The University of Texas at Austin
Electrical and Computer Engineering
Cockrell School of Engineering

Fall 2021

ADVANCED TOPICS IN COMPUTER VISION

Atlas Wang
Assistant Professor, The University of Texas at Austin

Visual Informatics Group@UT Austin
https://vita-group.github.io/

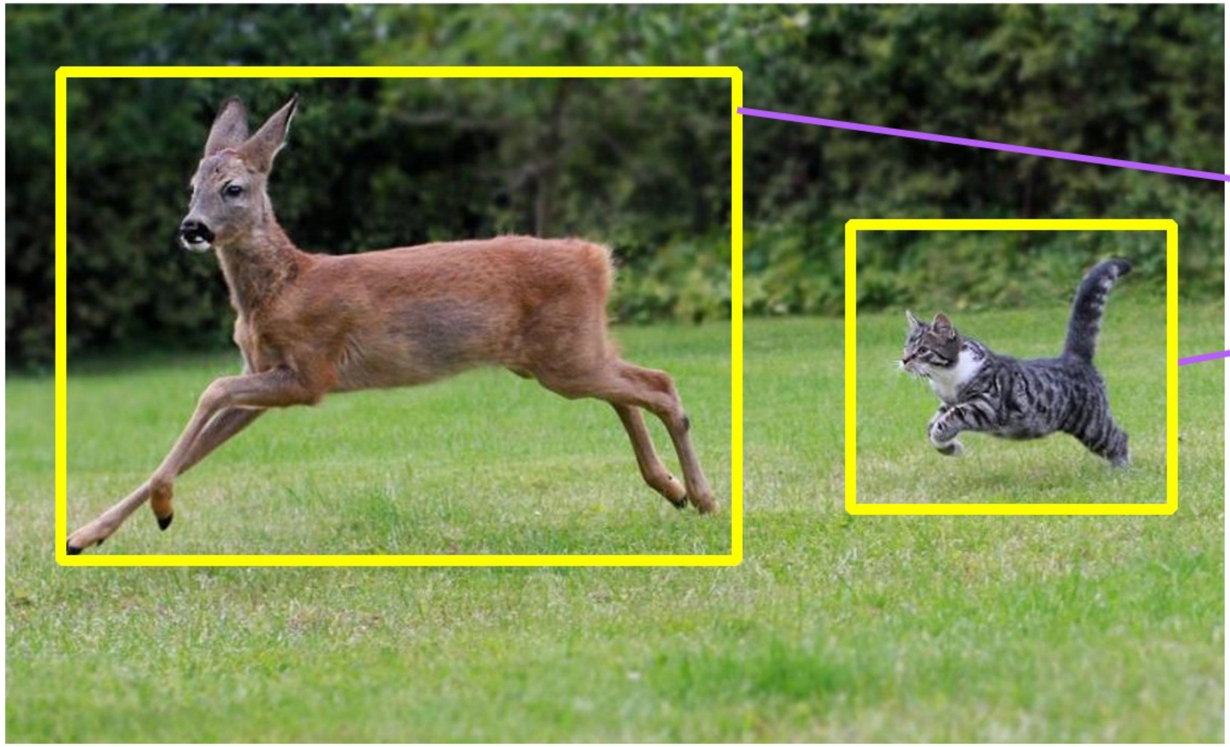
What is Visual Recognition?

Image tagging

deer
cat
trees
grass

What is Visual Recognition?

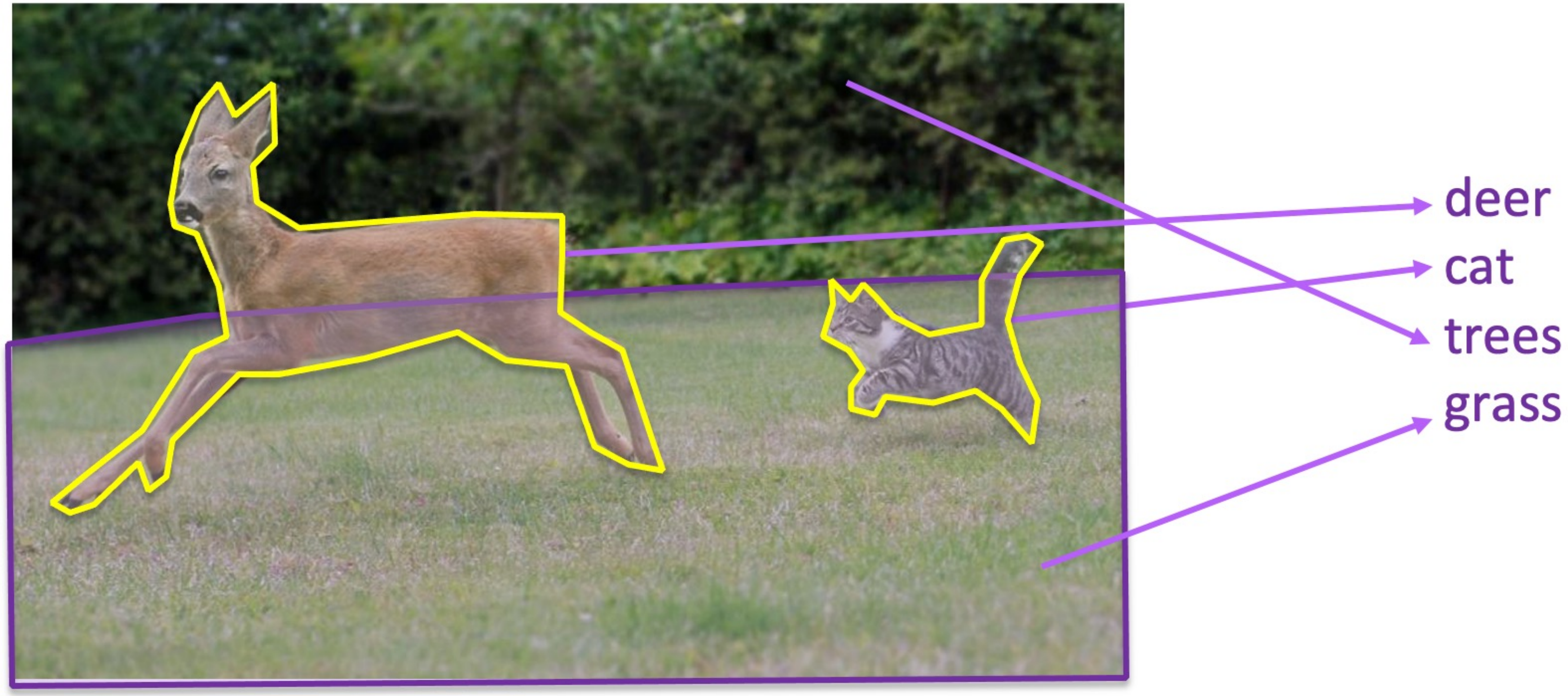
Object detection



deer
cat
trees
grass

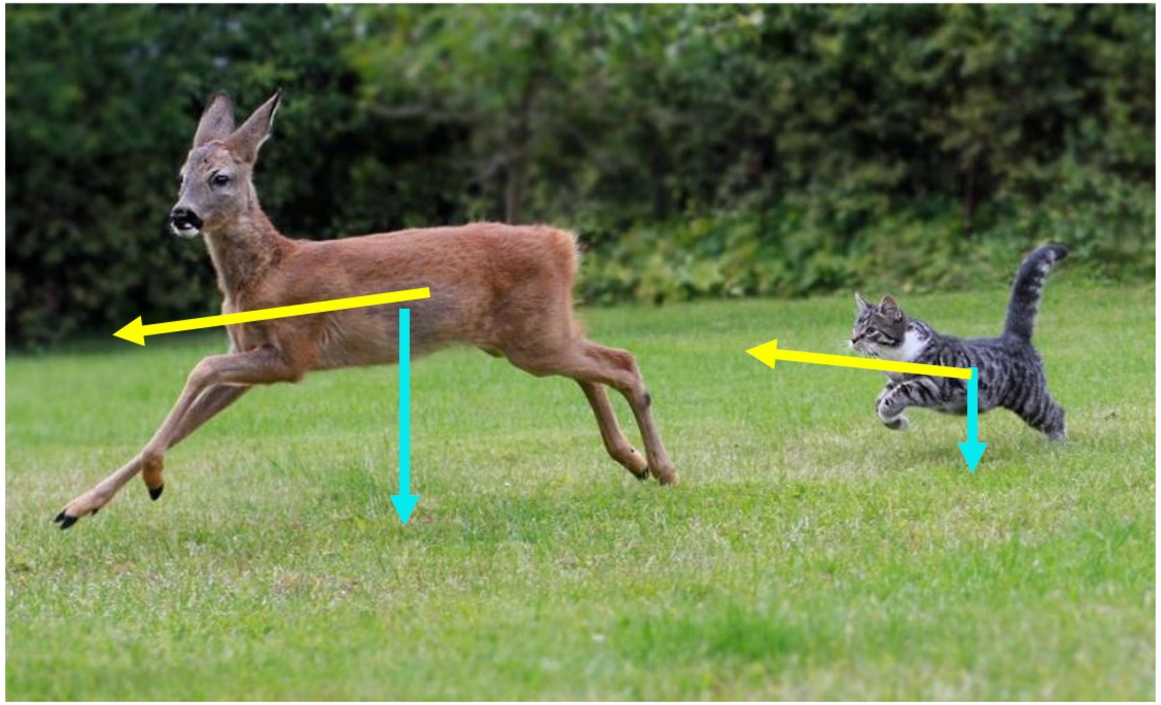
What is Visual Recognition?

Object segmentation



What is Visual Recognition?

Physics / Intuition



Gravity

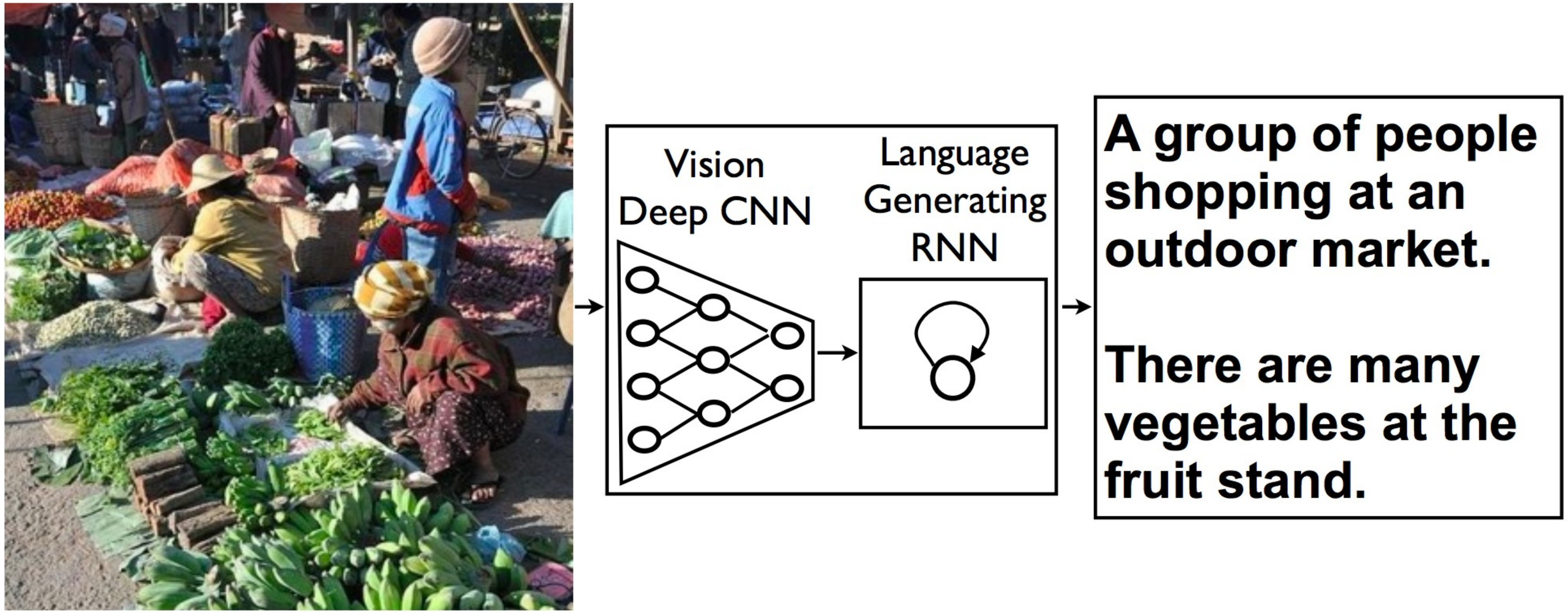
Velocity

Pushing the Limits of Visual Recognition

Reasoning about Language!

→ a cat is chasing a young deer

Vision + Language: Applications (1)

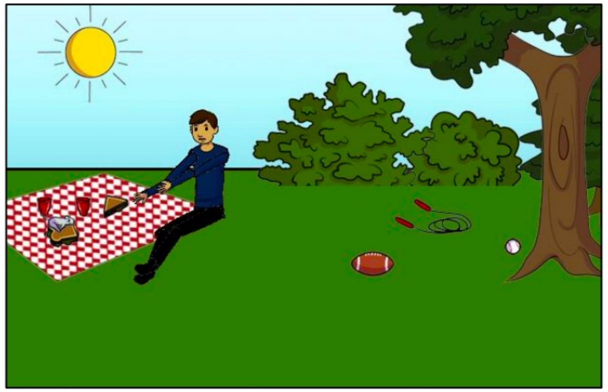


Visual Captioning: Vinyals et al. 2015

Vision + Language: : Applications (2)

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

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

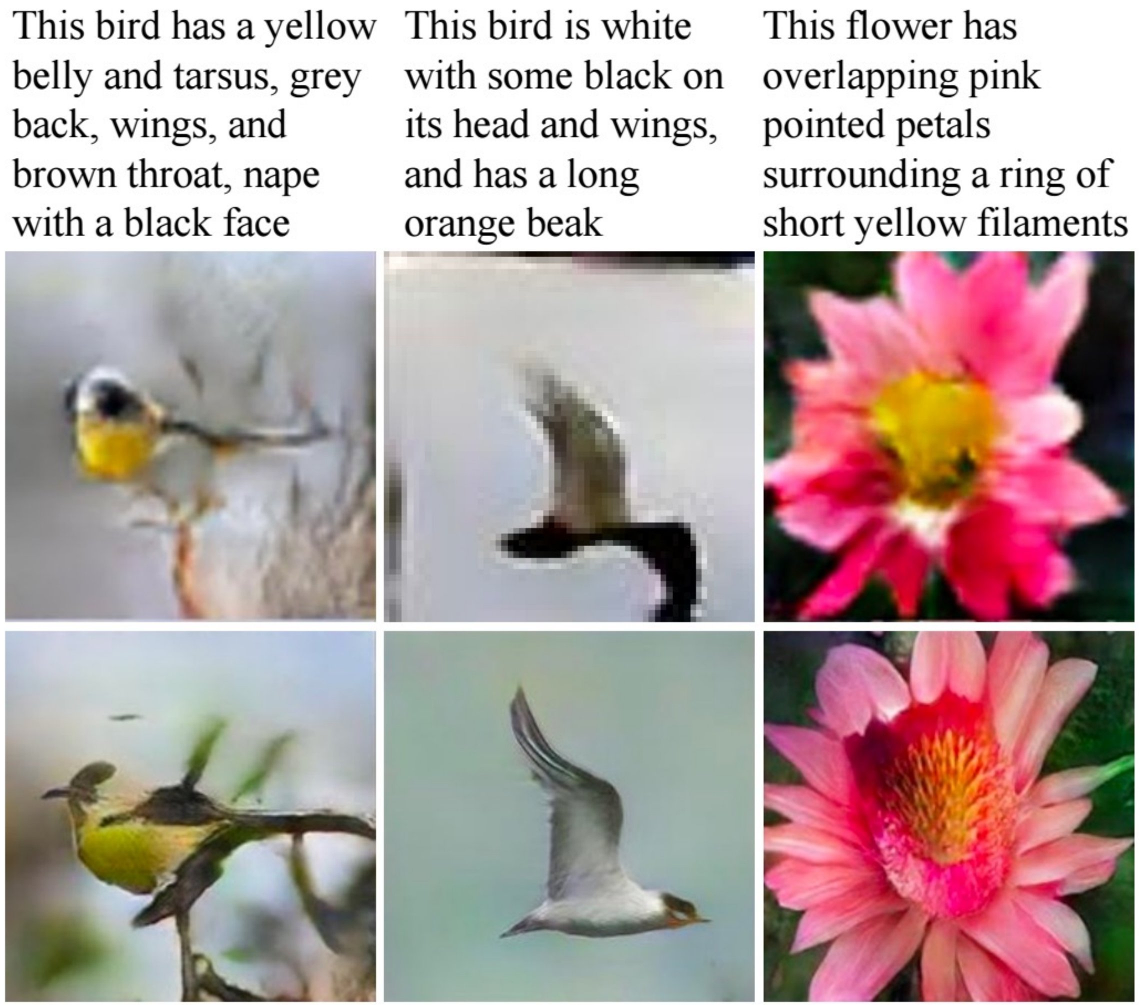


Is this person expecting company?
What is just under the tree?

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

Visual Question Answering: Agrawal et al. 2015

Vision + Language : Applications (3)



Text to Images: Zhang et al. 2016

Problem Overview (1): Visual Captioning

- Describe the content of an image or video with a natural language sentence.

A cat is sitting next to a pine tree, looking up.

A dog is playing piano with a girl.

Applications of Visual Captioning

- Alt-text generation (from PowerPoint)
- Content-based image retrieval (CBIR)
- Helping the visually impaired
- Or just for fun!

A fun video running visual captioning model real-time made by Kyle McDonald. Source: https://vimeo.com/146492001

Image Captioning with CNN-LSTM

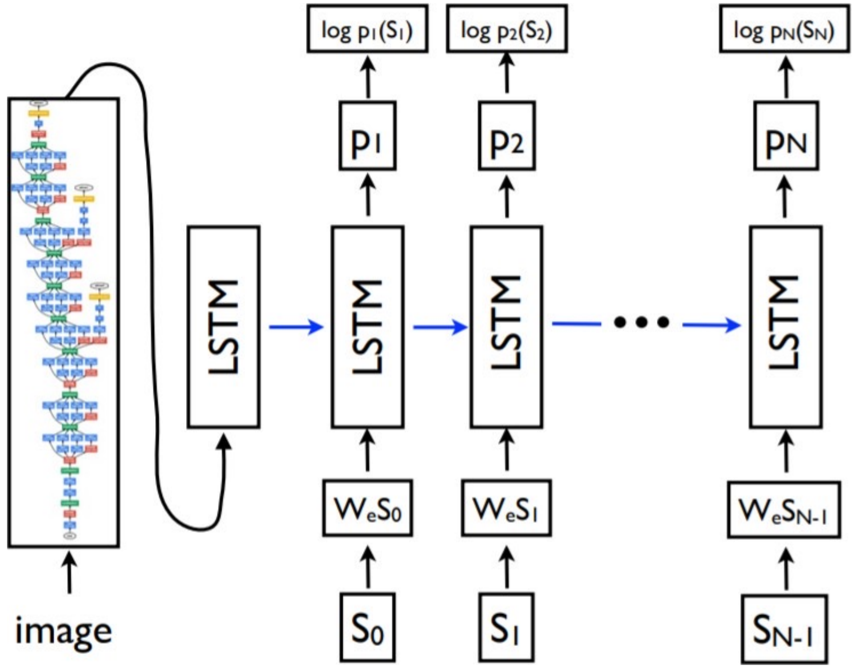
- Problem Formulation

$$\theta^\star = \arg\max_\theta \sum_{(I,S)} \log p(S|I;\theta)$$

$$\log p(S|I) = \sum_{t=0}^{N} \log p(S_t|I, S_0, \ldots, S_{t-1})$$

- The Encoder-Decoder framework

"Show and Tell"



→ Visual Encoder → Language Decoder → "Cat sitting outside"

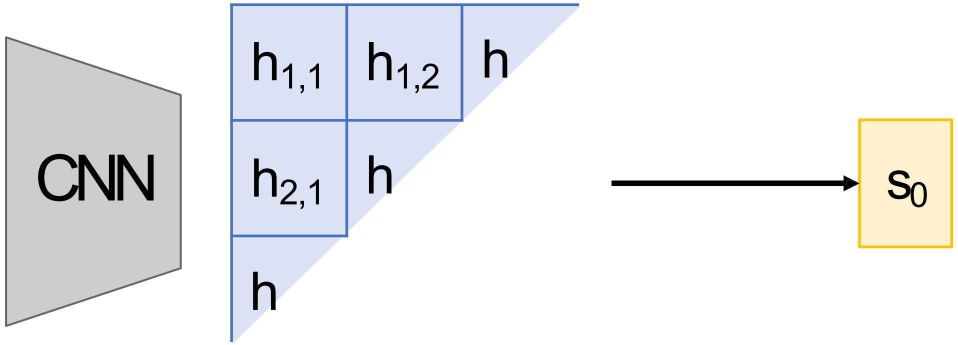
Image credit: Vinyals et al. "Show and Tell: A Neural Image Caption Generator", CVPR 2015.

Image Captioning with Soft Attention

- Soft Attention – Dynamically attend to input content based on query.
- Basic elements: query – q, keys - K, and values – V
- In our case, keys and values are usually identical. They come from the CNN activation map.
- Query q is determined by the global image feature or LSTM's hidden states.

Bahdanau et al. "Neural Machine Translation by Jointly Learning to Align and Translate", ICLR 2015.
Xu et al. "Show, Attend and Tell", ICML 2015.

