

Universal Graph Transformer Self-Attention Networks

Dai Quoc Nguyen*
Oracle Labs
Australia
dai.nguyen@oracle.com

Tu Dinh Nguyen
VinAI Research
Vietnam
v.tund21@vinai.io

Dinh Phung
Monash University
Australia
dinh.phung@monash.edu

ABSTRACT

We introduce a transformer-based GNN model, named UGformer, to learn graph representations. In particular, we present two UGformer variants, wherein the first variant (publicized in September 2019) is to leverage the transformer on a set of sampled neighbors for each input node, while the second (publicized in May 2021) is to leverage the transformer on all input nodes. Experimental results demonstrate that the first UGformer variant achieves state-of-the-art accuracies on benchmark datasets for graph classification in both inductive setting and unsupervised transductive setting; and the second UGformer variant obtains state-of-the-art accuracies for inductive text classification. The code is available at: <https://github.com/daiquocnguyen/Graph-Transformer>.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing; Neural networks**; • **Information systems** → **Social networks**.

KEYWORDS

graph neural networks, graph classification, inductive text classification, graph transformer, unsupervised transductive learning

ACM Reference Format:

Dai Quoc Nguyen, Tu Dinh Nguyen, and Dinh Phung. 2022. Universal Graph Transformer Self-Attention Networks. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion)*, April 25–29, 2022, Virtual Event, Lyon, France. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3487553.3524258>

1 INTRODUCTION

A graph is a connected network of nodes and edges. This type of graph-structured data is a fundamental mathematical representation and ubiquitous. They found applications in virtually all aspects of our daily lives from pandemic outburst response, internet of things, drug discovery to circuit design, to name a few. In machine learning and data science, learning and inference from graphs have been one of the most trending research topics. However, as data grow unprecedentedly in volume and complexity in modern time,

*This work was done before Dai Quoc Nguyen joined Oracle Labs, Australia. UGformer Variant 1 was proposed when Dai Quoc Nguyen was a PhD student at Monash University, Australia.

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.

WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France

© 2022 Association for Computing Machinery.

ACM ISBN 978-1-4503-9130-6/22/04...\$15.00

<https://doi.org/10.1145/3487553.3524258>

traditional learning methods for graph are mostly inadequate to model increasing complexity, to harness rich contextual information as well as to scale with large-scale graphs. The recent rise of deep learning, and in turn, of representation learning field has radically advanced machine learning research in general, and pushing the frontier of graph learning. In particular, the notion of graph representation learning has recently emerged as a new promising learning paradigm, which aims to learn a parametric mapping function that embeds nodes, subgraphs, or the entire graph into low-dimensional continuous vector spaces [6, 15, 26]. The central challenge to this endeavor is to learn rich classes of complex functions to capture and preserve the graph structural information as much as possible and also be able to geometrically represent the structural information in the embedded space.

Recently, graph neural networks (GNNs) become an essential strand, forming the third direction to learn low-dimensional continuous representations for nodes and graphs [6, 21, 26]. In general, GNNs use an aggregation function to update the vector representation of each node by transforming and aggregating the vector representations of its neighbours [5, 10, 16, 24]. Then GNNs apply a graph-level readout function such as simple sum pooling to obtain graph embeddings [4, 25, 28, 32, 33]. GNN-based approaches provide faster and practical training and state-of-the-art results on benchmark datasets for downstream tasks such as node classification [10], graph classification [28], knowledge graph completion [17], vulnerability detection [18], and text classification [31]. To further improve the performance, it is worth developing advanced GNNs to better update vector representations of nodes from their neighbours. Currently, there are novel applications of the transformer [23] recognized, published, and used successfully in natural language processing. Inspired by this fact, we consider the use of the transformer to a new domain such as GNNs as a novelty and present UGformer, a transformer-based GNN model, to learn graph representations. Our main contributions in this paper are as follows:

- We propose a transformer-based GNN model, named UGformer, to learn graph representations. In particular, we consider two model variants of (i) leveraging the transformer on a set of sampled neighbors for each input node and (ii) leveraging the transformer on all input nodes. *We publicized the first UGformer variant in September 2019 and the second UGformer variant in May 2021.*

- The unsupervised learning is essential in both industry and academic applications, where expanding unsupervised GNN models is more suitable to address the limited availability of class labels. Thus, we present an *unsupervised transductive* learning approach to train GNNs.

