OPTFM: A Scalable Multi-View Graph Transformer for Hierarchical Pre-Training in Combinatorial Optimizations

Hao Yuan, Wenli Ouyang, Changwen Zhang, Congrui Li, Yong Sun

Abstract

Foundation Models (FMs) have demonstrated remarkable success in fields like computer vision and natural language processing, yet their application to combinatorial optimization remains underexplored. Optimization problems, often modeled as graphs, pose unique challenges due to their diverse structures, varying distributions, and NP-hard complexity. To address these challenges, we propose OPTFM, the first graph foundation model for general combinatorial optimization. OPTFM introduces a scalable multi-view graph transformer with hybrid self-attention and cross-attention to model large-scale heterogeneous graphs in O(N) time complexity while maintaining semantic consistency throughout the attention computation. A dual-level pre-training framework integrates node-level graph reconstruction and instance-level contrastive learning, enabling robust and adaptable representations at multiple levels. Experimental results across diverse optimization tasks show that models trained on OPTFM embeddings without fine-tuning consistently outperform task-specific approaches, establishing a new benchmark for solving combinatorial optimization problems.

Why Foundation Models for CO?

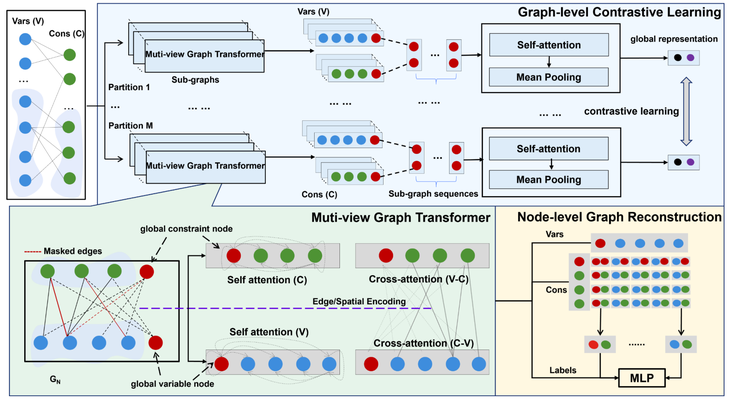
- Combinatorial Optimization (CO) problems are **NP-hard** and often modeled as graphs with complex, heterogeneous structures. Existing methods are task-specific and lack **generalization**.
- **Our goal**: Build a foundation model that learns universal representations for variables, constraints, and entire instances — enabling zero-shot transfer to unseen problems.

Methods	Set Covering (SC2)		Maximal Independent Set (MIS2)		Combinatorial Auction (CA2)		Maximum Cut (MC2)	
	Gap%	PI	Gap%	PI	Gap%	PI($\times 10^3$)	Gap%	PI
SCIP	4.51	14953	3.45	9542.1	17.87	12312	8.38	30039
RL-LNS [14]	1.66	13007	0.51	1524.7	4.13	5933.4	3.20	8449.6
Branching [68]	1.53	12916	0.55	1769.4	4.52	6142.7	3.19	7857.3
CL-LNS [15]	1.41	12914	0.41	1298.5	3.51	5621.7	2.83	7184.1
Fast-T2T [69]	-	-	0.49	1482.5	-	-	-	-
AnySCP [70]	-	-	-	-	-	-	3.51	10327
Pretrain:GCN	2.03	13983	0.79	2681.9	5.15	7095.8	3.65	10112
Pretrain:SGFormer	1.85	13425	0.55	1772.5	4.42	6716.9	3.17	8502.7
Pretrain:OPTFM-Nocross	1.49	13112	0.29	1210.0	2.95	5529.3	2.33	5521.9
Pretrain:OPTFM-WGNN	1.66	13309	0.24	1004.8	3.58	6439.7	2.33	5539.0
Pretrain:OPTFM	1.03	12699	0.15	872.16	**2.19**	**5027.3**	1.95	4339.4
Gurobi	**0.71**	**12528**	**0.01**	**495.88**	3.60	5723.5	**1.01**	**2195.6**