Image Captioning with Soft Attention

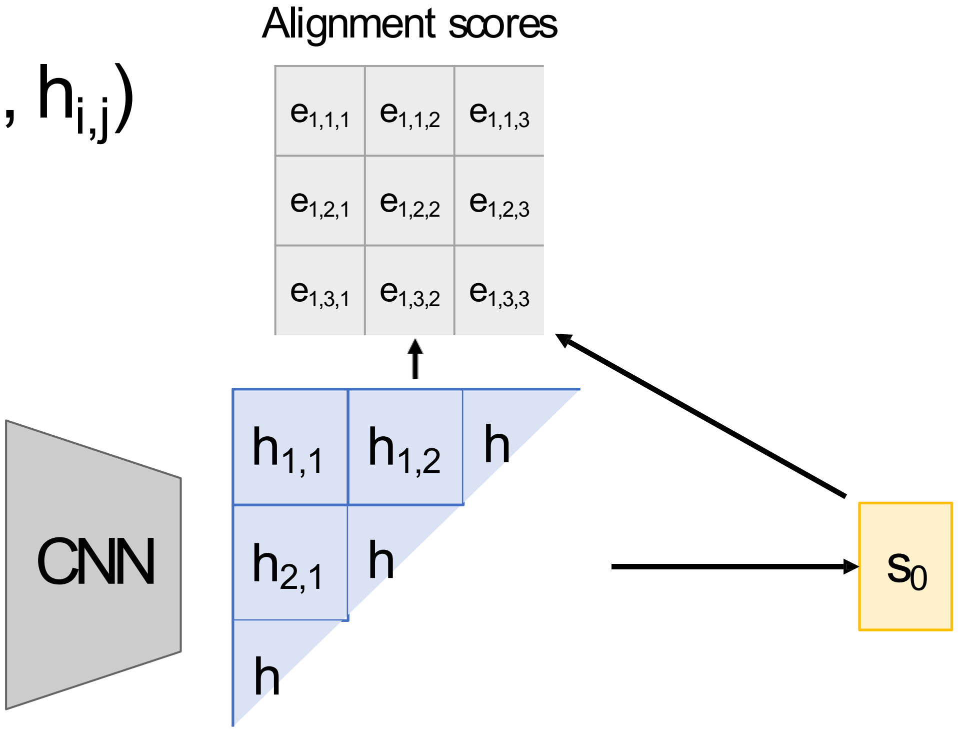


Use a CNN to compute a grid of features for an image

Slide credit: UMich EECS498/598 DeepVision course by Justin Johnson. Method: “Show, Attend and Tell” by Xu et al. ICML 2015.

Image Captioning with Soft Attention

$$e_{t,i,j} = f_{att}(s_{t-1}, h_{i,j})$$

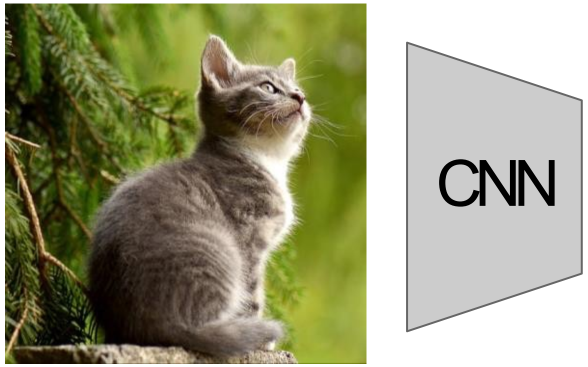


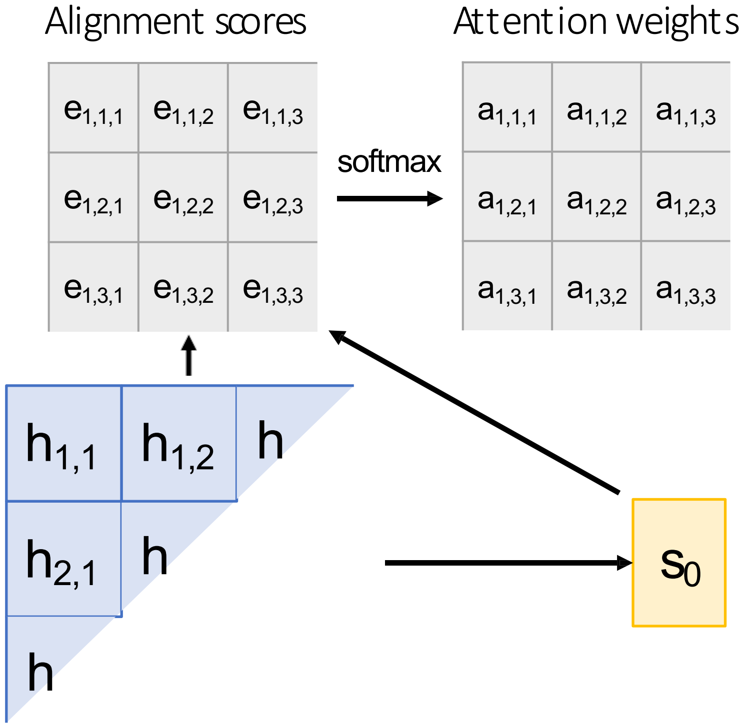
Use a CNN to compute a grid of features for an image

Slide credit: UMich EECS498/598 DeepVision course by Justin Johnson. Method: "Show, Attend and Tell" by Xu et al. ICML 2015.

Image Captioning with Soft Attention

$e_{t,i,j} = f_{att}(s_{t-1}, h_{i,j})$
$a_{t,:,:} = \text{softmax}(e_{t,:,:})$



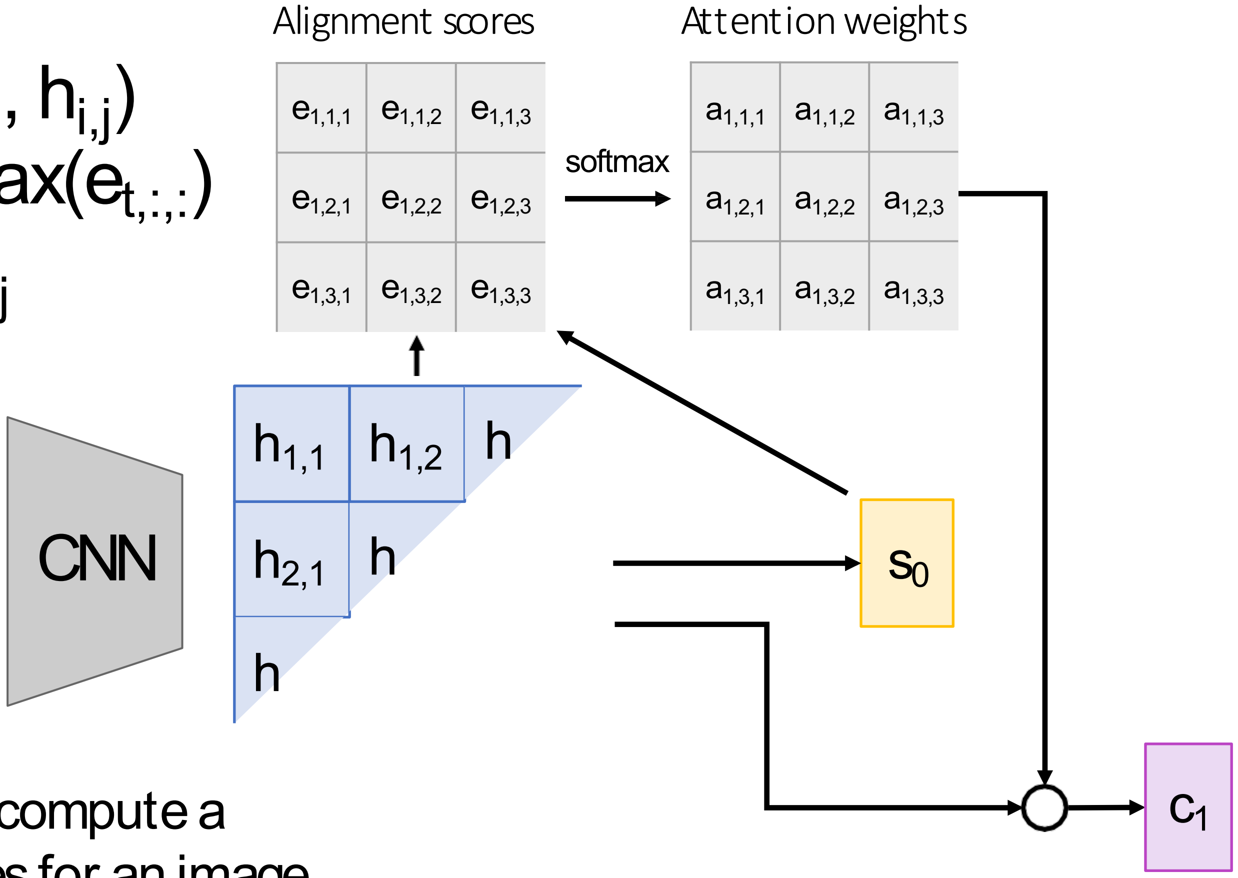


Use a CNN to compute a
grid of features for an image

Slide credit: UMich EECS498/598 DeepVision course by Justin Johnson. Method: “Show, Attend and Tell” by Xu et al. ICML 2015.

Image Captioning with Soft Attention

$e_{t,i,j} = f_{att}(s_{t-1}, h_{i,j})$
$a_{t,:,:} = \text{softmax}(e_{t,:,:})$
$c_t = \sum_{i,j} a_{t,i,j} h_{i,j}$

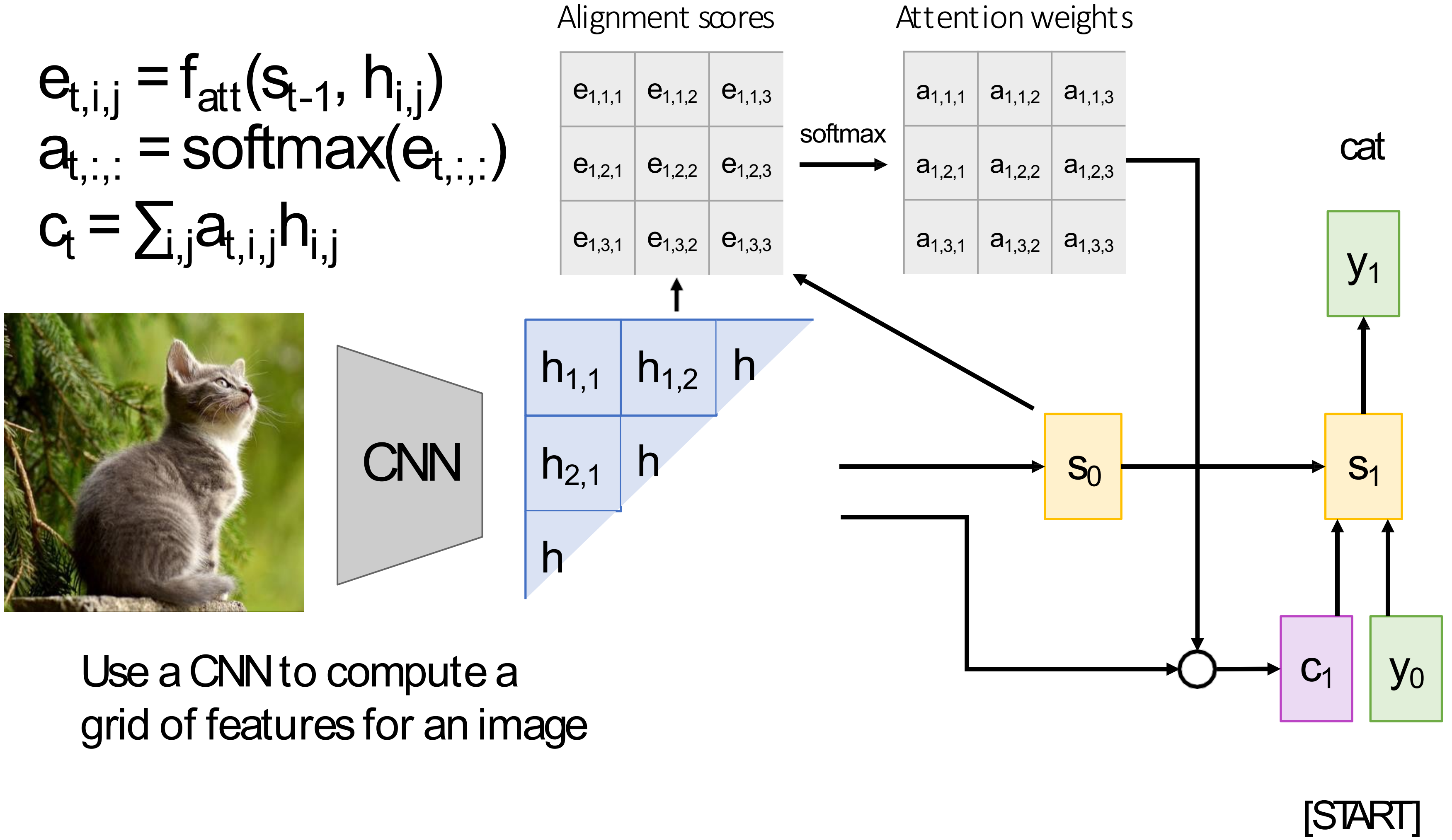


Use a CNN to compute a grid of features for an image

Slide credit: UMich EECS498/598 DeepVision course by Justin Johnson. Method: “Show, Attend and Tell” by Xu et al. ICML 2015.

Image Captioning with Soft Attention

$$e_{t,i,j} = f_{att}(s_{t-1}, h_{i,j})$$
$$a_{t,:,:} = \text{softmax}(e_{t,:,:})$$
$$c_t = \sum_{i,j} a_{t,i,j} h_{i,j}$$

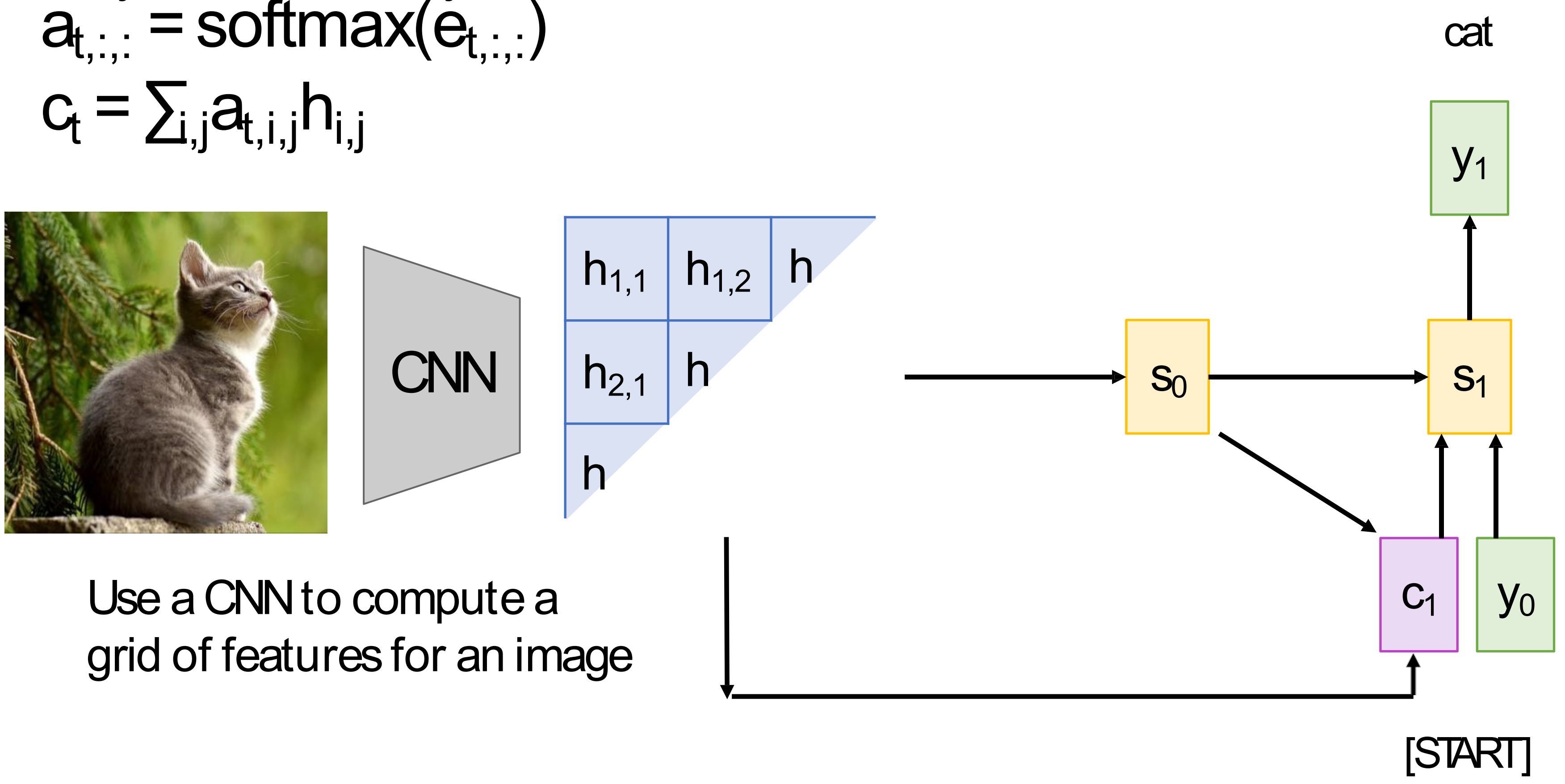


Use a CNN to compute a grid of features for an image

Slide credit: UMich EECS498/598 DeepVision course by Justin Johnson. Method: “Show, Attend and Tell” by Xu et al. ICML 2015.

Image Captioning with Soft Attention

$e_{t,i,j} = f_{att}(s_{t-1}, h_{i,j})$
$a_{t,:,:} = \text{softmax}(e_{t,:,:})$
$c_t = \sum_{i,j} a_{t,i,j} h_{i,j}$

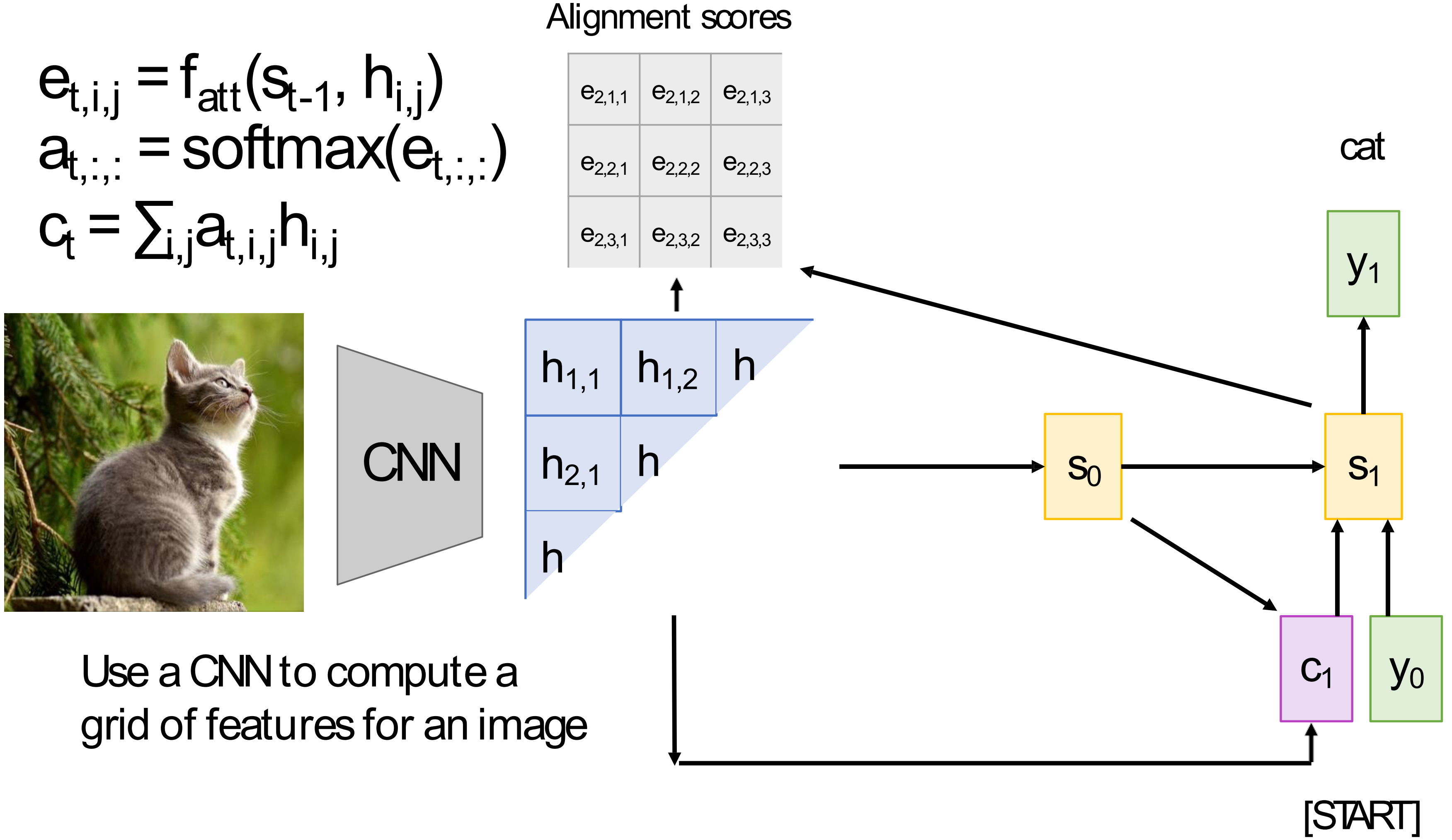


Use a CNN to compute a grid of features for an image

Slide credit: UMich EECS498/598 DeepVision course by Justin Johnson. Method: “Show, Attend and Tell” by Xu et al. ICML 2015.

