ_0 does not improve the performance except that $\epsilon_0 = 0.5$ works better on Amazon than $\epsilon_0 = 0$. We also show the performance of single-level optimization which optimizes both ω and λ with the union of training set and validation set, and

it matches bi-level optimization. To validate the second motivation of our proposed search algorithm, we study a variant which optimizes all architecture parameters of a mixed operation simultaneously, like DARTS [17]. Its performance is worse than our search algorithm.

Table 6: Ablation study results (%).

	Amazon	Yelp	Douban
DiffMG ($\epsilon_0 = 0$)	74.80 \pm 0.09	68.77 \pm 0.13	83.78 \pm 0.09
DiffMG ($\epsilon_0 = 0.3$)	74.51 \pm 0.17	68.42 \pm 0.09	83.58 \pm 0.08
DiffMG ($\epsilon_0 = 0.5$)	75.28 \pm 0.08	67.90 \pm 0.15	83.45 \pm 0.09
DiffMG (single-level)	74.80 \pm 0.09	68.94 \pm 0.12	83.70 \pm 0.03
DiffMG (DARTS)	74.54 \pm 0.22	66.72 \pm 0.20	83.68 \pm 0.05

6 CONCLUSION

In this work, we explore the direction of addressing representation learning on HINs with automated machine learning. We present key contributions on both the design of the search space and the search algorithm. For future work, we will study the convergence of our proposed search algorithm.

