Towards Fusion of Semantic Knowledge into Deep Learning Models

Agenda

I. **Neural-Symbolic systems: Overview and Ongoing work** (M. Ebrahimi, WSU)

II. **What is Multimodal Machine Learning?** (J. Francis, CMU & Bosch)

III. **Multimodal Sense-Making: a natural ground for neural-symbolic systems** (A. Oltramari, Bosch)

IV. **Discussion**

- Questions/Comments
- "I'm working on it!": share your experience in 60 seconds

US2TS

MULTIMODAL SENSE-MAKING
A NATURAL GROUND FOR NEURO-SYMBOLIC SYSTEMS

ALESSANDRO OLTRAMARI

BOSCH RESEARCH AND TECHNOLOGY CENTER
PITTSBURGH, PA (USA)

BOSCH

Multimodal Sense-Making
Multimodality across Bosch IoT Use Cases

Mobility

Increased convenience, efficiency and driver safety through in/out-vehicle sensing (driver's attention, obstacles, road friction,...)

Buildings

Simplification of people's everyday lives thanks to intelligent, interconnected buildings that learn from human behavior, energy consumption, appliances, ...)

Smart City

Increase in energy efficiency, quality of life, security optimized use of resources, through environmental sensing (cameras, traffic, pollution, water quality, ...)

Industry

Closely interconnected industry systems increase productivity and speed during the processing of joint tasks, enabling predictive maintenance

Semantic Scene Understanding
Multimodal applications: from Challenges to Opportunities

	Challenges				
Applications	Representation	Translation	Alignment	Fusion	Co-learning
Speech recognition and synthesis					
Audio-visual speech recognition	✓		✓	✓	✓
(Visual) speech synthesis	✓	✓			
Event detection					
Action classification	✓			✓	✓
Multimedia event detection	✓			✓	✓
Emotion and affect					
Recognition	✓		✓	✓	✓
Synthesis	✓	✓			
Media description					
Image description	✓	✓	✓		✓
Video description	✓	✓	✓	✓	✓
Visual question-answering	✓		✓	✓	✓
Media summarization	✓	✓		✓	
Multimedia retrieval					
Cross modal retrieval	✓	✓	✓		✓
Cross modal hashing	✓				✓

- Semantic Web Technologies can provide methods and tools to take up some of the challenges in multimodal applications, in particular wrt REPRESENTATION, ALIGNMENT, FUSION.

- An ontology that represents and grounds intrinsic and extrinsic multimodal properties is much needed
- Such ontology can serve as semantic model for
  - *joint representation* of multimodal space
  - characterizing the direct relations between elements from *n* different modalities (*explicit alignment*)
  - early fusion: multi-modal knowledge graph to perform prediction tasks (*model-agnostic approach*)
    - (see 1st presentation for approaches on knowledge graph embeddings)