Image Captioning with Soft Attention

$e_{t,i,j} = f_{att}(s_{t-1}, h_{i,j})$
$a_{t,:,:} = \text{softmax}(e_{t,:,:})$
$c_t = \sum_{i,j} a_{t,i,j} h_{i,j}$

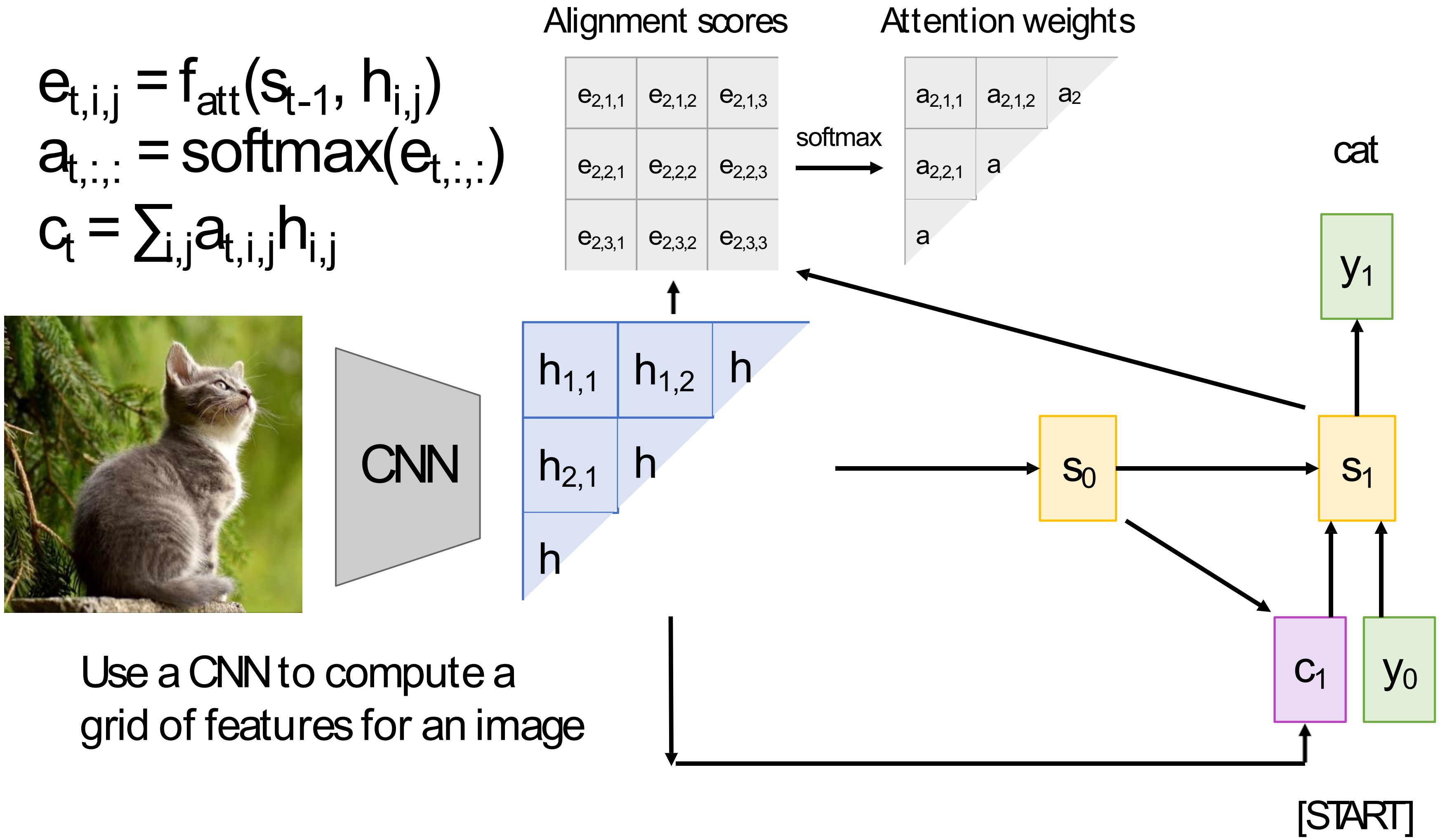


Use a CNN to compute a grid of features for an image

Slide credit: UMich EECS498/598 DeepVision course by Justin Johnson. Method: “Show, Attend and Tell” by Xu et al. ICML 2015.

Image Captioning with Soft Attention

$e_{t,i,j} = f_{att}(s_{t-1}, h_{i,j})$
$a_{t,:,:} = \text{softmax}(e_{t,:,:})$
$c_t = \sum_{i,j} a_{t,i,j} h_{i,j}$

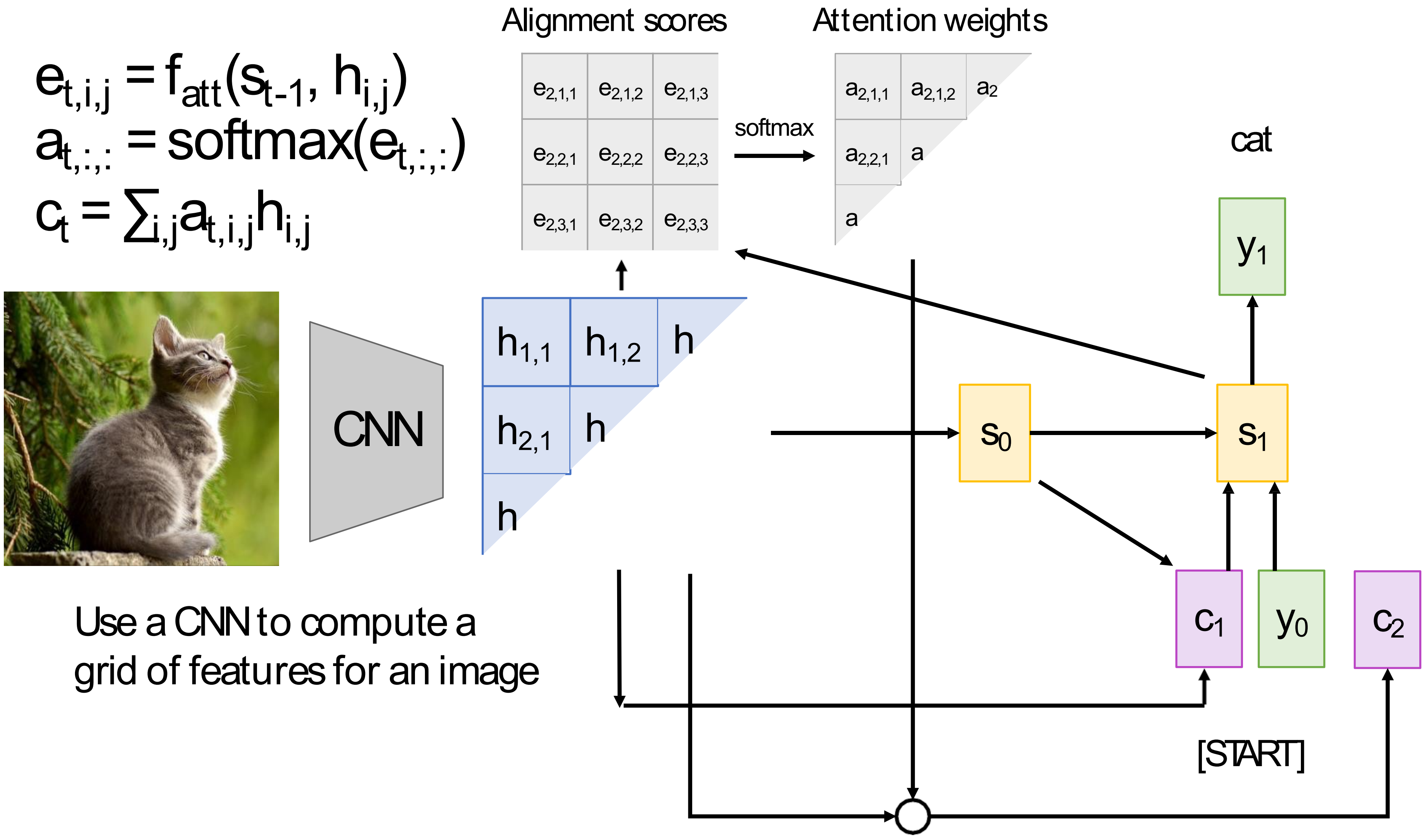


Use a CNN to compute a grid of features for an image

Slide credit: UMich EECS498/598 DeepVision course by Justin Johnson. Method: "Show, Attend and Tell" by Xu et al. ICML 2015.

Image Captioning with Soft Attention

$e_{t,i,j} = f_{att}(s_{t-1}, h_{i,j})$
$a_{t,:,:} = \mathrm{softmax}(e_{t,:,:})$
$c_t = \sum_{i,j} a_{t,i,j} h_{i,j}$

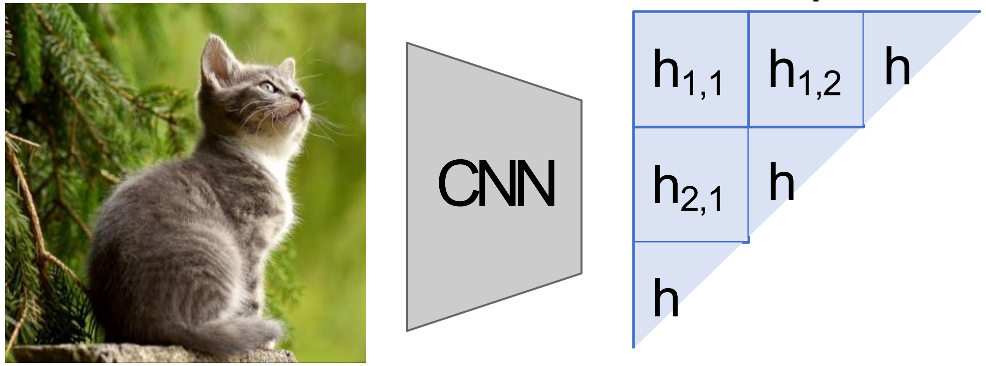


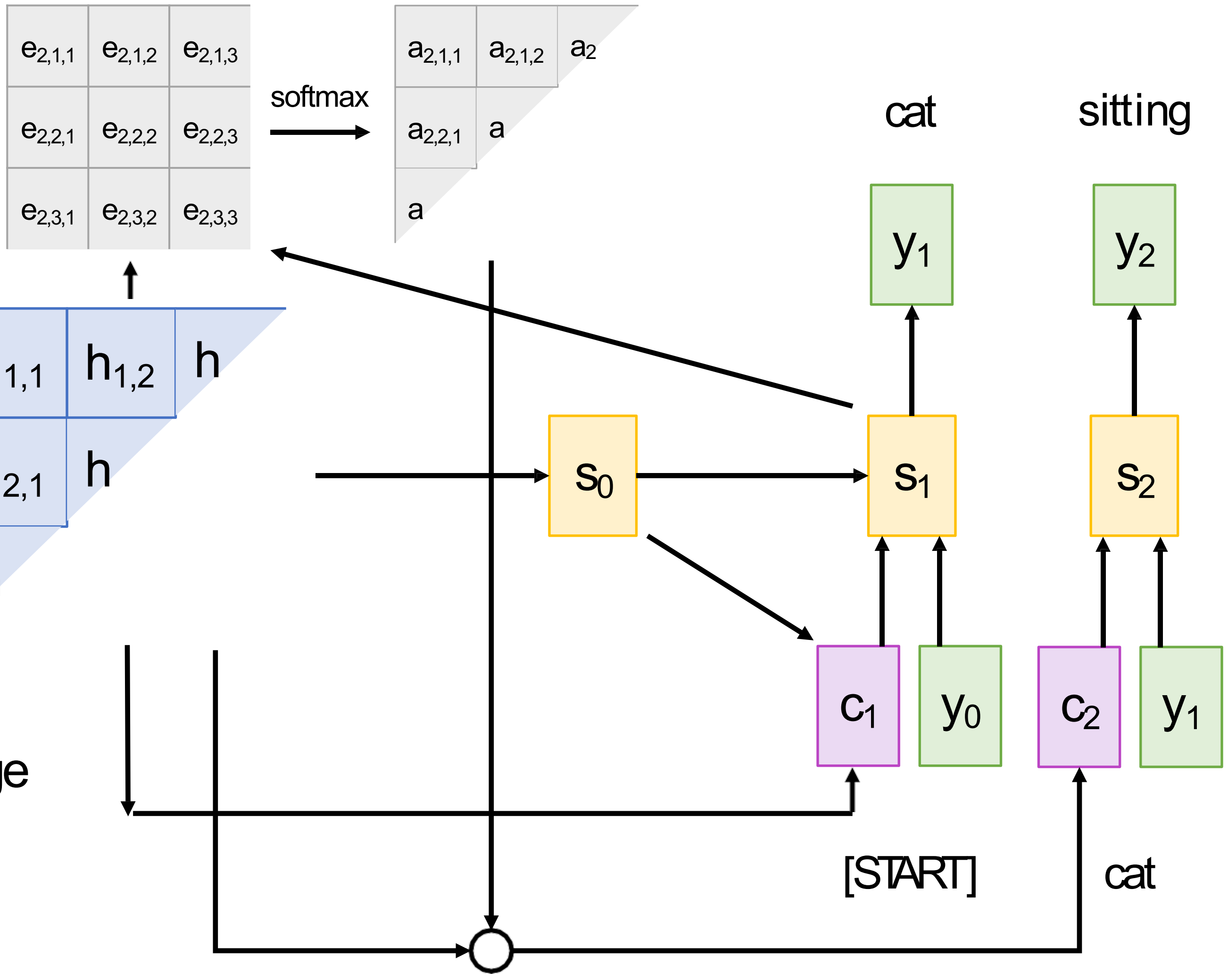
Use a CNN to compute a grid of features for an image

Slide credit: UMich EECS498/598 DeepVision course by Justin Johnson. Method: “Show, Attend and Tell” by Xu et al. ICML 2015.

Image Captioning with Soft Attention

$$e_{t,i,j} = f_{att}(s_{t-1}, h_{i,j})$$
$$a_{t,:,:} = \text{softmax}(e_{t,:,:})$$
$$c_t = \sum_{i,j} a_{t,i,j} h_{i,j}$$





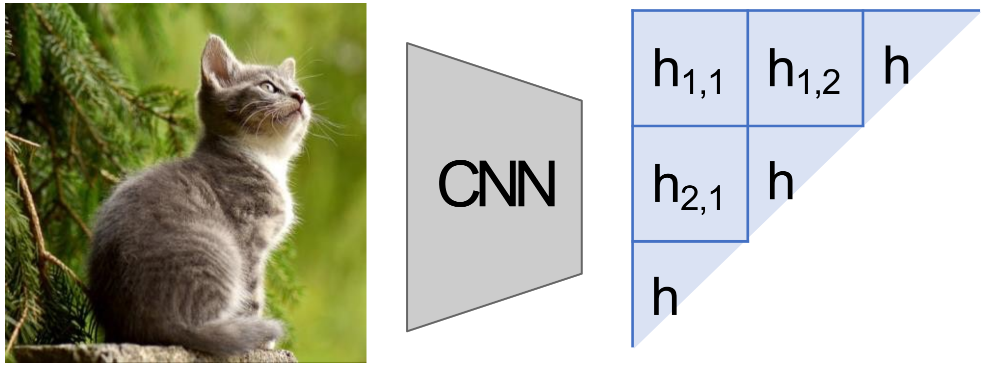
Use a CNN to compute a grid of features for an image

Slide credit: UMich EECS498/598 DeepVision course by Justin Johnson. Method: "Show, Attend and Tell" by Xu et al. ICML 2015.

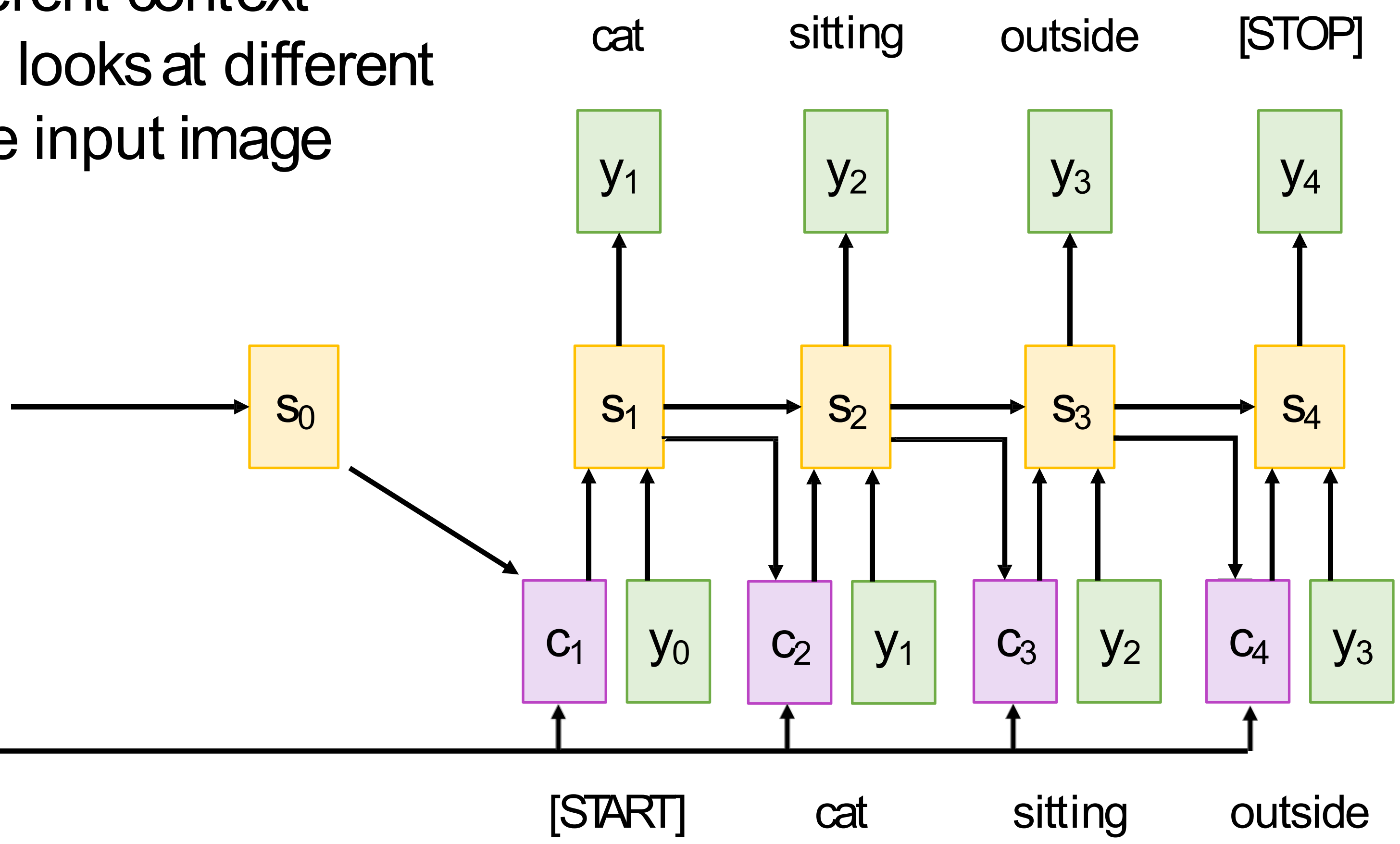
Image Captioning with Soft Attention

$e_{t,i,j} = f_{att}(s_{t-1}, h_{i,j})$
$a_{t,:,:} = \text{softmax}(e_{t,:,:})$
$c_t = \sum_{i,j} a_{t,i,j} h_{i,j}$

Each timestep of decoder uses a different context vector that looks at different parts of the input image

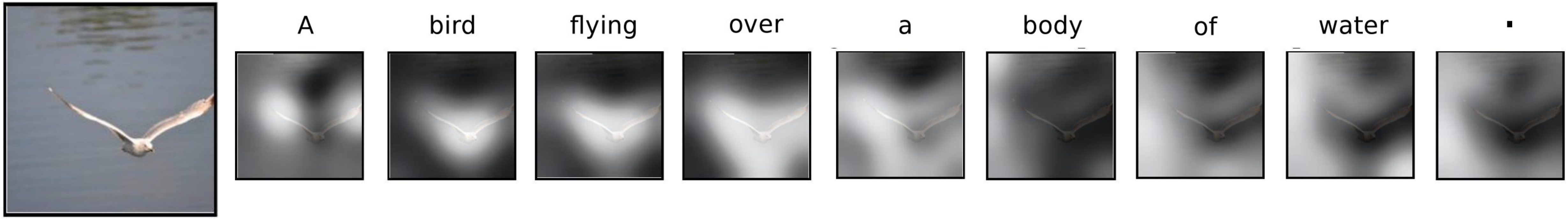


Use a CNN to compute a grid of features for an image



Slide credit: UMich EECS498/598 DeepVision course by Justin Johnson. Method: "Show, Attend and Tell" by Xu et al. ICML 2015.

Image Captioning with Soft Attention



Slide credit: UMich EECS498/598 DeepVision course by Justin Johnson. Method: “Show, Attend and Tell” by Xu et al. ICML 2015.

