End to End Learning for Self-Driving Cars

[Bojarski et al 2016]

Tingwu Wang, Dylan Turpin, Animesh Garg

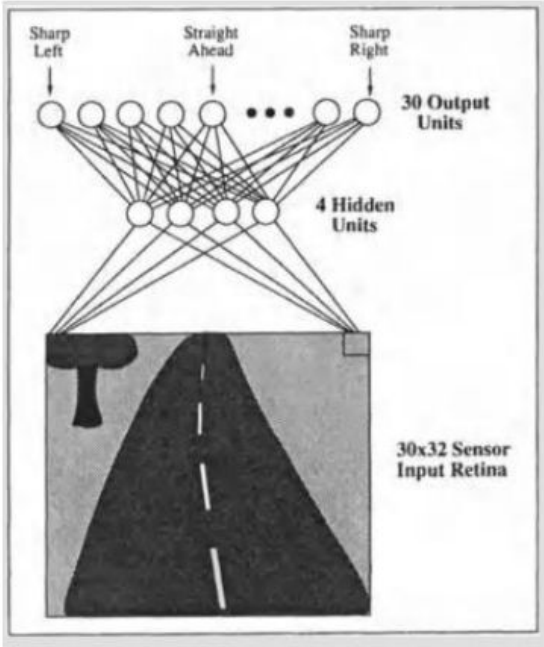

In-vehicle camera

Motivation: End to End Learning

- **Why end-to-end instead of a modular system?**
  - Gives up interpretability (at least in this case), so needs strong motivation
  - Better performance
    - system optimizes to maximize overall performance, not human chosen intermediate criteria
  - Smaller systems
    - system learns to use minimum number of steps instead of contrived human chosen ones
  - “The Bitter Lesson” [Rich Sutton 2019]
    - General approaches that incorporate minimal prior human knowledge and scale with computing power will eventually outperform handcrafted approaches

Motivation: related works

- Not the first end-to-end lane follower, why build another
- **ALVINN** (Dean Pomerleau 1989)
  - 30 x 32 pixels, 3 layer fully-connected network
  - 2016 what's changed?
    - "more data and computational power"
    - CNN instead of tiny FC network



[1] Dean A. Pomerleau. ALVINN, an autonomous land vehicle in a neural network. Technical report, Carnegie Mellon University, 1989. URL: http://repository.cmu.edu/cgi/viewcontent.cgi?article=2874&context=compsci

Motivation: related works

- Not the first, why build another
- **ALVINN** (Dean Pomerleau 1989) [1]
- **DAVE** (DARPA Autonomous Vehicle 2004) [2]
  - RC car
  - Trained on 2 hours of human time
  - Not reliable: crashed every ~20m in complex environments

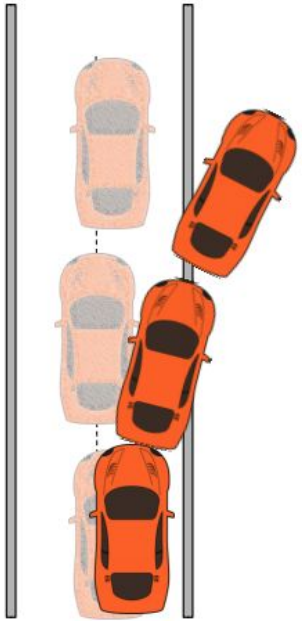
[2] Net-Scale Technologies, Inc. Autonomous off-road vehicle control using end-to-end learning, July 2004. Final technical report. URL: http://net-scale.com/doc/net-scale-dave-report.pdf.

Contributions

- **Problem:** Learn to steer from pixels with end-to-end system, learn from (augmented) human driving data
- **Why is this problem important?**
  - Autonomous driving is high impact
  - End-to-end systems should maximize performance
    - (no trade offs against interpretability, contrived human features)
  - Learn in the target domain
- **Why is the problem hard?**
  - Driving conditions vary
  - Compounding error, covariate shift
- **Key contribution:**
  - A road-tested autonomous lane following system that achieves 98% autonomy in realistic driving conditions is trained from human driving data

Imitation learning ≠ Supervised learning

- **Supervised learning**
  - Assumes train/test data are i.i.d.
- **Imitation learning**
  - Train/test data are not i.i.d.
  - Test distribution is different from training distribution
  - Your actions affect future observations/data



