Language and Vision: tasks, models and what they learn?

Ravi Shekhar

Queen Mary University of London

2019 Intelligent Sensing Summer School

Visual Grounding of Natural Language

Q: Is the door close?

Visual Grounding of Natural Language

Q: Is the door close?

A: Yes (Visual)

Visual Grounding of Natural Language

Caption: Mary is going to the office.
Q: Where is Mary going?

Visual Grounding of Natural Language

Caption: Mary is going to the office.

Q: Where is Mary going?

A: Office (Linguistic)

Visual Grounding of Natural Language

Q: Is the door close?

A: Yes (Visual)

Caption: Mary is going to the office.

Q: Where is Mary going?

A: Office (Linguistic)

Caption: Mary is going to the office.

Q: Is Mary's office close?

Visual Grounding of Natural Language

Q: Is the door close?

A: Yes (Visual)

Caption: Mary is going to the office.

Q: Where is Mary going?

A: Office (Linguistic)

Caption: Mary is going to the office.

Q: Is Mary's office close?

A: Yes (Multi-modal)

The Plan

- Language and Vision (LaVi) Tasks
- LaVi Models
- What they learn?

The Plan

- **Language and Vision (LaVi) Tasks**
- LaVi Models
- What they learn?

Language and Vision Tasks

- Non-interactive tasks
- Interactive tasks

LaVi tasks

- Non-interactive tasks
  - Image Captioning (IC)
  - Visual Question Answering (VQA)
  - Visual reference resolution (ReferIt)

LaVi tasks

- Non-interactive tasks
  - Image Captioning (IC)
  - Visual Question Answering (VQA)
  - Visual reference resolution (ReferIt)
- Interactive tasks
  - Goal oriented dialogue: GuessWhat (GW)
  - Free form dialogue: Visually Dialogue(VisDial)

LaVi Non-interactive tasks

- **IC**
  Two red mugs and a towel on a kitchen counter.

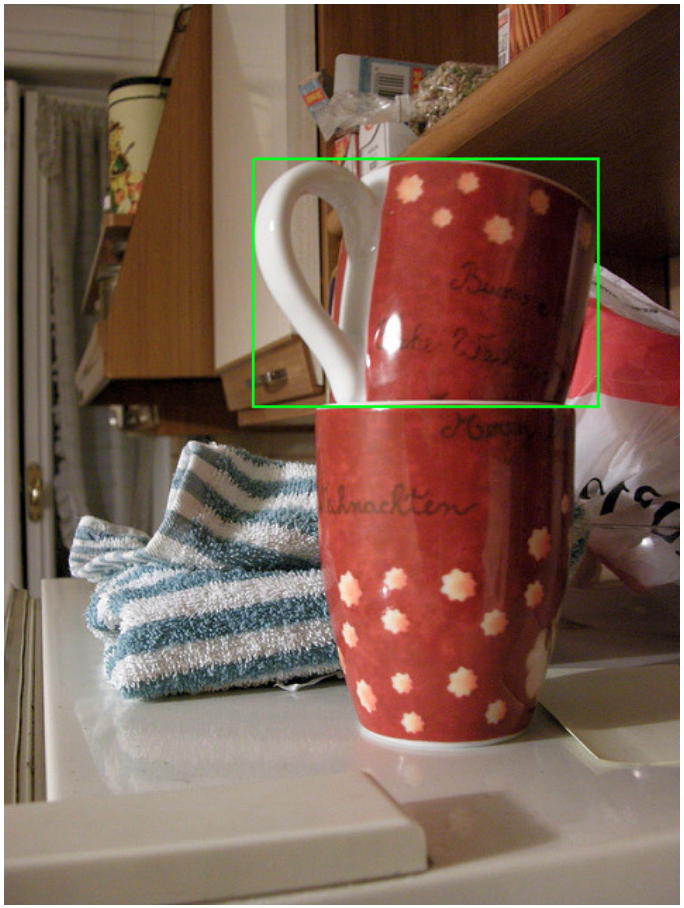
LaVi Non-interactive tasks

- **IC**
  Two red mugs and a towel on a kitchen counter.
- **VQA**
  Q: How many cups are there?

LaVi Non-interactive tasks

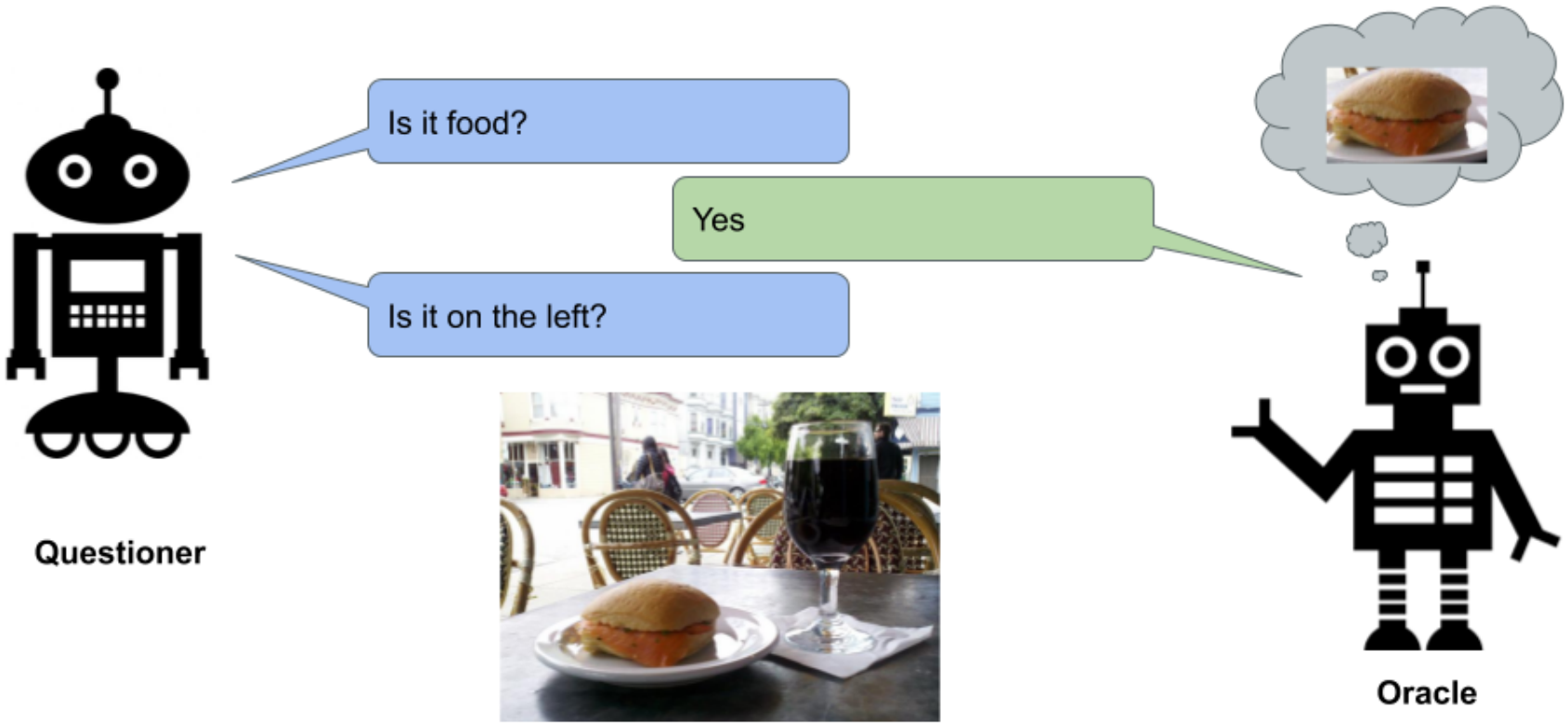
- **IC**
  Two red mugs and a towel on a kitchen counter.
- **VQA**
  Q: How many cups are there?
  A: *Two*

