View Reviews

Paper ID
13193

Paper Title
Distilling Autoregressive Models to Obtain High-Performance Non-Autoregressive Solvers for Vehicle Routing Problems with Faster Inference Speed

Track Name
AAAI2024

Reviewer #1

Questions

1. Summary Please briefly summarize the main claims/contributions of the paper in your own words. (Please do not include your evaluation of the paper here).
This paper addresses Vehicle Routing Problems (VRPs) by exploring a compromise between Autoregressive (AR) and Non-Autoregressive (NAR) models. AR models produce better solutions but are slower due to their sequential nature, while NAR parallel inference allows for faster solutions but with subpar quality.

The main contribution of this paper is a method, GNARKD, to transform an AR model into an NAR one. This transformation involves a modification of the AR input and output layers to enable parallel inference and using the AR solutions to supervise the training of the NAR model. GNARKD is applied on three AR models and evaluated on standard synthesized and real-world benchmarks, showing that the distilled NAR model are 4-5 times faster while only dropping solution quality by 2-3%

The paper also presents theoretical evidence that the MLE training used in NAR models is behind their generally lower performance.

2. Strengths and Weaknesses Please provide a thorough assessment of the strengths and weaknesses of the paper, touching on each of the following dimensions: - Novelty (how novel are the concepts, problems addressed, or methods introduced in the paper) - Quality (is the paper technically sound?) - Clarity (is the paper well-organized and clearly written?) - Significance (comment on the likely impact of the paper on the AI research community, i.e., explain how the paper might impact its own sub-field or the general AI community.).
The paper is well-written and clear. In particular, the authors did a great job at motivating their work by explaining, both through textual descriptions and visualizations, the advantages and disadvantages of AR and NAR models and how they intend to leverage their strengths with GNARKD.

The VRP problem addressed here is very relevant and likely to become ubiquitous with the expansion of automation. The empirical results presented here suggest that GNARKD is a great option when such problems must be solved with limited resources or under strict time constraints.

The experimental evaluation is quite extensive and includes several analyses on a testbed of 10000 instances and against many diverse SOTA baselines. The results convincingly demonstrate that GNARKD achieves its goal of significantly improving time performance with little impact on solution quality.

Minor issues (but ones that should be fixed):
The paper does not seem to follow the formatting guidelines. Most notably, in regards to some section headers using underline and line spacing.
The font in Figure 3 is not readable.
There are lots of instances of incorrect hyphenation all throughout the paper.

3. Questions for the Authors Please carefully list the questions that you would like the authors to answer during the author feedback period. Think of the things where a response from the author may change your opinion, clarify a confusion or address a limitation. Please number your questions.
Q1. Some scenarios where GNARKD could be deployed, like in warehouse automation that you bring up in the paper, present multiple agents and are often formulated as multi-agent path finding problems. Can GNARKD be deployed in such settings? If not, do you think it could be extended to accommodate the MA aspect?

4. Reproducibility Does the paper provide enough information to be reproducible? If not, please explain (It may help to consult the paper's reproducibility checklist.)
Yes

5. Resources Are there novel resources (e.g., datasets) the paper contributes? (It might help to consult the paper's reproducibility checklist)
No

6. Ethical considerations Does the paper adequately address the applicable ethical considerations, e.g., responsible data collection and use (e.g., informed consent, privacy), possible societal harm (e.g., exacerbating injustice or discrimination due to algorithmic bias, etc.)? If not, explain why not. Does it need further specialized ethics review?
not applicable

7. OVERALL EVALUATION Please provide your overall evaluation of the paper, carefully weighing the reasons to accept and the reasons to reject the paper.
Weak accept: Technically solid paper where reasons to accept, e.g., good novelty, outweigh reasons to reject, e.g., fair quality.

8. CONFIDENCE How confident are you in your evaluation?

Not very confident. I am able to defend my evaluation of some aspects of the paper, but it is quite likely that I missed or did not understand some key details, or can't be sure about the novelty of the work.

