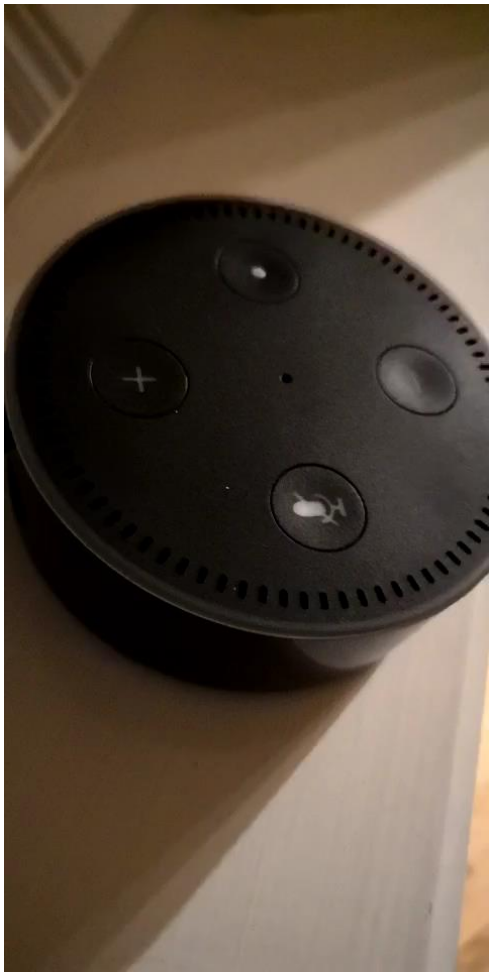
Multi-modal Reasoning: Bridging Vision and Language

Heming Zhang

Media Communications Lab
University of Southern California

Personal Assistant – AI Touchstone



The mass of an electron is approximately 9.109×10^{-31} kg.

Has Personal Assistant Come True?

Illustration by Fiona Carswell

Vision & Language in MCL

Vision

- Object detection
- Semantic segmentation
- Video segmentation

Language

- Text classification
- Language graph learning

Vision & Language

- **Visual dialogue**
- Vision & Language navigation
- Multi-modal machine translation

What is Visual Dialogue?

- Dialogue that is grounded in vision

A man wearing leather jacket standing next to a motorcycle

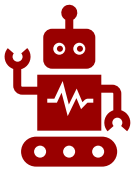
Is it colored leather?

Yes, it is.

What color is his leather?

Why Visual Dialogue?

- Aiding visually impaired users



Daisy just sent you some pictures of her new house.

Yes, there is a large living room with fireplace

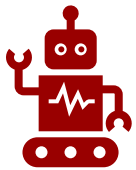
Great, is the living room large?

- Aiding analysts

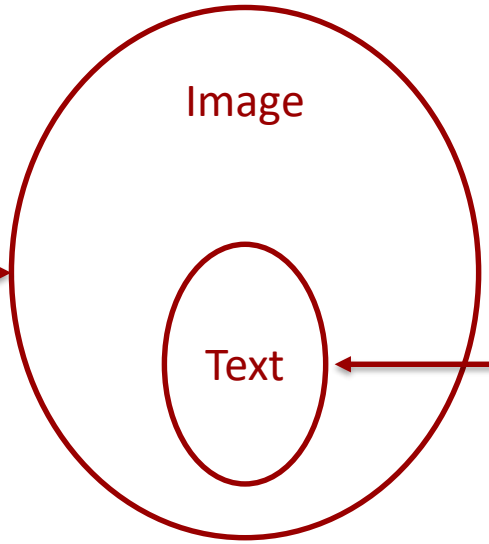
Did anyone pass the gate yesterday?

Were any of them carrying a cardboard box?

Yes, 45 instances logged on camera.



From Information Point of View



A man wearing leather jacket standing next to a motorcycle

Is it colored leather?

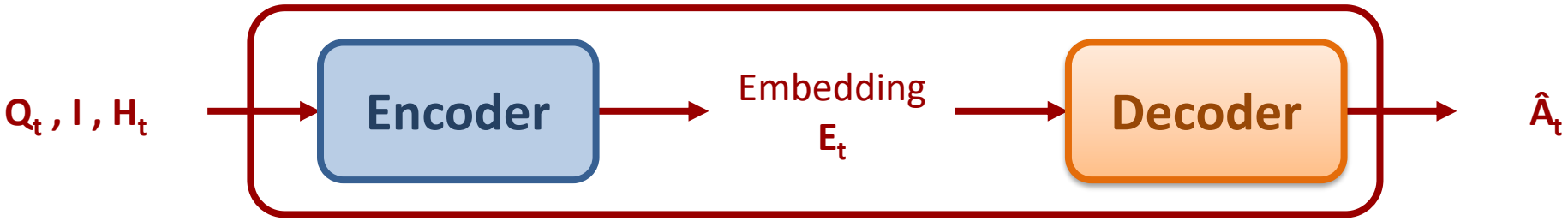
Yes, it is.

What color is his leather?

Previous Work

- Encoder-decoder framework

(Das et al., 2017, Lu et al., 2017, Wu et al., 2018, etc.)



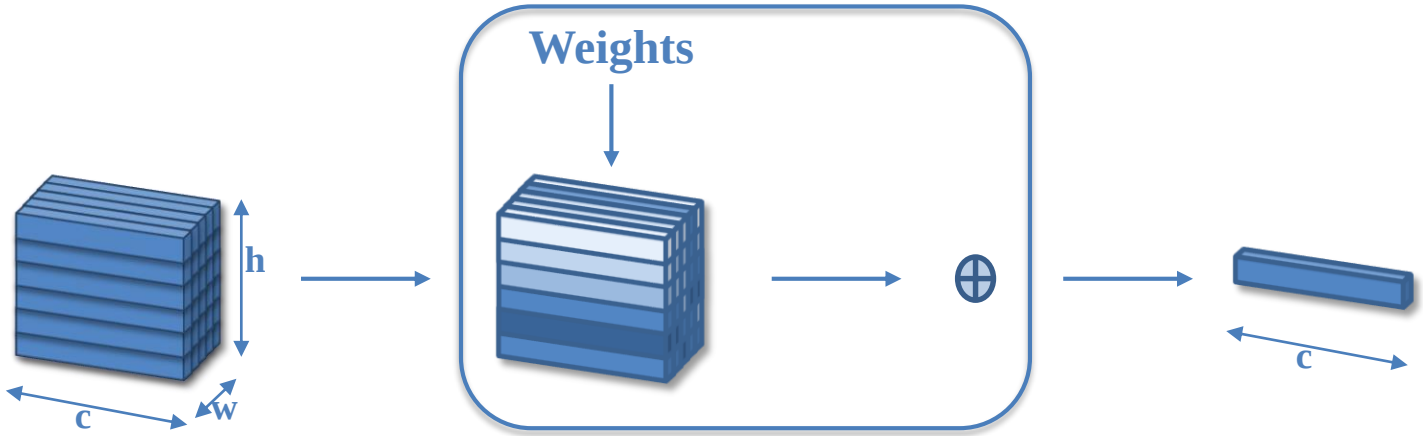
- Encoder
  - Embeds image, question and dialogue history
- Decoder
  - Decodes the embedding to answers in natural language