LaVi Non-interactive tasks



- **IC**
  Two red mugs and a towel on a kitchen counter.
- **VQA**
  Q: How many cups are there?
  A: Two
- **ReferIt**
  The top mug.

LaVi Interactive tasks



Goal oriented dialogue: GuessWhat(GW) Game

LaVi Interactive tasks

Questioner

Caption: A girl and a dog in a car.

Q1: Is dog sitting?

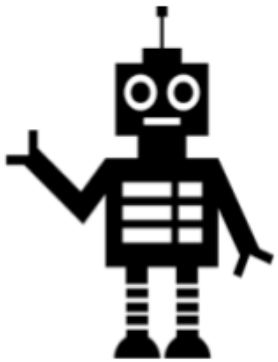
A1: No, dog is half outside the car.

Q2: What girl is doing?

A2: She is sitting on the driver's seat.

Q3: Is she old enough to drive?

A3: Yes.



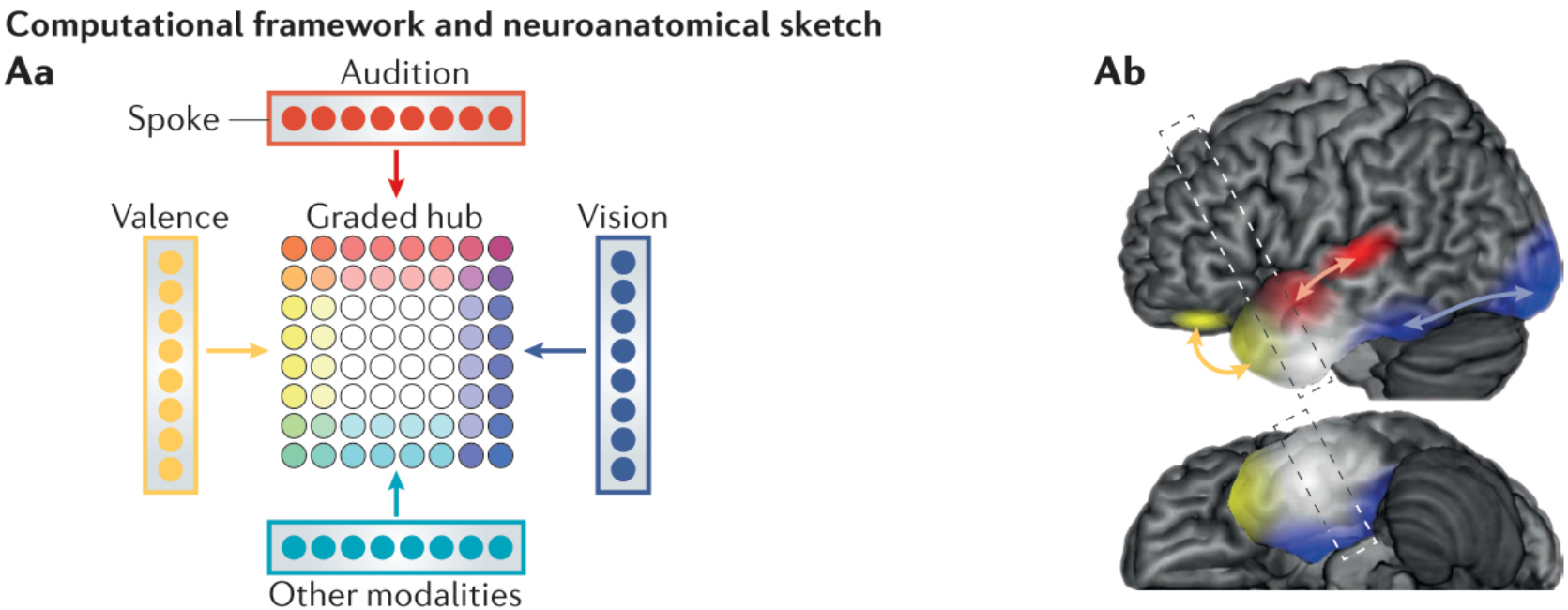
Answerer

Free form dialogue: Visually Dialogue(VisDial)

