Deep-Learning for Robotics & Autonomous Vehicles

Pr. Fabien Moutarde
Centre de Robotique
MINES ParisTech
PSL Université

Fabien.Moutarde@mines-paristech.fr
http://people.mines-paristech.fr/fabien.moutarde

Outline

• Introduction: Artificial Intelligences & Machine-Learning
• AIs for robotics & Autonomous Vehicles
• What can Deep-Learning perform with images?
• Recognition of Gestures/Actions for Human-Robot Collaboration
• Imitation Learning & Deep Reinforcement Learning for Autonomous Driving and design of Robots behavior
What is (human) intelligence??

• Intelligence = reasoning? or Intelligence = adaptation?
• In fact, MANY DIFFERENT TYPES OF INTELLIGENCE

A possible typology:
– Perception Intelligence
– Prediction Intelligence
– Reasoning Intelligence
– Behavior Intelligence
– Interaction Intelligence
– Curiosity

What is AI?

Artificial Intelligence, a vast and heterogeneous domain:
– Rule-based reasoning, expert systems
– Algorithms for playing games (chess, Go, etc..)
– Multi-agents, emergence of collective behavior
– ...
– Optimization, Operational Research, Dynamic Programming
– Planning (of trajectories, tasks, etc...)
– Computer vision, pattern recognition
– Machine-Learning = empirical data-driven modelling
  (optimization, based on examples, of a parametric model)
Artificial Intelligence (AI)

Outline

- Introduction: Artificial Intelligences & Machine-Learning
- AIs for robotics & Autonomous Vehicles
  - What can Deep-Learning perform with images?
  - Recognition of Gestures/Actions for Human-Robot Collaboration
  - Imitation Learning & Deep Reinforcement Learning for Autonomous Driving and design of Robots behavior
"Traditional" (industrial) Robots

Repetitive actions, fast, strong, …
BUT dangerous and NOT VERY ADAPTIVE
(simple "automatons")

"Intelligents" robots
(≈ adaptive and/or interactive)

React adaptively to environment…

… and/or interact with Humans
General principle of robots

Perception → Reasoning & decision

ROBOTIC LOOP

Action

Autonomous Vehicles are mobile robots!
What types of AIs are needed for Autonomous Vehicles?

- "Semantic" interprétation 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/gestures/gazes

- **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"

Intelligent Perception for Autonomous Vehicles

Essential need: real-time "understanding" of surroundings

From camera

From LIDAR

Strong real-time constraint: process $\geq$ 15 frames/seconde
What types of "intelligence" do ROBOTS need?

- **Analyze & interpret** a dynamic environment
  - Recognize a place & ego-localize
  - Detect/localize & categorize "objects"
  - Track & predict their movements
  - Guess "intentions"
- **Choose** action/movement to be performed
  - Decision logics
- **Adapt/optimize** chosen action/movement
  - Having a BEHAVIOR rather than rigid rules
- **Interact** with humans or other robots
  - Speech Recognition
  - Natural Language Processing, ability to dialog
  - Recognition of Gestures/Actions, of emotions?
  - Coordination/collaboration between robots

Intelligent Perception for Collaborative Robots

Strong need: monitoring and interpreting movements, actions & activities of Humans around

Action recognition for Human-Robot Collaboration

[centre de Robotique de MINES ParisTech, Chaire PSA "Robotique et Réalité Virtuelle"]
Major challenges for Intelligent Robots & Autonomous Vehicles

- Inference of INTENTIONS of Humans
- Human activity understanding
- Learning of adaptive BEHAVIOR
  - Learning by demonstration/imitation
  - Learning by reinforcement
  - Abstraction of task rather than recording of trajectory
  - One/few shot(s) learning
- Coordination/collaboration
  - between robots (cooperative planning, etc...)
  - with Humans:
    - Non-verbal communication (gestures, movement, gaze)
    - Learning of implicit "social rules"

Coordination with Humans: "human-aware" AI

Challenge: learn implicit "social rules" of interaction
Outline

• Introduction: Artificial Intelligences & Machine-Learning
• AIs for robotics & Autonomous Vehicles
• What can Deep-Learning perform with images?
  • Recognition of Gestures/Actions for Human-Robot Collaboration
  • Imitation Learning & Deep Reinforcement Learning for Autonomous Driving and design of Robots behavior

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
Visually detect & categorize objects

Visual objects Simultaneous Detection and Categorization with Faster_RCNN

Beyond bounding-boxes: getting \textit{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


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

[Realtme Multi-Person 2D Pose Estimation using Part Affinity Field, Cao et al., CVPR'2017 [CMU]]

OpenPose
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]

Autonomous learning of a task by a robot

Robot autonomously learns bin picking without human instruction

Supervised learning, but with success/failure easily estimated automatically, for a bin-picking task
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]

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
Object visual detection without proposal

Solve detection as a regression problem
("single-shot" detection)

YOLO and SSD

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

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.
Deep Learning for Robotics & Autonomous Vehicles, Pr. Fabien Moutarde, Centre de Robotique, MINES ParisTech, PSL, 29/1/2019

Convolutional Encoder-Decoder

Outline

- Introduction: Artificial Intelligences & Machine-Learning
- AIs for robotics & Autonomous Vehicles
- What can Deep-Learning perform with images?
- Recognition of Gestures/Actions for Human-Robot Collaboration
- Imitation Learning & Deep Reinforcement Learning for Autonomous Driving and design of Robots behavior
Human-Robot/Autonomous Vehicle interactions

