Gradient-free Policy Architecture Search and Adaptation

Sayna Ebrahimi[1], Anna Rohrbach[1,2], Trevor Darrell[1]
[1]UC Berkeley, [2]MPII

BAIR Berkeley Artificial Intelligence Research
CoRL 2017 International Conference on Robot Learning

Abstract

We explore policy architecture search and adaptation via gradient-free optimization for autonomous driving tasks. Using both demonstration and environmental reward our method can learn with relatively few early catastrophic failures. Our approach has two steps:

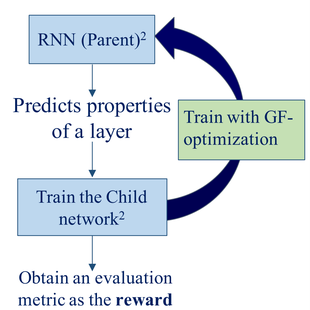
1. We first learn an architecture to perceive aspects of world state relevant to the expert demonstration
2. We then adapt a policy demonstrated in a source domain to rewards obtained in a target environment.

We show that our approach allows *safer learning* than baseline methods, offering a reduced cumulative crash metric over the agent's lifetime as it learns to drive in a realistic simulated environment.

Architecture Search w. GF-opt

The *parent* predicts the layers of the *child* network and is trained with GF-optimization algorithm

performance and **cost** of the child network are optimized



Algorithm 1 Gradient-free Optimization Algorithm

for $t = 1$ to T **do**
  Sample $\Delta_t \sim \text{Laplace}(0, 0.07)$
  $y_t^{(+)} = R(\theta + c_t\Delta_t)$
  $y_t^{(-)} = R(\theta - c_t\Delta_t)$
  $\nabla_{\theta_t} = \frac{y_t^{(+)} - y_t^{(-)}}{2c_t\Delta_t}$
  $\theta \leftarrow \theta + \alpha_t \nabla_{\theta_t}$

Behavioral Cloning w. Arch Search

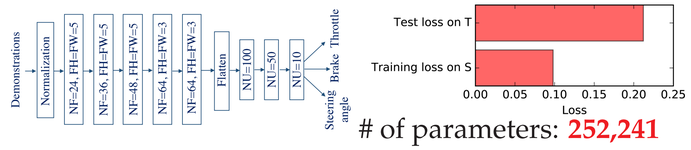
- 2.2M images collected by expert policy on GTA-V
- Labeled with steering angle, brake, throttle

Sample images from **source** domain:

Sample images from **target** domain:

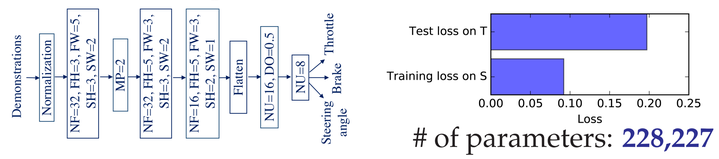
Results obtained for architecture search:

- Baseline[1]



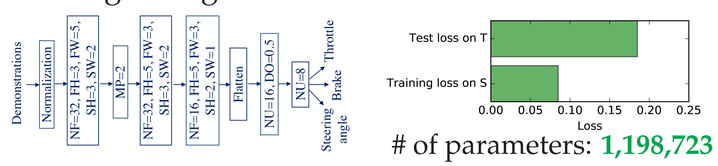
# of parameters: **252,241**

- Our small designed architecture



# of parameters: **228,227**

- Our large designed architecture



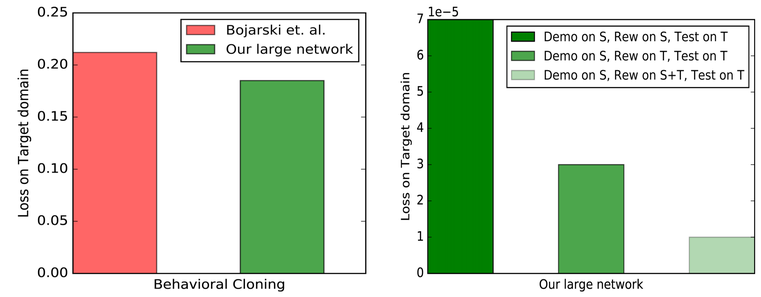
# of parameters: **1,198,723**

Reward for Adaptation

Binary rewards received from the game based on:

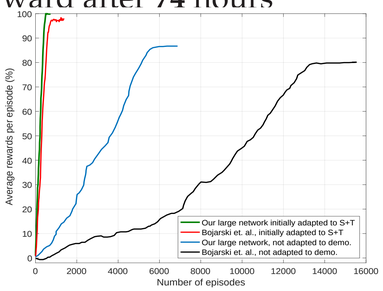
- Lane keeping
- No crash of any kind

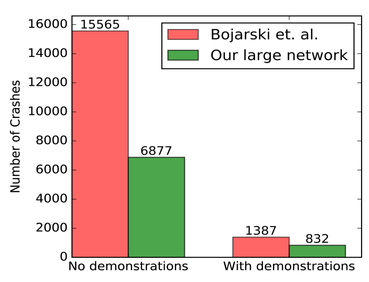
The behaviorally cloned model is shown to adapt with the target domain using rewards as the loss goes down:



Safe Learning

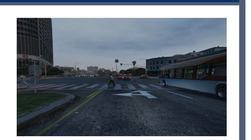
Learning with no demo. causes 8-10 times more crashes! Our large network reaches to **100%** of averaged reward after **53** hours training (**90** minutes with no mistake in its last episode) while baseline reaches **97.3%** of averaged reward after **74** hours





Predictions for a complex driving scene

	BC	Demo on S Rew on S Test on T	Demo on S Rew on S+T Test on T	Demo on S Rew on T Test on T	BC	Demo on S Rew on S Test on T	Demo on S Rew on S+T Test on T	Demo on S Rew on T Test on T
	Our large network				**Baseline[1]**			
Angle	-0.006	0.003	0.005	0.002	-0.005	-0.002	0.002	-0.001
Brake	0.191	0.889	0.931	0.956	0.183	0.567	0.677	0.778
Throttle	0.665	0.083	0.052	0.010	0.775	0.223	0.121	0.156



Pedestrian passing while the light is green

More example videos can be found at ***https://saynaebrahimi.github.io/corl.html***

References

[1] Bojarski, Mariusz, et al. "End to end learning for self-driving cars." *arXiv preprint arXiv:1604.07316 (2016).*

[2] Zoph, Barret, and Quoc V. Le. "Neural architecture search with reinforcement learning." *arXiv preprint arXiv:1611.01578 (2016).*