Multimodal Sense-Making
SSN as basis of a Scene Ontology

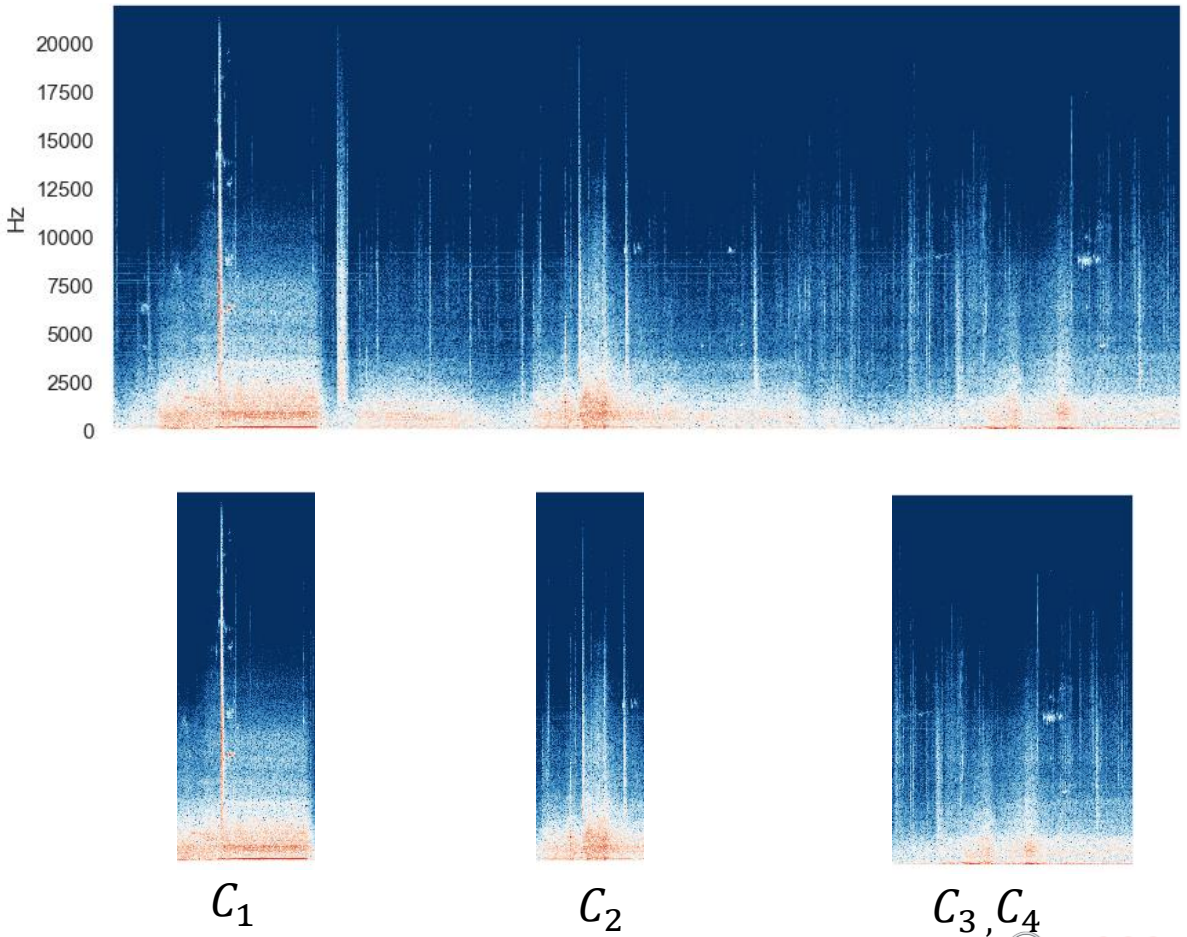
- W3C's Semantic Sensor Network as reference model for Scene Ontology
- Short definition through SSN: "scene" is *constituted-by* multimodal OBSERVATIONS of some FEATURES OF INTEREST *made-by* some SENSORS
- Spatiotemporally co-located entities form a scene according to a context of reference
  - "People running out from the office because of a fire"
    - temperature-observation X of office Y made by thermostat Z
    - number-of-people-observation W of office Y made by occupancy-sensor S
    - trajectory-observation(s) T of person(s) (P_1,..P_i) located-in office Y made by camera C
    - CO_2-concentration-observation O of office Y made by smoke-detector K
  - "Car stopping before hitting pedestrian"
    - Position-observation I of pedestrian B made by short-range-radar G
    - Harsh-braking-observation H of car A made by accelerator-sensor L
    - Friction-coefficient-observation F of road R made by force-sensor E

Multimodal Sense-Making
Integrating symbolic and sub-symbolic level

- As the previous slide hinted at, each observation class is instantiated by data in a specific modality
- Each instantiated observation is part of a sensing event
- Instances of different types of sensor-based observation are typically sub-symbolic, e.g. audio signals
  - Interesting problem: how to represent a mathematical object, e.g. a Short-term Fourier Transform. See integration of MathML and OWL (stackoverflow.com and semantiweb.com)
- Bundles of instantiated observations denote a specific scene
- Scene ontology can be used to:
  - Ground multimodal observations on a context
  - Improve high-level reasoning and disambiguation
  - supervise multi-modal representation learning algorithms

Multi-modal Sense Making
Towards Scene Understanding: the case of audio modality /1

- Current **Deep Learning** (DL) approach: feed large temporal windows to pipeline
  - **Complexity degrades performance**: long inputs include different classes of event, some being causally linked, some being background noise, some being common across scene types
  - **Lack of Explainability**: classifiers learn something, but what exactly? Which events are important in the detection process?
- **Ontologies** can represent classes of observations that define a scene, helping to understand it
  - We want to classify scene S where each scene is a collection of classes, i.e., $C_S \subseteq C$



Multi-modal Sense Making
Towards Scene Understanding: the case of audio modality /2

