Unifying Computer Vision and Natural Language Understanding for Autonomous Systems

Mumtaz Vauhkonen
Lead Distinguished Scientist
AI & D
Verizon

Computer Vision (CV) and Natural Language Processing (NLP)

Combining computer vision and NLP will lead to more integrated intelligent autonomous systems

CV and NLP Integration:

- **Generate reasoning in language from visual input**
- **From language input generate visual representation**

Example Applications

1. Robotic delivery systems
2. Service industry robots (hospitality, medical)
3. Educational systems
4. Disaster recovery systems

Approaches in Integration of CV and NLP

Some of the integration methods

- Visual description generation
- Visual reasoning
- Visual question answering (VQA)
- Visual generation from text
- Visual dialog
- Visual storytelling
- Multi modal machine translation

Areas of Focus Today

VISUAL DESCRIPTION

Identifying objects and providing captions on their attributes and sometimes action of individual objects

VISUAL REASONING

In addition to visual description adding the most possible interactions associated between each object and their attributes

A rule-based combined with Deep Learning approach for Visual reasoning is presented

Visual Description

1. Generate basic description or captions

Process -> Detect objects -> Generate captions

2. Generate sentence level description of a scene

{desc: Airplane in the sky. Sky is partly cloudy}

1. Sample images from Coco Dataset

History of models:

N-grams, templates, dependency parsing, Sequence to sequence models, Encoder-decoder models

Visual Reasoning

- Visual reasoning involves:
  - Detecting the objects
  - Identifying the attributes of the objects (size, color, shape, features, etc.)
  - Localizing the objects in relation to each others
  - Using reasoning and giving a logical explanation of the scene in the image

Object Detection

{Airplane is on runway in a city. Airplane is about to take off or approach the airport gate.}

A System with Visual Reasoning Should Be Capable Of

- Detecting objects
- Identifying objects' attributes
- Establishing meaningful relationships between objects
- Understanding language
- Translating those relationships into sentences

Example Model on Visual Reasoning

- TbD-Net Transparency by Design Network (1) creates multiple submodules in a series of steps that when combined create a logical sequence of reasoning.
- Example: *"What color is the cube to the right of the large metal sphere?"*
  - Identify the large metal sphere (attributes, color, shape, etc.)
  - Identify the cube's
    - Attributes (color, size)
    - location / direction (what is right vs. left)

