Gradient-free Policy Architecture Search and Adaptation

Sayna Ebrahimi, Anna Rohrbach, Trevor Darrell
UC Berkeley

Overview

ALVINN (1989)

Bojarski, et. al. 2016

Xu *et. al.* (2016)

Our contributions:

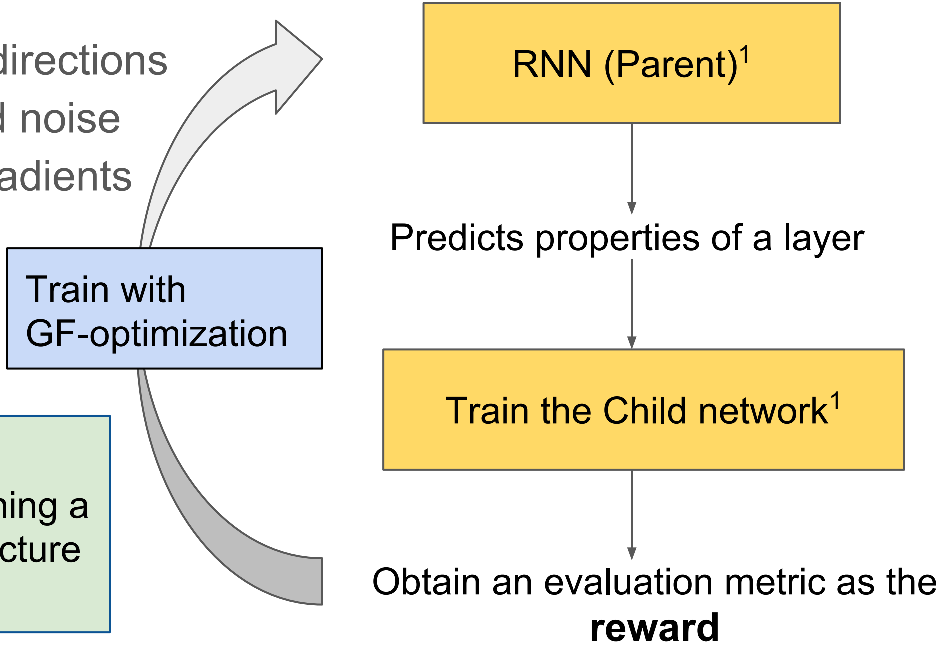
Using **gradient-free optimization** for:

- Architecture search on demonstration to mimic expert policy
- Adapting the learnt policy to the target domain by giving rewards resulting in a ***safe learning*** method

Architecture Search using GF-optimization

Optimizing the reward function by:

- Perturbing the parameters in random directions
- Evaluate the reward due to the applied noise
- Use finite difference to estimate the gradients and update!

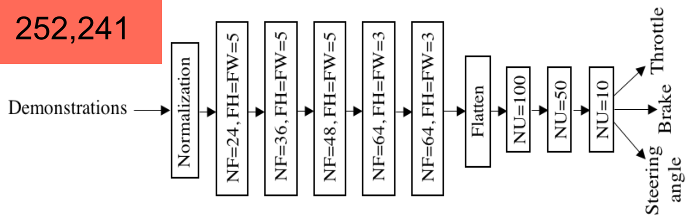


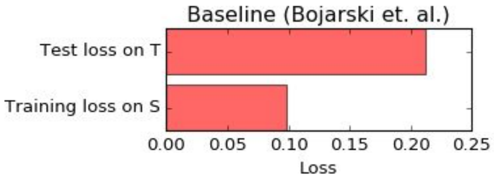
Reward function:
Optimizes **both cost and performance** by reaching a certain performance first and growing the architecture if performance improves

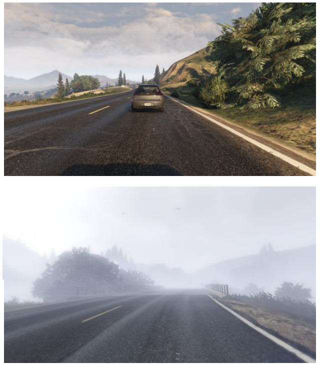
[1] Zoph, B., & Le, Q. V. (2016). Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*.