REFERENCES

- [1] Gabriel Bender, Pieter-Jan Kindermans, Barret Zoph, Vijay Vasudevan, and Quoc Le. 2018. Understanding and simplifying one-shot architecture search. In *ICML*. 550–559.
- [2] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *SIGKDD*. 135–144.
- [3] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. 2019. Neural Architecture Search: A Survey. *JMLR* 20, 55 (2019), 1–21.
- [4] Yuan Fang, Wenqing Lin, Vincent W Zheng, Min Wu, Kevin Chen-Chuan Chang, and Xiao-Li Li. 2016. Semantic proximity search on graphs with metagraph-based learning. In *ICDE*. 277–288.
- [5] Matthias Fey and Jan E. Lenssen. 2019. Fast Graph Representation Learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.
- [6] Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazi, and Massimiliano Pontil. 2018. Bilevel Programming for Hyperparameter Optimization and Meta-Learning. In *ICML*. 1568–1577.
- [7] Xinyu Fu, Jiani Zhang, Ziqiao Meng, and Irwin King. 2020. MAGNN: Metapath Aggregated Graph Neural Network for Heterogeneous Graph Embedding. In *WebConf*. 2331–2341.
- [8] Yang Gao, Hong Yang, Peng Zhang, Chuan Zhou, and Yue Hu. 2020. Graph Neural Architecture Search. In *IJCAI*. 1403–1409.
- [9] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. In *ICML*. 1263–1272.
- [10] William L Hamilton, Zitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *NeurIPS*.
- [11] Zhenyu Han, Fengli Xu, Jinghan Shi, Yu Shang, Haorui Ma, Pan Hui, and Yong Li. 2020. Genetic Meta-Structure Search for Recommendation on Heterogeneous Information Network. In *CIKM*. 455–464.
- [12] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. Heterogeneous graph transformer. In *WebConf*. 2704–2710.
- [13] Zhipeng Huang, Yudian Zheng, Reynold Cheng, Yizhou Sun, Nikos Mamoulis, and Xiang Li. 2016. Meta structure: Computing relevance in large heterogeneous information networks. In *SIGKDD*. 1595–1604.
- [14] Jiarui Jin, Jiarui Qin, Yuchen Fang, Kounianhua Du, Weinan Zhang, Yong Yu, Zheng Zhang, and Alexander J Smola. 2020. An Efficient Neighborhood-based Interaction Model for Recommendation on Heterogeneous Graph. In *SIGKDD*. 75–84.
- [15] Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.
- [16] Kwei-Herng Lai, Daochen Zha, Kaixiong Zhou, and Xia Hu. 2020. Policy-GNN: Aggregation Optimization for Graph Neural Networks. In *SIGKDD*. 461–471.
- [17] Hanxiao Liu, Karen Simonyan, and Yiming Yang. 2019. DARTS: Differentiable Architecture Search. In *ICLR*.
- [18] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS*.
- [19] Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, and Jeff Dean. 2018. Efficient Neural Architecture Search via Parameters Sharing. In *ICML*. 4095–4104.
- [20] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V Le. 2019. Regularized evolution for image classifier architecture search. In *AAAI*, Vol. 33. 4780–4789.
- [21] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S Yu, and Tianyi Wu. 2011. PathsSim: Meta path-based top-k similarity search in heterogeneous information networks. *PVLDB* 4, 11 (2011), 992–1003.
- [22] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *ICLR*.
- [23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NeurIPS*.
- [24] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. Graph attention networks. In *ICLR*.
- [25] Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies* 1, 1 (2020), 396–413.
- [26] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. 2019. Heterogeneous graph attention network. In *WebConf*. 2022–2032.
- [27] Sirui Xie, Hehui Zheng, Chunxiao Liu, and Liang Lin. 2019. SNAS: stochastic neural architecture search. In *ICLR*.
- [28] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2019. How Powerful are Graph Neural Networks?. In *ICLR*.
- [29] Carl Yang, Yuxin Xiao, Yu Zhang, Yizhou Sun, and Jiawei Han. 2020. Heterogeneous Network Representation Learning: A Unified Framework with Survey and Benchmark. *IEEE Transactions on Knowledge and Data Engineering* (2020).
- [30] Quanming Yao, Ju Xu, Wei-Wei Tu, and Zhanxing Zhu. 2020. Efficient Neural Architecture Search via Proximal Iterations. In *AAAI*.
- [31] Jiaxuan You, Zitao Ying, and Jure Leskovec. 2020. Design Space for Graph Neural Networks. In *NeurIPS*.
- [32] Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. 2019. Graph Transformer Networks. In *NeurIPS*.
- [33] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. 2019. Heterogeneous graph neural network. In *SIGKDD*. 793–803.
- [34] Fanjin Zhang, Xiao Liu, Jie Tang, Yuxiao Dong, Peiran Yao, Jie Zhang, Xiaotao Gu, Yan Wang, Bin Shao, Rui Li, et al. 2019. Oag: Toward linking large-scale heterogeneous entity graphs. In *SIGKDD*. 2585–2595.
- [35] Jiani Zhang, Xingjian Shi, Shenglin Zhao, and Irwin King. 2019. STAR-GCN: Stacked and Reconstructed Graph Convolutional Networks for Recommender Systems. In *IJCAI*. 4264–4270.
- [36] Muhan Zhang and Yixin Chen. 2018. Link prediction based on graph neural networks. In *NeurIPS*.
- [37] Wentao Zhang, Yuan Fang, Zemin Liu, Min Wu, and Xinming Zhang. 2020. mg2vec: Learning relationship-preserving heterogeneous graph representations via metagraph embedding. *IEEE Transactions on Knowledge and Data Engineering* (2020).
- [38] Yizhou Zhang, Yun Xiong, Xiangnan Kong, Shanshan Li, Jinhong Mi, and Yangyong Zhu. 2018. Deep collective classification in heterogeneous information networks. In *WebConf*. 399–408.
- [39] Huan Zhao, Quanming Yao, Jianda Li, Yangqiu Song, and Dik Lun Lee. 2017. Metagraph based recommendation fusion over heterogeneous information networks. In *SIGKDD*. 635–644.
- [40] Huan Zhao, Quanming Yao, and Weiwei Tu. 2021. Search to aggregate neighborhood for graph neural network. In *ICDE*.
- [41] Marinka Zitnik, Francis Nguyen, Bo Wang, Jure Leskovec, Anna Goldenberg, and Michael M Hoffman. 2019. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion* 50 (2019), 71–91.
- [42] Barret Zoph and Quoc Le. 2017. Neural architecture search with reinforcement learning. In *ICLR*.
- [43] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. 2018. Learning transferable architectures for scalable image recognition. In *CVPR*. 8697–8710.