Image Captioning with Region Attention

- Variants of Soft Attention based on the feature input
  - Grid activation features (covered)
  - Region proposal features

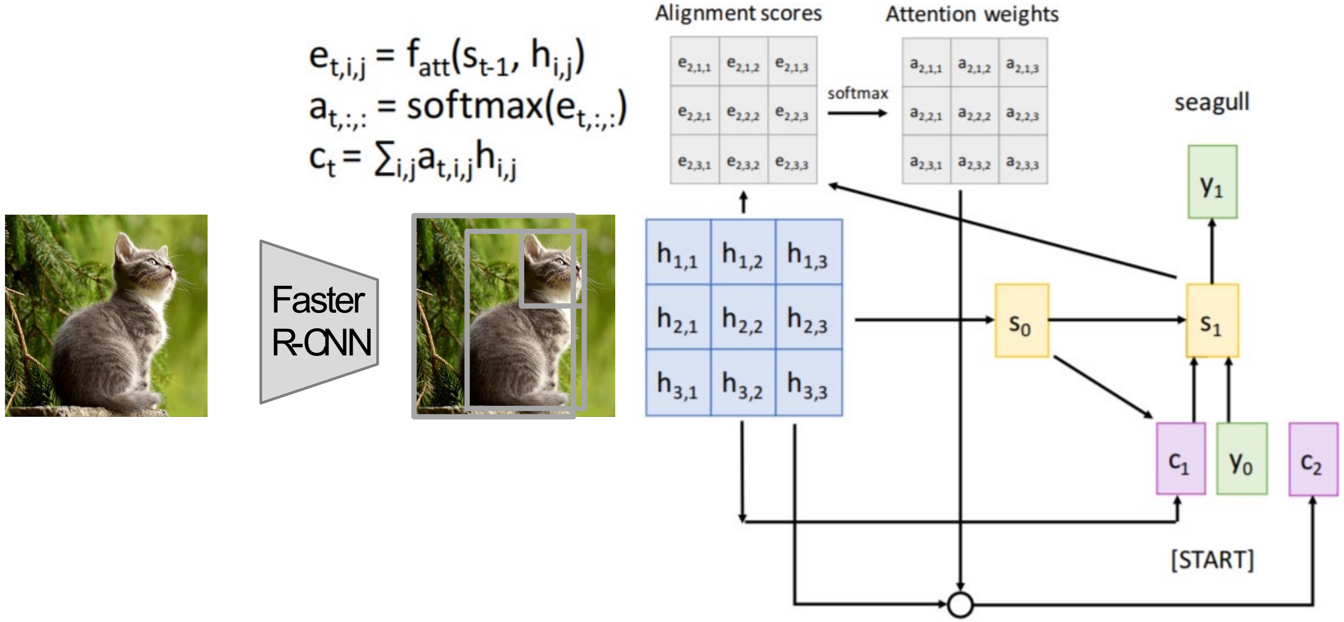


Image Captioning with Transformer

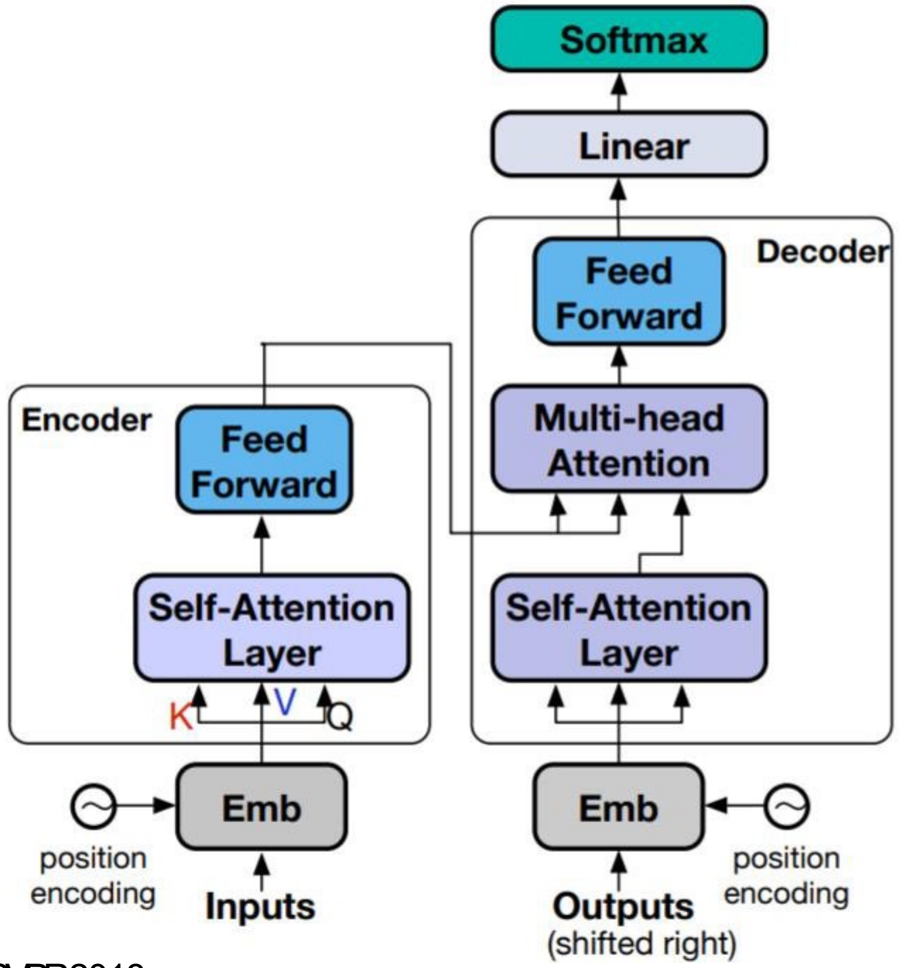
- Transformer performs sequence-to-sequence generation.
- Self-Attention – A type of soft attention that "attends to itself".
- Self-Attention is a special case of Graph Neural Networks (GNNs) that has a fully-connected graph.
- Self-attention is sometimes used to model relationship between object regions, similar to GCNs.

Vaswani et al. "Attention is all you need", NIPS 2017.
Yao et al. "Exploring visual relationship for image captioning", ECCV 2018.
Further readings: https://graphdeeplearning.github.io/post/transformers-are-gnns/

Image Captioning with Transformer

- Transformer is first adapted for captioning in Zhou et al.

- Others: Object Relation Transformer, Meshed-Memory Transformer



Zhou et al. "End-to-end dense video captioning with masked transformer", CVPR2018.
Herdade et al. "Image Captioning: Transforming Objects into Words", NeurIPS 2019.
Cornia et al. "Meshed-Memory Transformer for Image Captioning", CVPR 2020.

Vision-Language Pre-training (VLP)

- Two-stage training strategy: **pre-training** and **fine-tuning**.

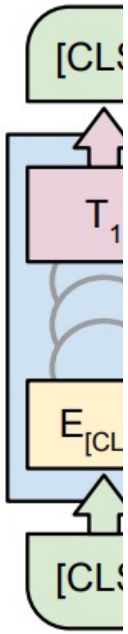
- **Pre-training** is performed on a large dataset. Usually with auto-generated captions. The training objective is *unsupervised*.

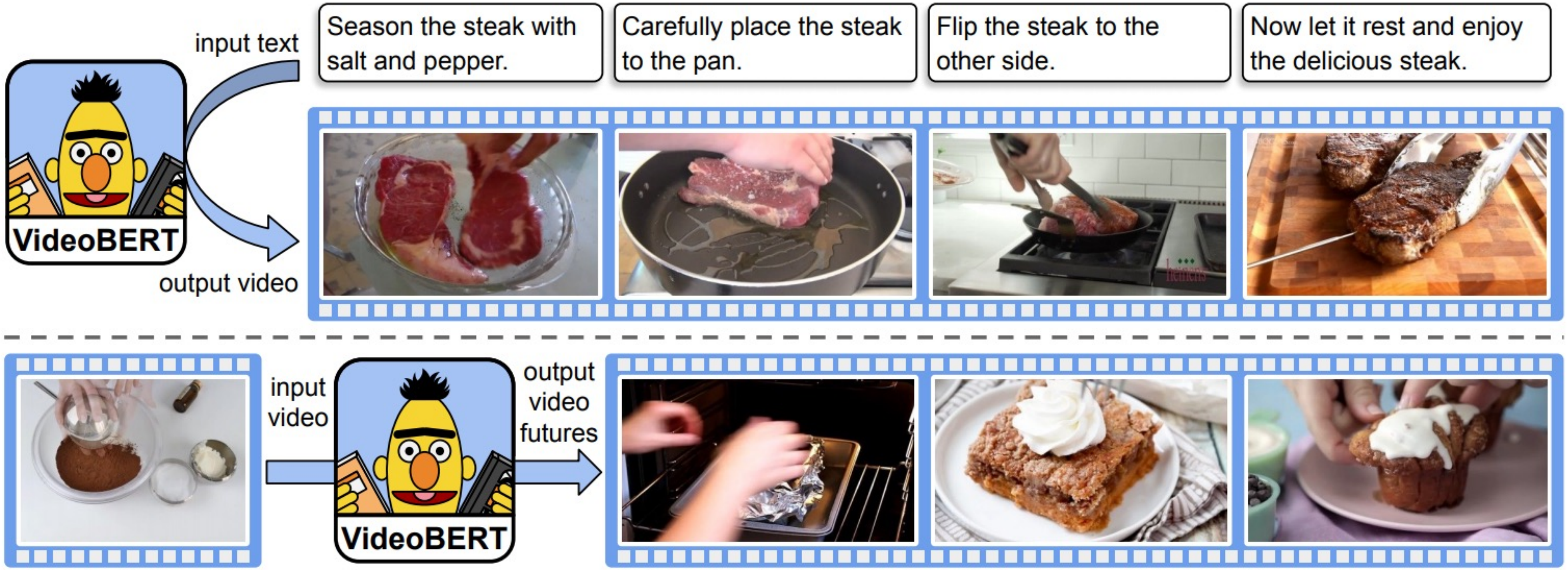
- **Fine-tuning** is task-specific *supervised* training on downstream tasks.

- All methods are based on BERT (a variant of Transformer).

Zhou et al. “Unified vision-language pre-training for image captioning and vqa”, AAAI 2020.
Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, NAACL 2019.

VideoBERT: A Joint Model for Video and Language Representation Learning





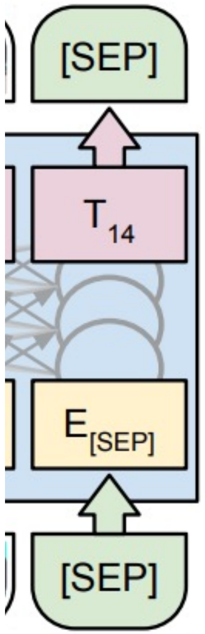


Figure 1: **VideoBERT text-to-video generation and future forecasting.** (Above) Given some recipe text divided into sentences, $y = y_{1:T}$, we generate a sequence of video tokens $x = x_{1:T}$ by computing $x_t^* = \arg\max_k p(x_t = k|y)$ using VideoBERT. (Below) Given a video token, we show the top three future tokens forecasted by VideoBERT at different time scales. In this case, VideoBERT predicts that a bowl of flour and cocoa powder may be baked in an oven, and may become a brownie or cupcake. We visualize video tokens using the images from the training set closest to centroids in feature space.

Grounded Visual Description

- Essentially, visual description + object grounding or detection

- To achieve better result interpretability, we need grounding!
  - Image domain: Neural Baby Talk, etc.
  - Video domain: Grounded Video Description, etc.

- Requires special dataset that has both description and bounding box

Lu et al. "Neural Baby Talk", CVPR 2018.
Zhou et al. "Grounded video description", CVPR 2019.

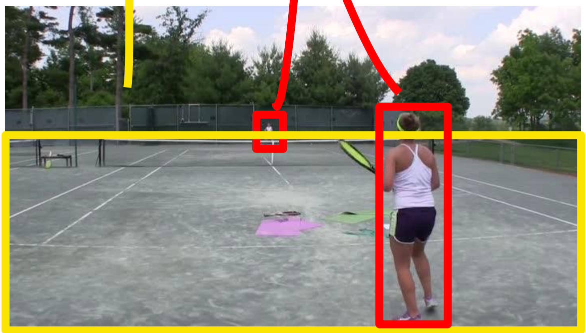
Single-Frame Annotation

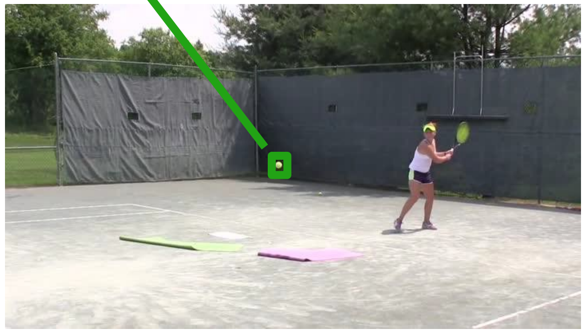
We see a man playing a saxophone in front of microphones.

From ActivityNet-Entities dataset. Zhou et al. “Grounded video description”, CVPR 2019.

Multi-Frame Annotation

Two women are on a tennis court, showing the technique to posing and hitting the ball.

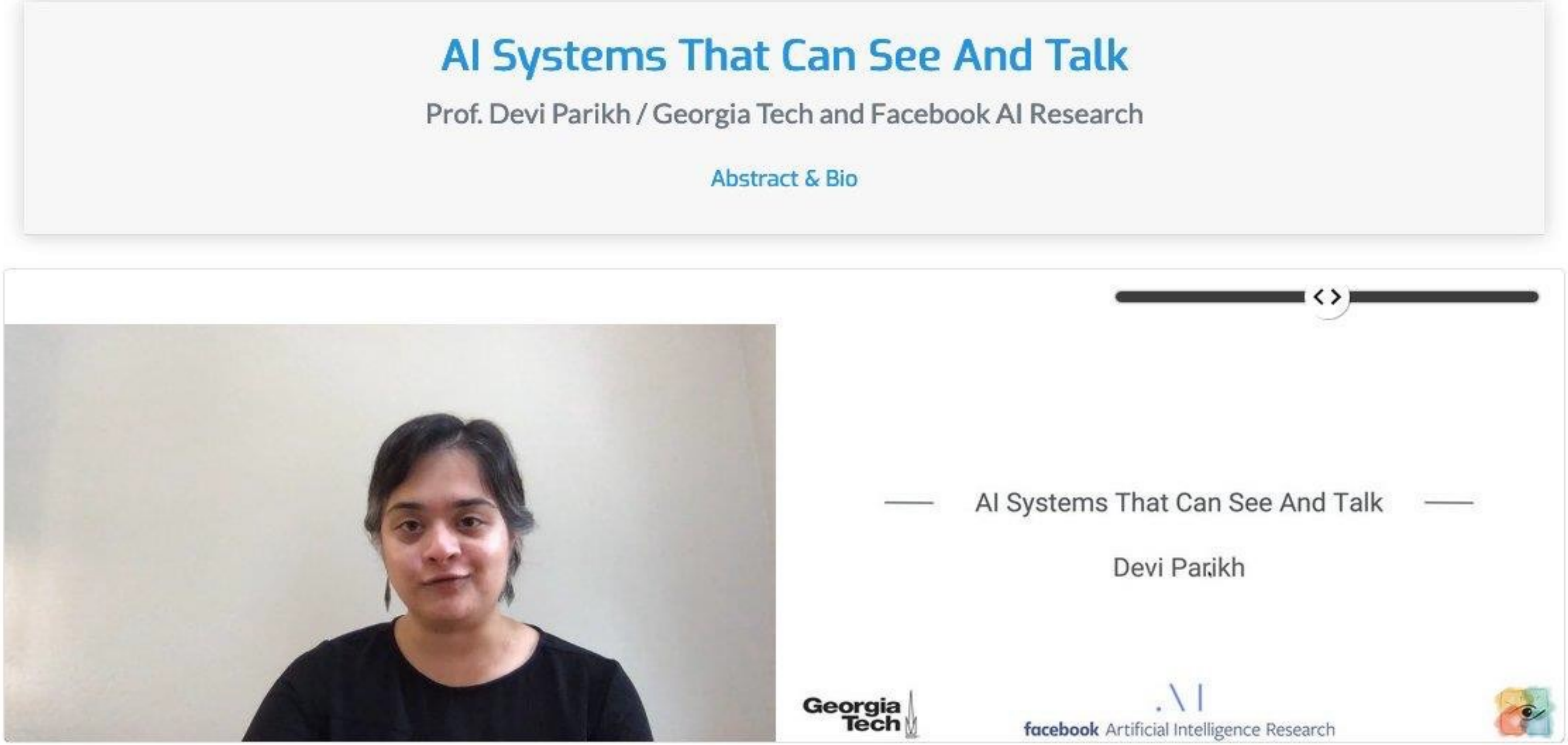




From ActivityNet-Entities dataset. Zhou et al. “Grounded video description”, CVPR 2019.

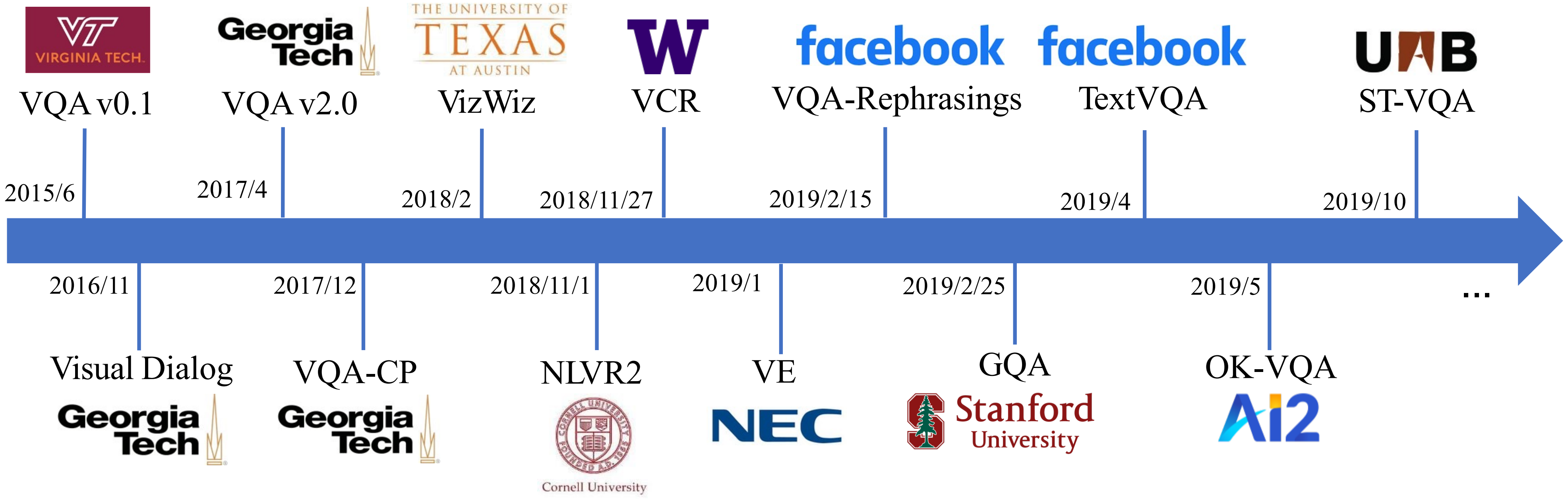
Problem Overview (2): VQA and Visual Reasoning

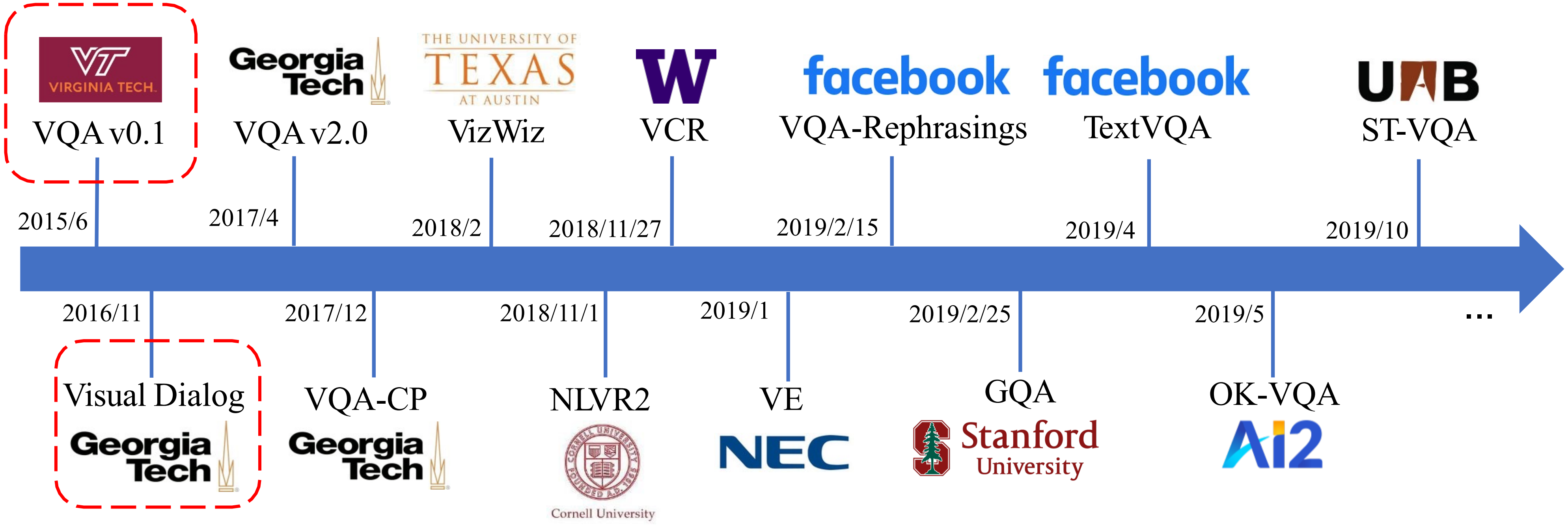
- How to train a smart multi-modal AI system that can both see and talk?