- Experimental results show that the first UGformer variant obtains state-of-the-art accuracies on social network and bioinformatics datasets for graph classification in both inductive setting and unsupervised transductive setting; and the second UGformer

variant produces state-of-the-art accuracies on benchmark datasets for inductive text classification.

2 THE PROPOSED UGFORMER

2.1 Variant 1: Leveraging the transformer on a set of sampled neighbors for each node

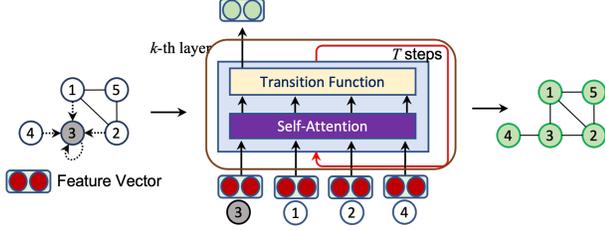


Figure 1: An illustration of UGformer Variant 1.

We publicized the first UGformer variant in September 2019. This variant is also suitable for large graphs. Given an input graph \mathcal{G} , we uniformly sample a set \mathcal{N}_v of neighbors for each $v \in \mathcal{V}$ and then input $\mathcal{N}_v \cup \{v\}$ to the UGformer learning process, as illustrated in Figure 1. Note that we sample a different \mathcal{N}_v for node v at each training batch. We utilize the transformer to construct multiple layers stacked on top of each other, wherein the weight matrices are shared in both the self-attention and transition functions across all positions and timesteps. In particular, regarding the k -th layer, given a node $v \in \mathcal{V}$, at each step t , we induce a transformer-based function to aggregate the vector representations for all nodes $u \in \mathcal{N}_v \cup \{v\}$ as follows:

$$\mathbf{x}_{t,u}^{(k)} = \text{LAYERNORMALIZATION} \left(\mathbf{h}_{t-1,u}^{(k)} + \text{ATT} \left(\mathbf{h}_{t-1,u}^{(k)} \right) \right) \quad (1)$$

$$\mathbf{h}_{t,u}^{(k)} = \text{LAYERNORMALIZATION} \left(\mathbf{x}_{t,u}^{(k)} + \text{TRANS} \left(\mathbf{x}_{t,u}^{(k)} \right) \right) \quad (2)$$

where $\mathbf{h}_{0,v}^{(0)} = \mathbf{h}_v^{(0)}$ is the feature vector of v ; $\text{TRANS}(\cdot)$ and $\text{ATT}(\cdot)$ denote a MLP network and a self-attention layer respectively. In particular, we have:

$$\text{ATT} \left(\mathbf{h}_{t-1,u}^{(k)} \right) = \sum_{u' \in \mathcal{N}_v \cup \{v\}} \alpha_{u,u'}^{(k)} \left(\mathbf{V}^{(k)} \mathbf{h}_{t-1,u'}^{(k)} \right) \quad (3)$$

where $\mathbf{V}^{(k)} \in \mathbb{R}^{d \times d}$ is a value-projection weight matrix; $\alpha_{u,u'}$ is an attention weight, which is computed using the softmax function over scaled dot products between nodes u and u' :

$$\alpha_{u,u'}^{(k)} = \text{softmax} \left(\frac{\left(\mathbf{Q}^{(k)} \mathbf{h}_{t-1,u}^{(k)} \right)^T \left(\mathbf{K}^{(k)} \mathbf{h}_{t-1,u'}^{(k)} \right)}{\sqrt{d}} \right) \quad (4)$$

where $\mathbf{Q}^{(k)} \in \mathbb{R}^{d \times d}$ and $\mathbf{K}^{(k)} \in \mathbb{R}^{d \times d}$ are query-projection and key-projection matrices, respectively.

$$\mathbf{H}_t^{(k)} = \text{ATTENTION}_{\mathcal{N}_v \cup \{v\}} \left(\mathbf{H}_{t-1}^{(k)} \mathbf{Q}^{(k)}, \mathbf{H}_{t-1}^{(k)} \mathbf{K}^{(k)}, \mathbf{H}_{t-1}^{(k)} \mathbf{V}^{(k)} \right) \quad (5)$$

After T steps, we feed $\mathbf{h}_{T,v}^{(k)} \in \mathbb{R}^d$ to the next $(k+1)$ -th layer as:

$$\mathbf{h}_v^{(k+1)} = \mathbf{h}_{0,v}^{(k+1)} = \mathbf{h}_{T,v}^{(k)}, \forall v \in \mathcal{V} \quad (6)$$

If we do not share the weight matrices in both the self-attention and transition functions across all positions and timesteps, T becomes the number of self-attention layers within each UGformer layer.

