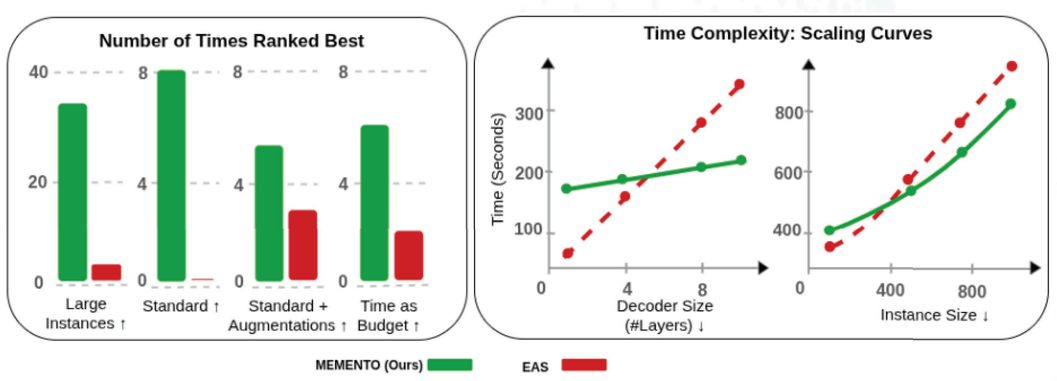
Memory-Enhanced Neural Solvers for Routing Problems

Felix Chalumeau, Refiloe Shabe, Noah De Nicola, Arnu Pretorius,
Thomas D. Barrett* & Nathan Grinsztajn*

InstaDeep®

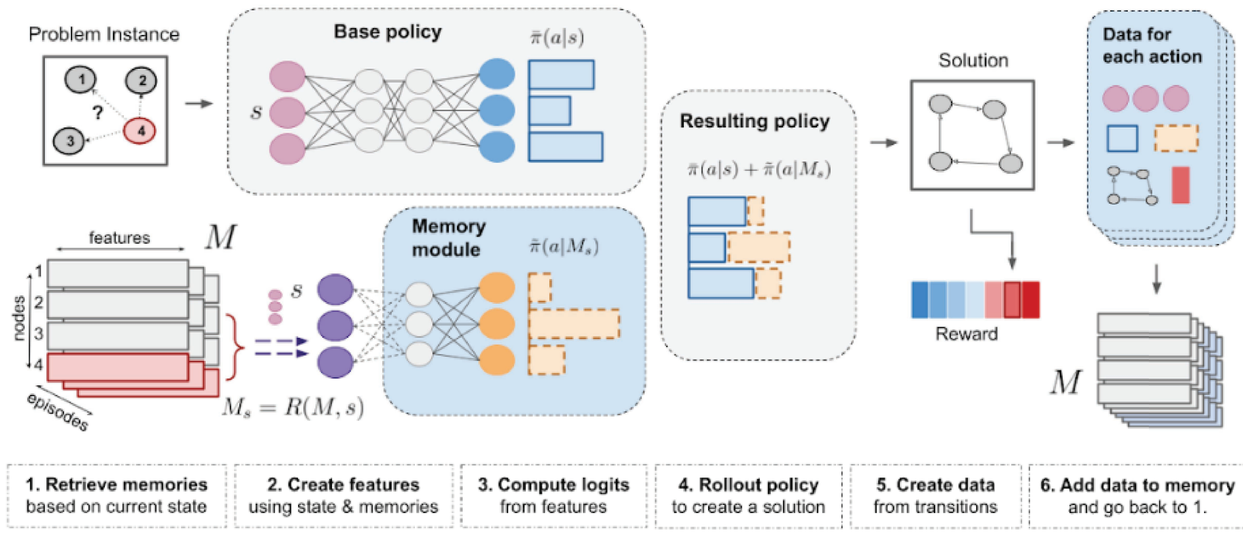
Motivation

- **Routing problems** span vital real-world domains (e.g., transportation & logistics).
- **Reinforcement learning** (RL) has emerged as a promising avenue to tackle routing problems.
- **The use of memory** has grown in modern deep learning, demonstrated by the success in Natural Language Processing (NLP), and in RL.
- Current best RL methods either rely on a collecting of pre-trained policies [3] (**limited to selecting the most appropriate one**), or RL fine-tuning [2] (suffers inherent **drawbacks of back-propagation**).
- Consequently, these methods **fail to fully utilise newly available information** within the constraints of a given search budget.
- To overcome this challenge, we propose an approach that **leverages memory of recently collected data** to improve the search of neural solvers at inference.