Slide from CSC2621 W2019 Florian Shkurti

Independent in time errors

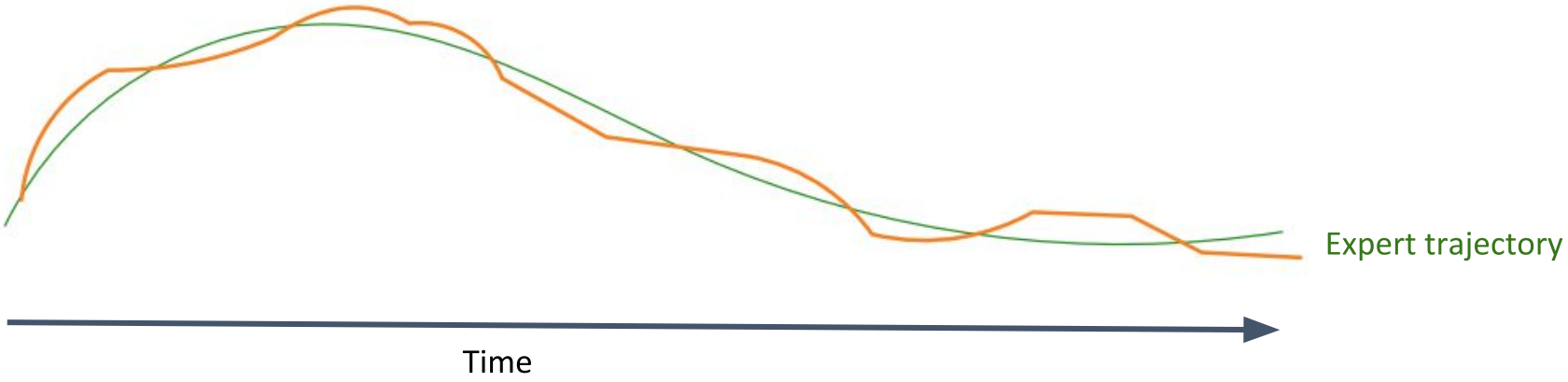


Image from Katerina Fragkiadaki CMU 10-703 slides

Compounding Errors

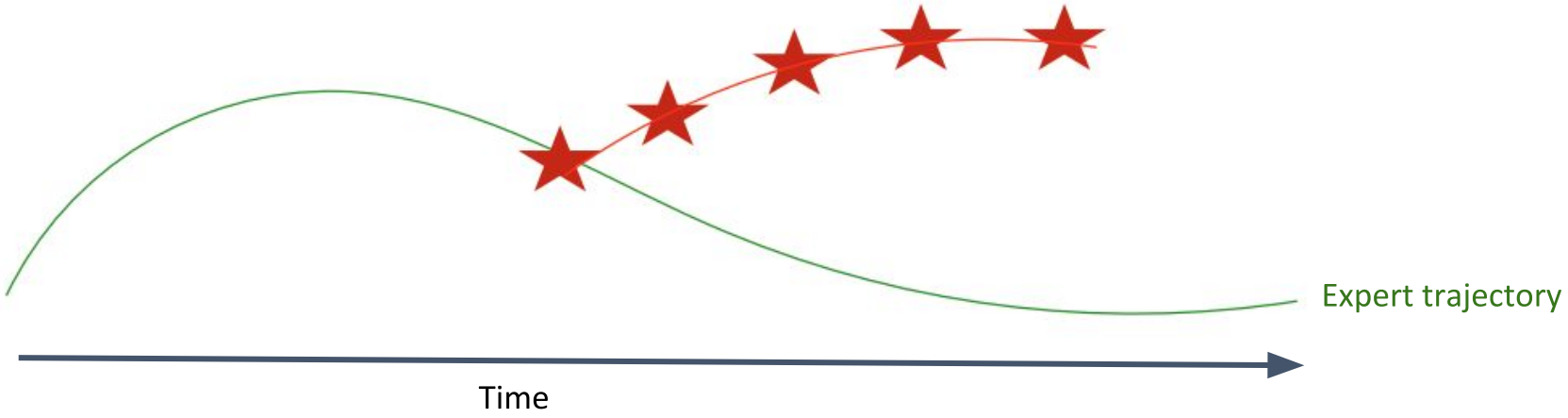
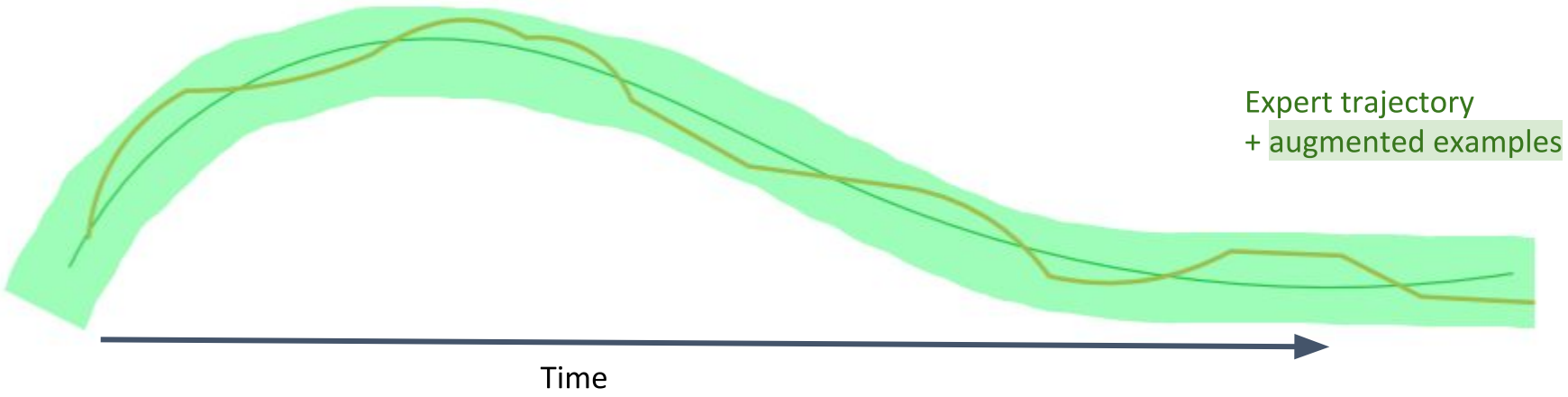