Previous multi-modal encoders

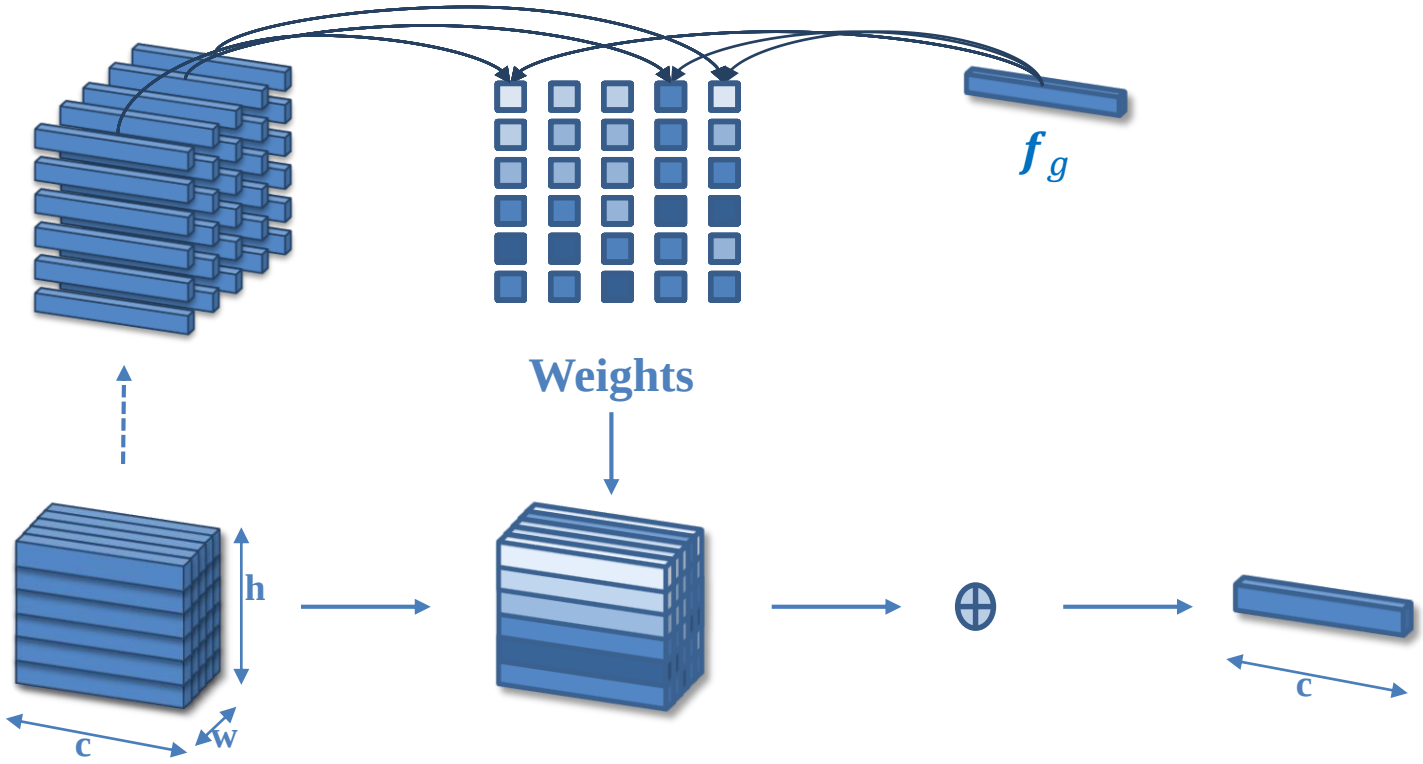
- Lu et al., 2017, Wu et al., 2018, etc.
  - Use one input as guidance to compute attention on another input

Attention

- Weighted-sum over features



Attention with Guidance



Previous multi-modal encoders

- Lu et al., 2017, Wu et al., 2018, etc.
  - Use one input as guidance to compute attention on another input
  - Process inputs sequentially in pre-defined orders