Methods	Set Covering (SC4)		Maximal Independent Set (MIS4)		Combinatorial Auction (CA4)		Maximum Cut (MC4)	
	Gap%	PI	Gap%	PI	Gap%	PI($\times 10^3$)	Gap%	PI
SCIP	5.41	15524	3.45	22745	16.61	25275	8.71	78510
RL-LNS [14]	3.73	14866	0.57	5365.1	3.52	13572	3.76	39645
Branching [68]	3.39	14689	0.64	5744.8	3.37	13349	4.21	42718
CL-LNS [15]	3.39	14325	0.45	4533.4	2.99	13025	3.29	37384
Fast-T2T [69]	-	-	0.63	5472.9	-	-	-	-
AnySCP [70]	-	-	-	-	-	-	4.29	38975
Pretrain:GCN	3.69	14895	0.92	8033.0	5.02	14789	3.98	41235
Pretrain:SGFormer	2.65	14098	0.55	5319.7	2.92	12997	3.95	41167
Pretrain:OPTFM-Nocross	1.55	13722	0.42	4937.5	2.25	12099	2.67	31397
Pretrain:OPTFM-WGNN	1.70	13815	0.41	4912.9	2.19	12173	2.45	30936
Pretrain:OPTFM	**1.09**	**13385**	**0.04**	2241.9	**1.98**	**11431**	**1.99**	**25213**
Gurobi	1.22	13795	**0.04**	**2215.7**	12.61	21959	5.38	51298

Key Insight & Architecture Overview

- The **First Graph Foundation Model** proposed for **Optimization**
- **Scalable multi-view** graph transformer structure for MILP
- **Dual-level self-supervised pre-training** --- node-level & graph-level



Overview of our OPTFM framework

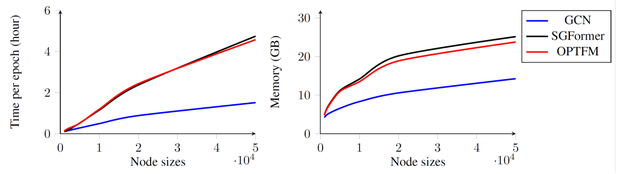
Zero-shot Performance on Downstream Tasks

	SCIP	RL-LNS	Pretrain:GCN	Pretrain:SGFormer	Pretrain:OPTFM	Gurobi
Gap%	15.15	8.07	8.95	5.19	2.92	**1.98**
Wins	96/240	114/240	99/240	121/240	171/240	**203/240**

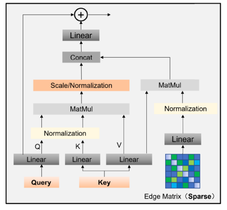
Methods	CA2 ($\times 10^3$)	CA3 ($\times 10^3$)	MIS2	MIS3	MVC2	MVC3	SC2	SC3
SCIP	11285.2	115117.8	18541.5	9086.6	31451.6	491084.8	25259.6	252199.6
GNN-GBDT [17]	13593.6	137035.9	22288.5	223295.2	27419.8	276235.0	17181.2	225725.9
LIGHT-MILPOPT [18]	13825.7	137529.5	22601.5	227198.4	27268.2	272941.2	17010.0	**165973.1**
GOAL [56]	13415.7	139471.5	22301.8	223126.1	27532.6	274472.8	17501.4	261195.3
MTL [57]	13389.3	139842.2	22284.2	224109.5	27498.5	275612.7	17443.7	258974.6
Pretrain:GCN	13022.5	136988.3	22075.9	221518.5	27965.4	279978.5	17765.9	273987.5
Pretrain:SGFormer	13787.5	140512.4	22457.6	226535.8	27398.5	275539.6	17223.5	220913.8
Pretrain:OPTFM-Nocross	14129.8	139863.5	22897.1	229614.9	27094.7	270123.5	16812.3	192945.4
Pretrain:OPTFM-WGNN	14275.6	140233.1	22935.4	228975.4	27118.4	269993.1	16793.5	181739.7
Pretrain:OPTFM	**14412.0**	**141529.7**	**23057.3**	**231528.1**	**26891.8**	**267974.5**	**16158.6**	**165972.6**
Time	2000s	30000s	2000s	8000s	2000s	8000s	2000s	12000s