Reviewer #2

Questions

1. Summary Please briefly summarize the main claims/contributions of the paper in your own words. (Please do not include your evaluation of the paper here).
A promising method to convert autoregressive (AR) NCO solvers into non-AR (NAR) counterparts based on knowledge distillation schemes. The proposed architecture for student network might be even possible to be used in a more general context (e.g., NAR solvers trained from scratch)

However, some of the seemingly important details and evaluations are omitted.

2. Strengths and Weaknesses Please provide a thorough assessment of the strengths and weaknesses of the paper, touching on each of the following dimensions: - Novelty (how novel are the concepts, problems addressed, or methods introduced in the paper) - Quality (is the paper technically sound?) - Clarity (is the paper well-organized and clearly written?) - Significance (comment on the likely impact of the paper on the AI research community, i.e., explain how the paper might impact its own sub-field or the general AI community.).
Strengths:
- A successful and efficient NAR approach to NCO
- Qualified technical novelties
- Clear writing

Weakness:
- A bit omitted details in the teacher network architecture.
- The evaluation scheme doesn't show the comparative advantages of the proposed methods.

I'm revisiting the weakness with further details in the 'Questions for the Authors' section.

3. Questions for the Authors Please carefully list the questions that you would like the authors to answer during the author feedback period. Think of the things where a response from the author may change your opinion, clarify a confusion or address a limitation. Please number your questions.
Q1: The detailed architecture of the student network is unclear. In the current manuscript, it's impossible to ascertain how the (masked) cross-attention modules are

implemented. Evaluating the technical details of the student network is vital, as it represents one of the core contributions of this work.

Q2: How is the first city selected during the guided decoding process, especially for the Traveling Salesman Problem (TSP)? Unlike the Capacitated Vehicle Routing Problem (CVRP), which has a designated depot, the TSP allows any node to serve as the starting point. Could the authors clarify how the initial node for the TSP is selected?

Q3: The data presented in the tables appears to be somewhat unclear:

Currently, Table 1 does not include data on the performance of the teacher network. I found that the greedy decoding of the original teacher networks performs similarly or even better than the greedy decoding of the student networks in TSP/CVRP 50/100, albeit at a slower inference speed. I recommend including the original performance data of the teacher network to facilitate a more comprehensive comparison.
In Table 2, there seems to be a discrepancy in the reported total runtime for Concorde. It appears that the "average" time required to solve a single instance is calculated as the total runtime (in seconds) divided by 10,000 (the number of instances). Therefore, the single runtime multiplied by 10,000 should equal the total runtime. However, none of the reported times seem to adhere to this relationship. Could the authors provide clarification on this?
Q4: The performance of the distilled NAR models on large-scale generalizations is interesting. These models appear to outperform the original student networks. Does this trend continue in larger-scale problems, especially in zero-shot evaluation settings? Evaluating this aspect is critical for NAR models, as they often trade off fine control of feasibility during training to achieve greater scalability for larger problems.

4. Reproducibility Does the paper provide enough information to be reproducible? If not, please explain (It may help to consult the paper's reproducibility checklist.)
No

5. Resources Are there novel resources (e.g., datasets) the paper contributes? (It might help to consult the paper's reproducibility checklist)
No

6. Ethical considerations Does the paper adequately address the applicable ethical considerations, e.g., responsible data collection and use (e.g., informed consent, privacy), possible societal harm (e.g., exacerbating injustice or discrimination due to algorithmic bias, etc.)? If not, explain why not. Does it need further specialized ethics review?
N/A

7. OVERALL EVALUATION Please provide your overall evaluation of the paper, carefully weighing the reasons to accept and the reasons to reject the paper.

Weak accept: Technically solid paper where reasons to accept, e.g., good novelty, outweigh reasons to reject, e.g., fair quality.