Encoders with Sequential Attention

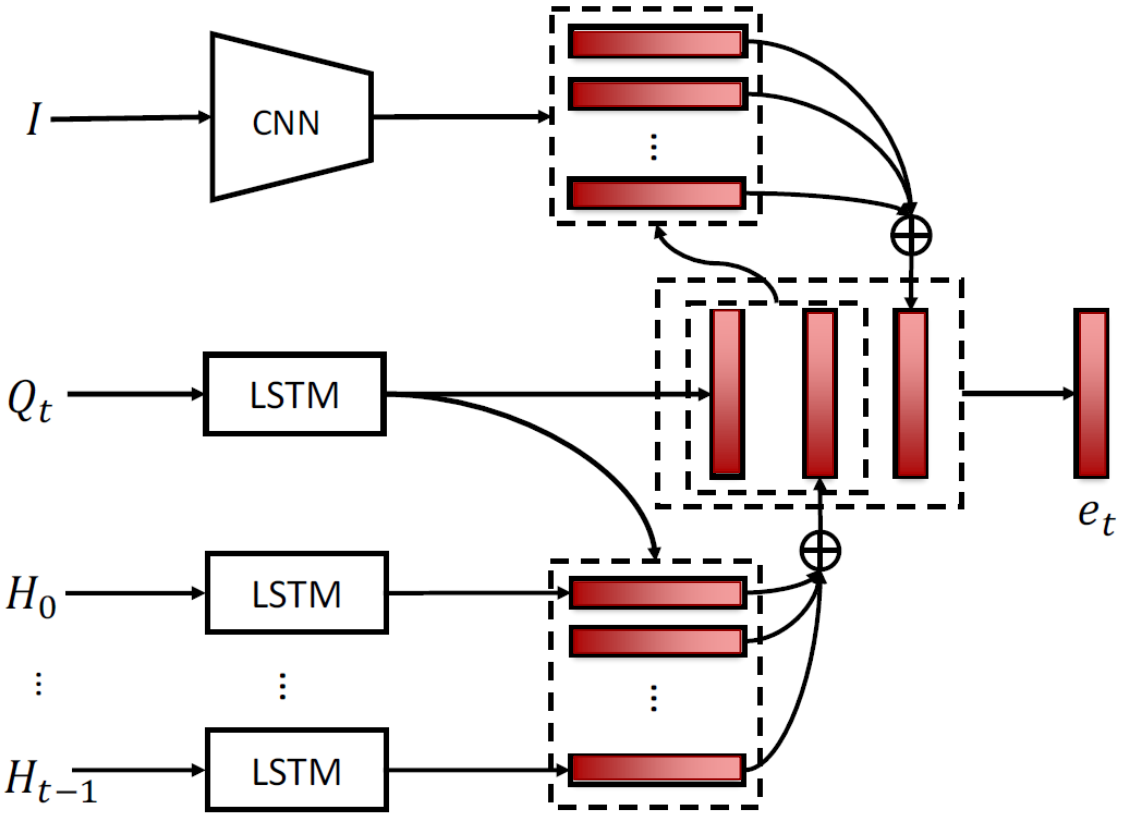
- Lu et al. 2017

What color is his leather?

A man wearing leather jacket standing next to a motorcycle

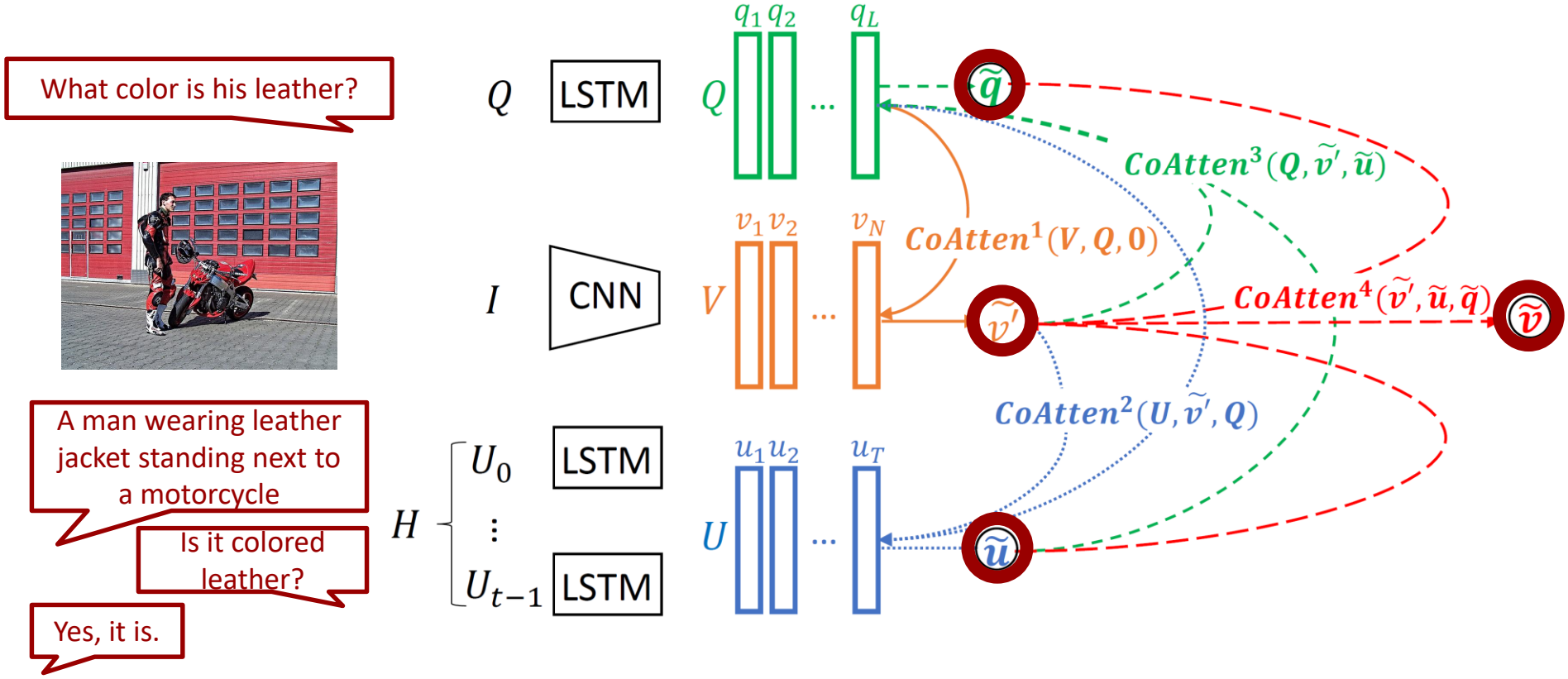
Is it colored leather?

Yes, it is.