- **Hypothesis**: support DL pipeline by integrating scene ontology (of specific domains), mapped-to/distilled-from external knowledge resources
  - Google AudioSet Ontologies have a similar goal
- **Testing**: based on DCASE 2018 challenge

http://dcase.community/challenge2018/task-acoustic-scene-classification

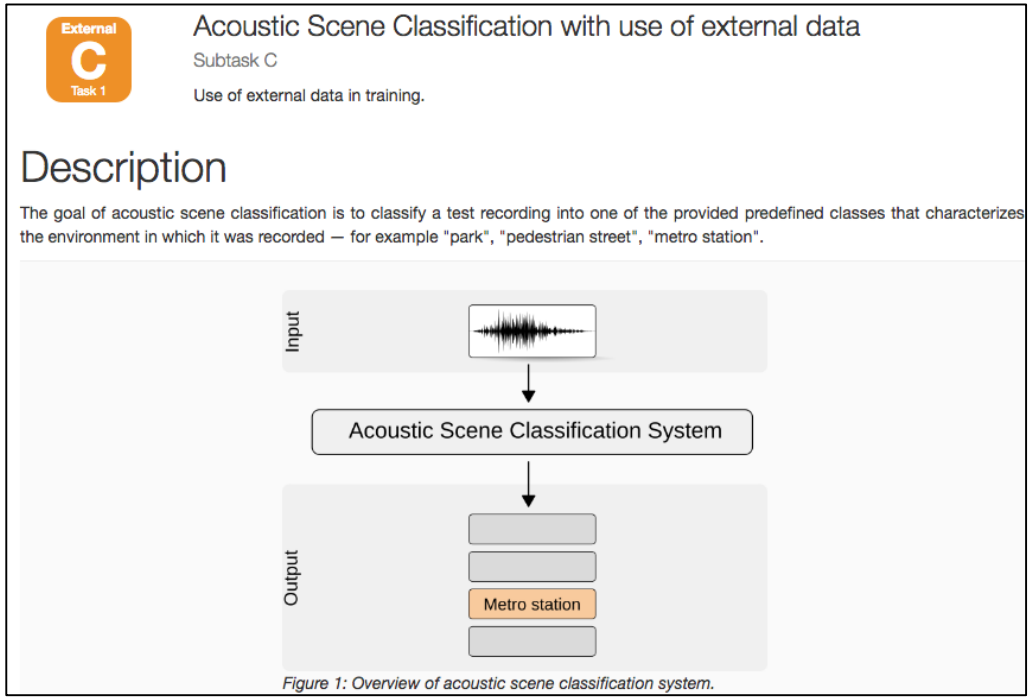


Figure 1: Overview of acoustic scene classification system.

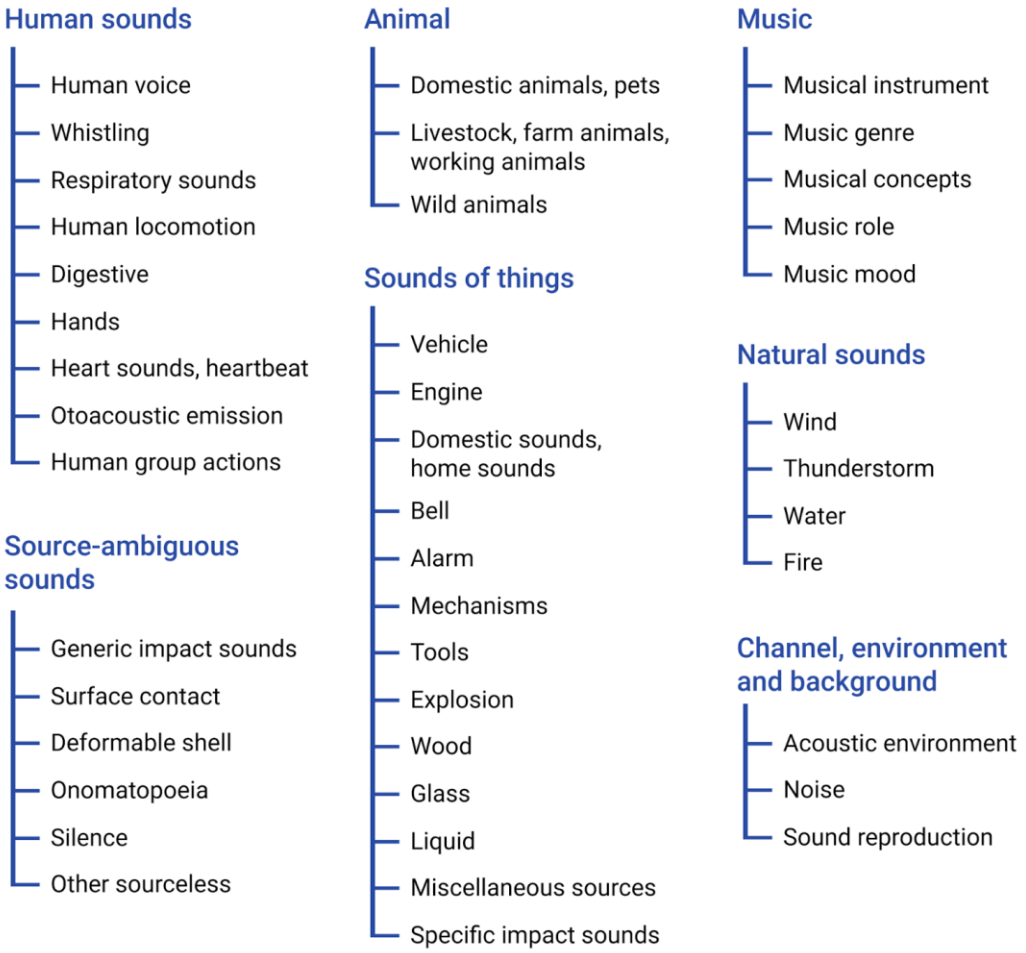
***DCASE classes**: Airport, Indoor shopping mall, Metro station, Pedestrian street, Public square, Street with medium level of traffic, Travelling by a tram, Travelling by a bus, Travelling by an underground metro, Urban park*