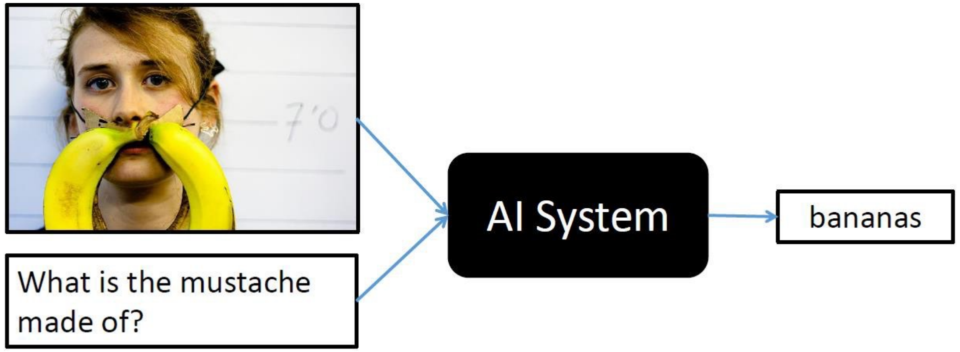


Problem Overview (2): VQA and Visual Reasoning

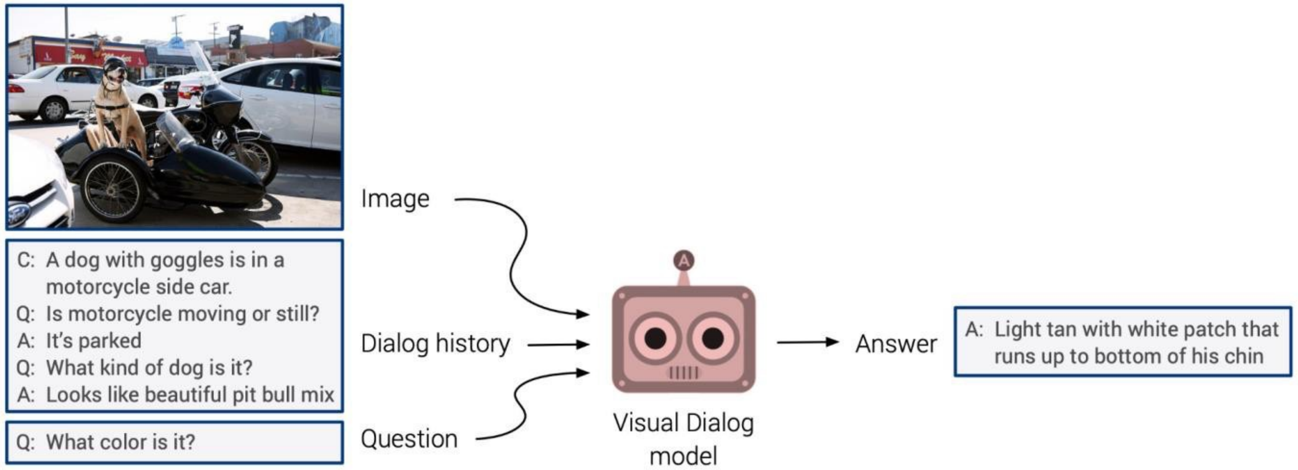
- Large-scale annotated datasets have driven tremendous progress in this field







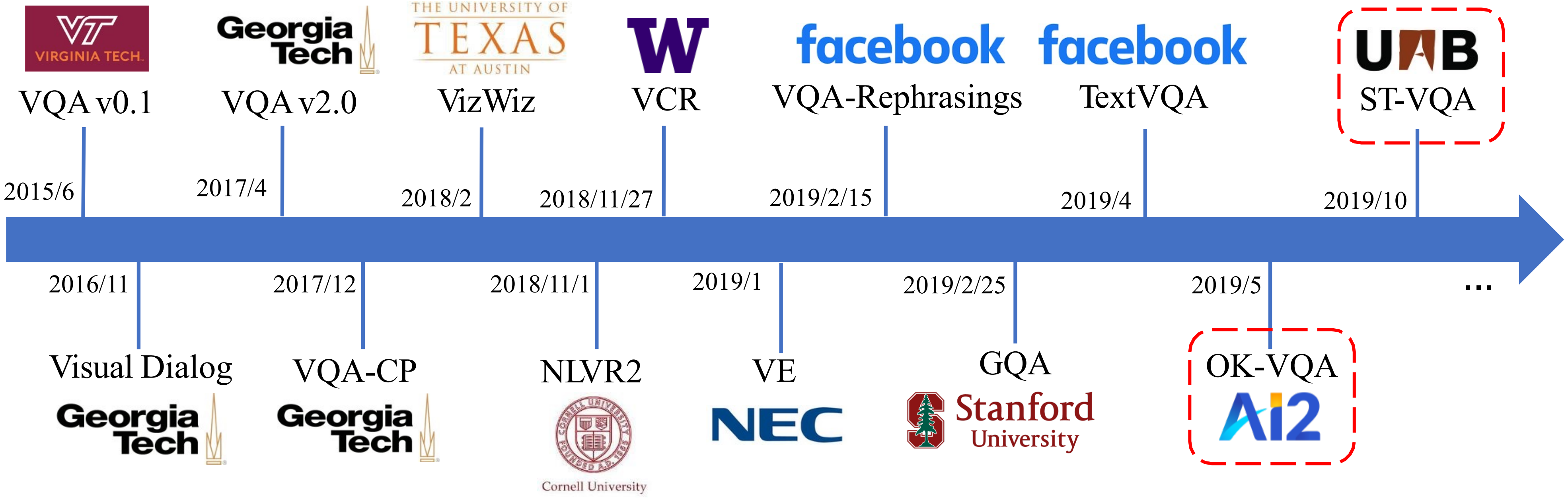
VQA



Visual Dialog

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

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



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

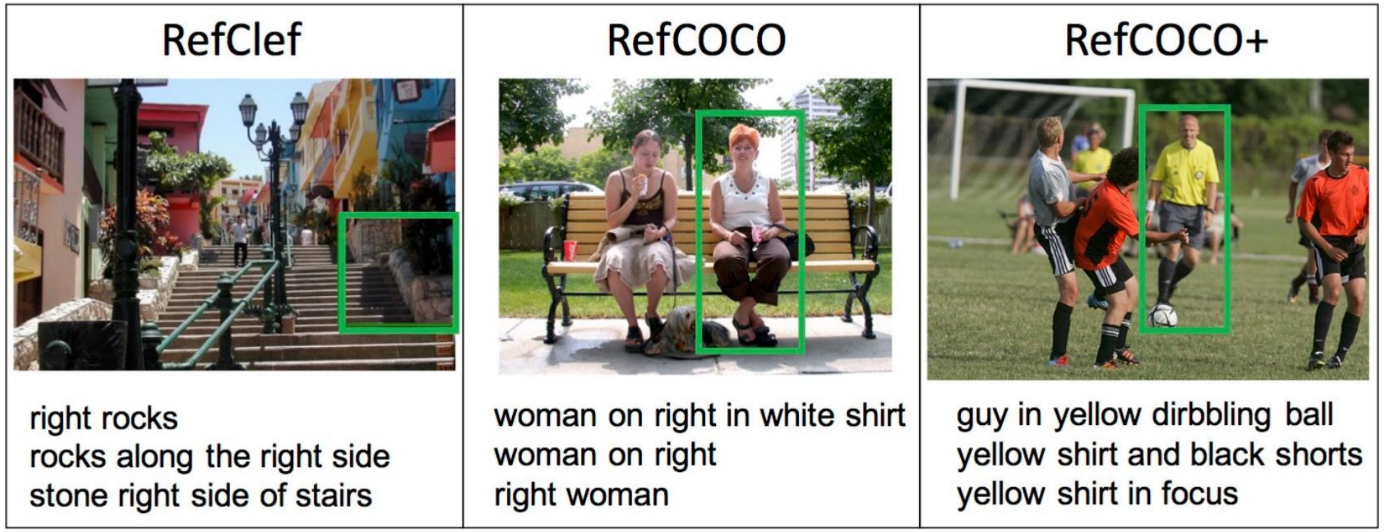
Q: What does the red sign say?
A: Stop

Scene Text VQA

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

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

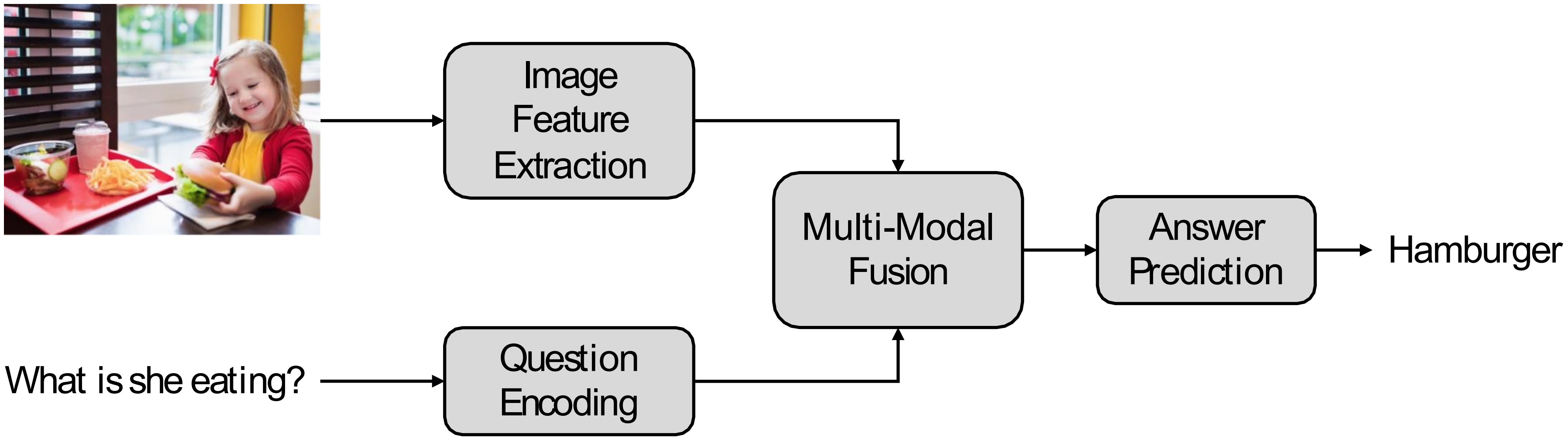
Beyond VQA: Visual Grounding

- PhraseCut: Language-based image segmentation

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

Approach Overview

- How a typical system looks like

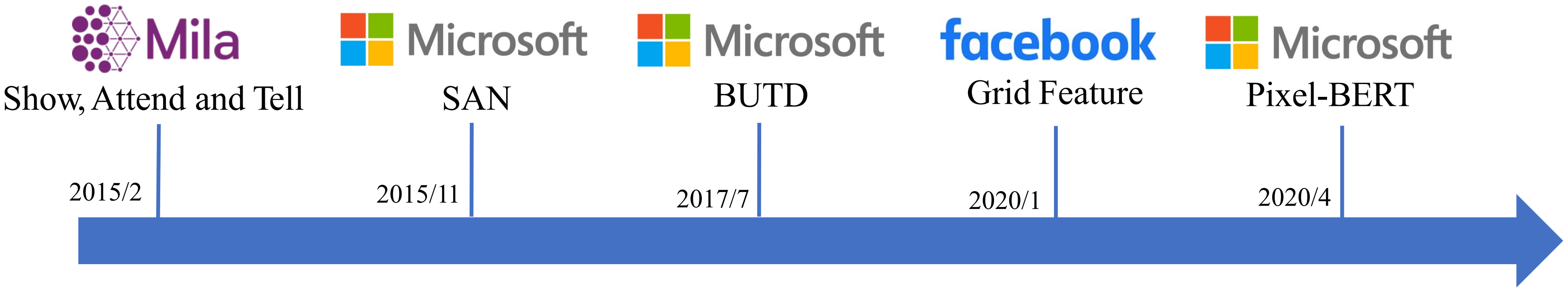


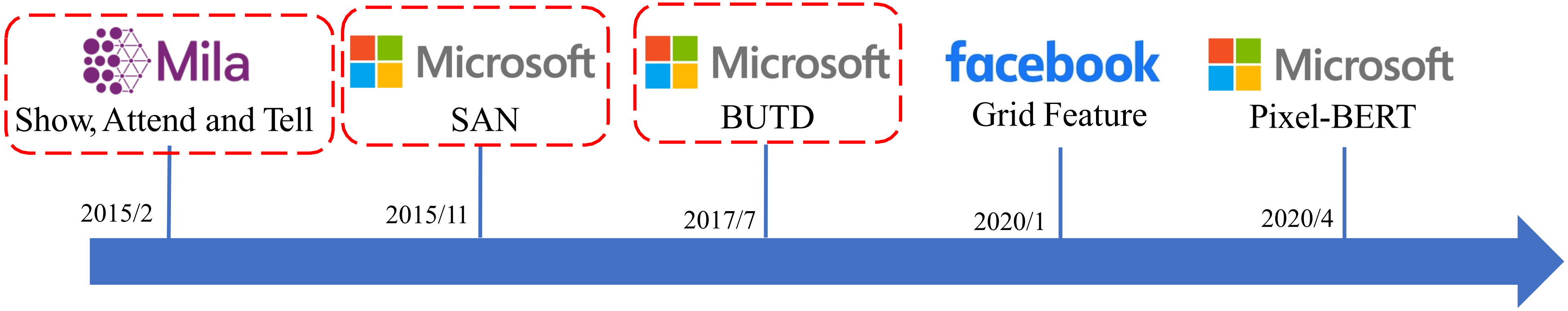
Research Challenges & Opportunities

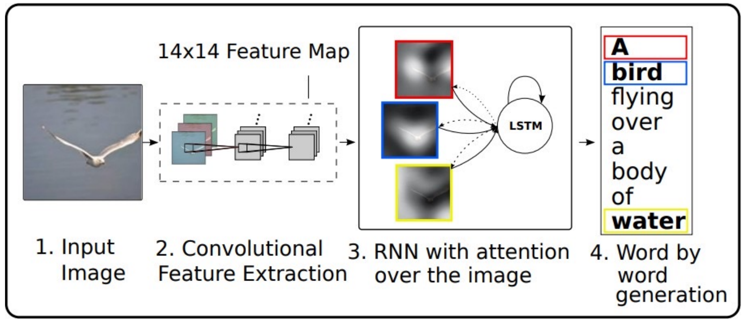
- Better image feature preparation
- Enhanced multimodal fusion
  - Bilinear pooling: how to fuse two vectors into one
  - Multimodal alignment: *cross-modal* attention
  - Incorporation of object relations: *intra-modal* self-attention, graph attention
  - Multi-step reasoning
- Neural module networks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

Better Image Feature Preparation

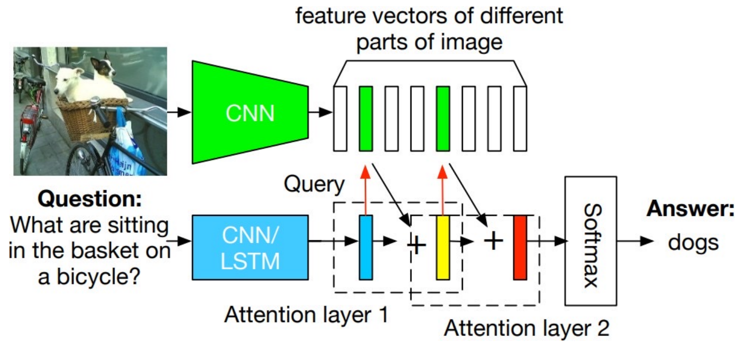
- From *grid* features to *region* features, and to *grid* features again



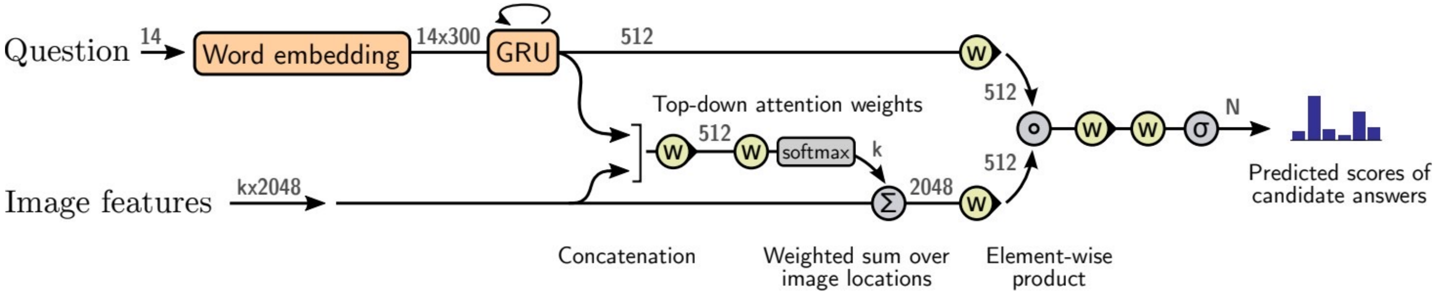




Show, Attend and Tell



Stacked Attention Network



2017 VQA Challenge Winner

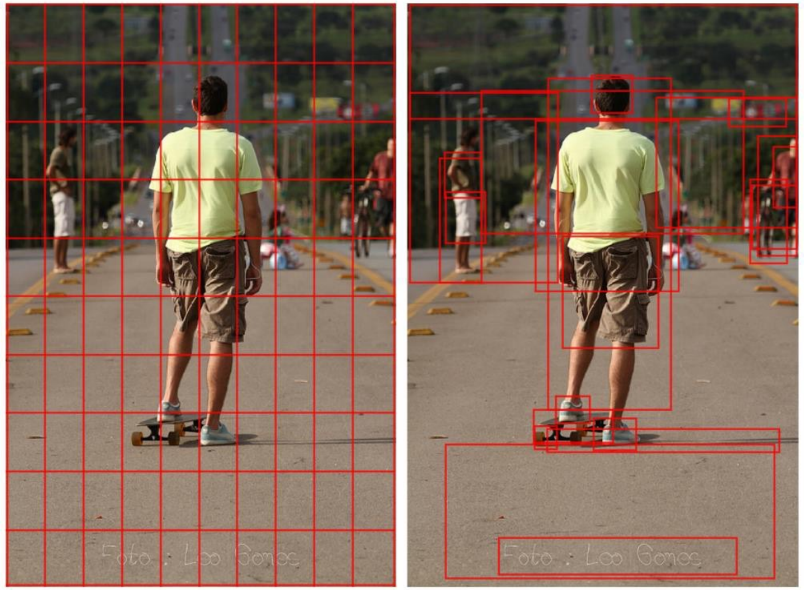
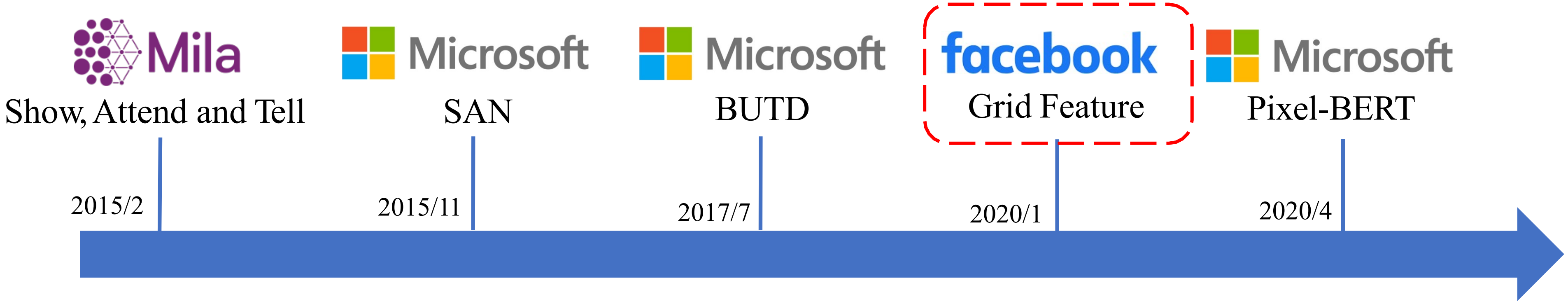
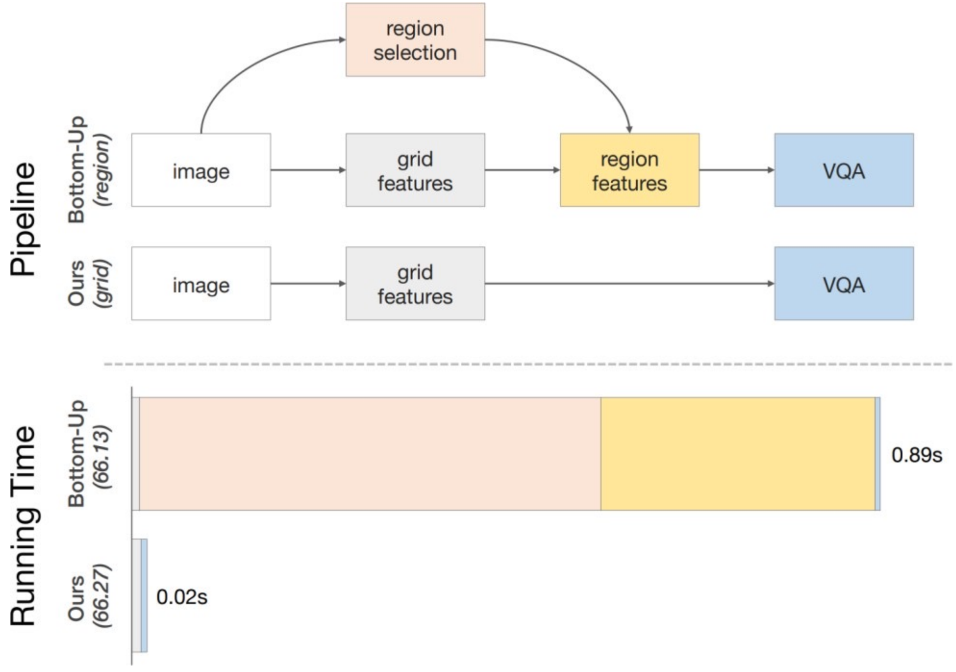
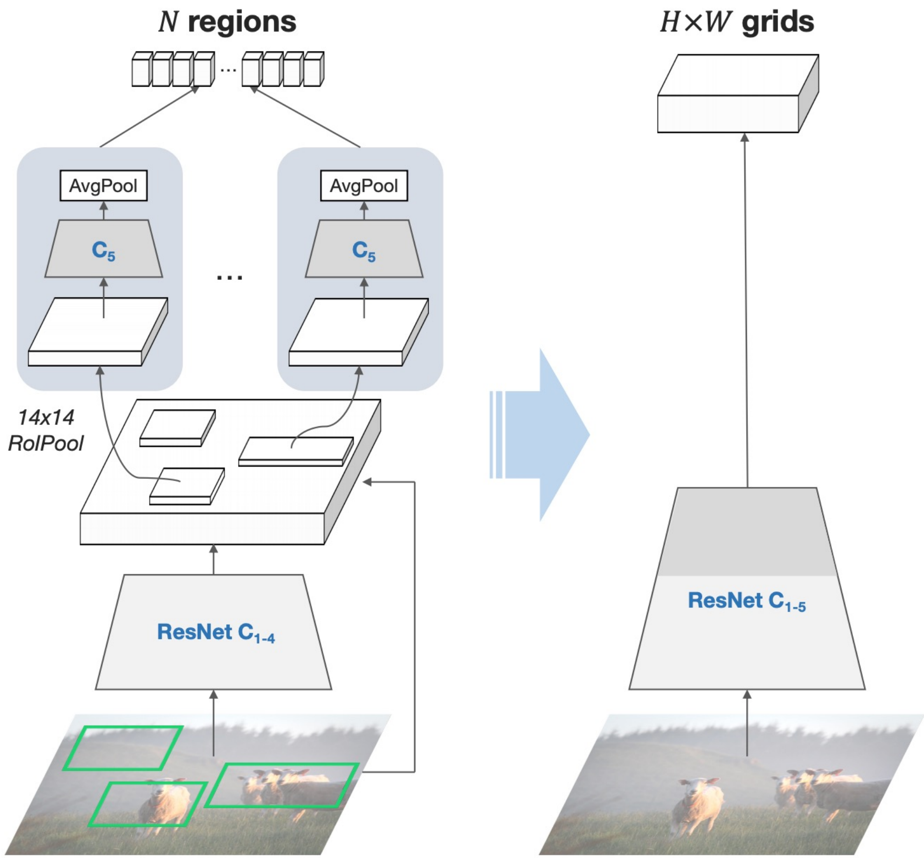


Figure 1. Typically, attention models operate on CNN features corresponding to a uniform grid of equally-sized image regions (left). Our approach enables attention to be calculated at the level of objects and other salient image regions (right).

