Direct Perception for Autonomous Driving

Chenyi Chen
DeepDriving

• http://deepdriving.cs.princeton.edu/
Direct Perception

CNN

estimate

control
Direct Perception (in Driving...)

- Ordinary car detection: Find a car! It’s localized in the image by a red bounding box.
- Direct perception: The car is in the right lane; 16 meters ahead
- Which one is more helpful for driving task?
Machine Learning
What’s machine learning?

• Example problem: face recognition
  
  ![Prof. K](image1.png)  ![Prof. F](image2.png)  ![Prof. P](image3.png)  ![Prof. V](image4.png)  ![Chenyi](image5.png)

• Training data: a collection of images and labels (names)
  
  ![Who is this guy?](image6.png)

• Evaluation criterion: correct labeling of new images
What’s machine learning?

- Example problem: scene classification

  - a few labeled training images

  - goal to label yet unseen image

  What’s the label of this image?
Supervised Learning

• System input: X
• System model: \( f_\theta() \), with parameter \( \theta \)
• System output: \( \hat{Y} = f_\theta(X) \),
• Ground truth label: Y
• Loss function: \( L(\theta) \)
• Learning rate: \( \gamma \)
Why machine learning?

• The world is very complicated
• We don’t know the exact model/mechanism between input and output
• Find an approximate (usually simplified) model between input and output through learning
Deep Learning: A sub-field of machine learning
Neural Networks

basic building blocks

\[ z = \sum_{i} x_i w_i + b, \quad y = f(z) \]

where \( f \) is an activation function:

\[ f(z) = \sigma(z) = \frac{1}{1 + \exp(-z)} \]

- sigmoid is bounded between 0 and 1
- monotonically increasing
- differentiation: \( \sigma'(z) = \sigma(z) \times (1 - \sigma(z)) \)
Neural Networks

representation

feed-forward

for each neuron in the next layer:

\[
z_i^{(2)} = \sum_{j=1}^{n} w_{ij}^{(1)} x_j + b_i^{(1)}, \quad a_i^{(2)} = f(z_i^{(2)})
\]

\[
z_1^{(2)} = w_{11}^{(1)} x_1 + w_{21}^{(1)} x_2 + w_{31}^{(1)} x_3 + b_1^{(1)}, \quad a_1^{(2)} = f(z_1^{(2)})
\]

compactly:

\[
z^{(2)} = W^{(1)} x + b^{(1)}, \quad a^{(2)} = f(z^{(2)})
\]

\[
z^{(3)} = W^{(2)} x + b^{(2)}, \quad a^{(3)} = f(z^{(3)})
\]

layer\_1 \quad \text{layer}\_2 \quad \text{layer}\_3

globally models a function:

\[
\hat{y} = h_{W,b}(x)
\]

where \(W\) and \(b\) are model parameters
Convolutional Neural Networks

Detection layers

Convolution

\[ Z_i = \sigma(W_i \ast X) \]
Convolutional Neural Networks

pooling:
- reduce the size of representations
- allow small translation invariance.

common pooling techniques:
max pooling, average pooling

figure from roger grosse tutorial
Convolutional Neural Networks

convolution network is just a combination of convolution layers, pooling layers and fully connected layers

(figure from LeNet tutorial, http://deeplearning.net/tutorial/lenet.html)
Why deep learning?

How do we detect a stop sign? It’s all about feature!
Why deep learning?

How does computer vision algorithm work? It’s all about feature!

Pedestrian found!!!
Why deep learning?

- We believe driving is also related with certain features
- Those features determine what action to take next
- Salient features can be automatically detected and processed by deep learning algorithm
Deep Convolutional Neural Network (CNN)

Convolutional layers

Fully connected layers

Parallel computing

Input image, 280*210*3, int8 (1 byte)

Deep feature, 4096D, float (4 bytes)

Output, 14D, float (4 bytes)

Figure courtesy and modified of Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton
How does CNN work?

- Input image convolves with a large set of filters, and this process is repeated in many layers.
- Compared to traditional computer vision feature extraction method, e.g. SIFT, HOG, more powerful feature representation of the input image is learnt automatically through convolutions.

A visualized example of the set of filters
Why deep learning?

- ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012), the most famous challenge in computer vision field
  - 1000 categories
  - 1.2 million training images
  - 50,000 validation images
  - 150,000 testing images
  - Top-5 error rate* of deep learning: 15.3%
  - Top-5 error rate of second best (which is non-deep learning): 26.2%

*Top-5 error rate*: the fraction of test images for which the correct label is not among the five labels considered most probable by the model
## Deep Learning

