What types of Intelligences are needed for Automated Vehicles?

- **"Semantic" interpretation of vehicle’s environment:**
  - Detect and categorize/recognize objects (cars, pedestrians, bicycles, traffic signs, traffic lights, ...)
  - Ego-localization
  - **Predict movements** of other road users
  - **Infer intentions** of other drivers and pedestrians (or policeman!) from their movements/expressions

- **Planning of trajectories** (including speed)
  In a dynamic and uncertain environment

- **Coordinated/cooperative planning** of multiple vehicles

- For Advanced Driving Assistance Systems (ADAS) and partial automated driving (level 3-4):
  - Analyze and understand attention and activities or gestures of the "driver-supervisor"
Outline

• What can Deep-Learning perform with images?
  • Visual Object detection & Semantic Segmentation
  • Image-based ego-localization
  • Human posture and movement analysis

Image-based Deep-Learning

• Image classification
• Visual object detection and categorization
• Semantic segmentation of images
• Realistic image synthesis
• Image-based localization
• Estimation of Human pose
• Inference of 3D (depth) from monocular vision
• Learning image-based behaviors
  • End-to-end driving from front camera
  • Learning robot behavior from demonstration/imitation
Visual objects Simultaneous Detection and Categorization with Faster_RCNN

Beyond bounding-boxes: getting contours of objects

Mask R-CNN extracts detailed contours and shapes of objects instead of just bounding-boxes
Example result of semantic segmentation by Deep-Learning


Semantic segmentation provides category information also for large regions (not only individualized « objects ») such as « road », « building », etc...

"Video-to-Video Synthesis", NeurIPS’2018 [Nvidia+MIT]
Using Generative Adversarial Network (GAN)
PoseNet: 6-DoF camera-pose regression with Deep-Learning

[Image of PoseNet diagram]

Input RGB Image → Convolutional Neural Network (GoogLeNet) → 6-DOF Camera Pose

Figure 4: Map of dataset showing training frames (green), testing frames (blue) and their predicted camera pose (red). The testing sequences are distinct trajectories from the training sequences and each scene covers a very large spatial extent.


Human pose estimation by Deep-Learning

Real-time estimation of Human poses on RGB video

[Realtime Multi-Person 2D Pose Estimation using Part Affinity Field, Cao et al., CVPR’2017 [CMU]]
**Inference of 3D (depth) from monocular vision**

Unsupervised monocular depth estimation with left-right consistency

*C Godard, O Mac Aodha, GJ Brostow* - CVPR’2017 [UCL]

End-to-end driving from camera by Deep-Learning

ConvNet input: Cylindrical projection of fisheye camera

ConvNet output: steering angle

Imitation Learning from Human driving on real data

End-to-end driving via Deep **Reinforcement** Learning

[thèse CIFRE Valeo/MINES-ParisTech en cours]
Outline

• What can Deep-Learning perform with images?
• Visual Object detection & Semantic Segmentation
• Image-based ego-localization
• Human posture and movement analysis

Visual objects Detection and Categorization: Faster RCNN

Region Proposal Network (RPN) on top of standard convNet.
End-to-end training with combination of 4 losses
Solve detection as a regression problem
(“single-shot” detection)

YOLO and SSD

YOLO: YOU ONLY LOOK ONCE
SSD: SINGLE SHOT MULTIBOX DETECTOR

Images from: https://www.slideshare.net/TaegyunJeon1/pr12-you-only-look-once-yolo-unified-realtime-object-detection

Both are faster, but less accurate, than Faster_R-CNN

Recent comparison of object detection convNets

- Faster R-CNN w/ResNet, Hi Res, 50 Proposals
- R-FCN w/ResNet, Hi Res, 100 Proposals
- Meta Architecture
  - Faster RCNN
  - R-FCN
  - SSD
  - Faster R-CNN w/Inception Resnet, Hi Res, 300 Proposals, Stride 8

YOLOv2 (different impl.; not 100% comparable)

SSD w/inception V2, Lo Res
SSD w/MobileNet, Lo Res

Feature Extractor
- Inception Resnet V2
- Inception V2
- Inception V3
- MobileNet
- Resnet 101
- VGG

Overall mAP vs. GPU Time

Slide from Ross Girshick's CVPR 2017 Tutorial, Original Figure from Huang et al
Mask R-CNN architecture (left) extracts detailed contours and shape of objects instead of just bounding-boxes.