Image from Katerina Fragkiadaki CMU 10-703 slides

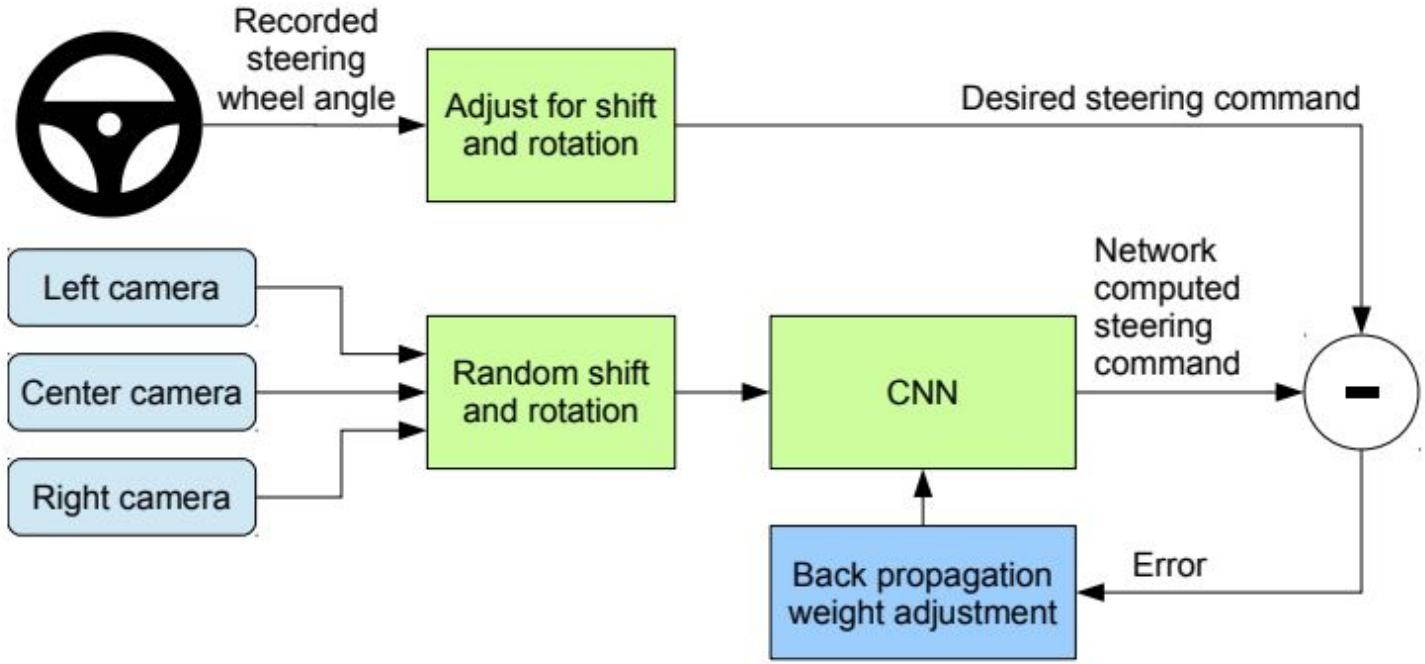
Data augmentation

- **Remind me to ask Professor Garg:**
  - Do you expect DAGGER would have worked better than data augmentation?



- **Problem:** Expert data is recorded from human drivers. How can it be augmented with nearby paths?
