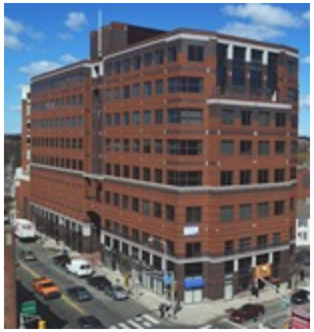
Human Perspective Scene Understanding via Multimodal Sensing

Chiori Hori
Mitsubishi Electric Research Laboratories (MERL)

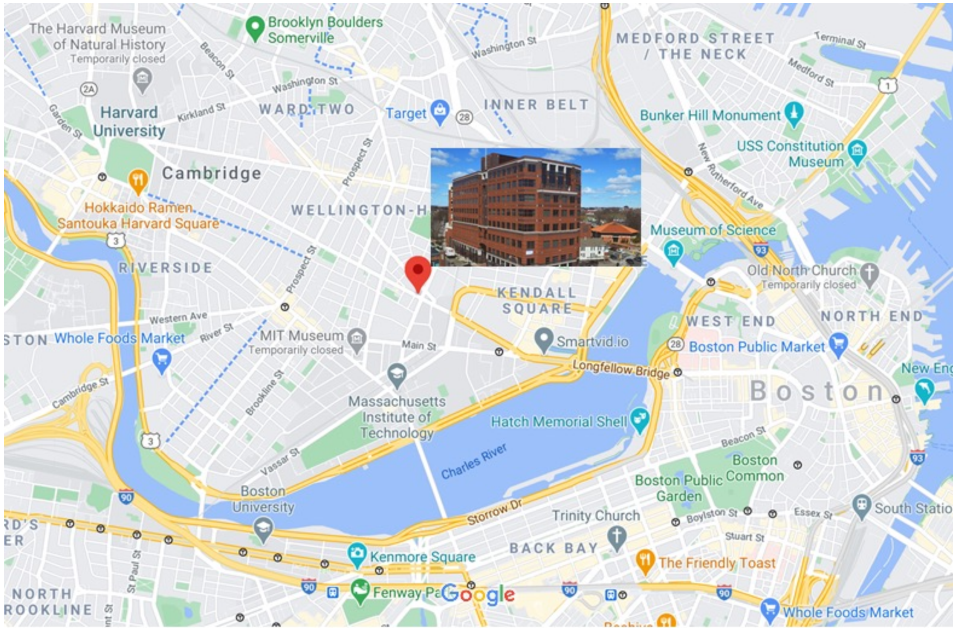


Introduction of MERL

- The North American arm of the Corporate R&D organization of **Mitsubishi Electric**

– **30 years** since **1991**
– 81 members (**More than 50 PhDs** pursue research)
– Mission: Advanced application-motivated basic R&D
– Intelligent properties: more than **700 patents**

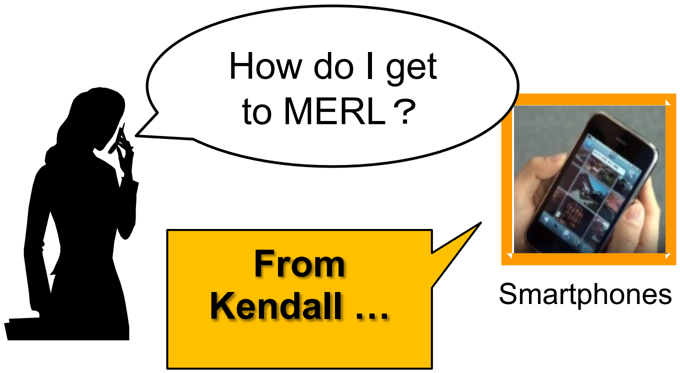
– Target areas
- Wired/wireless communications,
- Signal processing,
- **Audio and video processing,**
- **Spoken language interfaces,**
- **Computer vision,**
- Mechatronics,
- Fundamental algorithms



MITSUBISHI ELECTRIC
Changes for the Better

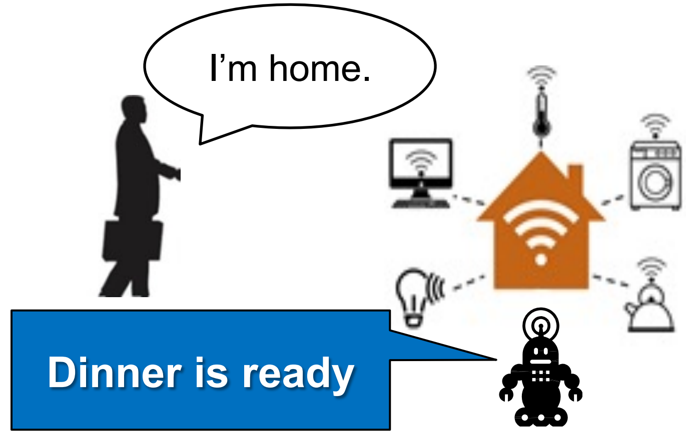
Target Knowledge to Talk

Static Knowledge DB

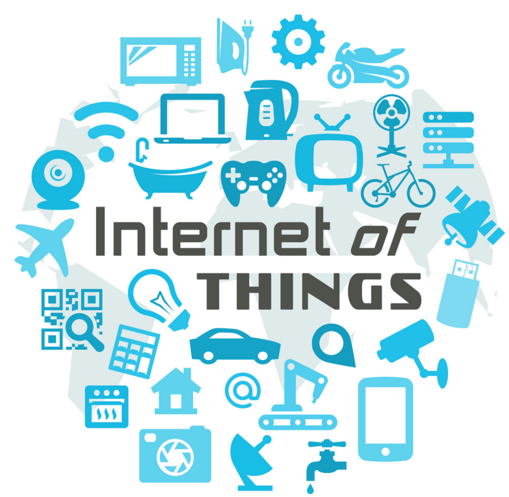


Smartphone Apps + cloud services

Dynamic Knowledge DB



Control appliances through internet



MITSUBISHI ELECTRIC
Changes for the Better

What are we talking about with machines?

I can **hear** !

Let's have a conversation freely.

I can **see** !

I can **walk** !

I can **speak** !

I can **dance**!

Human Machine Interaction: HMI

It's too high
for you.
Do you need
help?

Yes, Please!!!!
Take it down.

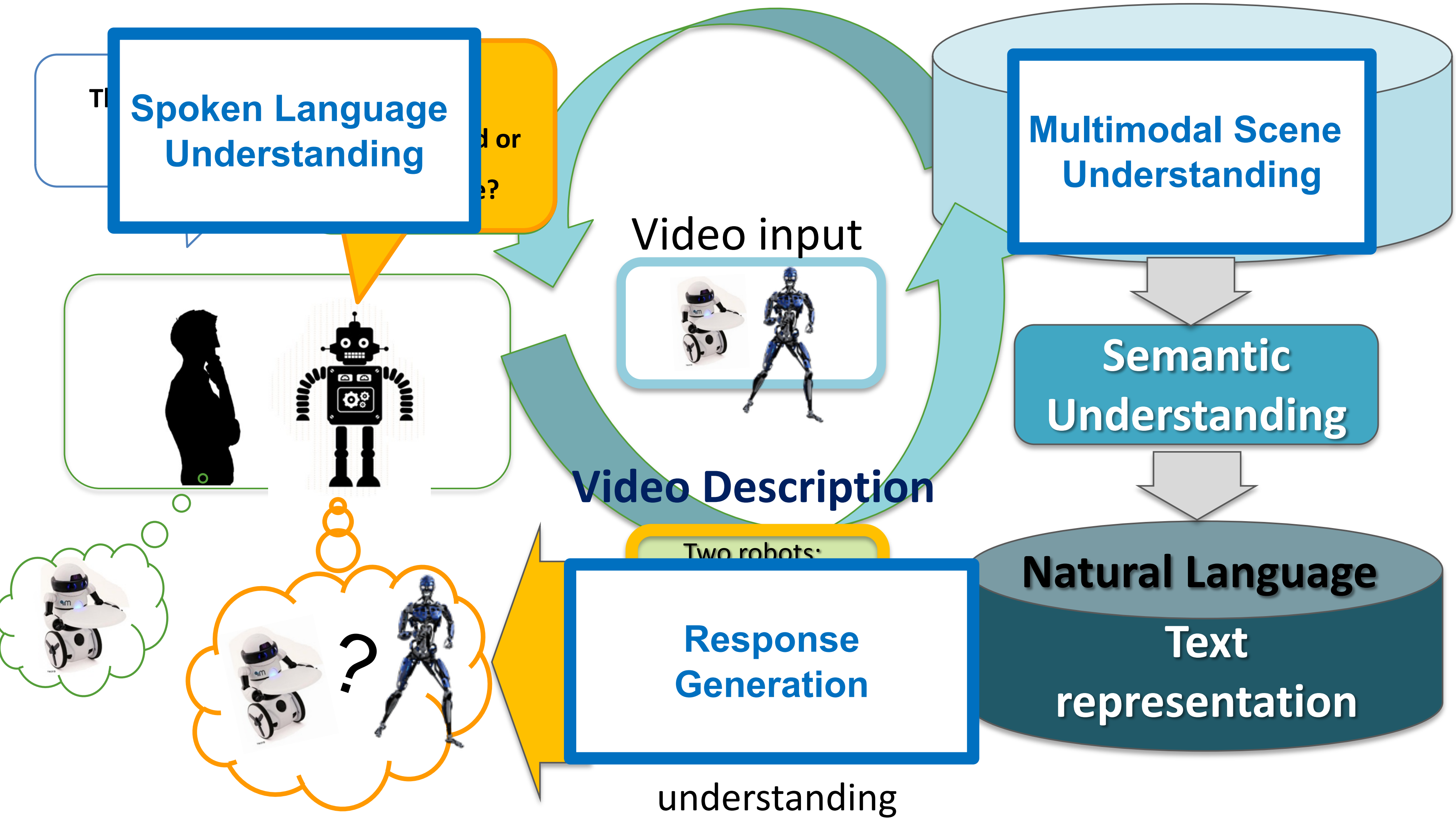
Essential Technologies for HMI

Human understanding

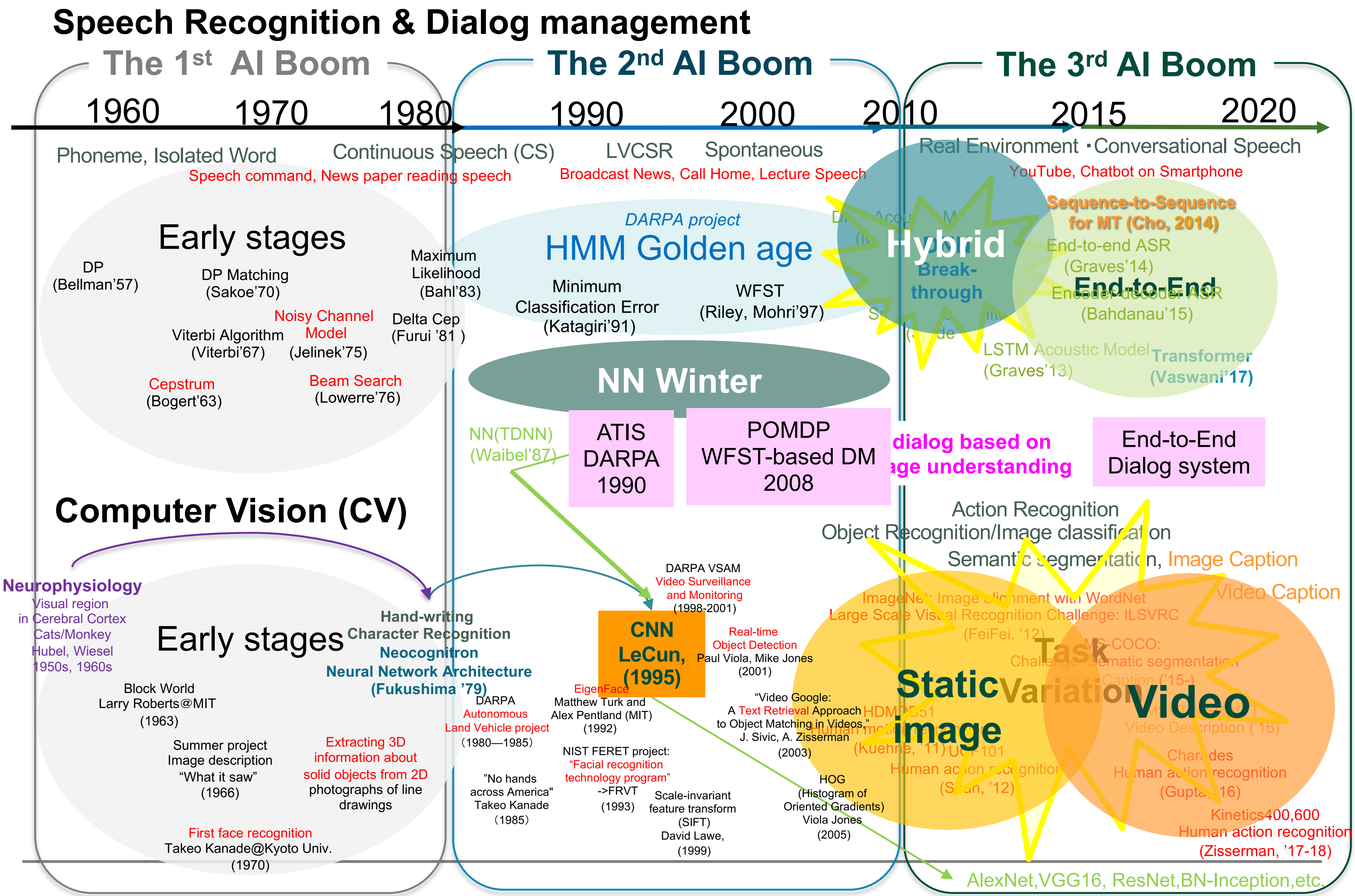
Machine understanding

- Humans understand scenes using natural language
- Machines understand scenes using multimodal sensing information
- To interact with humans in a natural and intuitive manner, machines need to translate the sensing information into natural language