MEMENTO outperforms Efficient Active Search on a wide range of instance sizes and inference budgets. Its time complexity makes it scalable to large instances and policies.

MEMENTO



The framework included 6 steps of **training phase**.
When taking a decision, **data from similar states is retrieved** (1), prepared (2), then processed by a MLP to derive correction logits for each action (3). **The action distribution is updated by summing the original and new logits**. The resulting policy is then rolled out (4), and transitions' data is stored in a memory (5, 6), including **node visited, action taken, log probability, and return obtained.**

At **inference**, the collected data stored in the memory is used to dynamically adapt neural solvers. MEMENTO is **architecture-agnostic, it can be used on top of any pre-trained policy.**

Results

Routing Problems: Travelling Salesman (TSP) and Capacitated Vehicle Routing (CVRP).
Baselines: POMO [1], SGBS [4] and EAS [2] -> all single agents with given equal search budget.
Evaluation: 8 tasks spanning in- and out-of distribution for both environments (routing problems).

MEMENTO outperforms all the baselines on all the categories with significant improvement over the base policy POMO, highlighting the efficacy of learned policy updates when compared to vanilla policy gradients.

(a) TSP

Method	Training distr. $n = 100$ Obj.	Gap	Generalization $n = 125$ Obj.	Gap	$n = 150$ Obj.	Gap	$n = 200$ Obj.	Gap
LKH3	7.765	0.000%	8.583	0.000%	9.346	0.000%	10.687	0.000%
POMO (greedy)	7.796	0.404%	8.635	0.607%	9.440	1.001%	10.933	2.300%
POMO (sampling)	7.779	0.185%	8.609	0.299%	9.401	0.585%	10.956	2.513%
SGBS *	*7.769**	0.058%	-	-	*9.367**	0.220%	*10.753**	*0.619%*
EAS-Emb	7.778	0.161%	8.604	0.238%	9.380	0.363%	10.759	0.672%
MEMENTO	**7.768**	**0.045%**	**8.592**	**0.109%**	**9.365**	**0.202%**	**10.758**	**0.664%**