2.2 Variant 2: Leveraging the transformer on all input nodes

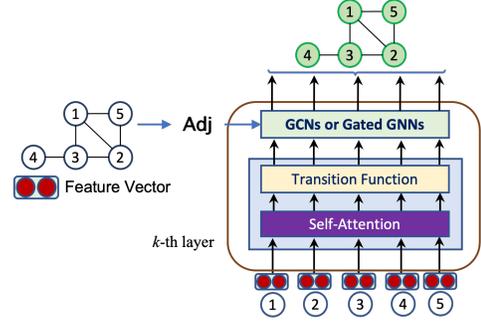


Figure 2: An illustration of UGformer Variant 2.

We publicized the second UGformer variant in May 2021, when we aimed to leverage the transformer for small and medium graphs. It is worth noting that all links/interactions among all positions in the self-attention layer build up a complete network. Hence, if we apply *only* the self-attention mechanism to all nodes of the input graph \mathcal{G} , we ignore the structure of \mathcal{G} . To overcome this limitation, we propose that each UGformer layer consists of a transformer self-attention network followed by a GNN layer such as GCNs [10] or Gated GNNs [12]. This model variant combines the transformer with the graph structure to learn better graph representations. Formally, we define a UGformer layer as illustrated in Figure 2 as:

$$\mathbf{H}'^{(k)} = \text{ATTENTION}_{\mathcal{V}} \left(\mathbf{H}^{(k)} \mathbf{Q}^{(k)}, \mathbf{H}^{(k)} \mathbf{K}^{(k)}, \mathbf{H}^{(k)} \mathbf{V}^{(k)} \right) \quad (7)$$

$$\mathbf{H}^{(k+1)} = \text{GNN} \left(\mathbf{A}, \mathbf{H}'^{(k)} \right) \quad (8)$$

where $\mathbf{H}^{(0)}$ is the feature matrix of all nodes in \mathcal{G} , and \mathbf{A} is the adjacency matrix. Equation 7 is with respect to Equations 1 and 2 when $T = 1$.

3 DEMONSTRATION OF UGFORMER

3.1 UGformer Variant 2 for inductive text classification

We demonstrate the advantage of UGformer for inductive text classification. Firstly, we follow [16, 34] to build a graph \mathcal{G} for each textual document by representing unique words as nodes and co-occurrences between words (within a fixed-size sliding window of length 3) as edges. We then employ our Variant 2, wherein we adapt a Gated GNN layer [12] for Equation 8 as the Gated GNN layer is more suitable for our data. After that, to produce the graph embedding $\mathbf{e}_{\mathcal{G}}$, we follow [18] to construct a graph-level readout function as:

$$\mathbf{e}_v = \sigma \left(\mathbf{w}^T \mathbf{h}_v^{(K)} + b \right) \odot g \left(\mathbf{W}_1 \mathbf{h}_v^{(K)} + \mathbf{b}_1 \right) \quad (9)$$

$$\mathbf{e}_{\mathcal{G}} = \sum_{v \in \mathcal{V}} \mathbf{e}_v \odot \max_pooling \{ \mathbf{e}_v \}_{v \in \mathcal{V}} \quad (10)$$

where $\mathbf{h}_v^{(K)}$ is the vector representation of node v at the last K -th layer; and $\sigma(\mathbf{w}^T \mathbf{h}_v^{(K)})$ acts as soft attention mechanisms over nodes. Finally, we also feed \mathbf{e}_G to a single fully-connected layer followed by a softmax layer to perform the text classification task.

