Visual Question Answering and Visual Reasoning

Zhe Gan

6/15/2020

Microsoft

Overview

- Goal of this part of the tutorial:
  - Use VQA and visual reasoning as example tasks to understand Vision-and-Language representation learning
  - After the talk, everyone can confidently say: “yeah, I know VQA and visual reasoning pretty well now”
  - Focus on high-level intuitions, not technical details
  - Focus on static images, instead of videos
  - Focus on a selective set of papers, not a comprehensive literature review

Agenda

- Task Overview
    - *What are the main tasks that are driving progress in VQA and visual reasoning?*
- Method Overview
    - *What are the state-of-the-art approaches and the key model design principles underlying these methods?*
- Summary
    - *What are the core challenges and future directions?*

Agenda

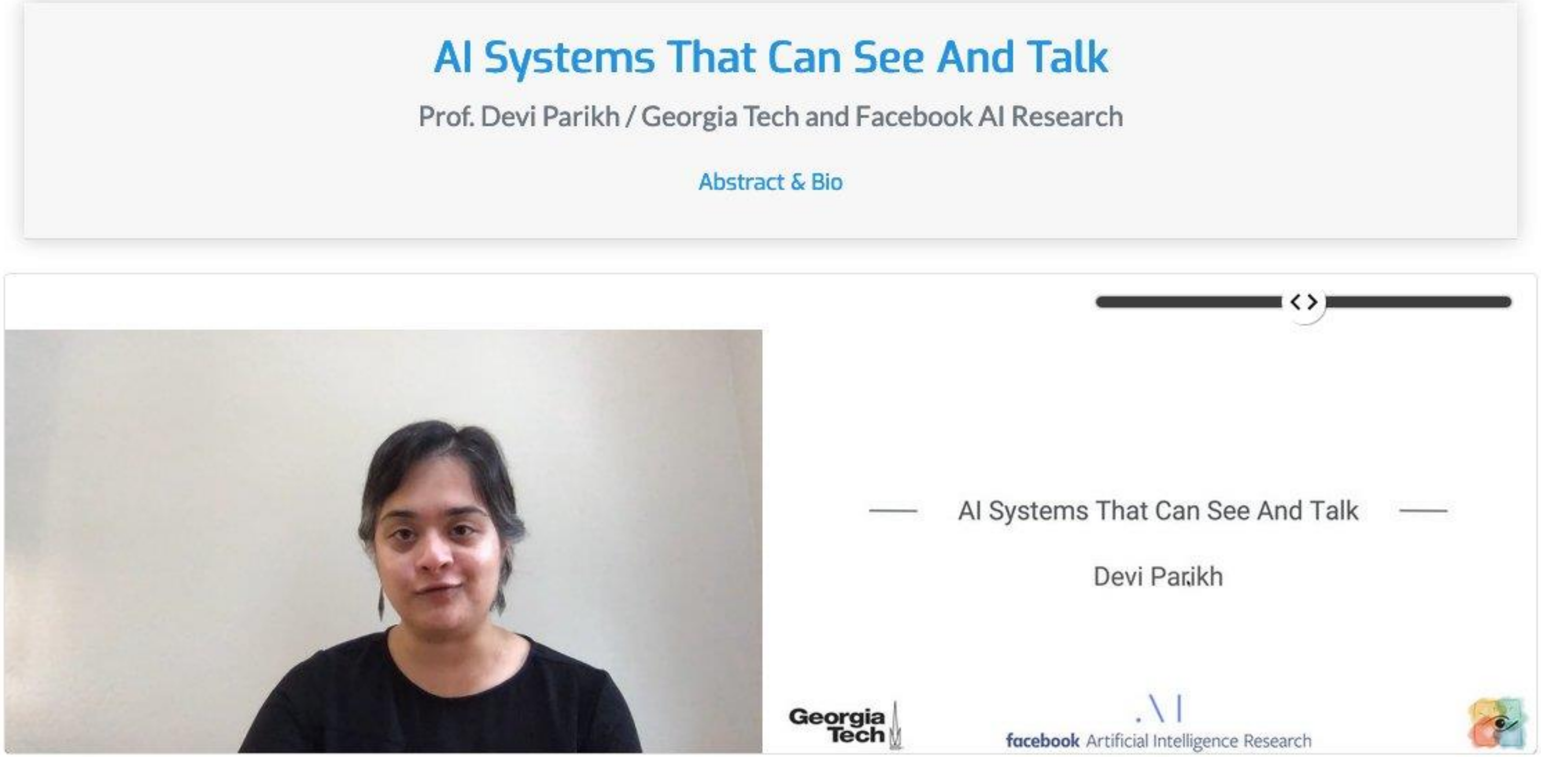
- Task Overview
  - *What are the main tasks that are driving progress in VQA and visual reasoning?*

- Method Overview
  - *What are the state-of-the-art approaches and the key model design principles underlying these methods?*

- Summary
  - *What are the core challenges and future directions?*

What is V+L about?

- V+L research is about how to train a smart AI system that can see and talk



What is V+L about?

- V+L research is about how to train a smart AI system that can see and talk

Prof. Yann LeCun’s cake theory

Reinforcement Learning

Supervised Learning

Unsupervised/Self-supervised Learning

In our V+L context

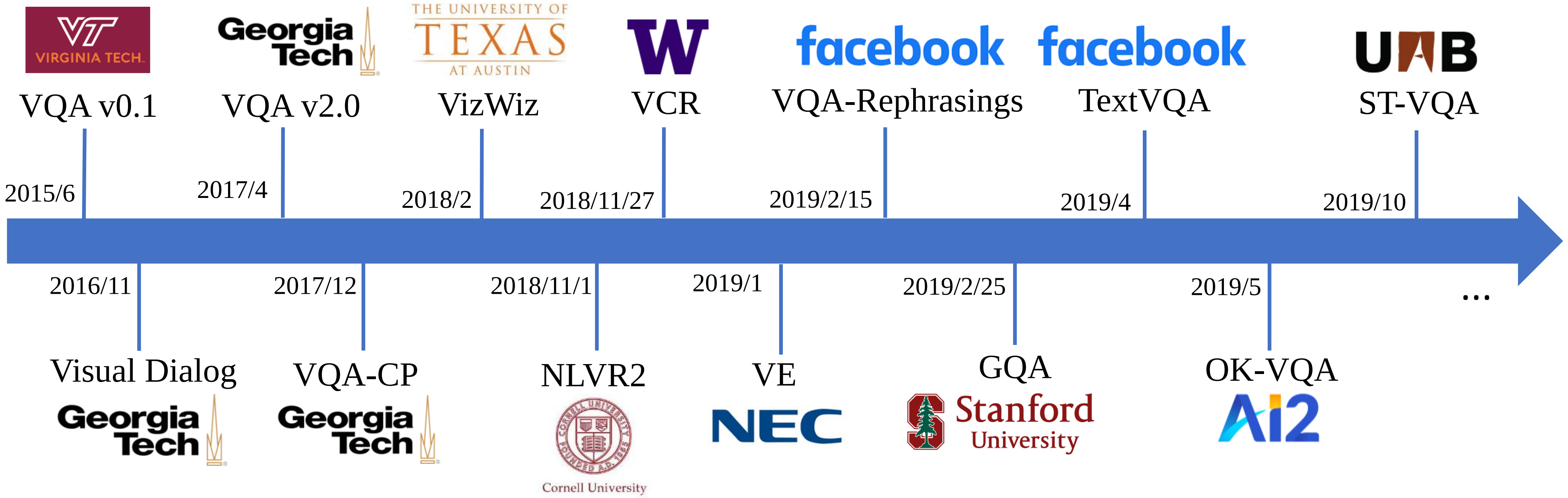
Multimodel Intelligence

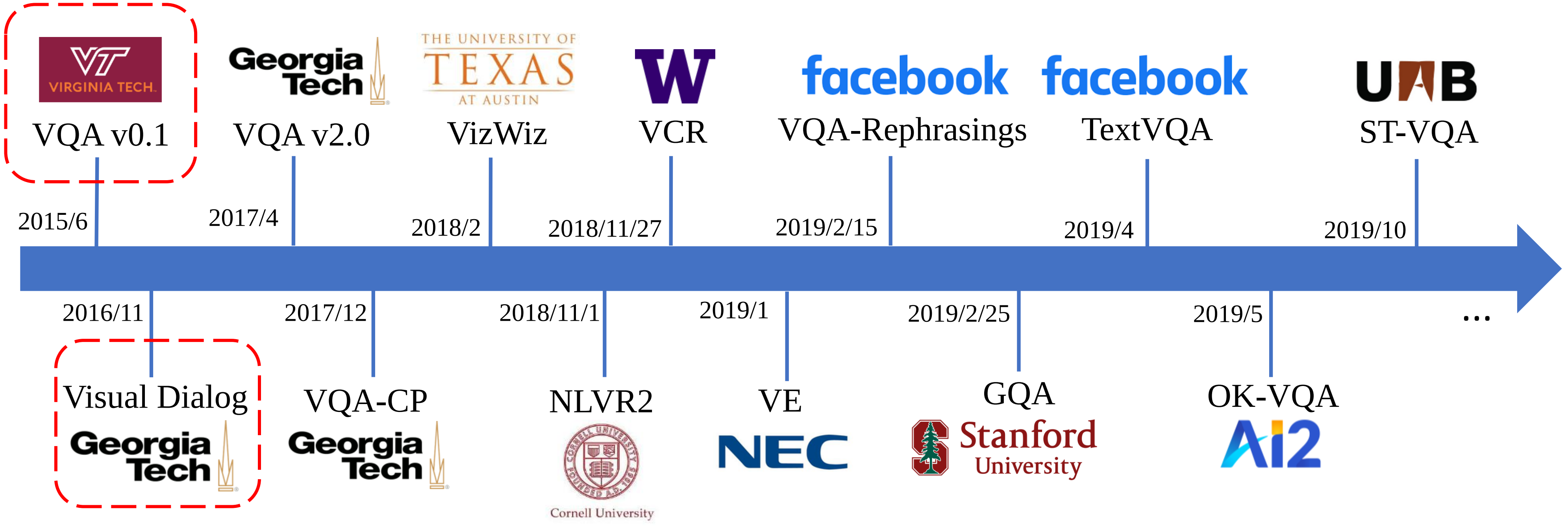
BERT Language Understanding

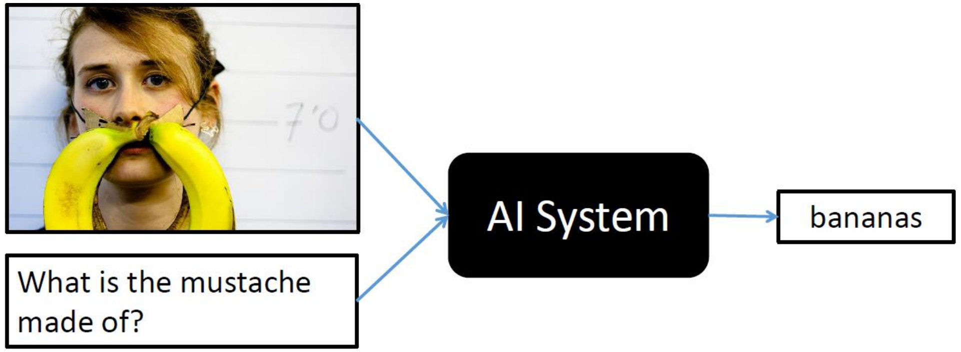
ResNet Visual Understanding

Task Overview: VQA and Visual Reasoning

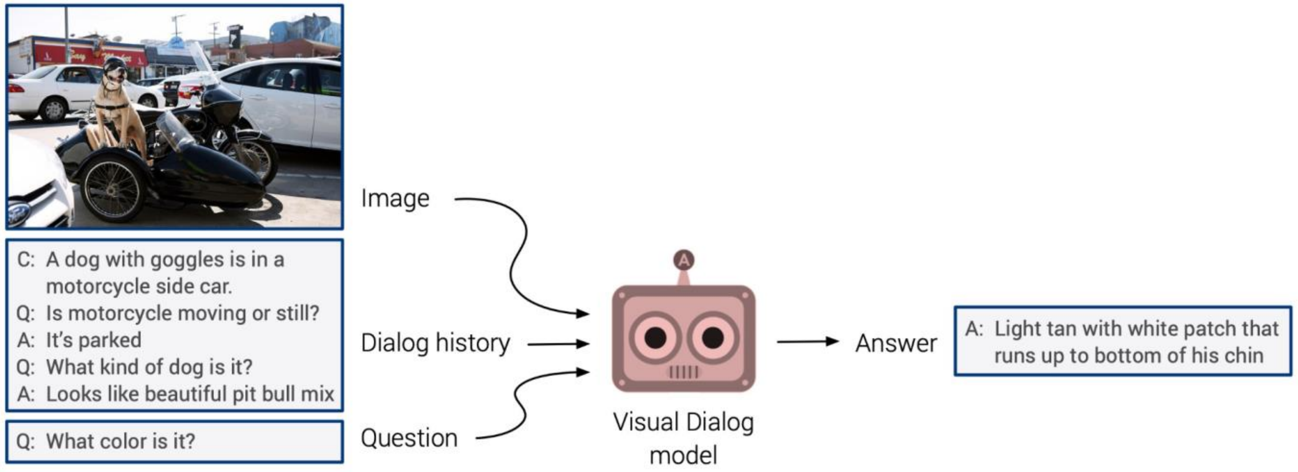
- Large-scale annotated datasets have driven tremendous progress in this field







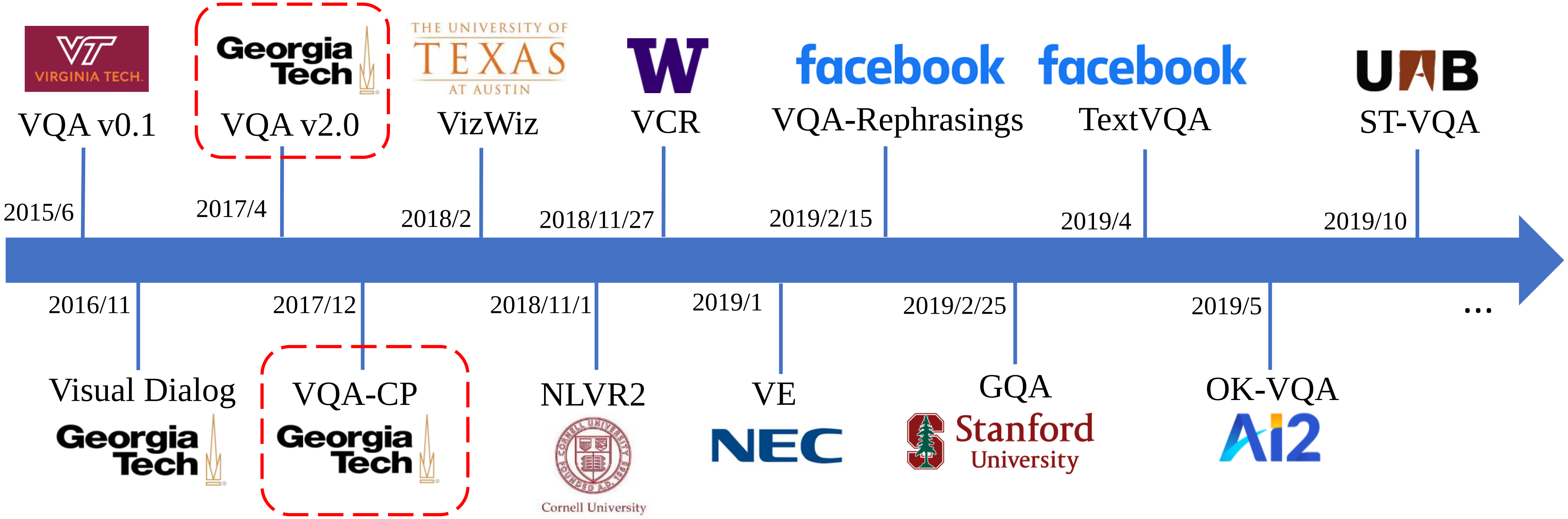
VQA



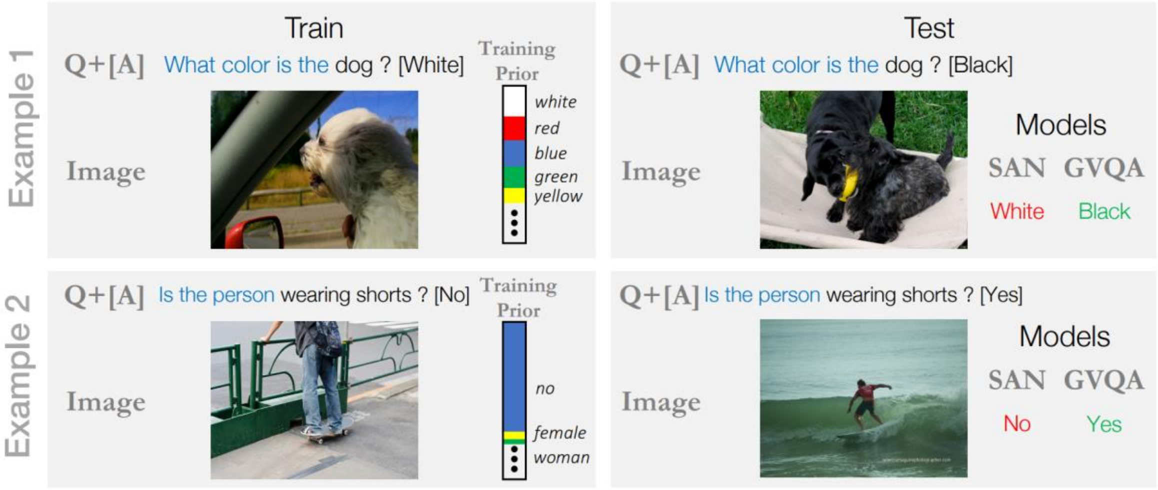
Visual Dialog

Image credit: https://visualqa.org/, https://visualdialog.org/

[1] VQA: Visual Question Answering, ICCV 2015
[2] Visual Dialog, CVPR 2017

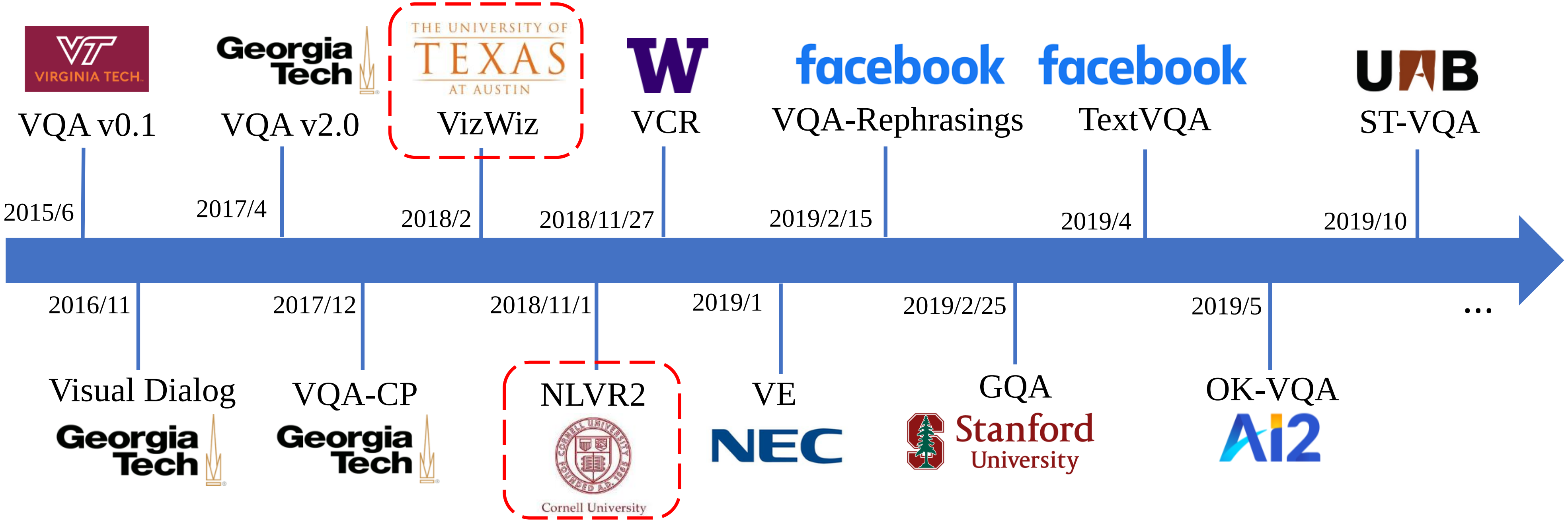


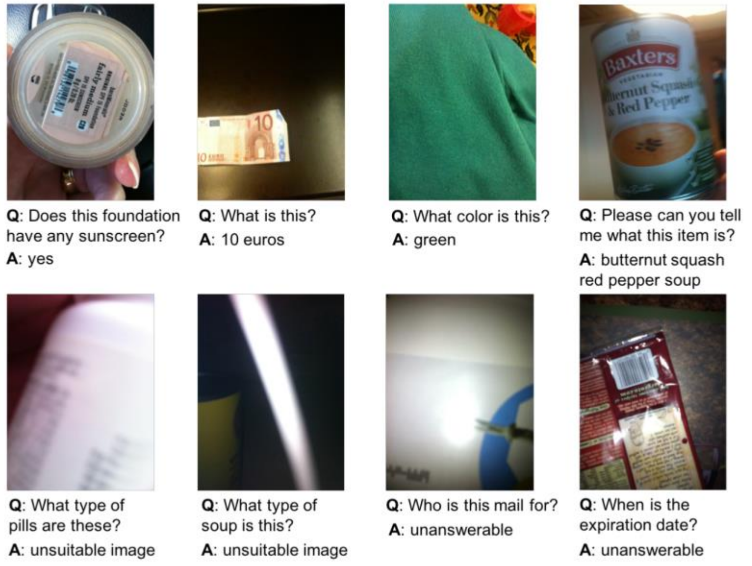
VQA v2.0



VQA-CP

[1] Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering, CVPR 2017
[2] Don’t Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering, CVPR 2018

