Deploying Deep Learning for Driving
Alex Kendall @ CVPR, Long Beach, June 2019

WAYVE
Outline of talk

1. Recipe for success for representation learning
2. Understanding what we don’t know
3. Strategy for training data
4. Interpretablility & verification
Machine Learning for Autonomous Driving

Some Historical 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
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.

Modelling Uncertainty

Understanding what we don’t know
What kind of uncertainty can we model?

**Epistemic uncertainty**
- Measures what you’re model doesn’t know
- Can be explained away by unlimited data

**Aleatoric uncertainty**
- Measures what you can’t understand from the data
- Can be explained away by unlimited sensing

---

Modeling Epistemic Uncertainty with Bayesian Deep Learning

- We can model epistemic uncertainty in deep learning models using Monte Carlo dropout sampling at test time.
- Dropout sampling can be interpreted as sampling from a distribution over models.

# Aleatoric Uncertainty with Probabilistic Deep Learning

<table>
<thead>
<tr>
<th></th>
<th>Deterministic Deep Learning</th>
<th>Probabilistic Deep Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model</strong></td>
<td>([\hat{y}] = f(x))</td>
<td>([\hat{y}, \hat{\sigma}^2] = f(x))</td>
</tr>
<tr>
<td><strong>Regression</strong></td>
<td>(Loss = |y - \hat{y}|^2)</td>
<td>(Loss = \frac{|y - \hat{y}|^2}{2\hat{\sigma}^2} + \log \hat{\sigma})</td>
</tr>
<tr>
<td><strong>Classification</strong></td>
<td>(Loss = \text{SoftmaxCrossEntropy}(\hat{y}_t))</td>
<td>(\hat{y}_t = \hat{y} + \epsilon_t, \quad \epsilon_t \sim N(0, \hat{\sigma}^2))</td>
</tr>
</tbody>
</table>

\[
Loss = \frac{1}{T} \sum_t \text{SoftmaxCrossEntropy}(\hat{y}_t)
\]

Train/Test Distribution Shift

- Aleatoric uncertainty remains constant while epistemic uncertainty increases for out of dataset examples!

<table>
<thead>
<tr>
<th>Train dataset</th>
<th>Test dataset</th>
<th>RMS</th>
<th>Aleatoric variance</th>
<th>Epistemic variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Make3D / 4</td>
<td>Make3D</td>
<td>5.76</td>
<td>0.506</td>
<td>7.73</td>
</tr>
<tr>
<td>Make3D / 2</td>
<td>Make3D</td>
<td>4.62</td>
<td>0.521</td>
<td>4.38</td>
</tr>
<tr>
<td>Make3D</td>
<td>Make3D</td>
<td>3.87</td>
<td>0.485</td>
<td>2.78</td>
</tr>
<tr>
<td>Make3D / 4</td>
<td>NYUv2</td>
<td>-</td>
<td>0.388</td>
<td>15.0</td>
</tr>
<tr>
<td>Make3D</td>
<td>NYUv2</td>
<td>-</td>
<td>0.461</td>
<td>4.87</td>
</tr>
</tbody>
</table>

Qualitative comparison

- Epistemic uncertainty is modeling uncertainty
- Aleatoric uncertainty is sensing uncertainty
Bayesian Deep Learning for Segmentation

Input Image  Semantic Segmentation  Uncertainty

Alex Kendall et al. Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. BMVC 2017
Bayesian Deep Learning for Stereo Vision

Training Data

How much and what type do we need?
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

(a) Reconstruction loss  (b) Cyclic recon. loss  (c) Control loss  (d) Cyclic control loss

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.

\[ X_d^{recon} = G_d(E_d(X_d)) \]
\[ X_d^{cyc} = G_d(E_d(G_d'(E_d(X_d)))) \]
\[ \hat{c} = C(E_d(X_d)) \]
\[ \hat{c}^{cyc} = C(E_d(G_d'(E_d(X_d)))) \]
### Comparison to Baseline Methods

<table>
<thead>
<tr>
<th></th>
<th>Simulation</th>
<th></th>
<th></th>
<th>Real</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>Bal-MAE</td>
<td>MAE</td>
<td>Bal-MAE</td>
<td>DPI (metres)</td>
</tr>
<tr>
<td>Drive-Straight</td>
<td>0.043</td>
<td>0.087</td>
<td><strong>0.019</strong></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><strong>Ours</strong></td>
<td><strong>0.017</strong></td>
<td><strong>0.018</strong></td>
<td><strong>0.081</strong></td>
<td><strong>0.087</strong></td>
<td><strong>&gt;3000</strong></td>
</tr>
</tbody>
</table>

† Auxiliary ground truth information can be provided with ease, and the vehicle can be put into situations that are difficult or dangerous to undertake in reality. Previously, with the substantial gap in complexity between the two domains, it was considered infeasible to transfer driving policies from simulation to the real world without a considerable additional investment in data gathering. This work provides evidence that end-to-end policy learning for autonomous driving systems.

Learning a driving policy from simulation has many advantages: training data is cheap, demonstrations in the real world driving scenarios are available, and auxiliary ground truth information can be provided with ease, and the vehicle can be put into situations that are difficult or dangerous to undertake in reality. Previously, with the substantial gap in complexity between the two domains, it was considered infeasible to transfer driving policies from simulation to the real world without a considerable additional investment in data gathering. This work provides evidence that end-to-end policy learning for autonomous driving systems.
Interpreting & Understanding Deep Learning Representations
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.

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

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

Inspecting the state for traffic light signal

\[ x_{\text{input}} \rightarrow z \rightarrow y_{\text{control}} \]

Learn to decode the high dimensional state

\[ y_{\text{aux}} \]
Inspecting the state for traffic light signal, semantics and depth
Conclusions
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