3.1.1 Experimental setup. We follow [16, 31, 34] to use four benchmarks – MR, R8, R52, and Ohsumed. We follow [16, 34] to use random vectors or pre-trained Glove [20] with the dimension size of 300 to initialize the feature vectors. We also follow [16, 31, 34] to build a 2-layer model. We set the number of attention heads to 2 and the hidden size to 384. We vary the learning rate in $\{1e^{-4}, 5e^{-4}, 1e^{-3}, 5e^{-3}\}$. Then, we utilize the Adam optimizer to train the model up to 150 epochs to evaluate our trained model. We also run 10 times and report the mean accuracy and standard deviation, wherein for each time, we randomly split 10% text from the training set to have the validation set for hyper-parameter turning.

Table 1: Text classification accuracies (%) on the test sets w.r.t UGformer Variant 2. Some baseline results are taken from [31]. Note that the original results of TextING [34] are reported for the best model on the test set; thus, we consider the results of TextING taken from [16] for a fair comparison. “w/o GNN” means that we do not use the GNN layer within the UGformer layer.

Dataset	MR	R8	R52	Ohsumed
Bi-LSTM	77.68 ± 0.86	96.31 ± 0.33	90.54 ± 0.91	49.27 ± 1.07
fastText	75.14 ± 0.20	96.13 ± 0.21	92.81 ± 0.09	57.70 ± 0.49
TextGCN	76.74 ± 0.20	97.07 ± 0.10	93.56 ± 0.18	68.36 ± 0.56
TextING	78.86 ± 0.26	96.90 ± 0.23	93.34 ± 0.24	69.72 ± 0.30
TextQGNN	78.93 ± 0.29	97.02 ± 0.28	94.45 ± 0.35	69.93 ± 0.31
UGformer	79.29 ± 0.46	97.05 ± 0.32	94.71 ± 0.42	70.63 ± 0.18
w/o GNN	77.47 ± 0.51	96.50 ± 0.42	93.85 ± 0.08	68.27 ± 0.52

3.1.2 Main results. Table 1 presents the classification accuracy results of our UGformer and the baseline models. In general, our UGformer outperforms the baseline models, produces state-of-the-art accuracies on three benchmark datasets R52, OHSUMED, and MR, and obtains a highly competitive accuracy on R8. In Table 1, we also compute and report our ablation results for using only the transformer w.r.t Equation 7, i.e., *not* using the GNN layer within the UGformer layer w.r.t Equation 8; hence we have:

$$\mathbf{H}^{(k+1)} = \text{ATTENTION}_{\mathcal{V}} \left(\mathbf{H}^{(k)} \mathbf{Q}^{(k)}, \mathbf{H}^{(k)} \mathbf{K}^{(k)}, \mathbf{H}^{(k)} \mathbf{V}^{(k)} \right)$$

It is worth noting that the scores degrade, thus showing the effectiveness of our proposed UGformer in integrating the transformer with the graph structure to learn better graph representations.

3.2 UGformer Variant 1 for graph classification in an inductive setting

Following [28, 29], we apply the vector concatenation across the layers to obtain the vector representations \mathbf{e}_v of nodes v as:

$$\mathbf{e}_v = \left[\mathbf{h}_v^{(1)} \parallel \mathbf{h}_v^{(2)} \parallel \dots \parallel \mathbf{h}_v^{(K)} \right], \forall v \in \mathcal{V} \quad (11)$$

where K is the number of layers. After that, we also follow [28] to sum all the final embeddings of nodes in \mathcal{G} to get the final

embedding \mathbf{e}_G of the entire graph \mathcal{G} . We feed \mathbf{e}_G to a single fully-connected layer followed by a softmax layer to predict the graph label as:

$$\hat{y}_G = \text{softmax}(\mathbf{W}\mathbf{e}_G + \mathbf{b}) \quad (12)$$

Finally, we learn the model parameters by minimizing the cross-entropy loss function.