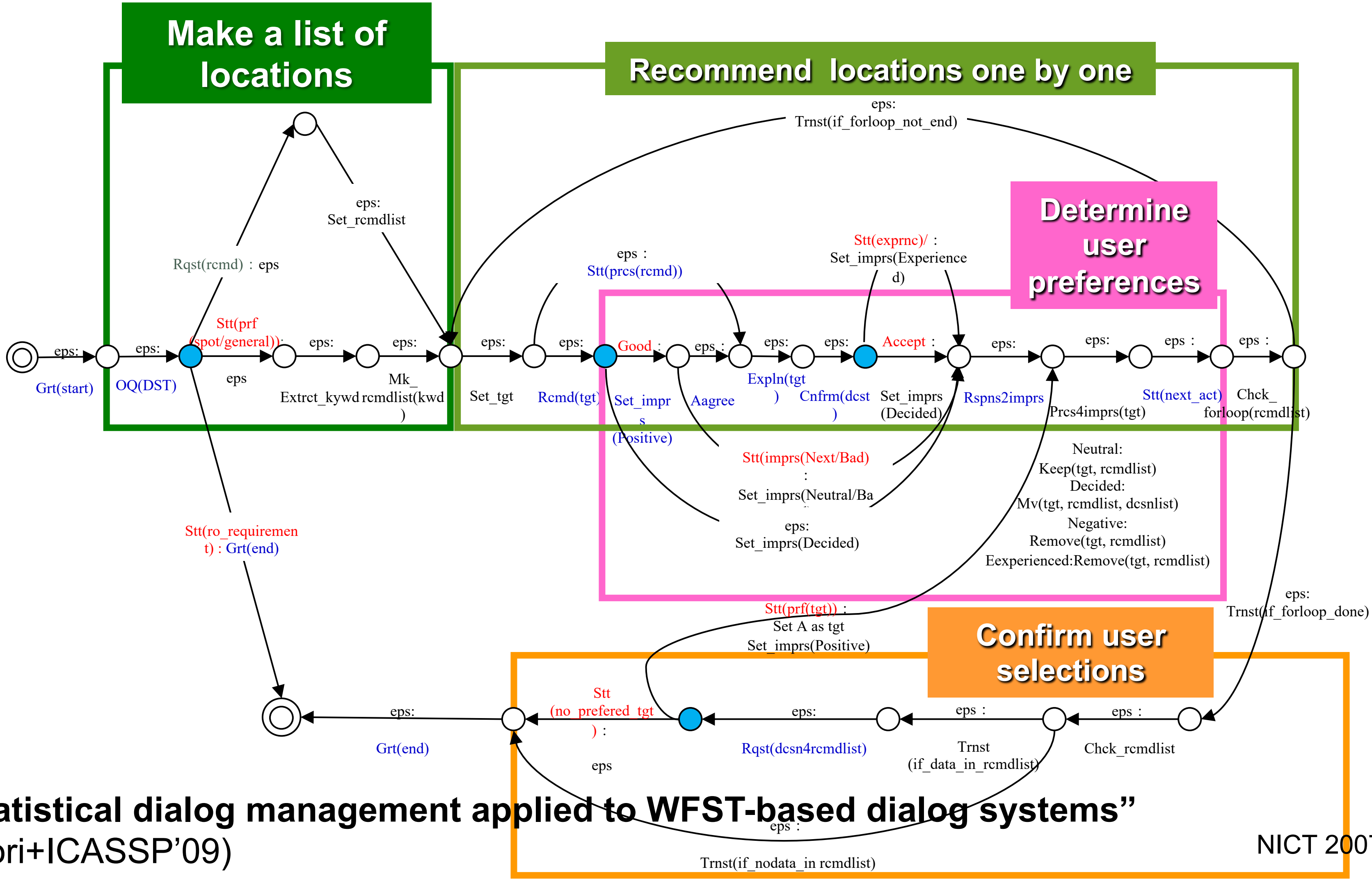
Machines Need to Understand Context Using Natural Language



History of AI Researches



Hand-Crafted Scenario for Tour Guide



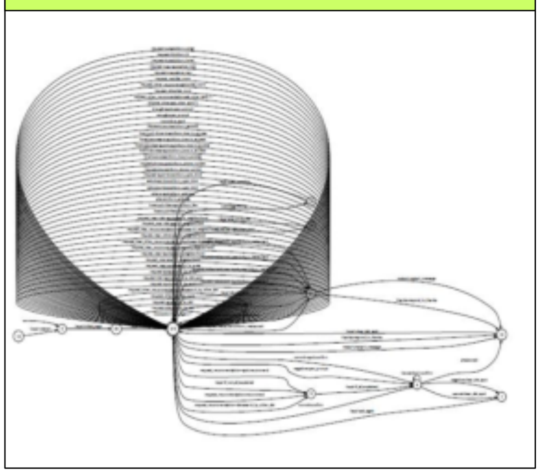
“Statistical dialog management applied to WFST-based dialog systems” (Hori+ICASSP’09)

NICT 2007

Statistical Dialog Technologies

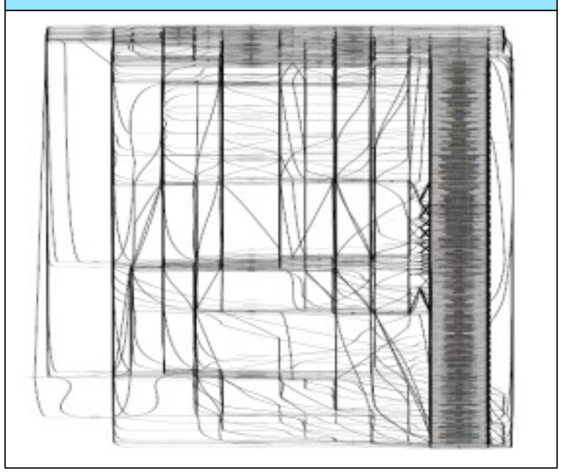
- Statistical dialog systems have been developed to provide greater robustness and flexibility, but rely on discrete dialog state graph to determine next system response
  - Many states and state transitions, esp for large problems

Rule-based model
(QA for Tour Guide)



Statistical model
(Guide action simulator)

Statistical model
(Hotel clerk simulator)



"Statistical dialog management applied to WFST-based dialog systems"
(Hori+ICASSP'09)

NICT 2007

How to scale up training data

- Language data is available with different levels of labels

Type of data & labels	Data size in words
Unlabeled documents	$N \approx 10^{12}$
Knowledge graph e.g., wikipedia	$N \approx 10^{9}$
Conversational data e.g., callhome	$N \approx 10^{5}$
Dialog with rich labels e.g., Kyoto tour guide	$N \approx 10^{5}$
Application intention understanding	$N \approx 10^{5}$
Application dialog data	$N \approx 10^{4}$

Learning embeddings

- Strategy:
  - Learn word/sentence embeddings for unlabeled data
  - Learn embeddings on smaller data + stronger labels based on embeddings from larger data + weaker labels

MITSUBISHI ELECTRIC
Changes for the Better

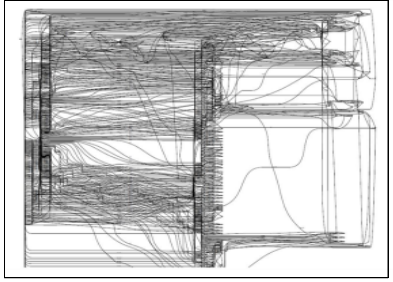
Transition to Deep Learning

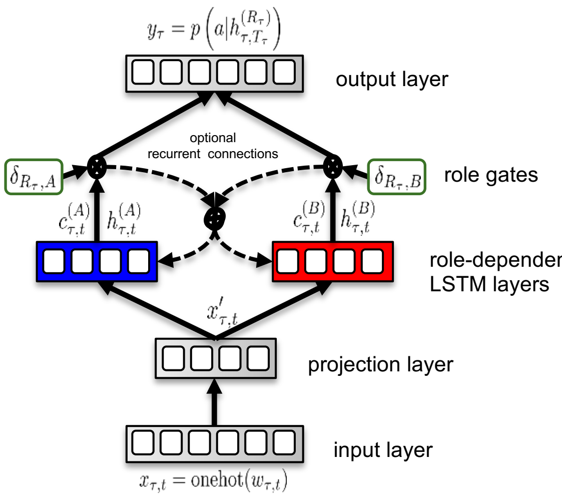
- Spoken language understanding:

  DNN models with continuous state space

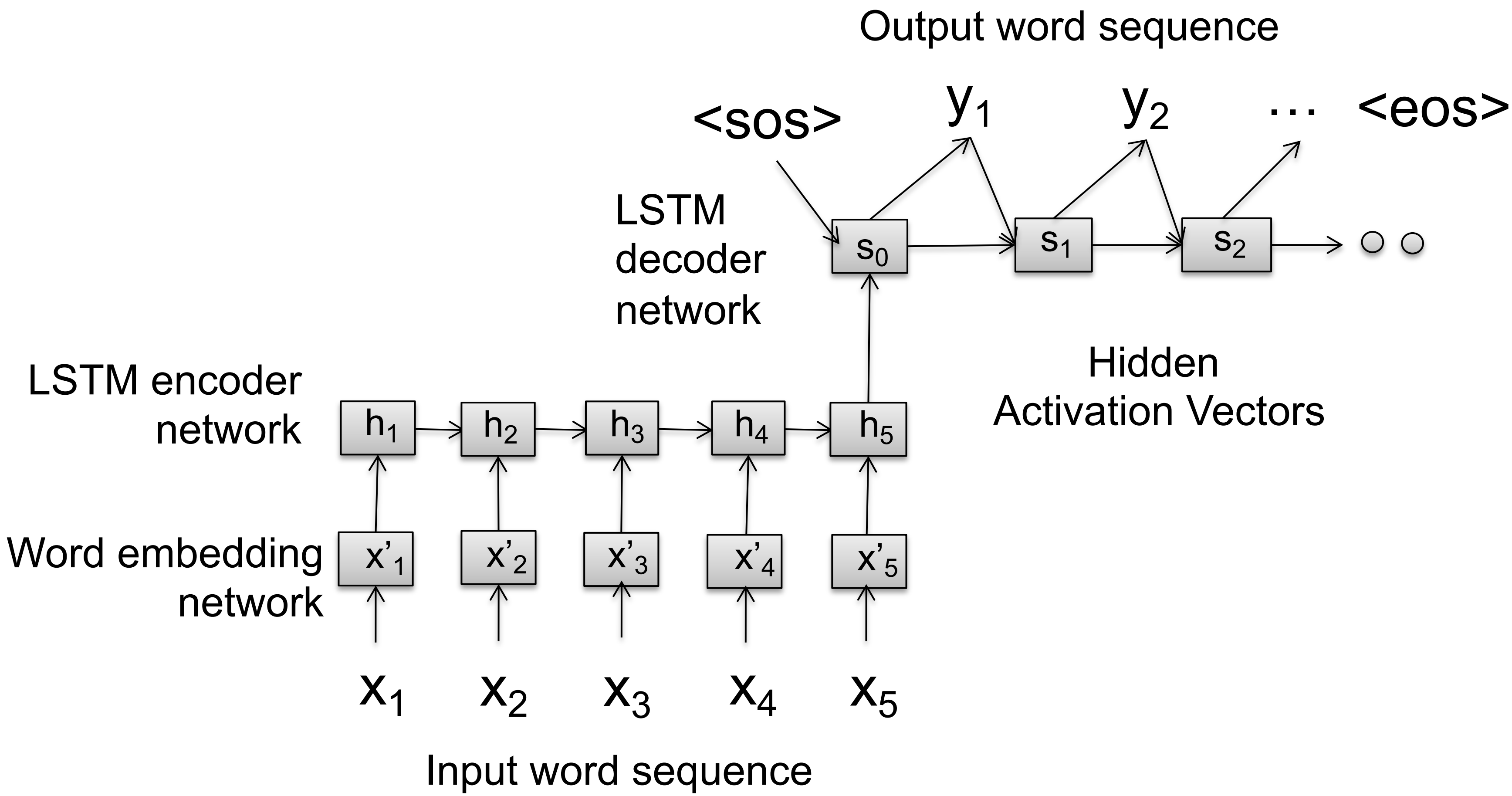
- End-to-end dialog systems to generate system responses directly from user inputs
  - Learn deep network model to generate system responses
    - **without annotating intermediate symbols**

Discrete State Graph





Neural Translation Models (Bahdanau+'14)
Sentence-to-Sequence Models



Neural Conversation Models (Vinyals+'15)

- Train from OpenSubtitles

A pair of two sentences were trained without context.

Various movie characters are mixed in the system role.

Conversation 4: Philosophical Q&A

Human: *what is the purpose of life ?*
Machine: *to serve the greater good .*
Human: *what is the purpose of living ?*
Machine: *to live forever .*
Human: *what is the purpose of existence ?*
Machine: *to find out what happens when we get to the planet earth .*
Human: *where are you now ?*
Machine: *i 'm in the middle of nowhere .*
Human: *what is the purpose of dying ?*
Machine: *to have a life .*
Human: *what is the purpose of being intelligent ?*
Machine: *to find out what it is .*
Human: *what is the purpose of emotions ?*
Machine: *i don 't know .*

MITSUBISHI ELECTRIC
Changes for the Better

DSTC6: Sentence-to-Sentence Generation Task (MERL 2016)

LoveHard @MyLovesLA · May 5
At the moment i am stuck in the middle of the road and the car will not turn on.. its not the battery! I shld not have this many issues with a NEW car!! @Honda

Honda Automobile Customer Service @HondaCustSvc · 1h
We are terribly sorry to hear about the battery concerns on your Civic. Have you been able to bring the vehicle to your Honda dealer for an inspection? Please feel free to DM us if there is anything we could look into for you. Your feedback is important to us. ^FC

1

LoveHard @MyLovesLA · 1h
Yes! Multiple times i feel like i live there! First they replaced the battery, then the alternator, now they are saying its the starter. A new car shouldnt be having this many issues!

1

Honda Automobile Customer Ser... Follow
@HondaCustSvc

Replying to @MyLovesLA

Thank you for the update. It is disappointing to hear this. Could you please share the most recent inspection details with us via DM? We would like to hear about this visit. Additionally, please provide us with your full name, VIN, mileage, phone #, & Honda dealer. ^FC

2:46 PM - 8 May 2018

Table 1: *Twitter data.*

	training	development	test
#dialog	888,201	107,506	2,000
#turn	2,157,389	262,228	5,266
#word	40,073,697	4,900,743	99,389

Context-dependent Response Generation

Evaluation:

- Comparison with 10 answers by humans
- Human rating

Context-Dependent System Response Generation

	New Phone	Bathroom Renovation	Bath product
USER	*I need to buy a new phone.*		*I love the new bath bombs!*
AGENT	*what phone are you looking for ?*		*we 're glad you like it !*
USER	*An android phone.*	*I want to renovate my bathroom.*	*Are the new flavors available yet?*
AGENT	*what phone do you have ?*	*we are the experts in bathroom remodeling . take a look ! no obligation consult : <URL>*	*yes !*
USER	*Where can I visit?*		
AGENT	*you can check out our new phones here : <URL>*	*you can check out our remodeling services : <URL>*	*you can find our store locator here : <URL>*

Milestones for Scene-aware Interaction

2015 2016 2017 2018 2019 2020 2021 2022

HMI:
Scene understanding based navigation " Follow silver car turning right "

1. Neural Text Dialog

Context based response generator
Customer service on Twitter
Input: dialog history + user input
* Data-driven + Annotation free