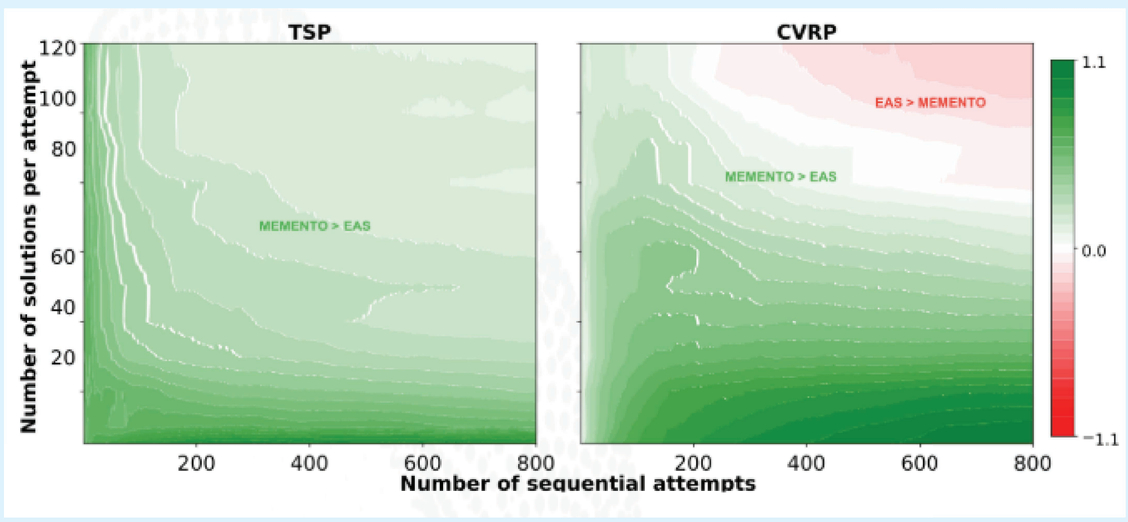
(b) CVRP

Method	Training distr. $n = 100$ Obj.	Gap	Generalization $n = 125$ Obj.	Gap	$n = 150$ Obj.	Gap	$n = 200$ Obj.	Gap
LKH3	15.65	0.000%	17.50	0.000%	19.22	0.000%	22.00	0.000%
POMO (greedy)	15.874	1.430%	17.818	1.818%	19.750	2.757%	23.318	5.992%
POMO (sampling)	15.713	0.399%	17.612	0.642%	19.488	1.393%	23.378	6.264%
SGBS *	*15.659**	0.08%	-	-	*19.426**	1.08%	*22.567**	2.59%
EAS-Emb	15.663	0.081%	17.536	0.146%	19.321	0.528%	22.541	2.460%
MEMENTO	**15.657**	**0.066%**	**17.521**	**0.095%**	**19.317**	**0.478%**	**22.492**	**2.205%**

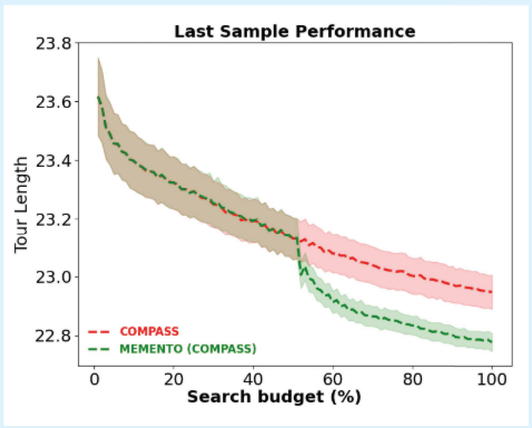
Scaling RL to larger instances

- Train MEMENTO on instances of size 500 starting from a checkpoint of instances of size 100.
- Use Efficient Attention to reduce memory cost of multi-head attention.
- Utilise gradient accumulation to keep good estimates despite small instance batch sizes.

MEMENTO outperforms EAS on instances of size 500 across batch sizes and sequential attempts. Green areas indicate settings where MEMENTO adapts more efficiently. It consistently outperforms EAS on TSP and in most CVRP settings, with strong gains under low budgets. MEMENTO achieves SOTA performance on TSP500 and overall SOTA when combined with COMPASS



Zero-shot combination with unseen solvers



- **No retraining**, zero-shot combination with COMPASS.
- **Activate MEMENTO at 50%** of the budget consumption.
- Clear and significant **drop in tour length** showing performance improvement.
- Narrower standard deviation indicating **confidence boost** and **narrowed down search.**

Conclusion

We present MEMENTO, a method to improve adaptation of neural CO solvers to unseen instances by conditioning a policy directly on data collected online during search and search budget. In practice, it proves to outperform stochastic sampling, tree search and policy-gradient fine-tuning, and shows zero-shot combination with an unseen solver. We open-source the algorithms and all the experimental data and results.

References

[1] Kwon et al., POMO: Policy Optimization with Multiple Optima for Reinforcement Learning. Neurips (2020)
[2] Hottung et al., Efficient active search for Combinatorial Optimization Problems, ICLR (2022)
[3] Chalumeau et al., Combinatorial Optimization with Policy Adaptation using Latent Space Search, Neurips (2023)
[4]Choo et al., Simulation-guided Beam Search for Neural Combinatorial Optimization, Neurips (2022)

Full paper

Deep Learning Indaba 2025, Kigali, Rwanda