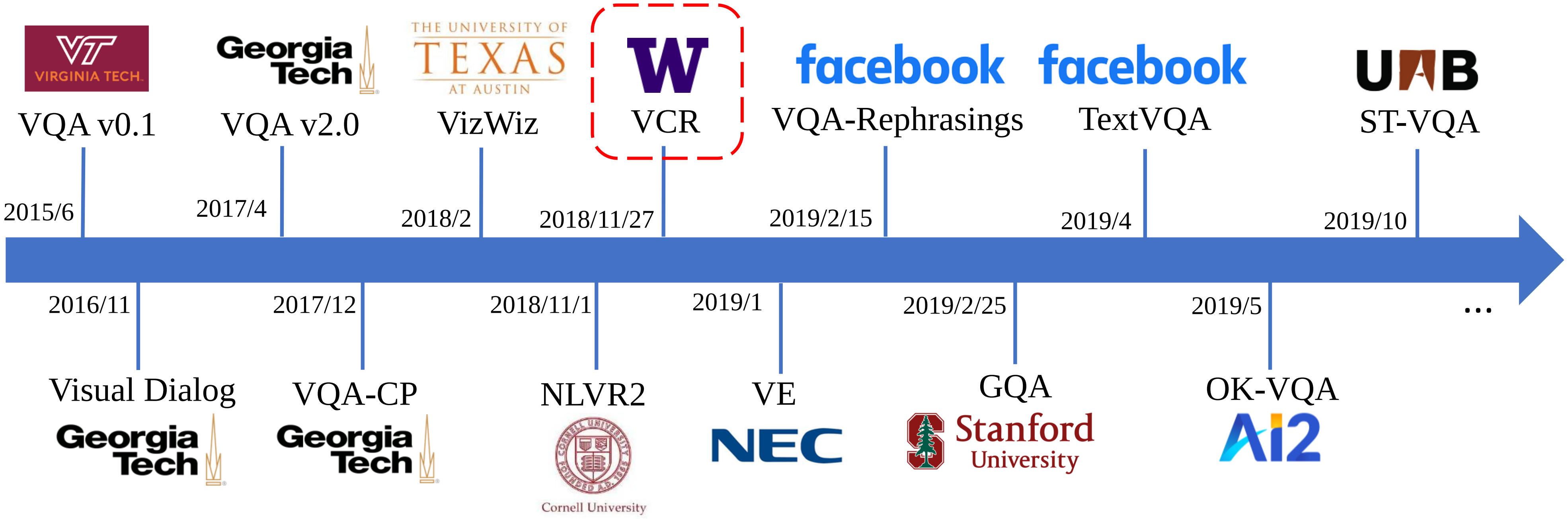


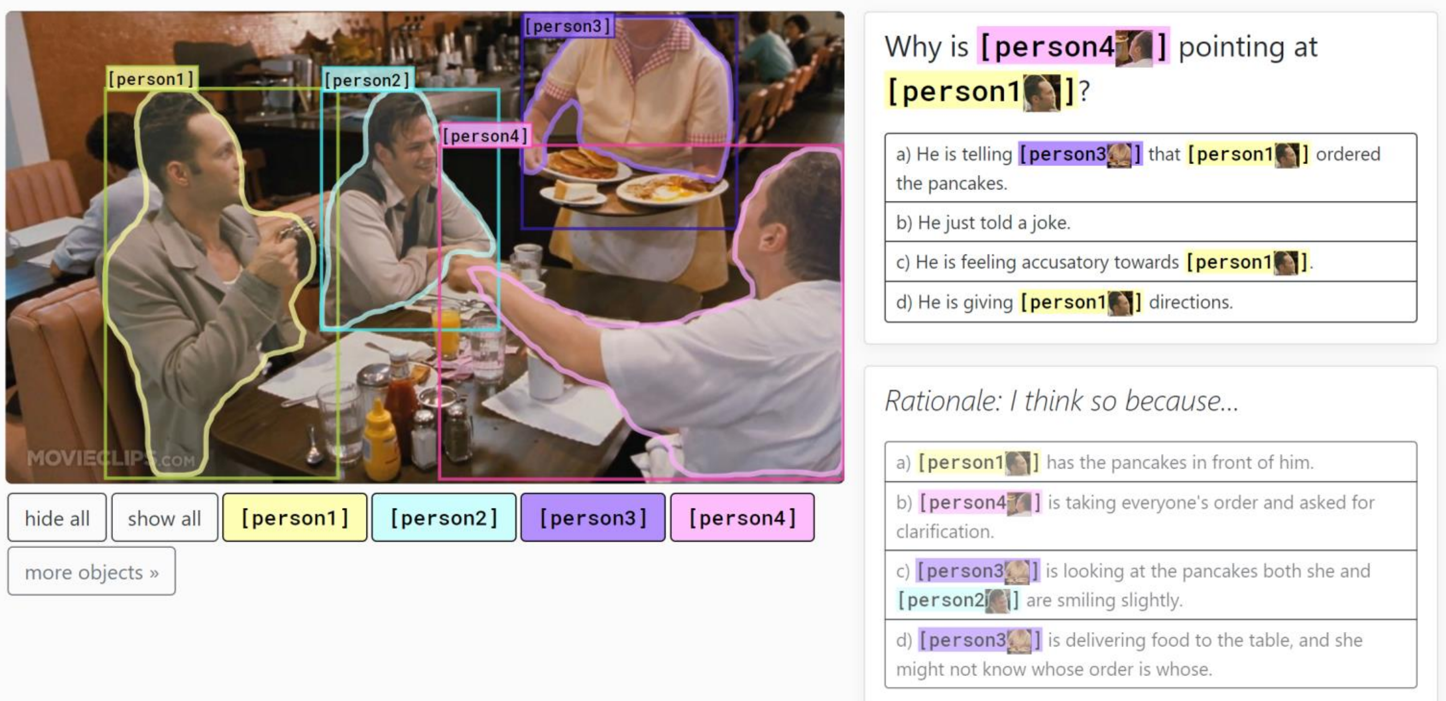


VizWiz

NLVR2

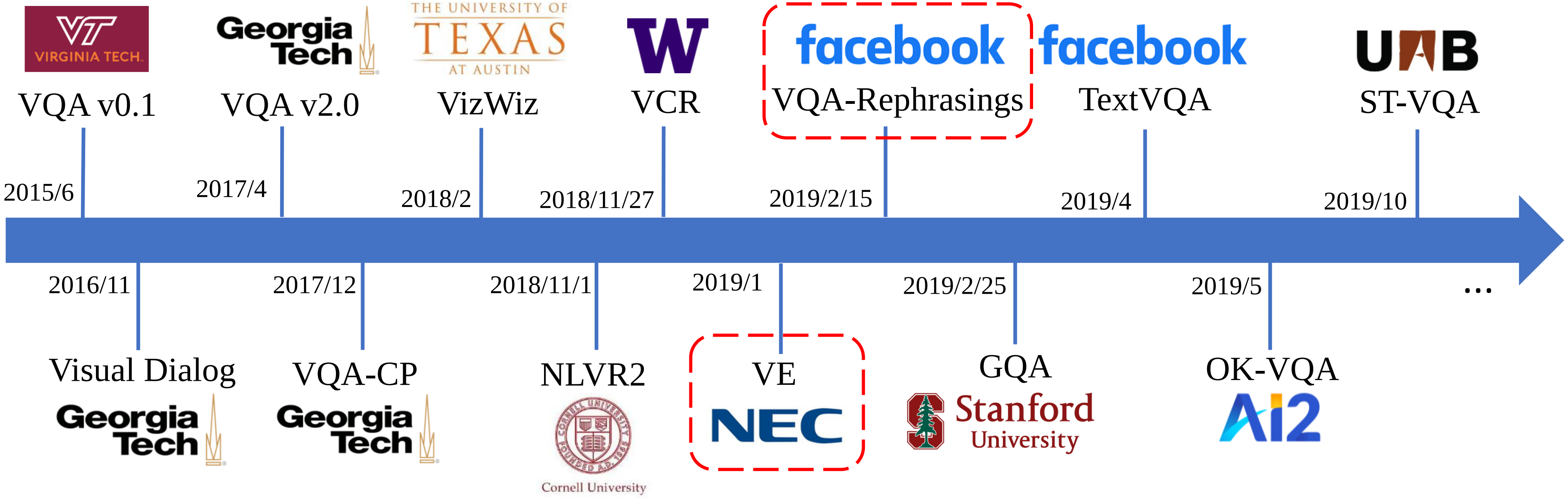
[1] VizWiz Grand Challenge: Answering Visual Questions from Blind People, CVPR 2018
[2] A Corpus for Reasoning About Natural Language Grounded in Photographs, ACL 2019

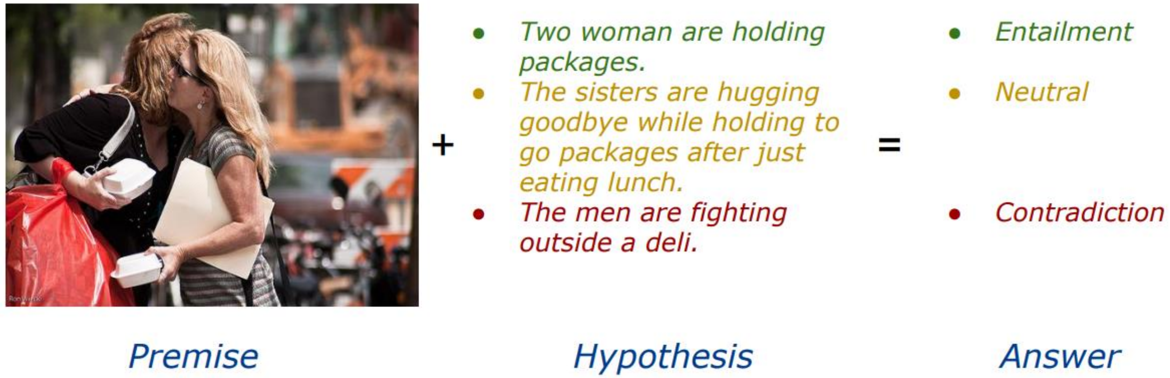




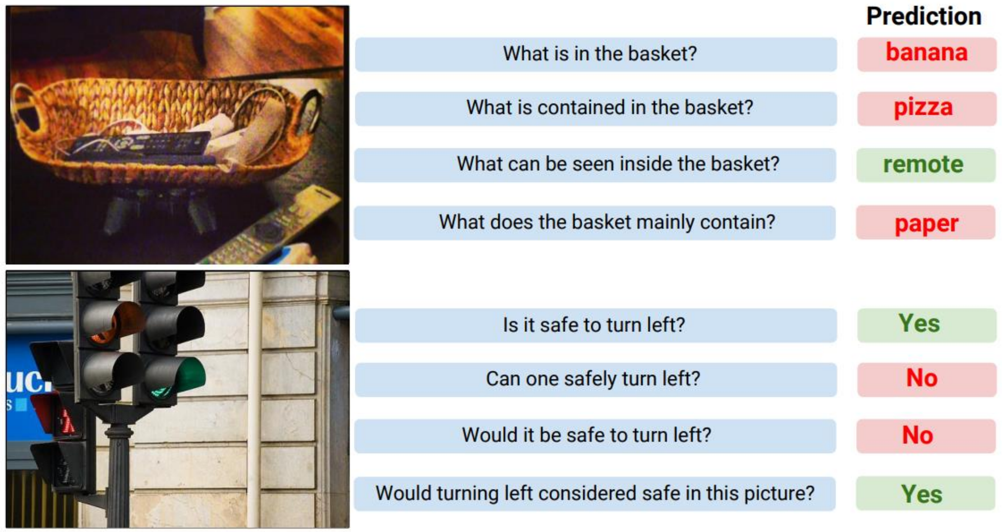
VCR
VISUAL COMMONSENSE REASONING

[1] From Recognition to Cognition: Visual Commonsense Reasoning, CVPR 2019



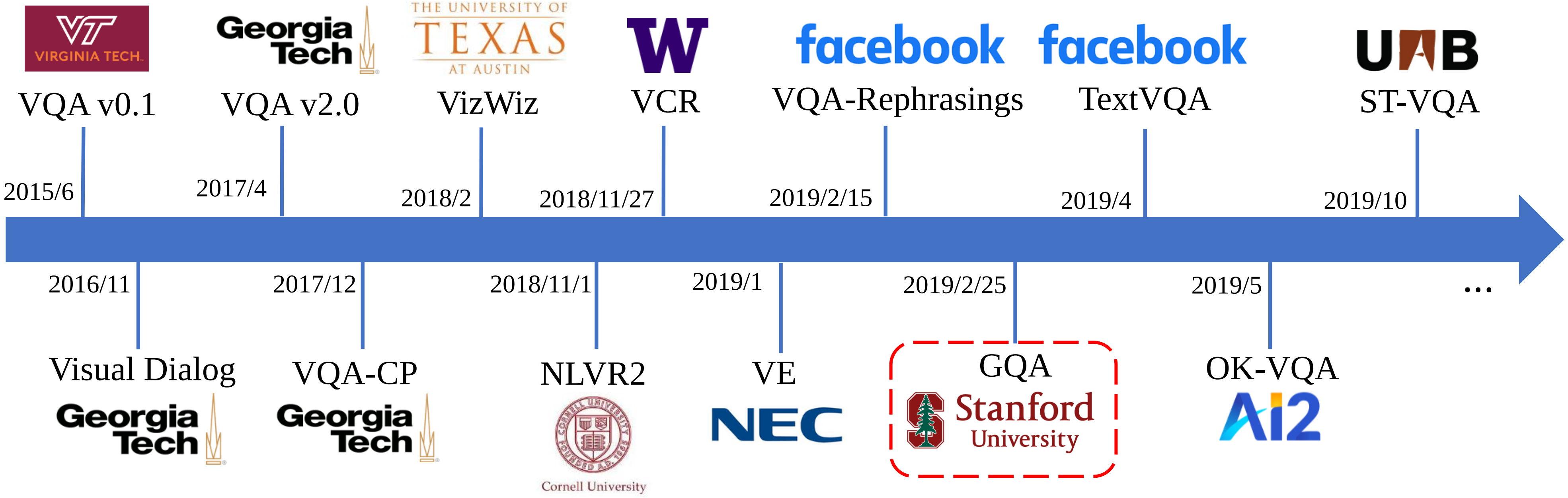


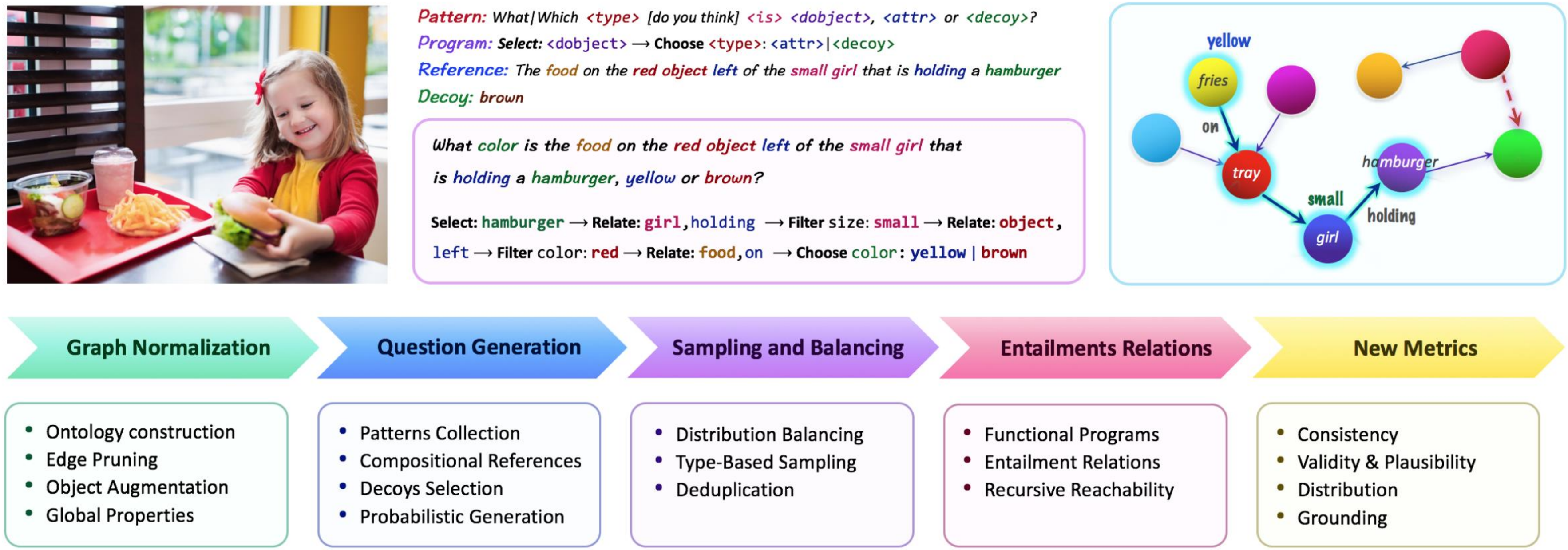
Visual Entailment



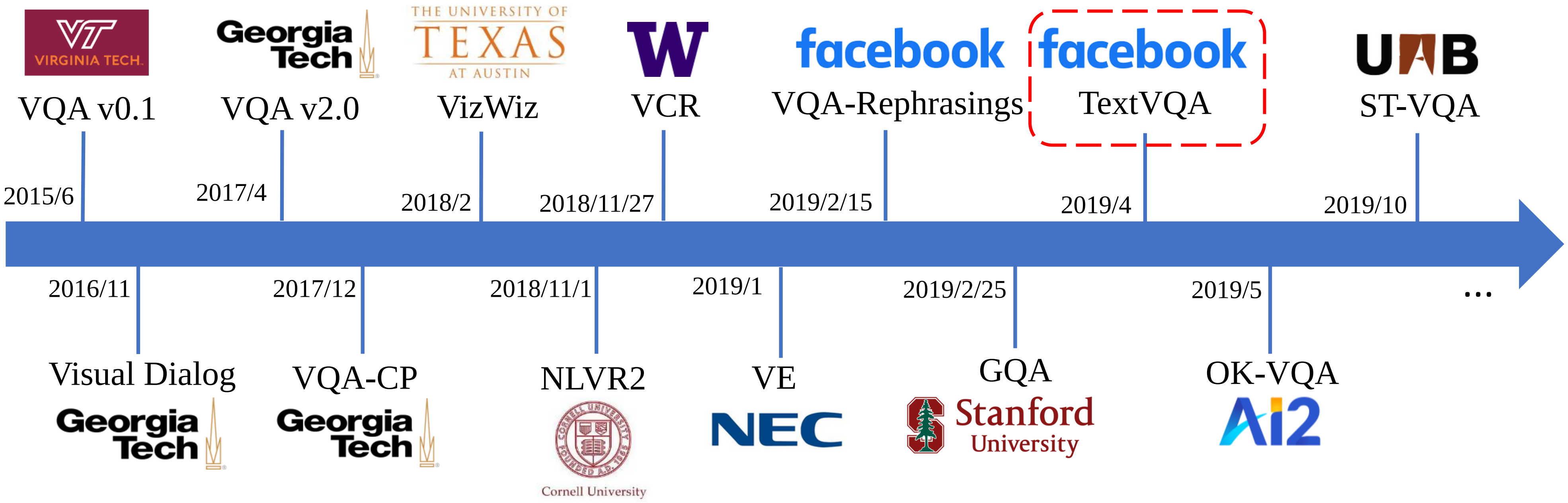
VQA-Rephrasings

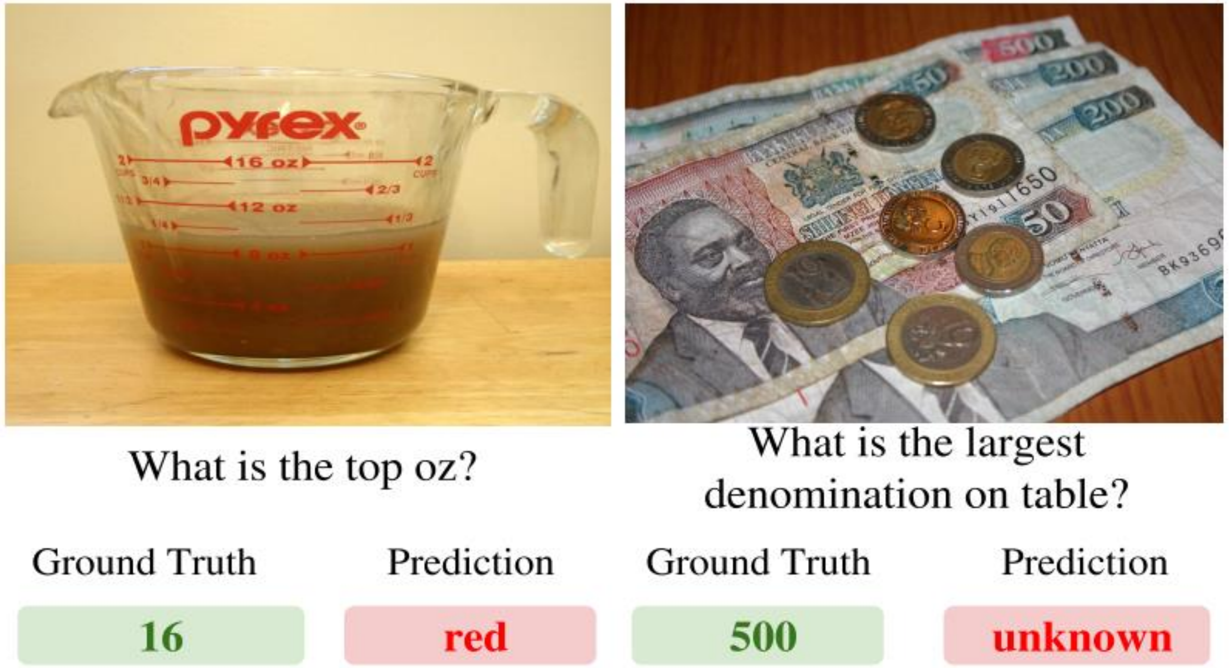
[1] Visual Entailment: A Novel Task for Fine-Grained Image Understanding, 2019
[2] Cycle-Consistency for Robust Visual Question Answering, CVPR 2019





[1] GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering, CVPR 2019

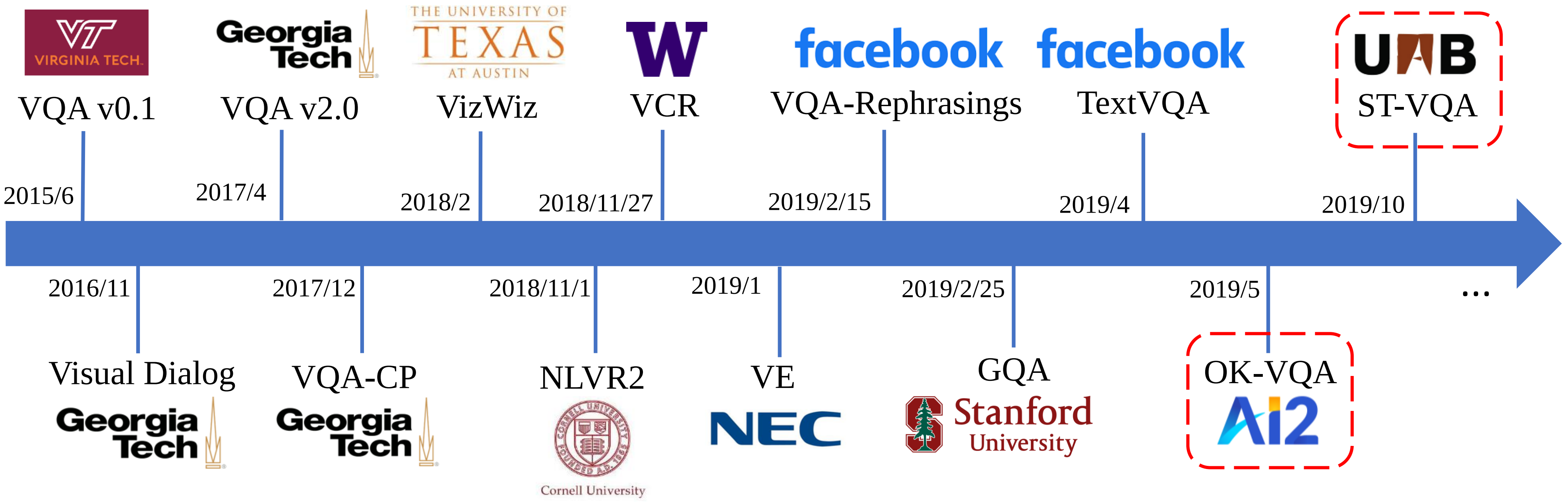




TextVQA

A dataset to benchmark visual reasoning based on text in images.

[1] Towards VQA Models That Can Read, CVPR 2019



Q: Which American president is associated with the stuffed animal seen here?

A: Teddy Roosevelt

Outside Knowledge

Another lasting, popular legacy of Roosevelt is the stuffed toy bears—teddy bears—named after him following an incident on a hunting trip in Mississippi in 1902.

Developed apparently simultaneously by toymakers ... and named after President Theodore "Teddy" Roosevelt, the teddy bear became an iconic children's toy, celebrated in story, song, and film.

At the same time in the USA, Morris Michtom created the first teddy bear, after being inspired by a drawing of Theodore "Teddy" Roosevelt with a bear cub.

OK-VQA

Q: What is the price of the bananas per kg?
A: $11.98

Q: What does the red sign say?
A: Stop

Scene Text VQA

[1] OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge, CVPR 2019
[2] Scene Text Visual Question Answering, ICCV 2019

More datasets…

SQuINTing at VQA Models: Interrogating VQA Models with Sub-Questions

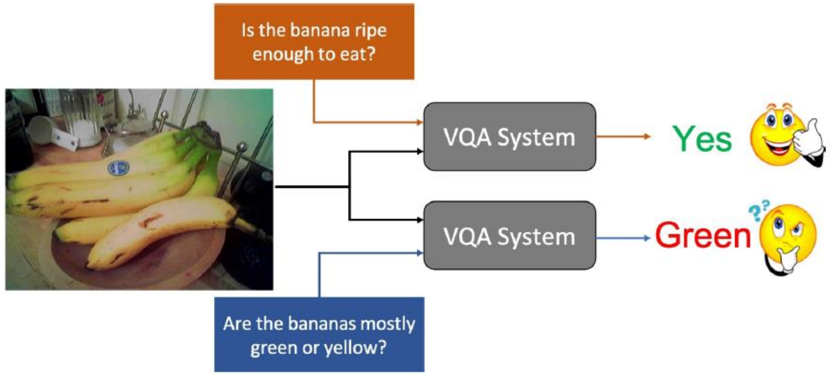
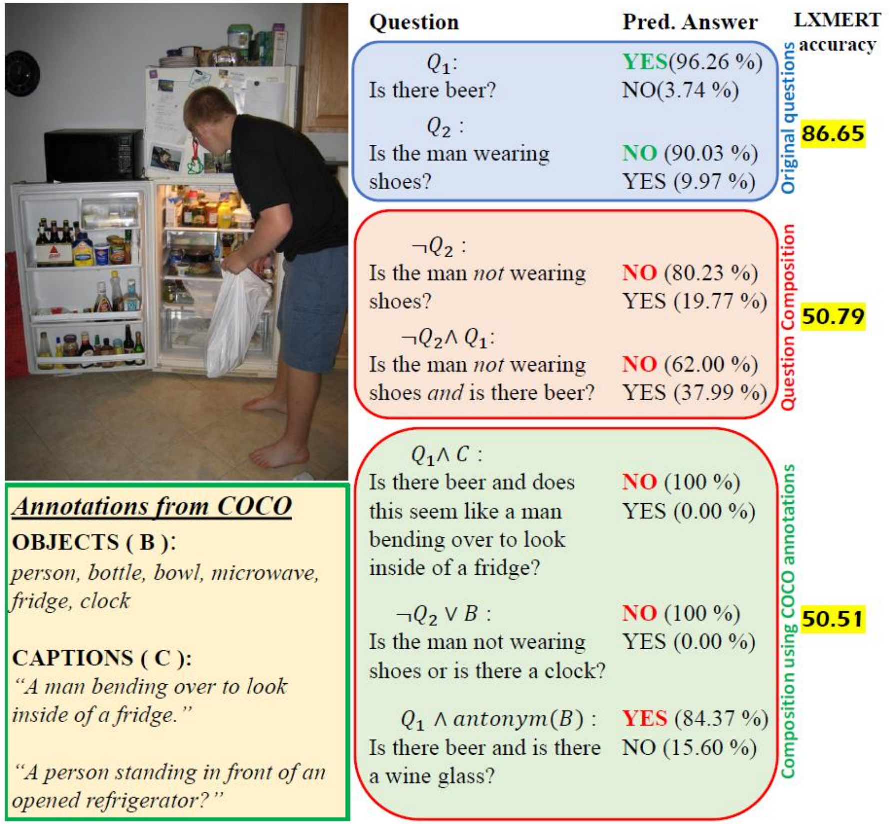


Figure 1: A potential reasoning failure: Current models answer "Yes" correctly to the Reasoning question "Is the banana ripe enough to eat?". We might assume that correctly answering the Reasoning question stems from perceiving relevant concepts correctly – perceiving yellow bananas in this example. But when asked "Are the bananas mostly green or yellow?", it answers "Green" incorrectly – indicating that the model possibly answered the original for the wrong reasons even if the answer was right. We quantify the extent to which this phenomenon occurs in VQA and introduce a new dataset aimed at stimulating research on well grounded reasoning.