2. Video Description

Attentional multimodal fusion
YouTube video, Input: visual + audio feature

3. Audio Visual Scene-aware Dialog

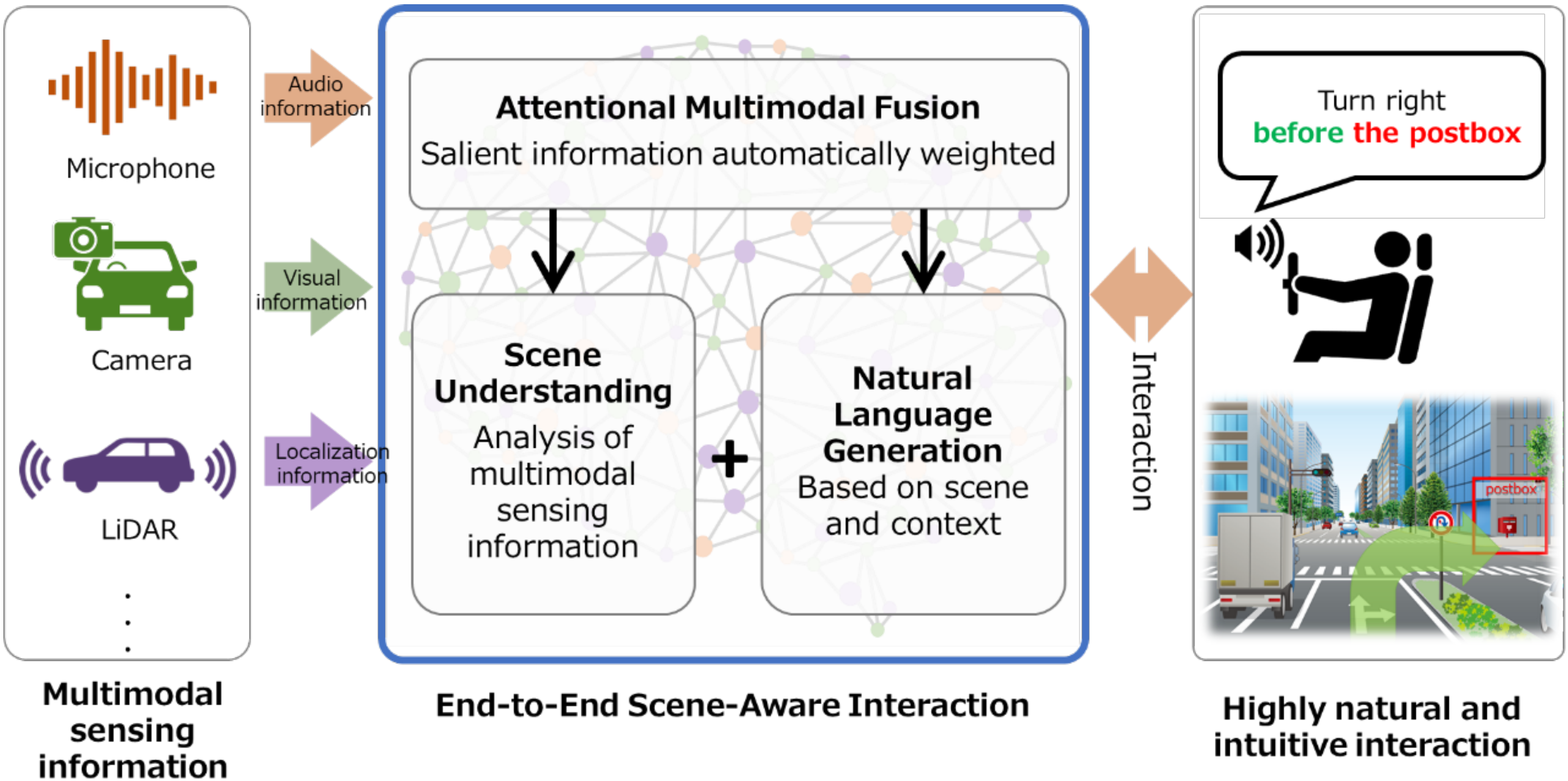
Joint Student-Teacher learning
Question answering system for action videos
Input: dialog history+ user input
+ audiovisual features

Scene-aware Interaction

Future Car navigation

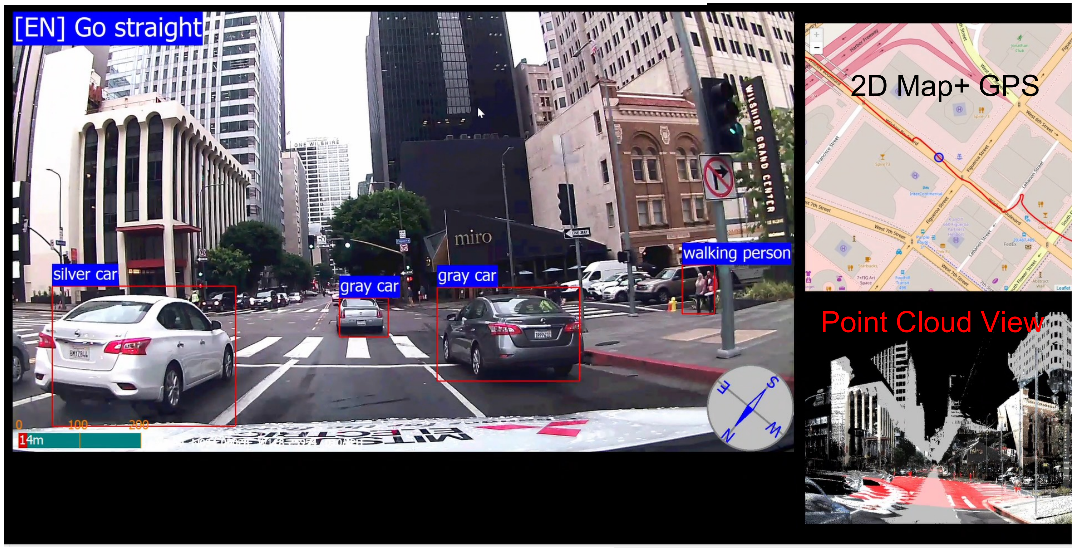
Scene-Aware Interaction

Car Navigation Use Case



Scene-aware Interaction for Car Navigation

Annotation 1: Bounding-box based Object recognition



Annotation 2: Sematic region segmentation

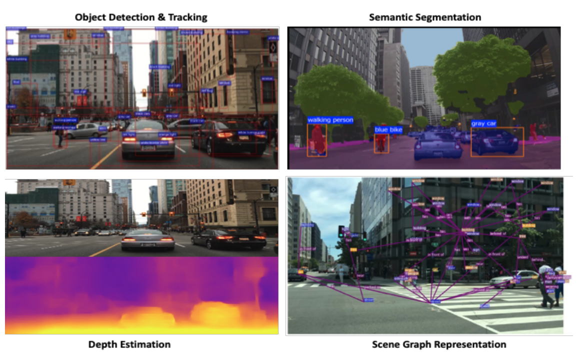
Scene-Aware Interaction

https://www.youtube.com/watch?v=t0izXoT_Aoc

Future Car Navigation

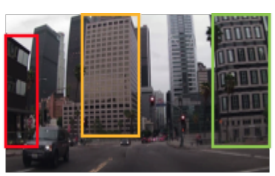
Turn right at the corner?

Yes!
After passing the gas station.