The Plan

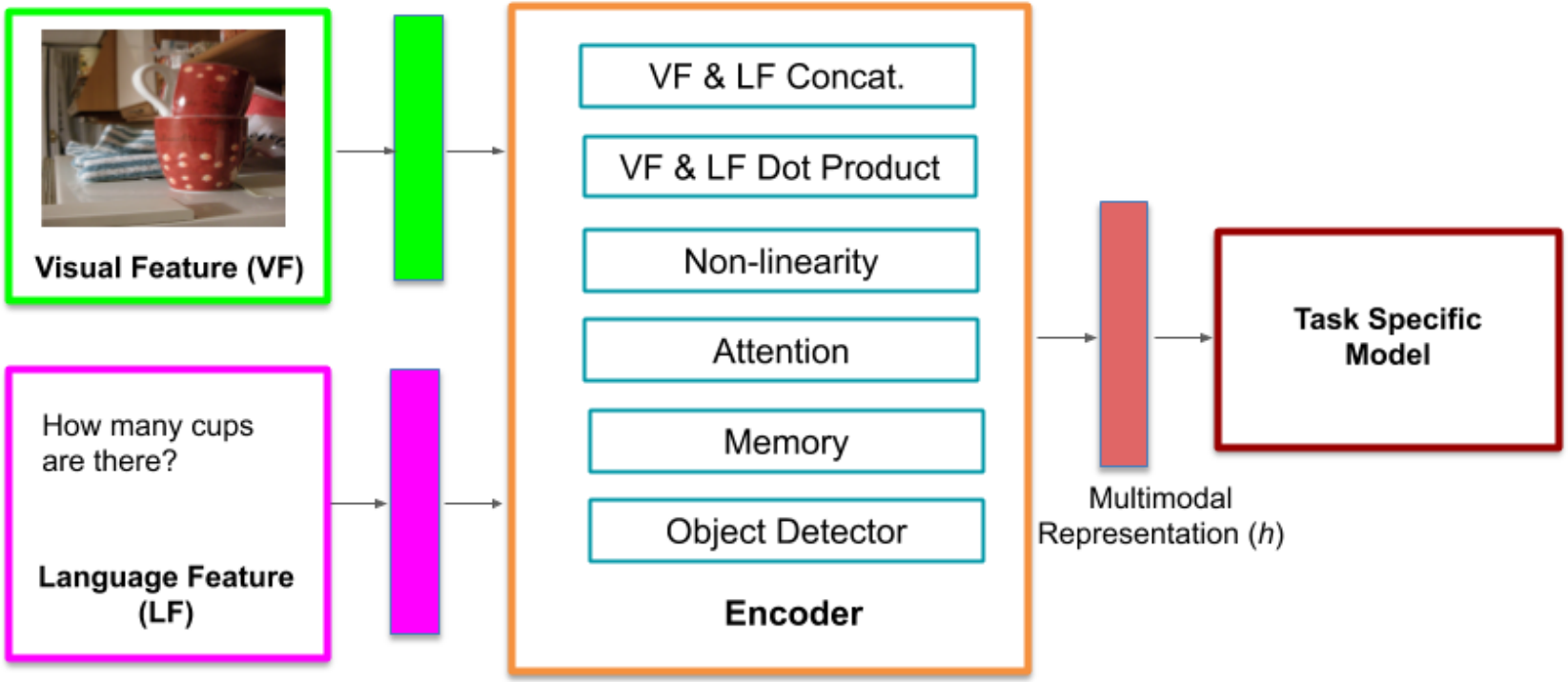
- Language and Vision (LaVi) Tasks
- **LaVi Models**
- What they learn?

The Hub-and-Spoke Model



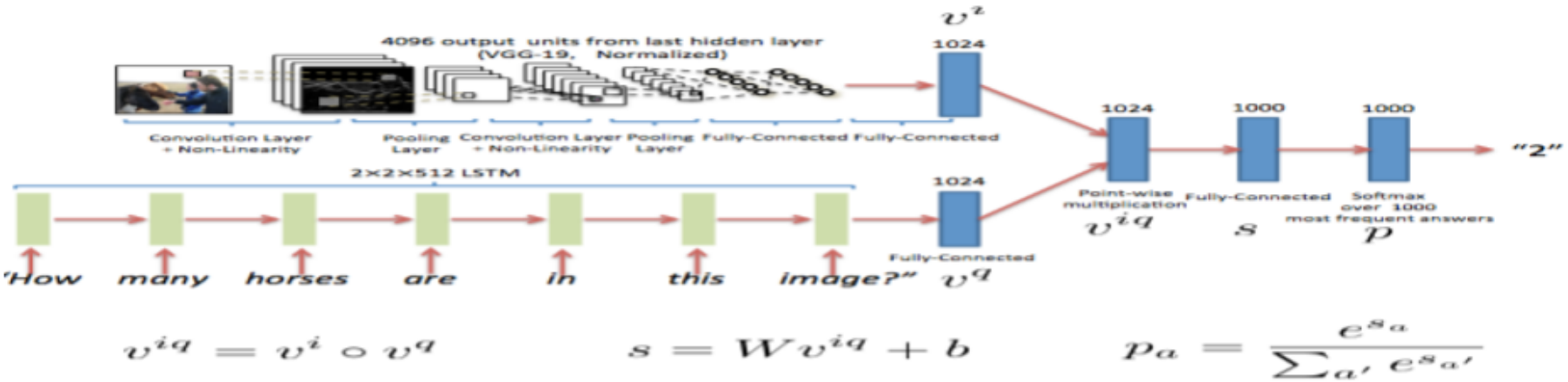
Lambon Ralph et al. The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, 2017.

Generic Architecture

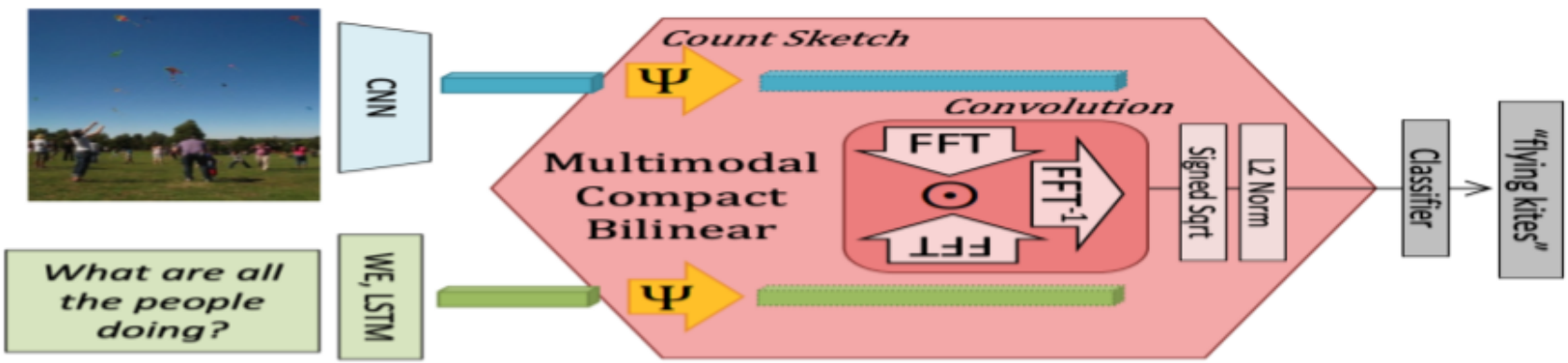


A Generic Language and Vision Model.