Architecture Search for Behavioral Cloning on GTA-V

252,241



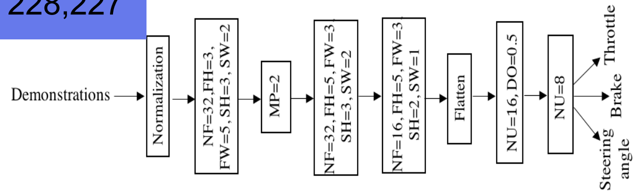


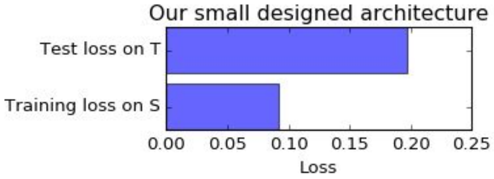


Samples from **source** domain

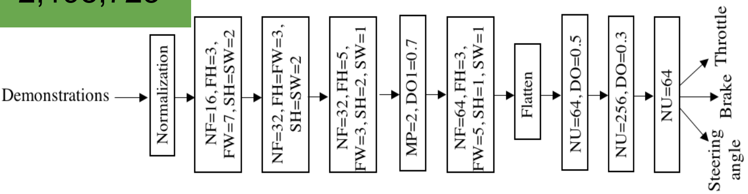
Samples from **target** domain

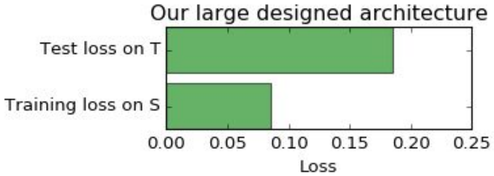
228,227





2,198,723

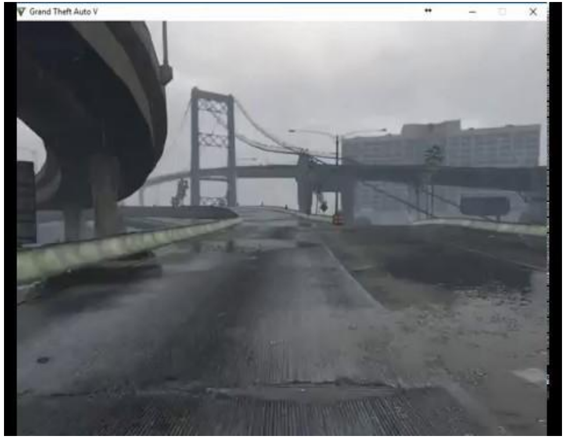




Demonstrations:

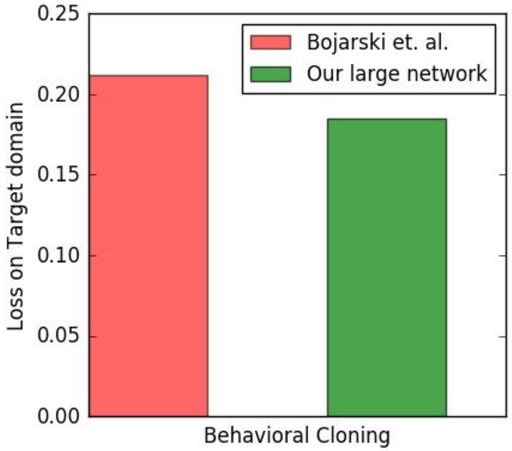
~2.2M images collected by expert policy,
Labeled with steering angle, brake,
throttle

Driving with Behaviorally Cloned Models



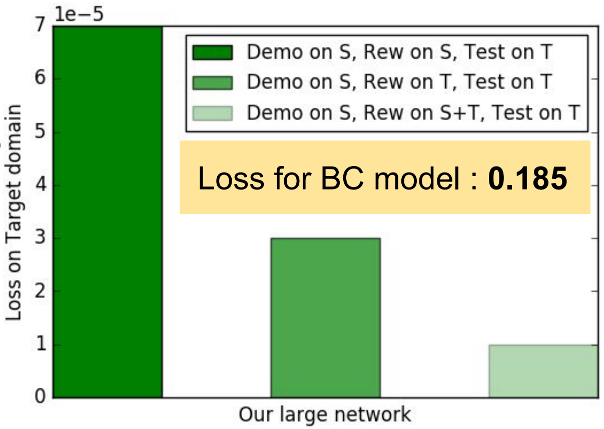
Baseline (Bojarski *et. al.*)
BC on demo. in source (S),
Testing on target (T) domain

Our large network
BC on demo. in source (S),
Testing on target (T) domain



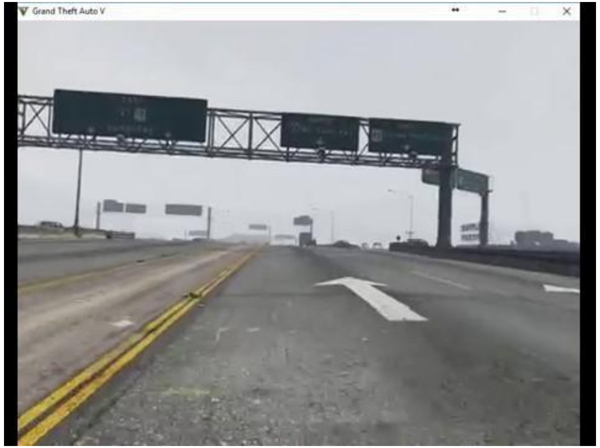
Comparison of total loss
between cloned models