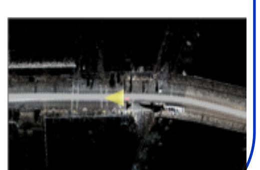


3D Point Cloud

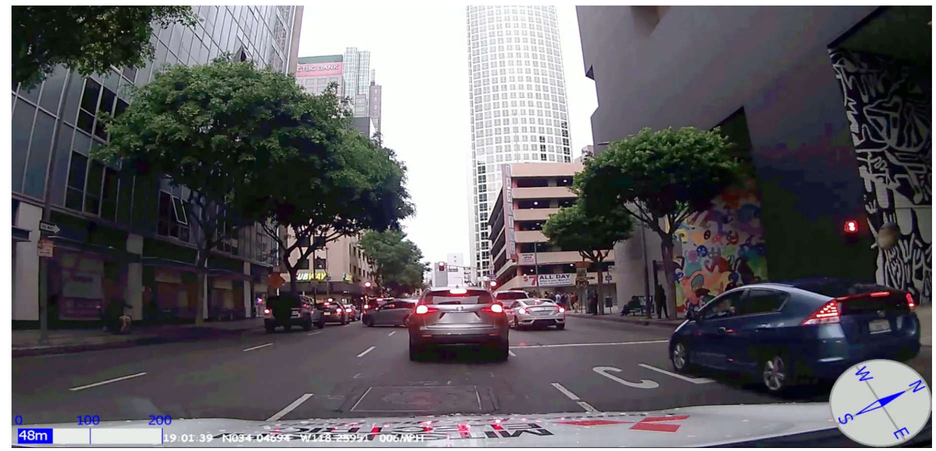
Camera View



Top-Down Map



Landmark-based Navigation

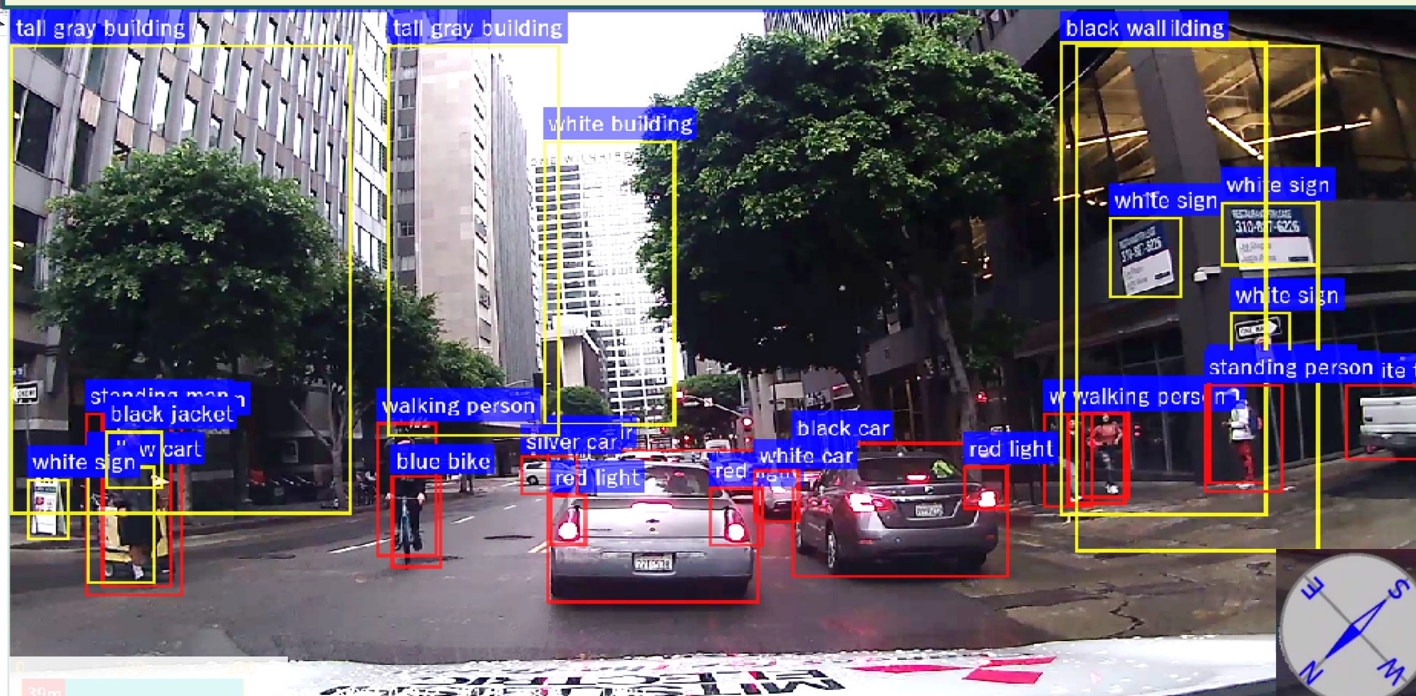


Tracking-based Voice Warning

Multimodal Feature Extraction 1

Original Video

Feature Extraction 1:
Bounding-box based object recognition

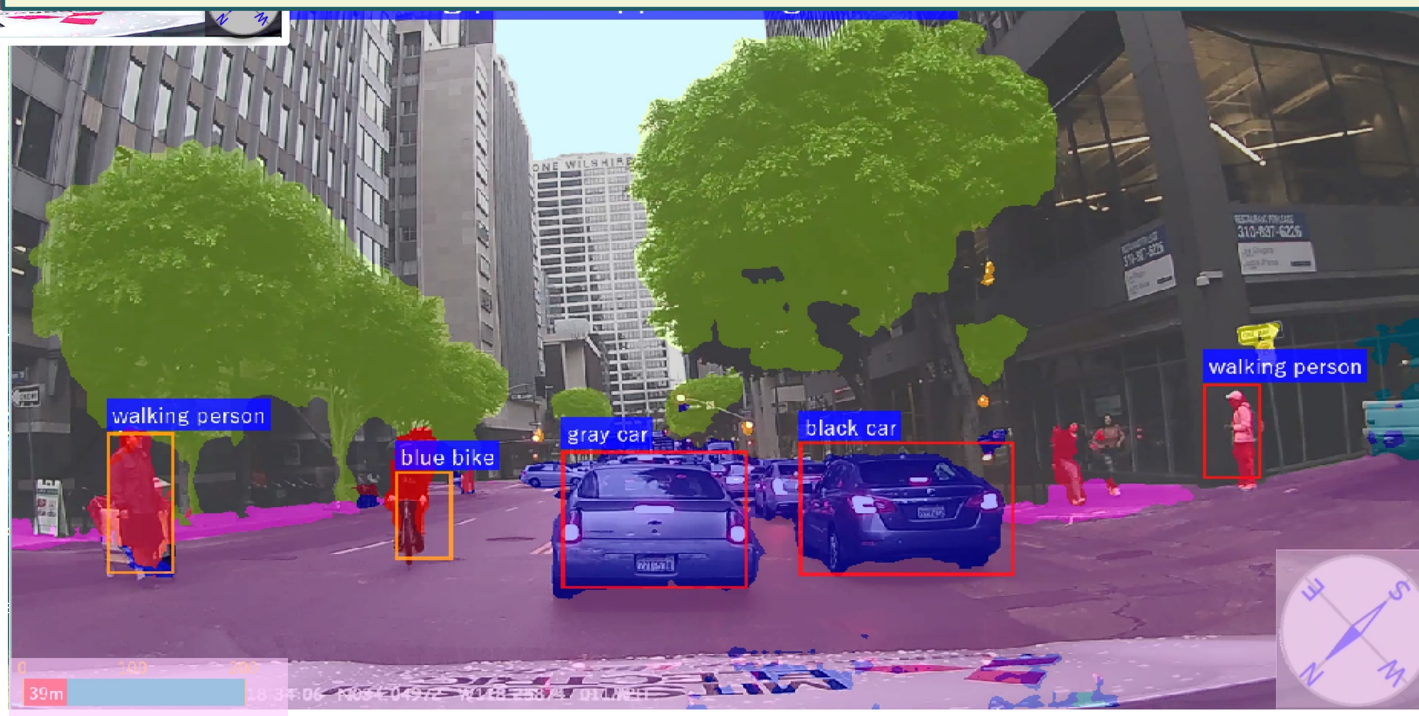


MITSUBISHI ELECTRIC
Changes for the Better

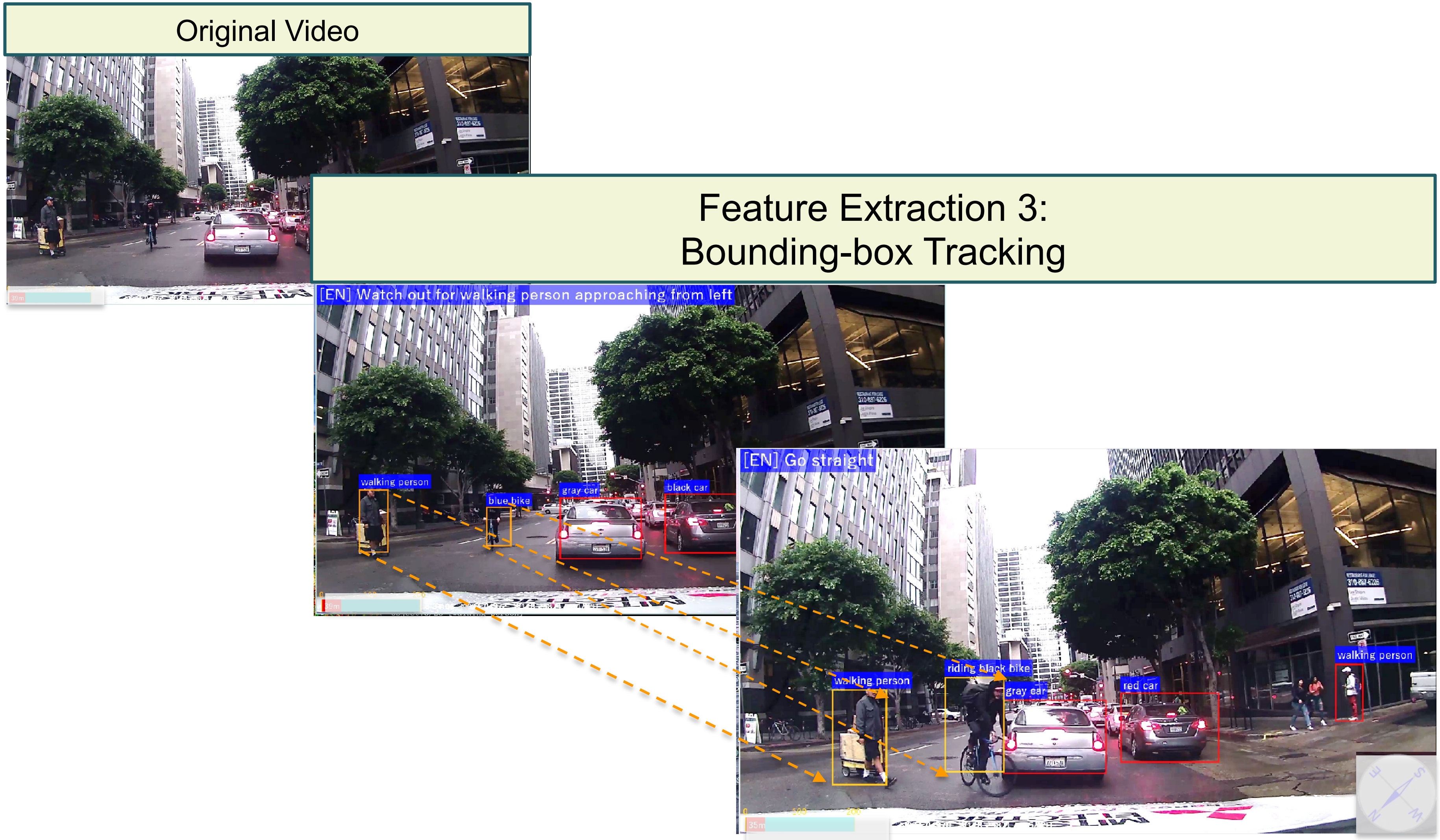
Multimodal Feature Extraction 2

Original Video

Feature Extraction 2
Sematic region segmentation

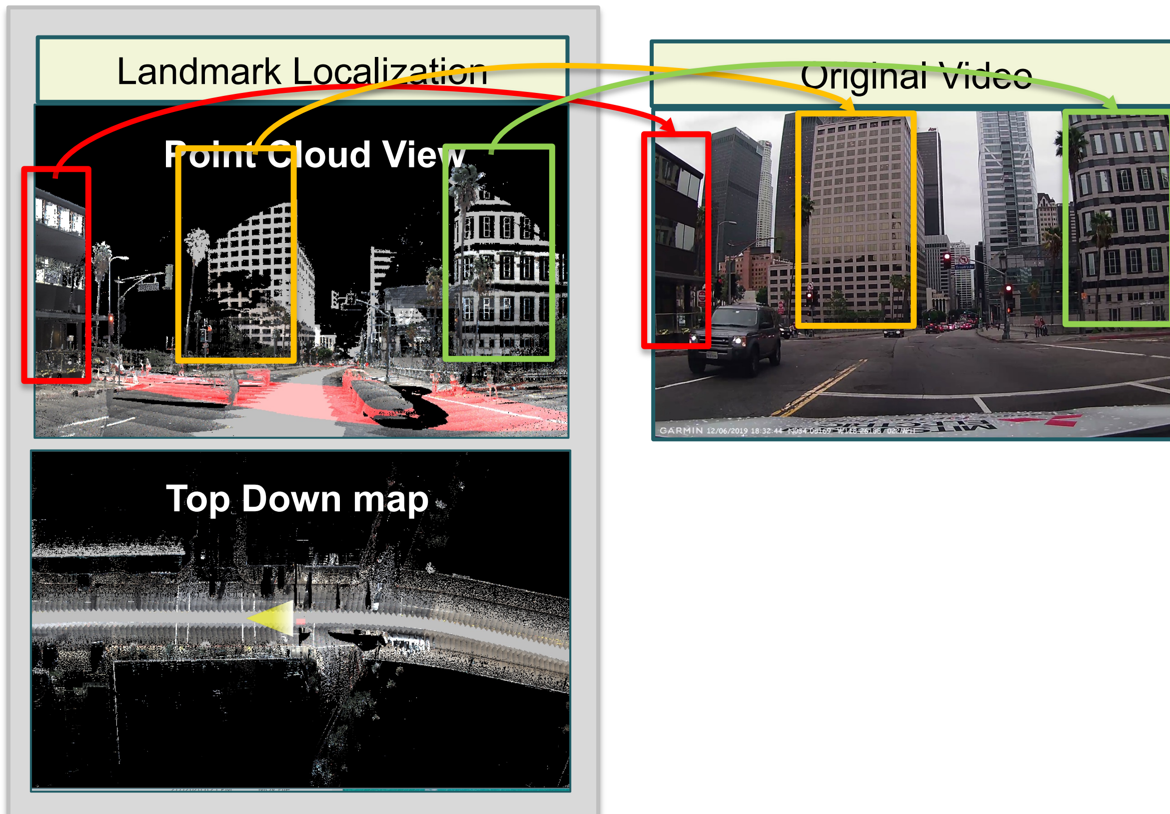


Multimodal Feature Extraction 3



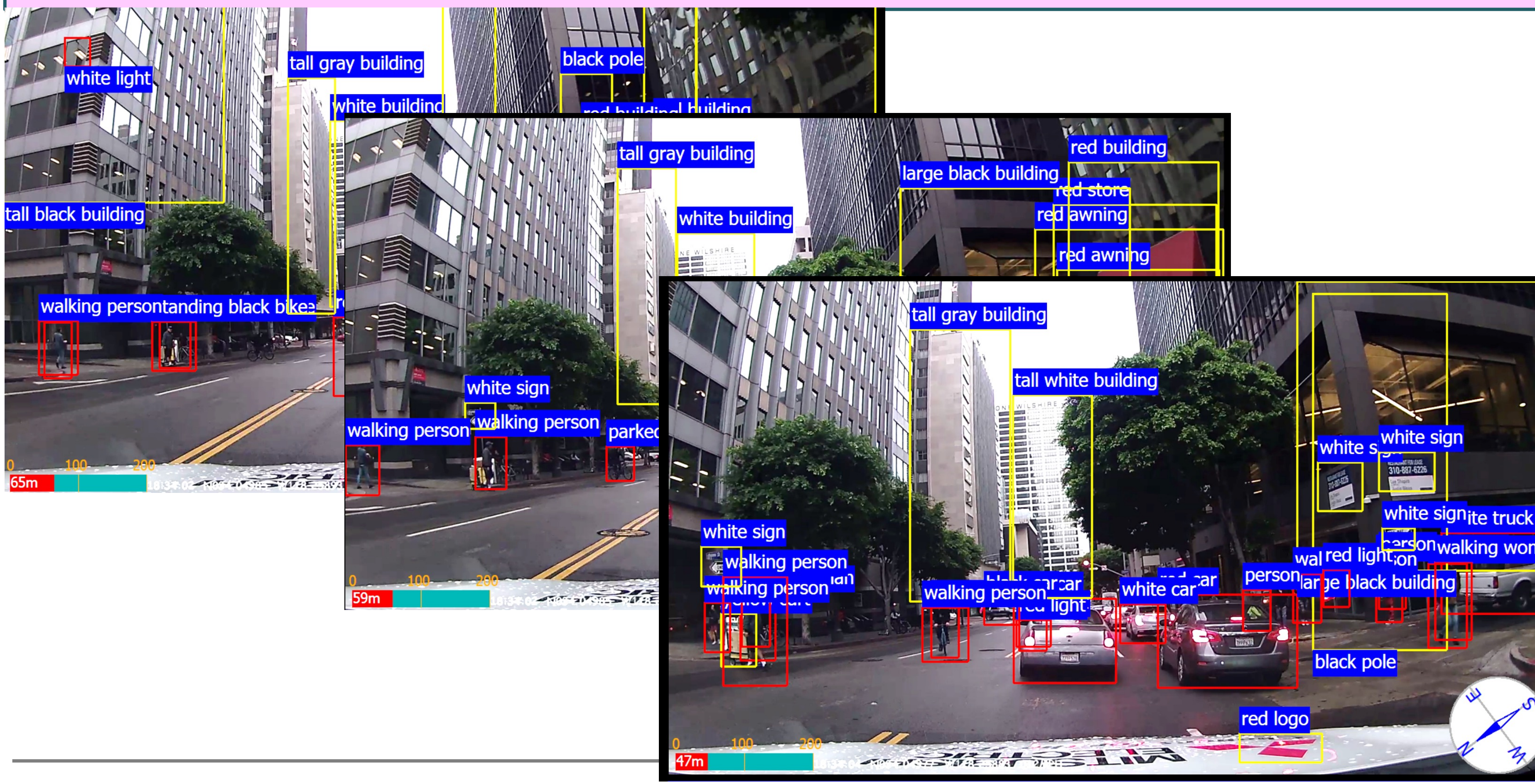
Multimodal Feature Extraction 4

Prerecorded data of Mobile Mapping System (MMS) provides object location in a view of streets. (http://www.mitsubishielectric.com/bu/mms/)



Object recognition results are changed

Feature Extraction 1:
Bounding-box based object recognition



MITSUBISHI ELECTRIC
Changes for the Better

Scene-aware Interaction for Daily Life Monitoring

Surveillance Camera Systems

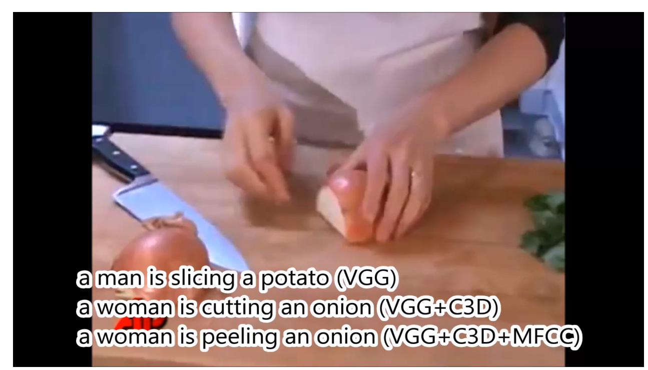
CNN: After Boston: The pros and cons of surveillance cameras

Ask seeking target:
“Find a small girl wearing a pink T-shirt”

Narrowing down by systems:
“Is she wearing a hat?”

Answer by users to add more information:
“Yes, she is wearing a straw hat.”

Scene understanding using video description



https://www.merl.com/demos/video-description

- Visual features: Object and event recognition
- Audio features: Audio event recognition
- Scene-understanding: Video description
- Dialog history: Context-based future prediction
- Response generation: Sequence-to-sequence generation

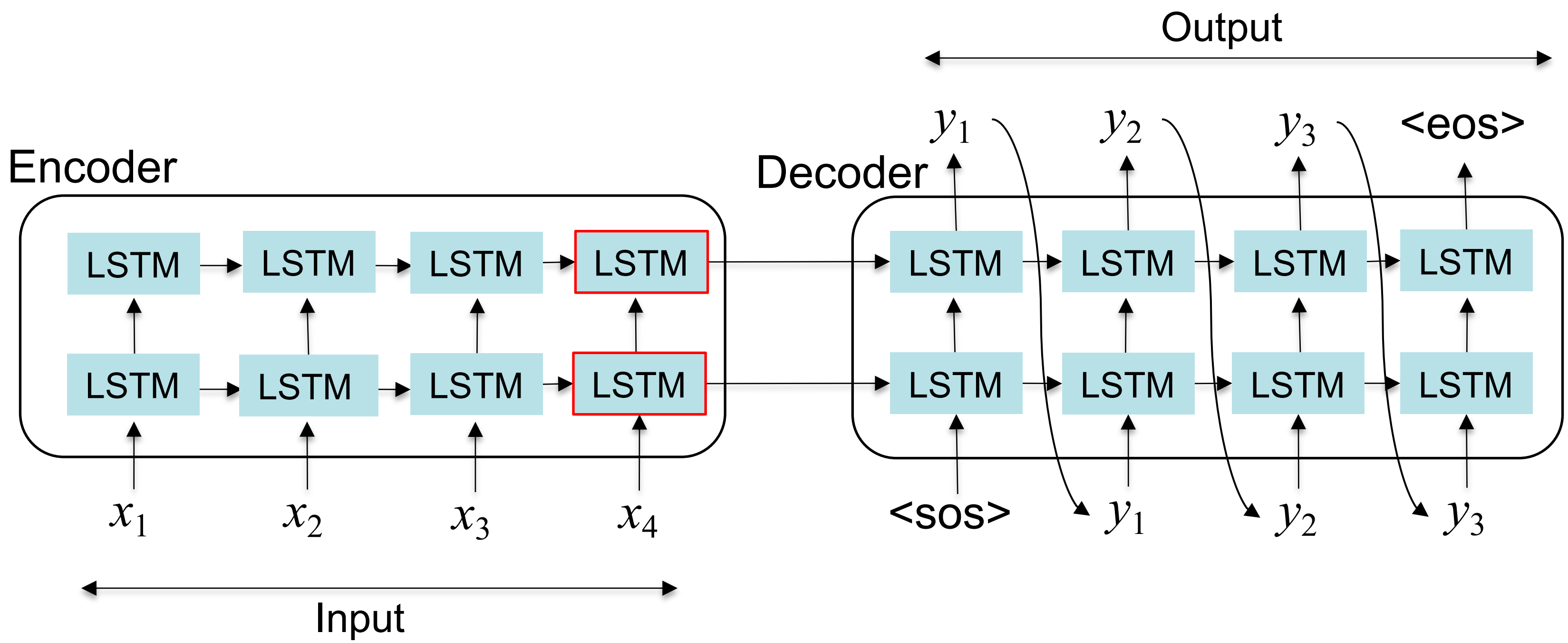
Sequence-to-sequence models

- Neural networks that can learn a mapping function between given input and output sequences in an end-to-end manner
  - LSTM encoder-decoder: Conversation+MT [Vinyals+'15]
  - Attention-based encoder decoder: MT [Bahdanau+'14]
  - Transformer: MT [Vaswani+'17]

- Widely used for various sequence-to-sequence tasks

Task	Input	Output
Speech recognition	Speech signal	Sentence text
Machine translation	Source language text	Target language text
Language understanding	Sentence text	Semantic label sequence
Dialog generation	User utterance	System response
Video description	Image sequence	Sentence text

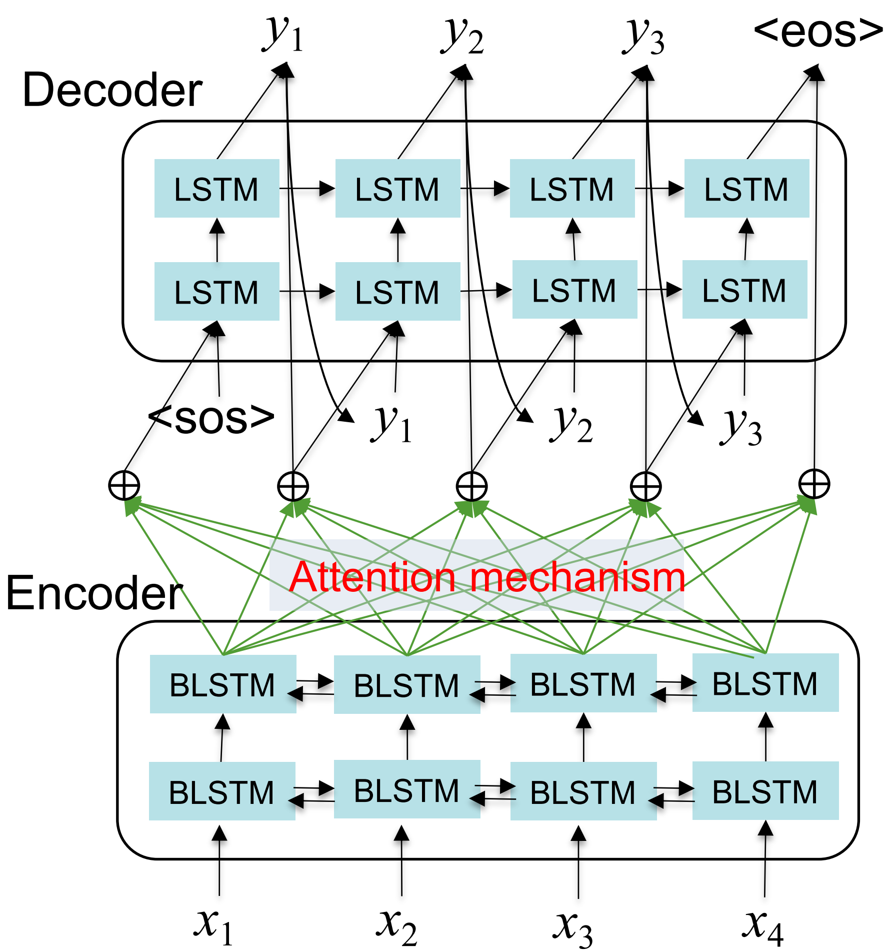
LSTM encoder decoder [Vinyals+'15]



LSTM: Long Short-Term Memory

Pros: Simple recurrent architecture with LSTM cells, which can memorize relatively long contextual information compared to vanilla RNNs losing contextual information exponentially.
Cons: Information of long input sequences may be lost by summarizing the sequence into a fixed dimensional vector in the last state.