BOSCH

Multi-modal Sense Making
Google AudioSet “Ontology”

- The AudioSet ontology is a collection of sound events organized in a hierarchy. The ontology covers a wide range of everyday sounds (human, animal, natural, environmental, musical, etc.
- Each ontology entry contains a description of the sound event, modified from Wikipedia, Wordnet, or written by Google.
- The ontology is meant to be expandable to meet to research needs of the academic community (GitHub)

https://research.google.com/audioset/



Multi-modal Sense Making
Benefits of Scene Ontology for Audio Scene Understanding

- **HIGH LEVEL REASONING**: semantic compositionality, dependencies, common sense inferences
- **FALL BACK STRATEGY**: When a classifier encounters ambiguity among several subcategories (e.g., a sound that is recognized ambiguously as “item scanning” “boarding pass scanning”, “transit card scanning”), it can fall back to a classification of “scanning” (super-class)
  - Domain coverage and granularity to be expanded/deepened on the basis of use cases.
    - Bosch has several use cases spanning from smart cities to industry 4.0
    - I envision a Scene Ontology W3C working group with a strong presence of industrial stakeholders in the consortium
- **BETTER ANNOTATIONS**: well-structured semantic models can aid human labeling tasks by allowing a labeler to quickly and directly find the set of terms that best describe a sound;
  - This is also important during the development of an event corpus when trying to add categories without overlap and duplication.

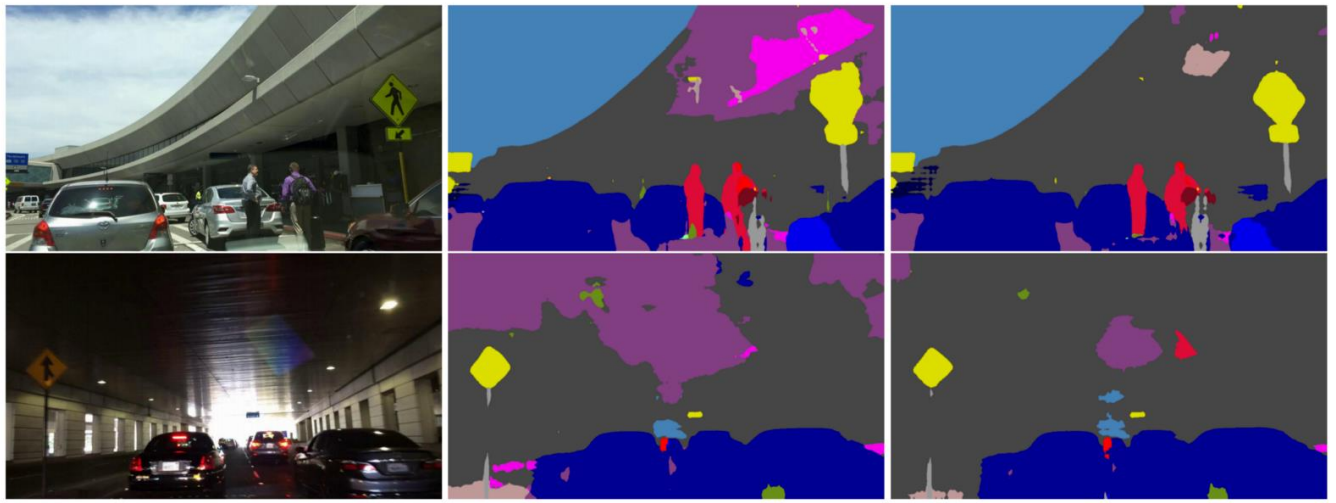
These benefits are not only applicable to audio, but to any modality.

