

# Memory Augmented Policy Optimization (MAPO) for Program Synthesis and Semantic Parsing

Chen Liang, Mohammad Norouzi, Jonathan Berant, Quoc Le, Ni Lao



## Program Synthesis / Semantic Parsing

how many more passengers flew to los angeles than to saskatoon?

REWARD

12,467

Sparse

Latent

```
(filter_in rows ['saskatoon'] r.city)
(filter_in rows ['los angeles'] r.city)
(diff v1 v0 r.passengers)
```

Rank	City	Passengers	Ranking	Airline
1	United States, Los Angeles	14,749		Alaska Airlines
2	United States, Houston	5,465		United Express
3	Canada, Calgary	3,761		Air Transat, WestJet
4	Canada, Saskatoon	2,282	4	
5	Canada, Vancouver	2,103		Air Transat
6	United States, Phoenix	1,829	1	US Airways
7	Canada, Toronto	1,202	1	Air Transat, CanJet
8	Canada, Edmonton	110		
9	United States, Oakland	107		

## Existing Solutions

Policy Gradient

Actor

On-policy Samples

Learner

Updated Policy

Unbiased => optimal solution

High variance => slow training

Only requires a reward signal

Imitation Learning

Actor

Demonstration

Learner

Updated Policy

Biased => suboptimal solution

Low variance => fast training

Requires human supervision

Importance Sampling

On-policy samples

Replay buffer

Importance sampling

Actor

Learner

Updated Policy

With truncation

W/o truncation

Biased

Unbiased

Low variance

High variance

Only requires a reward signal

## Memory Augmented Policy Optimization

- MAPO incorporates a memory of promising samples an unbiased gradient estimate with low variance.

MAPO

High-reward samples

Memory buffer

Samples inside Memory

Actor

Samples outside Memory

Learner

Updated Policy

Programs inside Memory

Programs outside Memory

Enumeration / Sampling

Sampling

Gradient Estimate

Unbiased => optimal solution

Low variance => fast training

Only requires a reward signal

- Decompose the expected return objective into weighted sum of two expectations inside and outside the memory.
$$\mathcal{O}_{ER}(\theta) = \pi_{\mathcal{B}} \mathbb{E}_{\mathbf{a} \sim \pi_{\theta}^+(\mathbf{a})} R(\mathbf{a}) + (1 - \pi_{\mathcal{B}}) \mathbb{E}_{\mathbf{a} \sim \pi_{\theta}^-(\mathbf{a})} R(\mathbf{a})$$

Expectation inside  $\mathcal{B}$       Expectation outside  $\mathcal{B}$

$$\nabla_{\theta} \mathcal{O}_{ER}(\theta) = \pi_{\mathcal{B}} \mathbb{E}_{\mathbf{a} \sim \pi_{\theta}^+(\mathbf{a})} \nabla \log \pi_{\theta}(\mathbf{a}) R(\mathbf{a}) + (1 - \pi_{\mathcal{B}}) \mathbb{E}_{\mathbf{a} \sim \pi_{\theta}^-(\mathbf{a})} \nabla \log \pi_{\theta}(\mathbf{a}) R(\mathbf{a})$$

$\mathcal{B}$  denotes the memory buffer.  $\pi_{\theta}^+$  and  $\pi_{\theta}^-$  denotes the renormalized probability.
- Memory weight clipping

WikiTable

WikiSQL

Clip fraction

Train accuracy
- Systematic exploration
  - Use a bloom filter to force the exploration to generate new programs.
  - Trade-off memory for more efficient exploration.
- Distributed sampling
  - Distribute the cost of computing  $\pi_{\mathcal{B}}$  and sampling into the actors.
  - Multiple actors each interacting with a shard of training set and send samples to a learner.

## Experiments

	E.S.	Dev.	Test
Pasupat & Liang (2015)	-	37.0	37.1
Neelakantan <i>et al.</i> (2017)	1	34.1	34.2
Neelakantan <i>et al.</i> (2017)	15	37.5	37.7
Haug <i>et al.</i> (2017)	1	-	34.8
Haug <i>et al.</i> (2017)	15	-	38.7
Zhang <i>et al.</i> (2017)	-	40.4	43.7
MAPO	1	42.4 ± 0.5	43.2 ± 0.5
MAPO (ensembled)	10	-	46.6

	Dev.	Test
<b>Fully supervised</b>		
Zhong <i>et al.</i> (2017)	60.8	59.4
Wang <i>et al.</i> (2017)	67.1	66.8
Xu <i>et al.</i> (2017)	69.8	68.0
Huang <i>et al.</i> (2018)	68.3	68.0
Yu <i>et al.</i> (2018)	74.5	73.5
Sun <i>et al.</i> (2018)	75.1	74.6
Dong & Lapata (2018)	79.0	78.5
<b>Weakly supervised</b>		
MAPO	71.6 ± 0.6	71.8 ± 0.4
MAPO (ensemble of 5)	-	74.9

- First RL-based state-of-the-art approach on **WikiTableQuestions**.
- Competitive to state-of-the-art on **WikiSQL**, which uses strong supervision (the ground truth programs), while MAPO only uses weak supervision (the final answers).

WikiTable

WikiSQL

- MAPO converges **slower** than **maximum likelihood training**, but reaches **a better solution**.
- REINFORCE doesn't make much progress (<10% accuracy).
- Spurious programs: right answer for the wrong reason

Which nation won the most silver medal?

Correct program:

(argmax rows "Silver")

(hop v1 "Nation")

Spurious programs:

(argmax rows "Gold")

(hop v1 "Nation")

(argmax rows "Bronze")

(hop v1 "Nation")

	Question 1		Question 2	
	correct	spurious	spurious	spurious
Iterative Maximum Likelihood (IML)	0.5	0.5	0.5	0.5
Maximum Marginal Likelihood (MML)	0.8	0.2	0.5	0.5
MAPO	0.6	0.15	0.1	0.1
Model Probability	0.6	0.15	0.1	0.1