Need to monitor and interpret Human movements, actions & activities:

– Action recognition for collaborative robots
– Inference of Human intentions (pedestrians and drivers) for Autonomous Vehicles
– Gestual communication with Humans for both
Deep-Learning for Robotics & Autonomous Vehicles, Pr. Fabien Moutarde, Centre de Robotique, MINES ParisTech, PSL, 29/1/2019  37

Example of Action Recognition
for HRC

Trajectories
of hands

PhD thesis of Eva Coupête at MINES_Paris (defended in 2016),
sponsored by Chaire PSA "Robotique et Réalité Virtuelle"

Deep-Learning for Robotics & Autonomous Vehicles, Pr. Fabien Moutarde, Centre de Robotique, MINES ParisTech, PSL, 29/1/2019  38

Example of Action Recognition result

Action recognition for Human-Robot Collaboration
[centre de Robotique de MINES ParisTech, Chaire PSA "Robotique et Réalité Virtuelle"]
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

Recognized action/gesture

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

Outline

• Introduction: Artificial Intelligences & Machine-Learning
• AIs for robotics & Autonomous Vehicles
• What can Deep-Learning perform with images?
• Recognition of Gestures/Actions for Human-Robot Collaboration
• Imitation Learning & Deep Reinforcement Learning for Autonomous Driving and design of Robots behavior
Idea of end-to-end driving

Capteurs (caméras, lidars, ultrasons...) ➔ Perception ➔ Planification ➔ Contrôle ➔ Angle volant, Accélération, frein

Capteurs (caméras, lidars, ultrasons...) ➔ Apprentissage par imitation, Renforcement ➔ Angle volant, Accélération, frein

Imitation Learning:
"copying" human driver

End-to-end driving convNet

ConvNet output: steering angle

ConvNet input: Cylindrical projection of fisheye camera

Real data + “simulator” with real images

Training+testing dataset = **10000 km** and **200 hours** of human driving in openroad (highways, urban streets, country roads, etc…) under various weather conditions

TrainSet = 10 million images, TestSet = 3 million images.

“ConvNet in-the-loop” simulator with real images
**End-to-end driving: closed loop evaluation**

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Urban Aut. (%)</th>
<th>Urban MAD (cm)</th>
<th>Highways Aut. (%)</th>
<th>Highways MAD (cm)</th>
<th>Sharp turns Aut. (%)</th>
<th>Sharp turns MAD (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>99.3</td>
<td>16</td>
<td>98.7</td>
<td>19</td>
<td>73.7</td>
<td>30</td>
</tr>
<tr>
<td>Sel. #1</td>
<td>98.9</td>
<td>15</td>
<td>97.7</td>
<td>25</td>
<td>83.7</td>
<td>27</td>
</tr>
<tr>
<td>Sel. #2</td>
<td>99.5</td>
<td>16</td>
<td>97.2</td>
<td>24</td>
<td>87.5</td>
<td>28</td>
</tr>
<tr>
<td>Oversamp.</td>
<td>98</td>
<td>18</td>
<td><strong>91.8</strong></td>
<td><strong>29</strong></td>
<td>82.5</td>
<td>29</td>
</tr>
<tr>
<td>Baseline</td>
<td>8</td>
<td>36</td>
<td>14</td>
<td>41</td>
<td>0</td>
<td>35</td>
</tr>
</tbody>
</table>

**Real vehicle end-to-end driving (learnt by imitation)**

[Work by Valeo using ConvNet trained by my CIFRE PhD student Marin Toromanoff]
Demonstration at CES’2018

[Work by Valeo using ConvNet trained by my CIFRE PhD student Marin Toromanoff]


Transferability from real-world to simulator

Test of driving convNet in GTA simulator

Note that learning was done only on real-world data (by human driving imitation)

[Work by my Valeo CIFRE PhD student Marin Toromanoff]
Reinforcement Learning (RL)

**Goal:** find a “policy” $a_t = \pi(s_t)$ that

Maximizes $R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$, $\gamma \in [0, 1[$

**Deep Reinforcement Learning (DRL)** if Deep NeuralNet used as model (for policy and/or its “value”): DQN, Actor-Critic A3C

End-to-end driving: policy $\pi$ searched as ConvNet(front-image)

---

End-to-end driving learning by RL in racing-car simulator

---

End-to-end driving learnt by RL (in a racing-car simulator)

Performance
Trained for 196 million steps

Test on training track

Snow (SE)

Network input and guided backpropagation

Game graphics

Activations

Layer 1

Layer 2


First RL experiment for end-to-end driving in urban environment

End-to-end driving via Deep Reinforcement Learning [thèse CIFRE Valeo/MINES-ParisTech en cours]

WORK IN PROGRESS...
Conclusions

• Most current AI challenges for Robotics and Autonomous Vehicles are related either to: Human-Robot Interaction, understanding of Human actions or behaviors, inference of Human intents, or learning of complex adaptive behaviors.

• 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.

• For Human movements or intents analysis, combining human-pose estimation by DL with Deep Temporal Convolution of time-series seems promising.

• For behavior learning, Deep Reinforcement Learning from images already produces interesting results.
Questions?