Models: Fusion

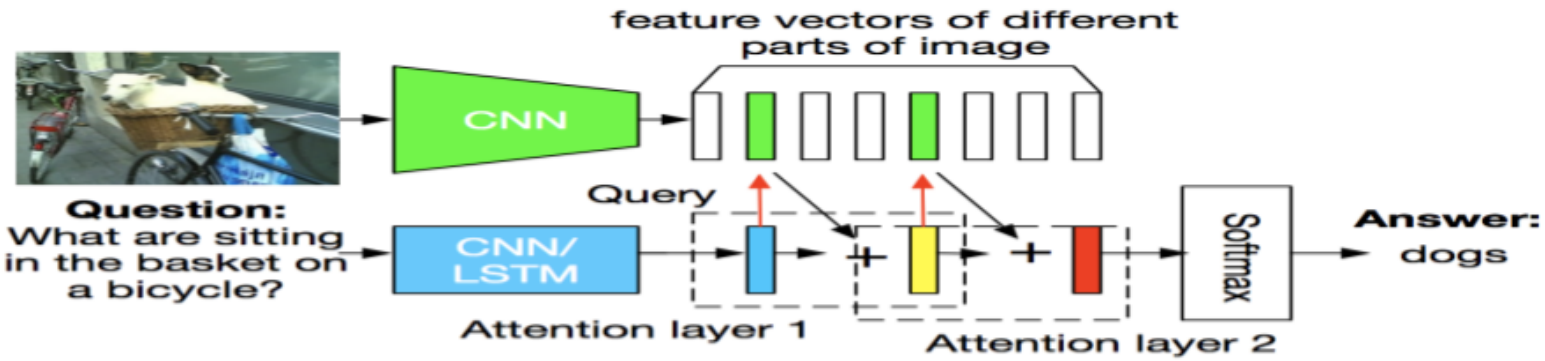


Point-wise Multiplication Antol et al ICCV'15

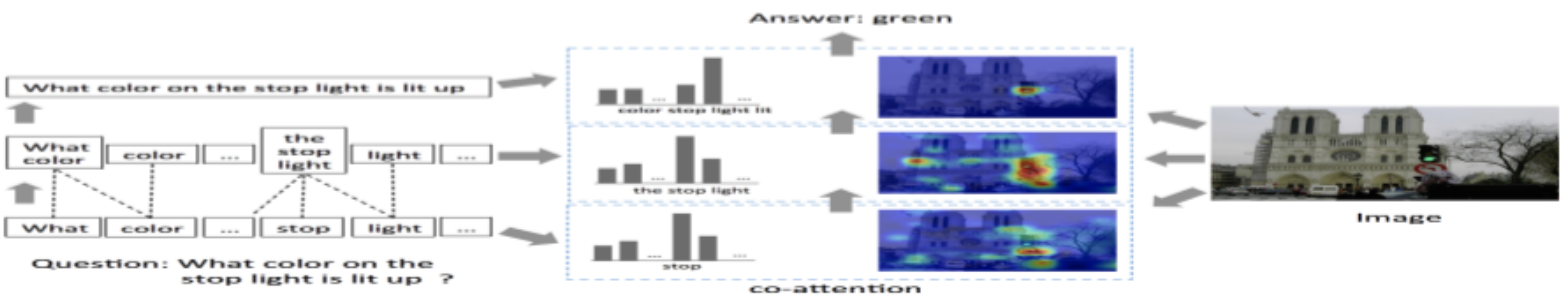


Multimodal Compact Bilinear Pooling Fukui et al., EMNLP 16

Models: Attention



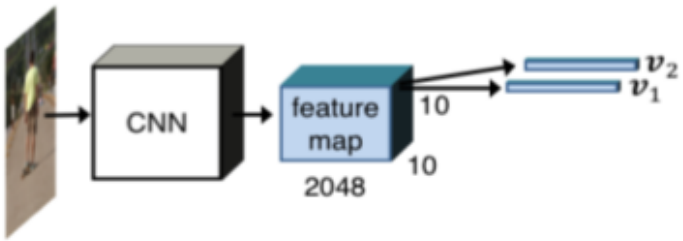
Stacked Attention Networks Yang et al., CVPR 16



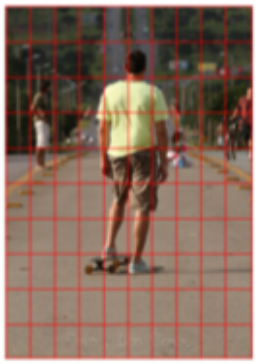
Hierarchical Question-Image Co-Attention Lu et al., NIPS 16

Models: Object Detector

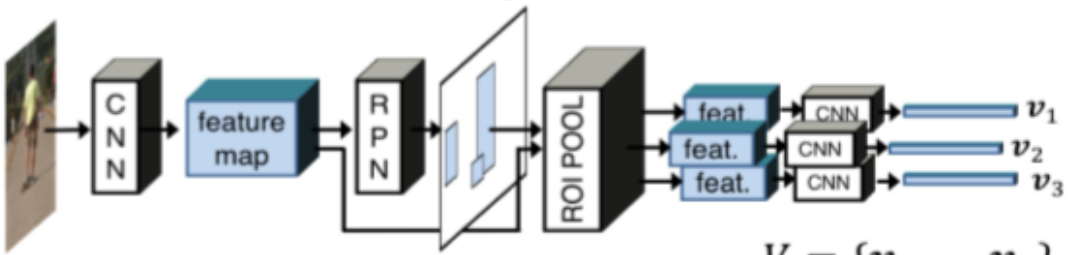
Spatial output of a CNN:



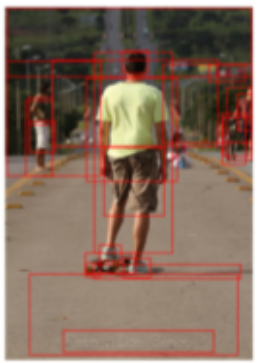
$$V = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{100}\}$$



Bottom-up attention (using Faster R-CNN):