### Result Page of ILSVRC2012 Website

#### Task 1

<table>
<thead>
<tr>
<th>Team name</th>
<th>Filename</th>
<th>Error (5 guesses)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SuperVision</td>
<td>test-preds-141-146.2009-131-137-146-140.2011-145f.</td>
<td>0.15315</td>
<td>Using extra training data from ImageNet Fall 2011 release</td>
</tr>
<tr>
<td>SuperVision</td>
<td>test-preds-131-137-146-136-145f.txt</td>
<td>0.16422</td>
<td>Using only supplied training data</td>
</tr>
<tr>
<td>ISI</td>
<td>pred_FVs_wLACs_weighted.txt</td>
<td>0.26172</td>
<td>Weighted sum of scores from each classifier with SIFT+FV, LBP+FV, GIST+FV, and CSIFT+FV, respectively.</td>
</tr>
<tr>
<td>ISI</td>
<td>pred_FVs_weighted.txt</td>
<td>0.26592</td>
<td>Weighted sum of scores from classifiers using each FV.</td>
</tr>
<tr>
<td>ISI</td>
<td>pred_FVs_summed.txt</td>
<td>0.26646</td>
<td>Naive sum of scores from classifiers using each FV.</td>
</tr>
<tr>
<td>ISI</td>
<td>pred_FVs_wLACs_summed.txt</td>
<td>0.26962</td>
<td>Naive sum of scores from each classifier with SIFT+FV, LBP+FV, GIST+FV, and CSIFT+FV, respectively.</td>
</tr>
<tr>
<td>OXFORD_VGG</td>
<td>test_techmix_classification_text.txt</td>
<td>0.26979</td>
<td>Mixed selection from High-Level SVM scores and Baseline Scores, decision is performed by looking at the validation performance</td>
</tr>
<tr>
<td>XRCE/INRIA</td>
<td>res_1M_svm.txt</td>
<td>0.27068</td>
<td></td>
</tr>
</tbody>
</table>
## Task 2a: Classification + localization with provided training data

Classification + localization with provided training data: Ordered by classification error

<table>
<thead>
<tr>
<th>Team name</th>
<th>Entry description</th>
<th>Classification error</th>
<th>Localization error</th>
</tr>
</thead>
<tbody>
<tr>
<td>GoogleNet</td>
<td>No localization. Top 5 val score is 8.66% error</td>
<td>0.06656</td>
<td>0.606257</td>
</tr>
<tr>
<td>VGG</td>
<td>a combination of multiple ConvNets, including a net trained on images of different size (fusion weights learnt on the validation set); detected boxes were not updated</td>
<td>0.07325</td>
<td>0.256167</td>
</tr>
<tr>
<td>VGG</td>
<td>a combination of multiple ConvNets, including a net trained on images of different size (fusion done by averaging); detected boxes were not updated</td>
<td>0.07337</td>
<td>0.255431</td>
</tr>
<tr>
<td>VGG</td>
<td>a combination of multiple ConvNets (by averaging)</td>
<td>0.07405</td>
<td>0.253231</td>
</tr>
<tr>
<td>VGG</td>
<td>a combination of multiple ConvNets (fusion weights learnt on the validation set)</td>
<td>0.07407</td>
<td>0.253501</td>
</tr>
<tr>
<td>MSRA Visual Computing</td>
<td>Multiple SPP-nets further tuned on validation set (B)</td>
<td>0.0806</td>
<td>0.354924</td>
</tr>
<tr>
<td>MSRA Visual Computing</td>
<td>Multiple SPP-nets further tuned on validation set (A)</td>
<td>0.08062</td>
<td>0.354799</td>
</tr>
<tr>
<td>Andrew Howard</td>
<td>Combination of Convolutional Nets with Validation set adaptation + KNN</td>
<td>0.08111</td>
<td>0.610365</td>
</tr>
<tr>
<td>MSRA Visual Computing</td>
<td>Multiple SPP-nets (B)</td>
<td>0.082</td>
<td>0.355568</td>
</tr>
<tr>
<td>Andrew Howard</td>
<td>Combination of Convolutional Nets with Validation set adaptation</td>
<td>0.08226</td>
<td>0.611919</td>
</tr>
<tr>
<td>MSRA Visual Computing</td>
<td>Multiple SPP-nets (A)</td>
<td>0.08307</td>
<td>0.3562</td>
</tr>
<tr>
<td>VGG</td>
<td>a single ConvNet (13 convolutional and 3 fully-connected layers)</td>
<td>0.08434</td>
<td>0.267184</td>
</tr>
<tr>
<td>Andrew Howard</td>
<td>Combination of Convolutional Nets + KNN</td>
<td>0.0853</td>
<td>0.612305</td>
</tr>
<tr>
<td>Andrew Howard</td>
<td>Baseline Combination of Convolutional Nets</td>
<td>0.08919</td>
<td>0.614307</td>
</tr>
<tr>
<td>MSRA Visual Computing</td>
<td>A single SPP-net</td>
<td>0.09079</td>
<td>0.36118</td>
</tr>
<tr>
<td>DeeperVision</td>
<td>Simple average ensemble</td>
<td>0.09503</td>
<td>1.0</td>
</tr>
<tr>
<td>DeeperVision</td>
<td>Simple average ensemble and box</td>
<td>0.09503</td>
<td>0.842953</td>
</tr>
<tr>
<td>DeeperVision</td>
<td>Weighted ensemble</td>
<td>0.09558</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Result Page of ILSVRC2014 Website

