Machine *Learning*

Fine-Grained Data Collection

Mostafa S. Ibrahim
Teaching, Training and Coaching for more than a decade!

Artificial Intelligence & Computer Vision Researcher
PhD from Simon Fraser University - Canada
Bachelor / MSc from Cairo University - Egypt
Ex-(Software Engineer / ICPC World Finalist)

Hands-off wheel Detection

Hands-off wheel Detection

- Some recent cars allows the customers to have their hands-off
- However, at some point it might ask them to take control
- The system needs to know the driver has hands-off or not
- If the hands is off and s/he must put his hands on, an alert is generated after X seconds
- Users try to trick the system
  - E.g. sleep in the car and use artificial hand
- Definition: **hand is on** if one of the 2 hands is touching the steering wheel (even basic touch)

Solutions

- There are 2 popular solutions
- **Capacitive** sensor that knows the user is touching the wheel
  - The user tricks by attaching a toy to the wheel
- Computer vision system: **Camera** to monitor the driver
- In practice, we may use both together

Fine-grained Collection

- Imagine, you want to build vehicle vision detector
- You don't want your system be surprised that never seen
- You want to think in all **possible** cases
  - Different cars vendors: toyota, honda, etc
  - All cars: black, white, etc
  - All sizes and styles
- Collect data to cover **specifically all scenarios** you identified
- If you did random collection, you may fail on what you miss

Data Collection

- In practice, companies seeks models that are sooooo accurate
- To do so, we collect a lot of data
- Before collecting data, we need really to be **specific** about collecting data with **variance** as much as we can
- We need to identify every single DIMENSION
- And it is dimension collect data
- We can collect as **fine-grained** as we can rather than **coarse** data
  - Car (coarse) - Honda Car (fine-grained), Honda CRV Car (More fine-grained)
- Goal accurately detect HOW status and consider users that tries tricking the system

Ongoing activity

- From week to another, I would like to see what more details you found
- Examples of dimensions:
- **Weather**
  - Different Types of Weather Conditions: sunny, cloudy, windy, rainy, and stormy
  - Then, we need to make sure to cover all such weather
  - Otherwise, the model might fail in a weather not in the database
- **Gender**
  - Male, Females
- Your turn, brainstorm as much as you can. Identify scenarios and deimsions to specifically collect rather than randomly collect!

Data Collection in Practice

- Collect diverse data (like this task)
  - Consult with domain experts who have a deep understanding of the problem and data
- Be aware of data imbalance and handle it (later)
  - For example, dataset is 90% dogs and 5% cats and 5% other 10 animals
- Data augmentation to artificially enhance the data
  - For example, images can be *rotated, scaled, or flipped* to create new variations without changing the underlying meaning
- Monitoring and updating: Regularly monitor the model's performance in real-world scenarios and gather feedback
- Active data collection
  - Based on distribution changes or model performance, collect more data!

- Solution [doc](#)
- Good relevant read: [Why Data Will Disappoint You](#)

"Acquire knowledge and impart it to the people."

"Seek knowledge from the Cradle to the Grave."