  - Augmentation is easy in a driving sim.
  - But IRL, the car went where it went, can’t go back and take nearby turns

Method: data augmentation



Pomerleau had a similar idea in his 1993 thesis.

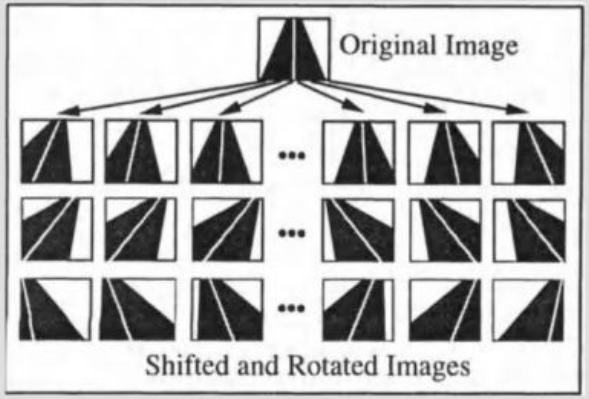
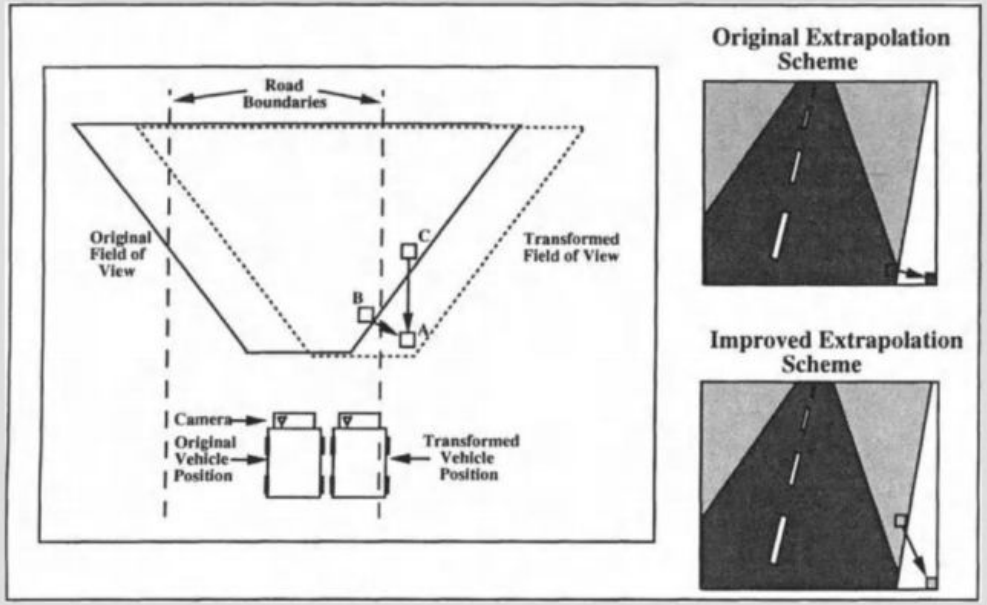