Encoders with Sequential Attention

- Wu et al., 2018



Previous multi-modal encoders

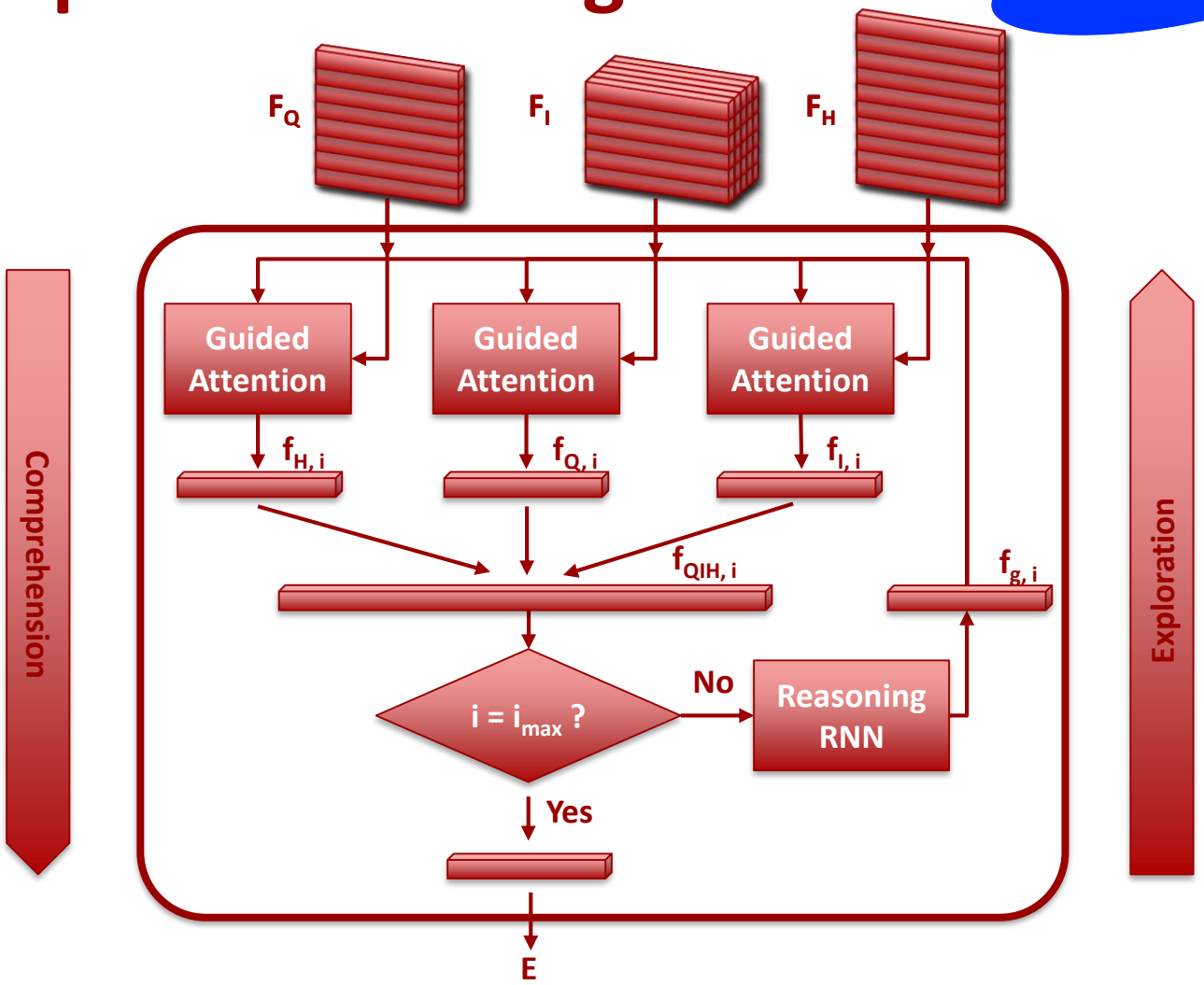
- Cannot accommodate to different scenario's
  - How many people are there in the image?

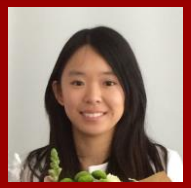
question → image
the word 'people' regions of people

  - Is there anything else on the table?

question → image → question → history
the word 'table' regions of table the word 'else' context for 'else'

Adaptive reasoning





Attention Visualization

Is the little boy on a beach?

How old does he look?

Attention Visualization

What color hair does he have?

How old does he look?

Attention Visualization

What color hair does he have?

Is he dressed for summer?

Attention Visualization

What color is the airplane?

Time step i=1

Attention Visualization

What color is the airplane?

Time step i=2

Qualitative results

4 ducks are in a grassy island of a parking lot with their heads down