Adapting to Target Domain with Rewards

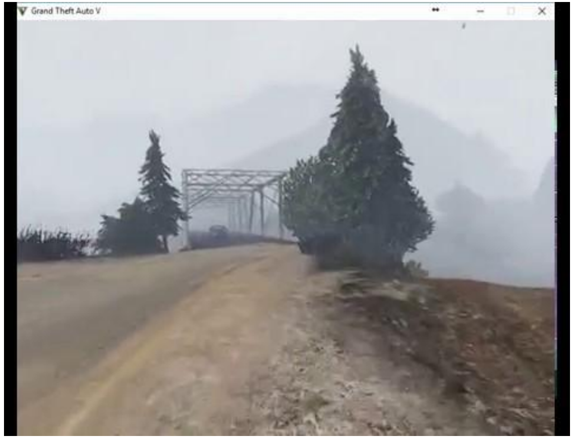


Binary rewards received from the game environment based on:

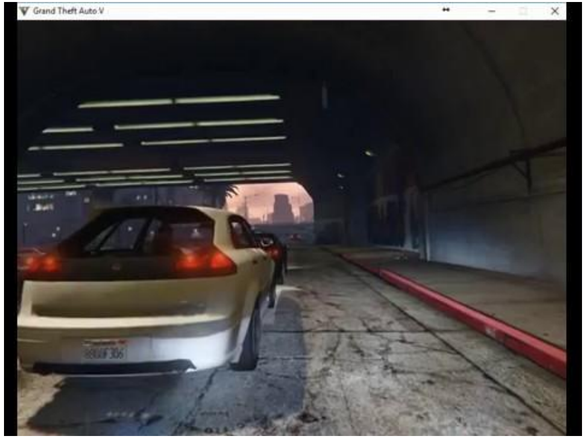
- Lane keeping
- No crash of any kind



Our large network
Demo on S, Rewards on S,
Testing on T



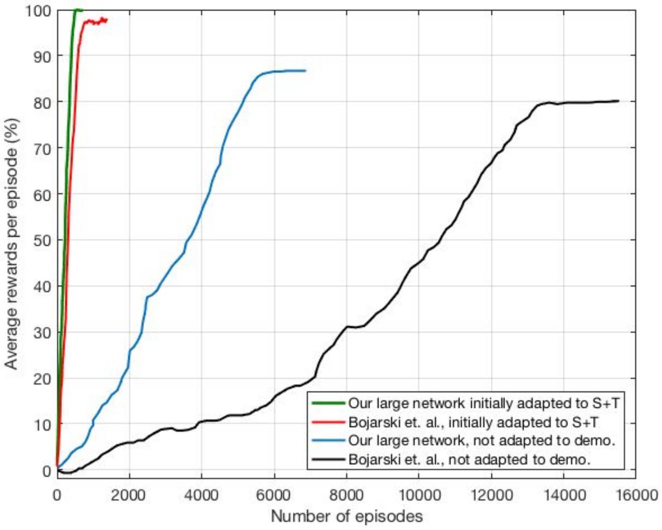
Our large network
Demo on S, Rewards on T,
Testing on T

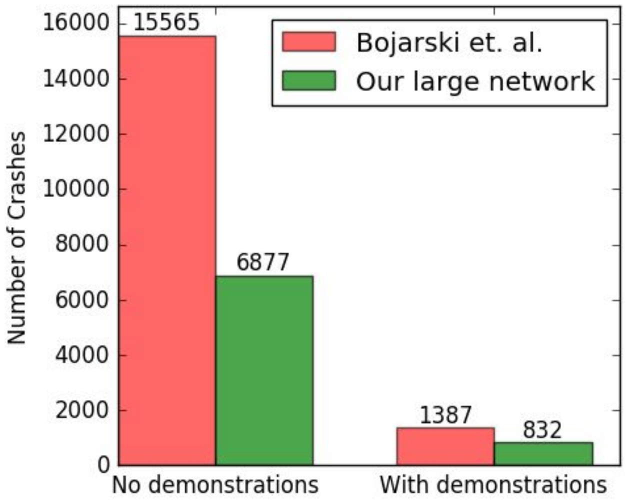


Our large network
Demo on S, Rewards on S+T,
Testing on T

Safe Learning

- Learning with no demonstrations is a hassle!
- 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





This is how it looks like to learn with no demonstrations!

Final Remarks

- ❖ We performed architecture search to mimic expert policy, optimizing both performance and computational cost and outperformed the baseline (Bojarski *et. al.*)
- ❖ We successfully adapted the learned policy to a new domain by using rewards received from the environment.
- ❖ We showed that combining imitation learning with a reward-based approach can achieve remarkably better results, faster convergence, as well as starting with less number of crashes (safe learning).

Thank you!

https://saynaebrahimi.github.io/corl.html