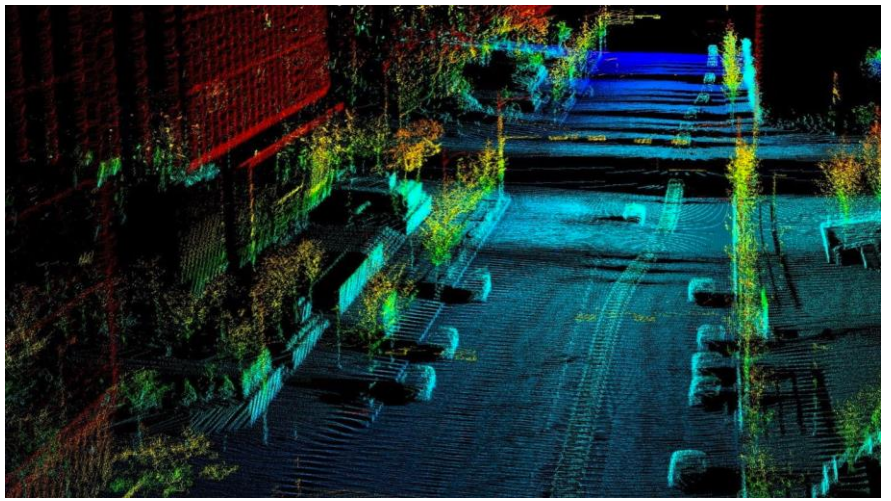
Multi-modal Sense Making
What *understanding a scene* means to a computational system /1

- As humans, we are good at making sense of the environment: we use knowledge to generalize over experience, and reason over and explain what we perceive
- "Sense-making" can be reproduced at the machine level by implementing a human-like pipeline
  - Perception-based Machine-Learning Classification
    - person#1, person#2, person#3
    - extended arm (person#1)
    - Hand/Gun
  - Knowledge-based reasoning
    - Person#1's extended arm
    - Gun co-located with hand
    - Hand <grab> gun
      → Person#1 pointing the gun
      → Person#2 and Person#3 are gun pointed at
    - Large parallelepiped between persons on a retail store → "counter"
    - **Scene: Armed Robbery in a Retail Store**

Multi-modal Sense Making

What *understanding a scene* means to a computational system /2





BOSCH

Multi-modal Sense Making
What understanding a scene means to a computational system /3

"Autonomous AI gives machines the ability to sense and respond to the world around them, to move intuitively, and to manipulate objects as easily as a human can. Included in this are autonomous vehicles that can see the environment around them, recognizing patterns in the camera's pixels, figuring out what they correlate to (e.g. road signs), and then using that information to make decisions"

Kai Fu Lee, CEO of Sinovation Ventures, former president of Google China, author of *AI Superpowers: China, Silicon Valley and the New World Order"*

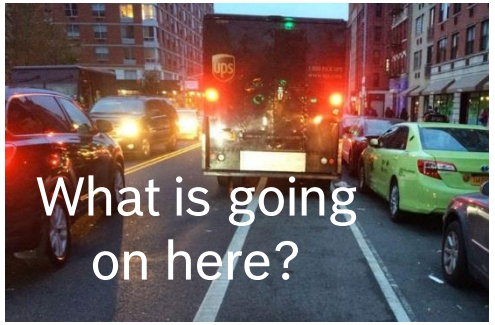
Multi-modal Sense Making
The case of Autonomous Driving

- Challenging environmental conditions make L-5 autonomy currently not achievable