TbD-Net MIT Mascharka et al 2018

Attention Module

And Module

Or Module

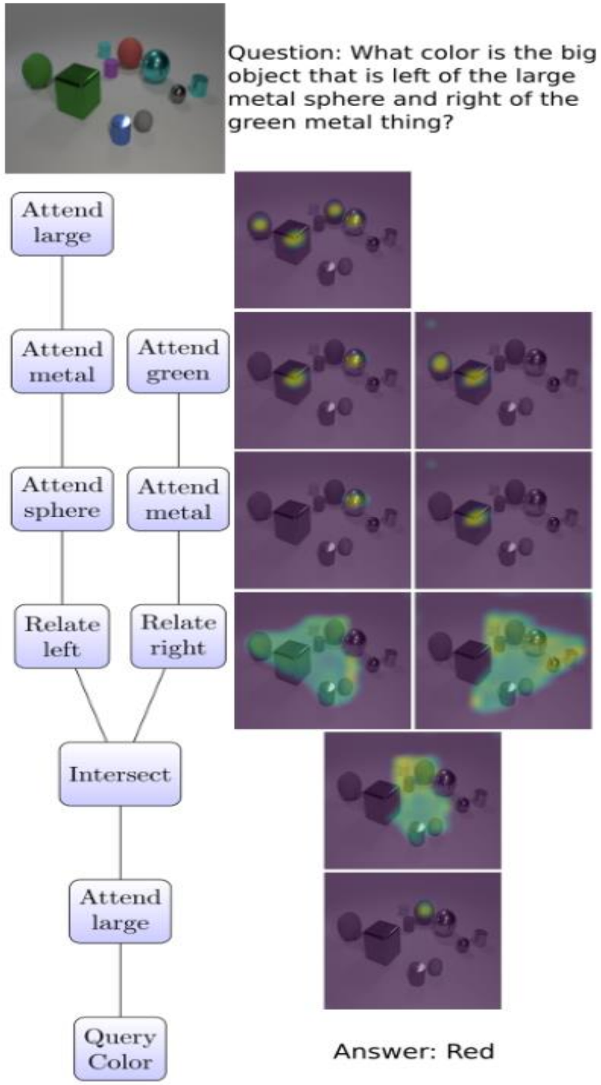
Relate Module

Same Module

Query module

Compare Module

Example Model of Visual Reasoning (continued)



Compositional Language and Elementary Visual Reasoning Dataset (CLEVR)

TbD-Net MIT Mascharka et al 2018

A New Approach: Rule-Based Lingual Model with Deep Learning

Present Work: Deep Learning Rule-based Approach for Visual Reasoning

- The method developed uses a combined CNN architecture and a rule-based approach to provide visual reasoning
- This method allows gaining confidence over object-to-object relationships to reason the interaction as more examples are encountered
- This is achieved by providing a Universal Lingual model
- This model has the ability to localize to specific domains using a distributed AI model with 5G (hospital, tourist centers, retail or manufacturing facility)

Architecture Modules

Module	Description
Derive object rules from corpus	Use method similar to word embeddings Identify the attributes, rules and actions of objects
Detect objects	A CNN-based Object Detector (Yolo) or other
Identify attributes Actions Functional rules	For each detected object identify attributes, actions rules from step 1 and also spatial mapping of each object in the frame
CNN to match the actions based on rules from corpus	Train a model to establish links between objects with their actions, rules and attributes
Create a Universal lingual model with codes for action and attributes	Mapping of actions and attributes to codes to establish reasoning connection vs lingual
Translate into language representation Transformer Model	Train a transformer language model to create sentences based on the attributes rules and actions matched

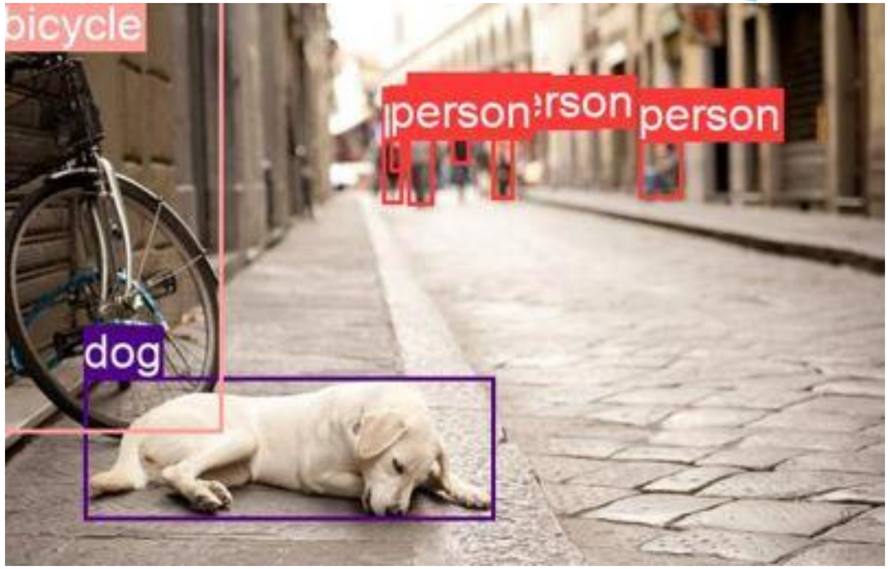
Derive Object Relationship from Language Corpus

Examples of derived rules, actions and attributes

- Dog walking with human
- Person biking on the road
- Attributes: (bicycle: two wheels, seat, pedal)