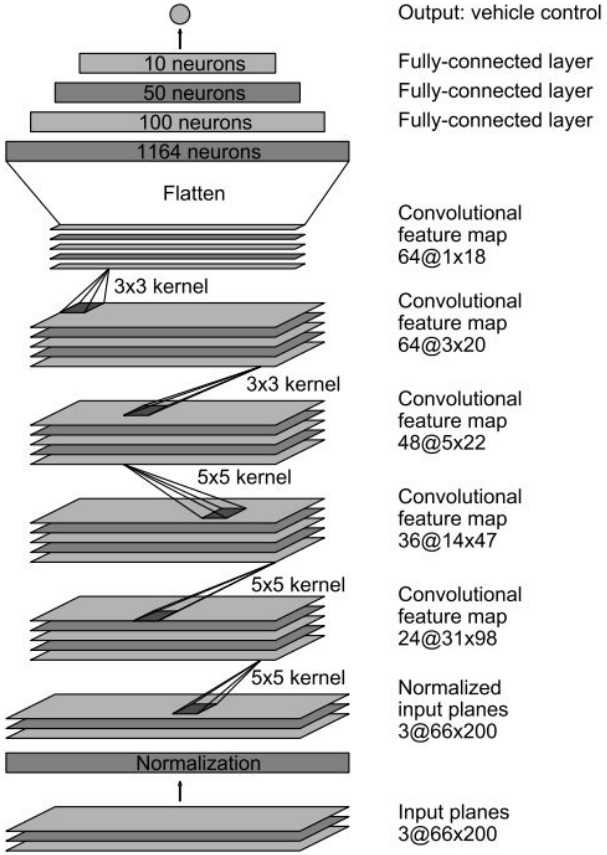
Figure 3.4: The single original video image is shifted and rotated to create multiple training exemplars in which the vehicle appears to be at different locations relative to the road.



Method: CNN architecture

(What Pomerleau didn't have…)

- ~27 million connections, ~250 thousand parameters
- "chosen empirically through a series of experiments that varied layer configurations"
- Design
  - Early layers designed to be feature extractors
  - Later layers designed to be a controller
  - But no real separation in an end-to-end system