Dynamic Environment

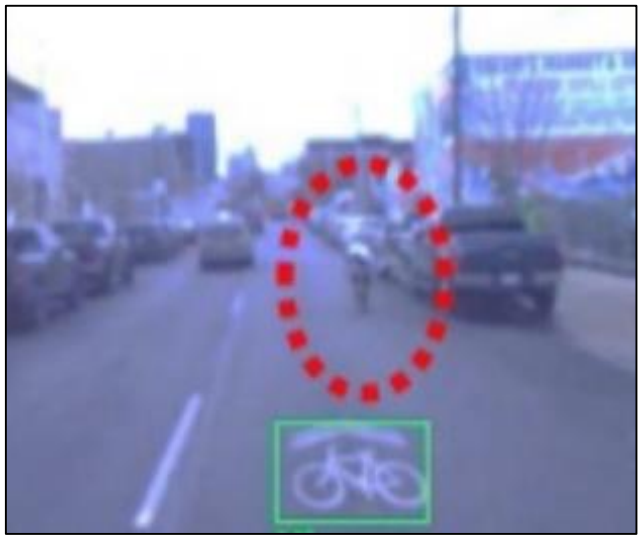
Perception

Behavior on the road

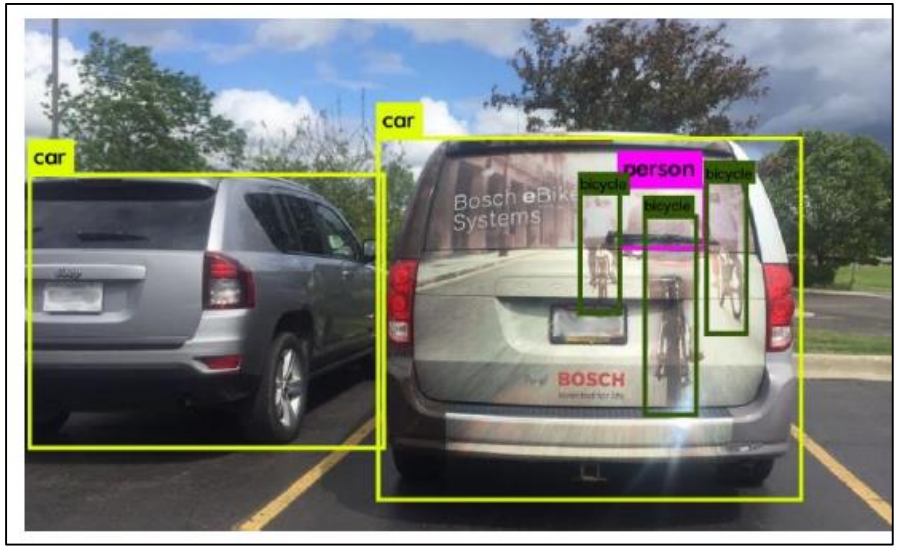


Multi-modal Sense Making
Lack of scene understanding exposes the vulnerability of ML /1

- Natural adversarial examples are far from being uncommon in the city landscape

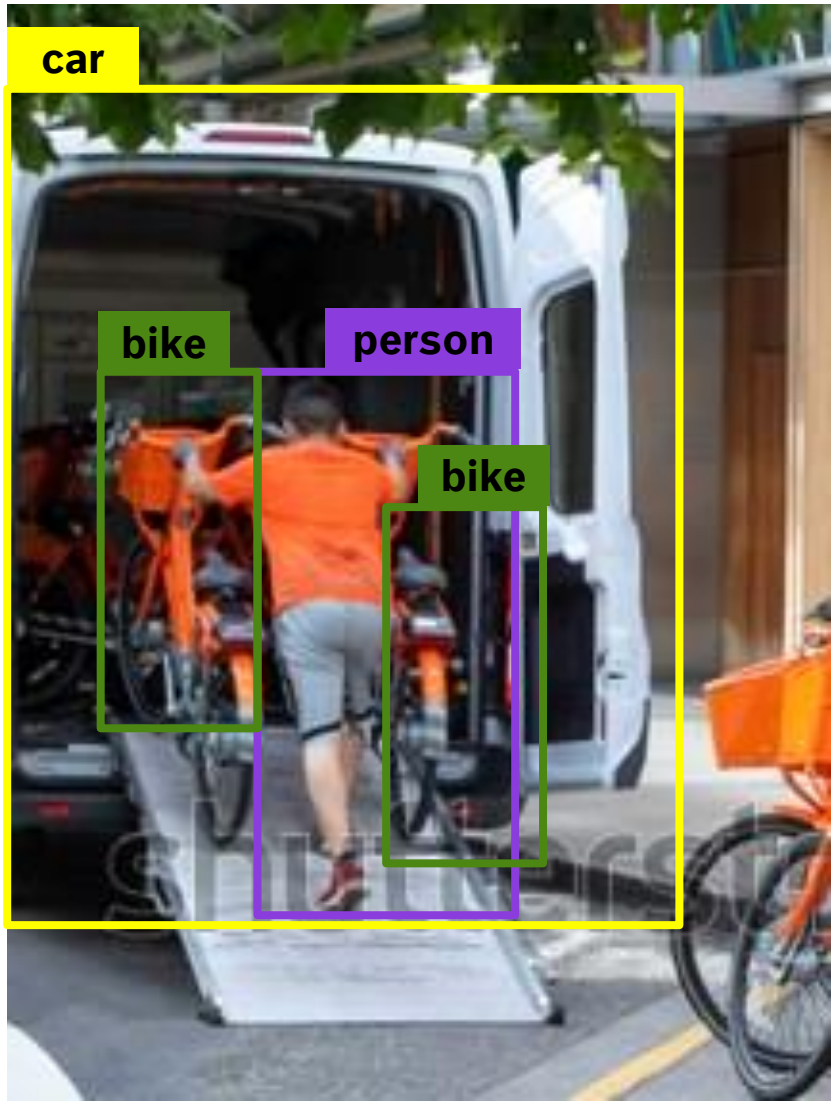


- Prototypical bike flattened on the road (false positive)
- Bike proceeding straight (false negative)
- The sign above the false positive changes the semantics of detection, from “object” to “road sign”