Object Detection and Attributes

- Use Yolo (or other) for object detection
- Match with derived object attributes
- For each object create an attribute table
- Identify orientation of each object in relation to each other
    - Relative pixel-wise distance between objects
- Map the actions of each object vs. the other



Dog	Bicycle
Attributes: Furry, legs, ears, tail	Two wheels, pedal, seat, handle

Creating the Universal Lingual Model

- Match with actions and functional rules of the detected object
  - Learn functional rules for each object
    - Examples: (bird can fly, car can drive forward or backward, park or crash, dog cannot fly)
- Derive functional rules of objects and encode with universal codes for actions
- Map code to each language

Derive Functional Rules of Objects from large corpus

Dog	Bicycle
Dog run(1)/runs(1a)/is running(1b) Dog walk/walks/walking Dog jump/jumps/jumping Dog sit/..../... Dog bark/...../	Bicycle ride(1a)/riding(1b) Bicycle parked Bicycle fall Bicycle broke Bicycle empty Bicycle rider

Establish Meaningful Relationships between Objects

- **Dog action + dog functional rule + location in relation to bicycle + bicycle functional rule**

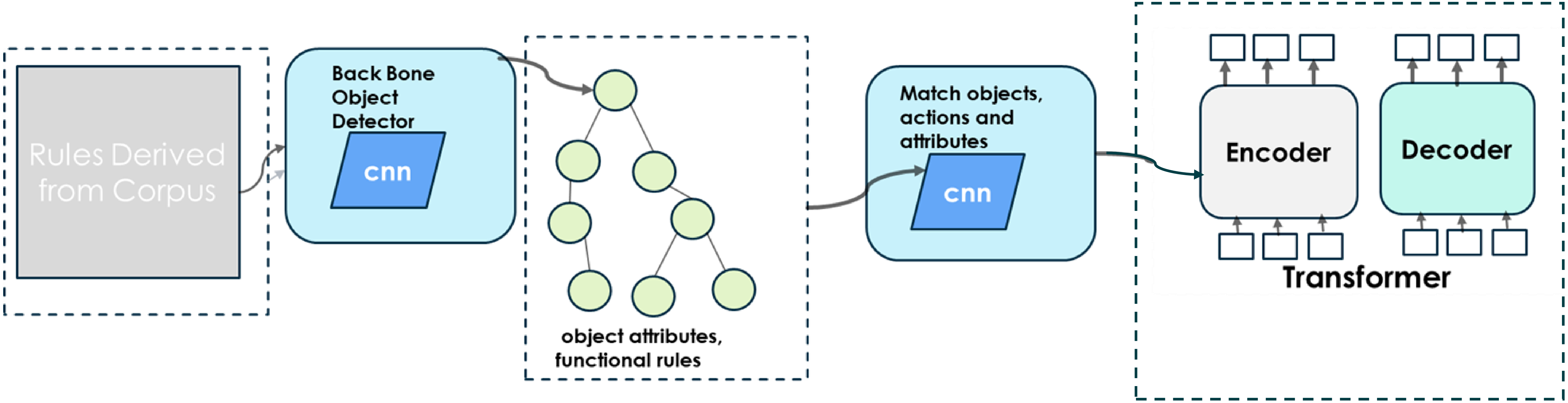
- Feed the attributes, actions and functional rules into CNN-based system and train to associate reasoning between objects based on highest probability matches
- The more relationships the model correctly identifies the higher the probability assigned to those associations for the future

System Understanding Language

- Map the match of attributes and functions onto to the universal codes
- Translate the match of attributes and function codes into lingual sentences