Task 1b: Object detection with additional training data

Ordered by mean average precision

<table>
<thead>
<tr>
<th>Team name</th>
<th>Entry description</th>
<th>Description of outside data used</th>
<th>mean AP</th>
<th>Number of object categories won</th>
</tr>
</thead>
<tbody>
<tr>
<td>GoogLeNet</td>
<td>Ensemble of detection models. Validation is 44.5% mAP</td>
<td>Pretraining on ILSVRC12 classification data</td>
<td>0.439329</td>
<td>142</td>
</tr>
<tr>
<td>CUHK DeepID-Net</td>
<td>Combine multiple models described in the abstract without contextual modeling. The training data includes the validation dataset 2.</td>
<td>ImageNet classification and localization data</td>
<td>0.406998</td>
<td>---</td>
</tr>
<tr>
<td>CUHK DeepID-Net</td>
<td>Combine multiple models described in the abstract without contextual modeling.</td>
<td>ImageNet classification and localization data</td>
<td>0.406592</td>
<td>29</td>
</tr>
<tr>
<td>DeepInsight</td>
<td>Combination of three detection models</td>
<td>Three CNNs from classification task are used for initialization.</td>
<td>0.404517</td>
<td>27</td>
</tr>
<tr>
<td>CUHK DeepID-Net2</td>
<td>Combine multiple models described in the abstract without contextual modeling. The training data includes the validation dataset 2.</td>
<td>ImageNet classification and localization data</td>
<td>0.40352</td>
<td>---</td>
</tr>
<tr>
<td>CUHK DeepID-Net2</td>
<td>Combine multiple models described in the abstract without contextual modeling.</td>
<td>ImageNet classification and localization data</td>
<td>0.403417</td>
<td>---</td>
</tr>
<tr>
<td>DeepInsight</td>
<td>A single detection model</td>
<td>A CNN from classification task is used for initialization.</td>
<td>0.401568</td>
<td>---</td>
</tr>
<tr>
<td>DeepInsight</td>
<td>Another single detection model</td>
<td>A CNN from classification task is used for initialization.</td>
<td>0.396982</td>
<td>---</td>
</tr>
<tr>
<td>GoogLeNet</td>
<td>Single detection model. Validation is 38.75% mAP</td>
<td>Pretraining on ILSVRC12 classification data</td>
<td>0.380277</td>
<td>---</td>
</tr>
<tr>
<td>CUHK DeepID-Net2</td>
<td>Multi-stage deep CNN without contextual modeling</td>
<td>ImageNet classification and localization data</td>
<td>0.377471</td>
<td>---</td>
</tr>
<tr>
<td>UVA-EuVision</td>
<td>Deep learning with outside data</td>
<td>ImageNet 1000</td>
<td>0.354213</td>
<td>1</td>
</tr>
<tr>
<td>CUHK DeepID-Net</td>
<td>A single deep CNN with deformation layers</td>
<td>ImageNet classification and localization data</td>
<td>0.340389</td>
<td>---</td>
</tr>
</tbody>
</table>
Why deep learning?

• Deep learning first impacted the computer vision field in ILSVRC2012
• Now it’s almost dominating the computer vision field
• It only takes less than three years!
How does our system work?
Basic Ideas

• Extract key parameters from driving scenes images with deep learning CNN

• Compute driving control (optimal control) based on those parameters
In our specific case...

- Let the deep learning algorithm tell us:
  - **angle**: the angle between the car’s heading and the tangent of the track;
  - **toMarking**: the distance between the center of the car and each lane marking;
  - **dist**: the distance between the car and the preceding car in each lane;
- ... 14 parameters in total
In our specific case...

- Illustration of key parameters for car pose and localization
The cases we are dealing with

- Monitoring current lane and left & right neighboring lanes
In our experiment

- Let the deep learning CNN drive in a racing game -- TORCS
How to train the system?
Seven Training Tracks
Multiple Asphalt Darkness Level for the Training Tracks
22 Training Cars
CNN during training phase

- Supervised learning, ground truth extracted from the game engine
How to run the system?
CNN during testing phase

• Unseen track with unseen cars
How does the successfully trained system drive a car?

- The system runs at 10Hz, real time
A Challenging Test: Perception during Night Driving
host car estimation  traffic car estimation
Next step: incorporating temporal information
Other Interesting Stuffs
A Taste of Stereo Vision
3D Scene Reconstruction

- Stereo images
- Color
- 1382*512
- 10 FPS
Visual Odometry

- Visual odometry computes the trajectory of the vehicle only based on image sequences
Depth Map

- Disparity map is computed from grayscale stereo image pairs
- Depth map can be derived from disparity map and camera model
Other Demos for Structure-from-Motion

• https://www.youtube.com/watch?v=i7ierVkXYa8
• https://www.youtube.com/watch?v=vpTEobpYoTg
Other Demos for Structure-from-Motion
Other Demos for Structure-from-Motion
Q & A