A PROOF

We provide the proof of Proposition 1 as follows. According to the chain rule, we have:

$$\frac{\partial \mathcal{L}_{\text{val}}}{\partial \alpha_{k,i}^m} = \frac{\partial \mathcal{L}_{\text{val}}}{\partial \mathbf{H}^{(k)}} \frac{\partial \mathbf{H}^{(k)}}{\partial \alpha_{k,i}^m} = \frac{\partial \mathcal{L}_{\text{val}}}{\partial \mathbf{H}^{(k)}} \sum_{q=1}^{|\mathcal{A}_{k,i}|} \frac{\partial \tilde{f}_{k,i}}{\partial \alpha_{k,i}^q} \frac{\partial \tilde{\alpha}_{k,i}^q}{\partial \alpha_{k,i}^m}. \quad (12)$$

Recall that in (10) we use a softmax function to approximate the discrete transformation of (9):

$$\tilde{\alpha}_{k,i}^m \approx \alpha_{k,i}^m \cdot h(m; t), \quad (13)$$

where $h(m; t) = \exp(\alpha_{k,i}^m/t) / \sum_{p=1}^{|\mathcal{A}_{k,i}|} \exp(\alpha_{k,i}^p/t)$, and $t > 0$ is a temperature parameter. Based on (13), we have:

$$\frac{\partial \tilde{\alpha}_{k,i}^q}{\partial \alpha_{k,i}^m} = \delta_{qm} h(q; t) + \alpha_{k,i}^q \frac{\partial h(q; t)}{\partial \alpha_{k,i}^m}, \quad (14)$$

where if q equals m , $\delta_{qm} = 1$, otherwise $\delta_{qm} = 0$. Combining (12) and (14), and note that $\lim_{t \rightarrow 0^+} \partial h(q; t) / \partial \alpha_{k,i}^m = 0$, then we get:

$$\lim_{t \rightarrow 0^+} \frac{\partial \mathcal{L}_{\text{val}}}{\partial \alpha_{k,i}^m} = \lim_{t \rightarrow 0^+} \frac{\partial \mathcal{L}_{\text{val}}}{\partial \mathbf{H}^{(k)}} \frac{\partial \tilde{f}_{k,i}}{\partial \tilde{\alpha}_{k,i}^m} h(m; t) \quad (15)$$

$$= \lim_{t \rightarrow 0^+} \frac{\partial \mathcal{L}_{\text{val}}}{\partial \tilde{\alpha}_{k,i}^m} h(m; t). \quad (16)$$

Since $\lim_{t \rightarrow 0^+} h(m^*; t) = 1$, and $\lim_{t \rightarrow 0^+} h(m; t) = 0$ for $m \neq m^*$, we finish the proof.

B STATISTICS OF DATASETS

Table 8: Statistics of HINs for node classification.

	DBLP	ACM	IMDB
# Nodes	18405	8994	12624
# Edges	67946	25922	37288
# Classes	4	3	3
# Edge types	4	4	4
# Training	800	600	300
# Validation	400	300	300
# Testing	2857	2125	2339

Table 7: Statistics of HINs for recommendation.

Dataset	Relations (A-B)	# A	# B	# A-B
Yelp	User-Business (U-B)	16239	14284	198397
	User-User (U-U)	16239	16239	158590
	User-Compliment (U-Co)	16239	11	76875
	Business-City (B-C)	14284	47	14267
	Business-Category (B-Ca)	14284	511	40009
Douban movie	User-Movie (U-M)	13367	12677	1068278
	User-Group (U-G)	13367	2753	570047
	User-User (U-U)	13367	13367	4085
	Movie-Actor (M-A)	12677	6311	33587
	Movie-Director (M-D)	12677	2449	11276
	Movie-Type (M-T)	12677	38	27668
Amazon	User-Item (U-I)	6170	2753	195791
	Item-View (I-V)	2753	3857	5694
	Item-Category (I-C)	2753	22	5508
	Item-Brand (I-B)	2753	334	2753