Qualitative results

4 ducks are in a grassy island of a parking lot with their heads down

Questions	Human	Ours
Any grass?	Yes	Yes, a lot of grass
What color grass?	It is green with brownish dead spots	Green and brown

Qualitative results

4 ducks are in a grassy island of a parking lot with their heads down

Questions	Human	Ours
Any vehicles on the lot?	Yes	Yes, there are a lot of cars
Do they look new or old?	They look new	They look new

IJCAI 2019

Generative Visual Dialogue System via Weighted Likelihood Estimation

Heming Zhang, Shalini Ghosh, Larry Heck, Stephen Walsh, Junting Zhang, Jie Zhang, C.-C. Jay Kuo

Thursday Aug. 15th 09:30 - 10:30 AM

CV|LV - Language and Vision 2 (2501-2502)

Vision-grounded Problems Revisited

- What is visual dialogue?
- Dialogue that grounded in vision

A man wearing leather jacket standing next to a motorcycle

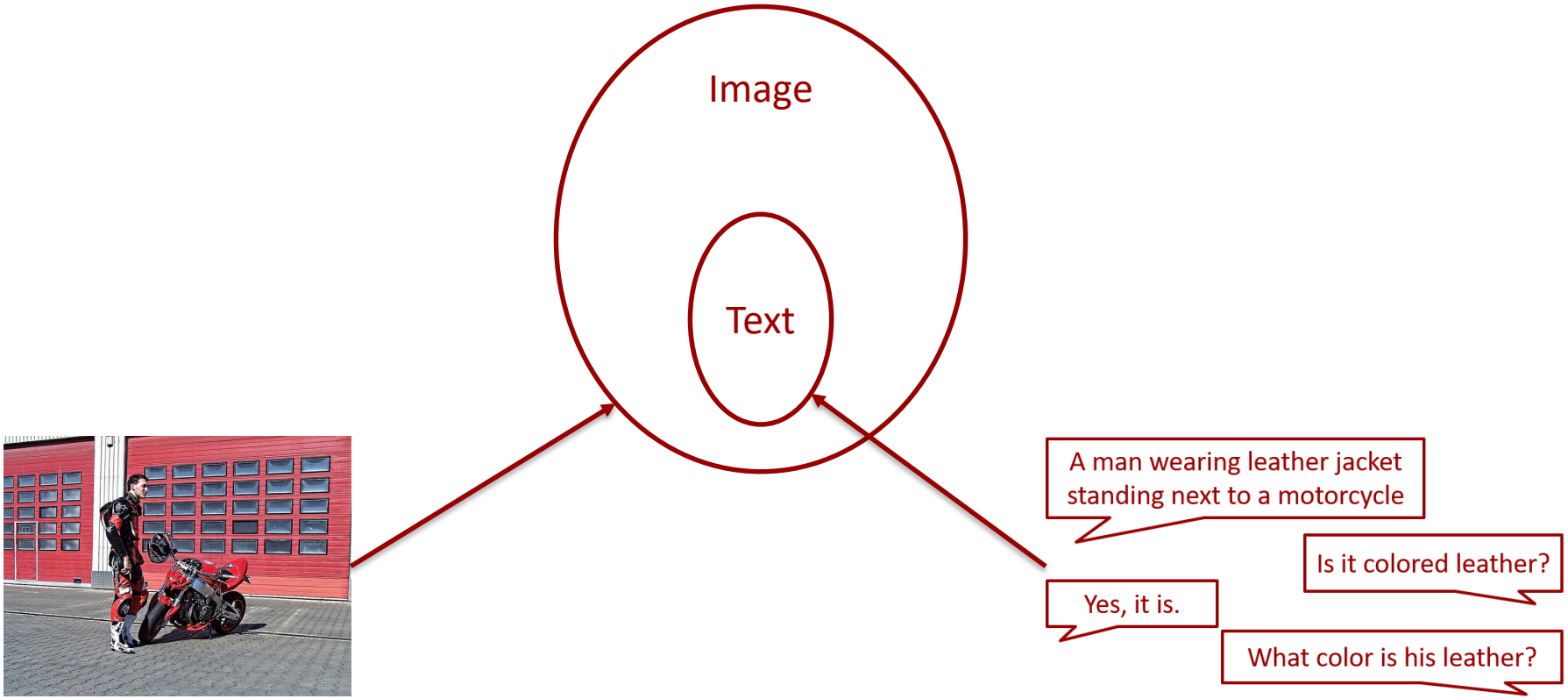
Is it colored leather?