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

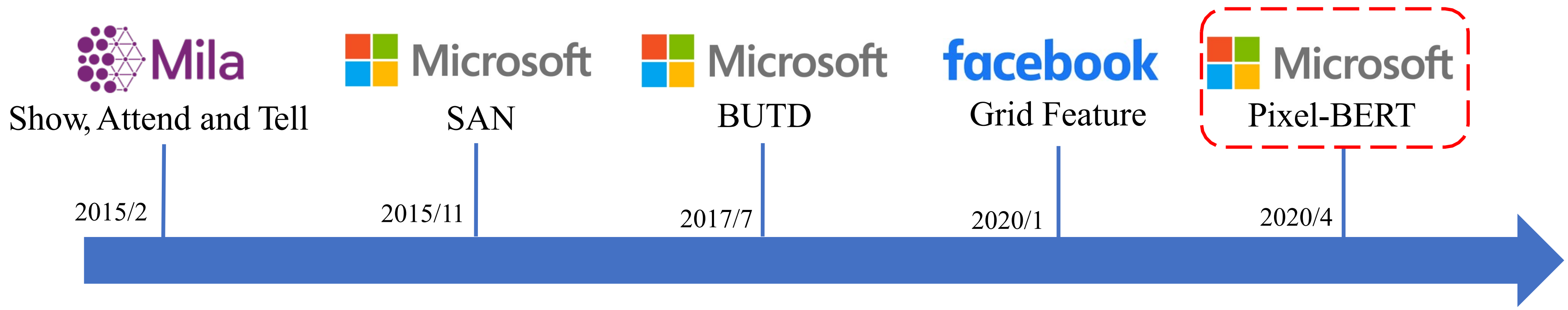


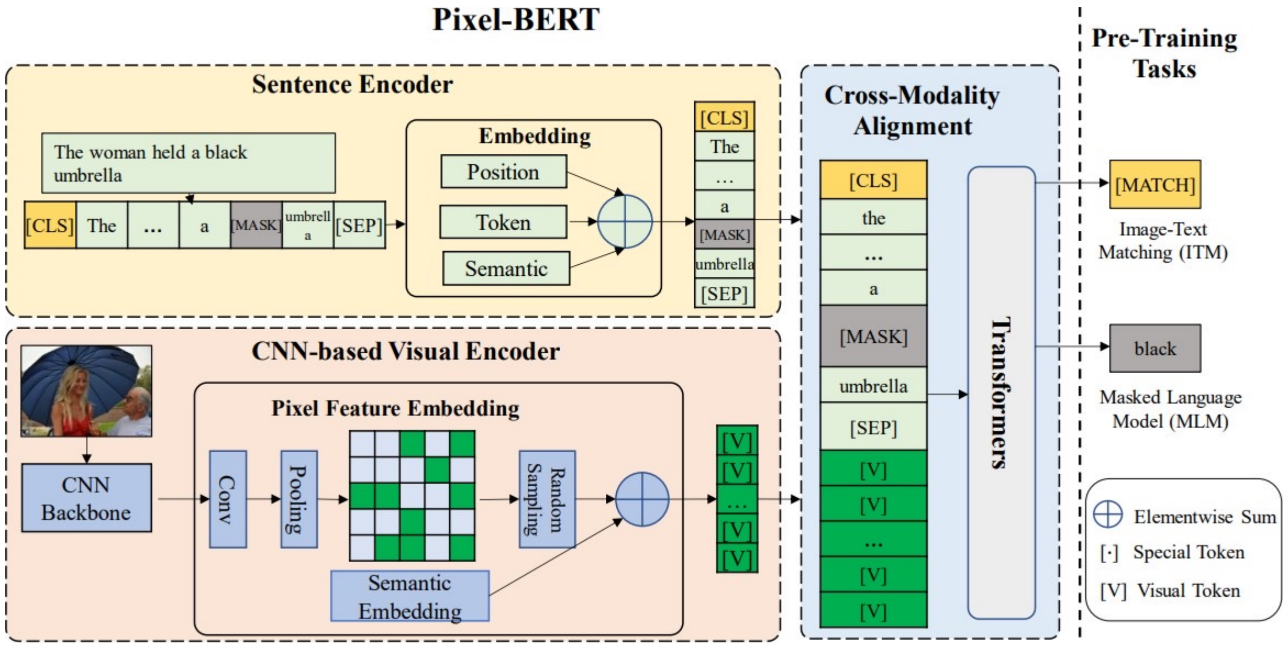




In Defense of Grid Features for VQA

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





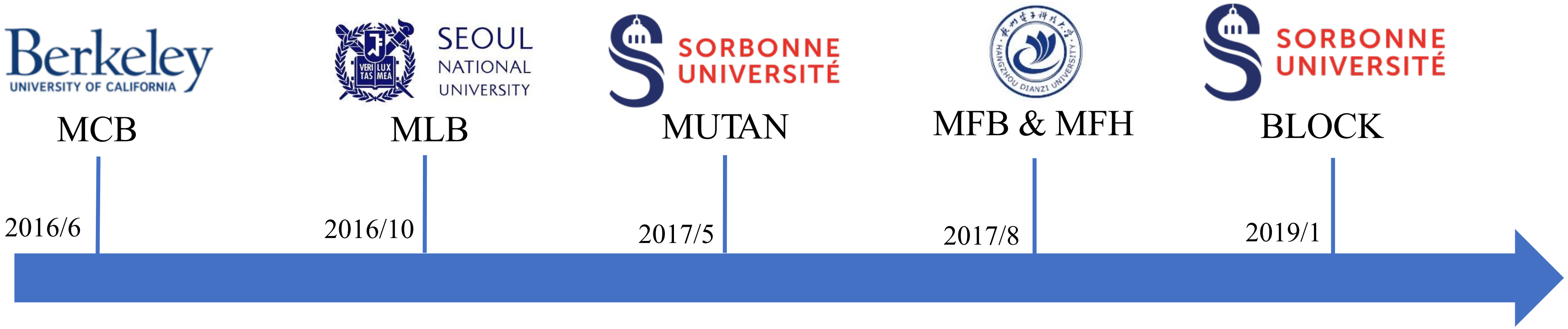
Model	test-dev	test-std
MUTAN[5]	60.17	-
BUTD[2]	65.32	65.67
ViLBERT[21]	70.55	70.92
VisualBERT[19]	70.80	71.00
VLBERT[29]	71.79	72.22
LXMERT[33]	72.42	72.54
UNITER[6]	72.27	72.46
Pixel-BERT (r50)	71.35	71.42
Pixel-BERT (x152)	**74.45**	**74.55**

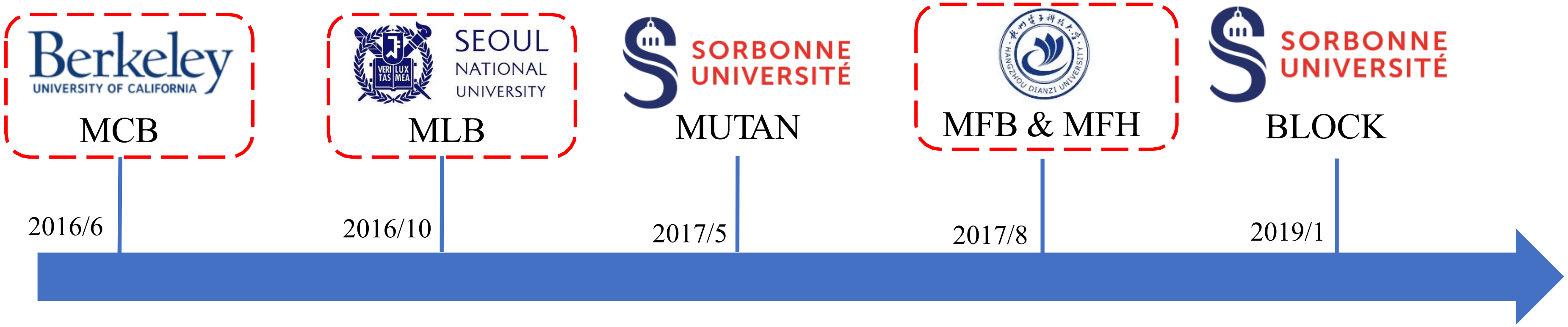
Table 2. Evaluation of Pixel-BERT with other methods on VQA.

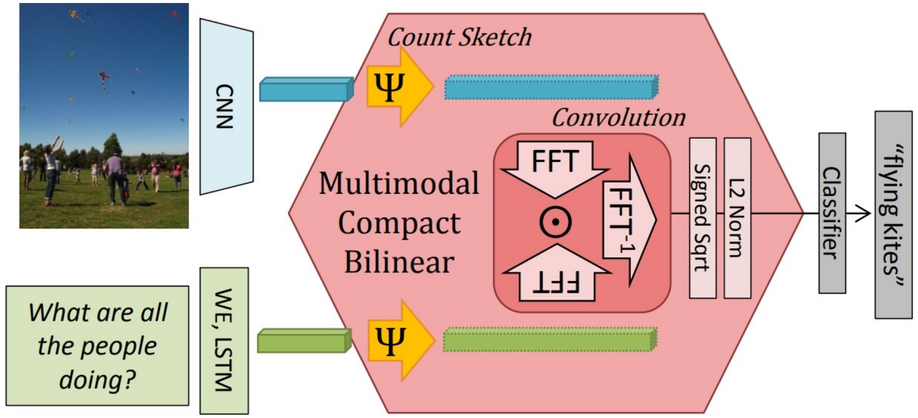
[1] Pixel-BERT: Aligning Image Pixels with Text by Deep Multi-Modal Transformers, 2020

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







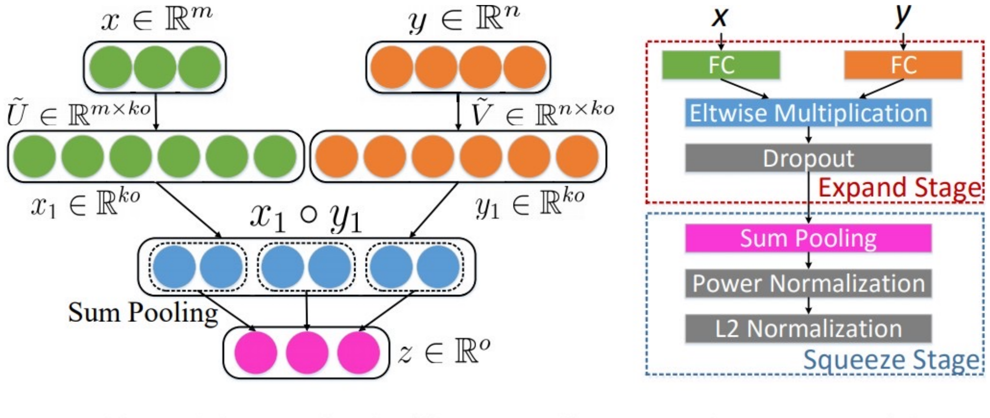
Multimodal Compact Bilinear Pooling

2016 VQA Challenge Winner

However, the feature after FFT is very high dimensional.

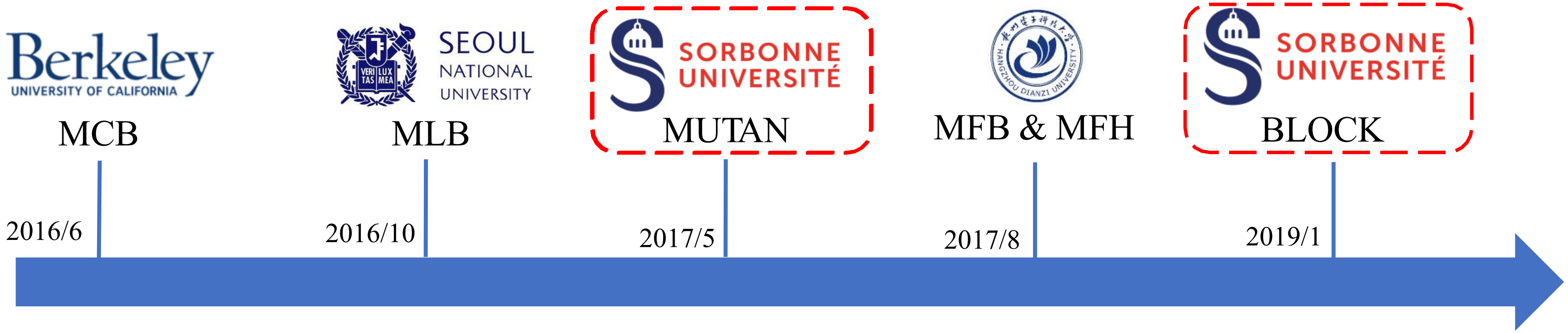
$$\mathbf{f} = \mathbf{P}^T(\mathbf{U}^T\mathbf{x} \circ \mathbf{V}^T\mathbf{y}) + \mathbf{b}$$

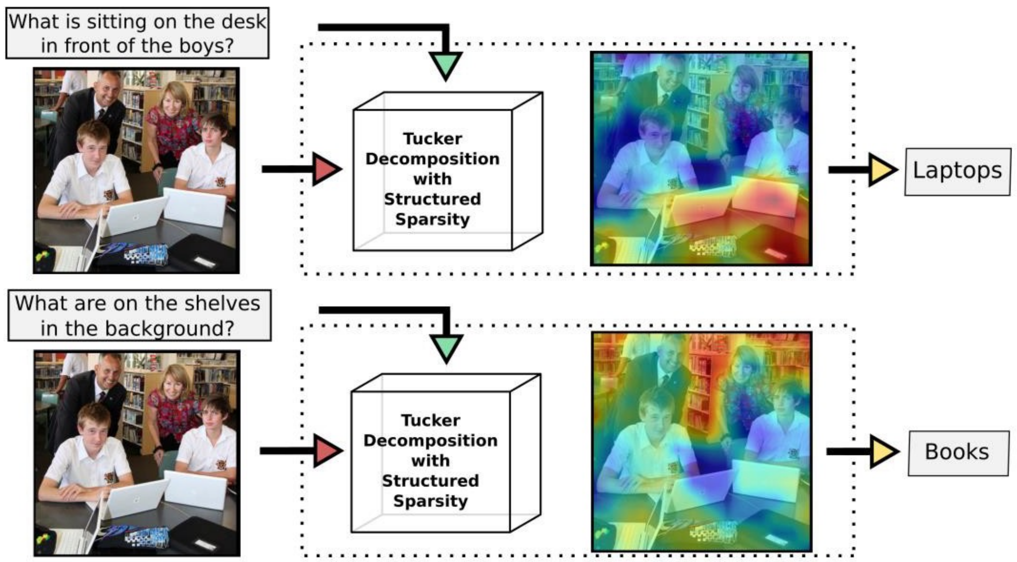
Multimodal Low-rank Bilinear Pooling



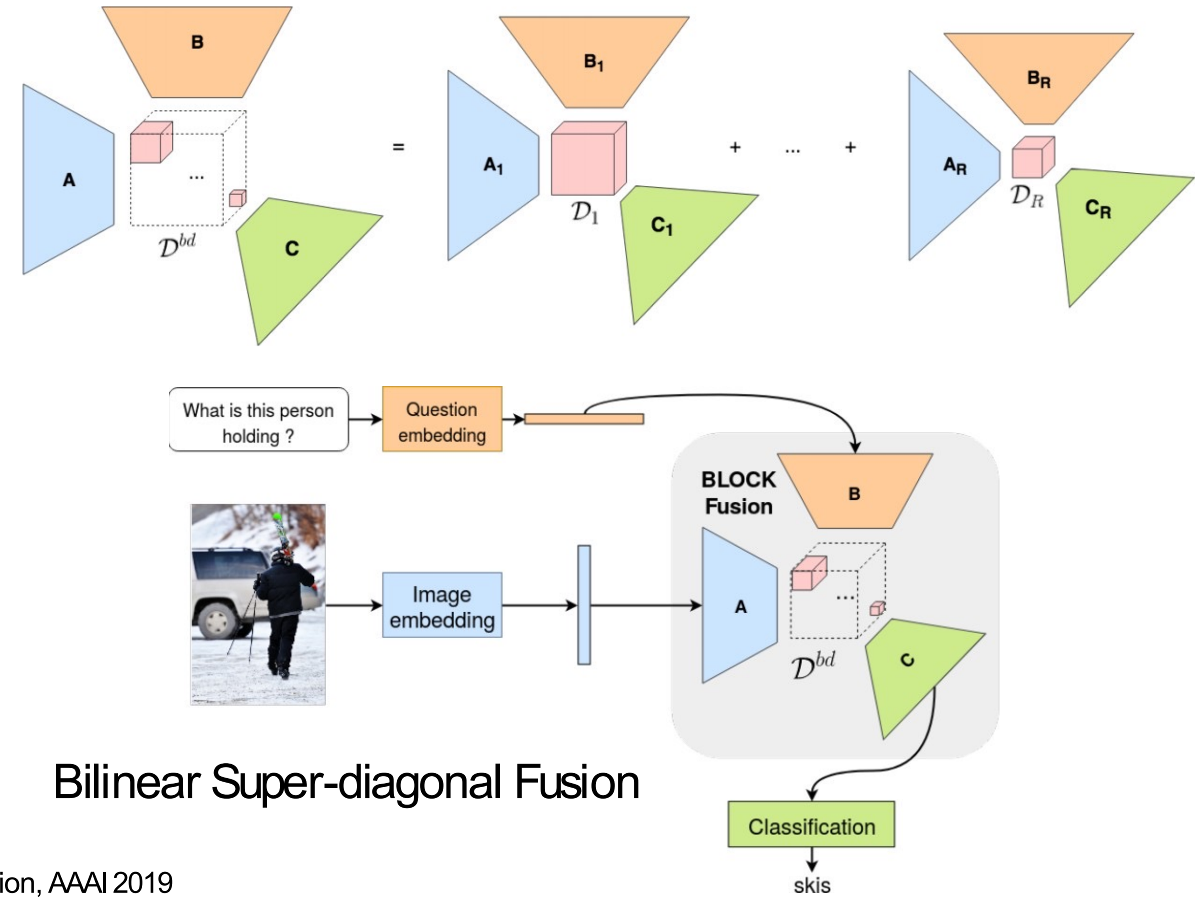
(a) Multi-modal Factorized Bilinear Pooling (b) MFB module

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





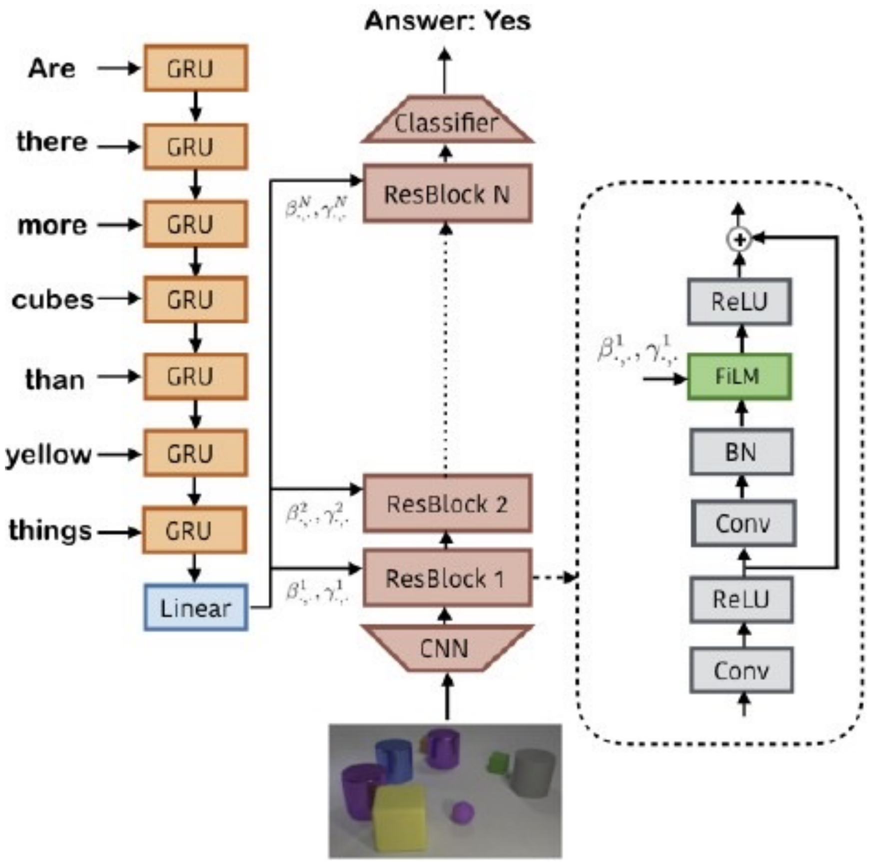
Multimodal Tucker Fusion



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

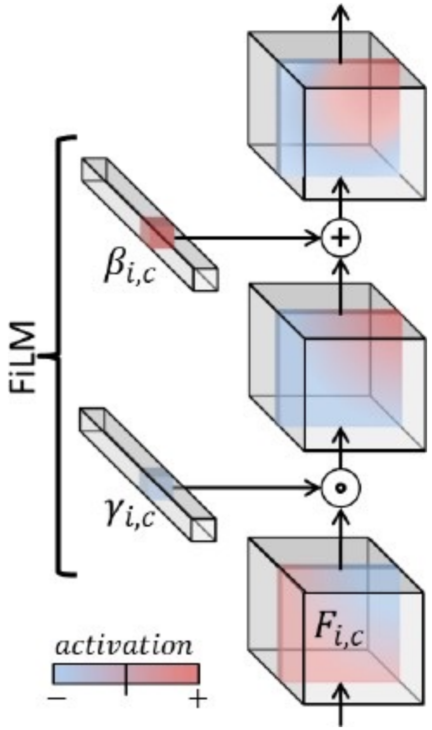
FiLM: Feature-wise Linear Modulation



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

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

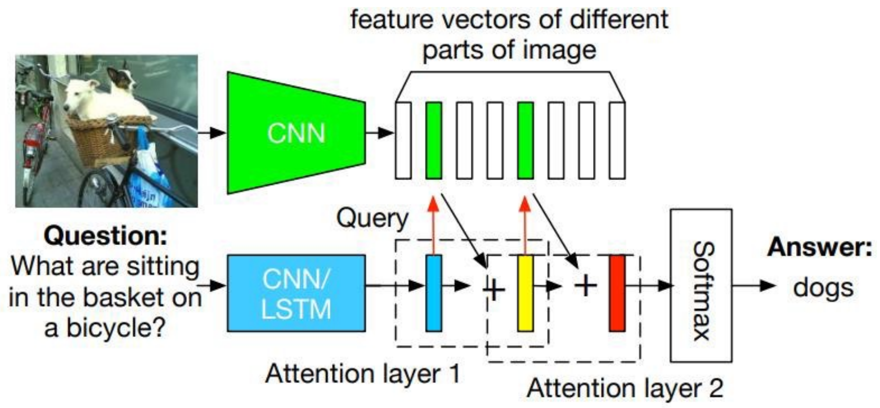
Something similar to conditional batch normalization