Attention-based encoder decoder [Bahdanau+'14]



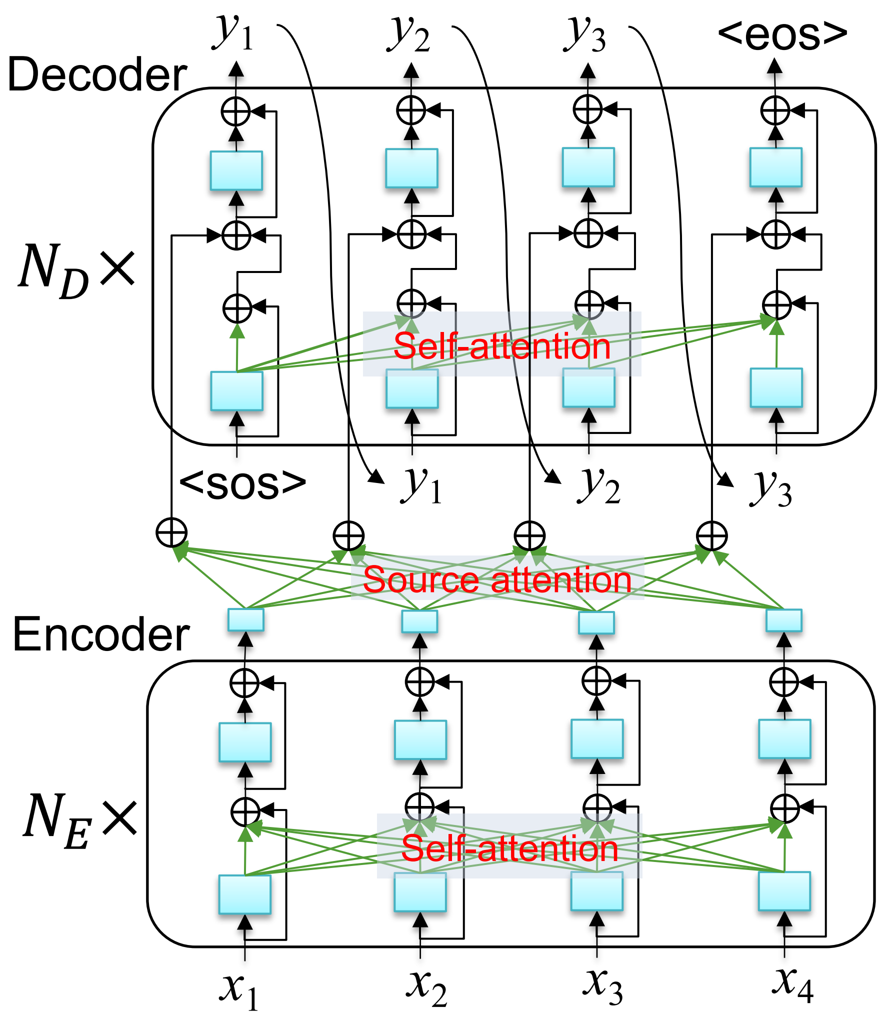
Pros

- BLSTM encoder captures bidirectional dependency within input sequence
- Attention mechanism allows the decoder to access full encoder outputs

Cons

- BLSTM (or LSTM) can utilize only adjacent state information, which may be insufficient to capture long context dependency over the input (or output) sequence

Transformer [Vaswani+'17]



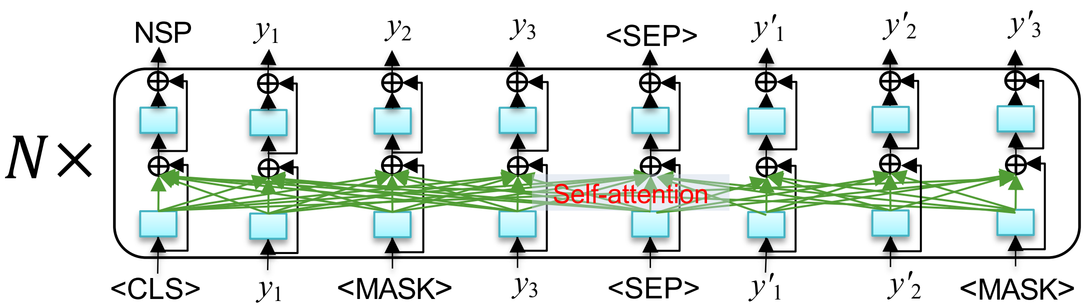
Pros

- Feed-forward network with residual connections enables to learn very deep architecture, which significantly improves the accuracy
- Self-attention mechanism allows to utilize full-sequence context in the encoder and the decoder
- The most successful model at the moment for many tasks

MITSUBISHI ELECTRIC
Changes for the Better

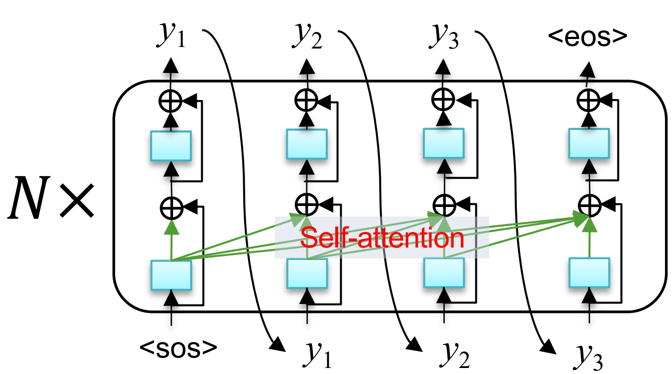
Large-scale pretrained language models (LMs) for various sequence-to-sequence tasks in NLP

- **BERT**: Transformer LM that predicts randomly masked words and next sentence or not [Devlin+'19]



- Feed sentence pairs with <SEP> symbol (e.g. QA pairs)
- Need task-specific labeled data for fine-tuning
- Achieve state-of-the-art performance on various NLP tasks

- **GPT-3**: Transformer LM that simply predicts next words [Brown+'20]



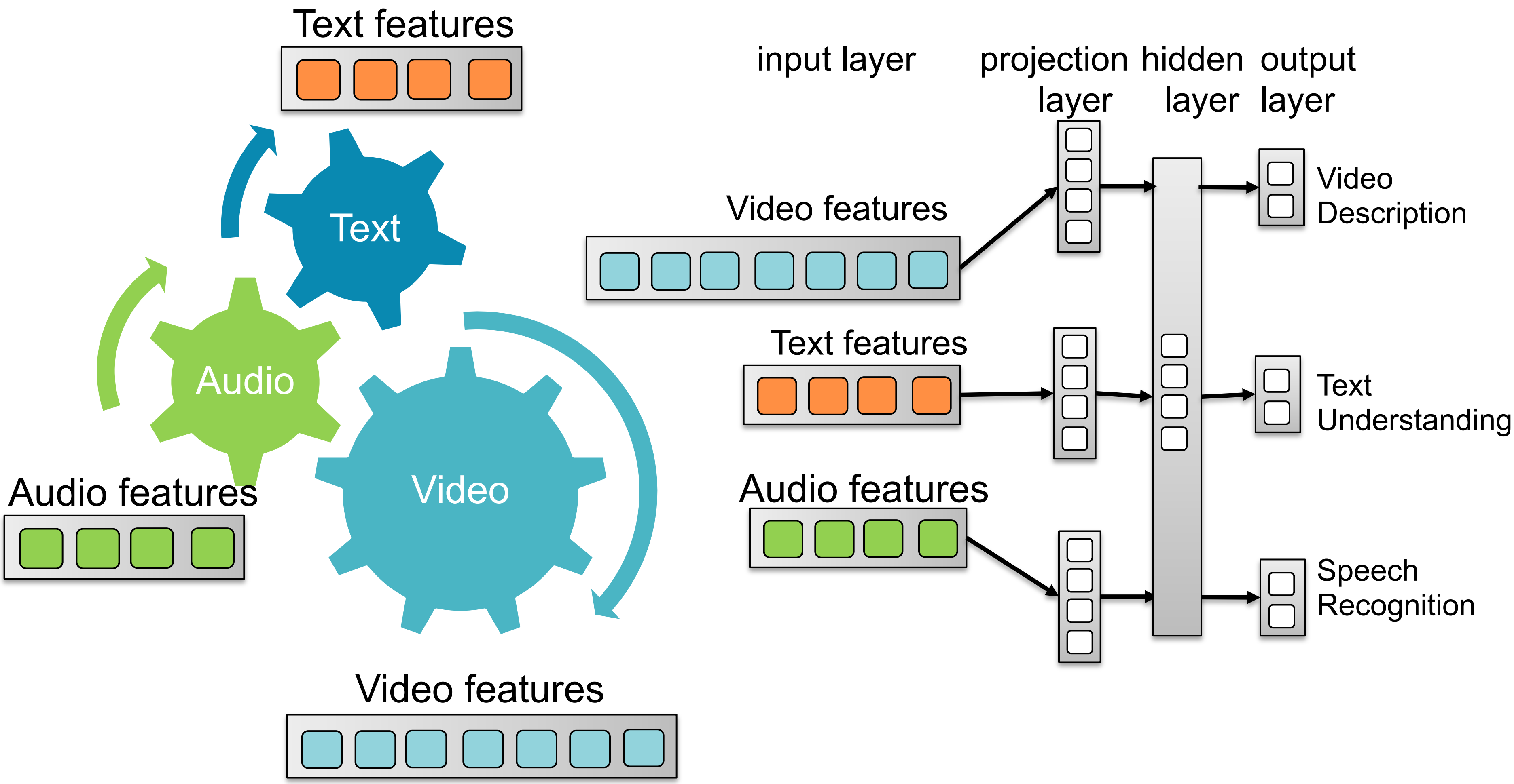
- Can be applied to various NLP tasks without fine-tuning
- Achieve state-of-the-art performance on several tasks by just providing a task specifying sentence and a few input/output examples in inference time

- The largest GPT-3 has 175 billion parameters!!
  ~= 500x larger than the largest BERT model
- Still difficult to generate natural long documents

Progress of AI platforms

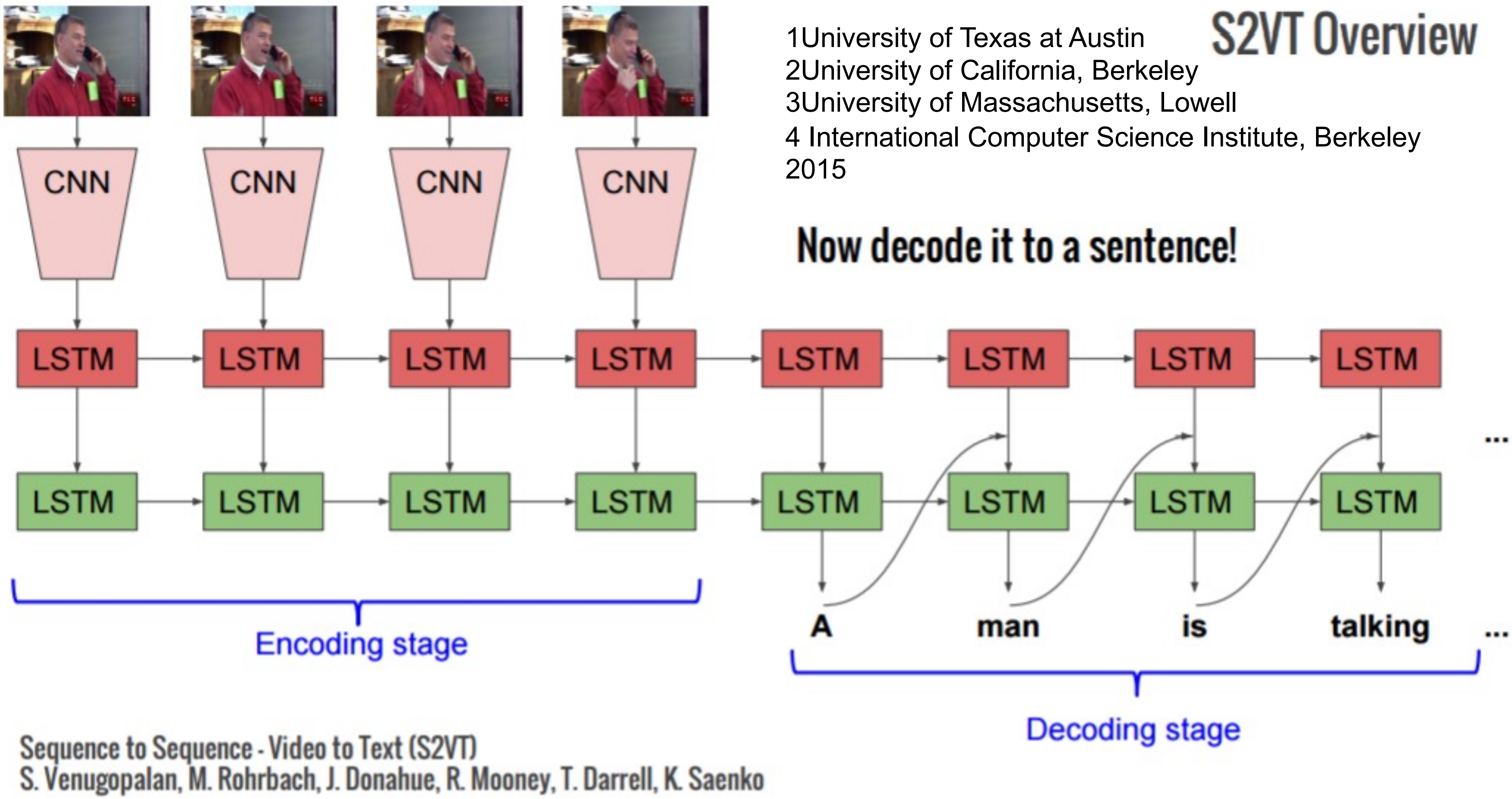
- Hardware
  - GPU (Nvidia, AMD, …), TPU, FPGA, … #cores, clocks, and memory are increasing (e.g. New Nvidia A40 has 10752 cores, 48GB memory, ...)
- GPU libraries
  - CUDA, CUDAToolkit, OpenCL, …, useful and efficient
- Deep learning Toolkits
  - Caffe, Theano, Torch, CNTK, Chainer, MXNet, ... TensorFlow, PyTorch
  - Easy implementation of complicated network architecture and training/testing procedure by Python scripting
    - Build computational graphs in advance (TensorFlow)
    - Define-by-run (PyTorch, originally from Chainer)
- Publicly available code (e.g. GitHub) and models (e.g. Model zoo)
- Nice ideas (+ computational resources and data) are important!