Yes, it is.

What color is his leather?

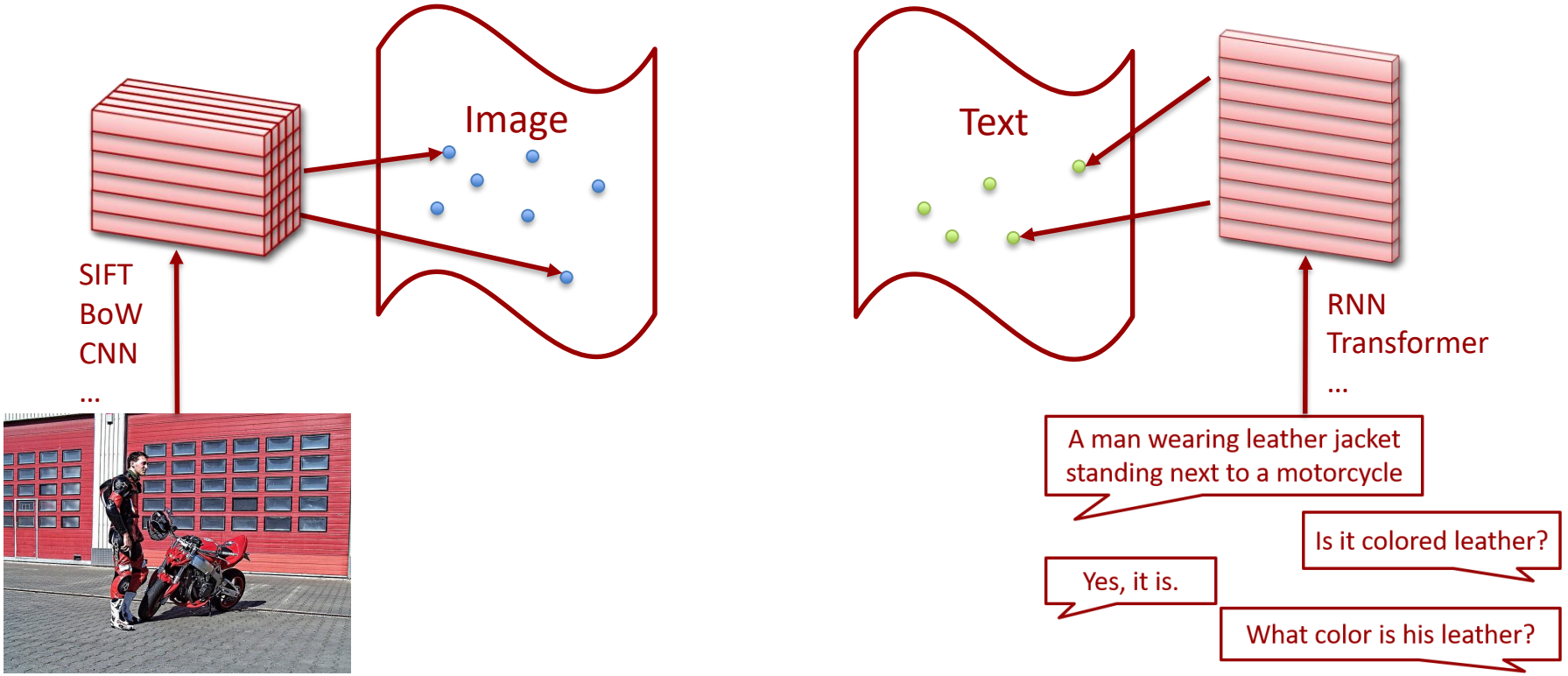
Vision-grounded Problems Revisited

- From information point of view

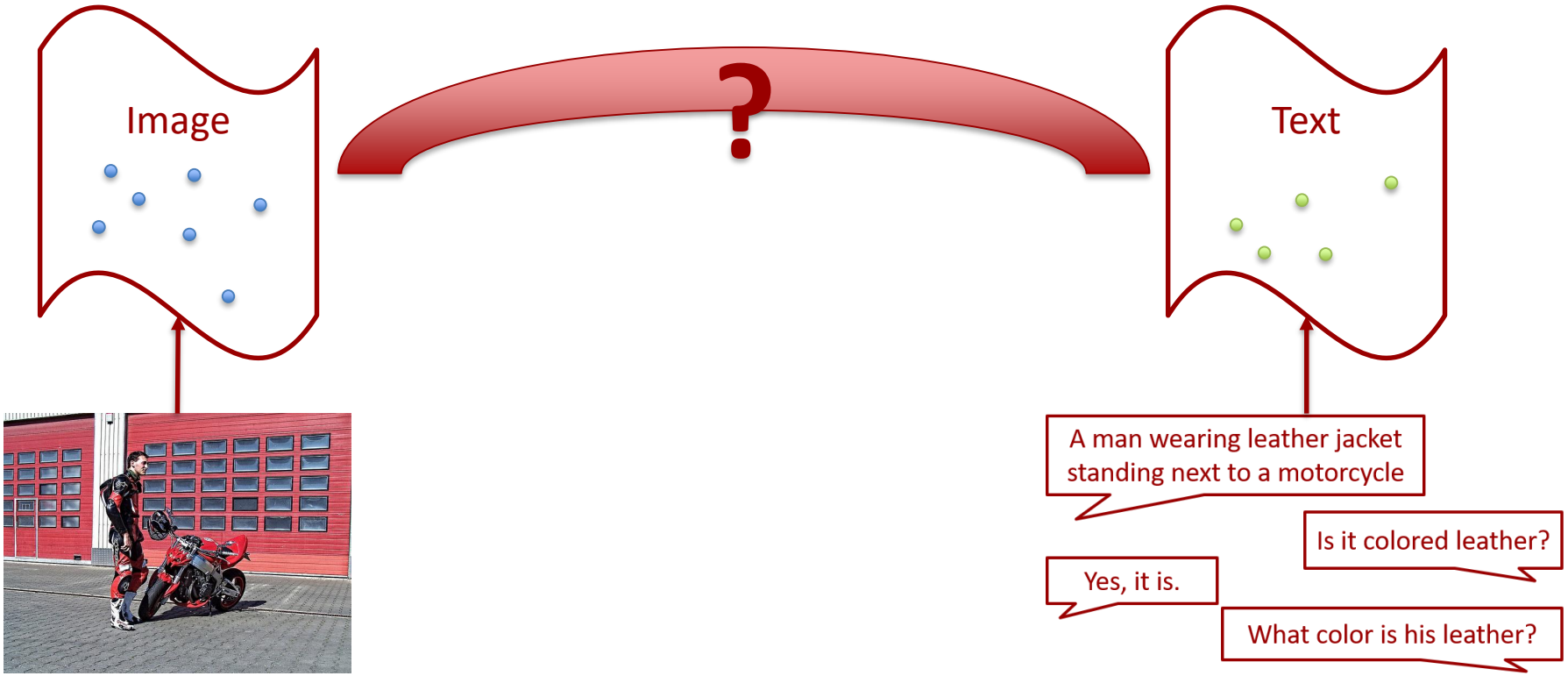


Vision-grounded Problems Revisited

- No alignment between image & text manifolds

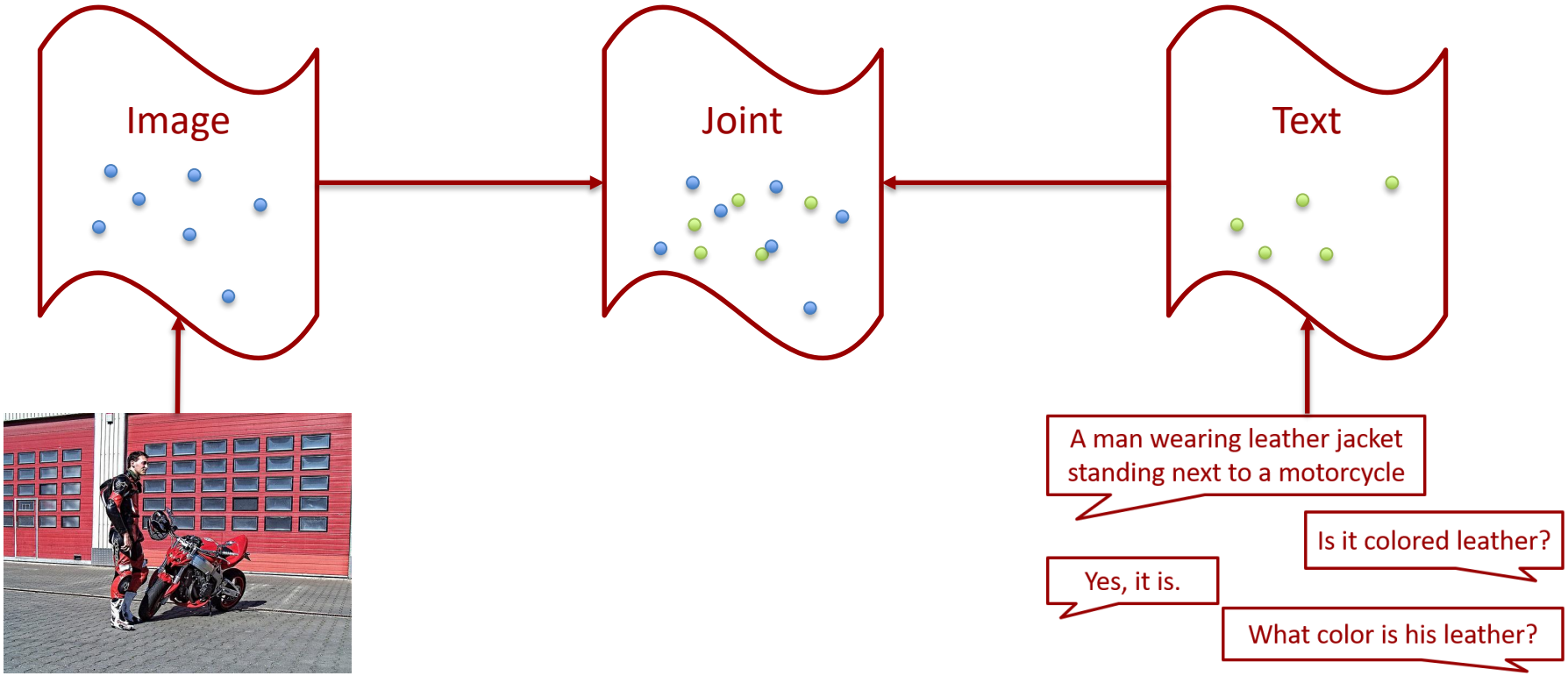


