Representation Learning for Driving
Alex Kendall  @  CVPR, Long Beach, June 2019
Outline of talk

1. Recipe for success for representation learning
2. Strategy for training data
3. Interpretability & verification
Machine Learning for Autonomous Driving

Some Background
1989 ALVINN: End-to-End Imitation Learning

2016 NVIDIA: Lane Following on Highways

Urban driving with end-to-end machine learning

Inputs: camera video and a sat-nav

End-to-end deep learning

Uncertainty propagation from sensing to action

Outputs: driving commands

Alex Kendall et al. Learning to Drive in a Day. ICRA, 2019
A recipe for representing driving
The Self-Driving State Representation Today

- 3D Object Detection
- Semantic Segmentation
- Agent Prediction
- Turning indicator detector
- HD Map
- Driving Affordability Prediction
- Traffic sign detection

Autonomous Driving Representation
We can’t enumerate the information we need for every last edge case

Billions of dollars and 10 years of commercial resources can’t do it in a constrained environment like Phoenix, Arizona.
A recipe for a good representation

1. Needs to encode information that we believe is necessary (but not sufficient) for the task
   - For driving, this includes semantics, motion and geometry

2. Should also be optimised w.r.t. the end task
   - Therefore we need an end to end learning signal

3. The decision must be observable in the input data
   - We need the right sensor type and configuration

4. Our representation must have a very good signal to noise ratio
   - We must transform the signal into a compressed, nuisance free & invariant representation
Progression of computer vision from 2015... to 2018

Training Data

How much and what type do we need?
How much data do we need?

• It’s not the amount, but the type of data!
• Not all data is created equal
• Important you create a driving curriculum and can seek the right data to improve
  • Off-policy / dash cam data is not good enough!
  • Beneficial to have control over what data is collected
  • Probably need to have on-policy data
Driving data is exceptionally biased

- How much of the state space do we need to explore to learn a good representation?
- If we need training examples densely across all state space, human driving data is not sufficient
- But exploration is dangerous in the real world...
Can we train real-world models in simulated worlds?

- Zero shot sim2real
- Learn to project to a latent space for domain translation and control jointly
- Demonstrate this method can drive 3km+ on public UK roads

Alex Bewley et al. Learning to Drive from Simulation without Real World Labels. ICRA, 2019.
Learning to Drive from Simulation without Real World Labels

Reconstruction Loss
Cyclic Reconstruction Loss
Control Loss
Cyclic Control Loss

Not shown: adversarial LSGAN loss, latent reconstruction loss, perceptual loss.

Alex Bewley et al. Learning to Drive from Simulation without Real World Labels. ICRA, 2019.
## Comparison to Baseline Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Simulation MAE</th>
<th>Simulation Bal-MAE</th>
<th>Real MAE</th>
<th>Real Bal-MAE</th>
<th>DPI (metres)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drive-Straight</td>
<td>0.043</td>
<td>0.087</td>
<td>0.019</td>
<td>0.093</td>
<td>23†</td>
</tr>
<tr>
<td>Simple Transfer</td>
<td>0.055</td>
<td>0.056</td>
<td>0.265</td>
<td>0.272</td>
<td>9†</td>
</tr>
<tr>
<td>Real-to-Sim Translation</td>
<td>-</td>
<td>-</td>
<td>0.261</td>
<td>0.234</td>
<td>10†</td>
</tr>
<tr>
<td>Sim-to-Real Translation</td>
<td>-</td>
<td>-</td>
<td>0.059</td>
<td><strong>0.045</strong></td>
<td>28†</td>
</tr>
<tr>
<td>Latent Feature ADA [3]</td>
<td>0.040</td>
<td>0.047</td>
<td>0.032</td>
<td>0.071</td>
<td>15†</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>0.017</strong></td>
<td><strong>0.018</strong></td>
<td>0.081</td>
<td>0.087</td>
<td>&gt;3000</td>
</tr>
</tbody>
</table>

† Indicates results not collected during operation on a private rural road.
Model-Based Saliency

Suppose $f(\cdot)$ is our driving model and $m(\cdot)$ is our saliency model and $L(\cdot)$ is our loss function for the driving model and the operator $x \cdot m$ degrades the image with noise.

$$L = \lambda_1 |m(x)| + \lambda_2 |\nabla m(x)| + \lambda_3 L_0 \left( f \left( x \cdot m(x) \right) \right) + \lambda_4 L_0 \left( f \left( x \cdot (1 - m(x)) \right) \right)^{-\lambda_5}$$

Sparse saliency mask | Informative saliency mask
---|---
Smooth saliency mask | Uninformative inverse saliency mask

Model-Based Saliency

Suppose $f(\cdot)$ is our driving model and $m(\cdot)$ is our saliency model and $L(\cdot)$ is our loss function for the driving model and the operator $x \cdot m$ degrades the image with noise.

$$L = \lambda_1 |m(x)| + \lambda_2 |\nabla m(x)| + \lambda_3 L_0 \left(f(x \cdot m(x))\right) + \lambda_4 L_0 \left(f(x \cdot (1 - m(x)))\right)^{-\lambda_5}$$

Sparse saliency mask  Informative saliency mask

Smooth saliency mask  Uninformative inverse saliency mask

Inspecting the state for traffic light signal

$x_{input} \rightarrow z \rightarrow y_{control}$

Learn to decode the high dimensional state

$y_{aux}$
Inspecting the state for traffic light signal, semantics and depth
Which metrics do we optimise?

Must move away from component based verification
Improving individual components is no longer a proxy for improving system performance

• It assumes the interface between components is sufficient
• E.g. most KITTI metrics are at 90%+, does improving these metrics increase autonomous driving performance?
Mean Scenario Success

We need to consider complexity of autonomy, not just intervention metrics.
Conclusions
<table>
<thead>
<tr>
<th>Machine Learning</th>
<th>Human Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low engineering effort to create demo</td>
<td>Possible to enumerate all scenarios</td>
</tr>
<tr>
<td>Brittle representation</td>
<td>Analytical safety guarantees</td>
</tr>
<tr>
<td>No performance guarantees</td>
<td>Limited complexity</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Excels with increasing data and scale</td>
<td>Unachievable to identify all edge-cases</td>
</tr>
<tr>
<td></td>
<td>Too complex for safety guarantees</td>
</tr>
<tr>
<td></td>
<td>Requires extremely large engineering effort</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Constrained Setting</th>
<th>Open World</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validate with statistical evidence</td>
<td>Can learn powerful representations which generalise</td>
</tr>
</tbody>
</table>

Alex Kendall © Wayve 2019
Games like Go & DOTA

- Incredibly difficult action space: long term strategy, cooperation
- Very basic state space, often discrete, fully observable and noise-free

Autonomous Driving

- Quite easy action space: stop, go, left, right motion primitives
- Super challenging state space: manifold of natural images!

This needs to be solved by the computer vision community!
A complete paradigm shift for AVs

• Low vehicle compute and sensor requirements
• Large training compute and data requirements
• Increased vehicle intelligence
• No reliance on HD-maps
• Ability to leverage simulation for training
• Abundance of open and interesting research questions!

Come work with our team wayve.ai/careers