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

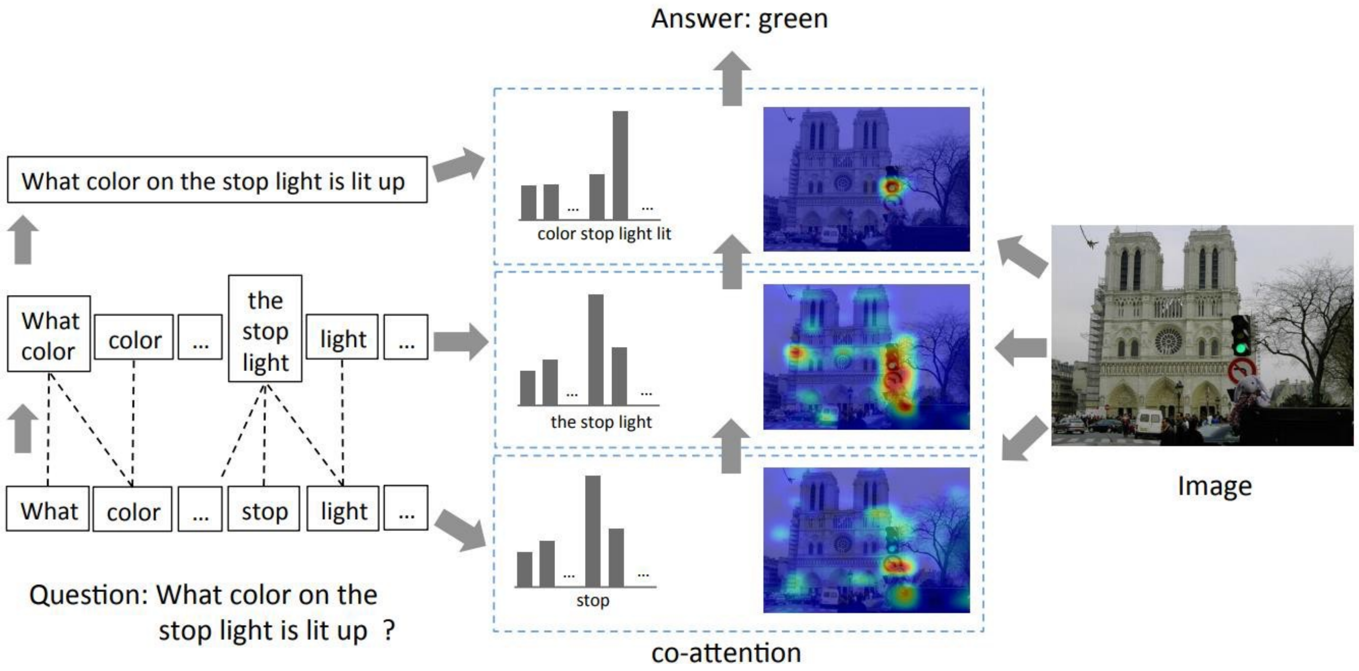
Multimodal Alignment

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



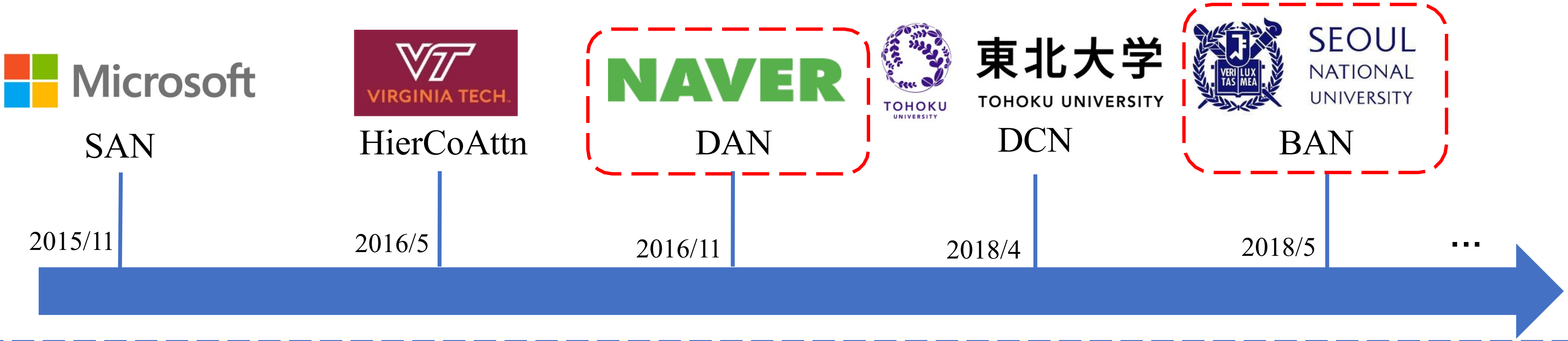
(a) Stacked Attention Network for Image QA

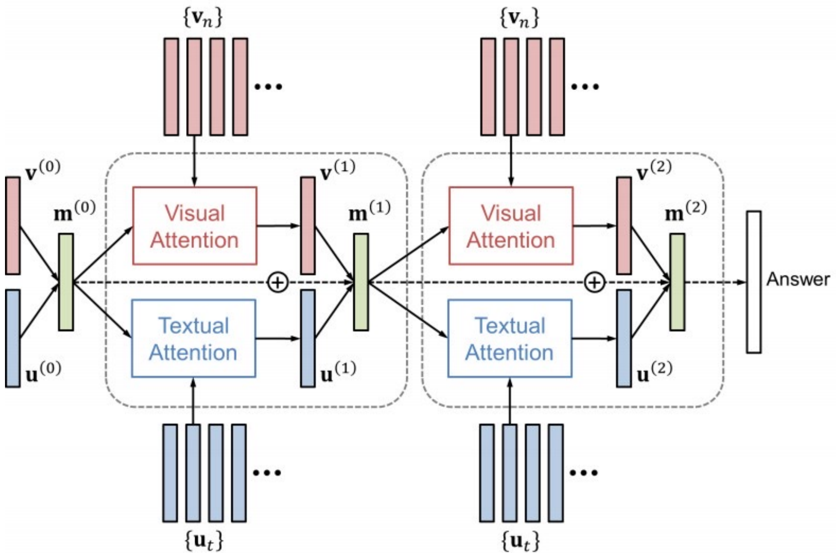
(b) Visualization of the learned multiple attention layers.

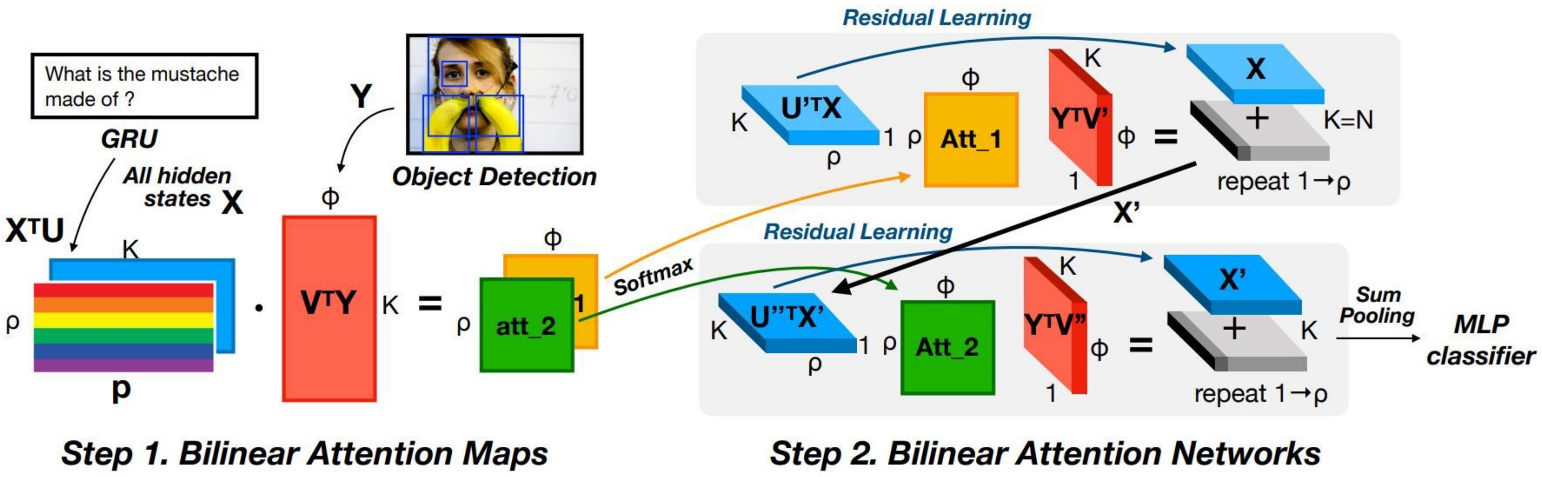


Parallel Co-attention and Alternative Co-attention

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

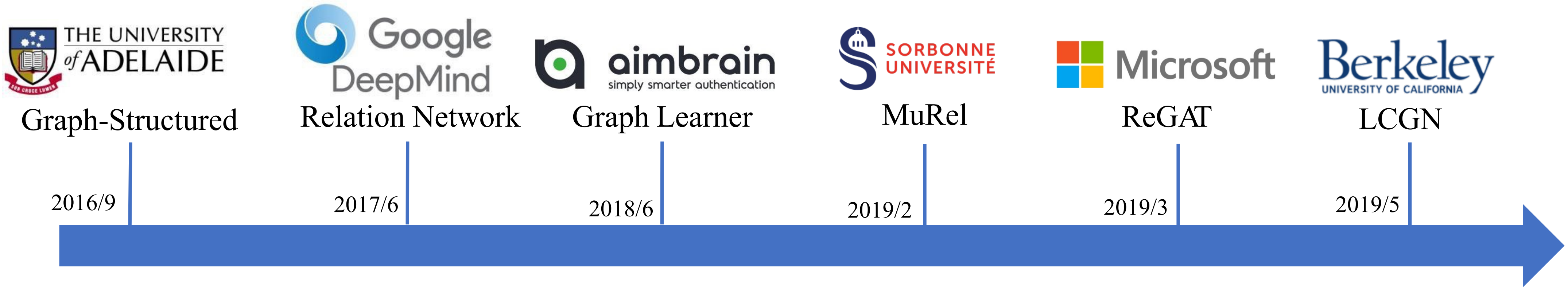
[2018 VQA Challenge Runner-Up](#)

- 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 — 2016/9

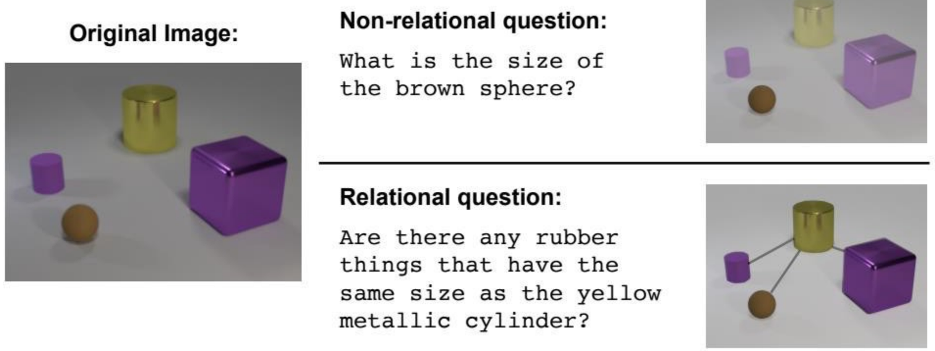
Relation Network — 2017/6

Graph Learner — 2018/6

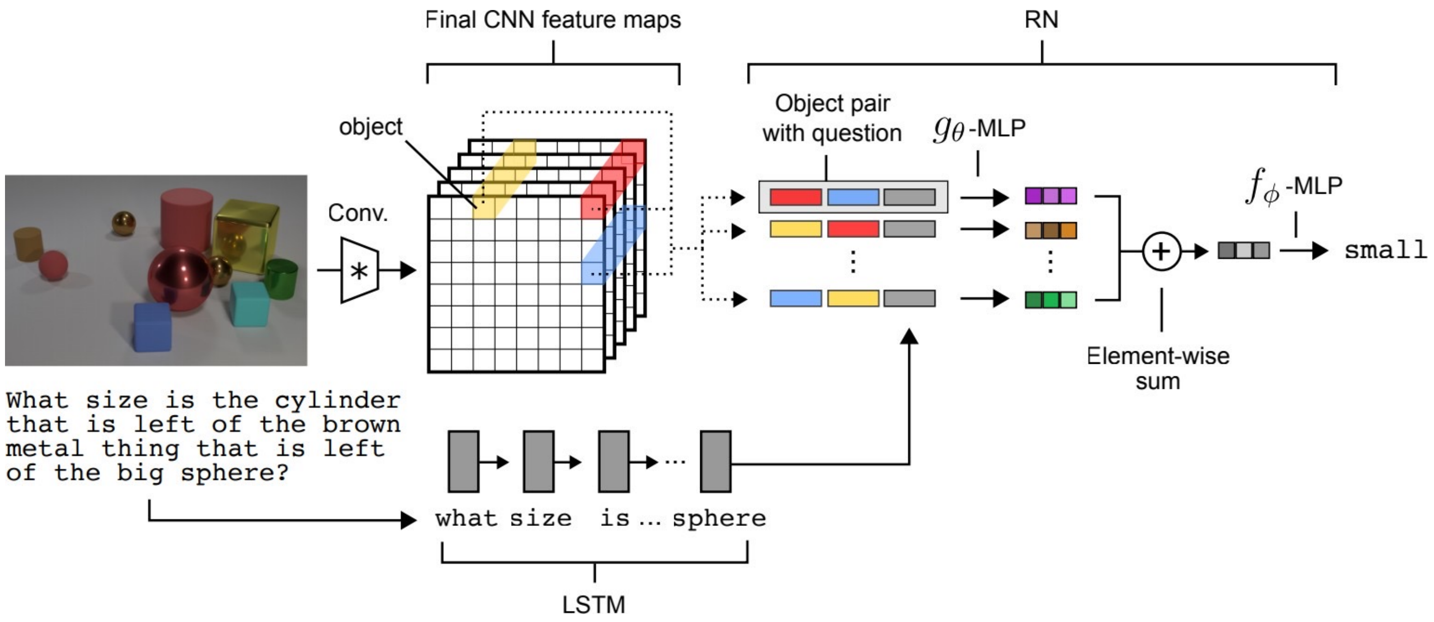
MuRel — 2019/2

ReGAT — 2019/3

LCGN — 2019/5

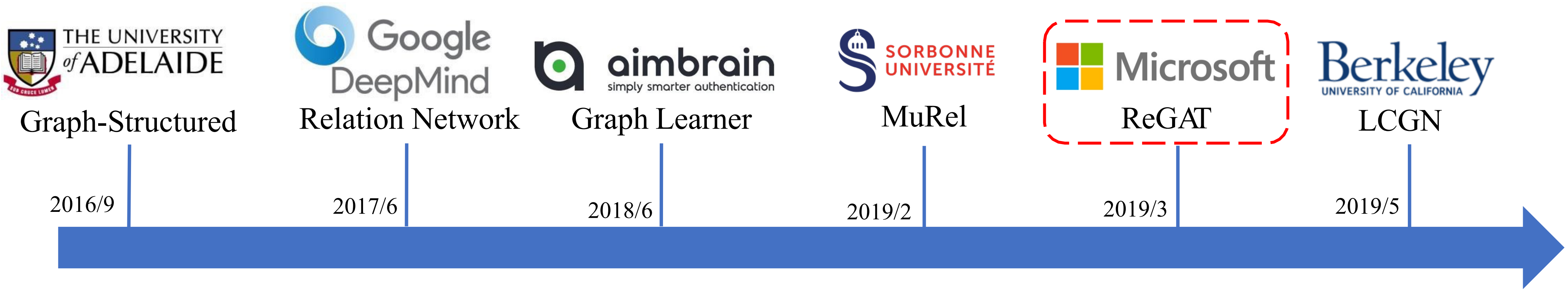


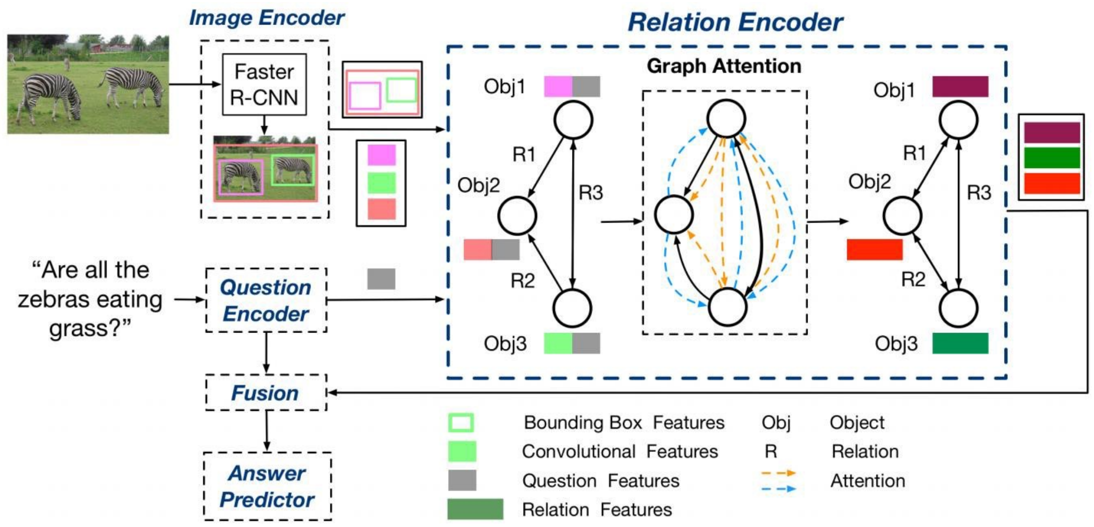
$$\mathrm{RN}(O) = f_\phi\left(\sum_{i,j} g_\theta(o_i, o_j)\right)$$



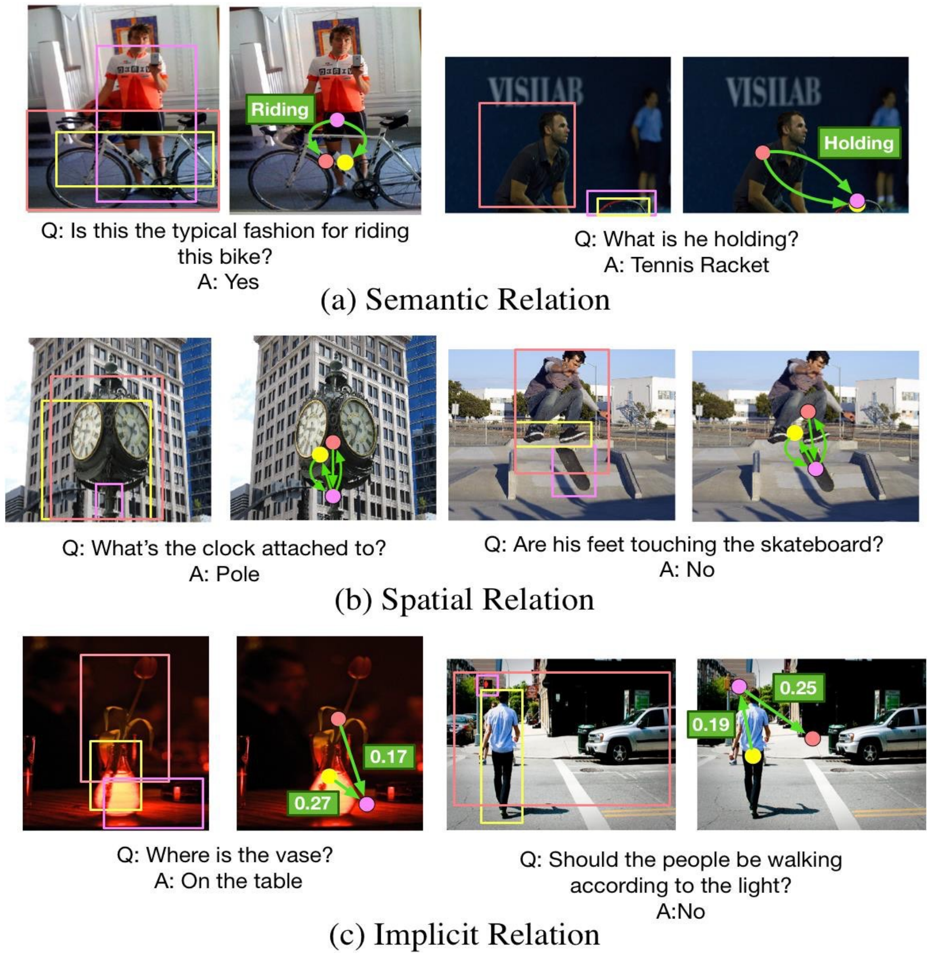
Relational Network: A fully-connected graph is constructed