Bridging Vision & Language



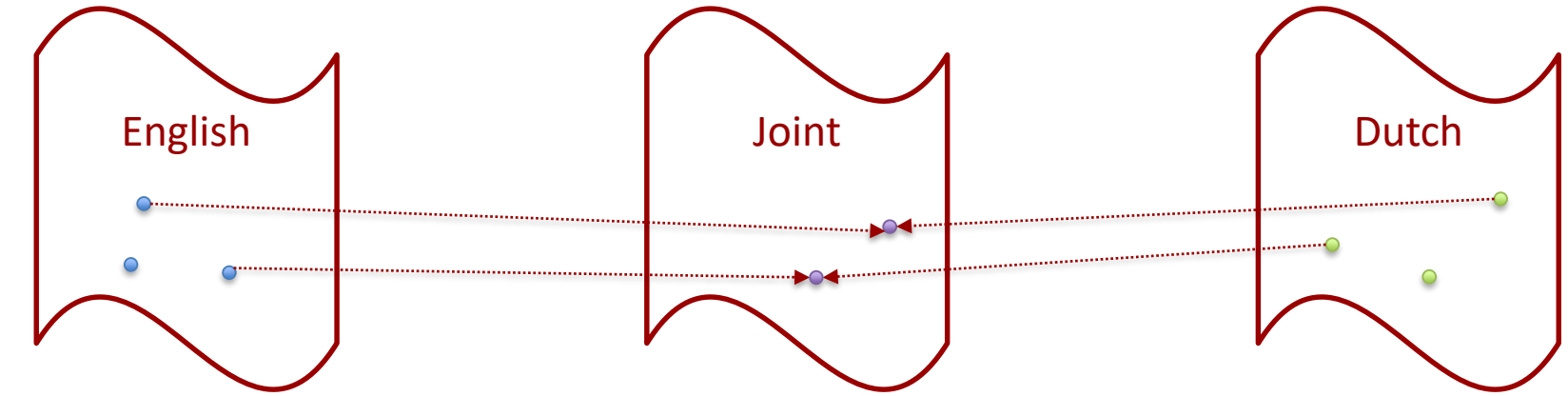
Bridging Vision & Language

- Manifold alignment



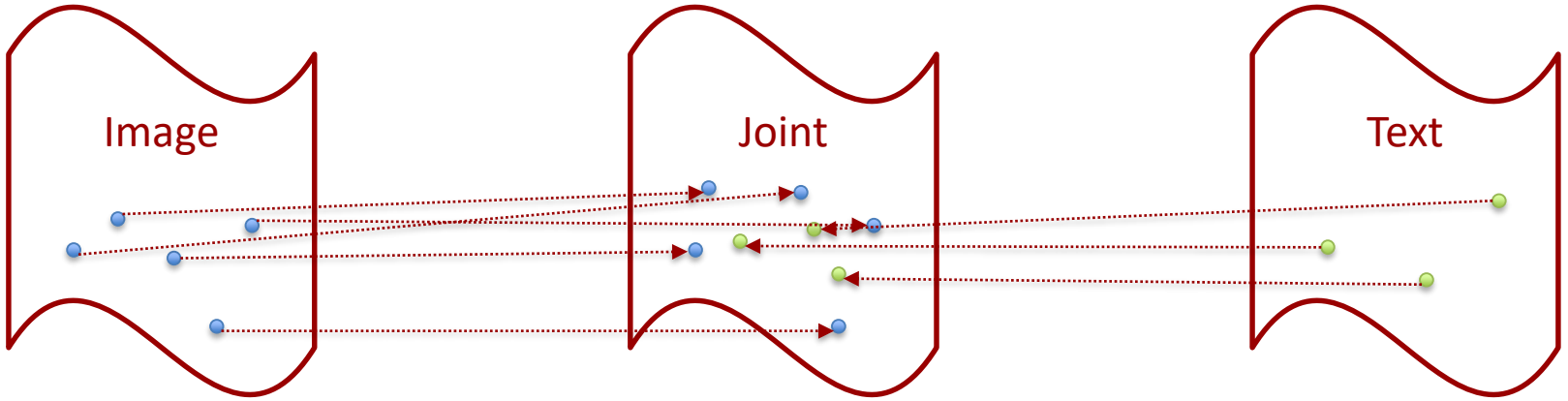
Bridging Vision & Language

- Usually one-to-one mapping in other manifold alignment problems
  - E.g. machine translation