8. CONFIDENCE How confident are you in your evaluation?
Quite confident. I tried to check the important points carefully. It is unlikely, though conceivable, that I missed some aspects that could otherwise have impacted my evaluation.

Reviewer #3

Questions

1. Summary Please briefly summarize the main claims/contributions of the paper in your own words. (Please do not include your evaluation of the paper here).
The paper presents GNARKD, a knowledge distillation approach to the vehicle routing problem (VRP) that attempts to marry the superior performance of autoregressive (AR) models with the computational efficiency of non-autoregressive (NAR) models. The approach involves converting the AR (teacher) model into its NAR equivalent and using the AR model to supervise the training of the NAR (student) model. This preserves the order dependency information that would otherwise be lost and degrade performance. The paper presents experiments applying GNARKD to three different AR models and measuring performance on randomly generated VRPs along a number of metrics, comparing the approach to both AR and NAR models as well as exact solvers. The results show that the GNARKD student model is not only more efficient but sometimes also outperforms its teacher model, and that GNARKD approaches the performance of exact solvers with substantially lower computation time.

2. Strengths and Weaknesses Please provide a thorough assessment of the strengths and weaknesses of the paper, touching on each of the following dimensions: - Novelty (how novel are the concepts, problems addressed, or methods introduced in the paper) - Quality (is the paper technically sound?) - Clarity (is the paper well-organized and clearly written?) - Significance (comment on the likely impact of the paper on the AI research community, i.e., explain how the paper might impact its own sub-field or the general AI community.).
The approach appears to be novel in its combination of AR teacher with NAR student models.

The paper seems technically sound. The authors make a good argument for why they chose to combine AR and NAR models, including a theoretical analysis (supplement) as to why NAR models are less confident in their predictions. The paper presents a comprehensive evaluation of GNARKD and the results are convincing. The authors first demonstrate GNARKD applied to different AR teachers and show how it usually results in student models that are nearly as good as the teacher, and sometimes even better. The authors then compare GNARKD to AR, NAR, and exact solvers. As expected, GNARKD does not outperform the exact solvers, but its solution quality is competitive

with other neural approaches while always having the lowest computation time, except for the smaller TSP problem, where the Concorde exact solver wins on all metrics. The authors provide further experiments that demonstrate how much better GNARKD scales to larger problems, including real-world TSPs, where GNARKD is faster than even Concorde, although with lower solution quality.

The paper is generally clear and the authors provide sufficient detail about the approach and the experiments. A couple of small comments: Figure 1 could do with a bit more elaboration--i.e., it is not immediately clear from looking at the diagrams that "TM produces a highly deterministic solution, while GCN achieves partial determinism." Also please check your hyphenation settings as there are numerous incorrectly hyphenated words in the paper that get distracting after a while--e.g., n-ode, studen-t, G-NARKD, T-SP, s-tudy, mod-e, differen-t, s-core, ...

The Vehicle Routing Problem is an important real-world problem, particularly the capacitated version, which applies to basically all delivery problems. GNARKD provides a promising ML-based solution that trades off some solution quality for much faster problem-solving. How acceptable and useful the compromise will be in real applications is a question for the future.

3. Questions for the Authors Please carefully list the questions that you would like the authors to answer during the author feedback period. Think of the things where a response from the author may change your opinion, clarify a confusion or address a limitation. Please number your questions.

1. How well-calibrated are the confidence estimates of GNARKD? I.e., It exhibits greater confidence, but is it always warranted?

2. Do you have an explanation/thoughts about the conditions under which the GNARKD student tends to perform less well (e.g., CVRP, AM vs. GNARKD-AM, Greedy)?

3. How well does GNARKD compare to approximate solvers (non-neural), which similarly attempt to balance solution quality with computation time?

4. Reproducibility Does the paper provide enough information to be reproducible? If not, please explain (It may help to consult the paper's reproducibility checklist.)