➢ **Dog sitting next to bicycle**

Overview of Architecture



Localization with 5G and Mobile Edge Compute

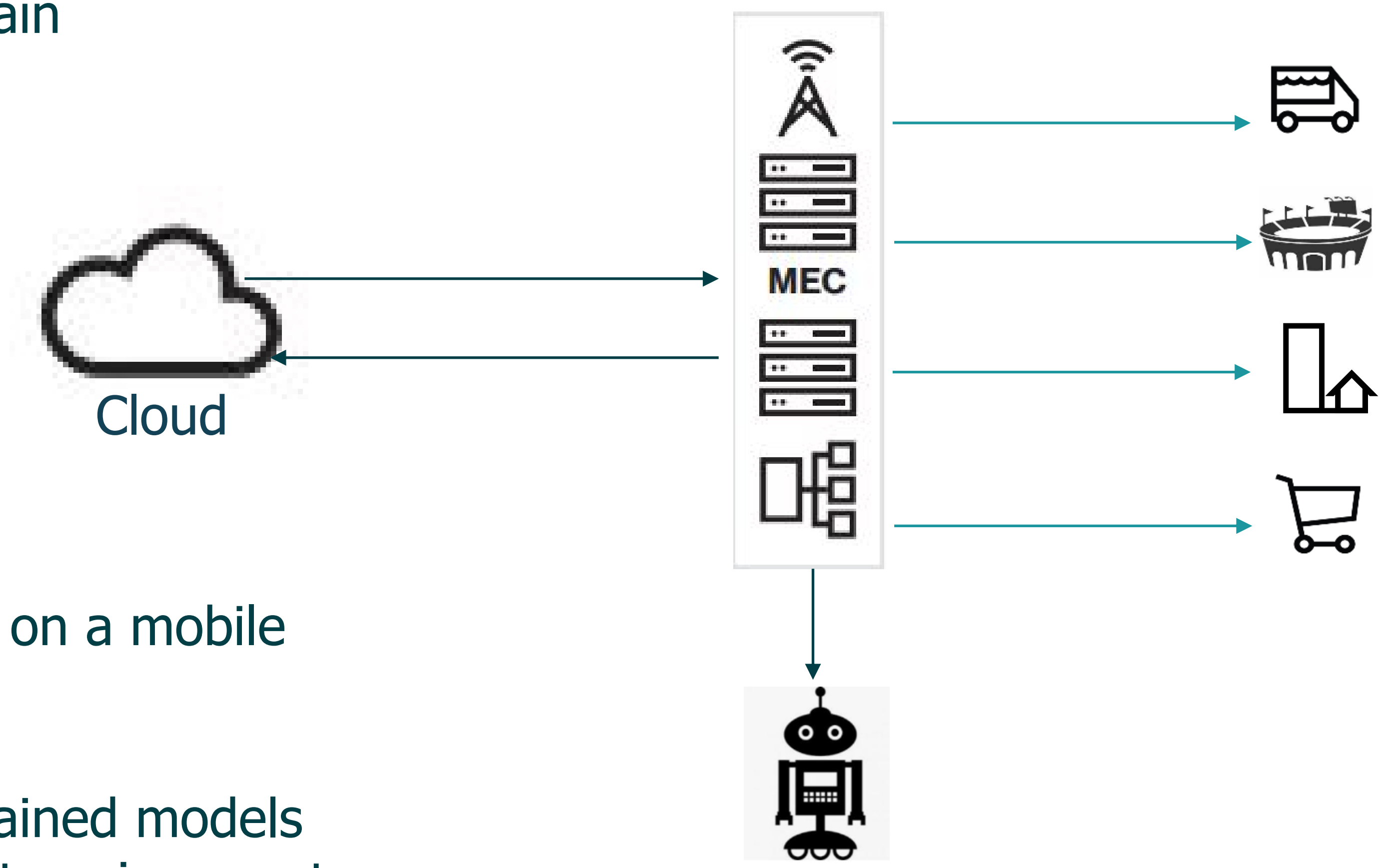
Localization: Can be localized to a domain

Training: Trained models for multiple domains can exist in cloud or Mobile Edge Compute (MEC)

Connectivity: 5G enables fast connectivity and data transfer with low latency and high compute accessibility