Semantic Representation using Audio, Video and Text Features



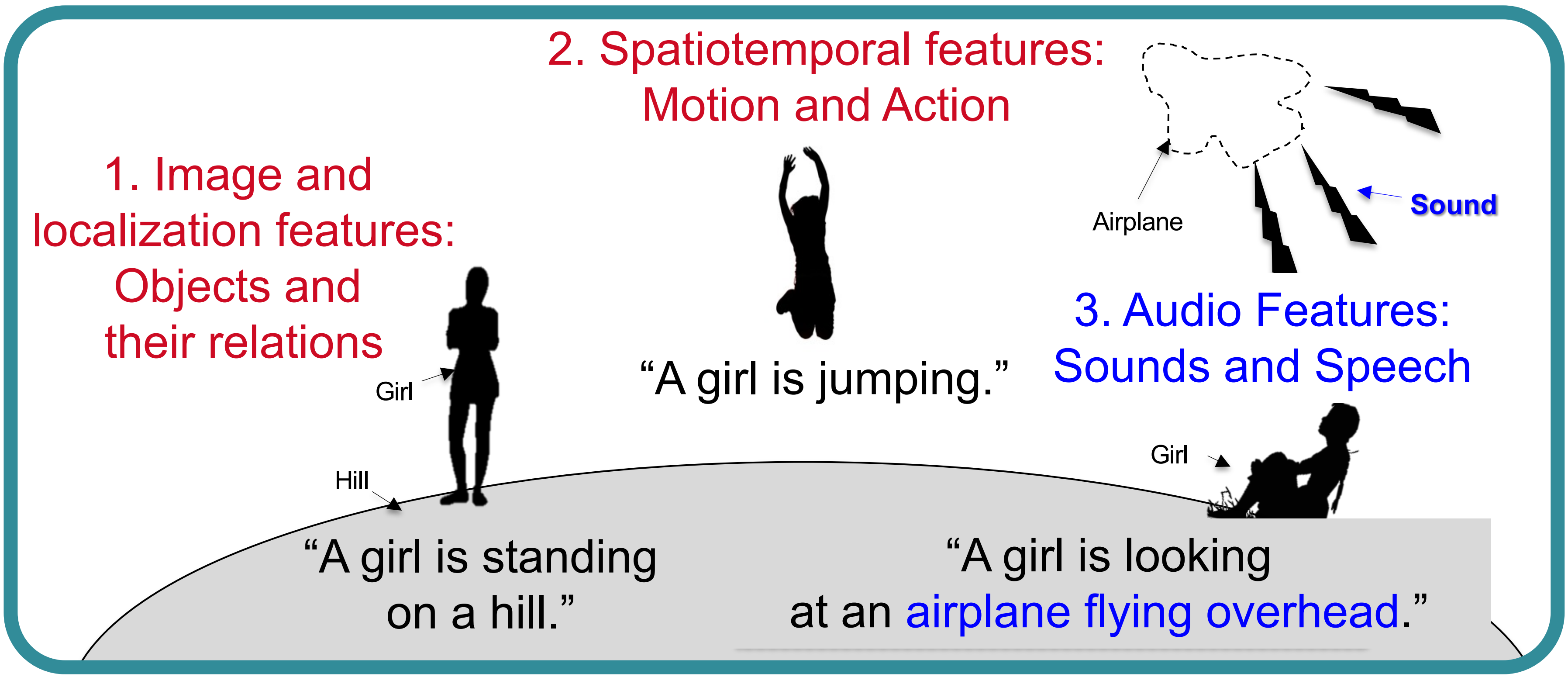
Encoder-decoder LSTM for Video Description

1University of Texas at Austin
2University of California, Berkeley
3University of Massachusetts, Lowell
4 International Computer Science Institute, Berkeley
2015



Multimodal Scene Understanding

Objects and events in scenes are recognized using multimodal information



Multimodal Fusion

Longstanding area of research:

"How to combine information from multiple modalities for machine perception?"

- Bayesian adaptation approaches

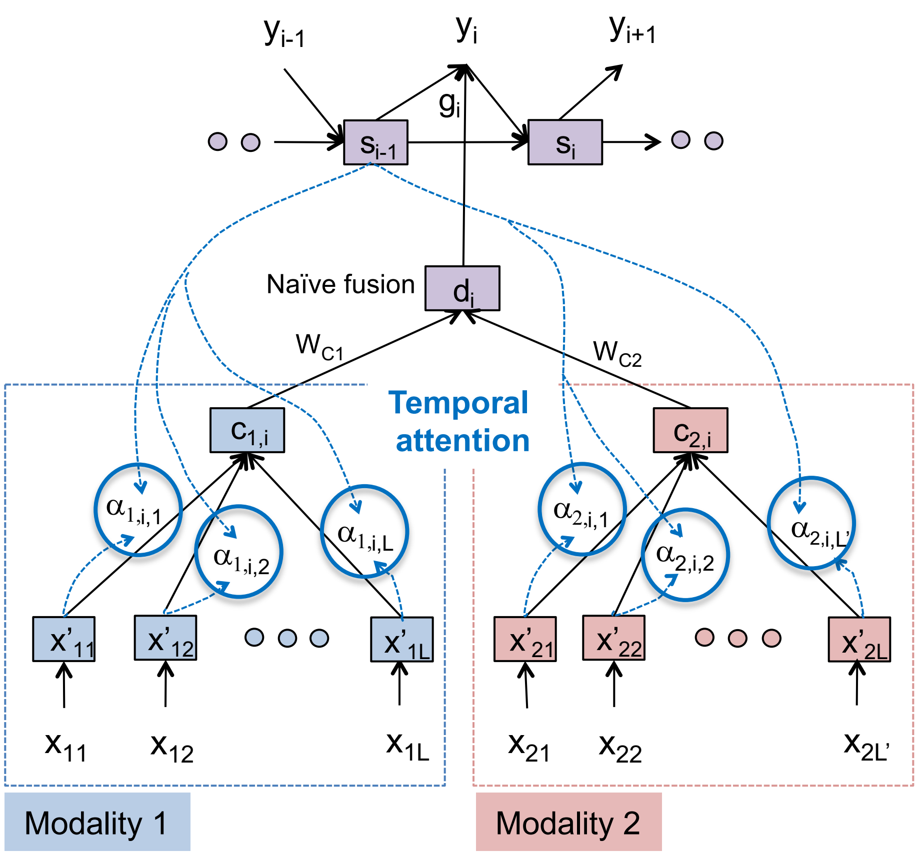
  J. R. Movellan and P. Mineiro. "Robust sensor fusion: Analysis and application to audio visual speech recognition. Machine Learning," 1998

- Stream weights

  G. Gravier, et al. "Maximum entropy and mce based hmm stream weight estimation for audio-visual asr," ICASSP, 2002

The first to fuse multimodal information using attention between modalities in a neural network

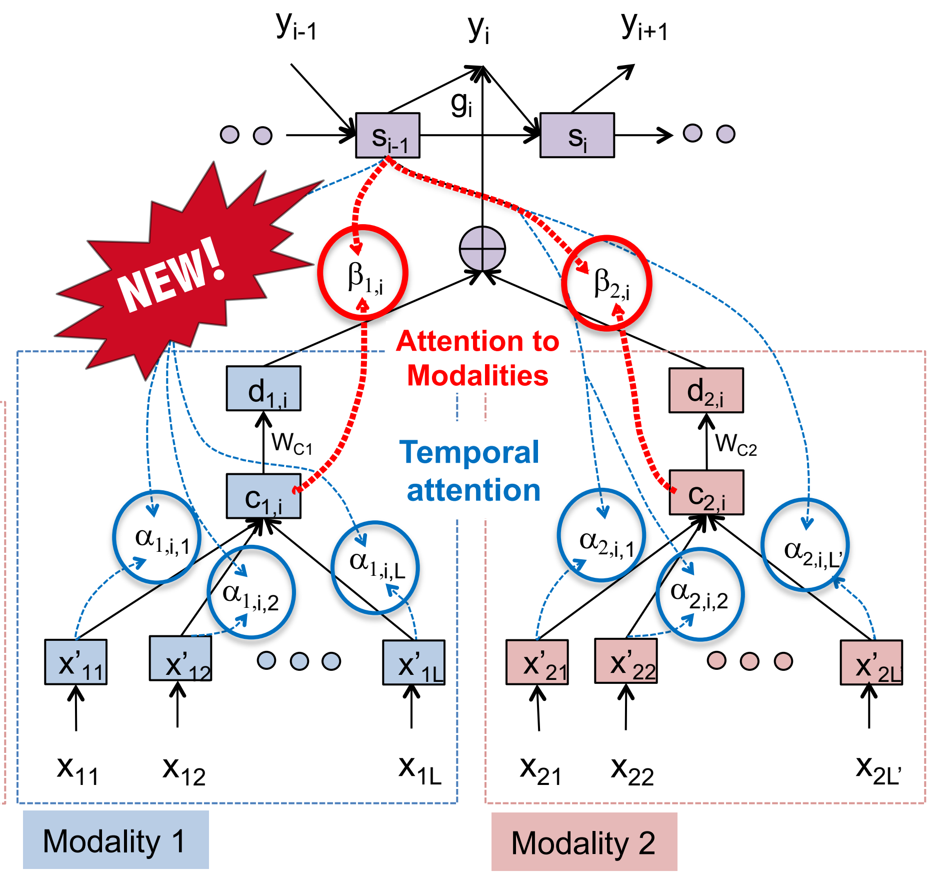
Naïve Fusion of Modalities



Context vector: weighted sum of frame features

Each modality projected into a common space

Attention-Based Multimodal Fusion



Attention weights for each input modality and input time

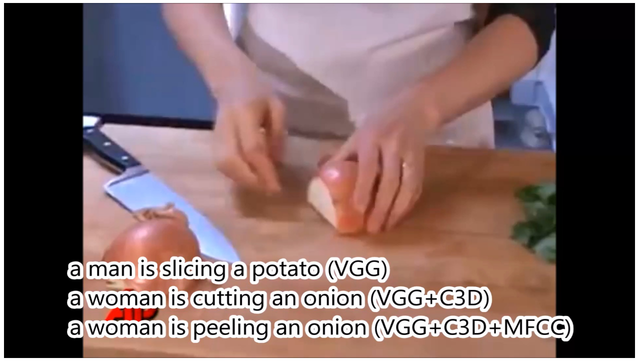
Selectively attends to specific modalities

Sample Videos with Automatic Description

Image: VGG, Motion: C3D, Audio: MFCC

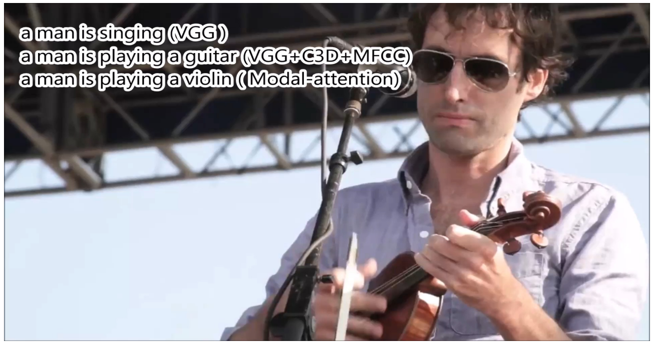
1. This video shows improvements due to our multimodal attention mechanism

2. Our use of audio features enable identification of peeling action.



3. Audio features make the description worse due to overdubbed music.

4. Our multimodal attention mechanism and audio features are complementary.



Unimodal < Naïve Multimodal Fusion < **Attentional Multimodal Fusion**

MITSUBISHI ELECTRIC
Changes for the Better

Words with strong average attention weights for each modality

Image (VGG-16)		Motion (C3D)		Audio (MFCC)	
bowl	0.9701	track	0.9887	talking	0.3435
pan	0.9426	motorcycle	0.9564	shown	0.3072
recipe	0.9209	baseball	0.9378	playing	0.2599
piece	0.9136	football	0.9275	singing	0.2465
paper	0.9098	horse	0.9212	driving	0.2284
kitchen	0.8827	soccer	0.9099	working	0.2004
toy	0.8758	basketball	0.9096	walking	0.1999

Our multimodal attention enables us to see which words rely most on each modality.

Image Features

Words describing **generic object type**

Motion Features:

Words describing scenes involving motion, such as **sports and vehicles**

Audio Features:

Action verbs associated with sound, such as talking, singing, and driving

AVSD

Audio Visual Scene-aware Dialog

Georgia Tech

Huda Alamri
Vincent Cartillier
Raphael G. Lopes
Abhishek Das
Irfan Essa
Dhruv Batra
Devi Parikh

MITSUBISHI ELECTRIC
Changes for the Better
MITSUBISHI ELECTRIC RESEARCH LABORATORIES, INC

Jue Wang
Gordon Winchern
Takaaki Hori
Anoop Cherian
Tim K. Marks
Chiori Hori

MITSUBISHI ELECTRIC
Changes for the Better

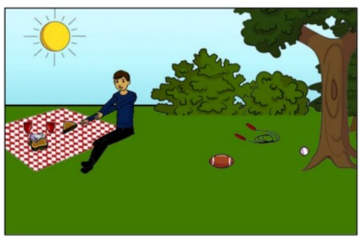
Georgia Tech

VQA to Visual Dialog

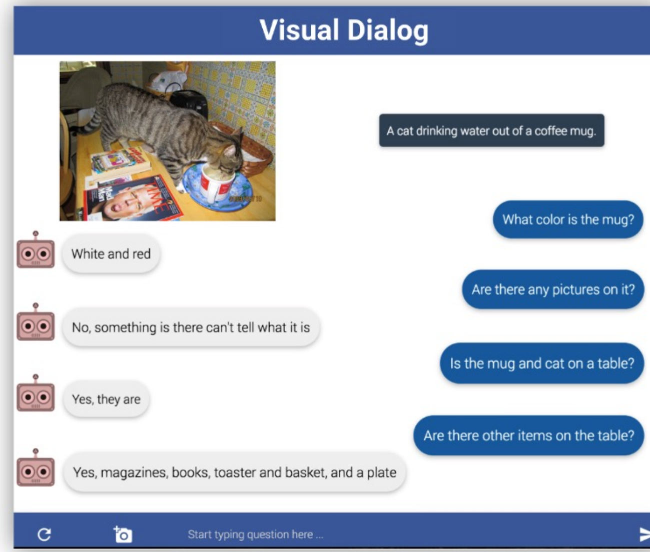
Devi Parikh, and Dhruv Batra


What color are her eyes?
What is the mustache made of?


How many slices of pizza are there?
Is this a vegetarian pizza?


Is this person expecting company?
What is just under the tree?