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





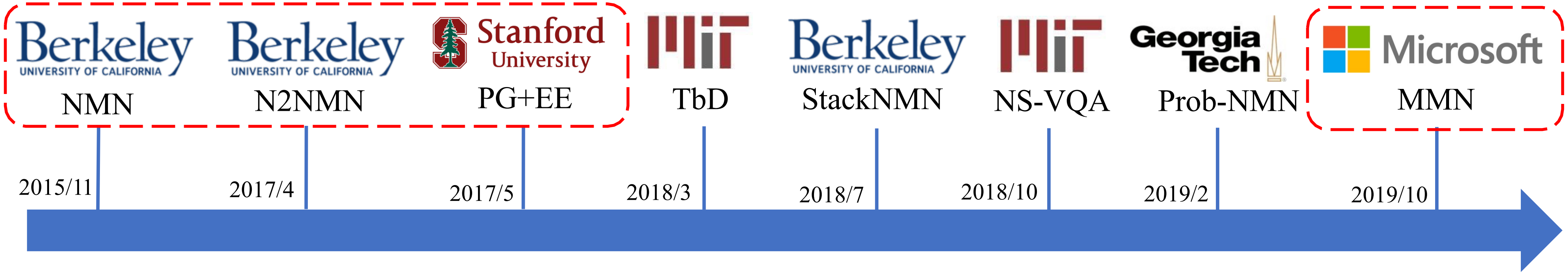
- *Explicit* Relation: Semantic & Spatial relation
- *Implicit* Relation: Learned dynamically during training



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

Neural Module Network (NMN)

- All the previously mentioned work can be considered as *Monolithic Network*
- Design *Neural Modules* for compositional visual reasoning – very “human like”



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, ECCV2018
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

Consider a compositional model

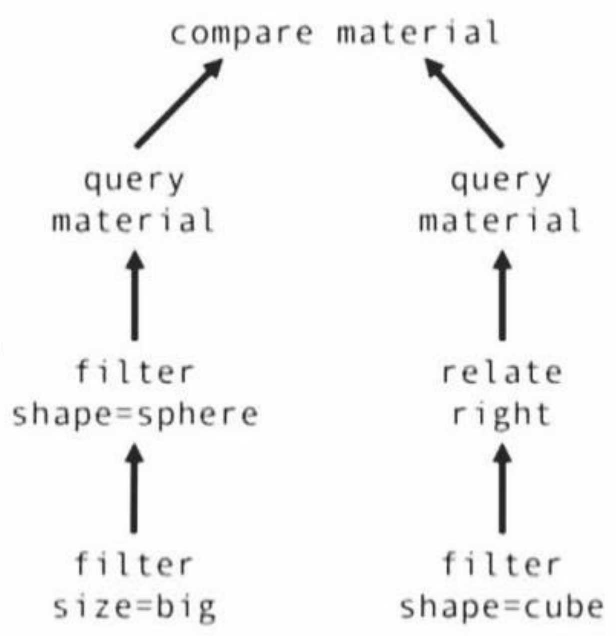
Q: How many spheres are the left of the big sphere and the same color as the small rubber cylinder?

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

Q: Is the big sphere the same material as the thing on the right of the cube?

Common operations

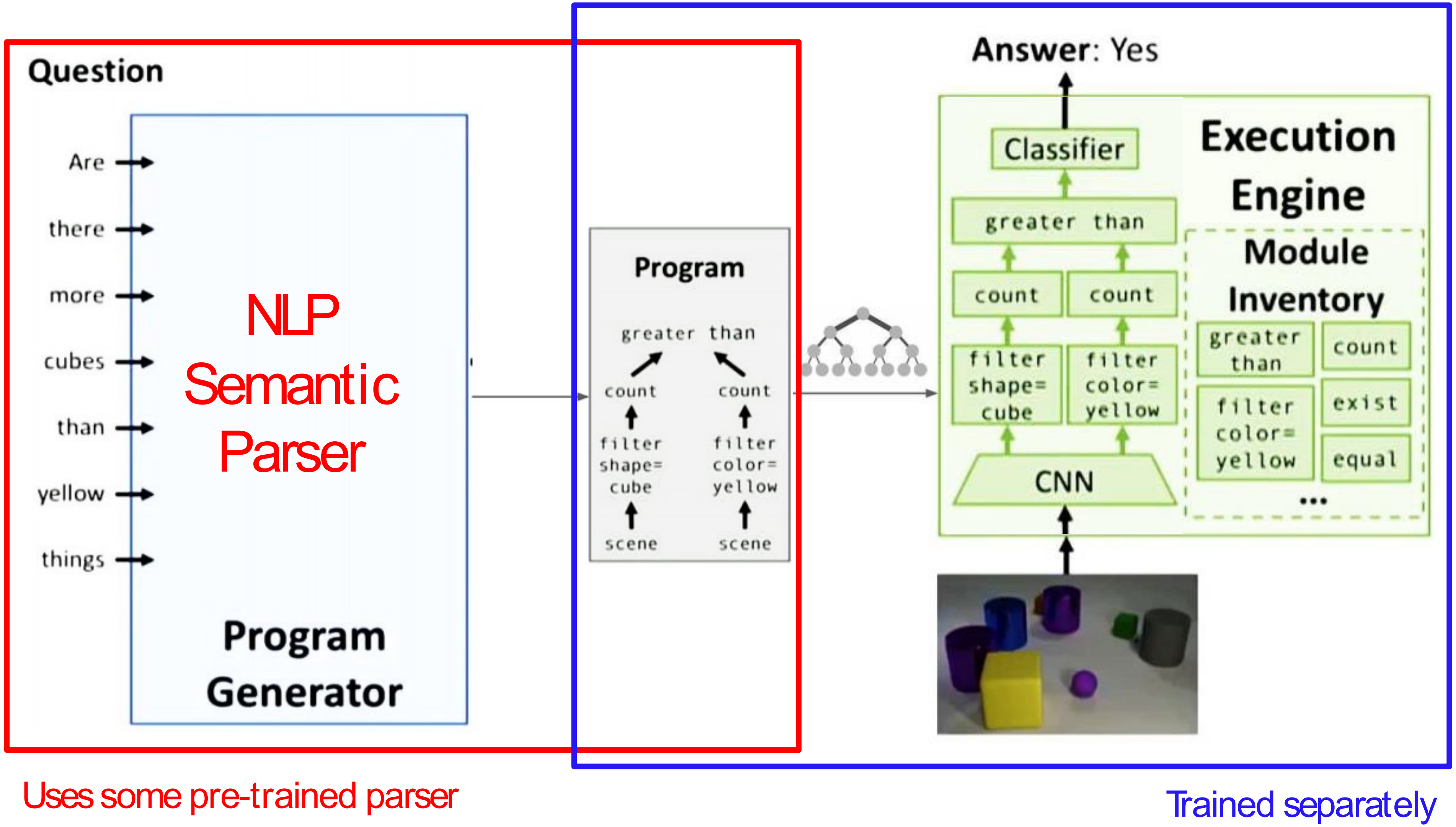
Attributes identification
Counting objects
Comparisons
Spatial relationships
Logical operations



Network architecture corresponding to the third question

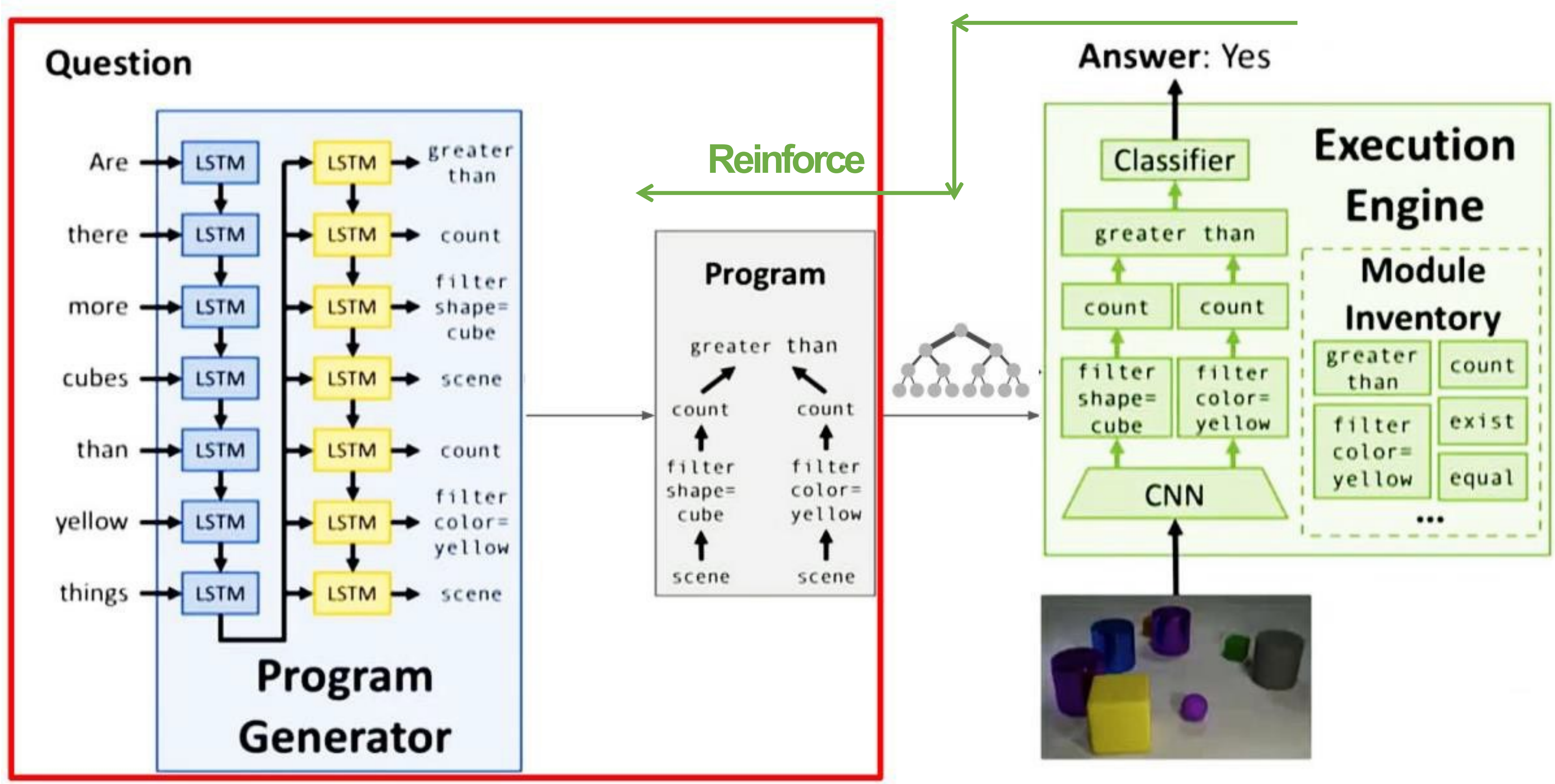
[1] Deep Compositional Question Answering with Neural Module Networks, CVPR, 2016

Overview of the NMN approach

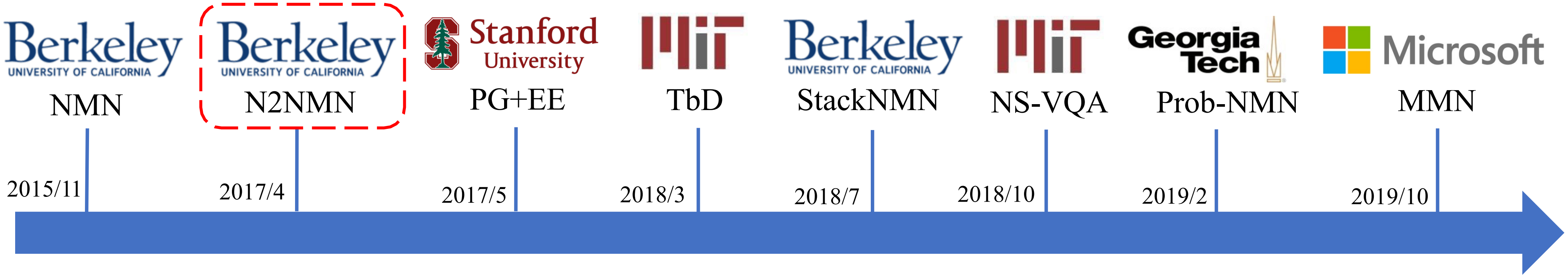


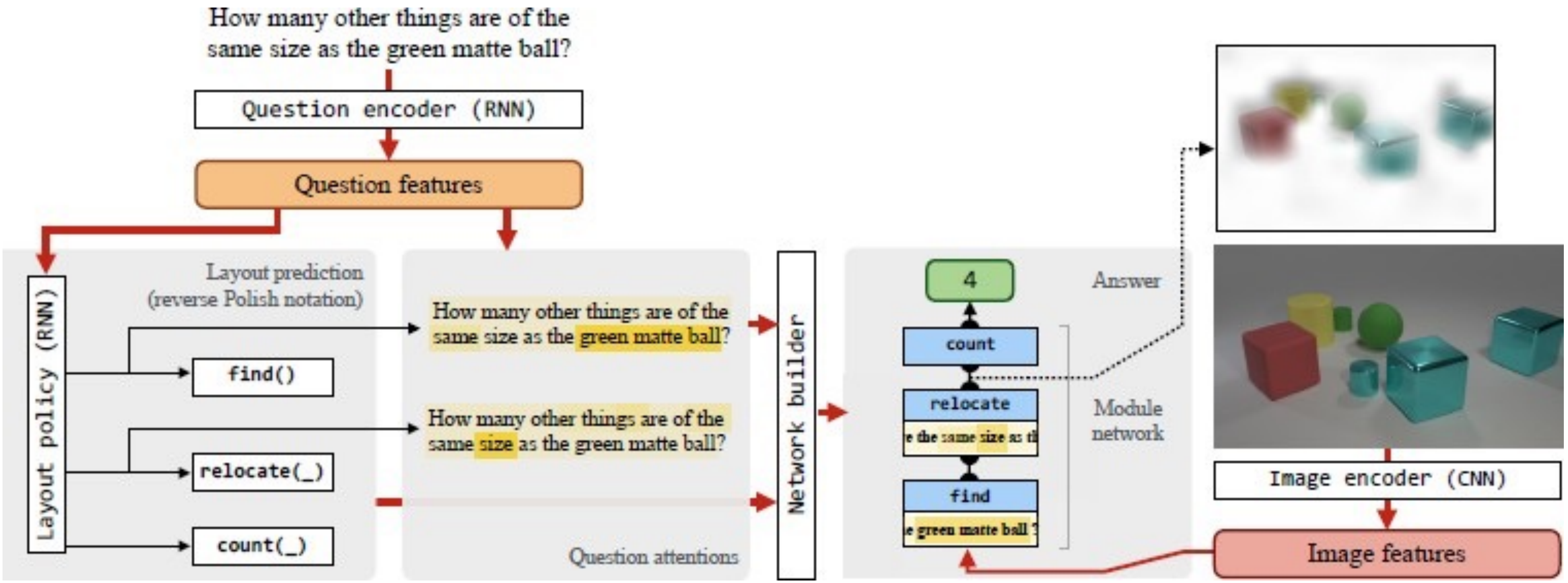
[1] Deep Compositional Question Answering with Neural Module Networks, CVPR, 2016

Inferring and Executing Programs



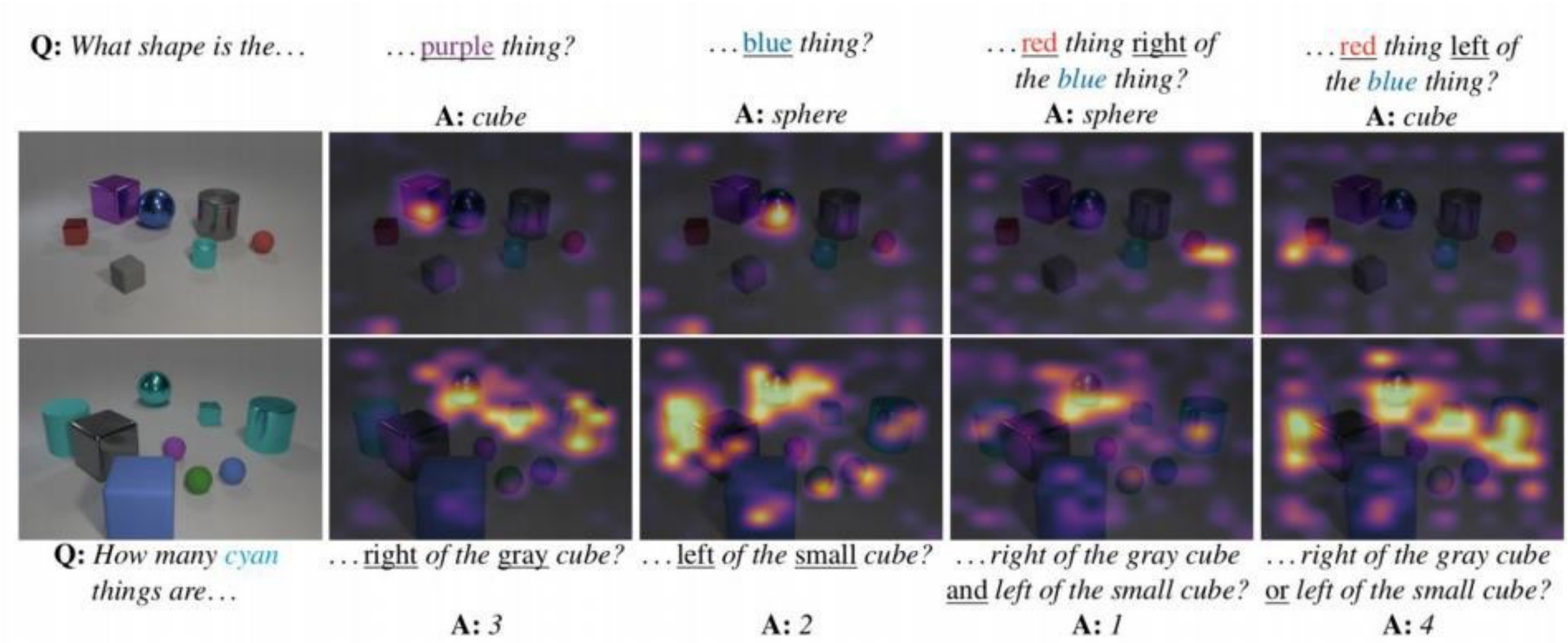
[1] Inferring and Executing Programs for Visual Reasoning, ICCV, 2017





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

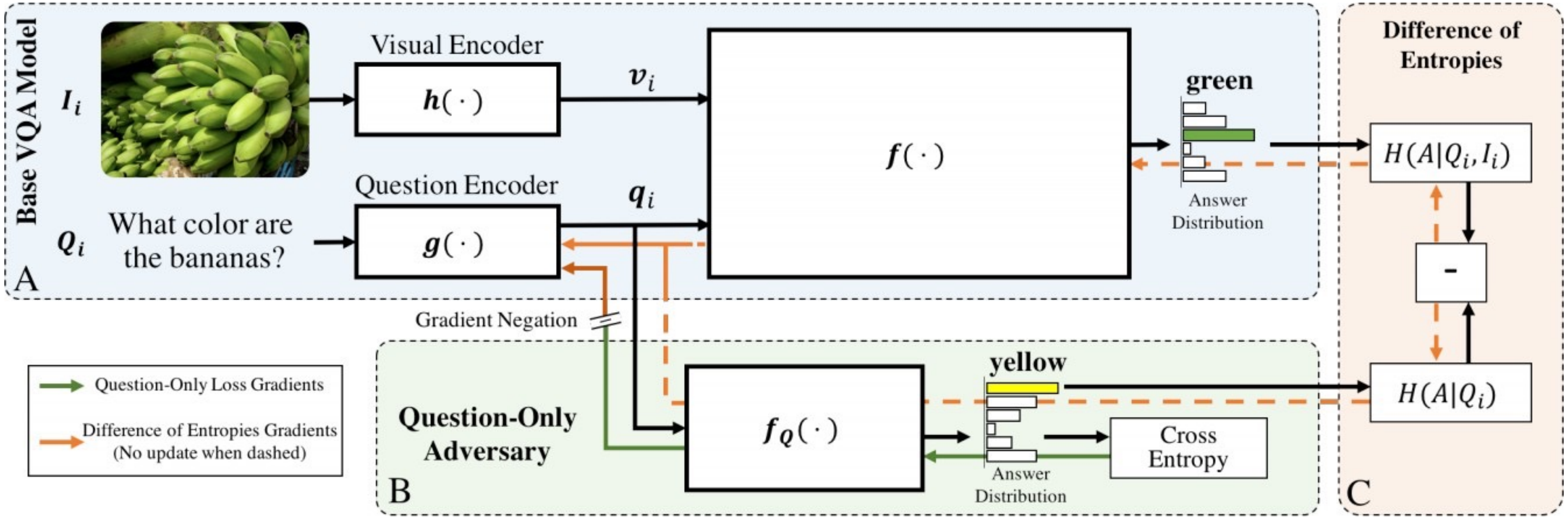
What do the modules learn?



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

Robust VQA: an example

- Overcoming language prior with adversarial regularization



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

The University of Texas at Austin
Electrical and Computer Engineering
Cockrell School of Engineering