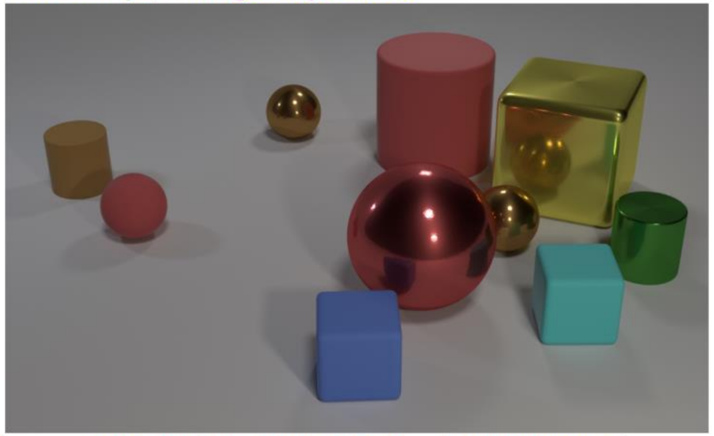
VQA-LOL: Visual Question Answering under the Lens of Logic



Diagnostic Datasets

- CLEVR (Compositional Language and Elementary Visual Reasoning)
  - Has been extended to visual dialog (CLEVR-Dialog), referring expressions (CLEVR-Ref+), and video reasoning (CLEVRER)

Questions in CLEVR test various aspects of visual reasoning including **attribute identification**, **counting**, **comparison**, **spatial relationships**, and **logical operations**.

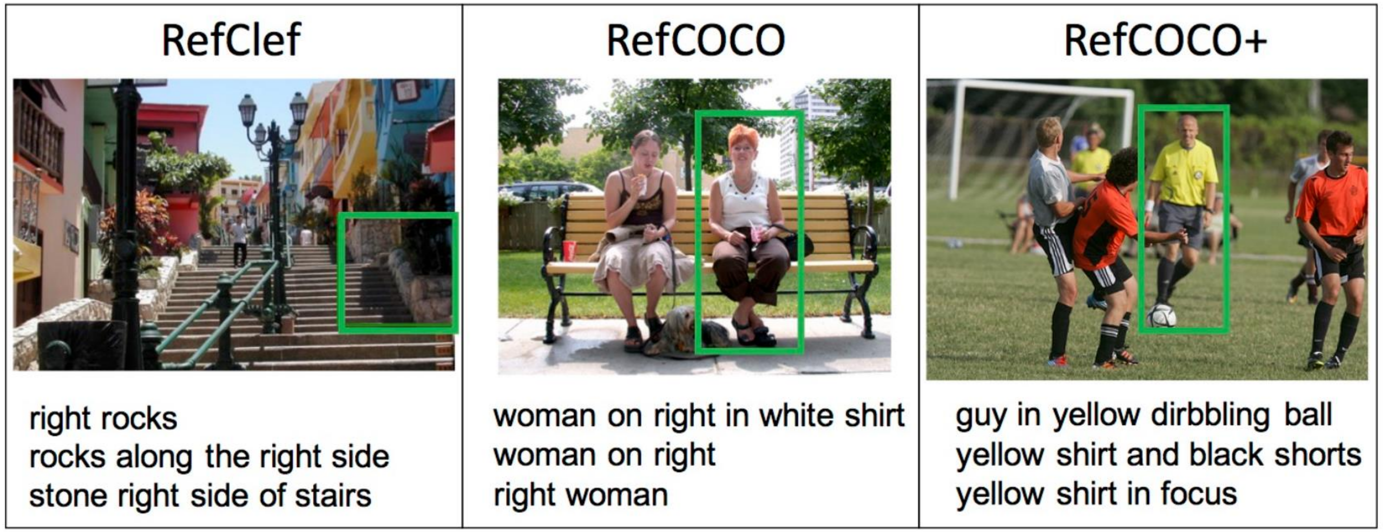


Q: Are there an **equal number** of **large things** and **metal spheres**?
Q: **What size** is the **cylinder** **that is left of** the **brown metal** thing **that is left of** the **big sphere**?
Q: There is a **sphere** with the **same size as** the **metal cube**; is it **made of the same material as** the **small red sphere**?
Q: **How many** objects are **either** **small cylinders** or **red** things?

[1] CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning, CVPR 2017
[2] CLEVR-Dialog: A Diagnostic Dataset for Multi-Round Reasoning in Visual Dialog, NAACL 2019
[3] CLEVR-Ref+: Diagnosing Visual Reasoning with Referring Expressions, CVPR 2019
[4] CLEVRER: CoLlision Events for Video REpresentation and Reasoning, ICLR 2020

Beyond VQA: Visual Grounding

- Referring Expression Comprehension: RefCOCO(+/g)
  - ReferIt Game: Referring to Objects in Photographs of Natural Scenes
- Flickr30k Entities



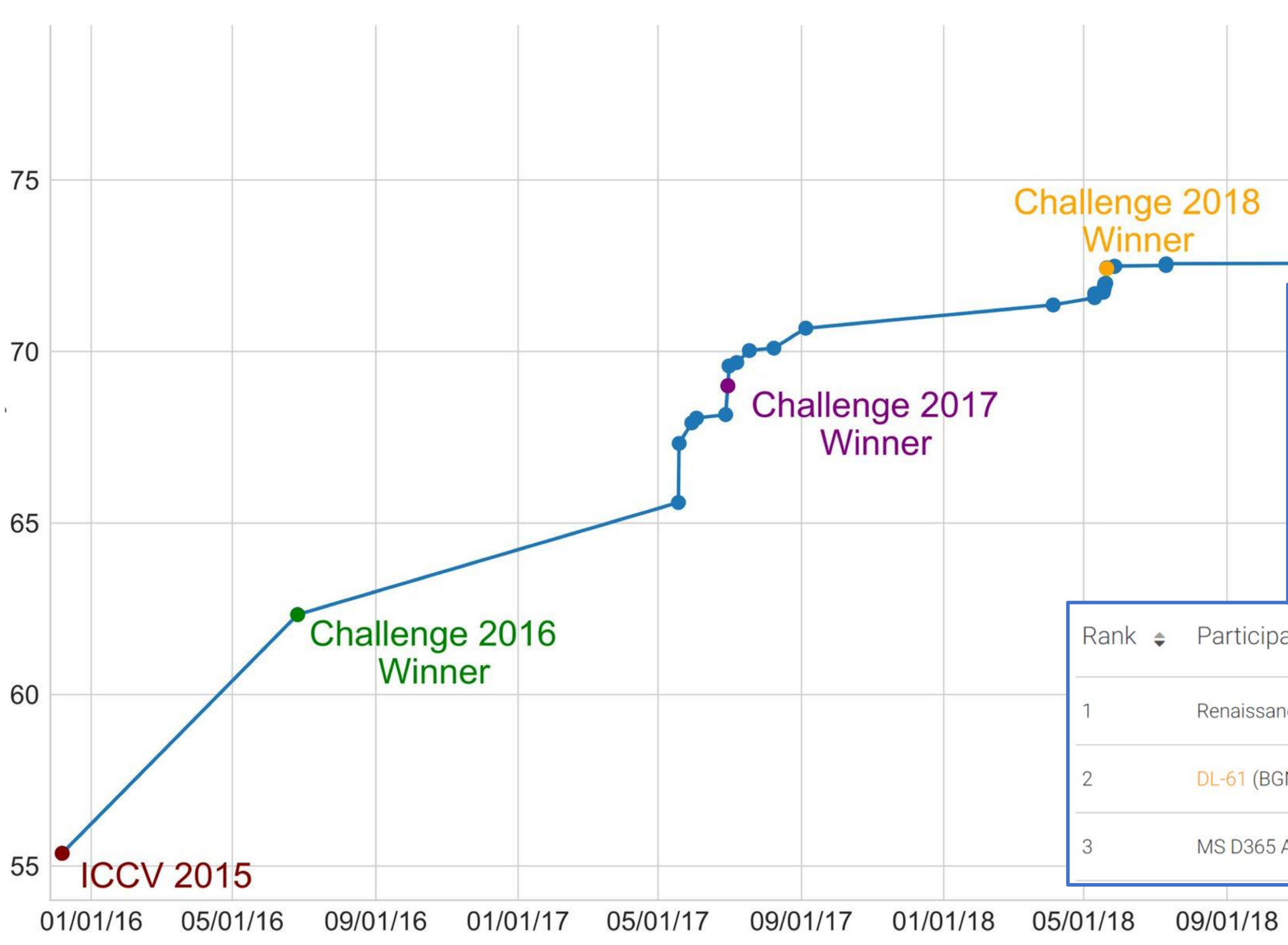
[1] OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge, EMNLP 2014
[2] Flickr30K Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models, IJCV 2017

Beyond VQA: Visual Grounding

- PhraseCut: Language-based image segmentation

[1] PhraseCut: Language-based Image Segmentation in the Wild, CVPR 2020

Visual Question Answering



Rank	Participant team	yes/no	number	other	overall
1	Renaissance (StructVBERT-base Ensemble)	90.71	59.80	66.92	76.01
2	DL-61 (BGN, ensemble)	90.89	61.13	66.28	75.92
3	MS D365 AI (VILLA Ensemble)	91.30	59.23	66.20	75.85

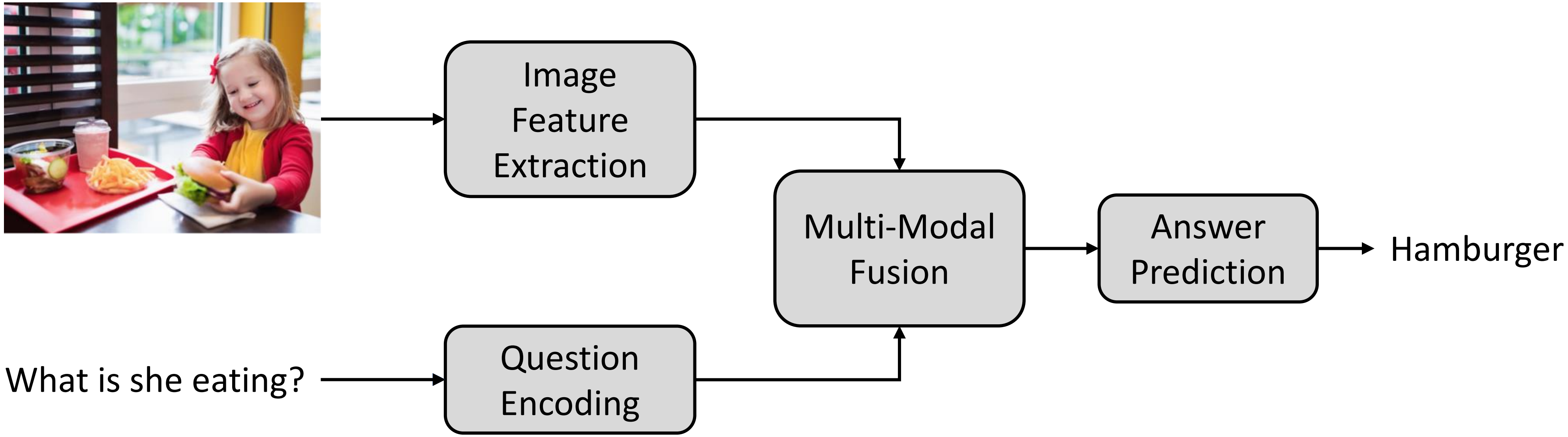
Image Credit: CVPR 2019 Visual Question Answering and Dialog Workshop

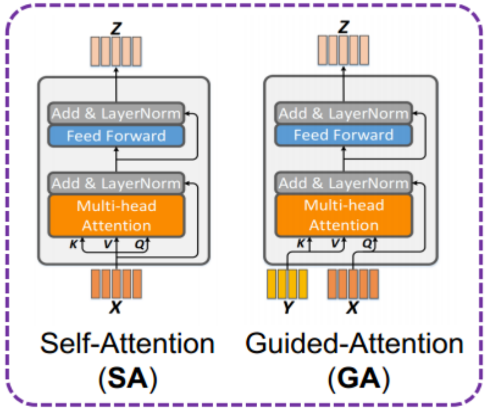
Agenda

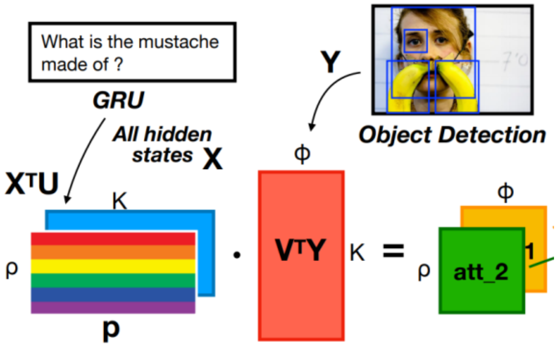
- Task Overview
  - *What are the main tasks that are driving progress in VQA and visual reasoning?*
- Method Overview
  - *What are the state-of-the-art approaches and the key model design principles underlying these methods?*
- Summary
  - *What are the core challenges and future directions?*

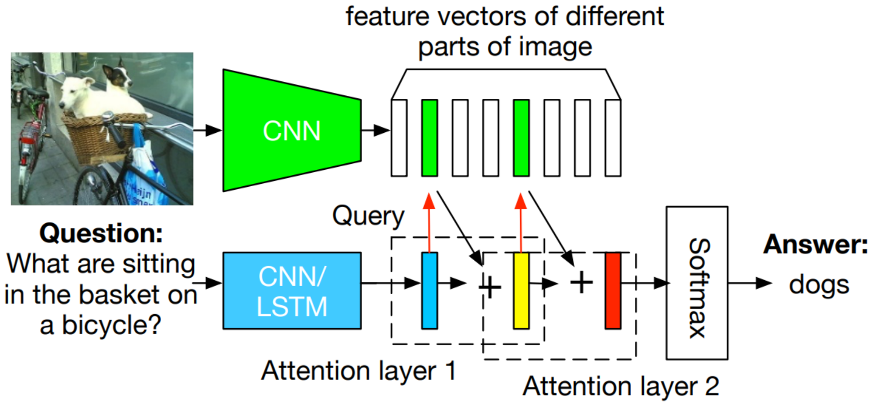
Overview

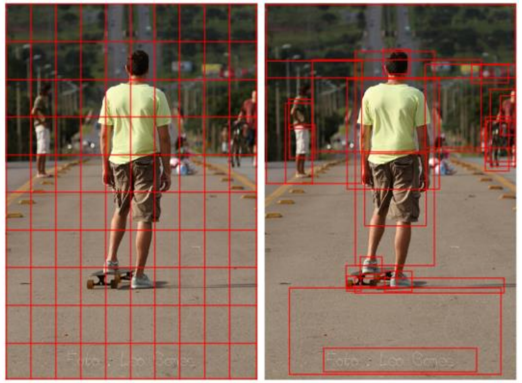
- How a typical system looks like

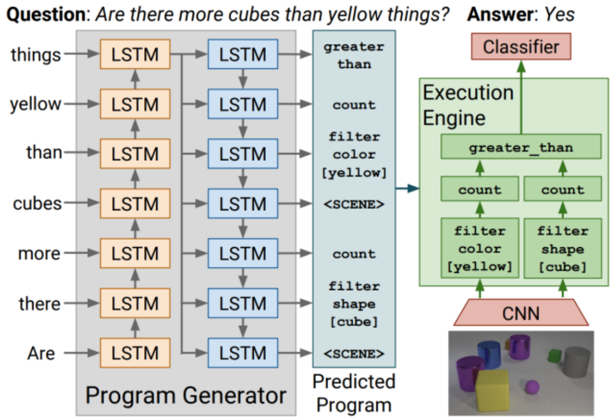


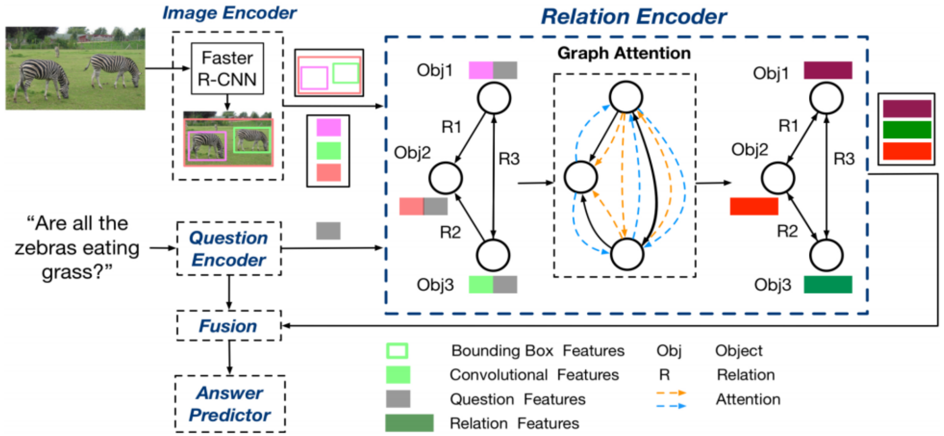


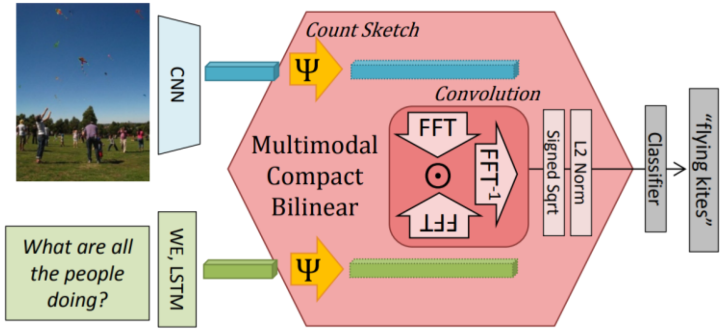


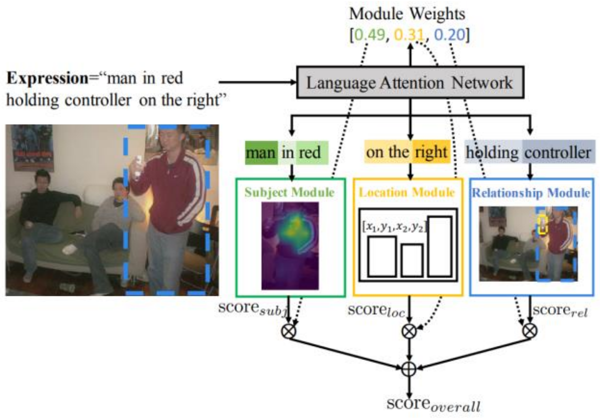












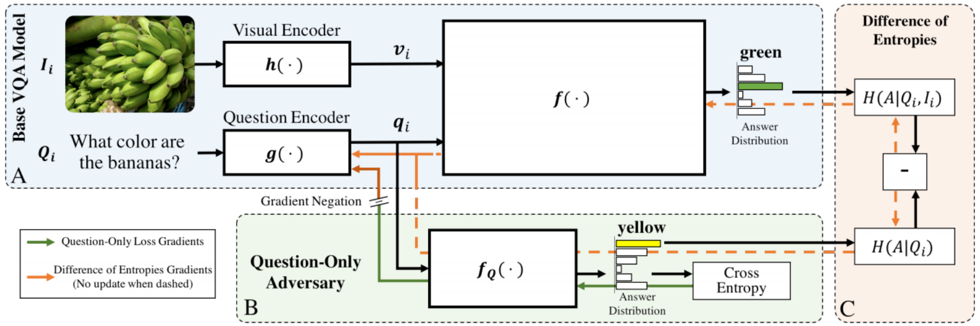


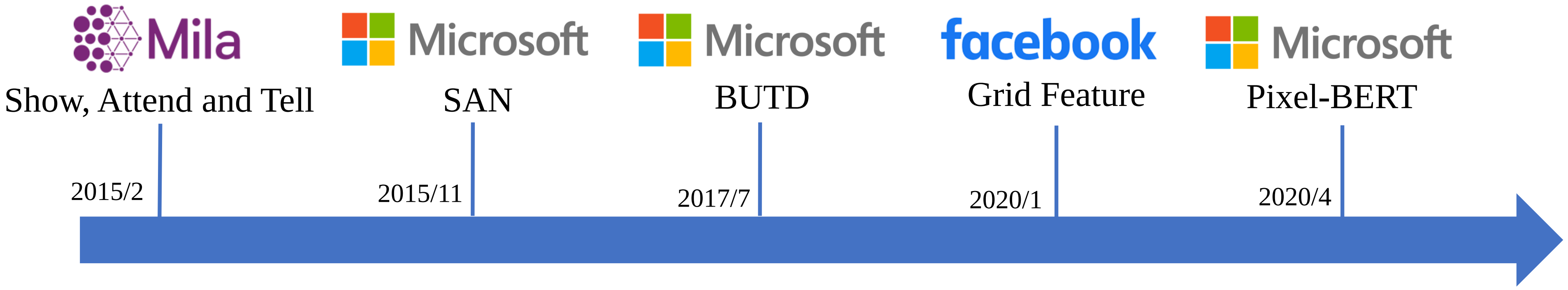
Image credit: from the original papers

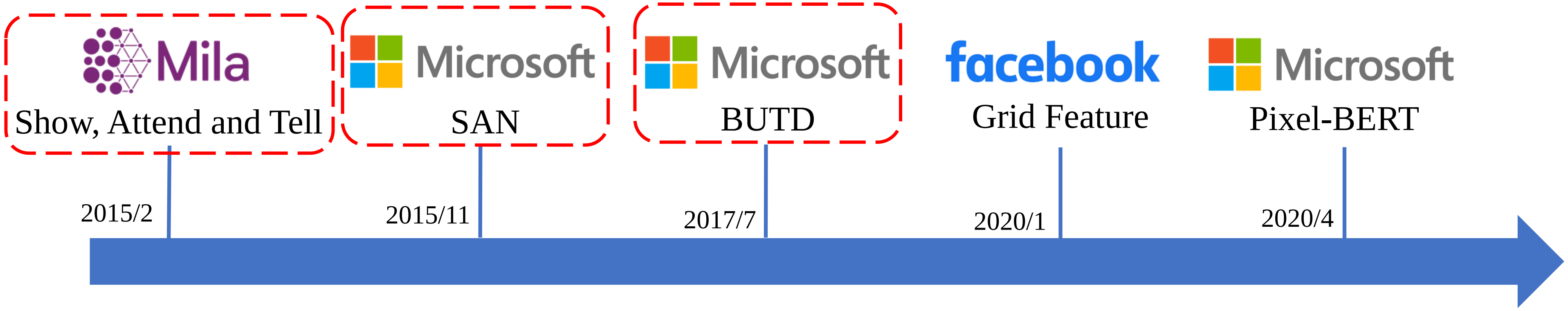
Overview

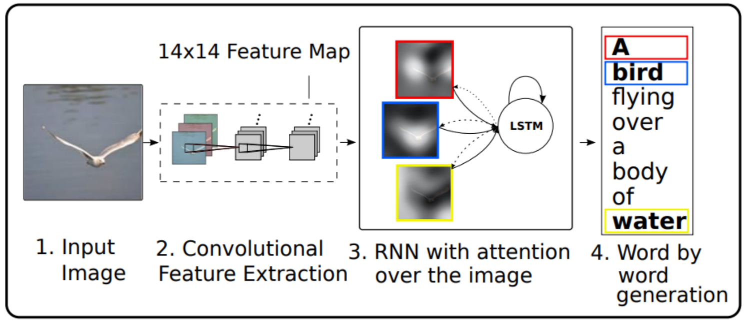
- Better image feature preparation
- Enhanced multimodal fusion
  - Bilinear pooling: how to fuse two vectors into one
  - Multimodal alignment: *cross-modal* attention
  - Incorporation of object relations: *intra-modal* self-attention, graph attention
  - Multi-step reasoning
- Neural module networks for compositional reasoning
- Robust VQA (briefly mention)
- Multimodal pre-training (briefly mention)

Better Image Feature Preparation

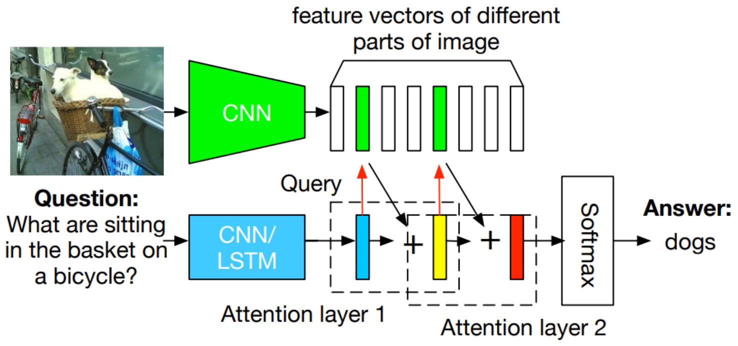
- From *grid* features to *region* features, and to *grid* features again



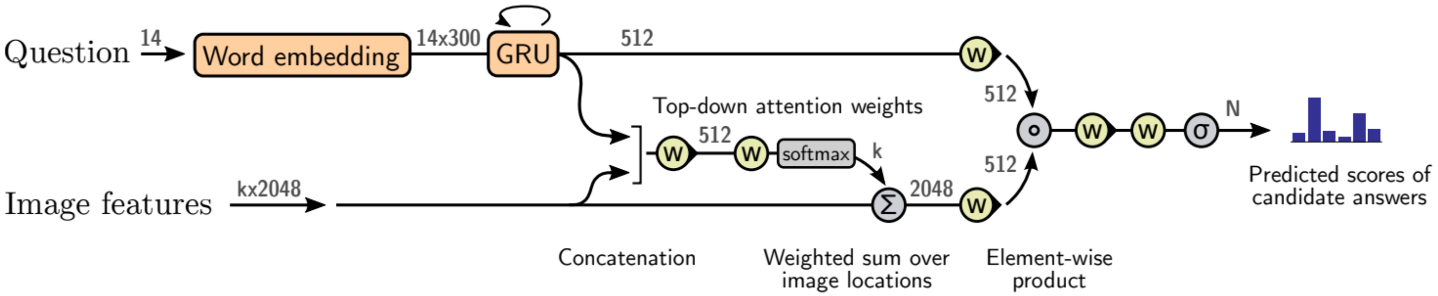




Show, Attend and Tell



Stacked Attention Network



2017 VQA Challenge Winner

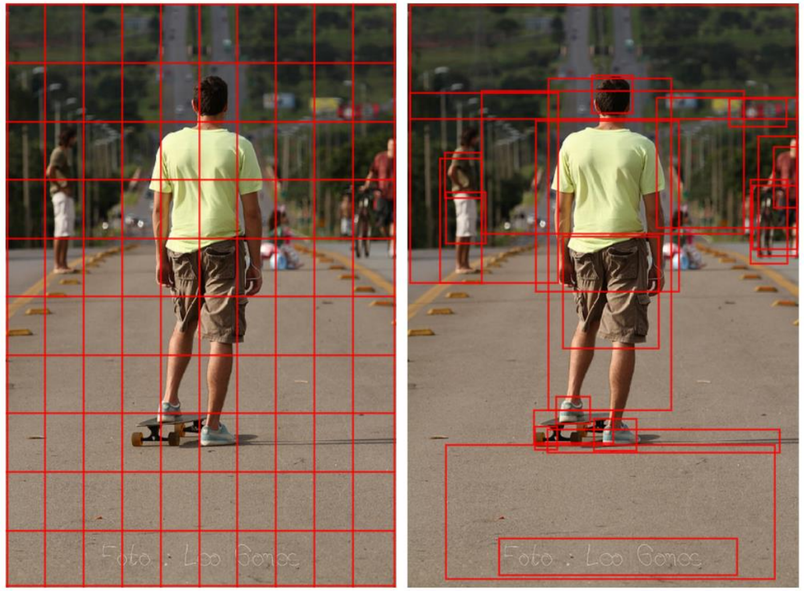


Figure 1. Typically, attention models operate on CNN features corresponding to a uniform grid of equally-sized image regions (left). Our approach enables attention to be calculated at the level of objects and other salient image regions (right).