Yes

5. Resources Are there novel resources (e.g., datasets) the paper contributes? (It might help to consult the paper's reproducibility checklist)

Yes

6. Ethical considerations Does the paper adequately address the applicable ethical considerations, e.g., responsible data collection and use (e.g., informed consent, privacy), possible societal harm (e.g., exacerbating injustice or

discrimination due to algorithmic bias, etc.)? If not, explain why not. Does it need further specialized ethics review?
Yes. The paper does not mention ethical considerations but I cannot think of any significant concerns that need mention.

7. OVERALL EVALUATION Please provide your overall evaluation of the paper, carefully weighing the reasons to accept and the reasons to reject the paper.
Accept: Technically solid paper, with high impact on at least one sub-area of AI or modest-to-high impact on more than one area of AI, with good to excellent quality, reproducibility, and if applicable, resources, and no unaddressed ethical considerations.

8. CONFIDENCE How confident are you in your evaluation?
Somewhat confident, but there's a chance I missed some aspects. I did not carefully check some of the details, e.g., novelty, proof of a theorem, experimental design, or statistical validity of conclusions.

Reviewer #4

Questions

1. Summary Please briefly summarize the main claims/contributions of the paper in your own words. (Please do not include your evaluation of the paper here).
The paper introudes GNARKD, which significantly
reduces the inference time with acceptable
performance drop for Vehicle Routing problems. It combines the strengths or AR and NAR models with a Teacher-Student model.

2. Strengths and Weaknesses Please provide a thorough assessment of the strengths and weaknesses of the paper, touching on each of the following dimensions: - Novelty (how novel are the concepts, problems addressed, or methods introduced in the paper) - Quality (is the paper technically sound?) - Clarity (is the paper well-organized and clearly written?) - Significance (comment on the likely impact of the paper on the AI research community, i.e., explain how the paper might impact its own sub-field or the general AI community.).
-Novelty: the motivation is well-written and the paper may give the community a new idea on how to combine the strengths of AR and NAR models.
-Quality and Clarity: The paper is easy to follow for me. But more details can be added in Guided NAR Knowledge Distillation like the position encoding? Does the encoding function matter?
- Significance: A through evaluation is conducted in terms of accuracy and speed. But some more interesting or detailed evaluation can be done as listed in my next comments.

3. Questions for the Authors Please carefully list the questions that you would like the authors to answer during the author feedback period. Think of the things

where a response from the author may change your opinion, clarify a confusion or address a limitation. Please number your questions.

1. A general question is that in the CVRP problem, how is the constraints on the capacity is incorporated in the teacher and student model? If it is incorporated in the teacher model, can it be preserved in the student model?
2. Is the model's performance consistent, which means can the model always give the same answer in the greedy mode?
3. It would be interest to see how the student and the teacher model perform on some boundary case. Let's consider if the optimal path is a circle, is the sequences of circle generated by the teacher and studnet the same?

4. Reproducibility Does the paper provide enough information to be reproducible? If not, please explain (It may help to consult the paper's reproducibility checklist.)

Yes

5. Resources Are there novel resources (e.g., datasets) the paper contributes? (It might help to consult the paper's reproducibility checklist)

No

6. Ethical considerations Does the paper adequately address the applicable ethical considerations, e.g., responsible data collection and use (e.g., informed consent, privacy), possible societal harm (e.g., exacerbating injustice or discrimination due to algorithmic bias, etc.)? If not, explain why not. Does it need further specialized ethics review?

No. The study is based on simulation.

7. OVERALL EVALUATION Please provide your overall evaluation of the paper, carefully weighing the reasons to accept and the reasons to reject the paper.

Weak accept: Technically solid paper where reasons to accept, e.g., good novelty, outweigh reasons to reject, e.g., fair quality.

8. CONFIDENCE How confident are you in your evaluation?

Quite confident. I tried to check the important points carefully. It is unlikely, though conceivable, that I missed some aspects that could otherwise have impacted my evaluation.