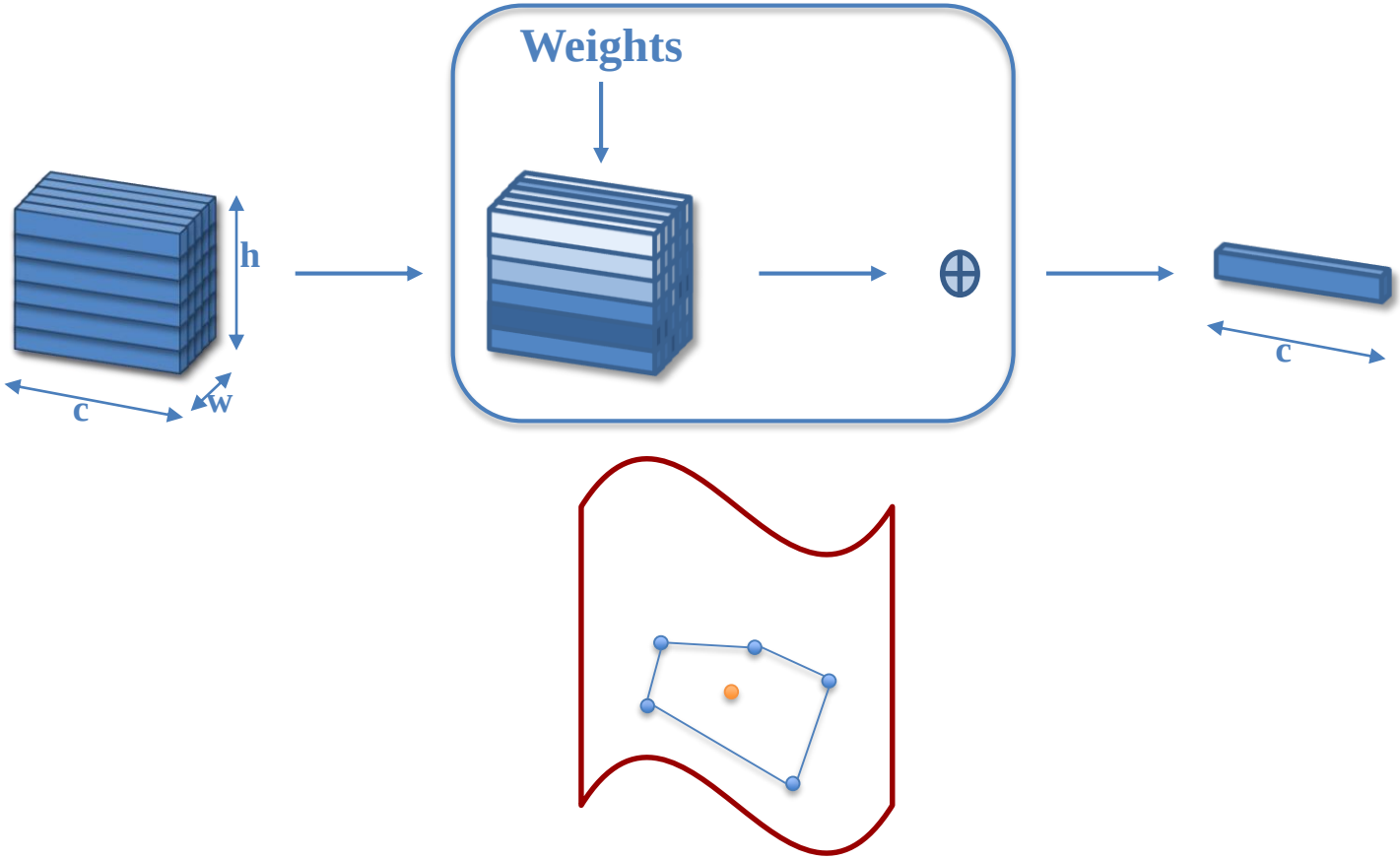
Bridging Vision & Language

- Alignment between vision and language
  - **No** one-to-one mapping



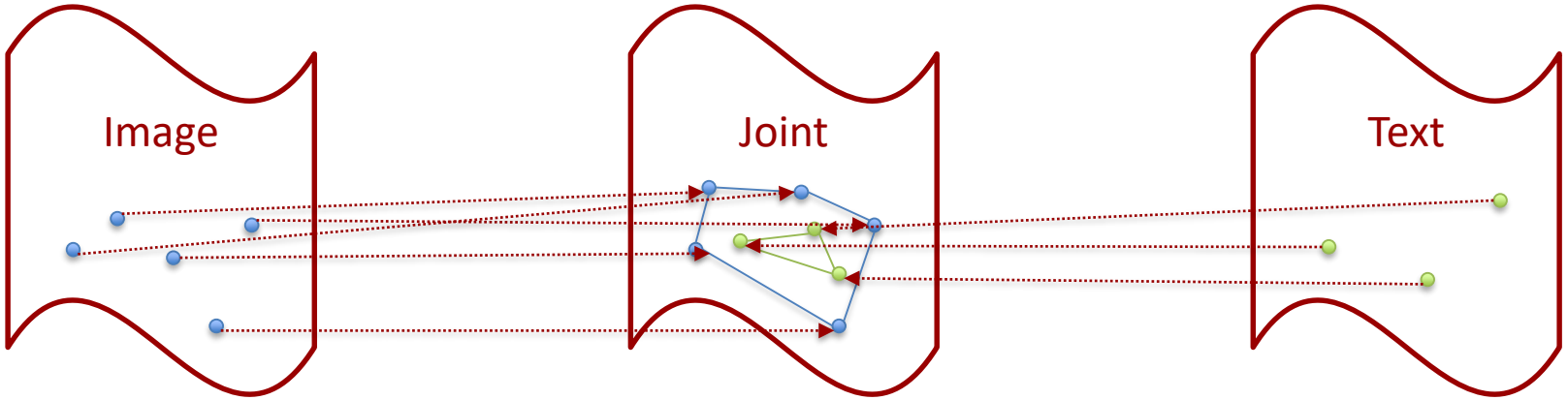
Attention Revisited

- Weighted-sum over features



Bridging Vision & Language

- Alignment by attention
  - Joint learning of attention and alignment



Related Research in MCL

Vision

- Object detection
- Semantic segmentation
- Video segmentation

Language

- Text classification
- Language graph learning

Vision & Language

- Visual dialogue
- Vision & Language navigation
- Multi-modal machine translation

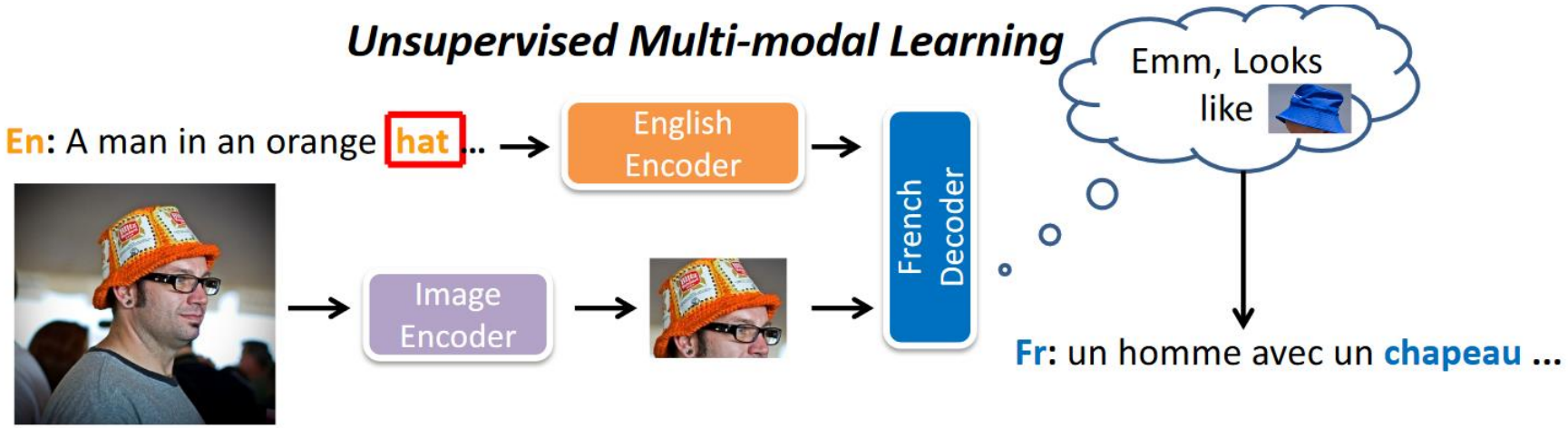
Vision-and-language Navigation

- Instructions in natural language
  - Walk down and turn right.
- Surrounding environment in vision

Co-attention between Vision & Language

- Leave the room into the hall and go straight.
- Head towards the stairs.
- Stop on the round rug next to the flowers.

Unsupervised Multi-modal Neural Machine Translation



Media Communication Lab

- Lab director: Prof. C.-C. Jay Kuo
- Visiting scholars

- PhD students

- Master students

Thank you for listening

Visit us at

http://mcl.usc.edu/