[1] Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, ICML 2015
[2] Stacked Attention Networks for Image Question Answering, CVPR 2016
[3] Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering, CVPR 2018

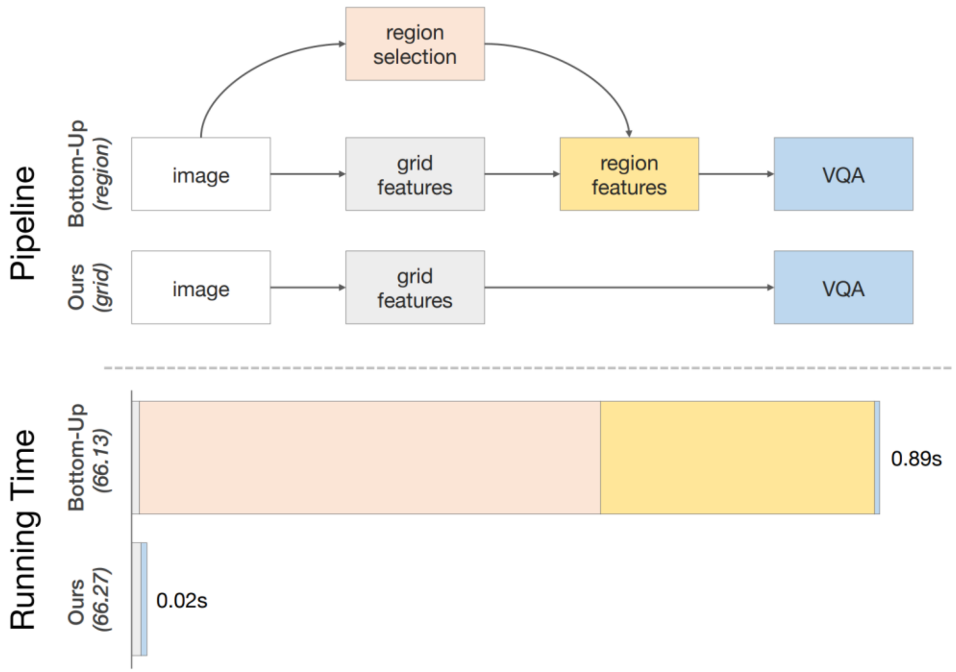
Mila
Show, Attend and Tell
2015/2

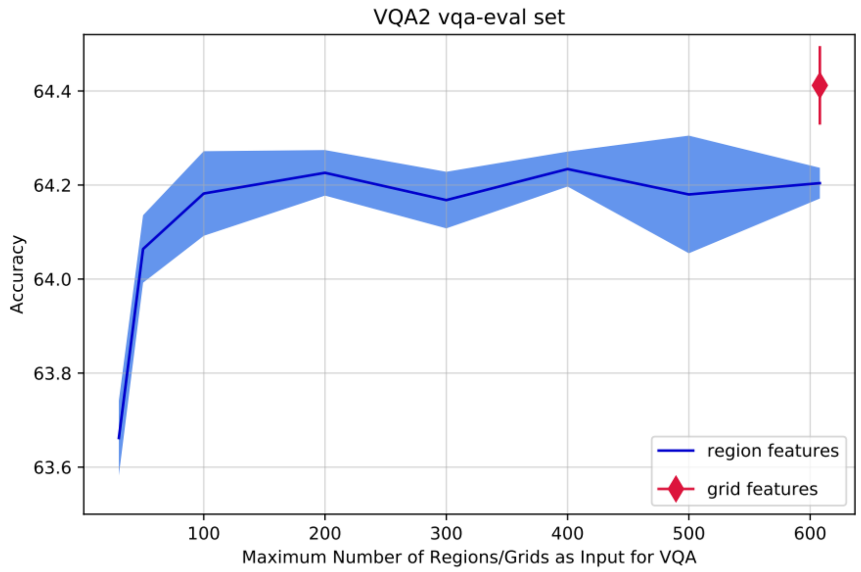
Microsoft
SAN
2015/11

Microsoft
BUTD
2017/7

facebook
Grid Feature
2020/1

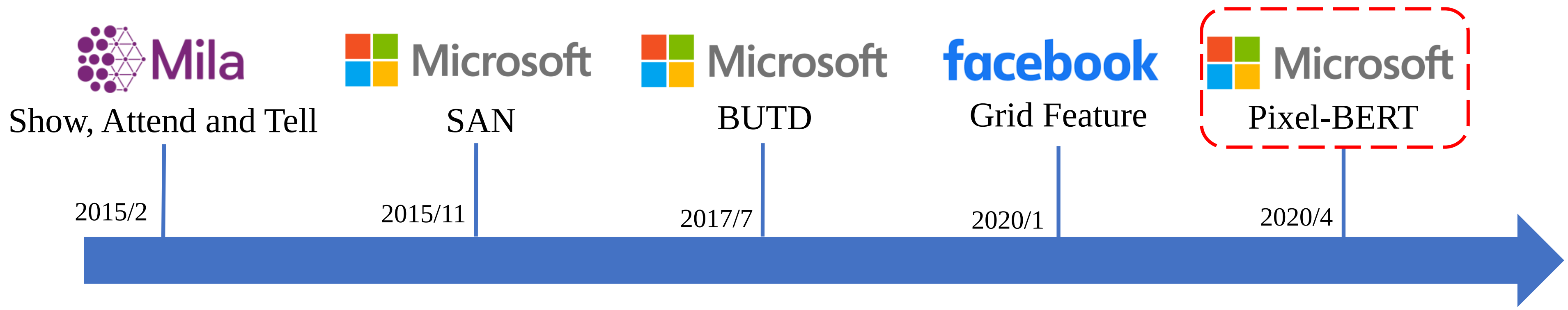
Microsoft
Pixel-BERT
2020/4

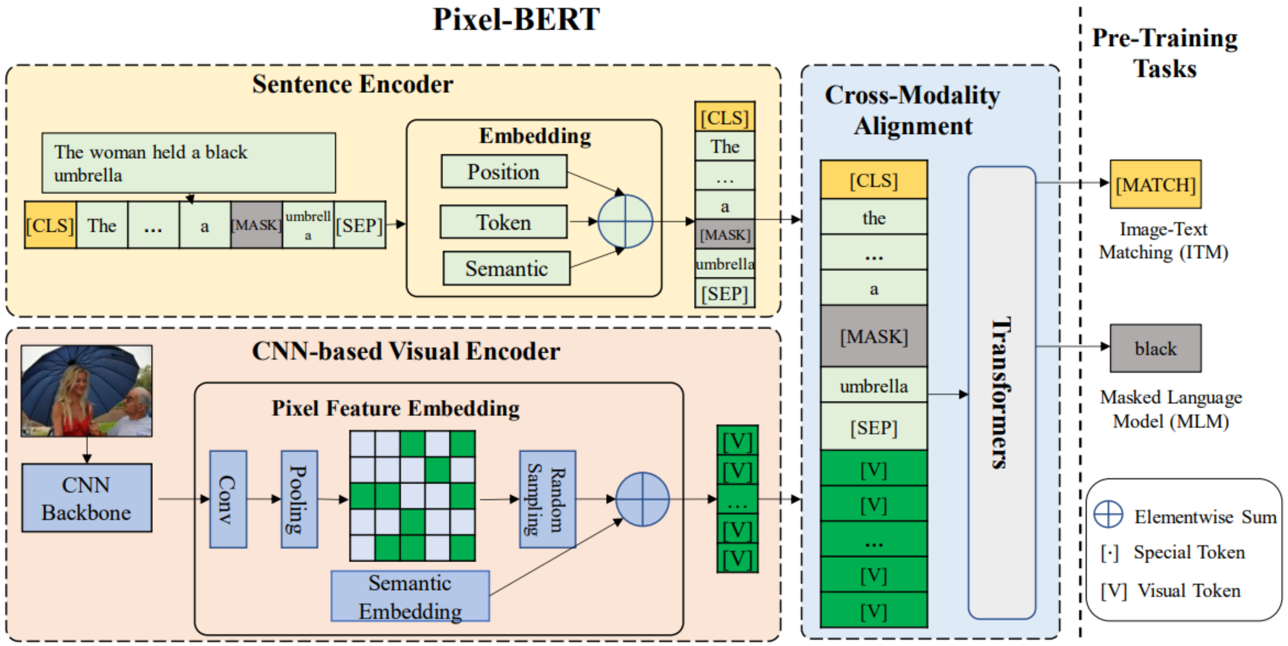




In Defense of Grid Features for VQA

[1] In Defense of Grid Features for Visual Question Answering, CVPR 2020





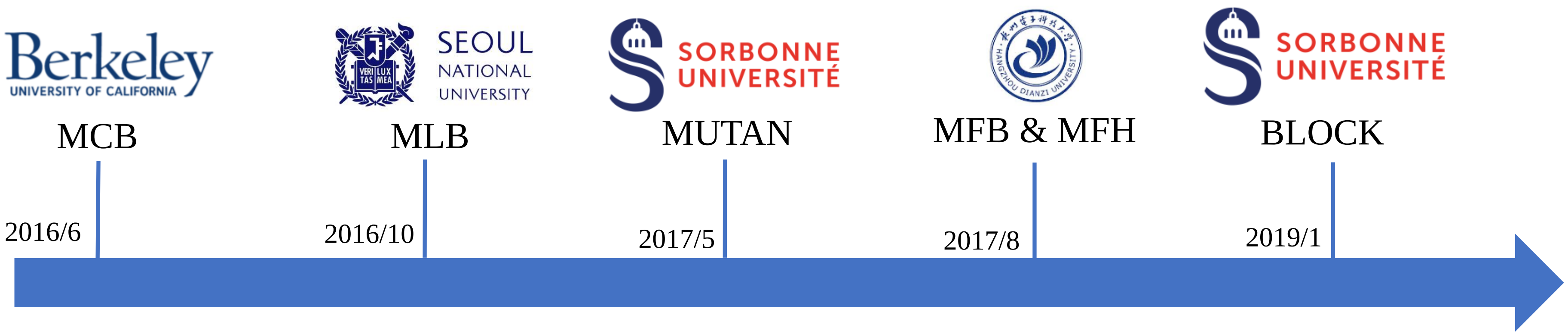
Model	test-dev	test-std
MUTAN[5]	60.17	-
BUTD[2]	65.32	65.67
ViLBERT[21]	70.55	70.92
VisualBERT[19]	70.80	71.00
VLBERT[29]	71.79	72.22
LXMERT[33]	72.42	72.54
UNITER[6]	72.27	72.46
Pixel-BERT (r50)	71.35	71.42
Pixel-BERT (x152)	**74.45**	**74.55**

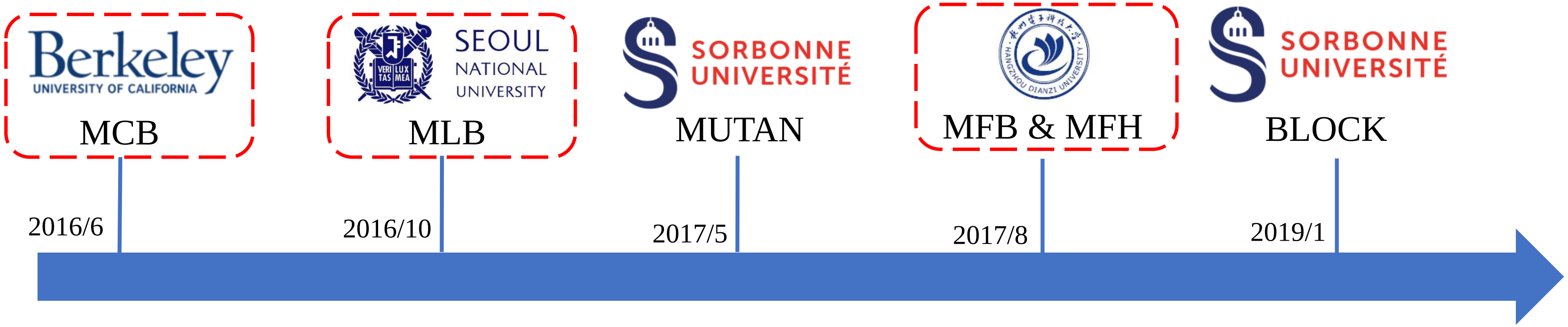
Table 2. Evaluation of Pixel-BERT with other methods on VQA.

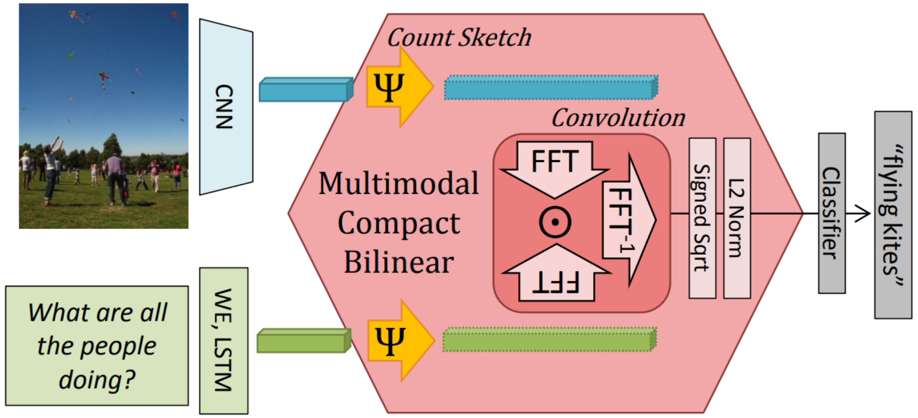
[1] Pixel-BERT: Aligning Image Pixels with Text by Deep Multi-Modal Transformers, 2020

Bilinear Pooling

- Instead of simple concatenation and element-wise product for fusion, bilinear pooling methods have been studied
- Bilinear pooling and attention mechanism can be enhanced with each other







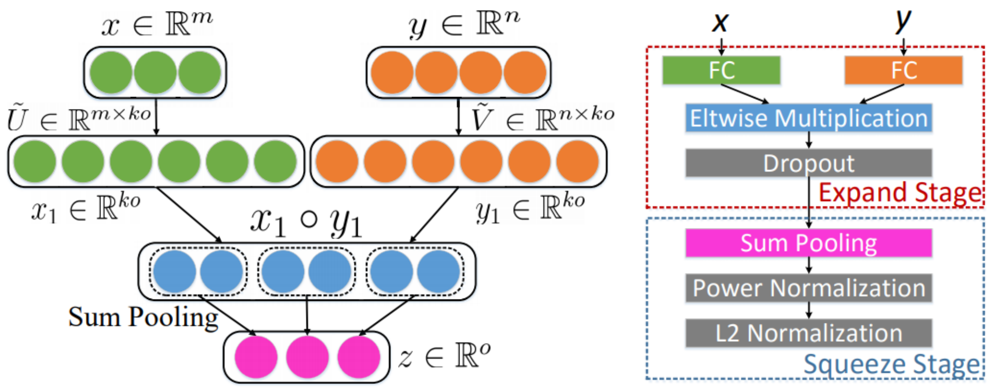
Multimodal Compact Bilinear Pooling

2016 VQA Challenge Winner

However, the feature after FFT is very high dimensional.

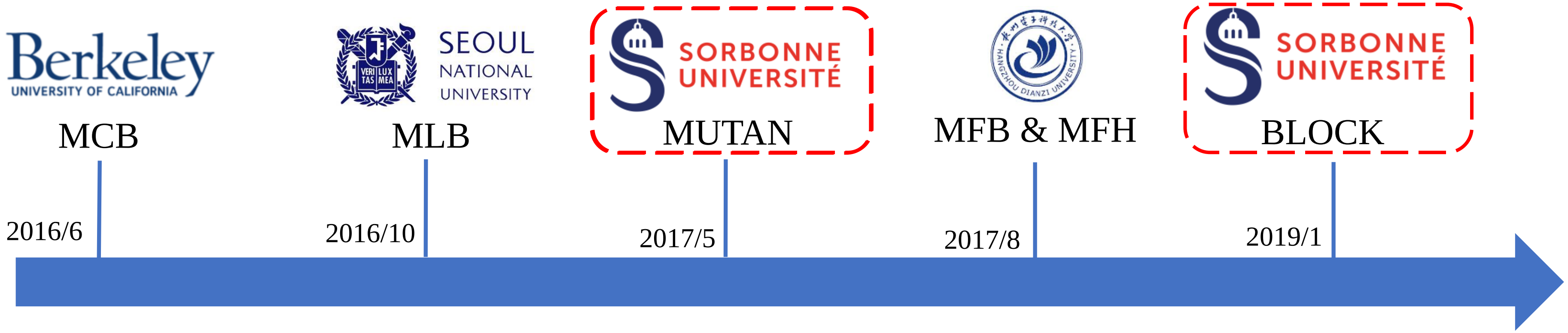
$$\mathbf{f} = \mathbf{P}^T(\mathbf{U}^T\mathbf{x} \circ \mathbf{V}^T\mathbf{y}) + \mathbf{b}$$

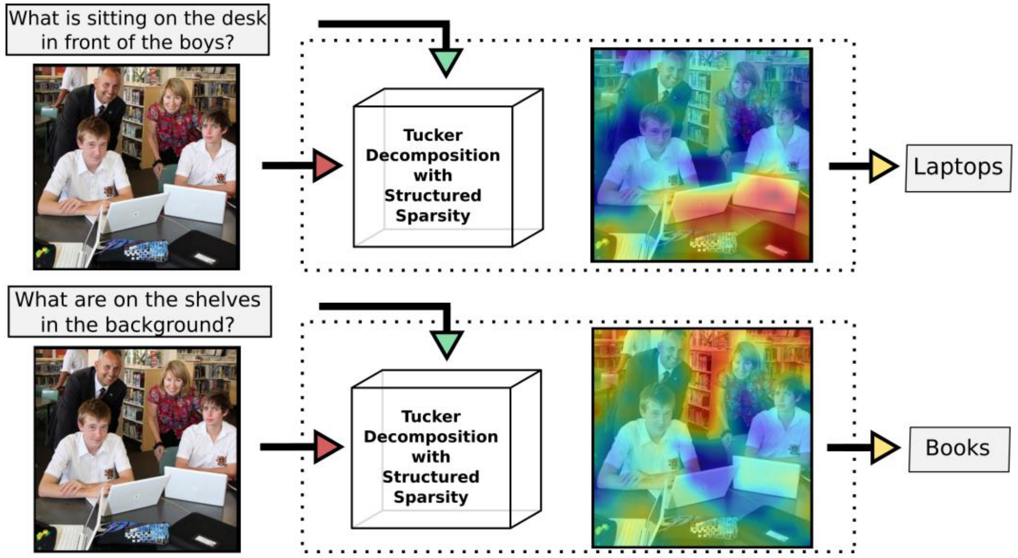
Multimodal Low-rank Bilinear Pooling



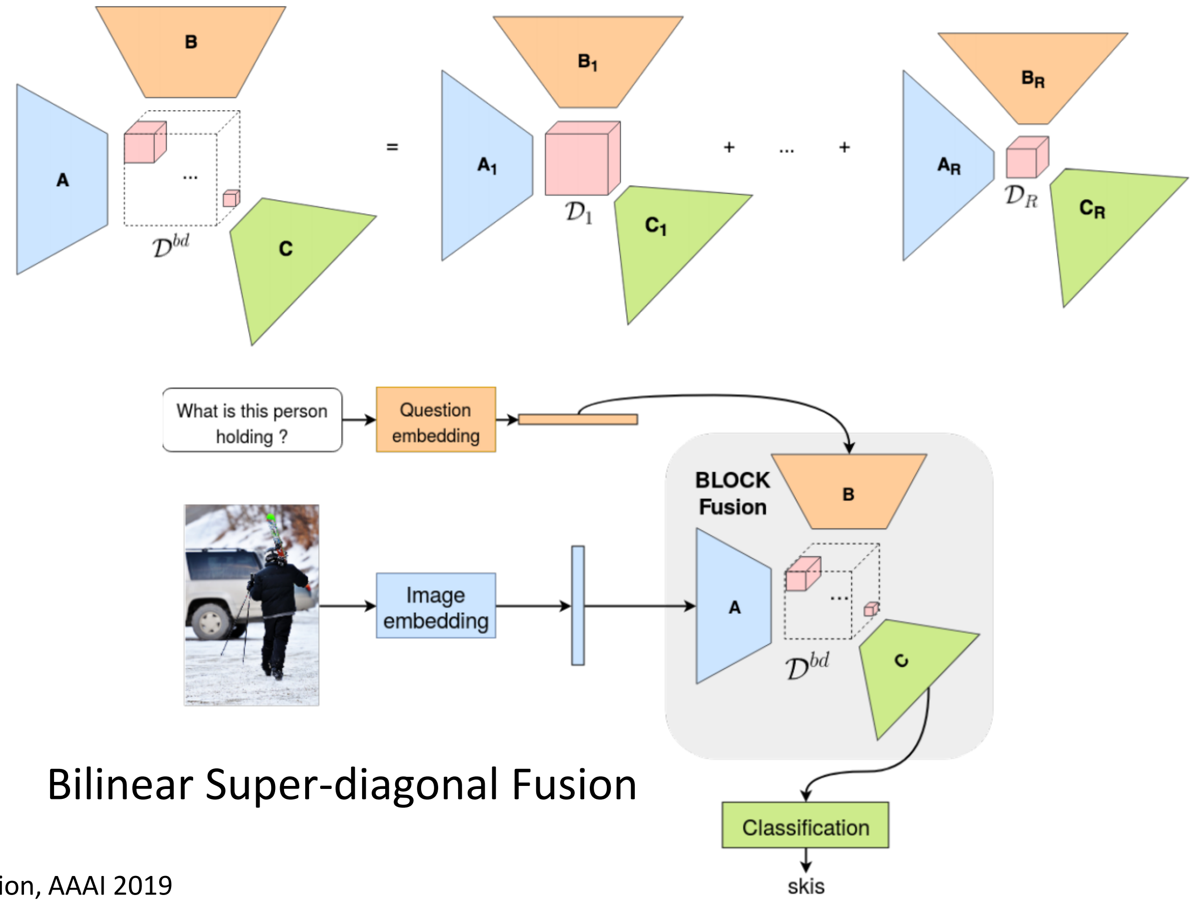
(a) Multi-modal Factorized Bilinear Pooling (b) MFB module

[1] Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding, EMNLP 2016
[2] Hadamard Product for Low-rank Bilinear Pooling, ICLR 2017
[3] Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering, ICCV 2017





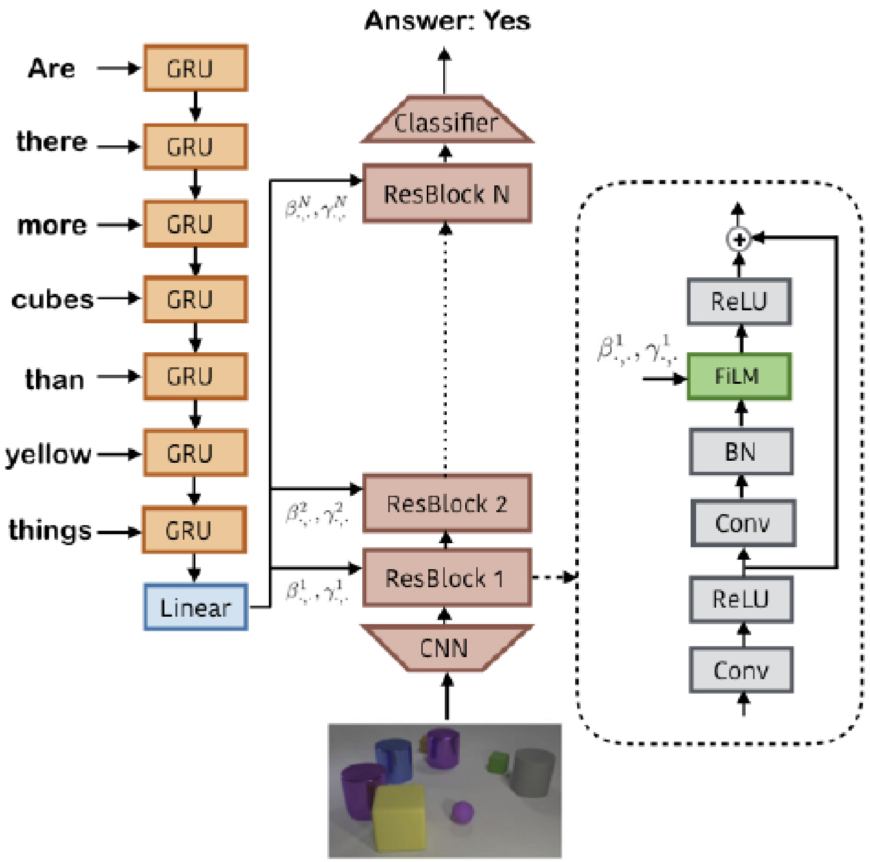
Multimodal Tucker Fusion



Bilinear Super-diagonal Fusion

[1] MUTAN: Multimodal Tucker Fusion for Visual Question Answering, ICCV 2017
[2] BLOCK: Bilinear Superdiagonal Fusion for Visual Question Answering and Visual Relationship Detection, AAAI 2019