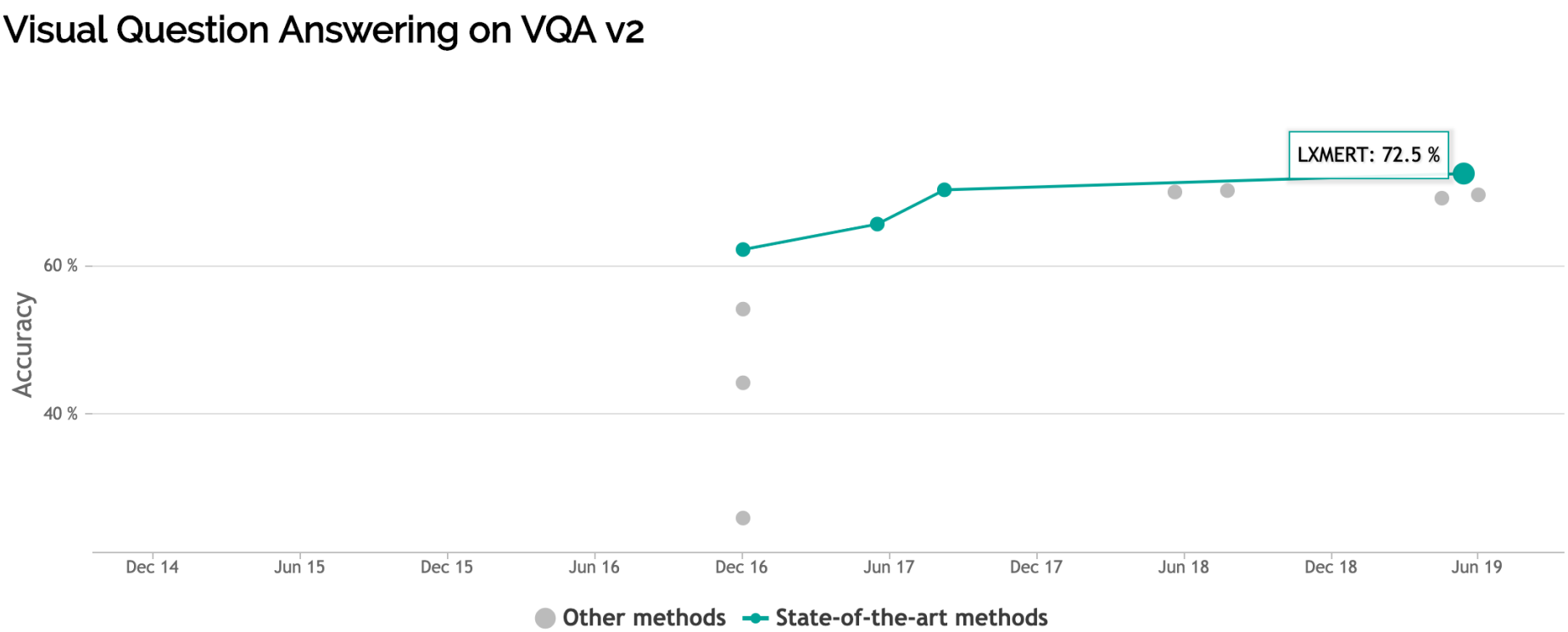


$$V = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$$



Bottom-Up and Top-Down Attention Anderson et al., CVPR 18

Models: VQA Progress



The Plan

- Language and Vision (LaVi) Tasks
- LaVi Models
- **What they learn?**

What they learn?

Do models trained on these tasks learn to properly represent multimodal information?

Making the Tasks Comparable

Originally: VQA, ReferIt and GW have different datasets & architectures

- ReferIt - originally an object detection task
- VQA and GW - defined as classification tasks

Making the Tasks Comparable

Originally: VQA, ReferIt and GW have different datasets & architectures

- ReferIt - originally an object detection task
- VQA and GW - defined as classification tasks

Our setup: for a fair comparison across tasks

- Create a **common dataset**
  - Get all the common images in all the datasets (14458) and create splits with respect to images (MS-COCO)
  - Make the linguistic items as similar as possible

Making the Tasks Comparable

Originally: VQA, ReferIt and GW have different datasets & architectures

- ReferIt - originally an object detection task
- VQA and GW - defined as classification tasks

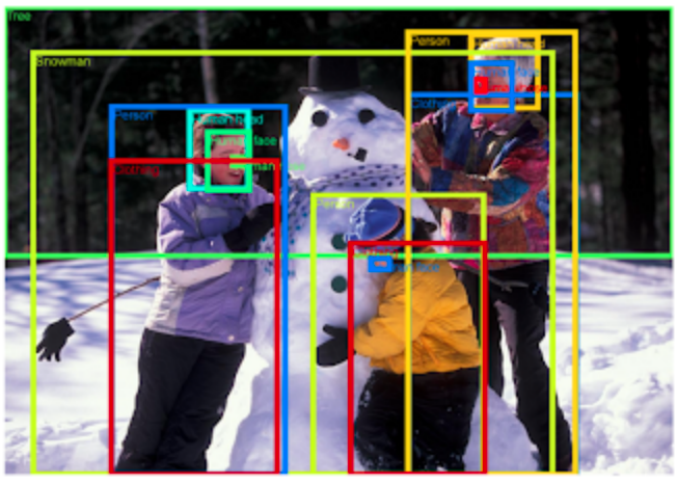
Our setup: for a fair comparison across tasks

- Create a **common dataset**
  - Get all the common images in all the datasets (14458) and create splits with respect to images (MS-COCO)
  - Make the linguistic items as similar as possible
- Pose the three tasks as **retrieval tasks**.