Derive High Value: Compute resources on a mobile autonomous system are limited

A basic robot can be connected to the trained models wherever needed and operate fully in that environment



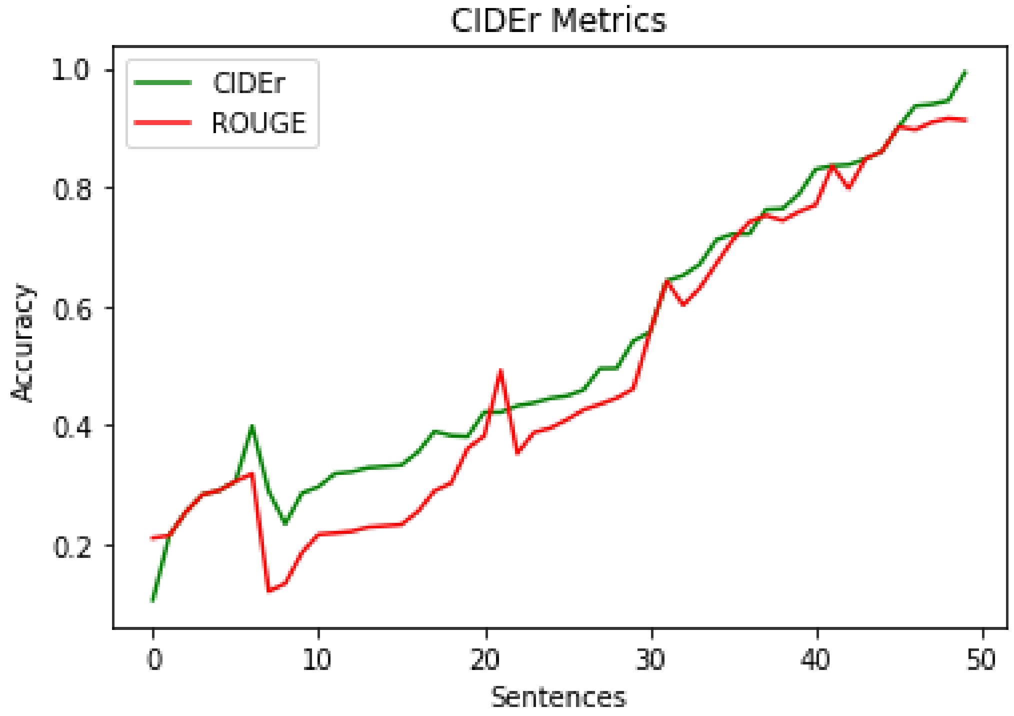
Advantages

- Deriving rules, attributes and functional rules and mapping to detected objects cuts down large computations needed for reasoning
- Specialized training for high accuracy in localized environments vs. overall generic training
- Rules can be constantly updated independently of object detection or mapping CNN architecture
- Encoding is from a perspective of reasoning vs. purely language based

Metrics and Results

- Metrics used for final output is Consensus based Image Description Evaluation (CIDEr)*. It measures the similarity of a generated sentence with a human derived ground truth sentence. This metric show consensus on how close to human generated sentences the auto generated sentences are.

- Compared to ROUGE: Uses n-gram method to
- compare automatic summarization to human created summary



* Ramakrishna Vedantam, C. Lawrence Zitnick, Devi Parikh. "CIDEr: Consensus-based Image Description Evaluation" Proceedings of IEEE conference on computer vision and pattern recognition 2015.

Thank You

References

1 . Mascharka, David and Tran, Philip and Soklaski, Ryan and Majumdar, Arjun. "Transparency by Design: Closing the Gap Between Performance and Interpretability in Visual Reasoning", The IEEE Conference on Computer Vision and Pattern Recognition - CVPR June, 2018TbD-Net MIT Mascharka et al 2018

2. Ramakrishna Vedantam, C. Lawrence Zitnick, Devi Parikh. "CIDEr: Consensus-based Image Description Evaluation" Proceedings of IEEE conference on computer vision and pattern recognition 2015.