- Person and bikes are detected on the reflective surface of the trunk of a parked car
- Persons and bikes cannot be detected inside of a parked car

….or can they?



- Person is loading two bikes into a van
- What is false positive in a scene becomes a true positive in another
- Knowledge can endow perception with a context of reference

- **Robust scene understanding can only be achieved by a hybrid reasoning that integrates common sense, domain knowledge, learning algorithms and high quality annotations**

Multi-modal Sense Making
Lack of scene understanding exposes the vulnerability of ML /2

Evtimov, I., Eykholt, K., Fernandes, E., Kohno, T., Li, B., Prakash, A., ... & Song, D. (2017). ***Robust physical-world attacks on deep learning models****. arXiv preprint arXiv:1707.08945, 1, 1. CVPR 2018*

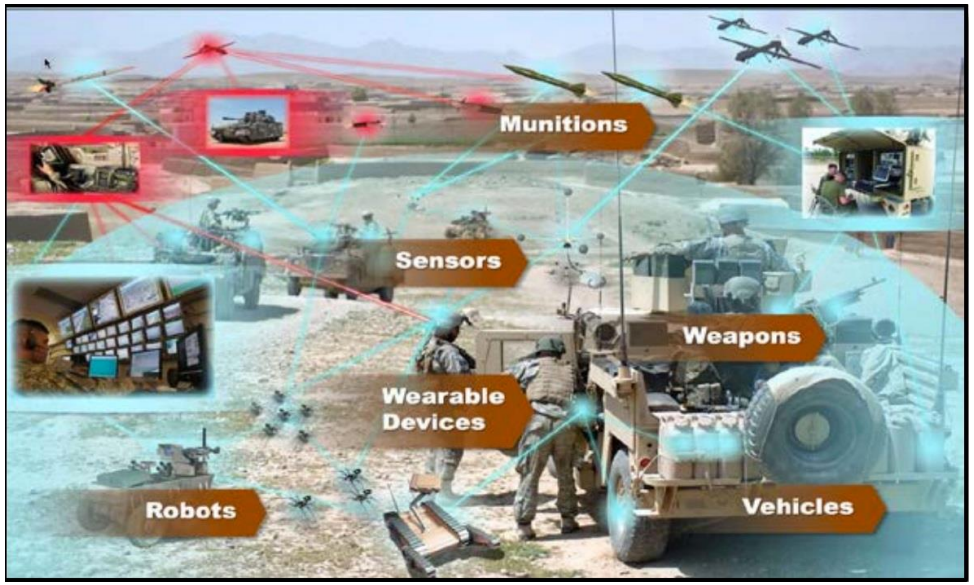
Table 1: Sample of physical adversarial examples against LISA-CNN and GTSRB-CNN.

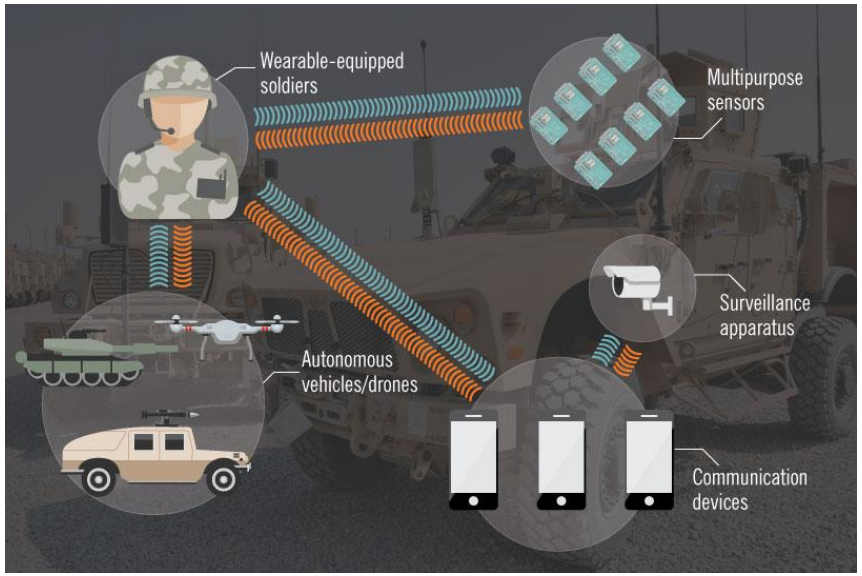
Distance/Angle	Subtle Poster	Subtle Poster Right Turn	Camouflage Graffiti	Camouflage Art (LISA-CNN)	Camouflage Art (GTSRB-CNN)
5′ 0°					
5′ 15°					
10′ 0°					
10′ 30°					
40′ 0°					
Targeted-Attack Success	100%	73.33%	66.67%	100%	80%