3.2.1 Experimental setup. We use seven well-known datasets consisting of three social network datasets (COLLAB, IMDB-B, and IMDB-M) and four bioinformatics datasets (DD, MUTAG, PROTEINS, and PTC). The social network datasets do not have node features; thus, we follow [19, 33] to use node degrees as features. We vary the number K of UGformer layers in $\{1, 2, 3\}$, the number of steps T in $\{1, 2, 3, 4\}$, the number of neighbors ($|\mathcal{N}_v| = N$) sampled for each node in $\{4, 8, 16\}$, and the hidden size in TRANS(.) in $\{128, 256, 512, 1024\}$. We set the batch size to 4. We apply the Adam optimizer [9] to train our UGformer and select the Adam initial learning rate $lr \in \{5e^{-5}, 1e^{-4}, 5e^{-4}, 1e^{-3}\}$. We run up to 50 epochs to evaluate our UGformer. We follow [1, 13, 22, 27, 28] to use the same data splits and the same 10-fold cross-validation scheme to calculate the classification performance for a fair comparison. We compare our UGformer with up-to-date strong baselines. We report the baseline results taken from the original papers or published in [1, 3, 22, 25, 27, 28].

Table 2: Graph classification results (% accuracy) w.r.t UGformer Variant 1 in an inductive setting.

Model	COLLAB	IMDB-B	IMDB-M	DD	PROTEINS	MUTAG	PTC
PSCN [19]	72.60 ± 2.15	71.00 ± 2.29	45.23 ± 2.84	77.12 ± 2.41	75.89 ± 2.76	92.63 ± 4.21	62.29 ± 5.68
GCN [10]	79.00 ± 1.80	74.00 ± 3.40	51.90 ± 3.80	–	76.00 ± 3.20	85.60 ± 5.80	64.20 ± 4.30
GFN [1]	81.50 ± 2.42	73.00 ± 4.35	51.80 ± 5.16	78.78 ± 3.49	76.46 ± 4.06	90.84 ± 7.22	–
GraphSAGE [5]	79.70 ± 1.70	72.40 ± 3.60	49.90 ± 5.00	65.80 ± 4.90	65.90 ± 2.70	79.80 ± 13.9	–
GAT [24]	75.80 ± 1.60	70.50 ± 2.30	47.80 ± 3.10	–	74.70 ± 2.20	89.40 ± 6.10	66.70 ± 5.10
DGCNN [33]	73.76 ± 0.49	70.03 ± 0.86	47.83 ± 0.85	79.37 ± 0.94	75.54 ± 0.94	85.83 ± 1.66	58.59 ± 2.47
SAGPool [11]	–	–	–	76.45 ± 0.97	71.86 ± 0.97	–	–
PPGN [13]	81.38 ± 1.42	73.00 ± 5.77	50.46 ± 3.59	–	77.20 ± 4.73	90.55 ± 8.70	66.17 ± 6.54
CapsGNN [27]	79.62 ± 0.91	73.10 ± 4.83	50.27 ± 2.65	75.38 ± 4.17	76.28 ± 3.63	86.67 ± 6.88	–
DSGC [22]	79.20 ± 1.60	73.20 ± 4.90	48.50 ± 4.80	77.40 ± 6.40	74.20 ± 3.80	86.70 ± 7.60	–
GCAPS [25]	77.71 ± 2.51	71.69 ± 3.40	48.50 ± 4.10	77.62 ± 4.99	76.40 ± 4.17	–	66.01 ± 5.91
IEGN [14]	77.92 ± 1.70	71.27 ± 4.50	48.55 ± 3.90	–	75.19 ± 4.30	84.61 ± 10.0	59.47 ± 7.30
GIN-0 [28]	80.20 ± 1.90	75.10 ± 5.10	52.90 ± 2.80	–	76.30 ± 2.80	89.40 ± 5.60	64.60 ± 7.00
UGformer	77.84 ± 1.48	77.04 ± 3.45	53.60 ± 3.53	80.23 ± 1.48	78.53 ± 4.07	89.97 ± 3.65	69.63 ± 3.60

3.2.2 Main results. Table 2 presents the experimental results of UGformer and other strong baseline models for the benchmark datasets. In general, our UGformer gains competitive accuracies on the social network datasets. Especially, UGformer produces state-of-the-art accuracies on IMDB-B and IMDB-M respectively, which outperform those of other existing models. On the bioinformatics datasets, UGformer obtains the highest accuracies on DD, PROTEINS, and PTC, respectively. Besides, there are no significant differences between our UGformer and the baselines on MUTAG as this dataset only consists of 188 graphs.