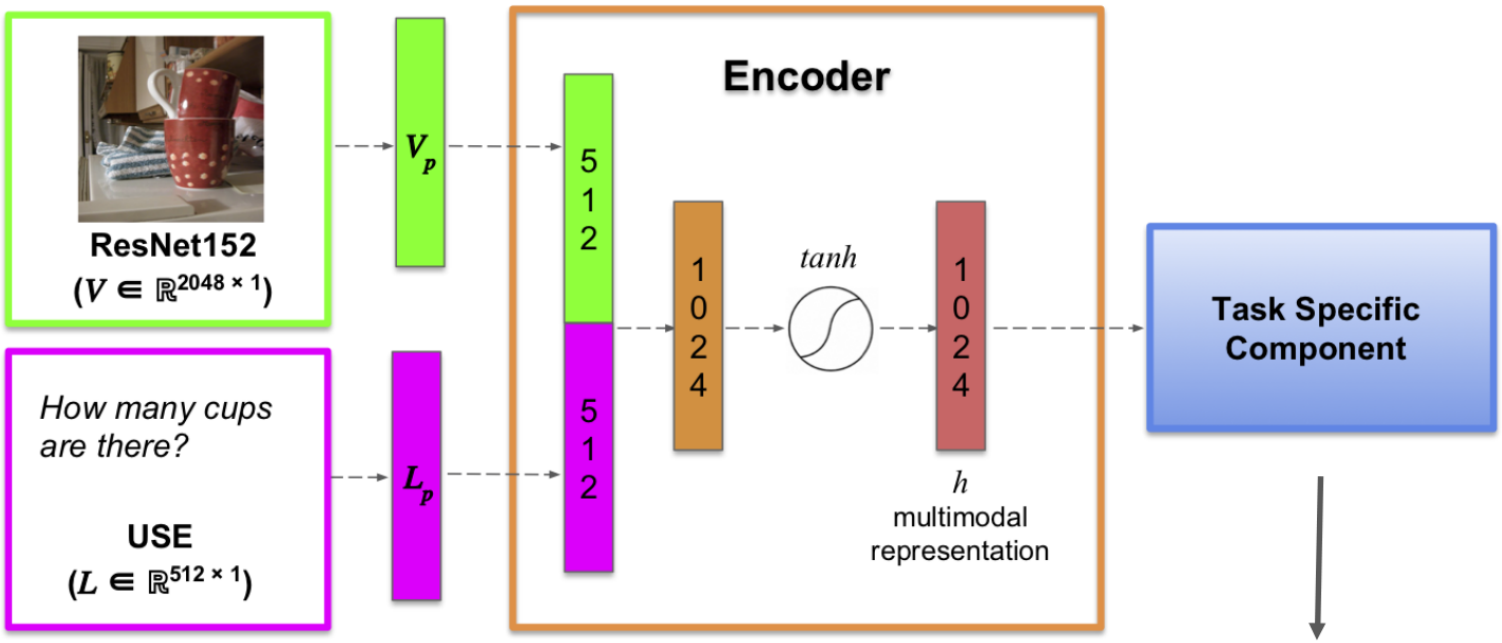
Formulation as Retrieval Tasks

- VQA:
  Retrieving the correct textual answer from a set of 18 possible natural language answers
  - {*Yes*, *No*, *Two mugs*, *Green*, *On the floor* . . . }

- ReferIt and GW:
  Retrieving the target object from a set of 20 possible objects in an image



General Architecture - The Hub



Pre-trained vectors:
- Visual features from ResNet152
- Language features from USE (Universal Sentence Encoder)

Multi-Layer Perceptron generating:
- Language embedding (VQA)
- Visual embedding (ReferIt, GW)

Trained with cosine similarity loss

What they learn?

Do models trained on these tasks learn to properly represent multimodal information?

- **Using a diagnostic task**
- Inspecting multimodal semantic spaces

FOIL as a Diagnostic Task

FOIL Diagnostic Task

Classify a caption as *original* or *foiled*

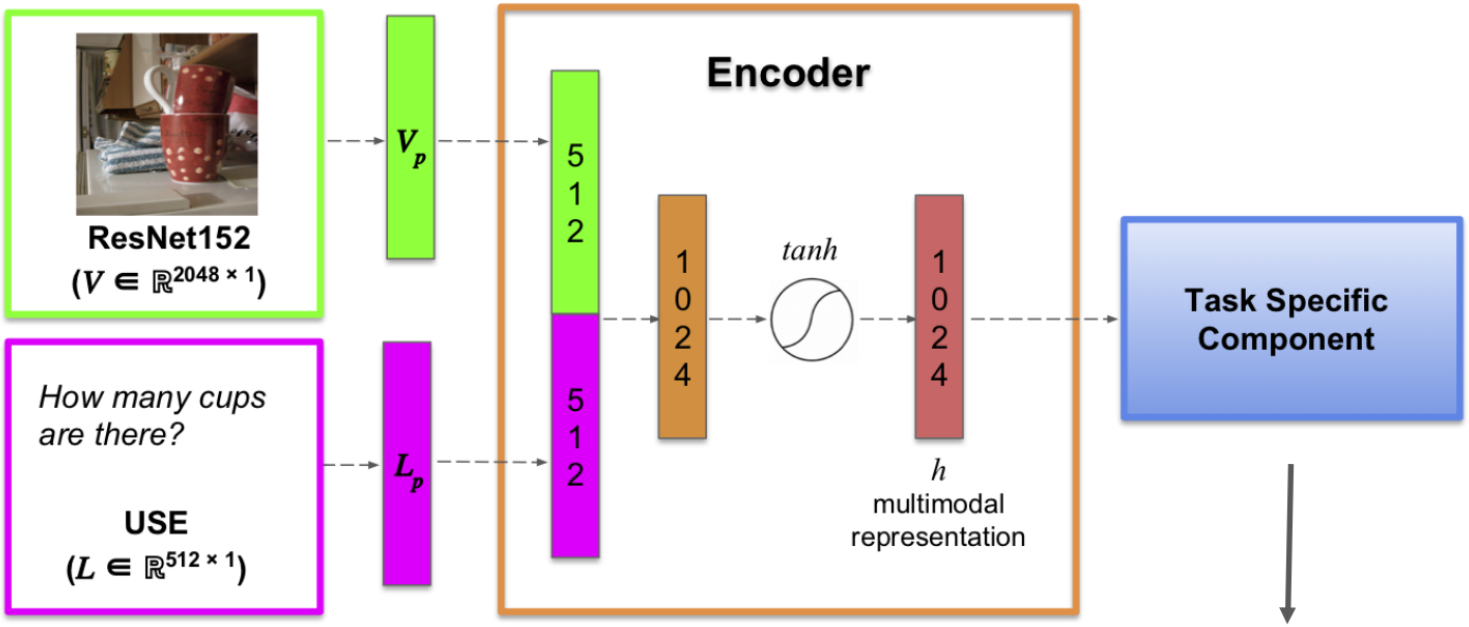
Bikers approaching a bird. **original**

Bikers approaching a dog. **foiled**

(Shekhar et al. ACL-2017)

- Spot semantic (in)congruence between an image and a caption.
- Requires compositional alignment between modalities.