Multi-modal Sense Making
Lack of scene understanding exposes the vulnerability of ML /3

- It's not far-fetched to envision a future where cyber-physical attacks will target AD systems, by injecting hallucinated multimodal observations in complex sensor networks

Kott, A., Swami, A. and West, B.J., 2016. **The internet of battle things**. Computer, 49(12), pp.70-75.

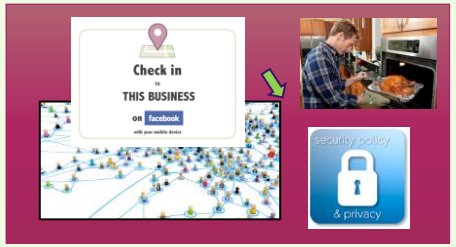




CONTEXT

Sensor Network

Social and Environment

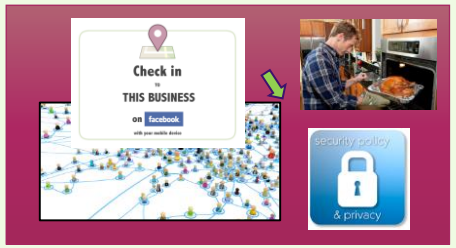


Cognition and Emotion

CONTEXT

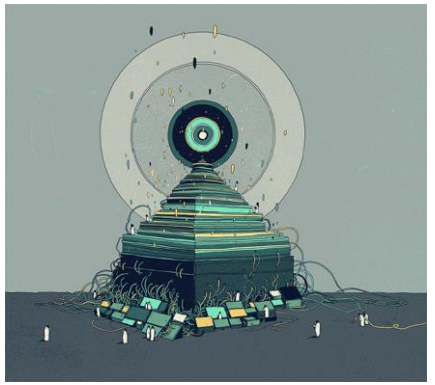
Sensor Network

Social and Environment



Cognition and Emotion

INTELLIGENCE



Adaptive IoT Systems

- Decision support
- Predictive analytics
- Autonomous agent

Actuator Network

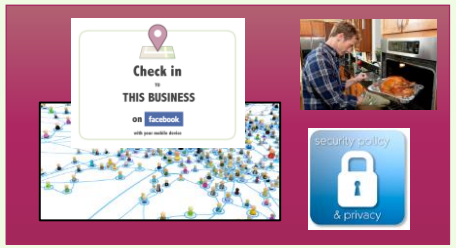
CONTEXT

KNOWLEDGE

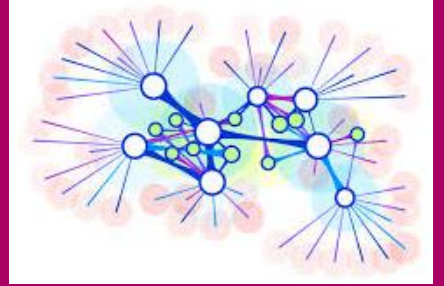
INTELLIGENCE

Sensor Network

Social and Environment



Cognition and Emotion



Cloud Services

Machine Reasoning

Adaptive IoT Systems

- Decision support
- Predictive analytics
- Autonomous agent

Actuator Network

BOSCH

Multi-modal Sense Making
Conclusions

- Integration between perception and knowledge is the key requirement for multi-modal sense making
  - Perception → multimodal (sub-symbolic) data
  - Deep neural networks trained with multi-modal data
  - Multi-modal data instantiate different observation (classes) in the Scene Ontology
  - Neuro-symbolic integration between knowledge graph that represent multimodal data and deep neural networks

- Scene Ontology, mapped to W3C SSN, can play a major role in tackling the representation, alignment and fusion challenges in multi-modal machine learning

THANK YOU!

INTERESTED IN THE TOPIC? LET'S STAY IN TOUCH:
ALESSANDRO.OLTRAMARI.EXT@US.BOSCH.COM

BOSCH