3.3 UGformer Variant 1 for graph classification in an “unsupervised transductive” setting

We propose an *unsupervised transductive* learning approach to train GNNs to address the limited availability of class labels. We consider a final embedding \mathbf{o}_v for each node v , and make the similarity between \mathbf{e}_v and \mathbf{o}_v higher than that between \mathbf{e}_v and the final embeddings of the other nodes. The goal is to guide GNNs to recognize and distinguish the sub-graph structural information within each

graph and also memorize the structural differences among graphs to produce better node and graph embeddings. We aim to minimize the sampled softmax loss function [8] applied to node v as:

$$\mathcal{L}_{\text{GNN}}(v) = -\log \frac{\exp(\mathbf{o}_v^T \mathbf{e}_v)}{\sum_{v' \in \mathcal{V}'} \exp(\mathbf{o}_v^T \mathbf{e}_{v'})} \quad (13)$$

where \mathcal{V}' is a subset sampled from $\{\cup_{m=1}^M \mathcal{V}_m\}$. Node embeddings \mathbf{o}_v are also learned as model parameters. We then sum all the final embeddings \mathbf{o}_v of nodes v in \mathcal{G} to obtain the graph embedding $\mathbf{e}_{\mathcal{G}}$.

3.3.1 Experimental setup. We use the same seven well-known datasets as used in the inductive setting. We follow some unsupervised approaches such as DGK [30] and AWE [7] to train our unsupervised UGformer on all nodes from the entire dataset (i.e., consisting of all nodes from the test set during training). The hyperparameters are also varied as same as in the inductive setting. We also train our GCN baseline following our unsupervised learning. We set the batch size to 4 and vary the number of GCN layers in {1, 2, 3} and the hidden layer size in {32, 64, 128, 256}. We also use the Adam optimizer [9] to train this unsupervised GCN up to 50 epochs.

We finally utilize the logistic regression classifier [2] with using the 10-fold cross-validation scheme to evaluate our models. We compare our unsupervised GCN (denoted as uGCN) and UGformer with the baselines: Deep Graph Kernel (DGK) [30] and Anonymous Walk Embedding (AWE) [7].

Table 3: Graph classification results (% accuracy) w.r.t UGformer Variant 1 in an unsupervised transductive setting. Note that we do not aim to directly compare the unsupervised transductive setting with the inductive setting.

Model	COLLAB	IMDB-B	IMDB-M	DD	PROTEINS	MUTAG	PTC
DGK [2015]	73.09 ± 0.25	66.96 ± 0.56	44.55 ± 0.52	73.50 ± 1.01	75.68 ± 0.54	87.44 ± 2.72	60.08 ± 2.55
AWE [2018]	73.93 ± 1.94	74.45 ± 5.83	51.54 ± 3.61	71.51 ± 4.02	-	87.87 ± 9.76	-
uGCN	93.28 ± 0.99	94.50 ± 2.79	81.66 ± 3.16	94.31 ± 1.71	89.09 ± 3.25	95.36 ± 2.64	92.67 ± 4.60
UGformer	95.62 ± 0.92	96.41 ± 1.94	89.20 ± 2.52	95.67 ± 1.89	80.01 ± 3.21	88.47 ± 7.13	91.81 ± 6.61

3.3.2 Main results. Table 3 presents the experimental results in the unsupervised transductive setting, wherein both our unsupervised UGformer and uGCN obtain new state-of-the-art accuracies on the benchmark datasets. The significant gains demonstrate a notable impact of our unsupervised transductive learning approach.

4 CONCLUSION

We present a transformer-based GNN model, named UGformer, to learn graph representations. We consider two UGformer variants of (i) leveraging the transformer on a set of sampled neighbors for each input node and (ii) leveraging the transformer on all input nodes. Experimental results show that our graph transformer UGformer produces state-of-the-art accuracies on well-known benchmark datasets for graph classification and text classification. Furthermore, we hope that future GNN works can consider the unsupervised transductive setting to address the limited availability of class labels.

REFERENCES

- [1] Ting Chen, Song Bian, and Yizhou Sun. 2019. Are Powerful Graph Neural Nets Necessary? A Dissection on Graph Classification. *arXiv:1905.04579* (2019).