Methods	Set Covering (SC)		Maximal Independent Set (MIS)		Combinatorial Auction (CA)		Maximum Cut (MC)	
	Gap%	PI	Gap%	PI	Gap%	PI ($\times 10^3$)	Gap%	PI
SCIP	3.23	20225	0.25	312.25	4.71	3312.4	8.01	15193
RL-LNS [14]	1.29	17623	0.07	182.63	2.36	2271.6	4.25	6538
Branching [68]	1.72	18007	0.07	183.44	3.09	2492.7	3.99	6104
CL-LNS [15]	0.92	17025	0.07	182.99	2.05	2198.5	3.03	3883.5
Fast-T2T [69]	-	-	0.13	241.72	-	-	-	-
AnySCP [70]	-	-	-	-	-	-	3.89	4981
Pretrain:GCN	1.95	18893	0.14	237.52	3.11	2507.8	4.73	8125
Pretrain:SGFormer	1.21	17633	0.05	181.97	2.59	2622.8	4.29	6601
Pretrain:OPTFM-Nocross	1.07	17428	0.06	182.44	2.35	2337.5	3.55	4856
Pretrain:OPTFM-WGNN	1.13	17506	0.05	181.95	2.19	2198.6	3.55	4901
Pretrain:OPTFM	0.93	**16782**	0.05	178.55	1.93	2099.5	3.02	3845
Gurobi	**0.75**	16796	**0**	**173.15**	**1.44**	**2075.4**	**0.62**	**842**

Linear-scaling attention backbone



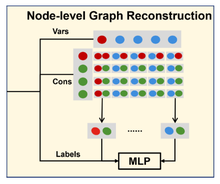
- Maintains **O(N)** complexity via linear attention across sizes;
- **Layer-wise graph encoding**
- Enable scale-transfer to **over 10M** nodes efficiently.
- Training time per graph with millions of nodes in **several seconds**.



Attention Backbone

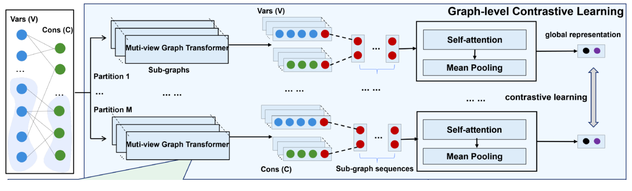
Dual-level training pipeline

- **Node-level graph reconstruction**



- Random **remove** a% edges in each sub-graph;
- **Reconstruct** the sub-graph;
- Perform on the sub-graph (node) level, limiting the **complexity**.
- Pair-wise edge prediction;

- **Graph-level contrastive learning**



- Perform on the graph-level, target on the graph embedding;
- **On top of node-level pretraining;**
- Complexity: O(K) depends on the sub-graph counts;
- **Positive** pairs: sub-graph sequences from the same graph;
- Otherwise the **negative** pairs;

Vars (V)
Cons (C)
...
...
Partition 1
...
Partition M
Muti-view Graph Transformer
Sub-graphs
Muti-view Graph Transformer
Vars (V)
... ...
... ...
Cons (C)
... ...
Sub-graph sequences
Graph-level Contrastive Learning
Self-attention
Mean Pooling
global representation
... ...
contrastive learning
Self-attention
Mean Pooling
Muti-view Graph Transformer
global constraint node
------ Masked edges
G_N
global variable node
Self attention (C)
Cross-attention (V-C)
Edge/Spatial Encoding
Self attention (V)
Cross-attention (C-V)
Node-level Graph Reconstruction
Vars
Cons
.....
Labels
MLP