General Architecture - The Hub

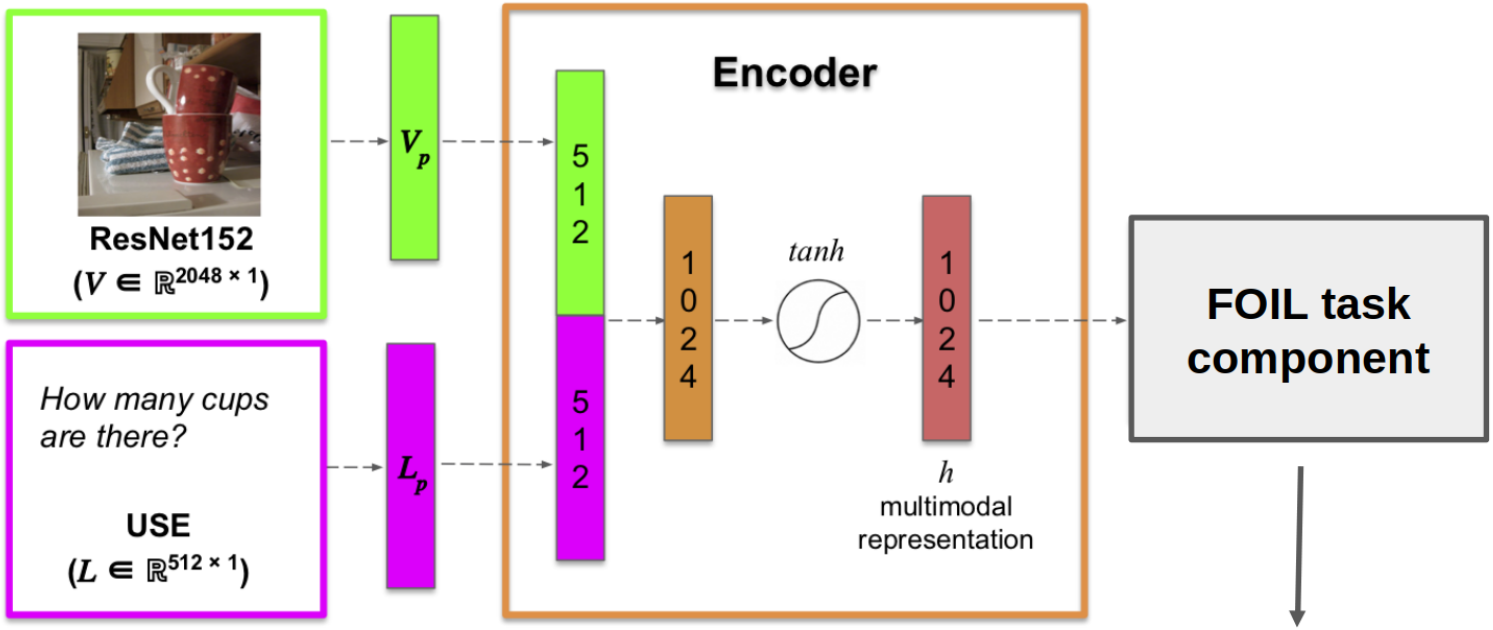


Pre-trained vectors:
- Visual features from ResNet152
- Language features from USE (Universal Sentence Encoder)

Multi-Layer Perceptron generating:
- Language embedding (VQA)
- Visual embedding (ReferIt, GW)

Trained with cosine similarity loss

Architecture of the FOIL Model

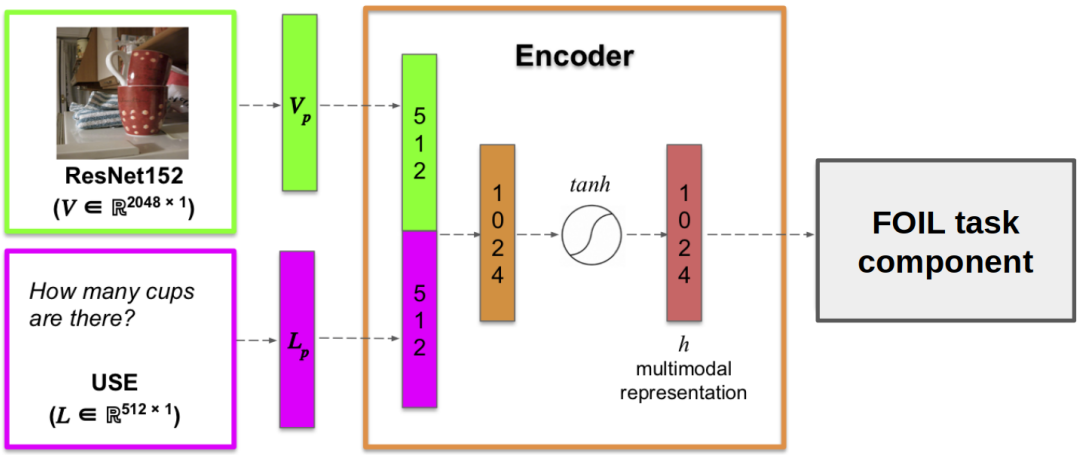


Pre-trained vectors:
- Visual features from ResNet152
- Language features from USE (Universal Sentence Encoder)

Task-specific MLPs are replaced by a fully connected layer.
Trained with cross-entropy loss.
Predicting original or foiled

Setups for the FOIL Diagnostic Task

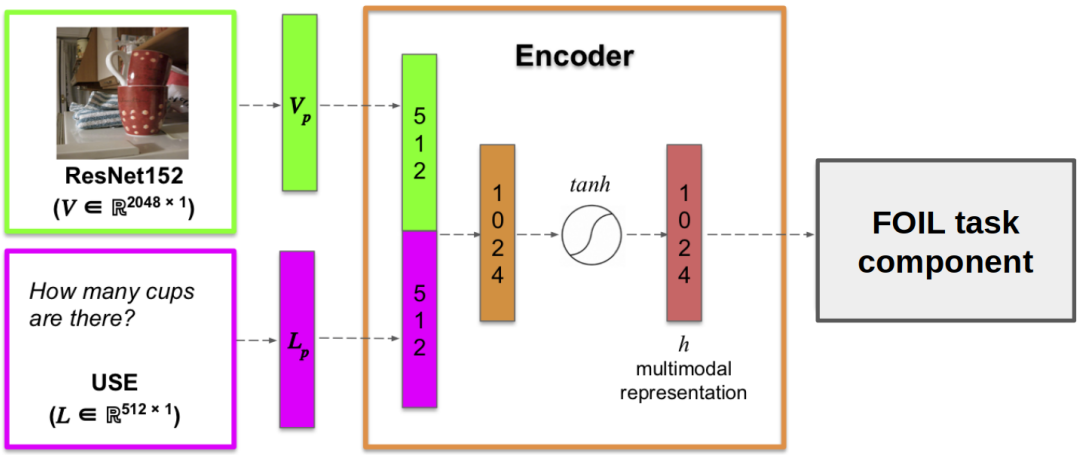
Encoder weights:



- **Trained on VQA/ReferIt/GW, then frozen**

Setups for the FOIL Diagnostic Task

Encoder weights:



- **Trained on VQA/ReferIt/GW, then frozen**
- Lower bound: frozen random initialisation
- Upper bound: trained on the FOIL task

