RL for one-shot DAG scheduling with comparability identification and dense reward

Xumai Qi, Dongdong Zhang*,
Taotao Liu, Hongcheng Wang

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction of our contribution

Many studies solve Directed Acyclic Graph (DAG) scheduling problem in one-shot by combining reinforcement learning (RL) and list scheduling heuristic. However, they suffer from biased estimation of sampling probabilities and inefficient guidance in training. To address these issues:

- We conducted a theoretical analysis about the limitations of existing RL-based one-shot DAG scheduling methods.
- We propose a novel RL-based one-shot solution generation method for DAG scheduling, including:
  (1) a comparable antichain identification mechanism to eliminate the biased policy probability estimation;
  (2) a heuristic-based dense reward signal for node level decision-making optimization in training.
- Comprehensive comparative and ablation experiments demonstrate the superiority of our method in terms of solution quality.

Analysis of existing methods' limitations

- In brief, we found that the biased probability estimation issue lies in the redundant comparison of the priorities of task nodes.
- Additionally, the sparse reward challenge: a global reward is generated only once in the whole decision-making process, which cannot effectively reflect the specific contribution of each local decision to the final result.

Our motivation

- Masking the invalid actions (i.e., removing those redundant priority comparison) would lead to a more accurate policy sampling probability, producing lower policy gradient variance, which is beneficial to the convergence of RL model.
- It is practical to estimate the cumulative reward (return) for each local decision in one-shot scheduling.

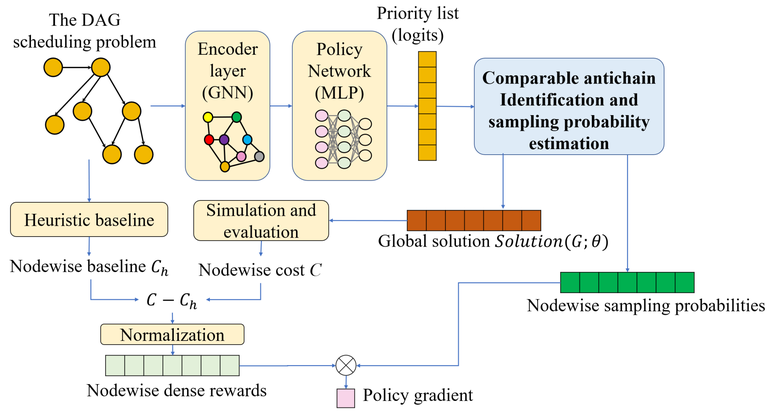


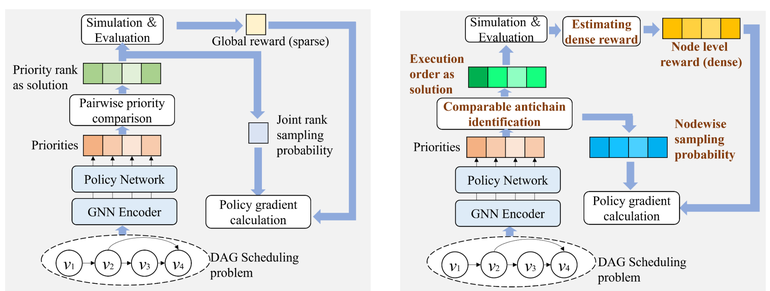
Illustration of our method

Method	SIPHT-100			SIPHT-200			SIPHT-300			SIPHT-400			SIPHT-1000		
	MS	Gap/%	Time/s	MS	Gap/%	Time/s	MS	Gap/%	Time/s	MS	Gap/%	Time/s	MS	Gap/%	Time/s
HEFT (baseline)	227.0	-	0.005	357.8	-	0.018	543.0	-	0.024	714.8	-	0.044	1821.4	-	0.196
Jeon et al. [2023]	218.5	-3.74	0.07	352.2	-1.57	0.15	550.6	1.40	0.26	712.7	-0.29	0.43	1898.1	4.21	2.44
POMO-DAG	214.0	-5.74	13.4	367.5	2.72	20.8	575.8	6.05	27.5	741.9	3.80	34.3	1875.1	2.95	50.6
EGS	200.6	-11.63	1.02	346.3	-3.21	2.35	542.8	-0.04	4.04	710.2	-0.42	10.2	1821.0	-0.02	64.5
Ours	**196.9**	**-13.3**	0.61	**338.4**	**-5.42**	0.97	**541.6**	**-0.25**	1.17	**708.3**	**-0.62**	1.22	1819.2	-0.13	2.59
Ours w/o DR	213.2	-6.07	0.35	345.6	-3.41	0.67	542.2	-0.14	1.38	710.3	-0.62	1.88	**1818.9**	**-0.13**	2.83
Ours w/o CAI	214.4	-5.55	0.11	345.8	-3.07	0.21	542.5	-0.09	0.34	710.4	-0.61	0.46	1867.7	2.54	2.22