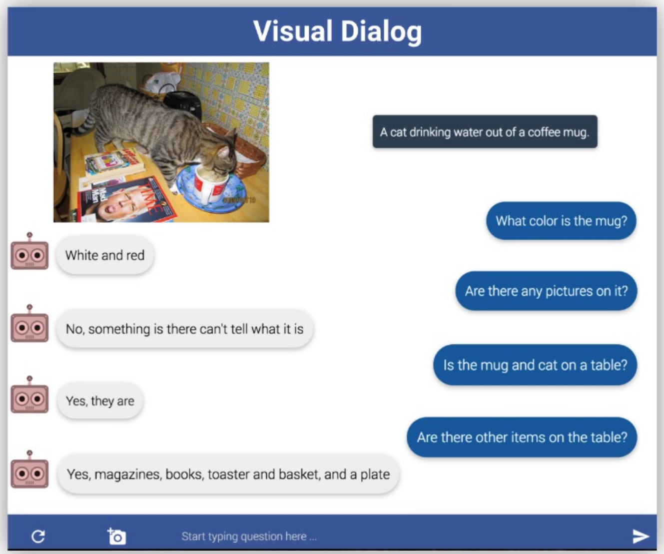

Does it appear to be rainy?
Does this person have 20/20 vision?

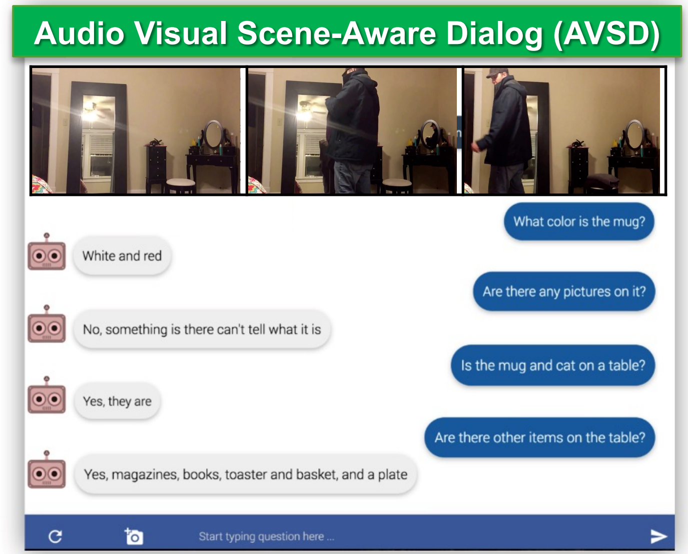


VQA:
Answering a question about an image

Visual Dialog:
Dialog discussing an image

Visual Dialog to
Audio Visual Scene-Aware Dialog





Visual Dialog:
Dialog discussing an image

AVSD:
Dialog discussing a video

http://www.video-dialog.com/

Audio Visual Scene-aware Dialog

A dataset with different modalities:

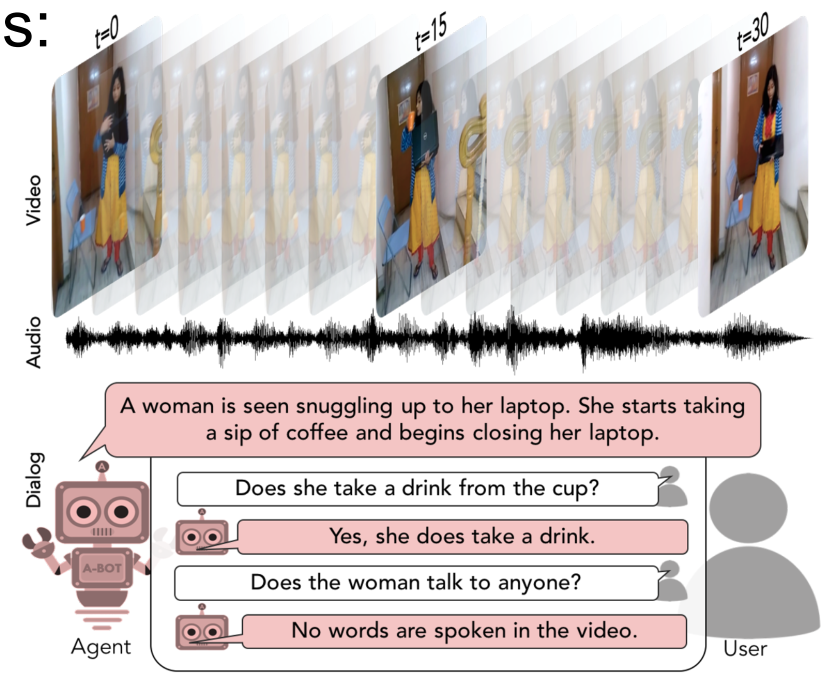
Video

Audio

Dialog history

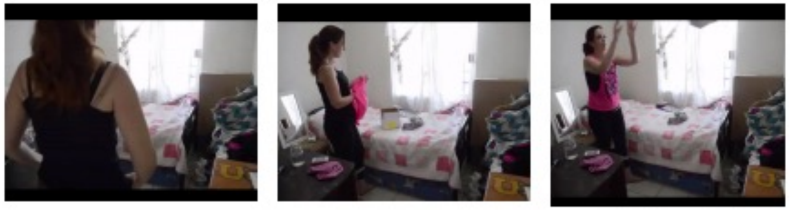
New question

→ Ground truth answer



AVSD data set

Questioner Interface



Answerer Interface

Script: A person is picking clothes out of a box and getting dressed and then throwing it across the room.

1. Is there just the one person in the scene?	1. Yes there is a woman.
2. Is she in the room the whole time?	2. Yes she walks towards the bed carrying a box.
3. What does she do with the box?	3. She puts it on the bed, opens it and pull out clothing.
4. What does she do next?	4. She fold up the empty box and tosses it.
5. Where does she toss the box?	5. She just tosses it towards the wall.
6. Does she say anything?	6. No she doesn't say anything
7. Do you hear anything in the background?	7. Just the closing of the box and it hitting the wall
8. Does she walk into the room carrying the box?	8. Yes it is in her hands
9. What does she do after she puts the clothes on?	9. She folds the box closed and tosses it
10. Does she walk away with the new shirt on?	10. The scene ends as the box hits the floor. She doesn't walk away

Questioner Summary

A person walks into a bedroom with a box. She opens the box, takes out a new shirt and puts it on.

Table 1: The dialog data for the AVSD track. The test videos for this challenge were selected from the official test data of the Charades dataset.

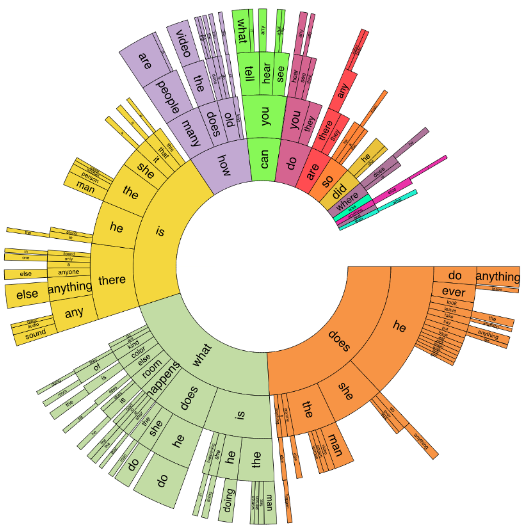
	Training	Validation	DSTC7 Test	DSTC8 Test
# of dialogs	7,659	1,787	1,710	1,710
# of turns	153,180	35,740	13,490	18,810
# of words	1,450,754	339,006	110,252	178,619

Sentence Selection or Sentence Generation

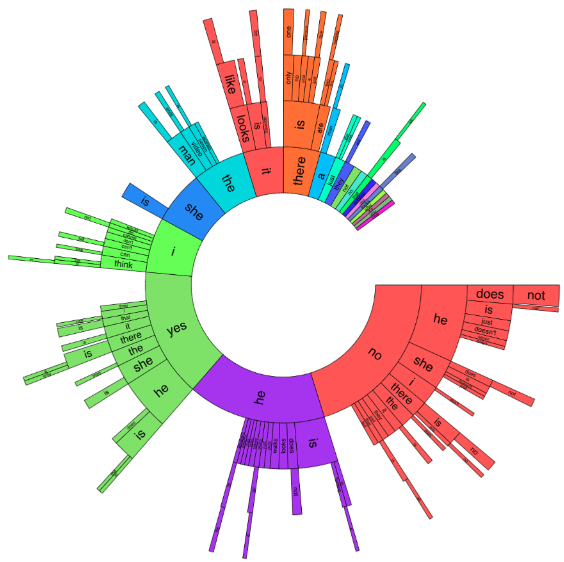
- Sentence selection:
  - Information Retrieval framework
  - Difficulty depends on how to prepare multiple candidates

- Sentence generation:
  - Speech recognition/Machine translation framework
  - Answers generation depends on language models

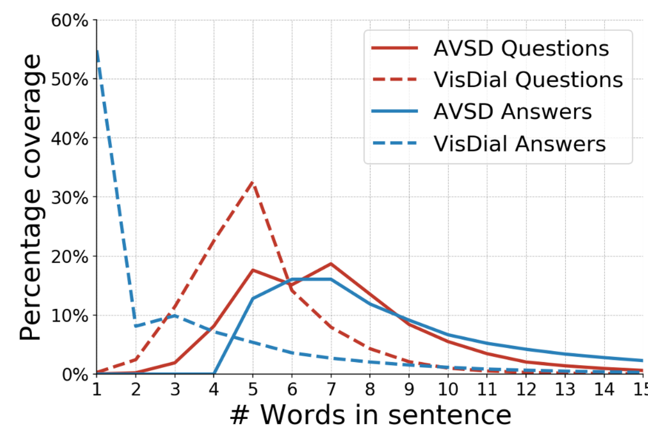
N-gram Distribution for Questions and Answers



AVSD Questions

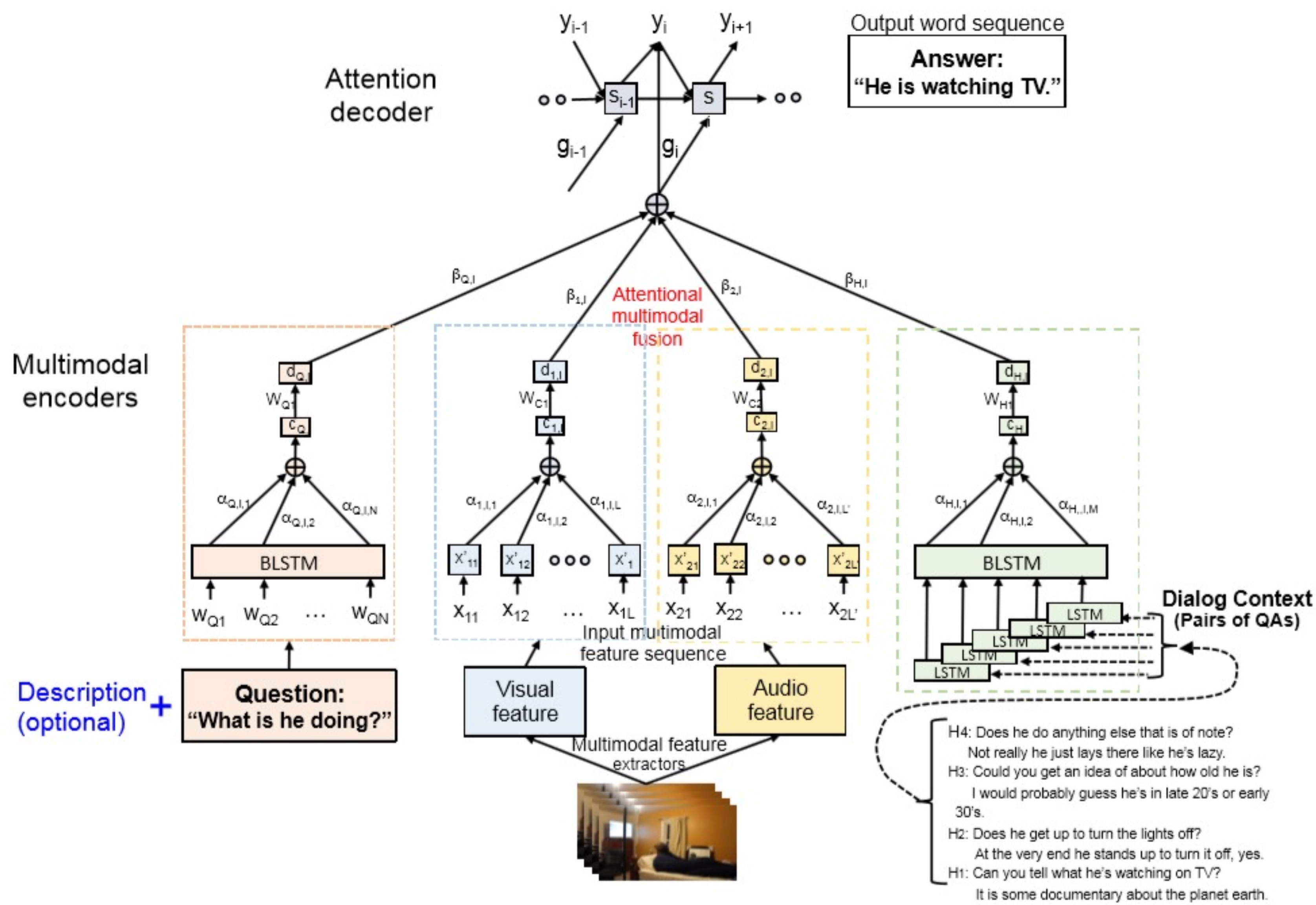


AVSD Answers



AVSD sentence lengths

Multimodal Dialog

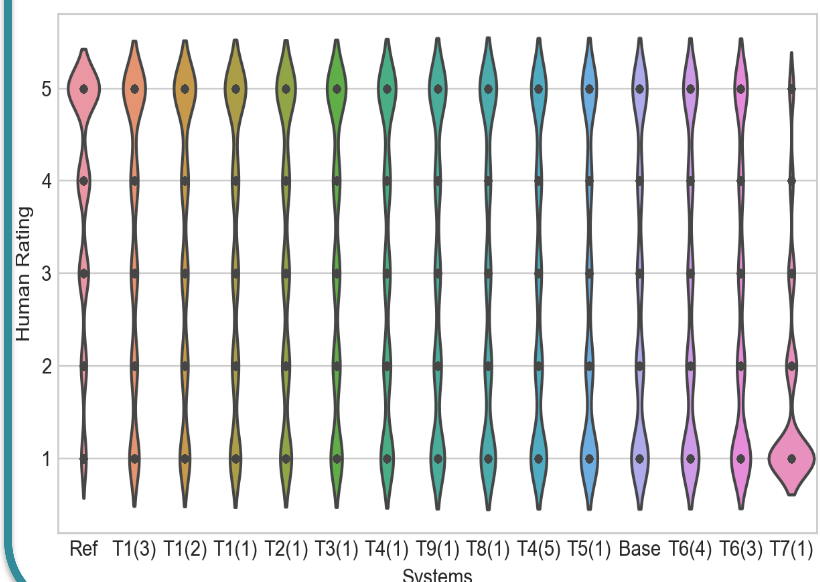


MITSUBISHI ELECTRIC
Changes for the Better

Open Research Platform

2018

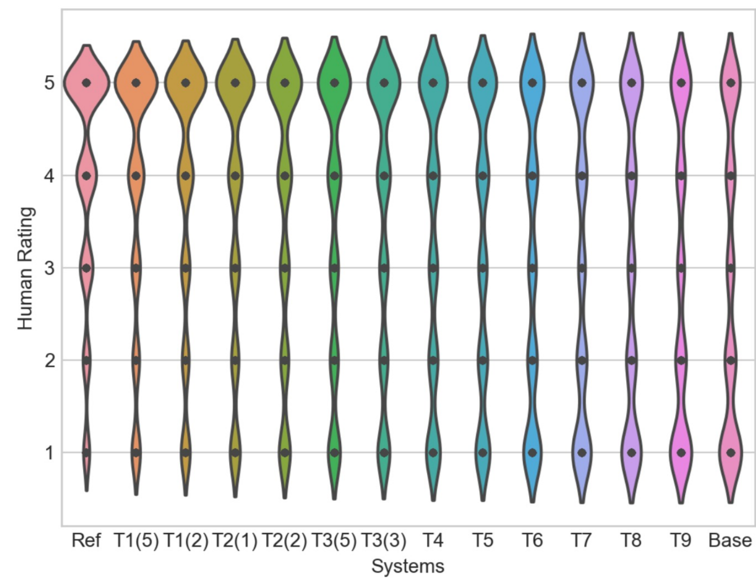
DSTC7



Special Issue on the "7th Dialog System Technology Challenge 2019" in Computer Speech and Language