Deep-Learning approach for semantic segmentation.
Convolutional Encoder-Decoder

Feature extractor

Shape generator

SegNet example results
Outline

- What can Deep-Learning perform with images?
- Visual Object detection & Semantic Segmentation
- Image-based ego-localization
- Human posture and movement analysis

PoseNet: 6-DoF camera pose regression with Deep-Learning

Trained with a naïve end-to-end loss function to regress camera position, $x$, and orientation, $q$

$$\text{loss}(l) = \|x - \hat{x}\|_2 + \beta \left\| q - \frac{\hat{q}}{||\hat{q}||} \right\|_2$$

PoseNet training data and test results

training data in green, test data in blue, PoseNet results in red


PoseNet results on other tests

Figure 4: Map of dataset showing training frames (green), testing frames (blue) and their predicted camera pose (red). The testing sequences are distinct trajectories from the training sequences and each scene covers a very large spatial extent.
PoseNet robustness

Tolerance to environment, unknown intrinsics, weather, etc.

Blur | Occlusion | Dusk | Night

PoseNet summary:
- Robust to scene change + very fast
- Robust to lighting, weather, dynamic objects
- Fast inference, <2ms per image on Titan GPU
- Scale not dependent on number of training images

Coarse accuracy
Difficult to learn both position vs orientation

PoseNet vs traditional methods

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>King’s College</td>
<td>0.88m, 1.04°</td>
<td>0.42m, 0.55°</td>
</tr>
<tr>
<td>Resolution</td>
<td>256 x 256 px</td>
<td>1920 x 1080 px</td>
</tr>
<tr>
<td>Inference Time</td>
<td>2 ms</td>
<td>78 ms</td>
</tr>
</tbody>
</table>

PoseNet less precise, but much faster and can work with much smaller images

Deep-Learning pose regression from GIS images

- Learn an only 3-DoF pose \((x, y, \theta)\)
- Start transfer learning from InceptionV3 model modified as follows:
  - final classifier replaced by a dropout layer
  - fully connected layer with 256 neurons added and connected to final 3-dimension pose regressor
- Use StreetView “augmented” with virtual views added 4m after each geo-tagged panorama

Work by Dr Li YU during his PhD thesis @ VeDeCom-MINES_ParisTech (defended in Apr.2018)
Results of Deep-Learning visual localization trained on GIS images

Results:

- Average error of 7.62m
- 54.2% within a 4m error

GIS-trained adapted PoseNet vs. Coarse-to-fine image matching

- Handcrafted feature method (2x) more accurate + smooth positions
- BUT convNet based method much faster to compute, and still reaches accuracy of a standard GPS.
Outline

• What can Deep-Learning perform with images?
• Visual Object detection & Semantic Segmentation
• Image-based ego-localization
• Human posture and movement analysis

Automated Vehicles
interactions with Humans

Need to monitor and interpret Human movements, actions & activities:
– Inference of Human intentions (pedestrians and drivers) for Automated Vehicles
– Gestual communication with Humans
Pose estimation now possible from RGB camera (openPose)

Deep-Learning for time-series

Two main approaches:
- Deep Recurrent Neural Network (RNN) e.g. LSTM or GRU
- Temporal Convolutions

"Convolutional Neural Networks for Multivariate Time Series Classification using both Inter- and Intra-Channel Parallel Convolutions", G. Devineau, W. Xi, F. Moutarde and J. Yang, RFIAP’2018.

[PhD thesis of Guillaume Devineau @ MINES_ParisTech, supervised by me]
**Combining DL pose estimation + Deep Temporal Convolution (or/and RNN)?**

---

**Camera**

- **DL pose estimation** (openPose/alphaPose)

---

**Deep Temporal Convolution (or/and Deep RNN?) for Multivariate Time-Series**

---

**Work in Progress (PhD thesis of Salwa El Kaddaoui at MINES_ParisTech, within H2020 European project COLLABORATE**

---

**Conclusions**

- **Deep Convolutional Neural Networks** already can perform many more things than just image classification: semantic segmentation, localization from vision, estimation of Human pose, inference of depth from monovision, generation of realistic synthetic images, and learning complex image-based adaptive behaviors.

- The above can be leveraged for many AI challenges for Automated Vehicles:
  - image-based ego-localization by convNet
  - for Human movements or intents analysis, combining human-pose estimation by DL with Deep Temporal Convolution of time-series seems promising
  - adaptive behavior learning as an image-based end-to-end driving task [NEXT DECK OF SLIDES]
Questions ?