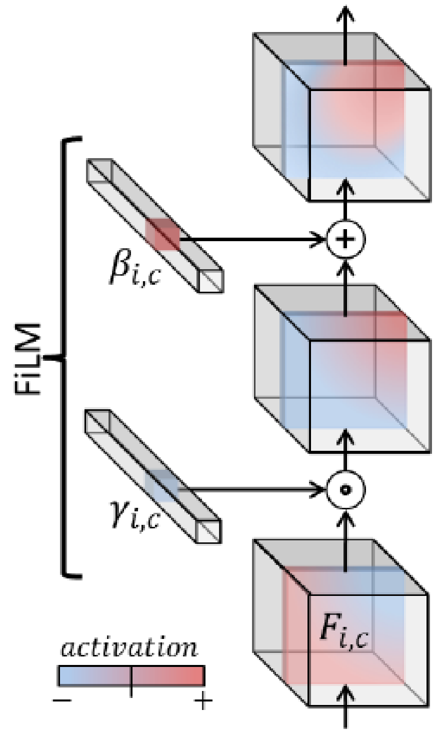
FiLM: Feature-wise Linear Modulation



$$\gamma_{i,c} = f_c(\boldsymbol{x}_i) \qquad \beta_{i,c} = h_c(\boldsymbol{x}_i),$$

$$FiLM(\boldsymbol{F}_{i,c}|\gamma_{i,c}, \beta_{i,c}) = \gamma_{i,c}\boldsymbol{F}_{i,c} + \beta_{i,c}.$$

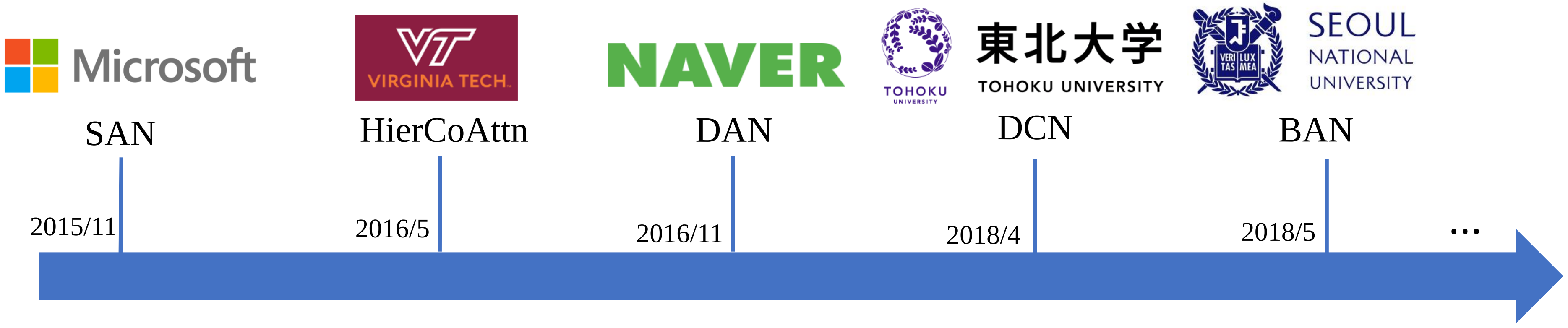
Something similar to conditional batch normalization

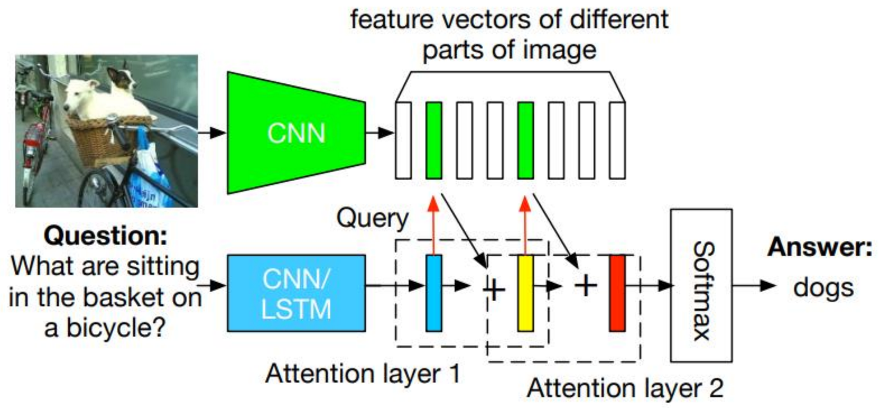


[1] FiLM: Visual Reasoning with a General Conditioning Layer, AAAI, 2018

Multimodal Alignment

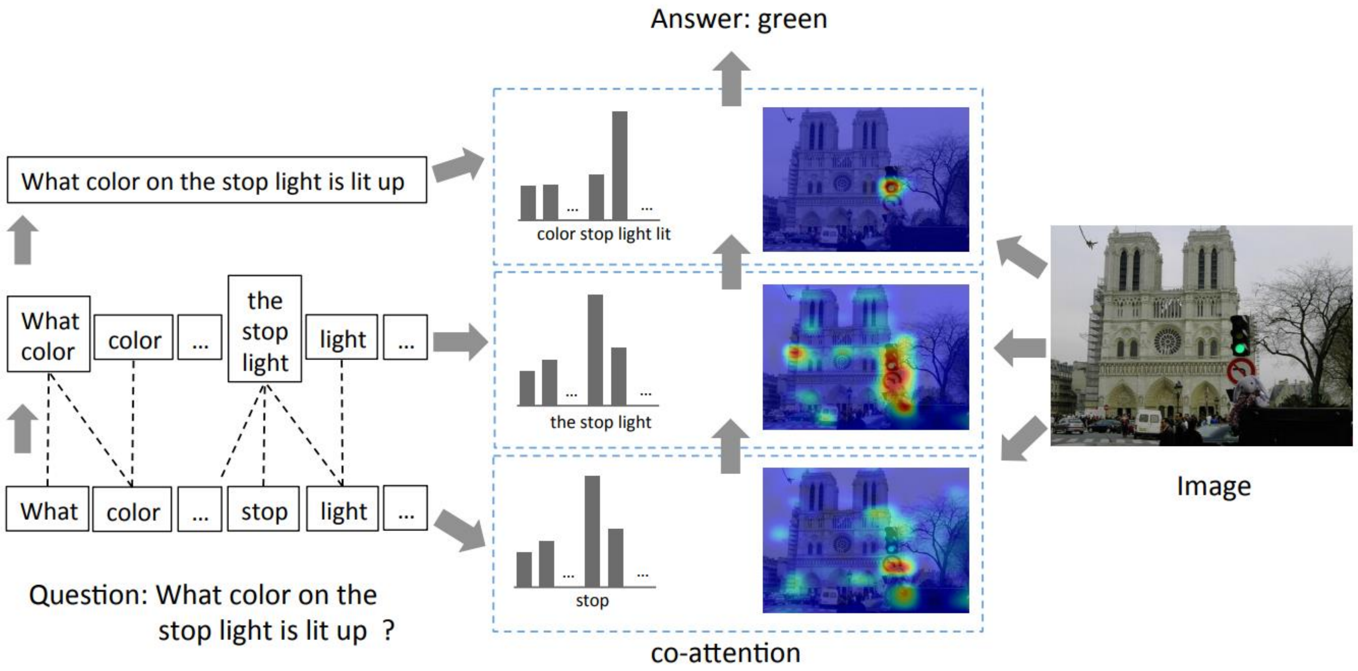
- Cross-modal attention:
  - Tons of work in this area
  - Early work: questions attend to image grids/regions
  - Current focus: image-text co-attention





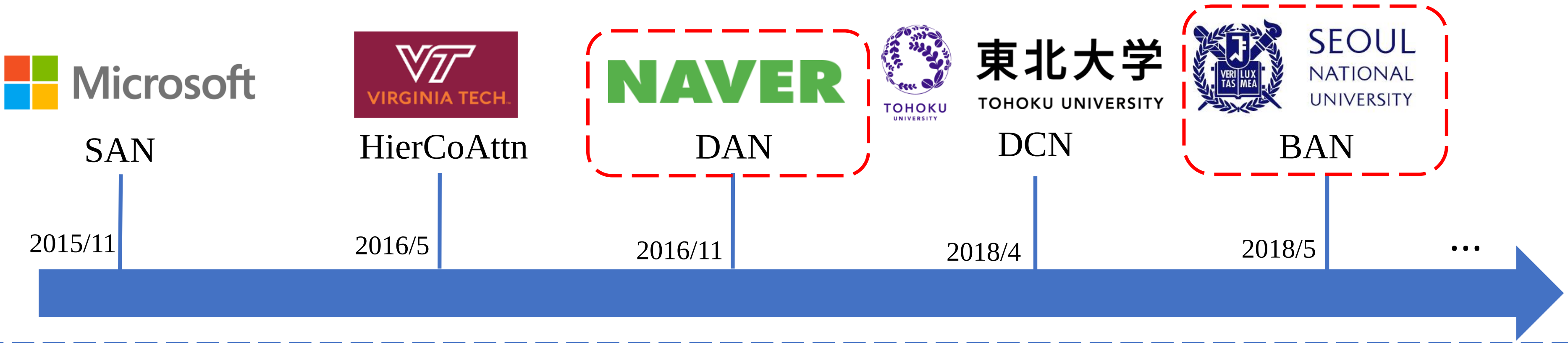
(a) Stacked Attention Network for Image QA

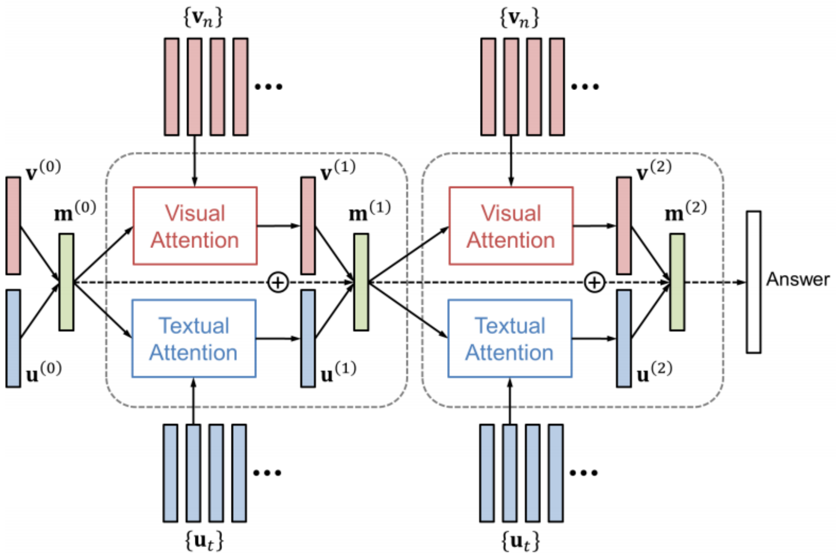
(b) Visualization of the learned multiple attention layers.

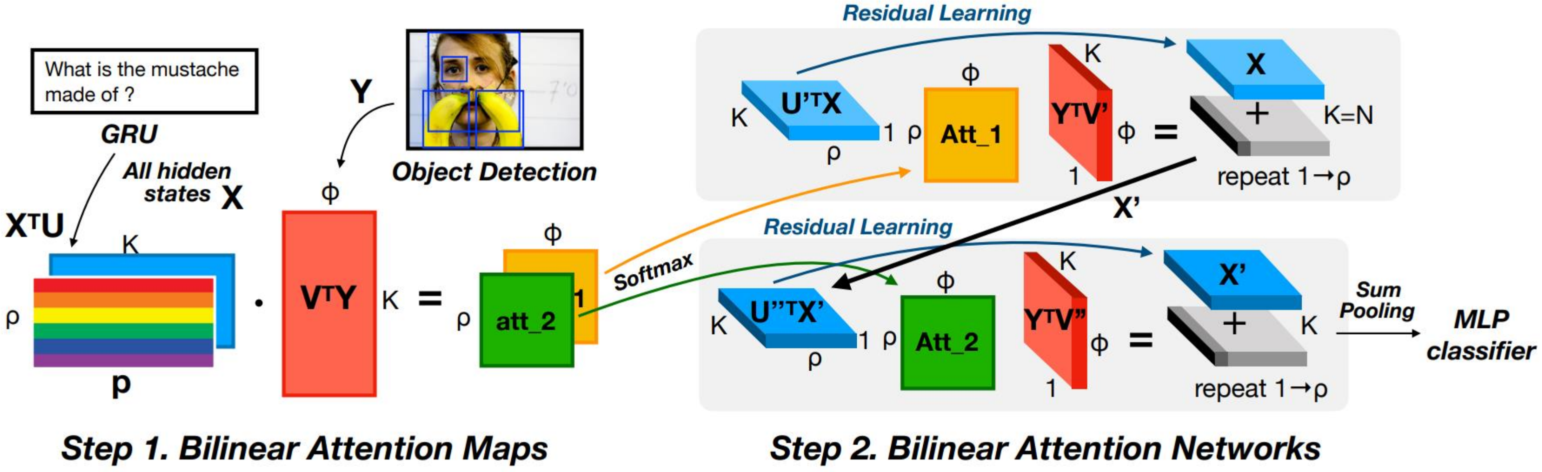


Parallel Co-attention and Alternative Co-attention

[1] Stacked Attention Networks for Image Question Answering, CVPR 2016
[2] Hierarchical Question-Image Co-Attention for Visual Question Answering, NeurIPS 2016







DAN: Dual Attention Network

DCN: Dense Co-attention Network

2018 VQA Challenge Runner-Up

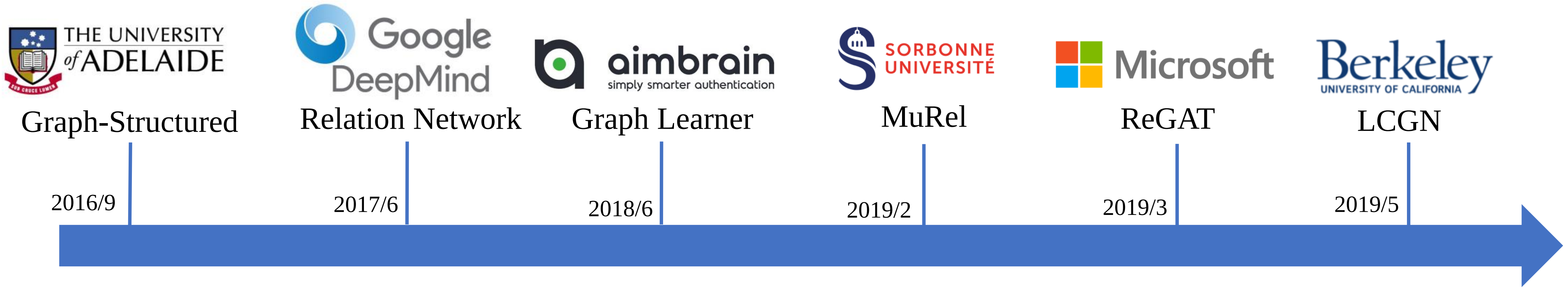
- Multiple Glimpses
- Counter Module
- Residual Learning
- Glove Embeddings

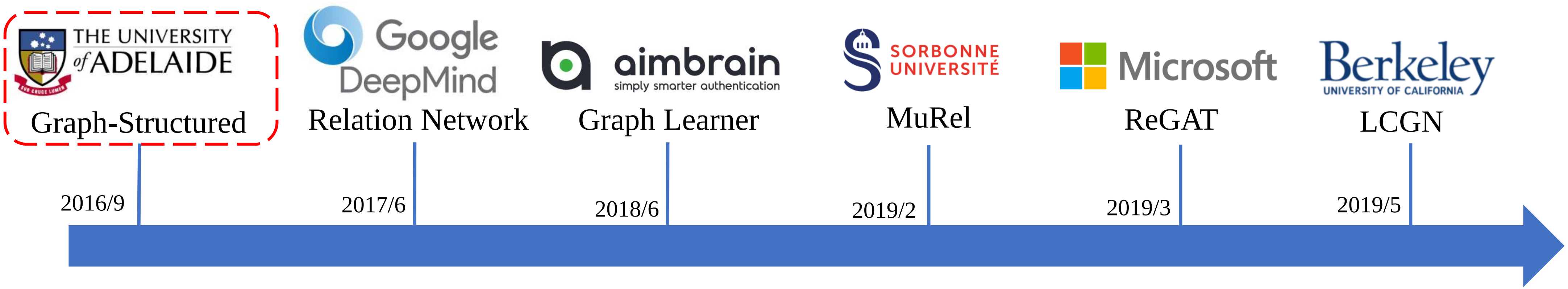
[1] Stacked Attention Networks for Image Question Answering, CVPR 2016

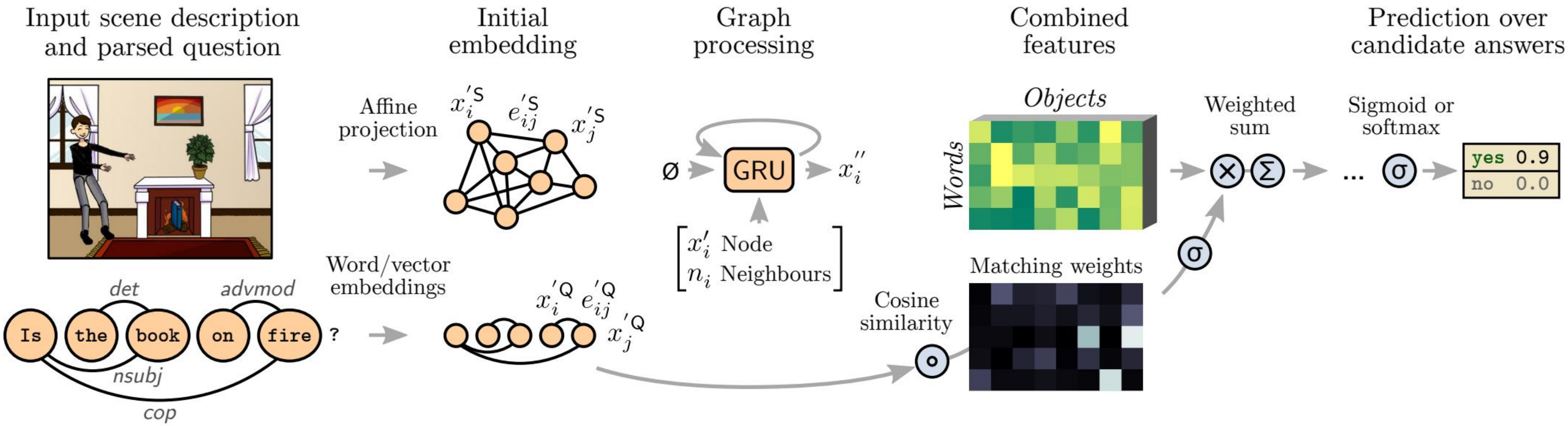
[2] Improved Fusion of Visual and Language Representations by Dense Symmetric Co-Attention for Visual Question Answering, CVPR 2018

Relational Reasoning

- Intra-modal attention
  - Recently becoming popular
  - Representing image as a graph
  - Graph Convolutional Network & Graph Attention Network
  - Self-attention used in Transformer



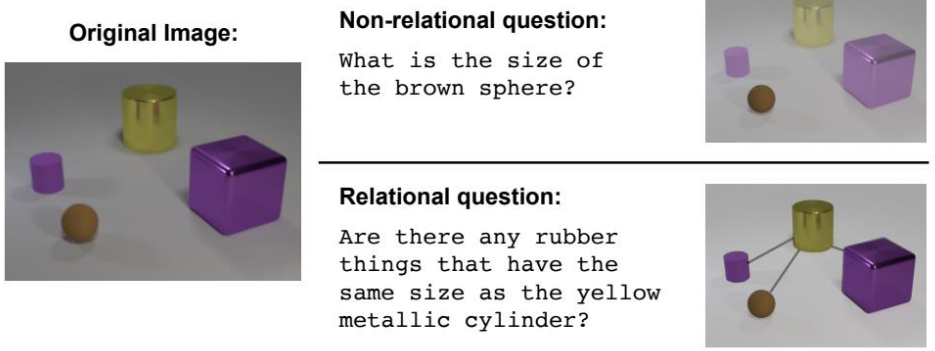




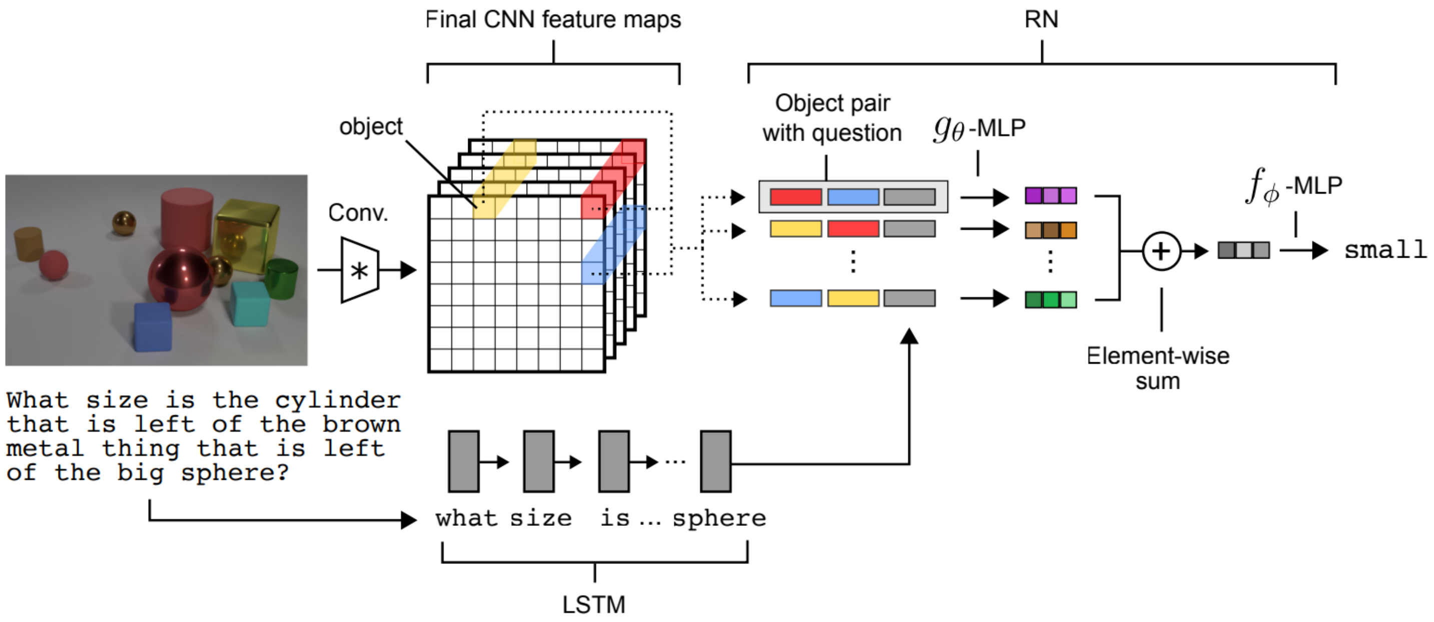
Graph-Structured Representations for Visual Question Answering

[1] Graph-Structured Representations for Visual Question Answering, CVPR 2017

- 2016/9 — Graph-Structured (The University of Adelaide)
- 2017/6 — Relation Network (Google DeepMind)
- 2018/6 — Graph Learner (aimbrain)
- 2019/2 — MuRel (Sorbonne Université)
- 2019/3 — ReGAT (Microsoft)
- 2019/5 — LCGN (Berkeley, University of California)



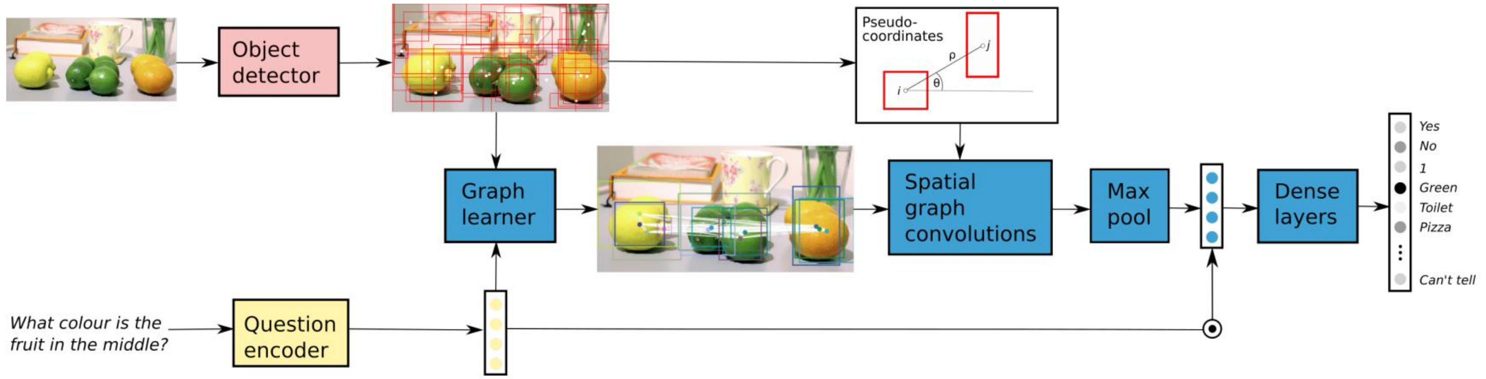
$$\mathrm{RN}(O) = f_\phi\left(\sum_{i,j} g_\theta(o_i, o_j)\right)$$