- [2] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research* 9 (2008), 1871–1874.
- [3] Xiaolong Fan, Maoguo Gong, Yu Xie, Fenglong Jiang, and Hao Li. 2019. Structured Self-attention Architecture for Graph-level Representation Learning. *Pattern Recognition* 100 (2019).
- [4] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. 2017. Neural Message Passing for Quantum Chemistry. In *ICML*.
- [5] William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *NeurIPS*. 1024–1034.
- [6] William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Representation learning on graphs: Methods and applications. *arXiv:1709.05584* (2017).
- [7] Sergey Ivanov and Evgeny Burnaev. 2018. Anonymous Walk Embeddings. In *ICML*. 2191–2200.
- [8] Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On Using Very Large Target Vocabulary for Neural Machine Translation. In *ACL*.
- [9] Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR* (2015).
- [10] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*.
- [11] Junhyun Lee, Inyeop Lee, and Jaewoo Kang. 2019. Self-Attention Graph Pooling. In *ICML*. 3734–3743.
- [12] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. 2016. Gated Graph Sequence Neural Networks. *ICML* (2016).
- [13] Haggai Maron, Heli Ben-Hamu, Hadar Serviansky, and Yaron Lipman. 2019. Provably Powerful Graph Networks. In *NeurIPS*. 2153–2164.
- [14] Haggai Maron, Heli Ben-Hamu, Nadav Shafir, and Yaron Lipman. 2019. Invariant and equivariant graph networks. *ICLR* (2019).
- [15] Dai Quoc Nguyen. 2021. *Representation Learning for Graph-Structured Data*. Ph.D. Dissertation. Monash University. <https://doi.org/10.26180/14450496.v1>
- [16] Dai Quoc Nguyen, Tu Dinh Nguyen, and Dinh Phung. 2021. Quaternion Graph Neural Networks. In *Asian Conference on Machine Learning*.
- [17] Dai Quoc Nguyen, Vinh Tong, Dinh Phung, and Dat Quoc Nguyen. 2022. Node Co-occurrence based Graph Neural Networks for Knowledge Graph Link Prediction. In *Proceedings of WSDM 2022 (Demonstrations)*.
- [18] Van-Anh Nguyen, Dai Quoc Nguyen, Van Nguyen, Trung Le, Quan Hung Tran, and Dinh Phung. 2022. ReGVD: Revisiting Graph Neural Networks for Vulnerability Detection. In *Proceedings of ICSE 2022 (Demonstrations)*.
- [19] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutsov. 2016. Learning Convolutional Neural Networks for Graphs. In *ICML*. 2014–2023.
- [20] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. 1532–1543.
- [21] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2009. The graph neural network model. *IEEE Transactions on Neural Networks* 20, 1 (2009), 61–80.
- [22] Younjo Seo, Andreas Loukas, and Nathanael Peraudin. 2019. Discriminative structural graph classification. *arXiv:1905.13422* (2019).
- [23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*. 5998–6008.
- [24] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. *ICLR* (2018).
- [25] Saurabh Verma and Zhi-Li Zhang. 2018. Graph capsule convolutional neural networks. *The Joint ICML and IJCAI Workshop on Computational Biology* (2018).
- [26] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. 2019. A comprehensive survey on graph neural networks. *arXiv:1901.00596* (2019).
- [27] Zhang Xinyi and Lihui Chen. 2019. Capsule Graph Neural Network. *ICLR* (2019).
- [28] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2019. How Powerful Are Graph Neural Networks? *ICLR* (2019).
- [29] Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka. 2018. Representation Learning on Graphs with Jumping Knowledge Networks. In *ICML*. 5453–5462.
- [30] Pinar Yanardag and SVN Vishwanathan. 2015. Deep Graph Kernels. In *The ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1365–1374.
- [31] Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *AAAI*, Vol. 33. 7370–7377.
- [32] Rex Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L. Hamilton, and Jure Leskovec. 2018. Hierarchical Graph Representation Learning with Differentiable Pooling. In *NeurIPS*. 4805–4815.
- [33] Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen. 2018. An End-to-End Deep Learning Architecture for Graph Classification. In *AAAI*.
- [34] Yufeng Zhang, Xueli Yu, Zeyu Cui, Shu Wu, Zhongzhen Wen, and Liang Wang. 2020. Every Document Owns Its Structure: Inductive Text Classification via Graph Neural Networks. In *ACL*. 334–339.