2019

DSTC8
The Eighth Dialog System Technology Challenge



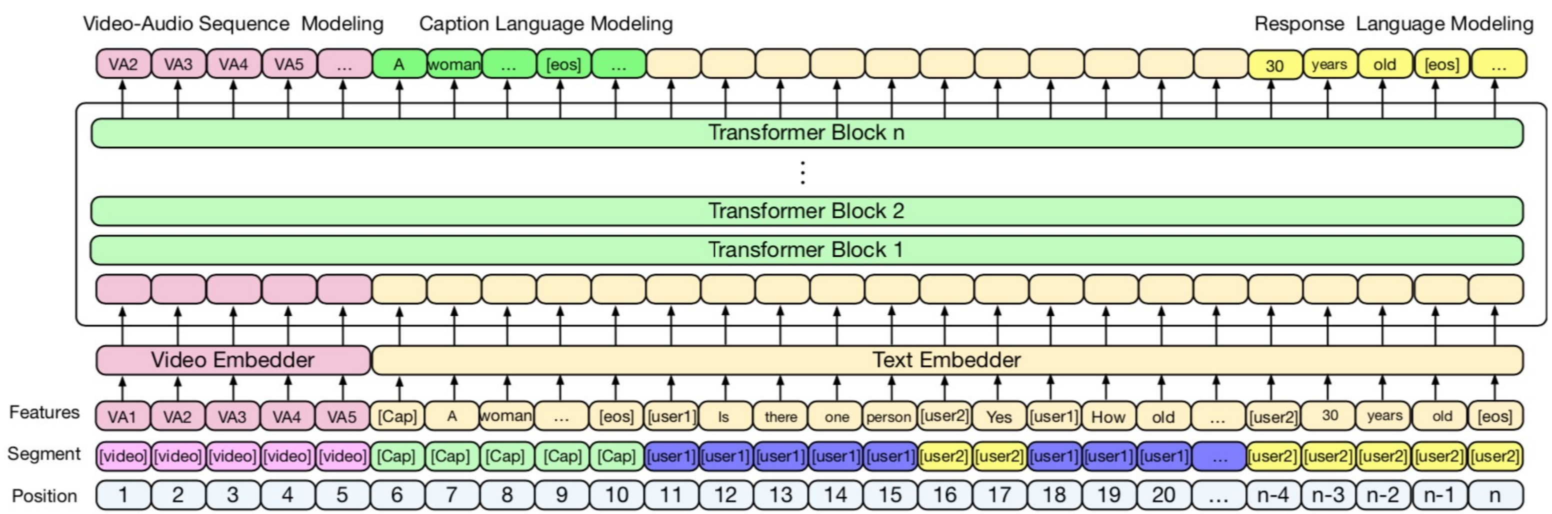
Special Issue on ”Eighth Dialog System Technology Challenge” in IEEE/ACM TASLP

DSTC8 Best System

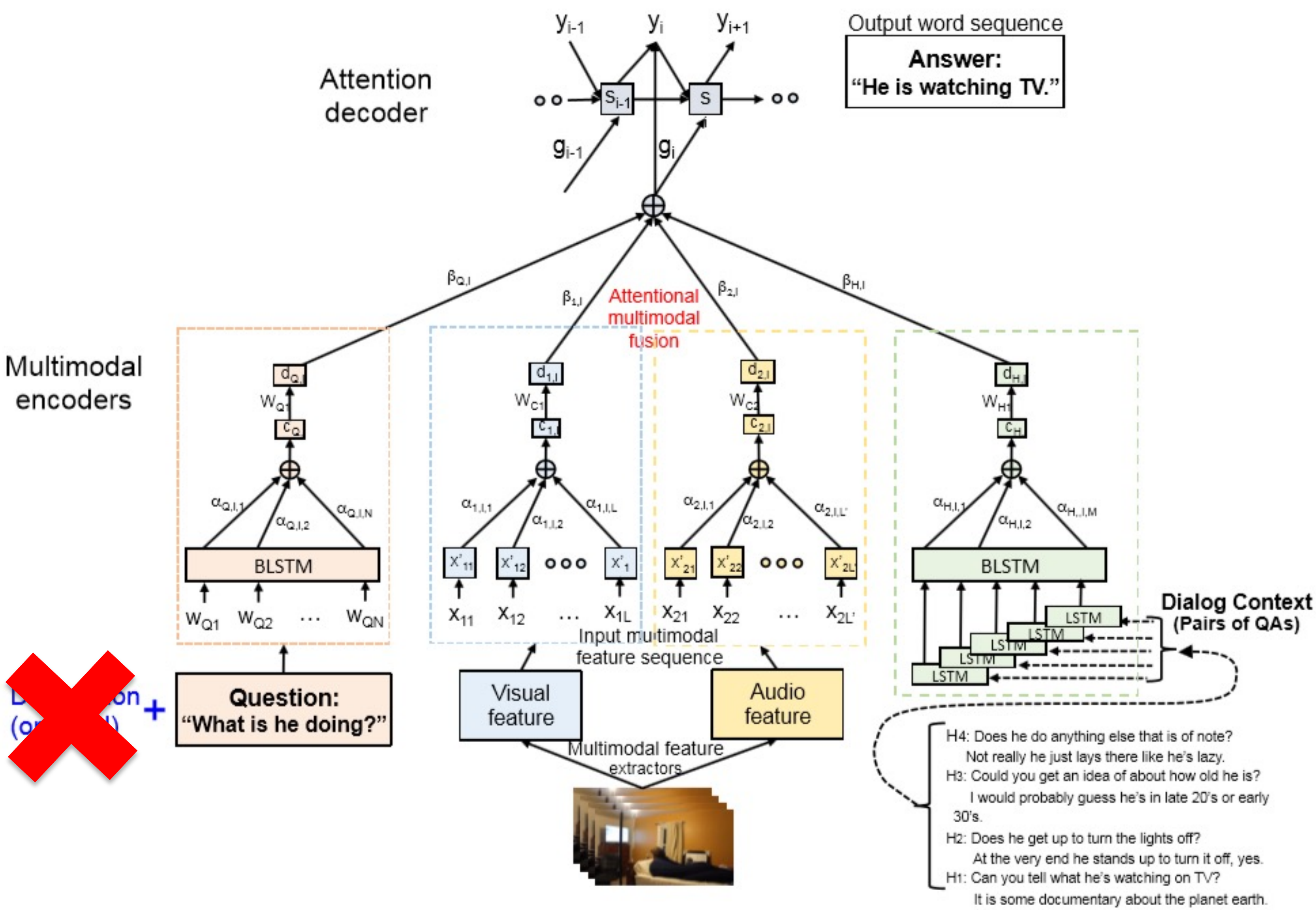
“Bridging Text and Video: A Universal Multimodal Transformer for Video-Audio Scene-Aware Dialog” by Zekang Li et al.

- Considering the similarity between the summary and the video caption, summary and caption are concatenated to be one sequence. U = {Q1, R1, Q2, R2, . . . QN , RN , } to denote the N turns of dialogue,
- Qn: n question, Rn: n response n containing m words.
- Probability to generate the response Rn for the given question Qn considering video V, audio A, dialogue history U

$$P(\mathbf{R}_n|\mathbf{V},\mathbf{A},\mathbf{C},\mathbf{U}_{<n},\mathbf{Q}_n;\theta) = \prod_{j=1}^{m} P(r_{nj}|\mathbf{V},\mathbf{A},\mathbf{C},\mathbf{U}_{<n},\mathbf{Q}_n,r_{n,<j};\theta) \quad (1)$$



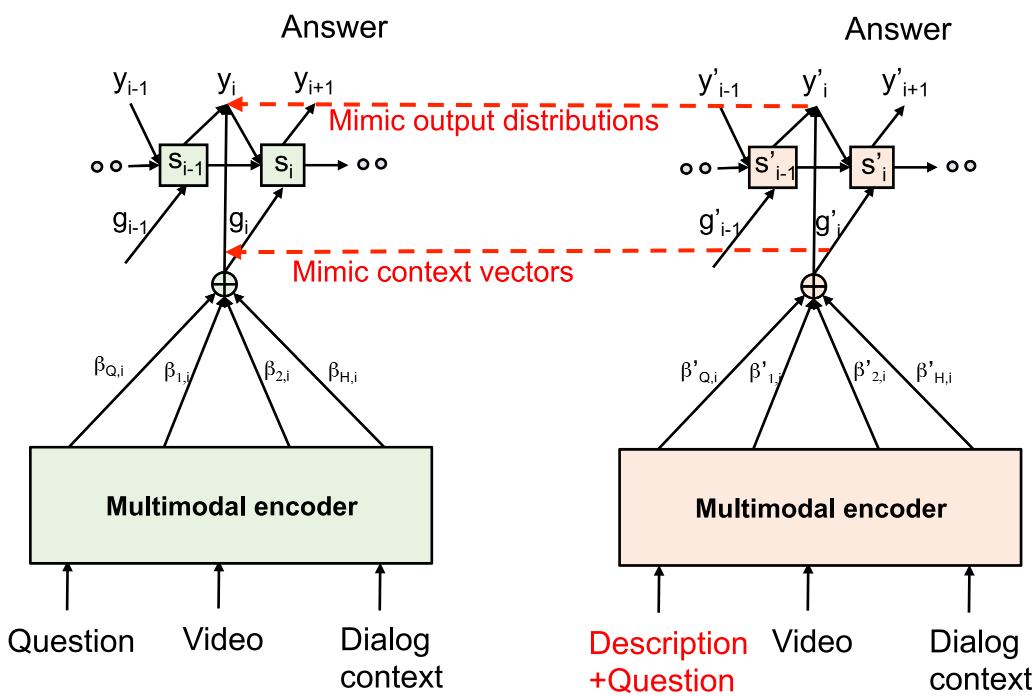
Multimodal Dialog



Essential Video Captioning Power

- Manual captions/summary is not available for real application.
- Answer generation models need video captioning power

Joint Student-Teacher Learning for Audio-Visual Scene-Aware



Student network

Teacher network

AVSD@DSTC10

3rd Edition of Audio Visual Scene-Aware Dialog Challenge

https://github.com/dialogtekgeek/AVSD-DSTC10_Official

Task 1: Video QA dialog

Goal: Answer generation without using manual descriptions for inference

You can train models using manual descriptions but CANNOT use them for testing.

Video description capability needs to be embedded within the answer generation models.

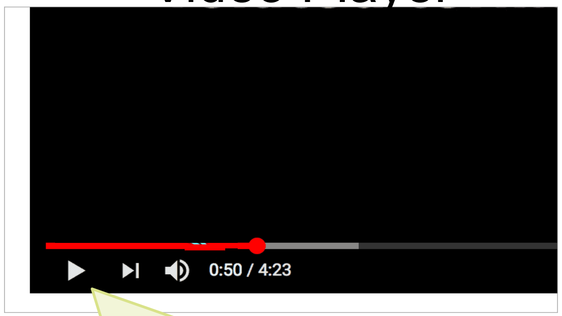
Task 2: Grounding Video QA dialog

Goal: Answer reasoning temporal Localization

To support answers, evidence is required to be shown without using manual descriptions.

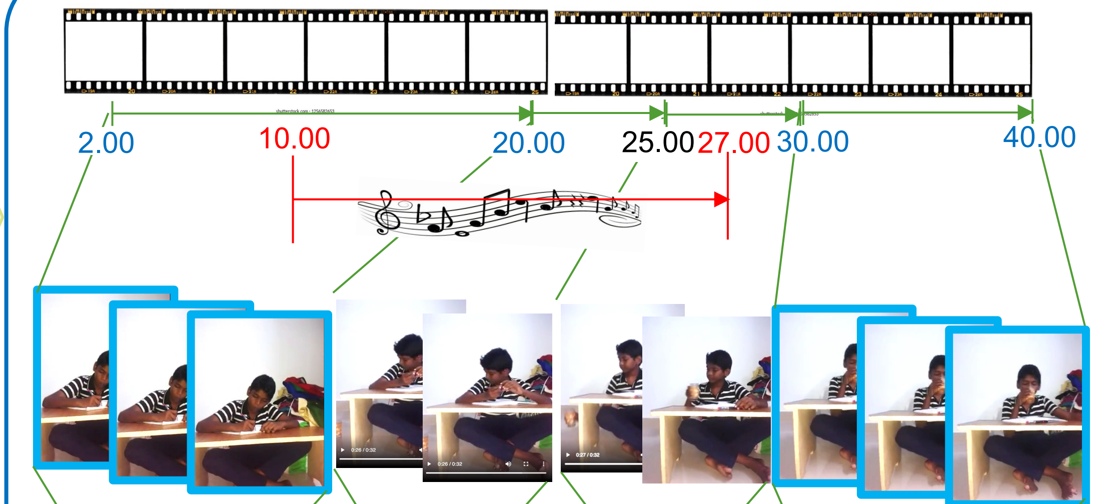
MITSUBISHI ELECTRIC
Changes for the Better

Video Player



Press the buttons to play and stop or slide the bar to find the begin and end timing for reasoning events.

Video Content



If the given answer is incorrect, please check "False" and write a correct answer with timestamp.

Press the "Begin" and "End" buttons to extract the current time from the video player.

Press the "+" button to add more evidence slots or press the "-" button to remove the last evidence.

Turn 1:

Question (Q): What actions are taken by the boy?
Answer (A): the boy is writing while singing and then drinks a glass of orange juice.

If the given answer is incorrect, please check the box and write a correct answer with timing.

False: ☐

Please select visual or audio evidence.

Evidence 1-1: visual ● audio ○
Begin 2.00 End 20.00 Jump to begin Jump to end

The boy is writing in the notebook.

Evidence 1-2: visual ● audio ○
Begin 30.00 End 40.00 Jump to begin Jump to end

The boy is drinking from the glass.

Evidence 1-3: visual ○ audio ●
Begin 10.00 End 27.00 Jump to begin Jump to end

The boy singing Rock music.

+ -

Explain reasons to extract the event to justify why the answer is correct.

Find as much evidence as possible you can.

AVSD@DSTC Challenge Schedule

June 14th, 2021: Answer generation data release

June 30th, 2021: Answer reasoning temporal localization data and baseline release

Sep. 13th, 2021: Test Data release

Sep. 21st, 2021: Test Submission due

Nov. 1st, 2021: Challenge paper submission due

Jan. or Feb., 2022: Workshop

Join us!!!