Relational Network: A fully-connected graph is constructed

[1] A simple neural network module for relational reasoning, NeurIPS 2017

Graph-Structured	Relation Network	Graph Learner	MuRel	ReGAT	LCGN
THE UNIVERSITY of ADELAIDE	Google DeepMind	aimbrain simply smarter authentication	SORBONNE UNIVERSITÉ	Microsoft	Berkeley UNIVERSITY OF CALIFORNIA
2016/9	2017/6	2018/6	2019/2	2019/3	2019/5

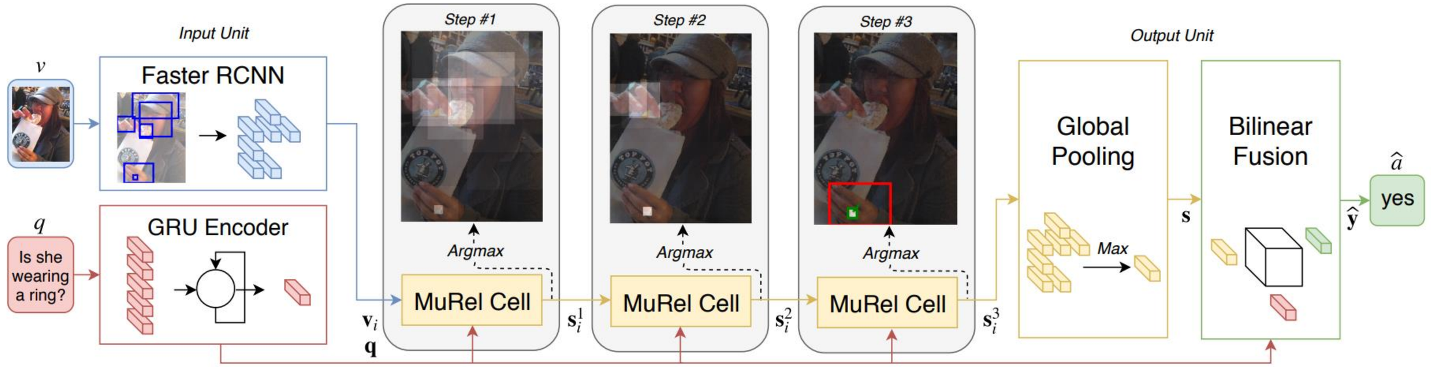


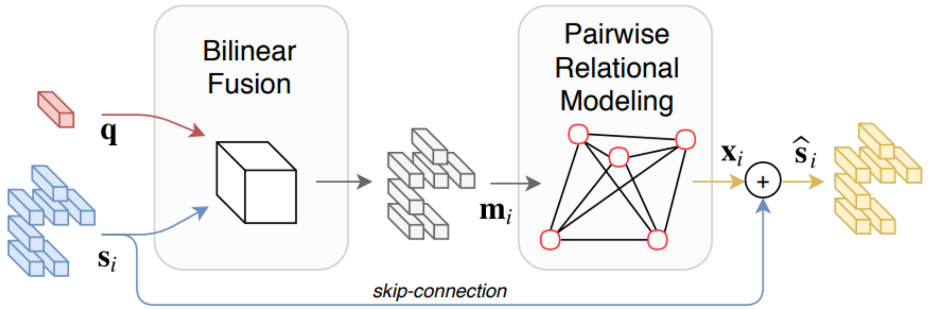
$$\mathbf{e}_n = F([\mathbf{v}_n \| \mathbf{q}]), \qquad n = 1, 2, ..., N$$

$$\mathbf{E} \in \mathbb{R}^{N \times d_e}$$

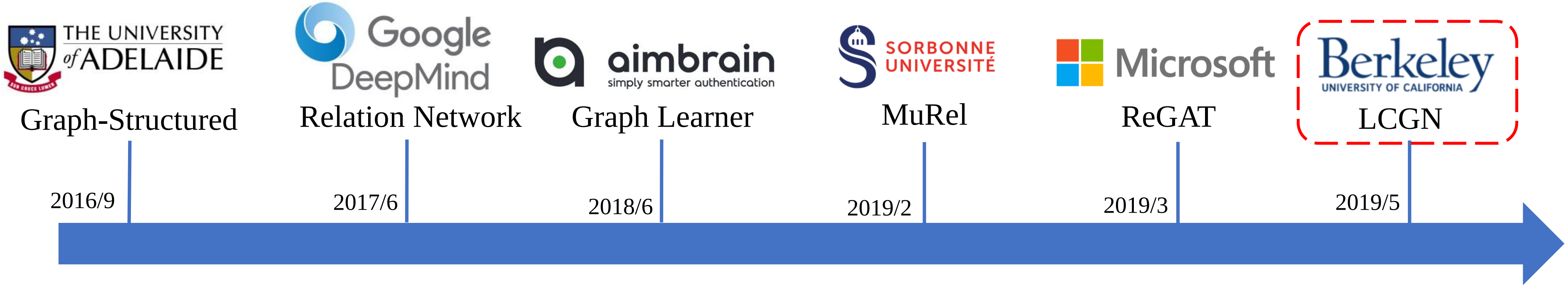
$$\mathbf{A} = \mathbf{E}\mathbf{E}^T \text{ so that } A_{i,j} = \mathbf{e}_i^T \mathbf{e}_j.$$

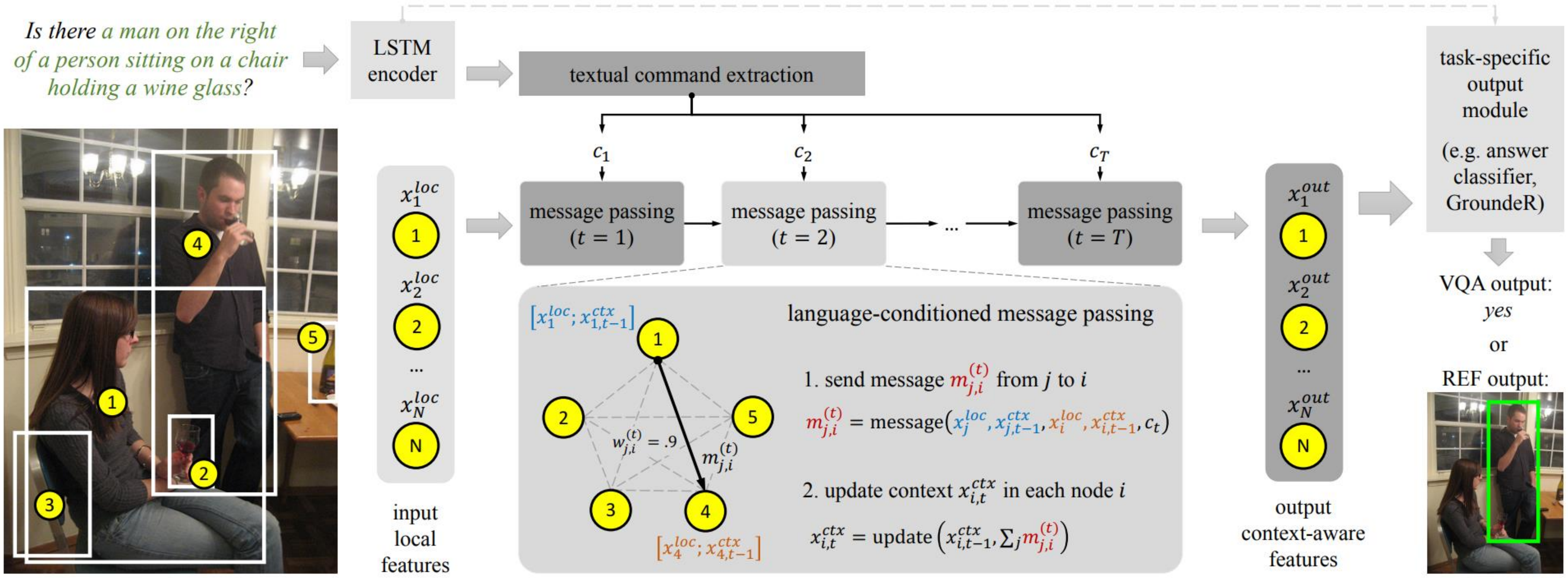
$$\mathcal{N}(i) = topm(\mathbf{a}_i)$$





[1] Learning Conditioned Graph Structures for Interpretable Visual Question Answering, NeurIPS 2018
[2] MUREL: Multimodal Relational Reasoning for Visual Question Answering, CVPR 2019





[1] Language-Conditioned Graph Networks for Relational Reasoning, ICCV 2019

Graph-Structured — 2016/9

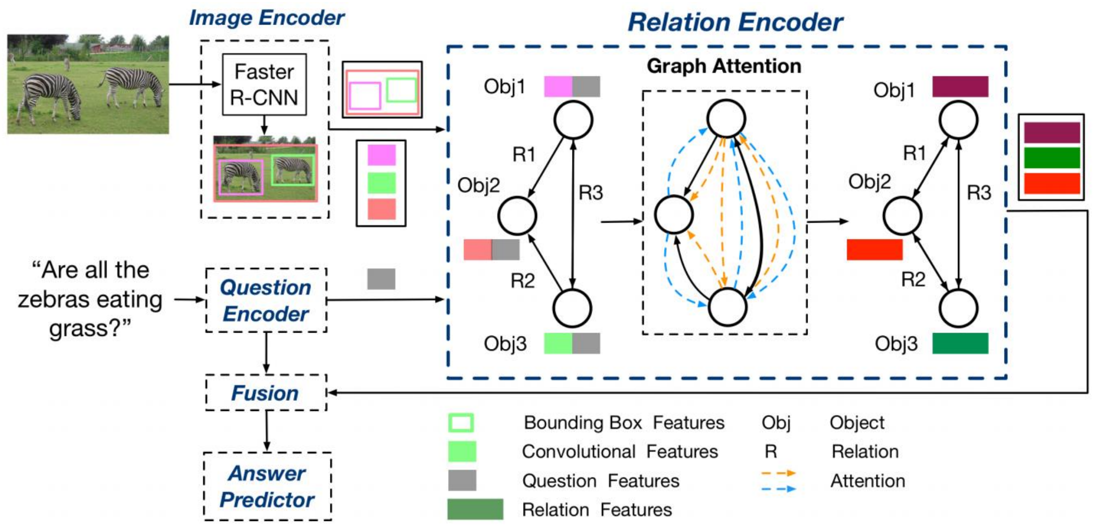
Relation Network — 2017/6

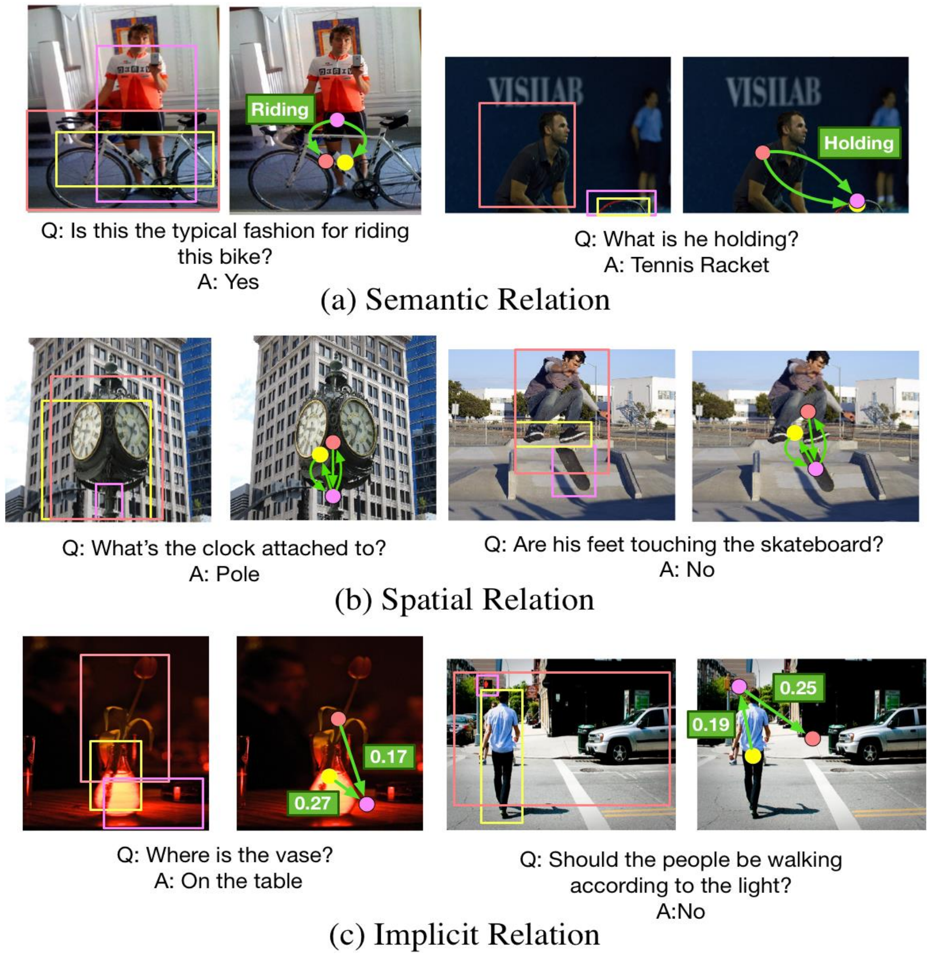
Graph Learner — 2018/6

MuRel — 2019/2

ReGAT — 2019/3

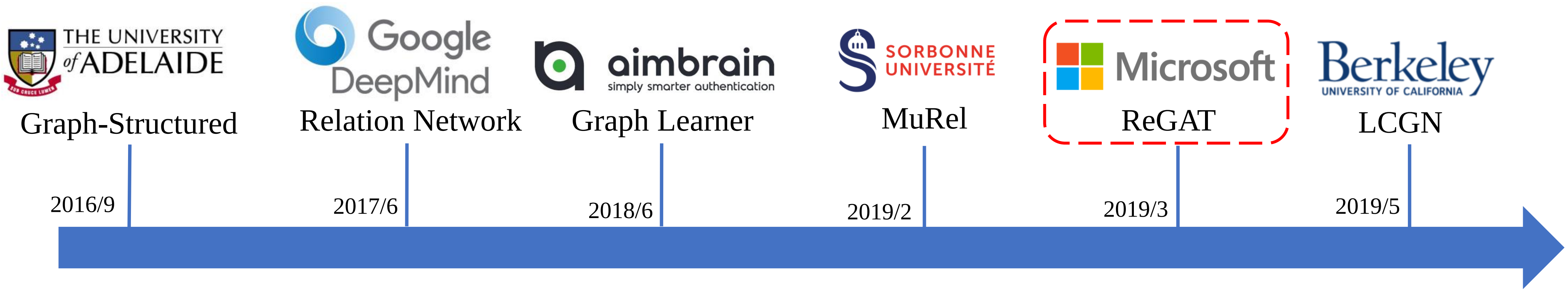
LCGN — 2019/5

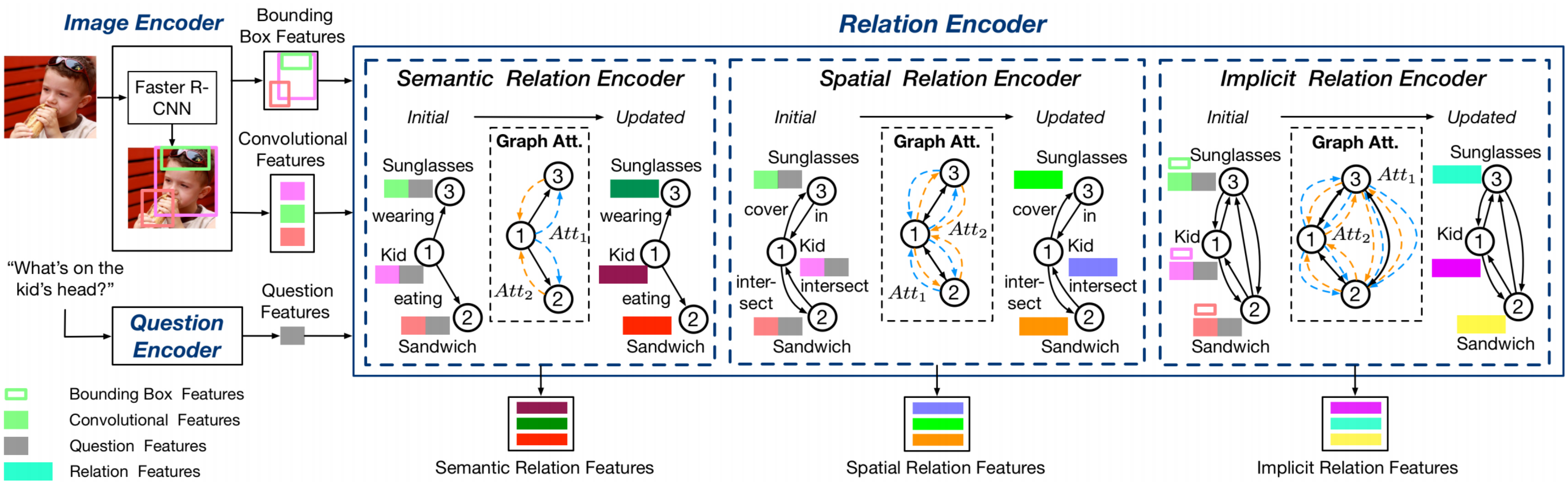




- *Explicit* Relation: Semantic & Spatial relation
- *Implicit* Relation: Learned dynamically during training

[1] Relation-Aware Graph Attention Network for Visual Question Answering, ICCV 2019

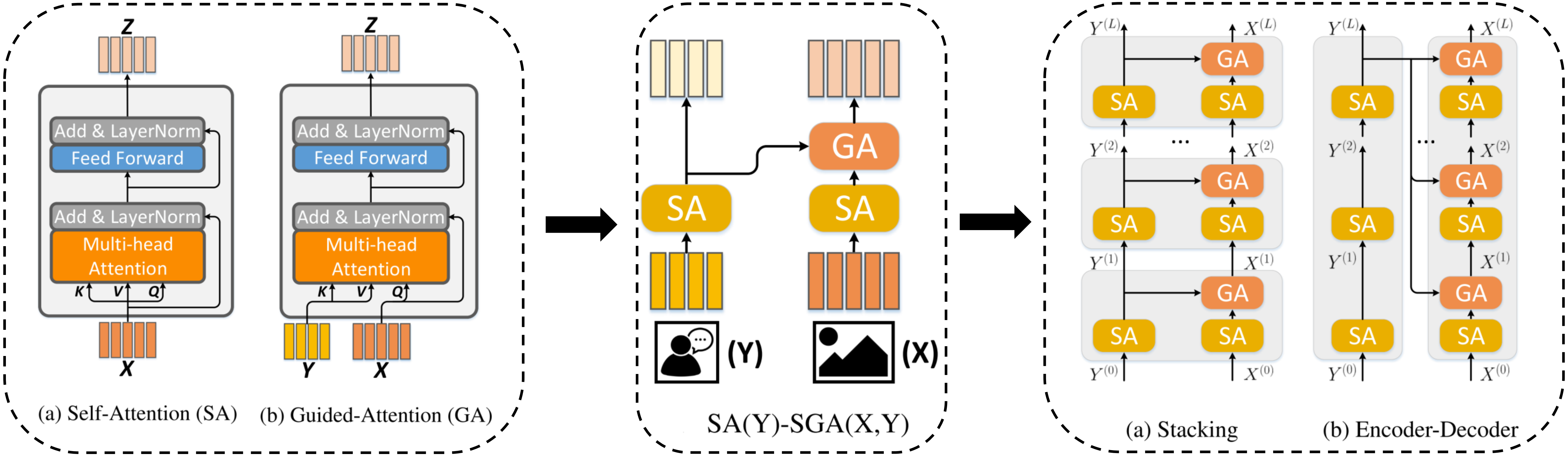




[1] Relation-Aware Graph Attention Network for Visual Question Answering, ICCV 2019

MCAN: Deep Modular Co-Attention Network

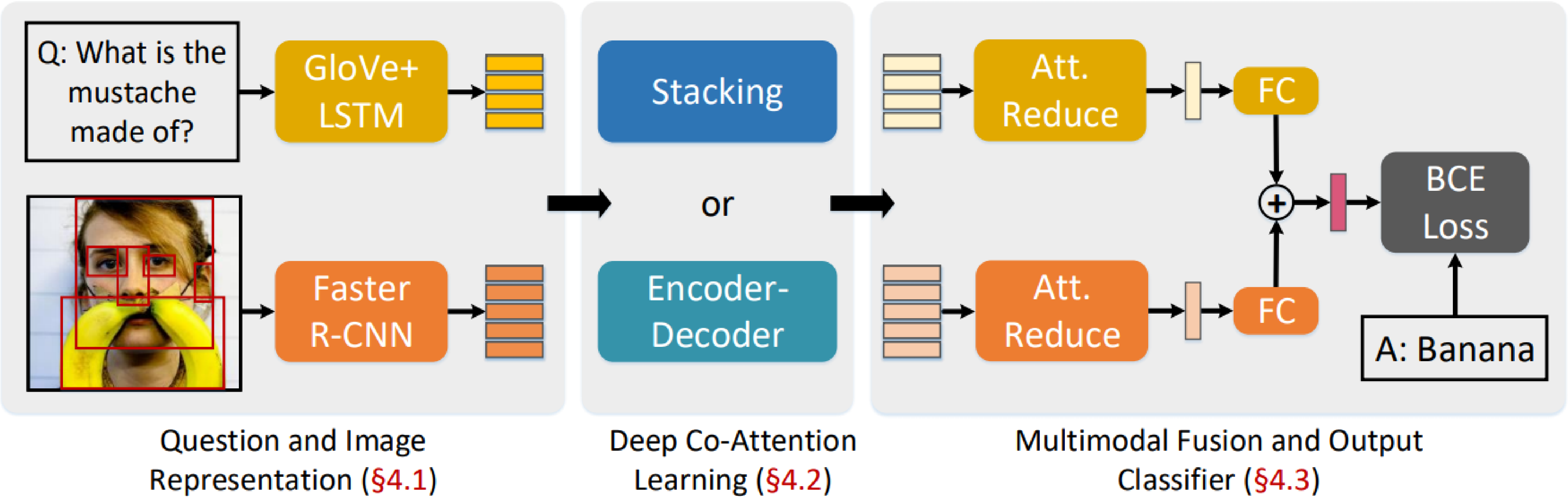
- Winning entry to VQA Challenge 2019
- Similar idea also explored in DFAF, close to *V+L pre-training* models



[1] Deep Modular Co-Attention Networks for Visual Question Answering, CVPR 2019
[2] Dynamic Fusion with Intra- and Inter- Modality Attention Flow for Visual Question Answering, CVPR 2019

MCAN: Deep Modular Co-Attention Network

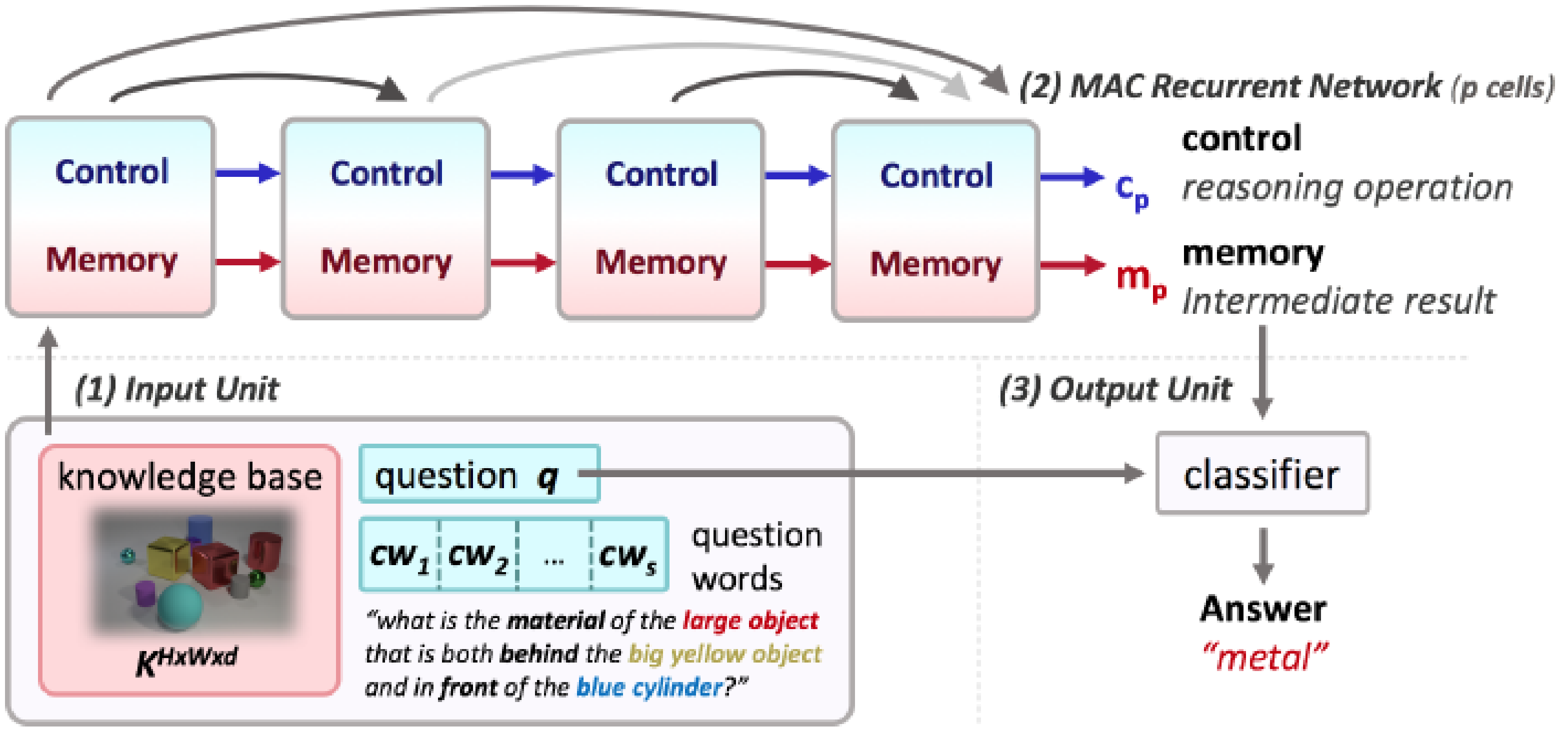
- Winning entry to VQA Challenge 2019
- Similar idea also explored in DFAF, close to *V+L pre-training* models