Accuracy Results on the FOIL Diagnostic Task

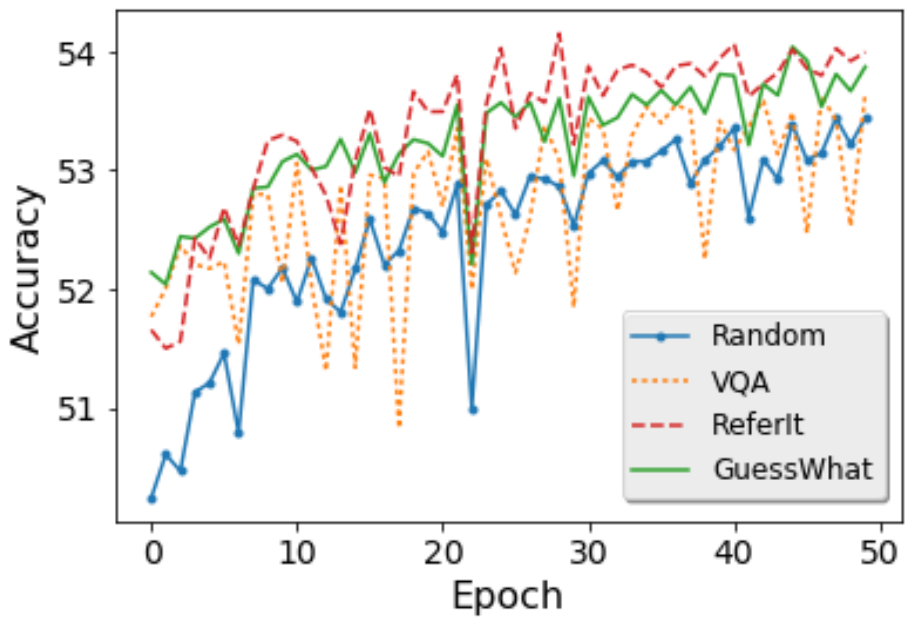
Chance accuracy: 50%

	overall	**original**	**foiled**
Lower bound	53.79	65.33	42.25
VQA	53.78	66.09	41.48
ReferIt	54.02	60.39	47.66
GuessWhat	54.18	59.02	49.34
Upper bound	67.59	87.66	47.52

- Low results overall, challenging task
- Better performance in original captions
- GW relatively better at foiled captions

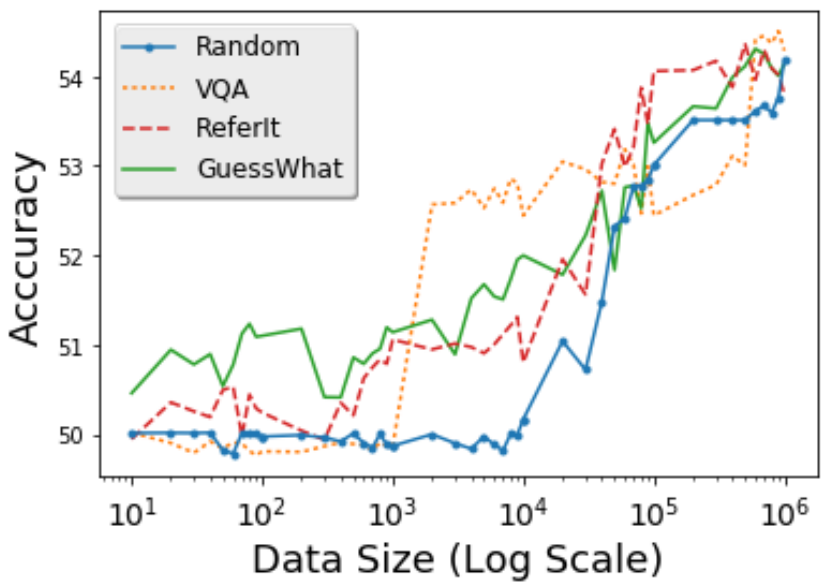
Analysis via FOIL Task: Learning over Time

- VQA is comparatively less stable over epochs
- GuessWhat and ReferIt show more stable increases in accuracy



Analysis via FOIL: Size of FOIL Training Data

- GuessWhat starts relatively better with little FOIL data: the task yields more transferable skills.
- Both GuessWhat and ReferIt increase their accuracy relatively smoothly as more FOIL data is provided.
- VQA requires more FOIL data to perform above chance and experiences a sudden accuracy increase.



What they learn?

Do models trained on these tasks learn to properly represent multimodal information?

- Using a diagnostic task
- **Inspecting multimodal semantic spaces**

What they learn?

Do models trained on these tasks learn to properly represent multimodal information?

- Using a diagnostic task
- **Inspecting multimodal semantic spaces**
  - Nearest Neighbour Overlap (NN)

Analysis via Nearest Neighbour Overlap

Is one of the modalities is given more weightage compare to other?

Analysis via Nearest Neighbour Overlap

Is one of the modalities is given more weightage compare to other?

- For each of the 80 MS-COCO object categories: average visual vector and linguistic vector of the category word.
- Check overlap of nearest neighbours in multimodal space vs. linguistic and visual space, respectively.
- v_{cat} $(v_{dog}, \boldsymbol{v_{tiger}}, \boldsymbol{v_{lion}})(v_{mouse}, \boldsymbol{v_{tiger}}, \boldsymbol{v_{lion}})$

Analysis via Nearest Neighbour Overlap

Is one of the modalities is given more weightage compare to other?

- For each of the 80 MS-COCO object categories: average visual vector and linguistic vector of the category word.
- Check overlap of nearest neighbours in multimodal space vs. linguistic and visual space, respectively.
- v_{cat} $(v_{dog}, \boldsymbol{v_{tiger}}, \boldsymbol{v_{lion}})(v_{mouse}, \boldsymbol{v_{tiger}}, \boldsymbol{v_{lion}})$

k=10	Vision	Language
VQA	0.703	0.365
ReferIt	0.780	0.386
GW	0.689	0.359
FOIL	0.246	0.291

FOIL multimodal space more abstract and balanced across modalities.
Multimodal spaces learned by other tasks are closer to the visual space.

Conclusion

- Introduces common Language and Vision Tasks and Models
- Common vision & language tasks do not lead to learning fine-grained multimodal understanding skills.
- Representations learned via VQA are less stable and transferable.
- Representations learned via GW dialogue seem to have more desirable properties
  - slightly higher accuracy on FOIL diagnostic task
  - with less FOIL data required
  - less bias/weightage towards the visual modality

Thank you!!!
Q&A

Collaborators

- **University of Trento:** Raffaella Bernardi, Alberto Testoni, Sandro Pezzelle, Yauhen Klimovich, Aurelie Herbelot, Moin Nabi, Enver Sangineto
- **University of Amsterdam:** Raquel Fernández, Aashish Venkatesh, Tim Baumgärtner, Elia Bruni
- **IT University of Copenhagen:** Barbara Plank