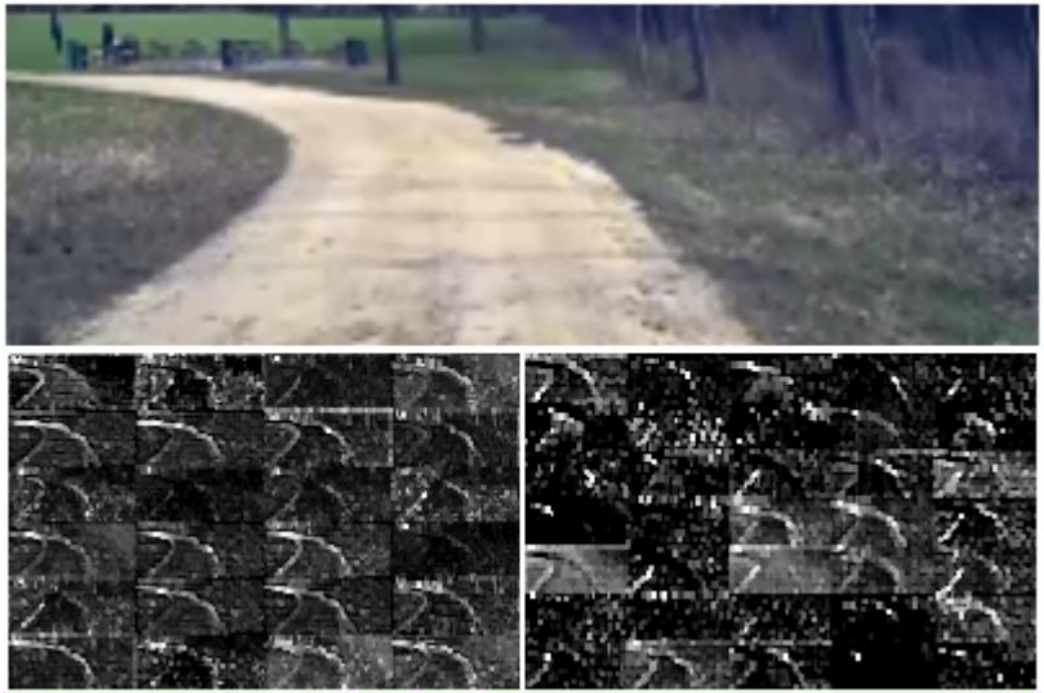


Evaluation: Autonomy

- **How should on-road performance be evaluated?**
  - Autonomy
    - What percentage of the time can the car drive itself without human intervention?
    - **On-road result:** 98% autonomy, about 1 intervention every 5 minutes
- **How should simulator performance be evaluated?**
  - Estimate autonomy
  - Intervene when vehicle departs from center line by more than one meter
  - Intervention takes ~6 seconds

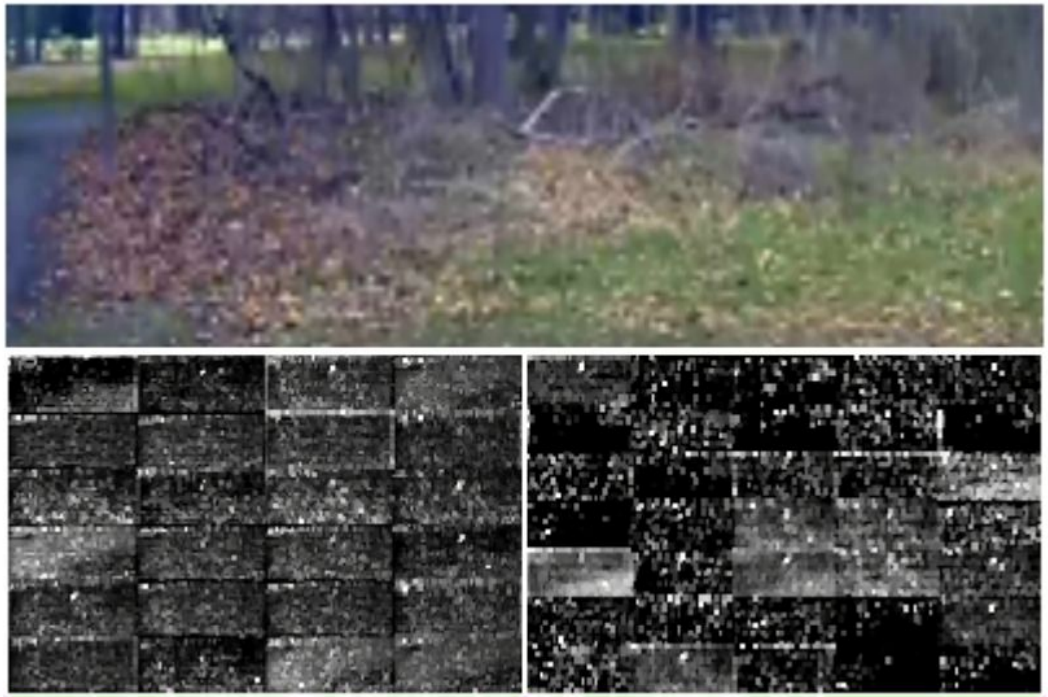
$$\text{autonomy} = \left(1 - \frac{(\text{number of interventions}) \cdot 6 \text{ seconds}}{\text{elapsed time [seconds]}}\right) \cdot 100$$

Evaluation: Feature visualization



- Detects outline of road
- Only training signal is human steering angle, never explicitly trained on these features

Evaluation: Feature visualization



- Features look like noise in forest seen
- Evidence that network learns driving-specific visual features and is not a generic edge or blob detector

Discussion

- Contrast with Dagger? Augmentation vs. Iterative expert labelling
- Existence proof that:
  - CNNs can learn road following without manual decomposition of the problem.
- Clever solution for augmenting human data
- Pros:
  - Simple, Bold
  - Feature visualization confirms driving related features are learned from sparse training signal
- Cons:
  - Not clear how to interpret 98% autonomy
  - No baselines
  - Strong claims about eventual better performance of end-to-end systems demand strong evidence