[1] Deep Modular Co-Attention Networks for Visual Question Answering, CVPR 2019
[2] Dynamic Fusion with Intra- and Inter- Modality Attention Flow for Visual Question Answering, CVPR 2019

MAC: Memory, Attention and Composition

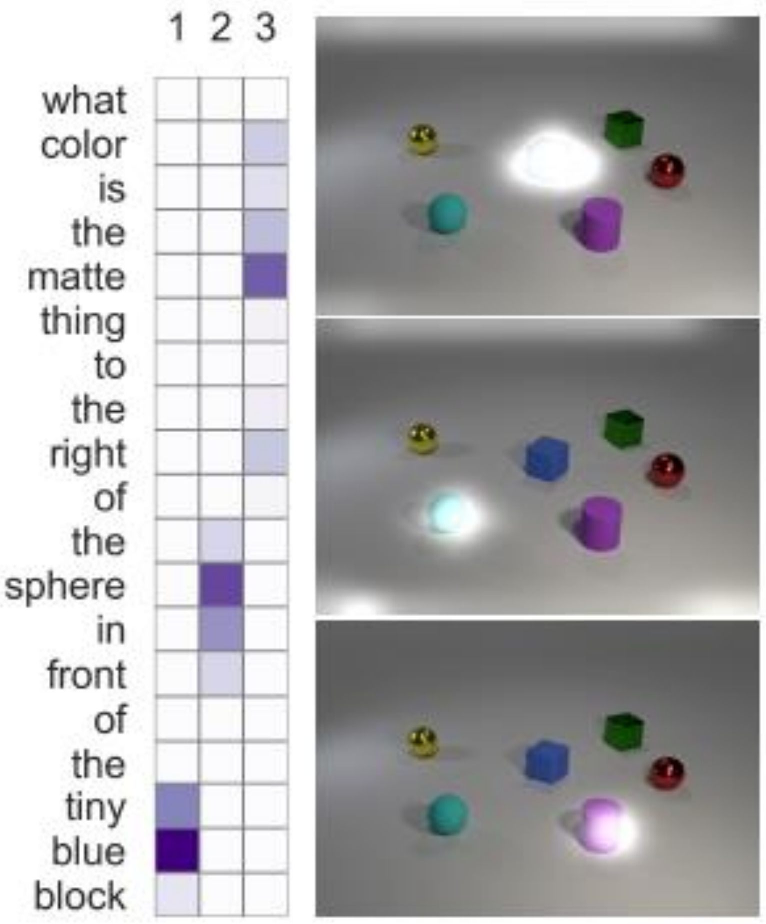
- Multi-step reasoning via recurrent MAC cells, while retaining end-to-end differentiability



[1] Compositional Attention Networks for Machine Reasoning, ICLR, 2018

MAC: Memory, Attention and Composition

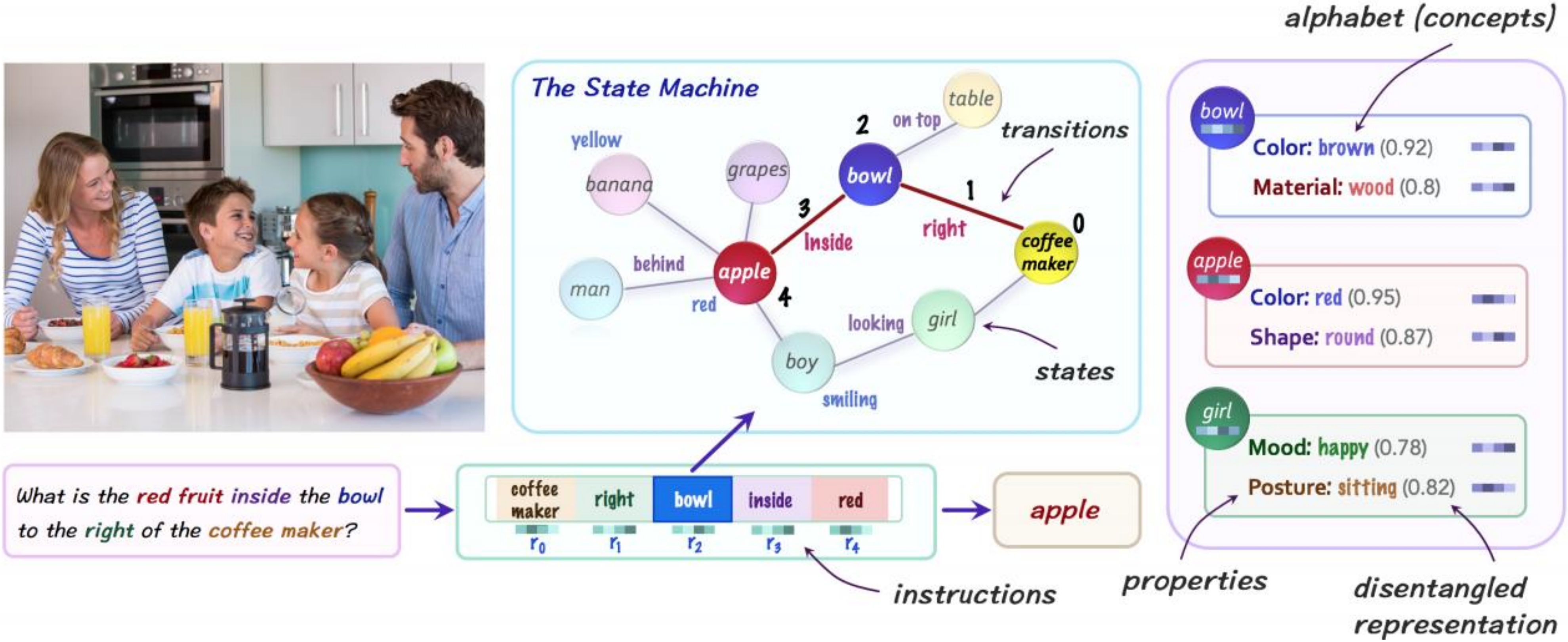
- Each cell maintains recurrent dual states:
  - *Control c_i*: the reasoning operation that should be accomplished at this step.
  - *Memory m_i*: the retrieved information relevant to the query, accumulated over previous iterations.
  - *Implementation-wise*:
    - **Attention-based average** of a given query (question)
    - **Attention-based average** of a given Knowledge Base (image)



[1] Compositional Attention Networks for Machine Reasoning, ICLR, 2018

Neural State Machine

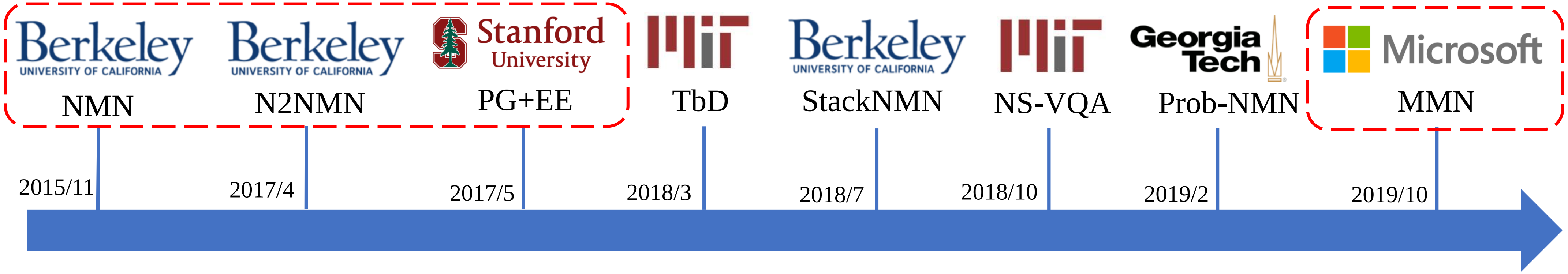
- We see and reason with concepts, not visual details, 99% of the time
- We build semantic world models to represent our environment



[1] Learning by Abstraction: The Neural State Machine, NeurIPS 2019

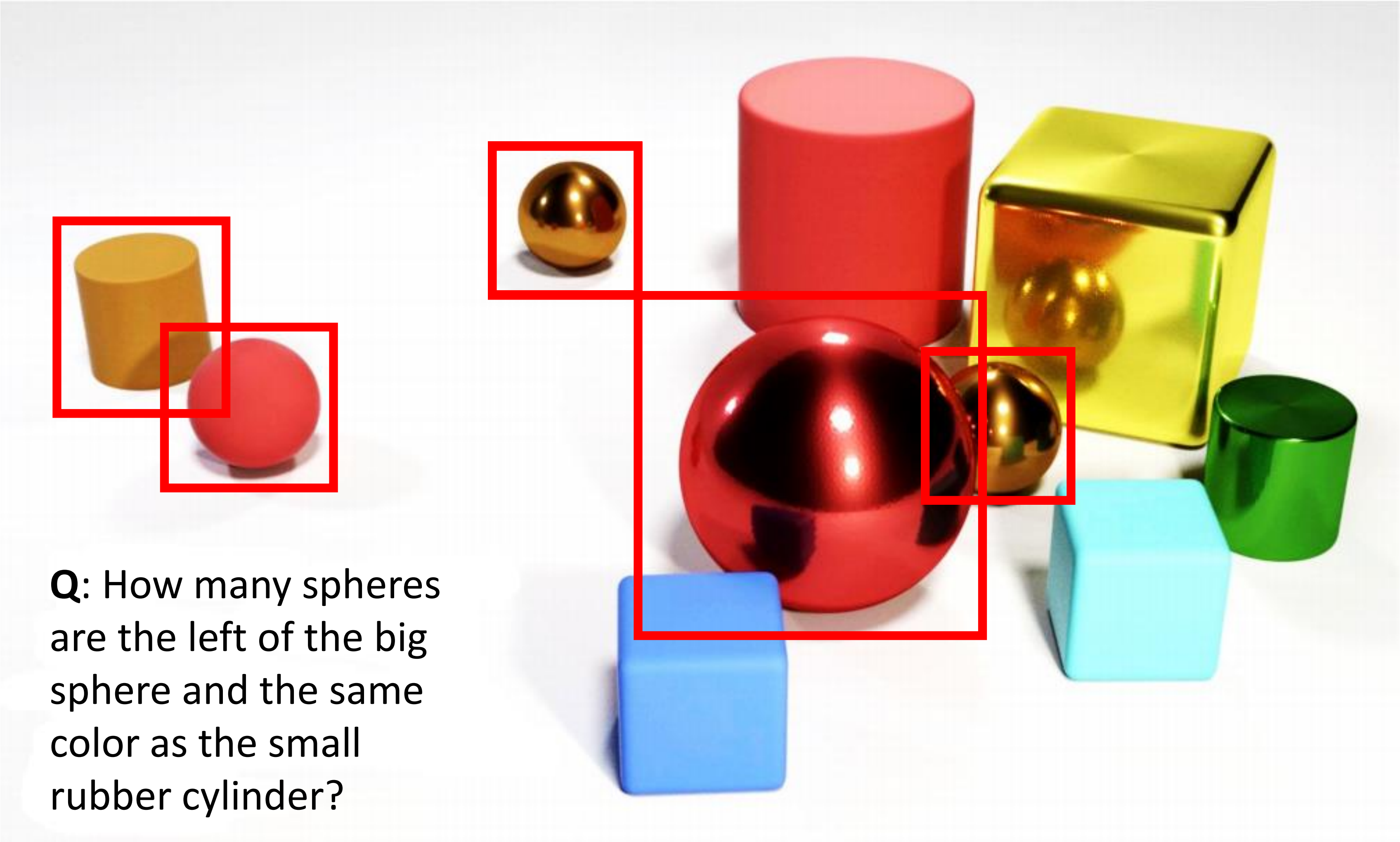
Neural Module Network

- All the previously mentioned work can be considered as *Monolithic Network*
- Design *Neural Modules* for compositional visual reasoning



[1] Deep Compositional Question Answering with Neural Module Networks, CVPR, 2016
[2] Learning to Reason: End-to-End Module Networks for Visual Question Answering, ICCV 2017
[3] Inferring and Executing Programs for Visual Reasoning, ICCV 2017
[4] Transparency by Design: Closing the Gap Between Performance and Interpretability in Visual Reasoning, CVPR 2018
[5] Explainable Neural Computation via Stack Neural Module Networks, ECCV 2018
[6] Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding, NeurIPS 2018
[7] Probabilistic Neural-symbolic Models for Interpretable Visual Question Answering, ICML 2019
[8] Meta Module Network for Compositional Visual Reasoning, 2019

Compositional Visual Reasoning



Q: How many spheres are the left of the big sphere and the same color as the small rubber cylinder?

Identify big sphere

↓

Spheres on left

↓

Rubber cylinder

↓

Sphere of same color

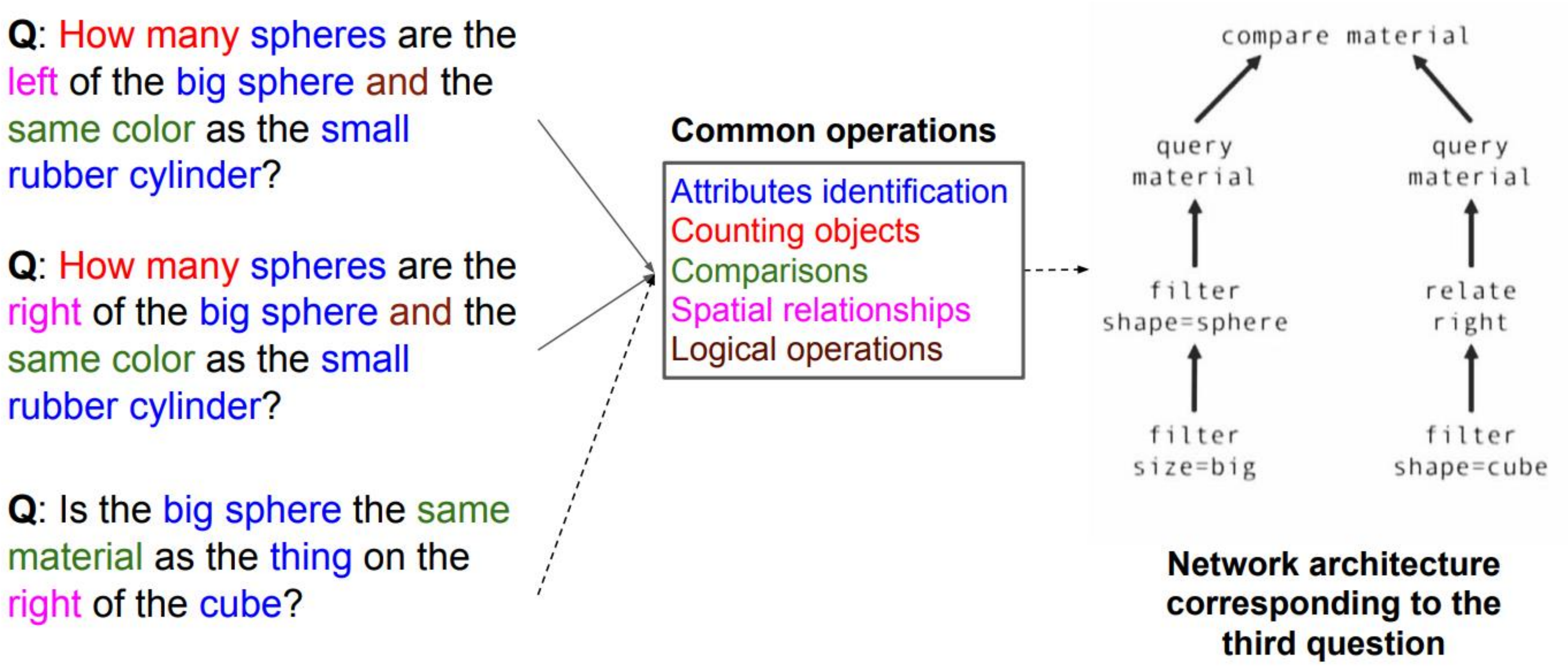
↓

Count

A: 1

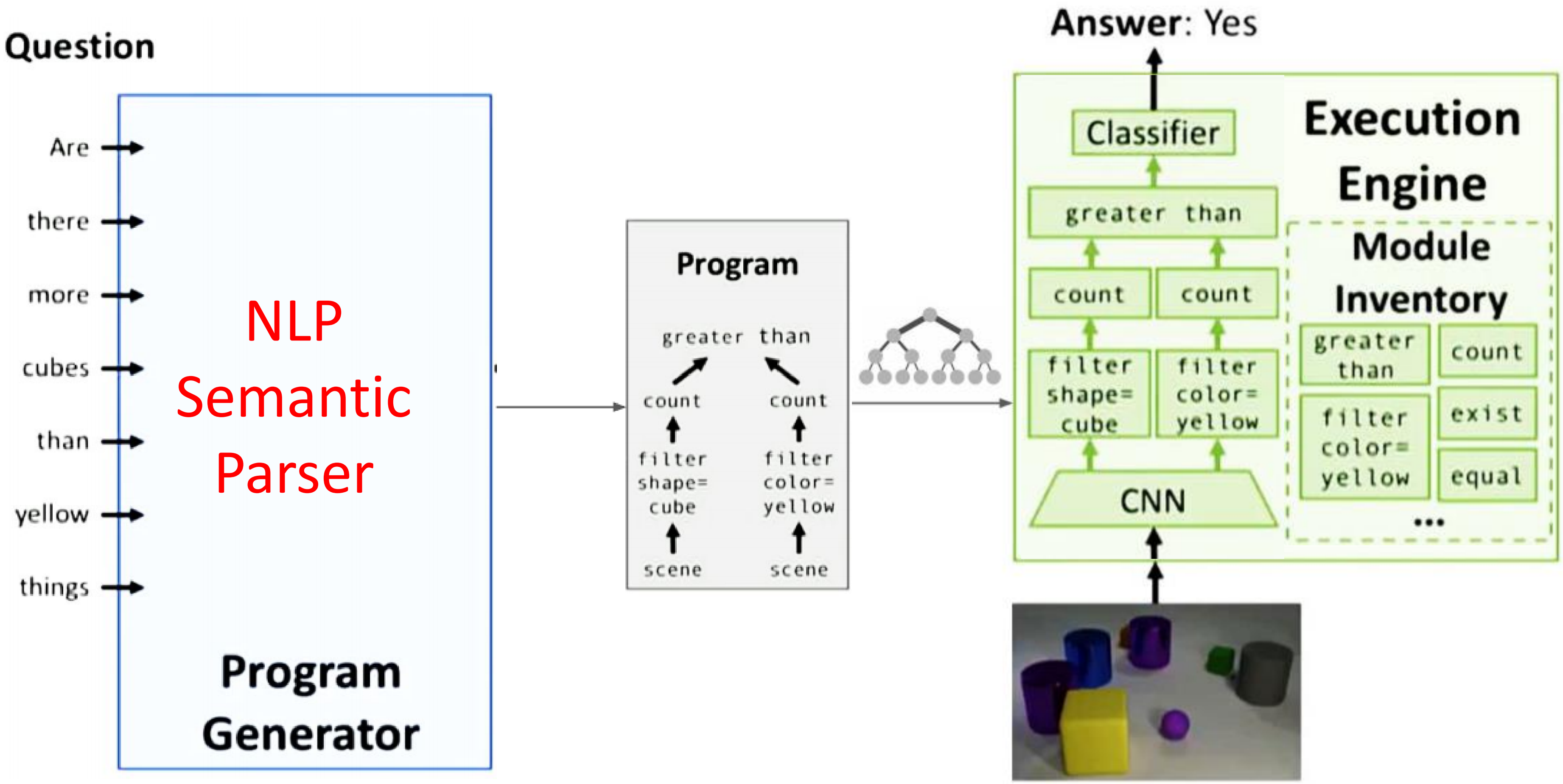
[1] CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning, CVPR, 2017

Consider a compositional model



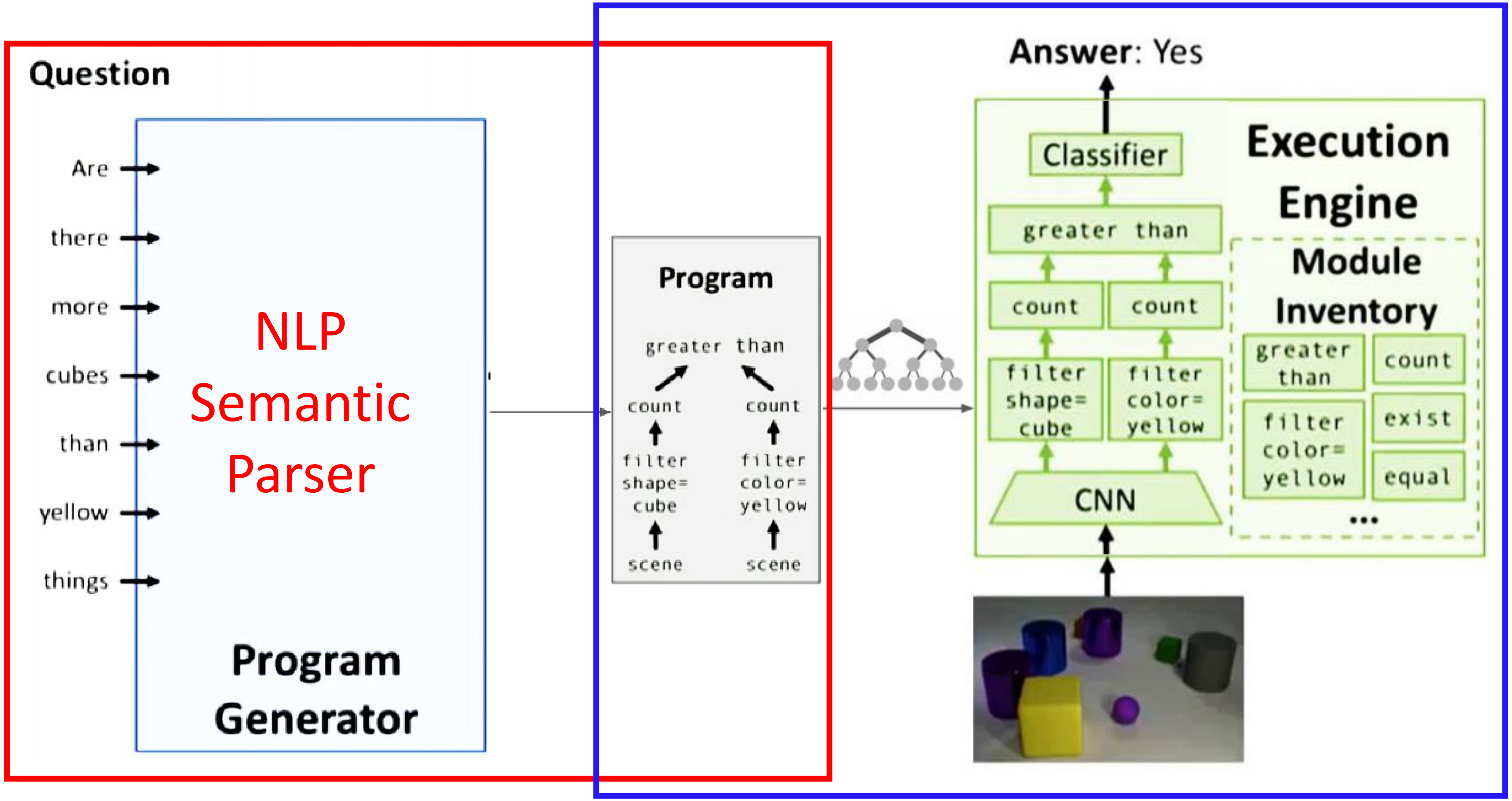
[1] Deep Compositional Question Answering with Neural Module Networks, CVPR, 2016

Overview of the NMN approach



[1] Deep Compositional Question Answering with Neural Module Networks, CVPR, 2016

Overview of the NMN approach

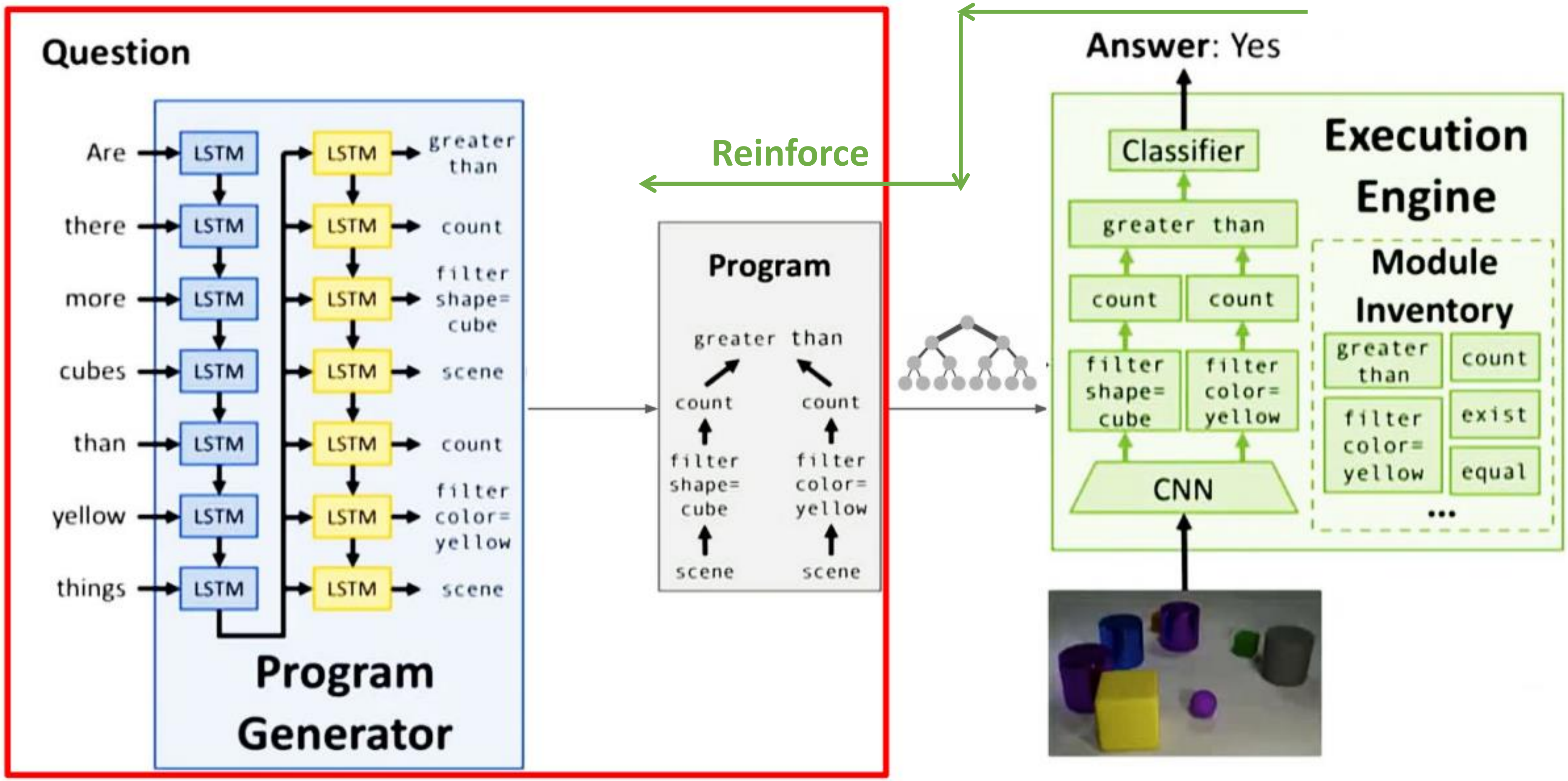


Uses some pre-trained parser

Trained separately

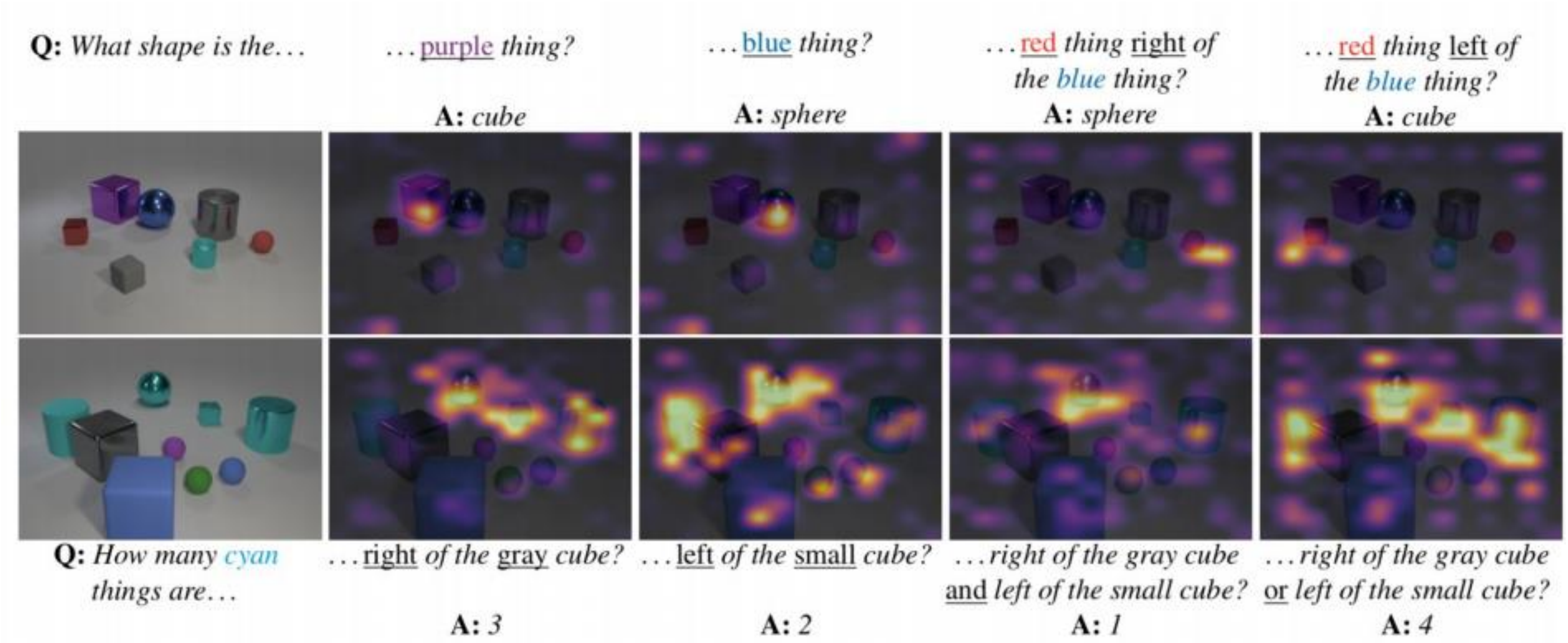
[1] Deep Compositional Question Answering with Neural Module Networks, CVPR, 2016

Inferring and Executing Programs

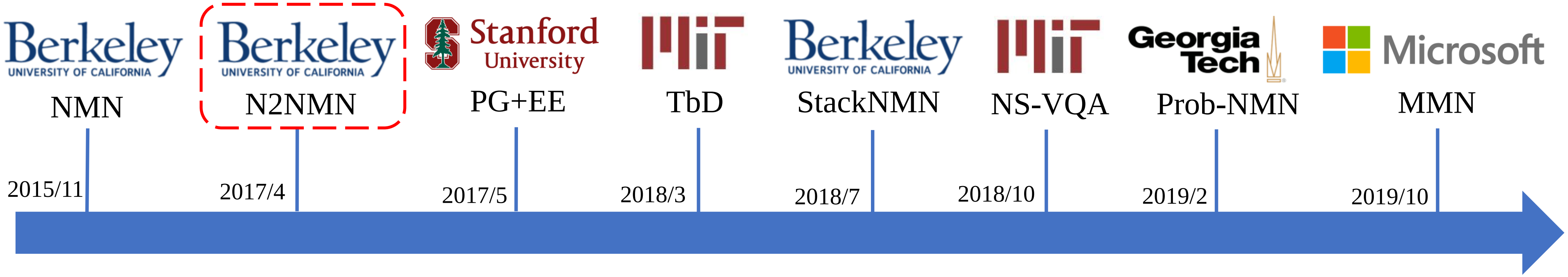


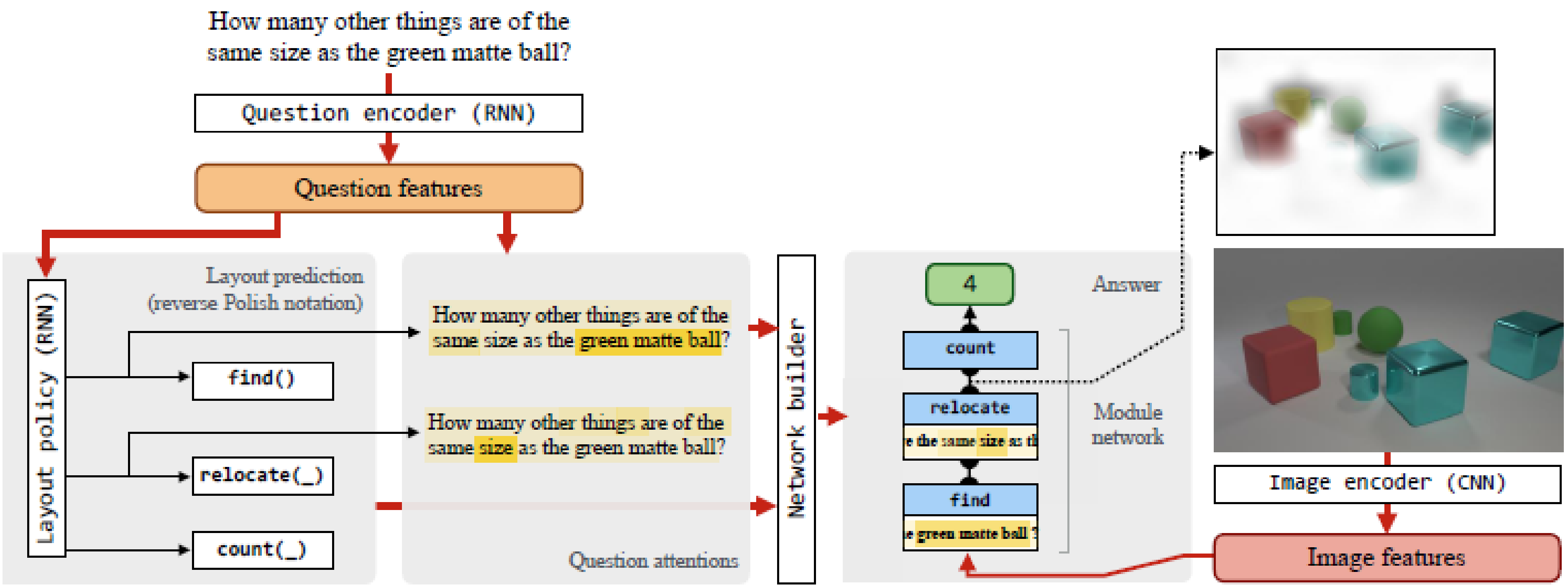
[1] Inferring and Executing Programs for Visual Reasoning, ICCV, 2017

What do the modules learn?

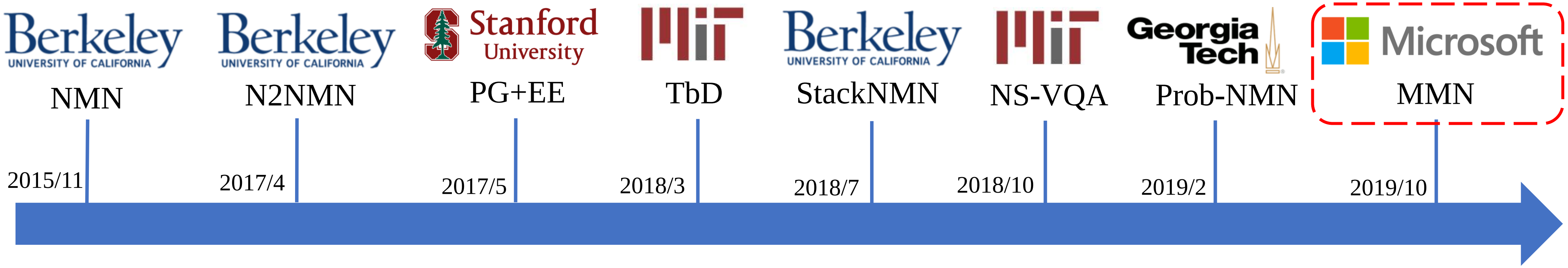


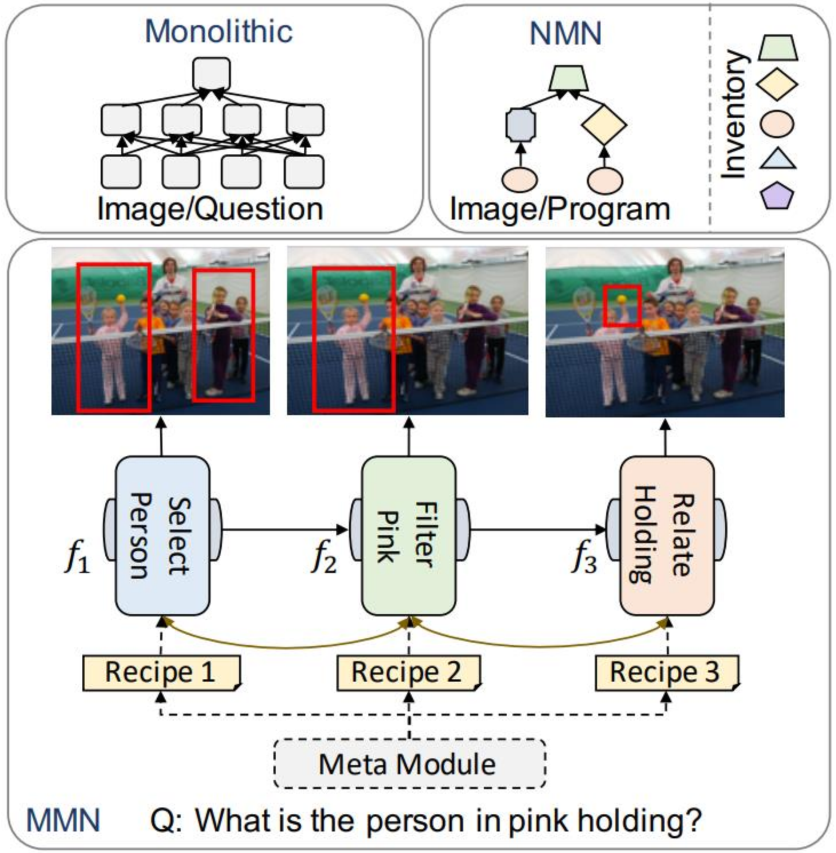
[1] Inferring and Executing Programs for Visual Reasoning, ICCV, 2017

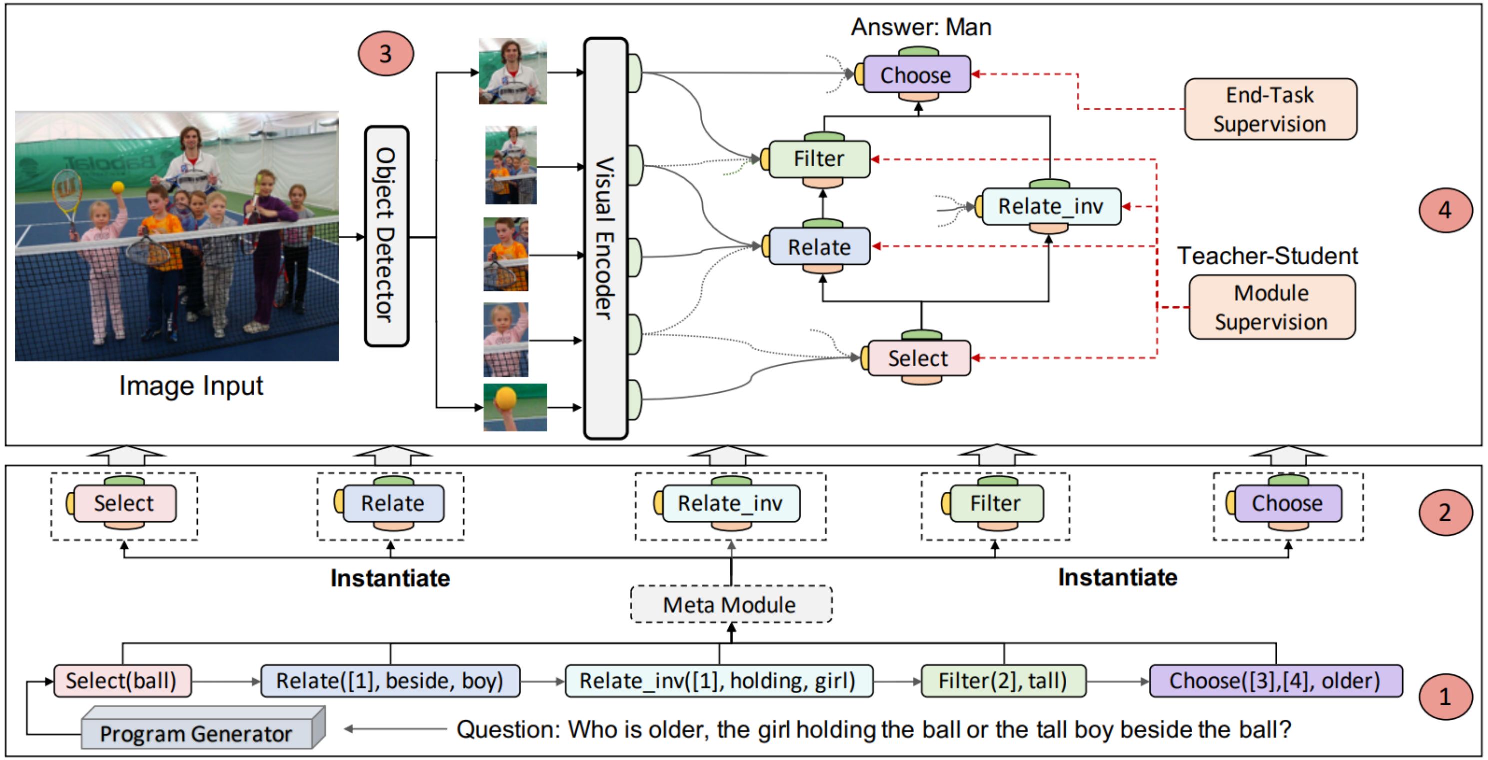




[1] Learning to Reason End-to-End Module Networks for Visual Question Answering, ICCV, 2017



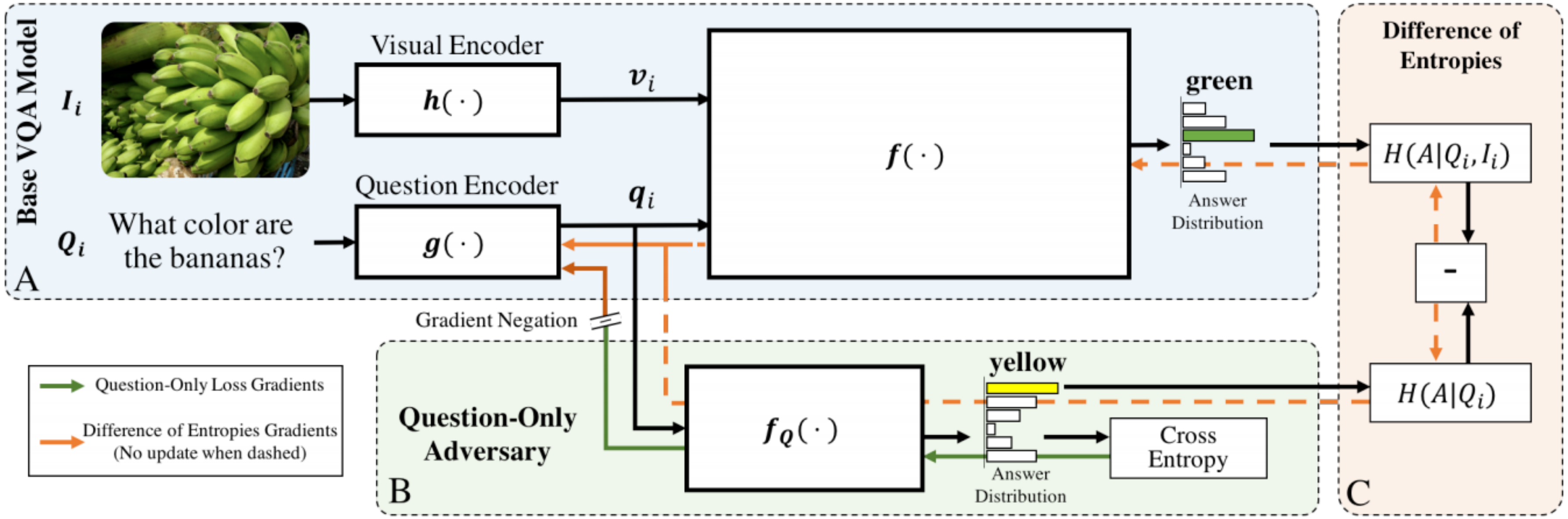




[1] Meta Module Network for Compositional Visual Reasoning, 2019

Robust VQA: two examples

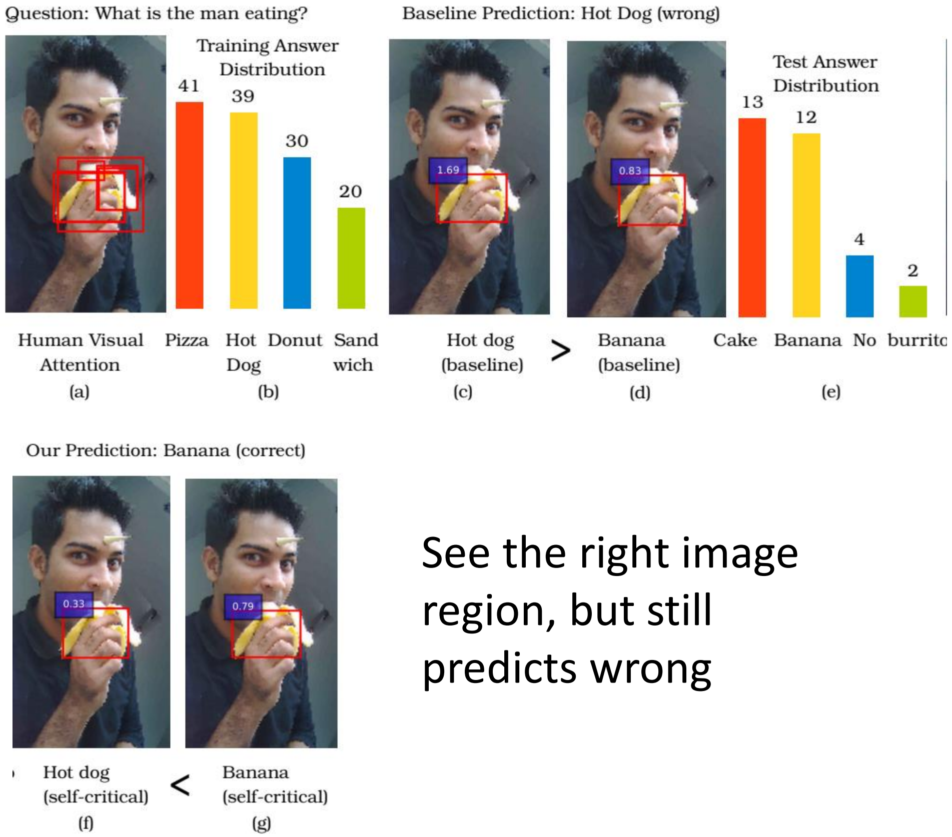
- Overcoming language prior with adversarial regularization



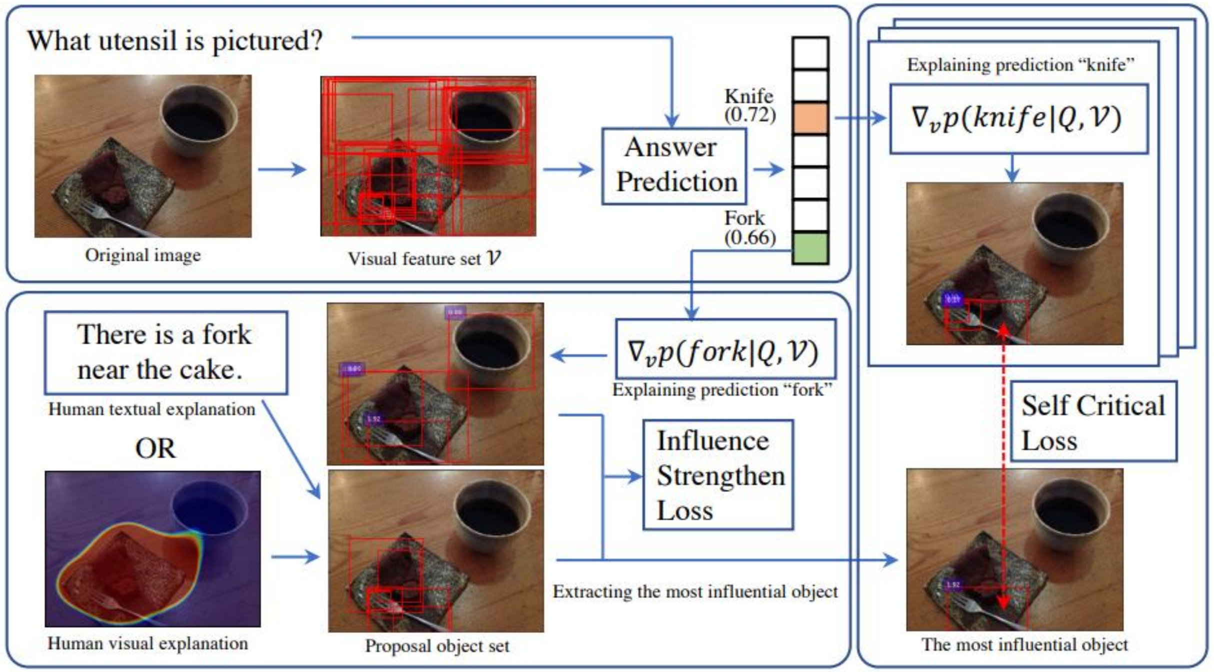
[1] Overcoming Language Priors in Visual Question Answering with Adversarial Regularization, NeurIPS 2018

Robust VQA: two examples

- Self-critical reasoning



See the right image region, but still predicts wrong



[1 Self-Critical Reasoning for Robust Visual Question Answering, NeurIPS 2019

Agenda

- Task Overview
  - *What are the main tasks that are driving progress in V+L representation learning?*

- Method Overview
  - *What are the state-of-the-art approaches and the key model design principles underlying these methods?*

- Summary
  - *What are the core challenges and future directions?*

Take-away Messages

- Popular tasks:
  - VQA, GQA, VCR, RefCOCO, NLVR2, etc.
- Methods:
  - Grid vs. region features
  - Bilinear pooling and FiLM
  - Multimodal alignment with cross-modal attention
  - Relational reasoning with intra-modal attention (self-attention, graph attention)
  - Transformer model becomes popular in the field
  - Multi-step reasoning
  - Neural state machine
  - Neural module network

Challenges & Future Directions

- Can we have something like GLUE and SuperGLUE?
- Can we use a Visual Transformer to encode images to train a large V+L Transformer model end-to-end?
- Instead of Transformer, can we perform FiLM-like fusion for multi-modal pre-training?
- Since all the reasoning is performed in the embedding/neural space, it is not clear whether the model “truly” learns how to reason
- Adversarial robustness of V+L models is less explored in the current literature

Thank you!
Any Questions?