Method	TPC-H 50			TPC-H 100			TPC-H 150		
	MS	Gap/%	Time/s	MS	Gap/%	Time/s	MS	Gap/%	Time/s
STF (baseline)	24.97	-	0.010	42.85	-	0.027	69.76	-	0.038
Jeon et al. [2023]	23.73	-4.95	0.23	41.22	-3.81	0.44	74.02	6.11	0.76
POMO-DAG	46.90	87.8	24.3	90.30	110.74	27.4	141.90	103.43	33.2
EGS	24.58	-1.55	10.8	42.18	-1.56	51.2	68.99	-1.10	156.8
Ours	**20.49**	**-17.73**	0.51	**39.22**	**-8.47**	0.82	73.47	5.32	1.49
Ours w/o DR	24.10	-3.47	0.59	39.68	-7.40	0.86	70.14	0.54	1.19
Ours w/o CAI	22.47	-10.00	0.23	42.96	0.16	0.43	**67.91**	**-2.65**	0.78

Method	JSSP 20*10			JSSP 20*20			JSSP 30*10			JSSP 30*20		
	MS	Gap/%	Time/s	MS	Gap/%	Time/s	MS	Gap/%	Time/s	MS	Gap/%	Time/s
SPT (Baseline)	516.7	-	<0.001	1096.2	-	<0.001	845.9	-	<0.001	1692.0	-	<0.002
Jeon et al. [2023]	445.2	-13.84	0.12	964.0	-12.06	0.15	7356	-13.04	0.24	1548.6	-8.48	0.15
POMO-DAG	**341.6**	**-33.88**	3.0	9362.6	-14.59	5.3	971.3	-6.45	4.2	1458.0	-13.83	6.7
EGS	465.9	-9.83	1.8	1034.5	-5.63	3.5	837.5	-0.99	6.7	1604.1	-5.20	17.0
Ours	397.3	-23.11	0.32	**813.8**	**-25.76**	0.41	**571.3**	**-32.46**	0.37	**1426.0**	**-15.72**	0.45
Ours w/o DR	394.7	-23.61	0.36	928.6	-15.29	0.41	661.1	-21.85	0.32	1481.9	-12.42	0.44
Ours w/o CAI	436.8	-15.46	0.12	939.2	-14.32	0.16	718.2	-15.10	0.12	1505.3	-11.03	0.14

Ablation studies and comparison experiments on various benchmarks



Existing method (left) and our method (right)

Comparable antichain identification

We provided formal definitions of *Comparable node pair* and *Comparable antichain,* to distinguish the set of nodes that actually influence the final scheduling. Then a method for identifying *comparable antichain* is proposed. We proved that this comparable antichain identification method for identifying the comparable antichain can eliminate redundant node priority comparisons.

Dense reward

We measure the distance between the target value of each local decision and the overall goal as a dense reward, and use a heuristic algorithm to evaluate the baseline.

Conclusion and future work

As a continuation of prior foundational researches, our method achieves theoretical guarantee and experimental performance.
For future work, we will expand this foundational study to more specific applications by considering domain. Exploring sequence-agnostic one-shot scheduling might also be a promising direction.

References

Wonseok Jeon et al. (2023). "Neural dag scheduling via one-shot priority sampling.". In: The Eleventh International Conference on Learning Representations, 2023.

Shengyi Huang and Santiago Ontañón (2020). "A closer look at invalid action masking in policy gradient algorithms.". In: arXiv preprint arXiv:2006.14171, 2020.

Our QR code: