Visually Grounded Language Understanding and Generation

Jiasen Lu

Georgia Institute of Technology

Image Credit: COCO dataset

Image Credit: COCO dataset

Image Credit: COCO dataset

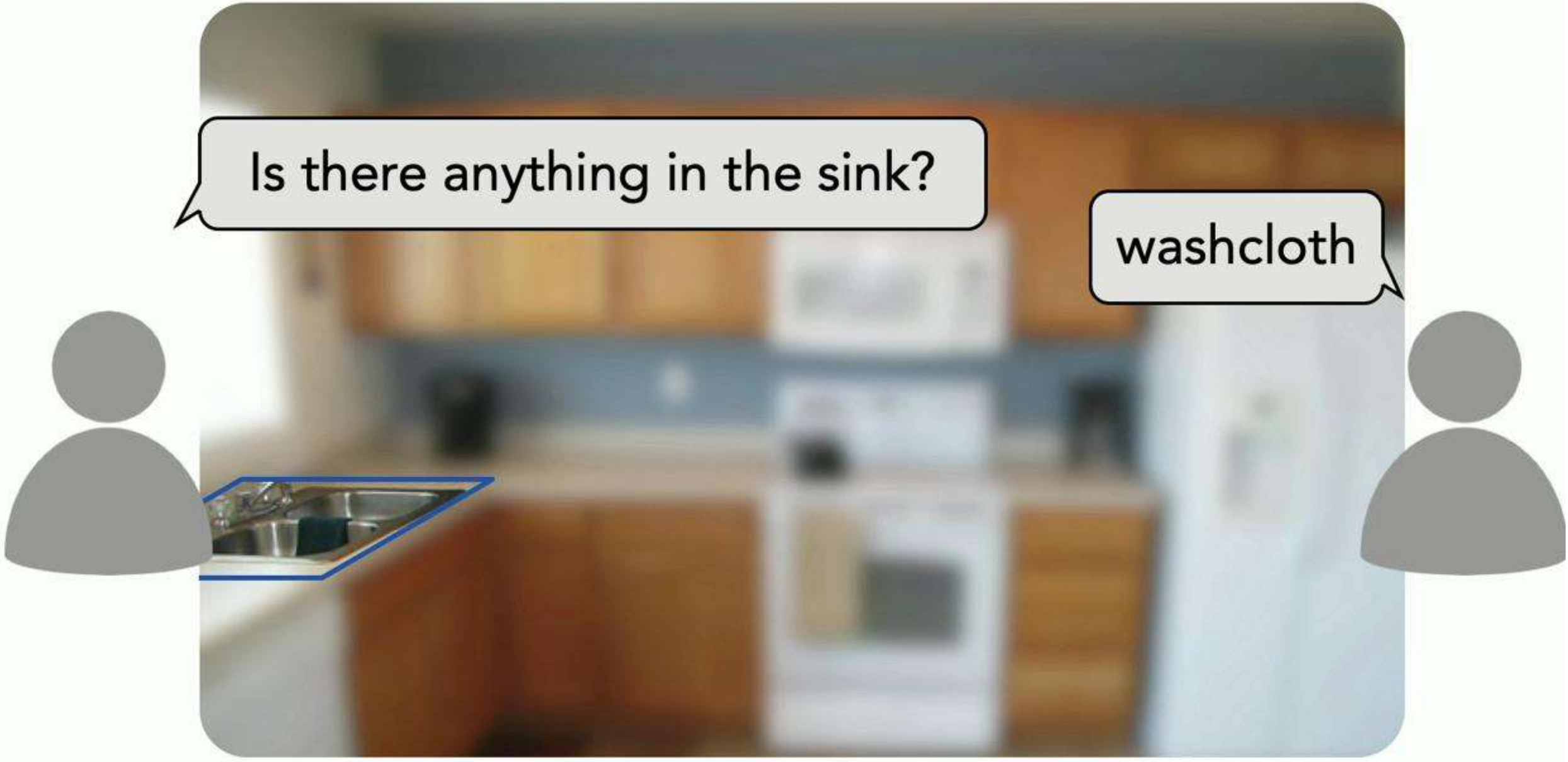


Image Credit: COCO dataset

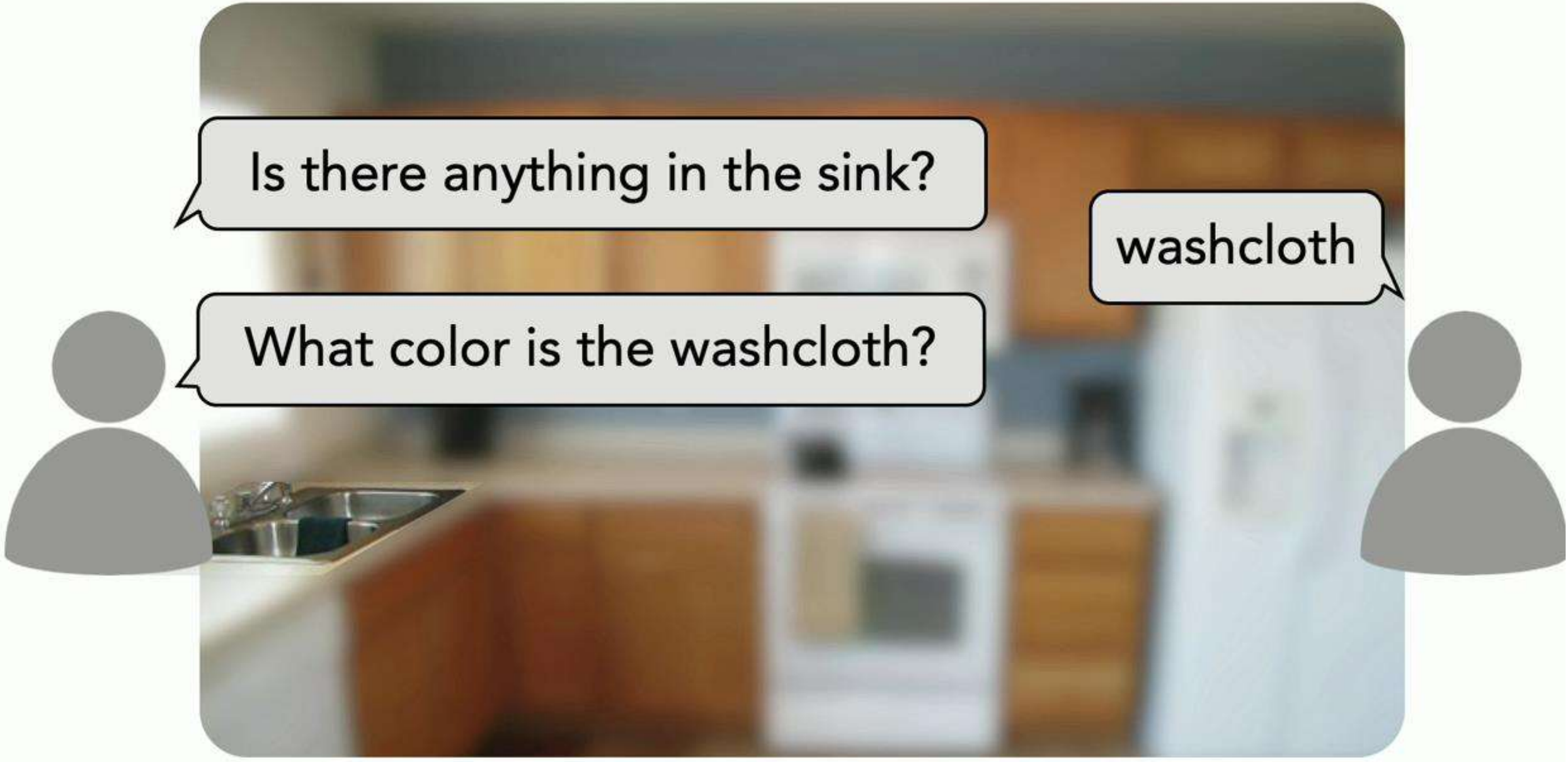


Image Credit: COCO dataset

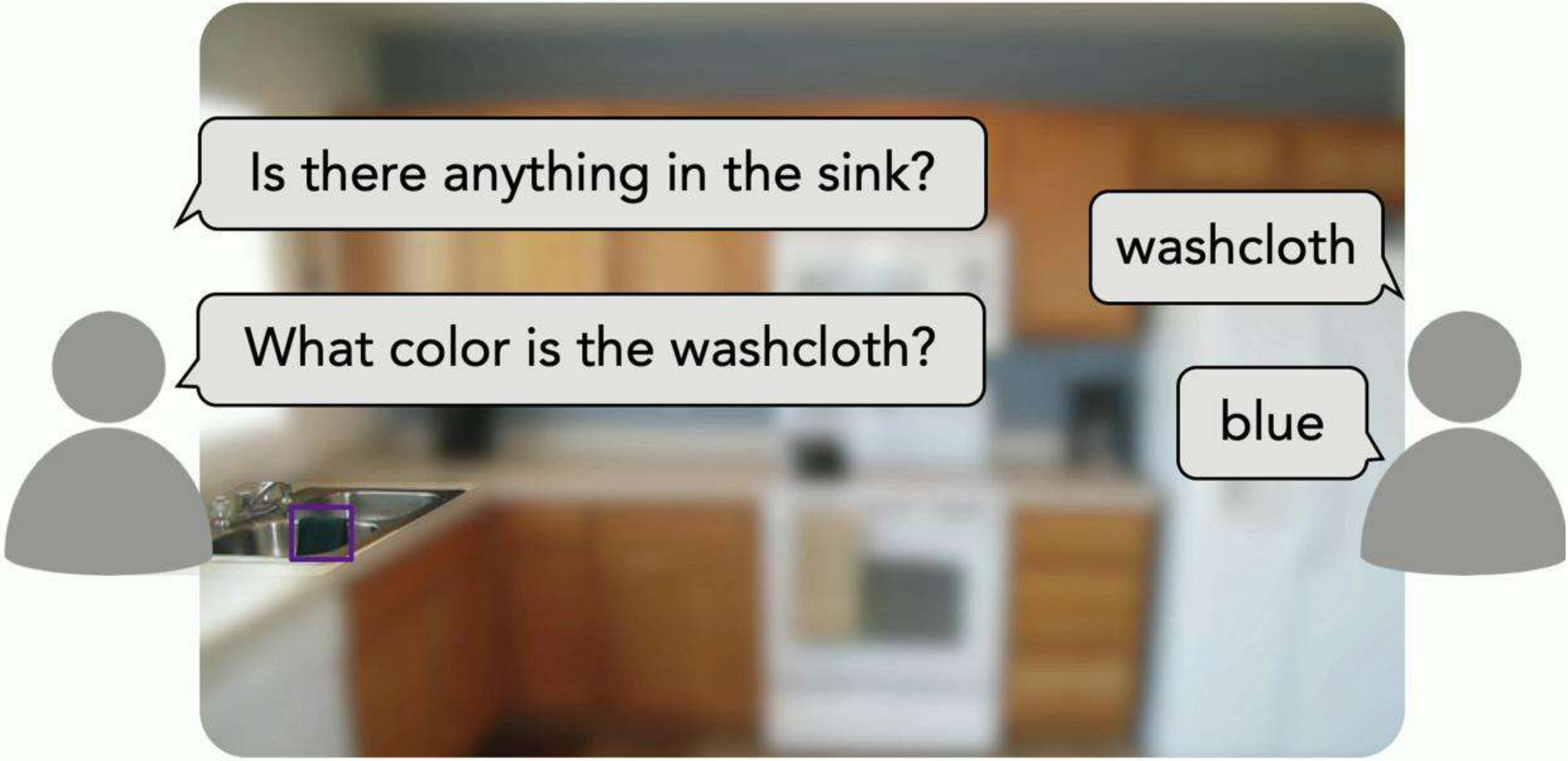


Image Credit: COCO dataset

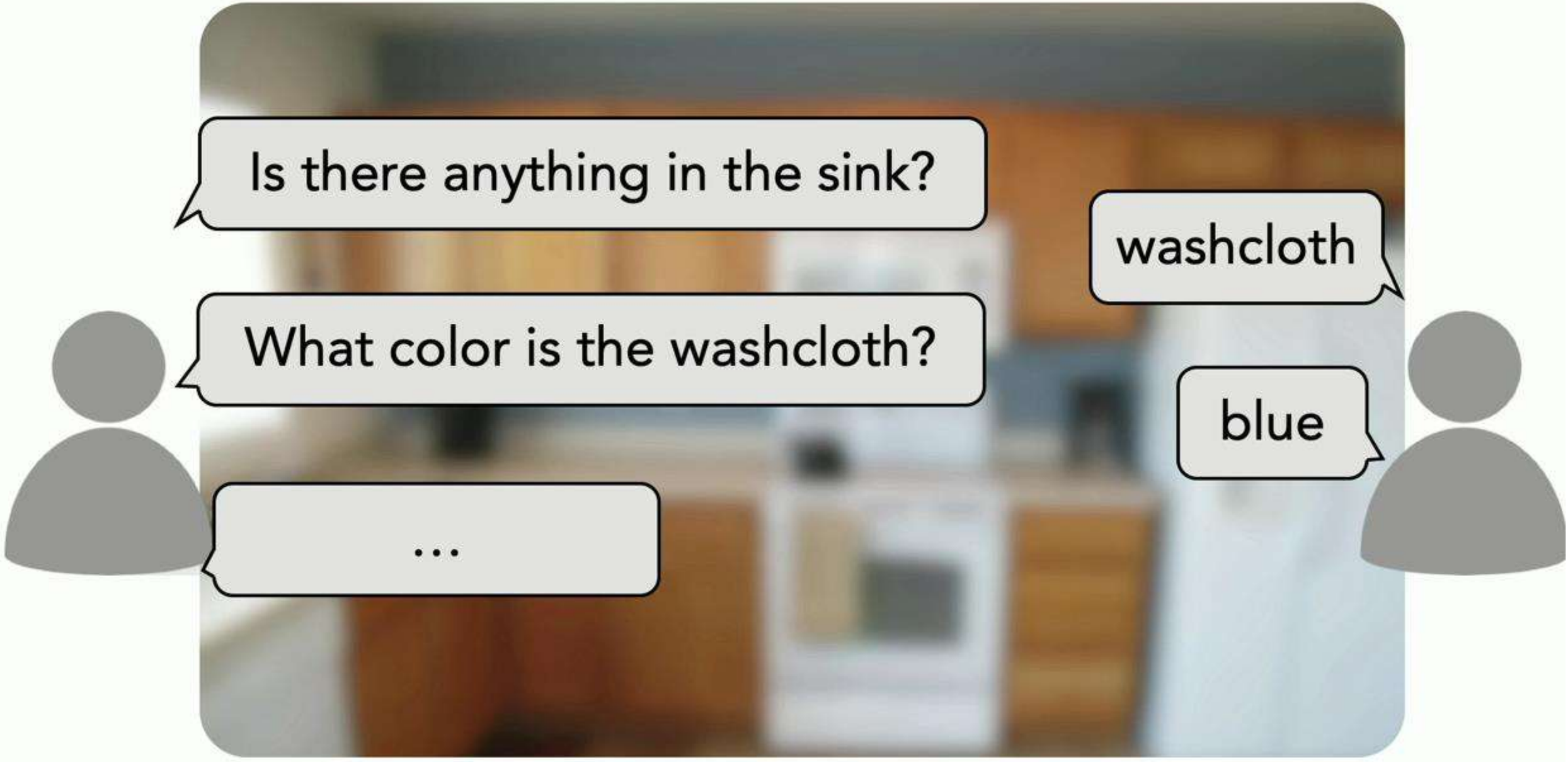


Image Credit: COCO dataset

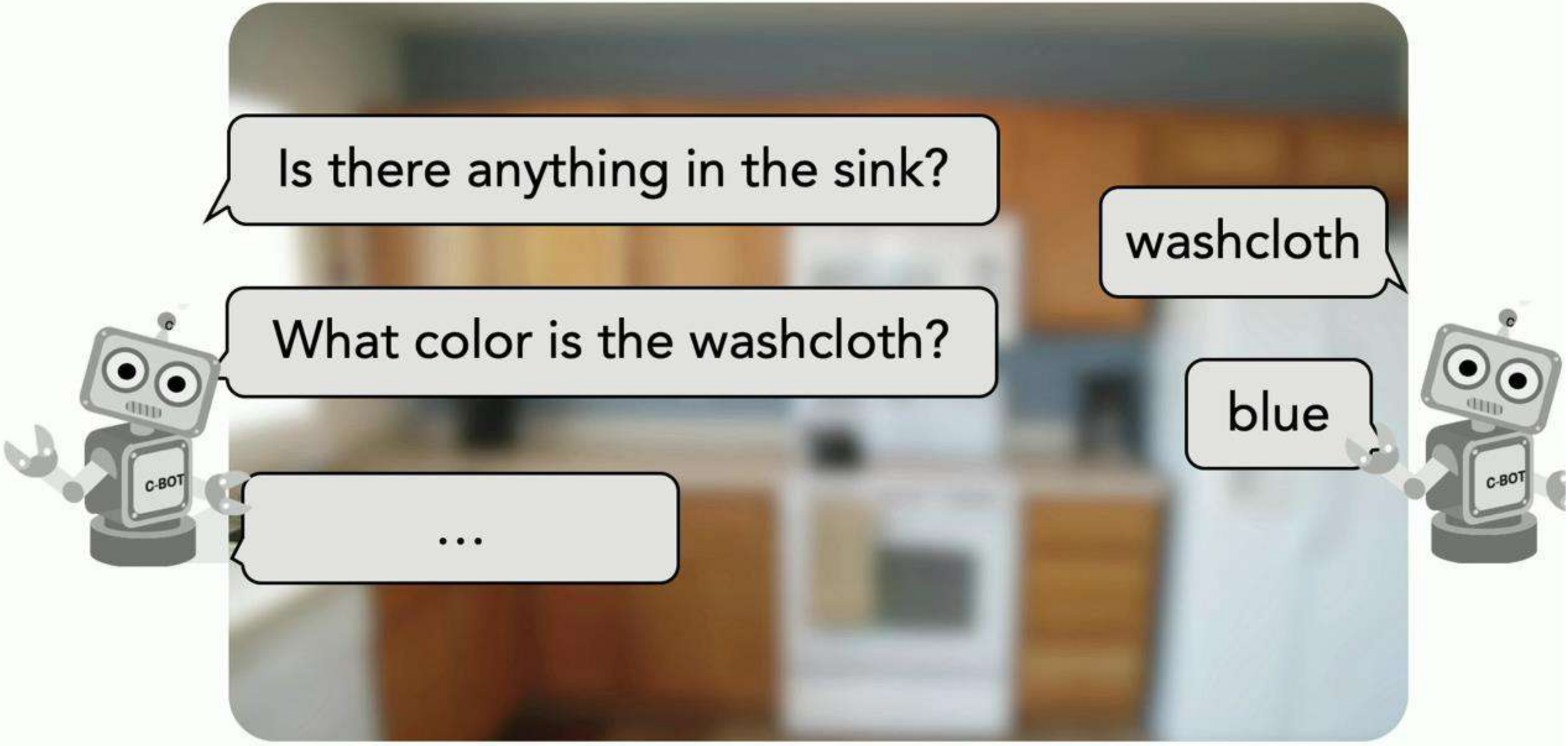


Image Credit: COCO dataset

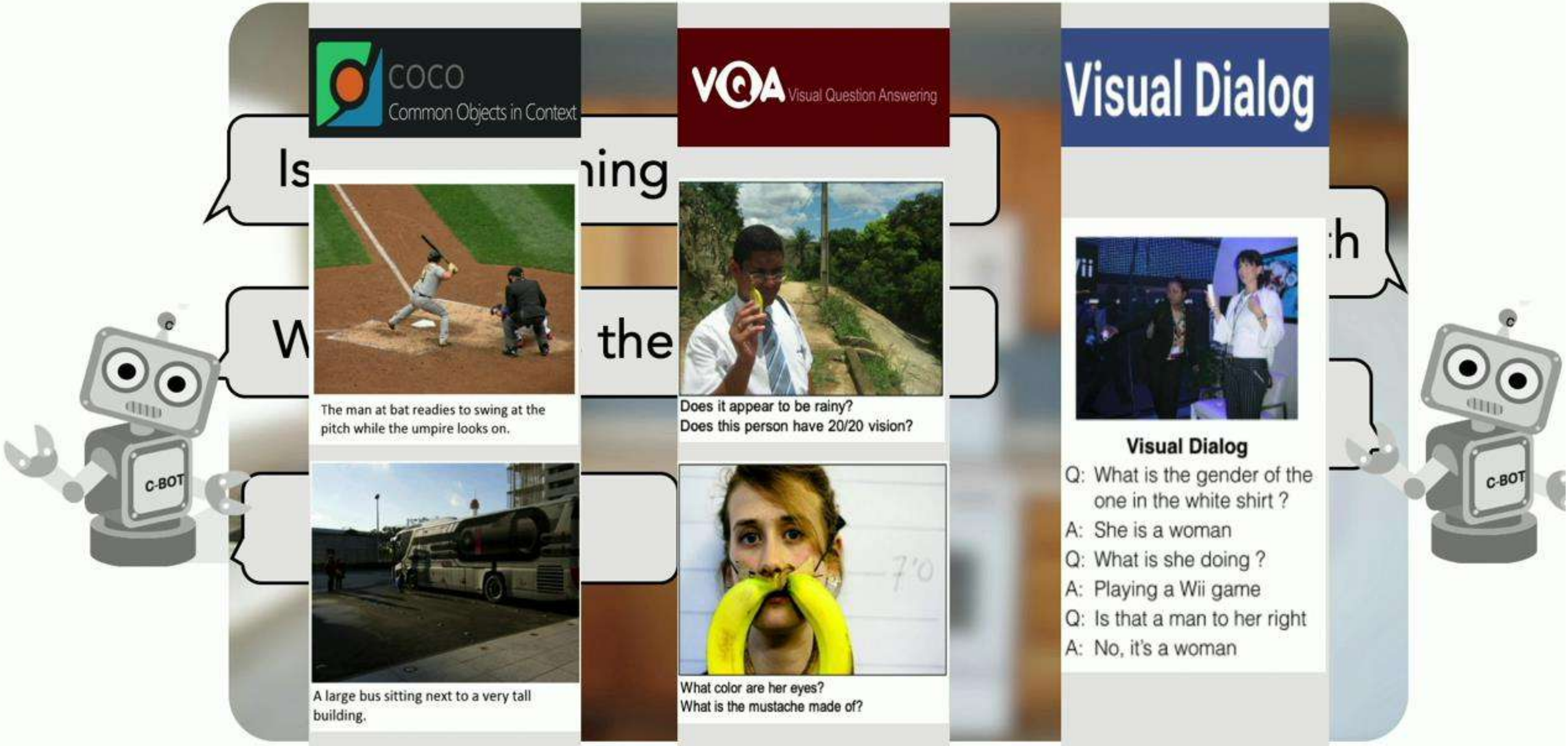


Image Credit: COCO dataset

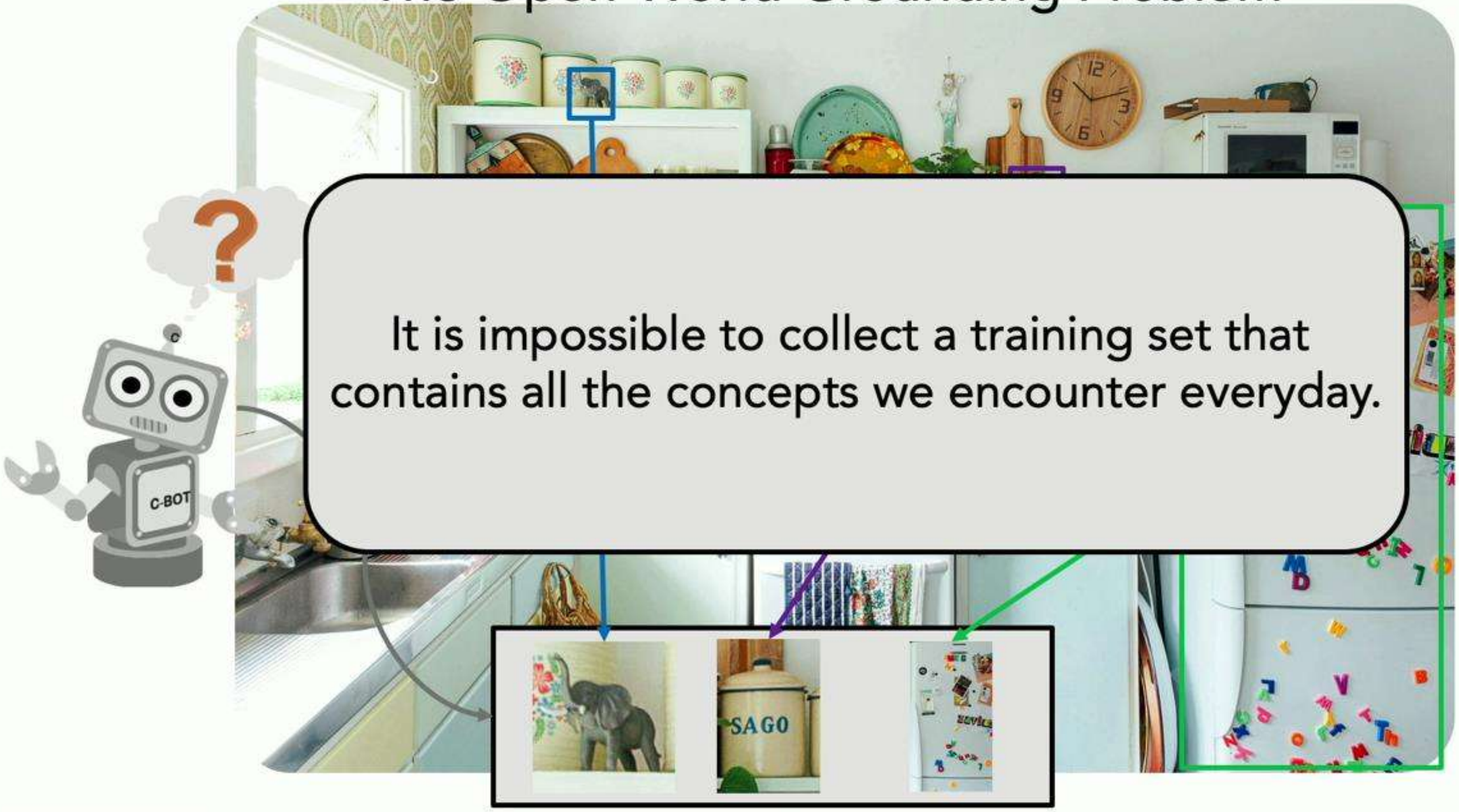
Image Credit: Ezgi Dirik

Image Credit: Ezgi Dirik

Image Credit: Ezgi Dirik

Image Credit: Ezgi Dirik

The Open-World Grounding Problem

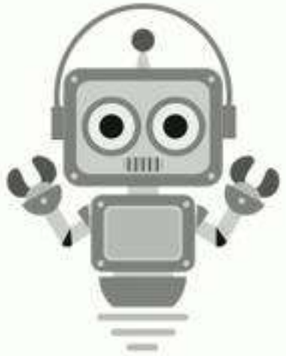


It is impossible to collect a training set that contains all the concepts we encounter everyday.

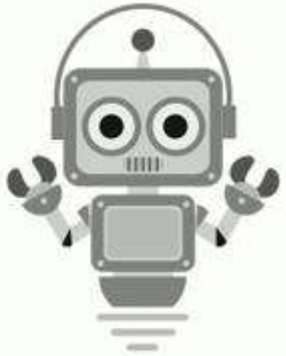
Image Credit: Ezgi Dirik

How to leverage grounding learned from other sources to improve multi-modal AI capabilities ?

Image Credit: Ezgi Dirik

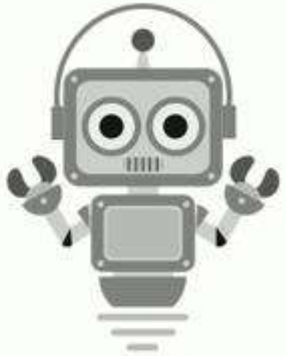


Virtual assistance



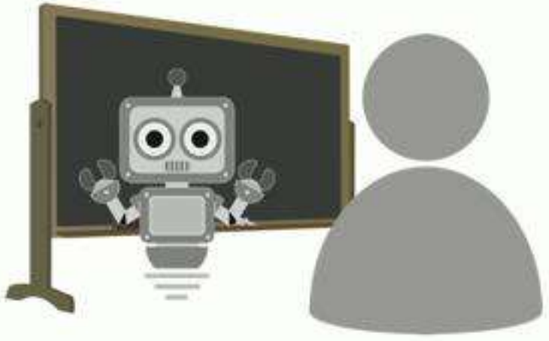
Virtual assistance

Assistive Technology

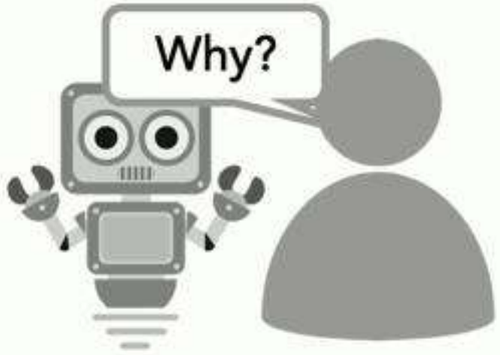


Virtual assistance

Assistive Technology



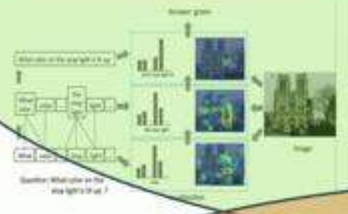
Machine Teaching



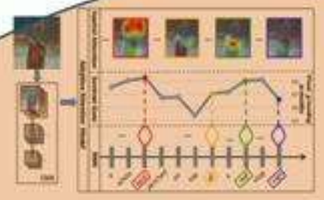
Explainable AI

Visual Question Answering
[ICCV 2015, IJCV]

VQA

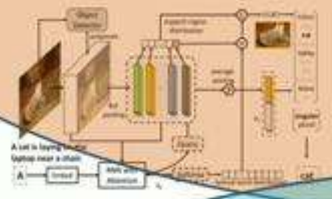


Hierarchical Question-Image Co-Attention for VQA.
[NeurIPS 2016]



Knowing when to Look: adaptive attention for Image captioning
[CVPR 2017]

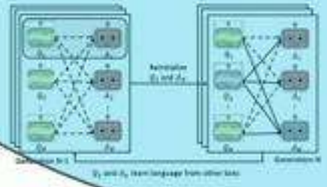
Image Captioning



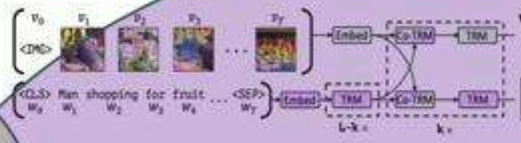
Neural Baby Talk
[CVPR 2018]

Discriminative Learning for visual dialog
[NeurIPS 2018]

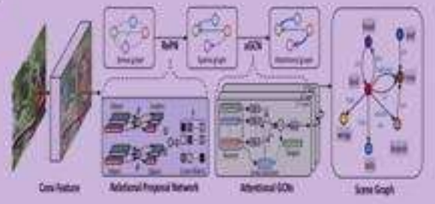
(Visual) Dialog



Emergence of Compositional Language with Deep Generational Transmission.
[In Submission]

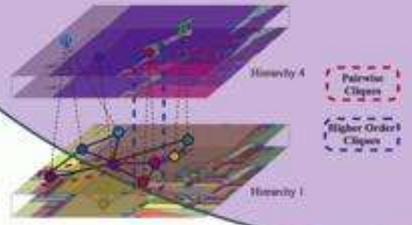


ViLBERT
[NeurIPS 2019]

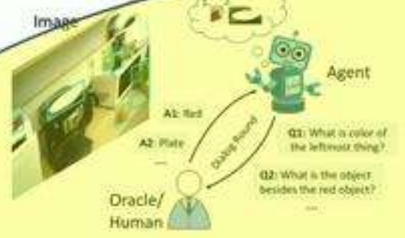


Graph RCNN
[ECCV 2018]

Image/Video understanding



Human Action Segmentation
[CVPR 2015]



Visual Curiosity
[CORL 2018]

Embodied Agent

Self-Monitoring Navigation Agent via Auxiliary Progress Estimation.
[ICLR 2019]

This Talk

Neural Baby Talk

CVPR 2018

ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks

NeurIPS 2019

Recent & Future Work

Image Captioning

The fundamental capability to describe what is seen.

Input:

[Chen *et.al. 2015*]

Image Captioning

The fundamental capability to describe what is seen.

Input:

Desired output:

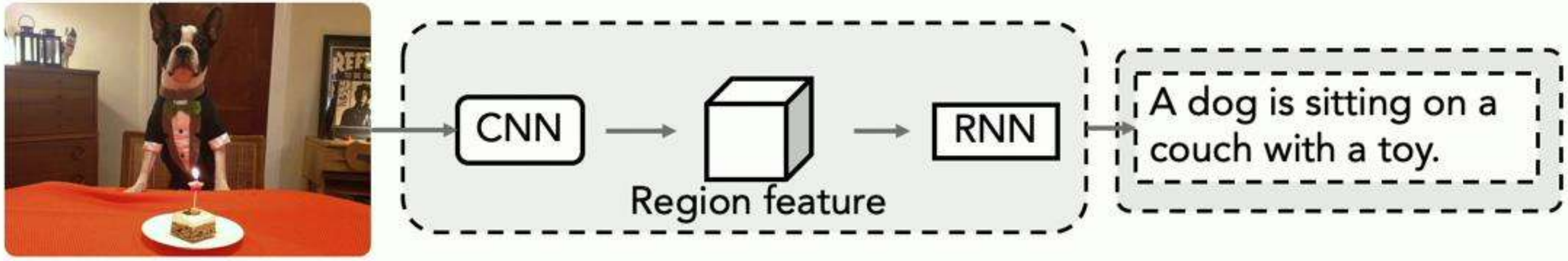
A table that has a mug and a plate with food on it.

A dog wearing goggles and leathers sitting on a motorcycle.

[Chen *et.al. 2015*]

Related Work

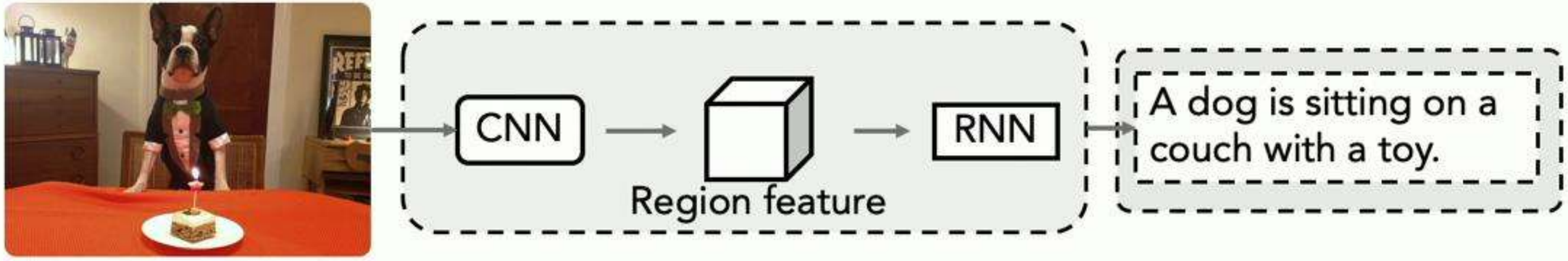
SOTA neural image captioning system.



[Anderson *et.al. 2018*]

Related Work

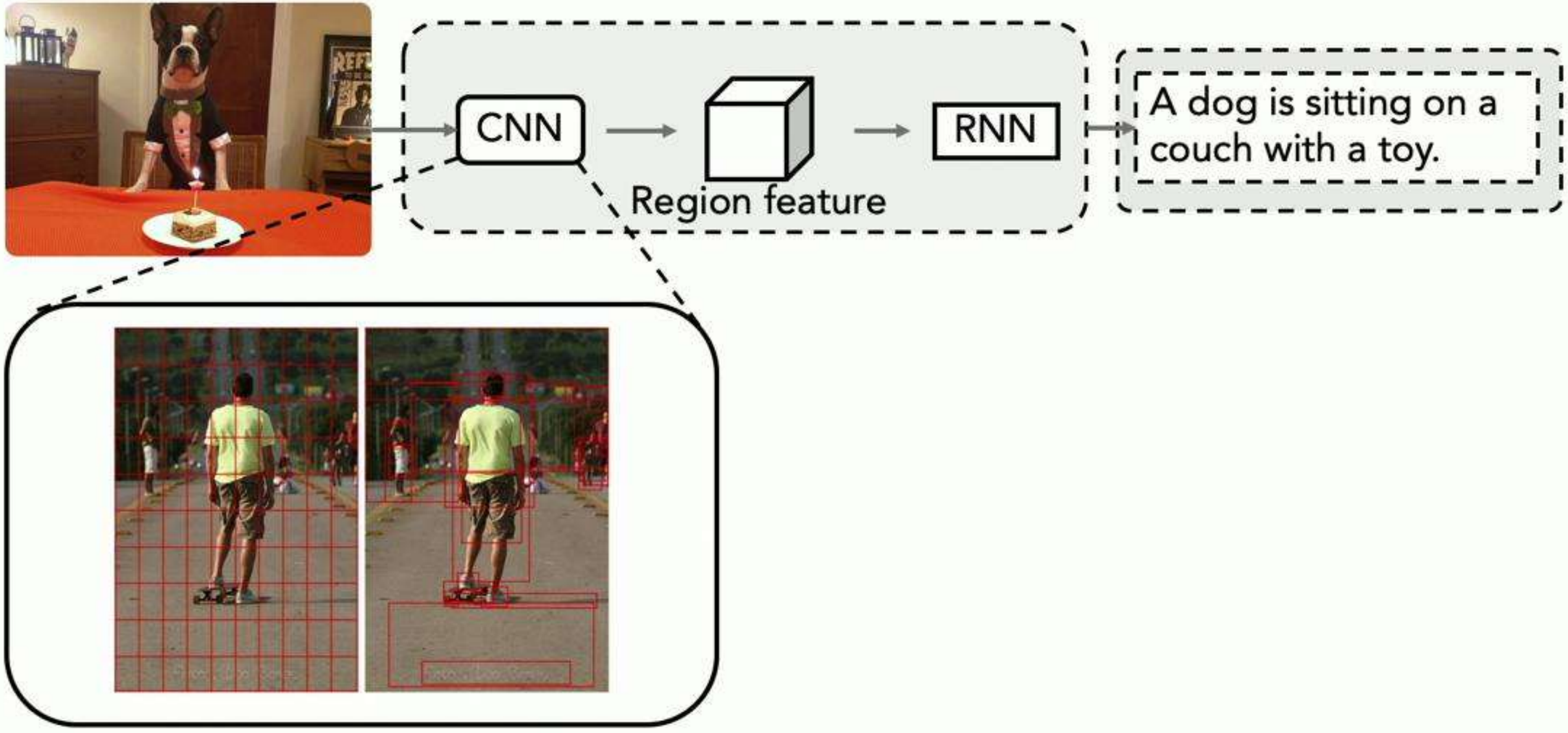
Bottom up and Top down attention for image captioning



[Anderson *et.al. 2018*]

Related Work

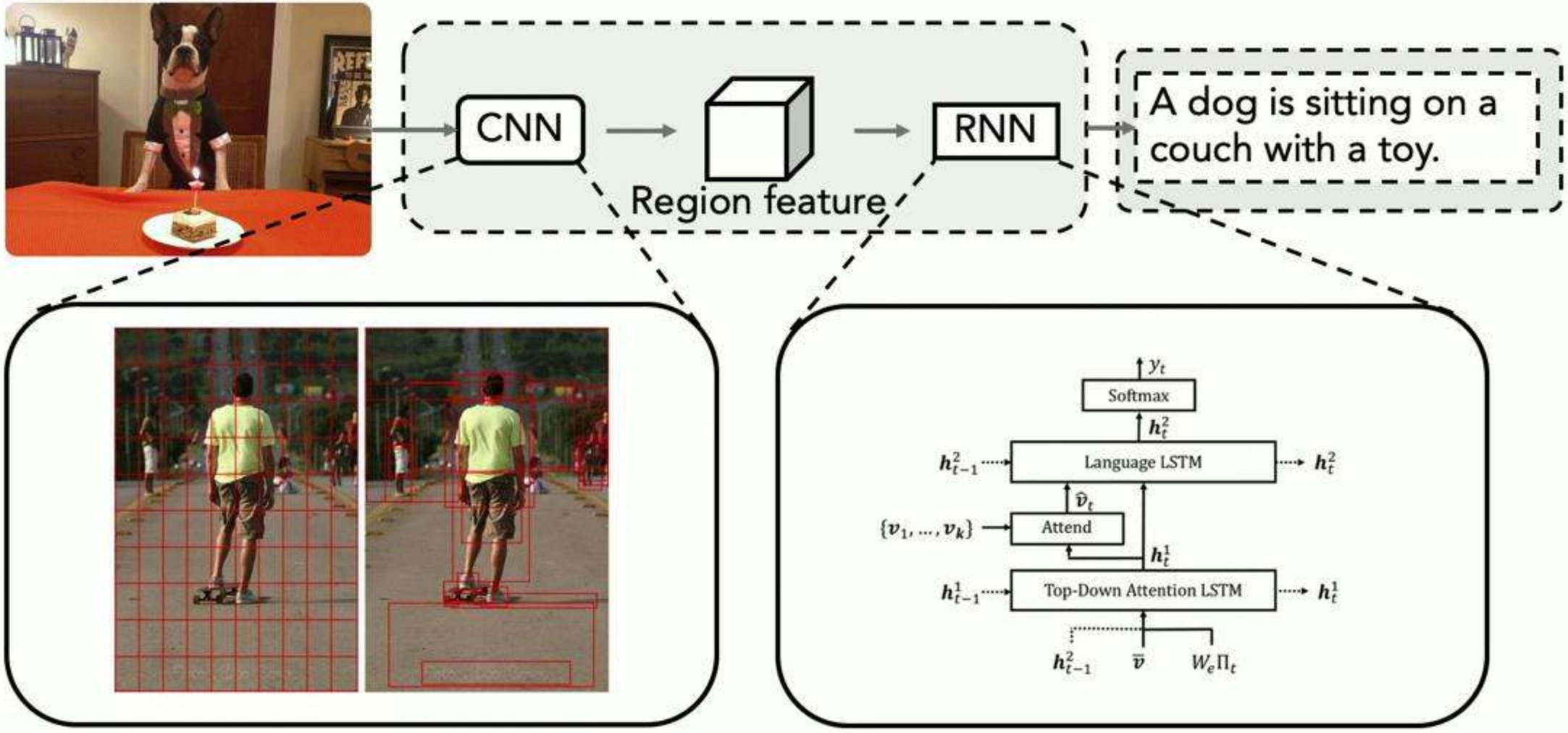
Bottom up and Top down attention for image captioning



[Anderson *et.al. 2018*]

Related Work

Bottom up and Top down attention for image captioning



[Anderson *et.al. 2018*]

Related Work

SOTA neural image captioning system.

Two elephants and a baby elephant walking together.

[Anderson *et.al. 2018*]

Related Work

SOTA neural image captioning system.

Two elephants and a baby elephant walking together.

A cat is standing on a sign that says "UNK".

[Anderson *et.al. 2018*]

Related Work

SOTA neural image captioning system.

Two elephants and a baby elephant walking together.

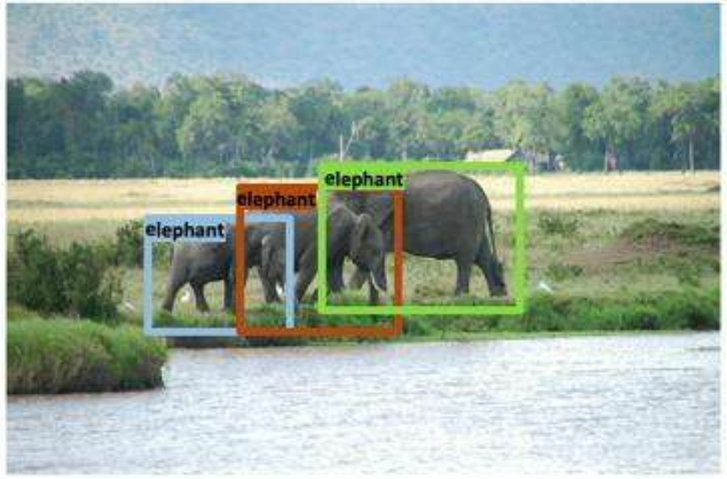
A cat is standing on a sign that says "UNK".

A man standing on a beach holding a surfboard.

[Anderson *et.al. 2018*]

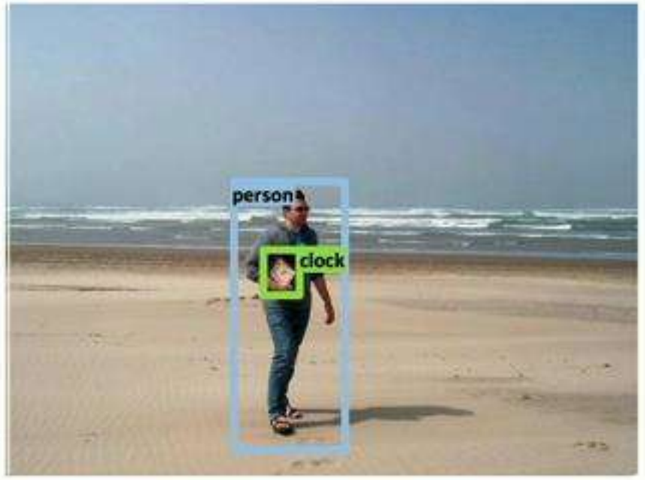
Related Work

Object detection and OCR output



Two elephants and a baby elephant walking together.

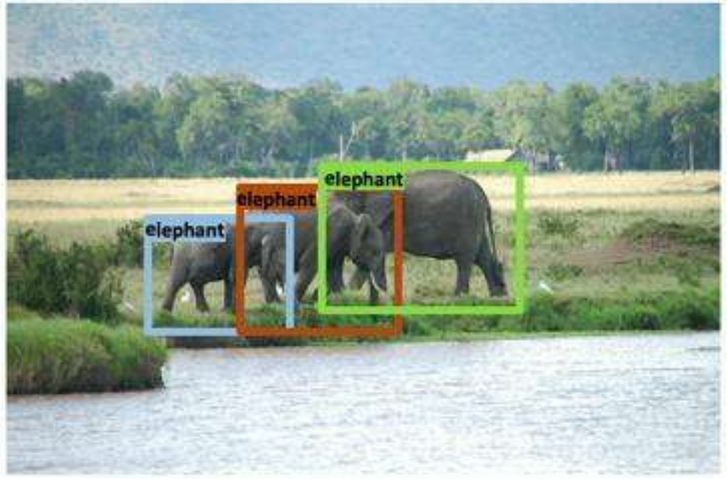
A cat is standing on a sign that says "UNK".



A man standing on a beach holding a surfboard.

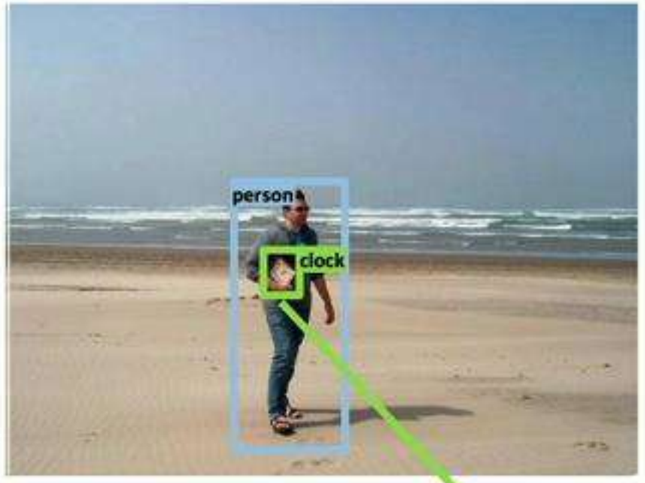
Related Work

Learn from object detection and OCR.



Two elephants and a baby elephant walking together.

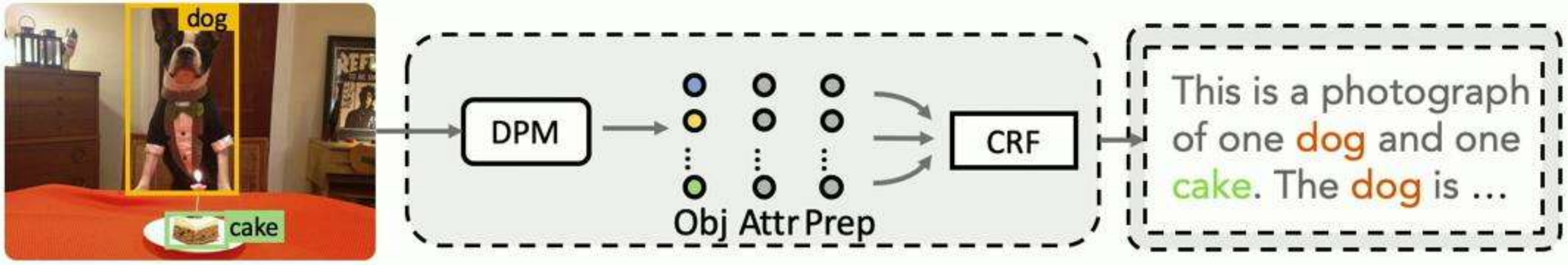
A cat is standing on a sign that says "~~UNK~~"-"Abundzu".



A man standing on a beach holding a ~~surfboard~~ - clock.

Related Work

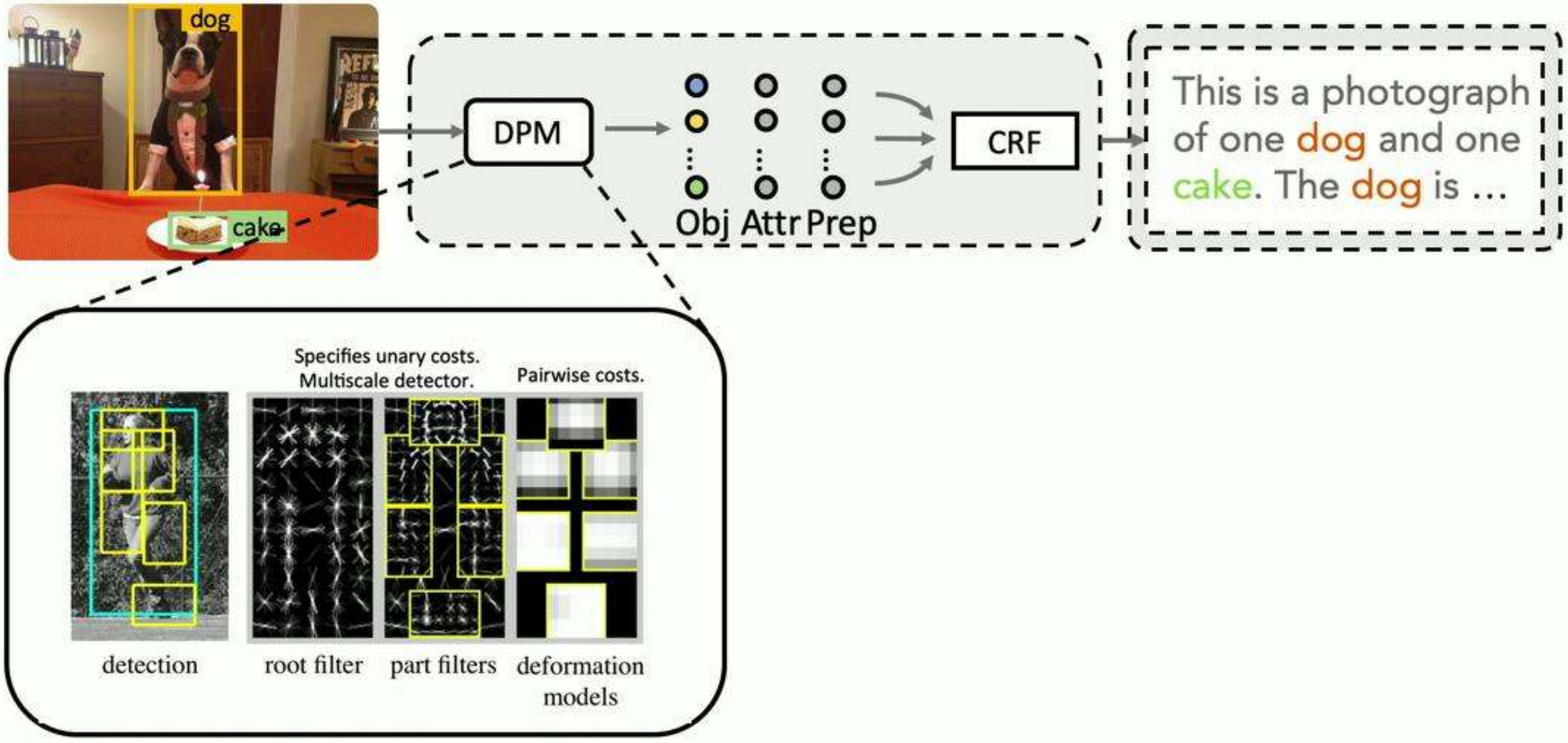
Image captioning system before deep learning "revolution".



[Kulkarni *et.al. 2011*]

Related Work

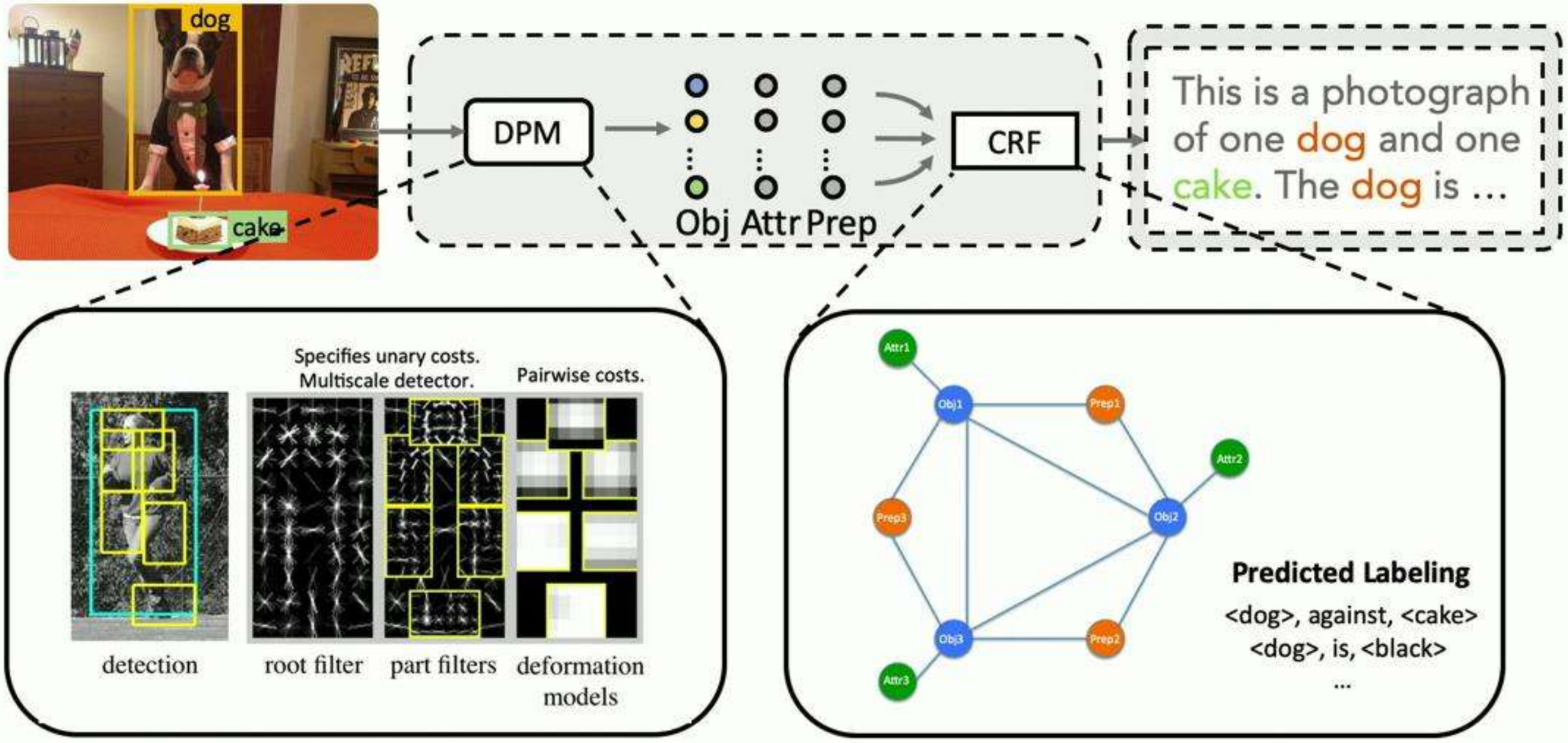
Image captioning system before deep learning "revolution".



[Kulkarni *et.al. 2011*]

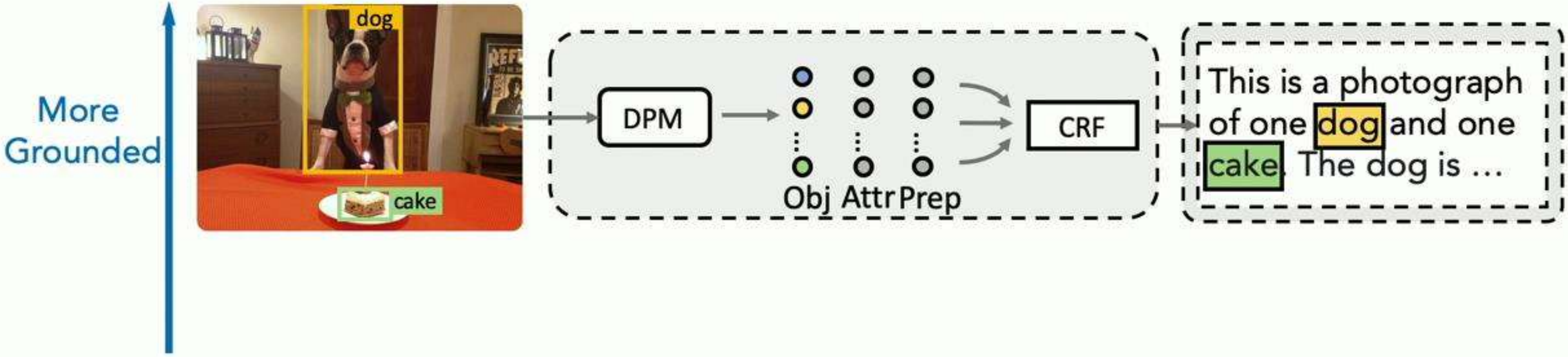
Related Work

Image captioning system before deep learning "revolution".



[Kulkarni *et.al. 2011*]

Motivation


More Grounded
dog
cake
DPM
Obj Attr Prep
CRF
This is a photograph of one dog and one cake. The dog is …

Motivation

More Grounded

DPM → Obj Attr Prep → CRF → This is a photograph of one dog and one cake. The dog is …

More Natural

CNN → Region feature → RNN → A dog is sitting on a couch with a toy.

Motivation

More Grounded

DPM → Obj Attr Prep → CRF → This is a photograph of one dog and one cake. The dog is …

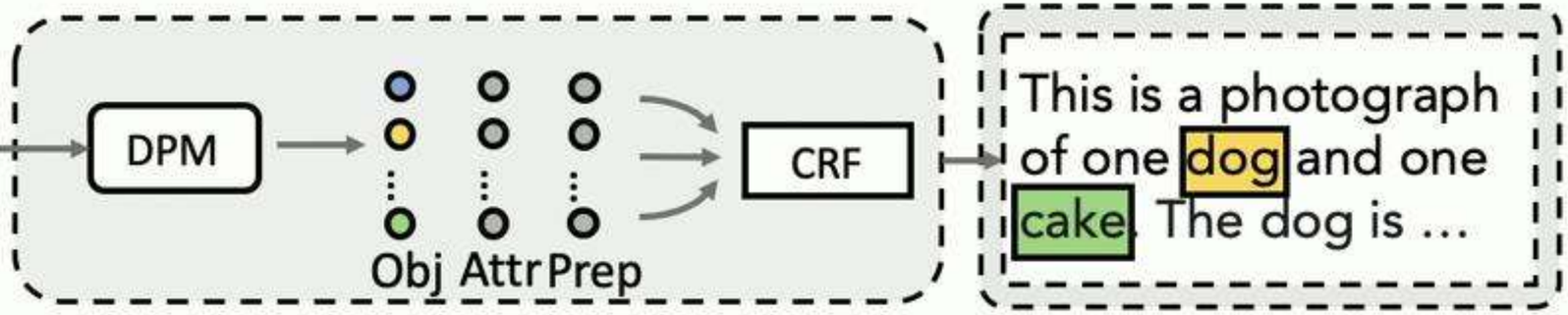
- Associate the named concept to the pixels / bounding box detections in the image.

More Natural

CNN → Region feature → RNN → A dog is sitting on a couch with a toy.

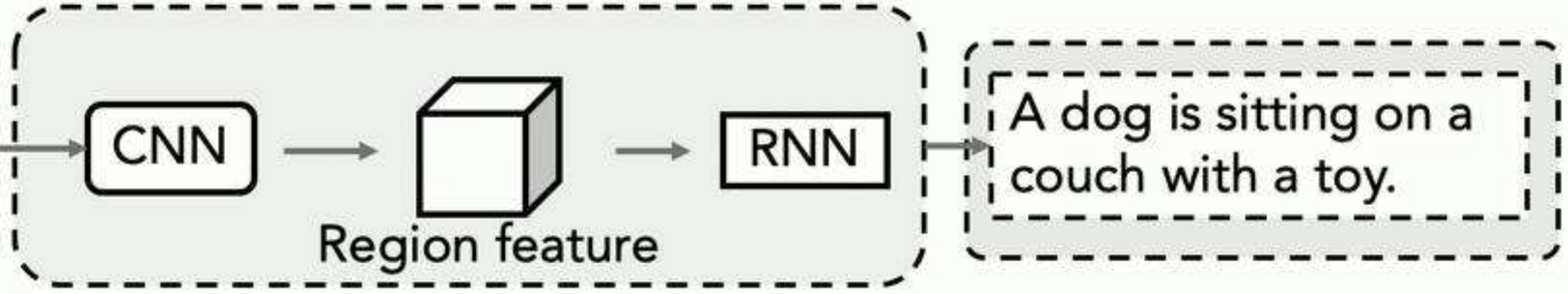
Motivation

More Grounded

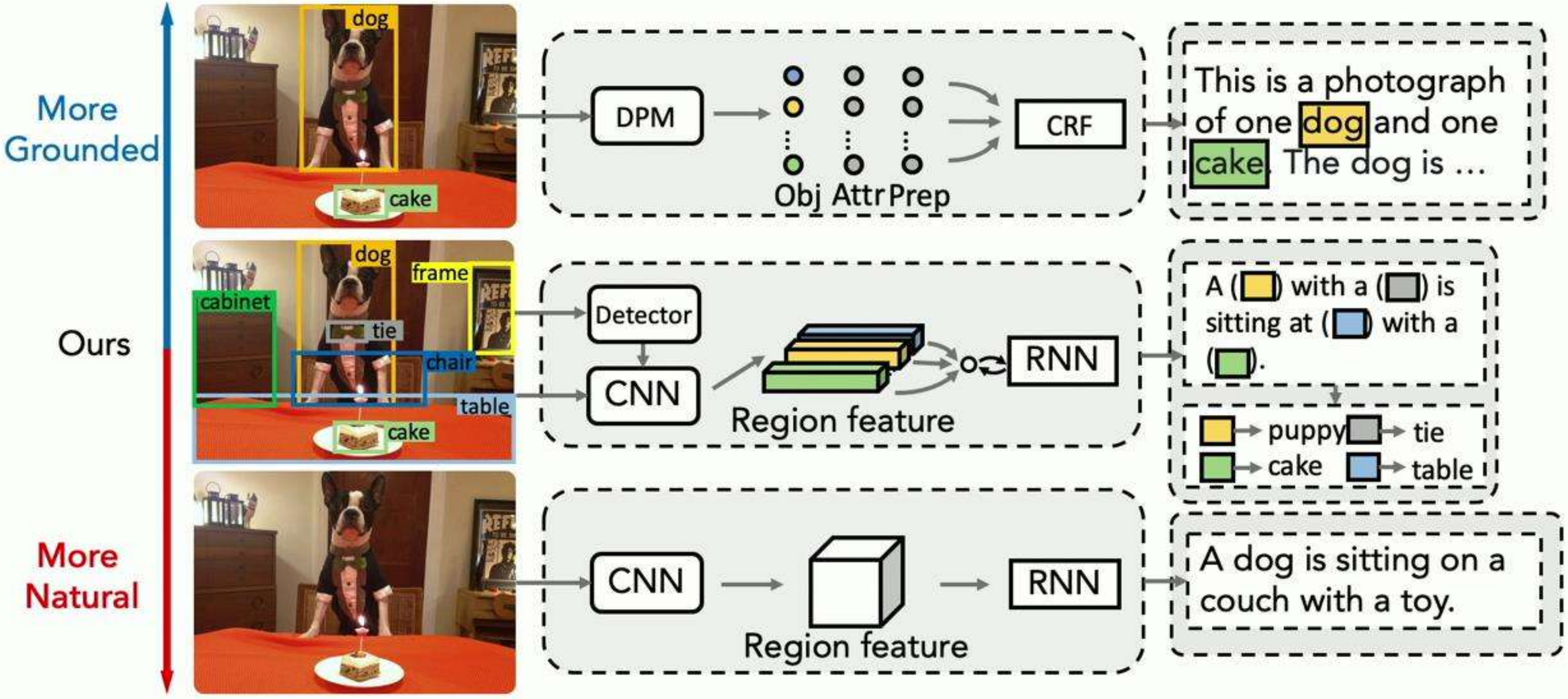


- Associate the named concept to the pixels / bounding box detections in the image.
- No human generated templated, language needs to be natural.

More Natural

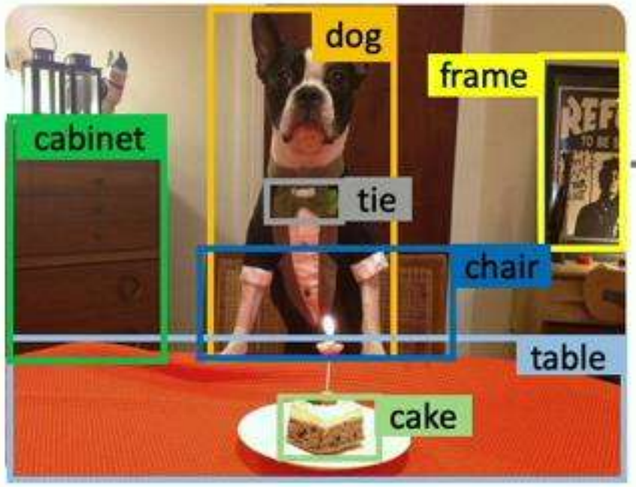


Motivation


More Grounded
Ours
More Natural
dog
cake
DPM
Obj Attr Prep
CRF
This is a photograph of one dog and one cake. The dog is …
dog
frame
cabinet
tie
chair
table
cake
Detector
CNN
Region feature
RNN
A ( ) with a ( ) is sitting at ( ) with a ( ).
puppy
tie
cake
table
CNN
Region feature
RNN
A dog is sitting on a couch with a toy.

Framework

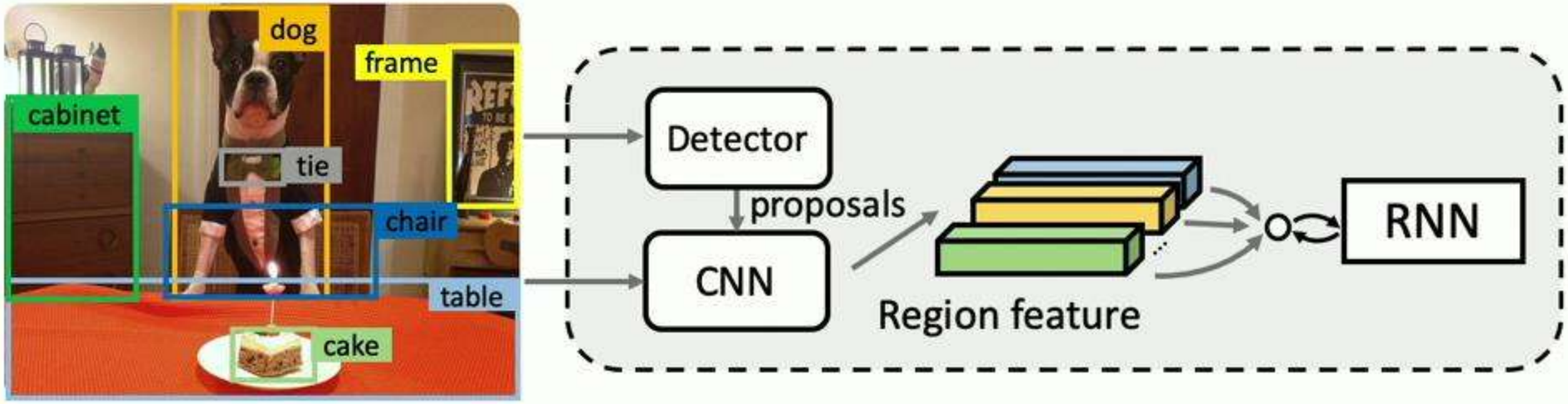
Neural Baby Talk



Detector

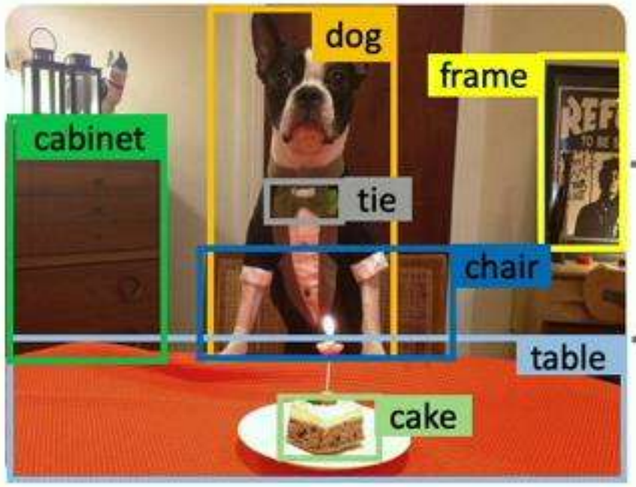
Framework

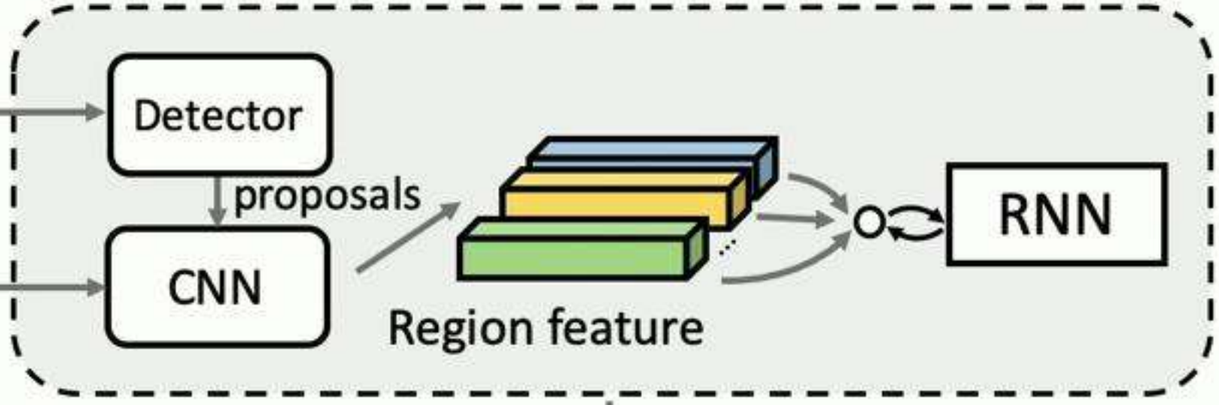
Neural Baby Talk



Framework

Neural Baby Talk



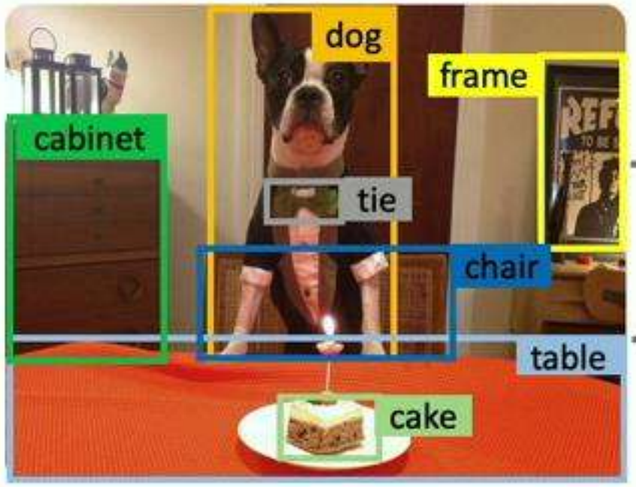


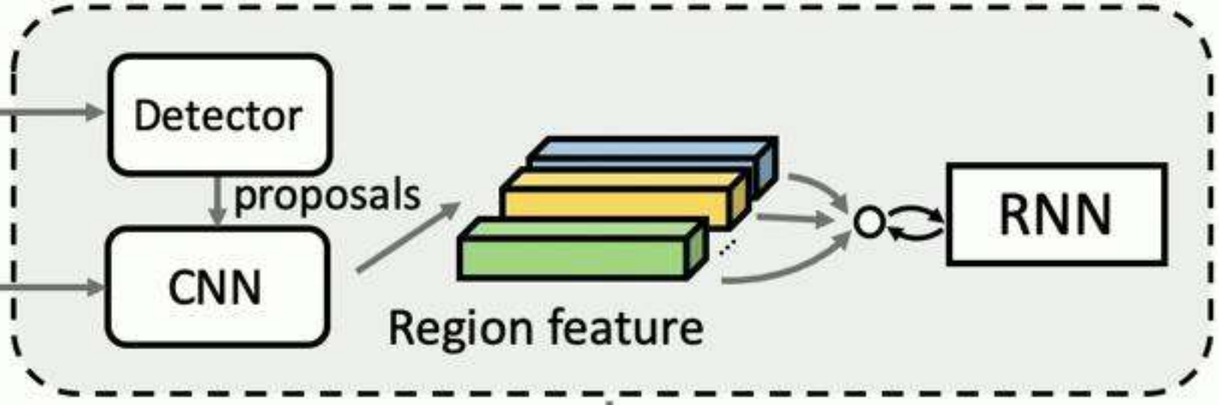
- Slotted caption template generation

A ( ) with a ( ) is sitting at ( ) with a ( ).

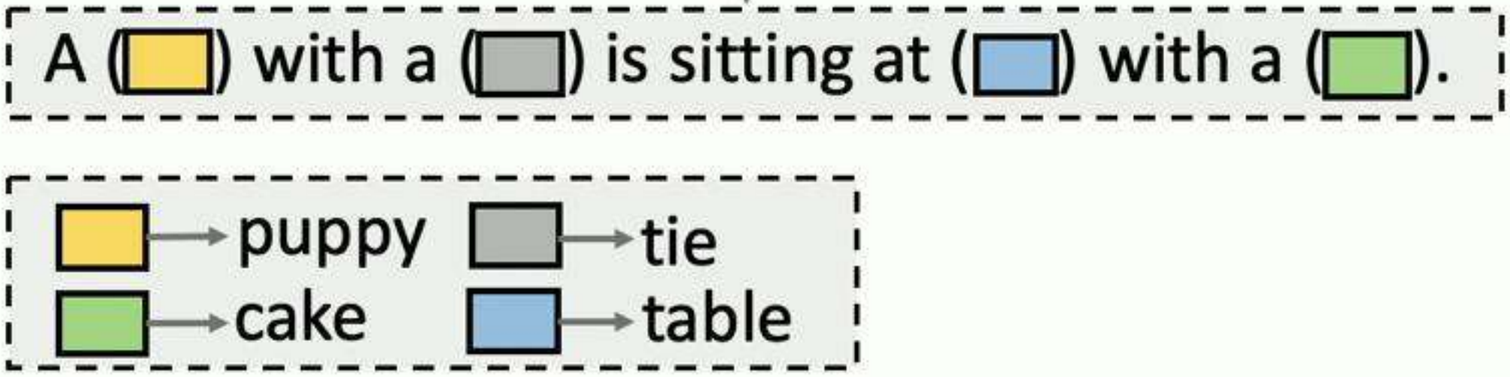
Framework

Neural Baby Talk



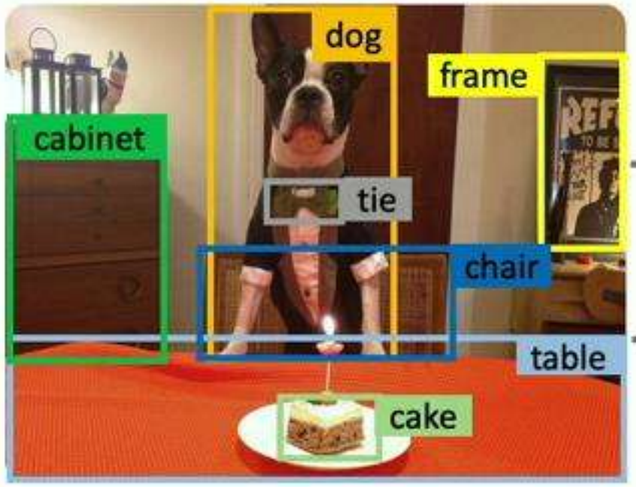


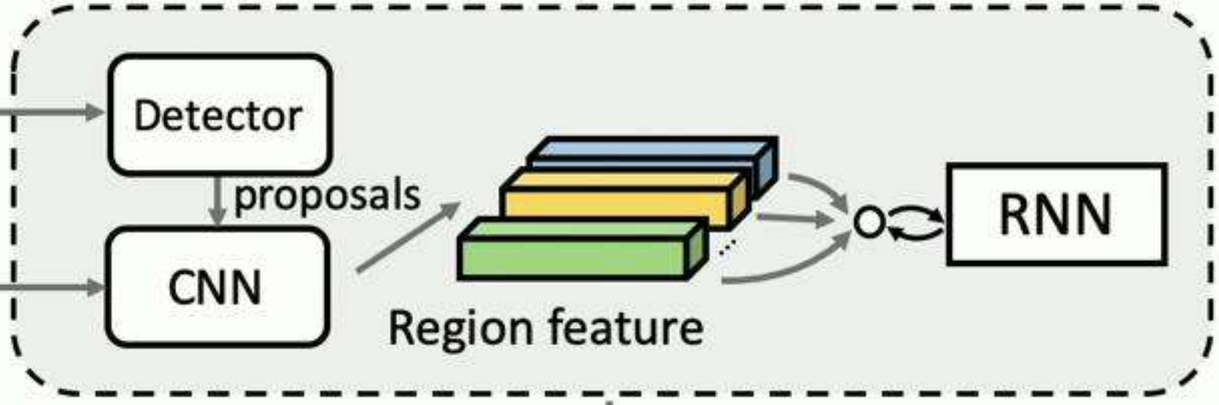
- Slotted caption template generation
- Filling in the slots



Framework

Neural Baby Talk





- Slotted caption template generation

A ( ) with a ( ) is sitting at ( ) with a ( ).

- Filling in the slots

- Final caption

A puppy with a tie is sitting at table with a cake.

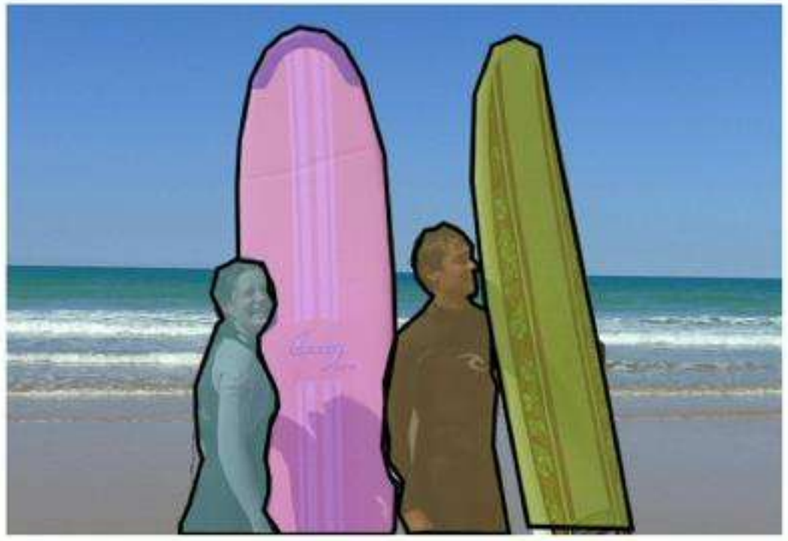
Model

How to construct the caption template?

A young woman standing next to a man holding a surfboard.

Model

How to construct the caption template?

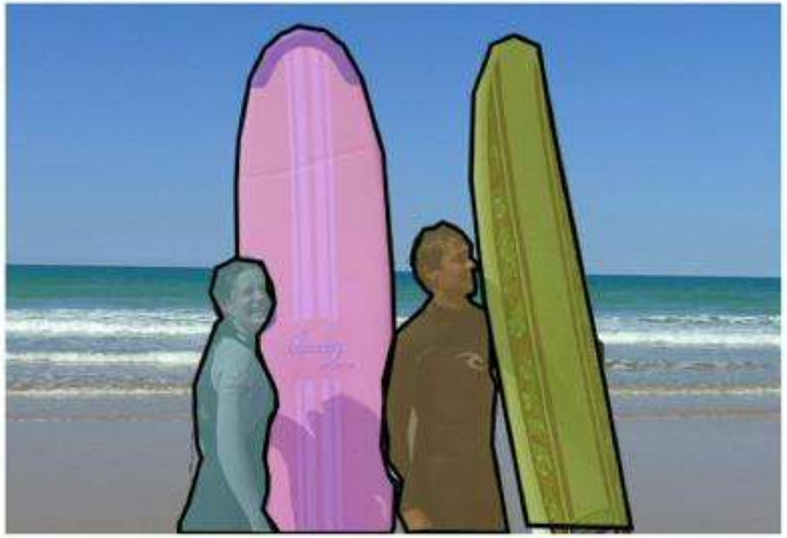


A young woman standing next to a man holding a surfboard.

Object label: <person>, <person>, <surfboard>, <surfboard>

Model

How to construct the caption template?



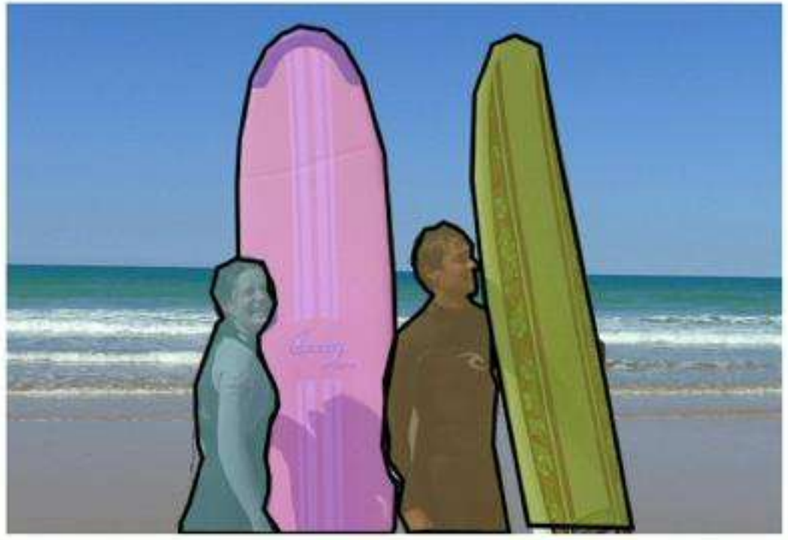
A young woman standing next to a man holding a surfboard.

Object label: <person>, <person>, <surfboard>, <surfboard>

A young woman standing next to a man holding a surfboard.

Model

How to construct the caption template?



A young woman standing next to a man holding a surfboard.

Object label: <*person*>, <*person*>, <*surfboard*>, <*surfboard*>

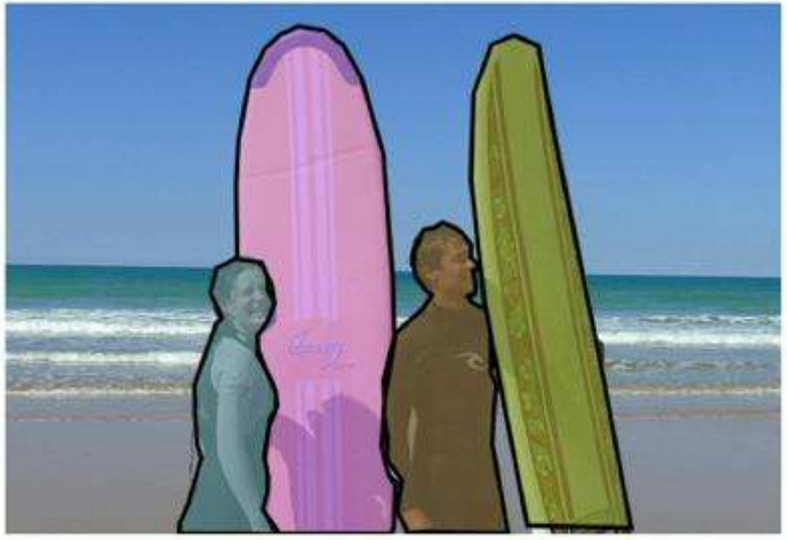
A young woman standing next to a man holding a surfboard.

A young [redacted] standing next to a [redacted] holding a [redacted].

[redacted] noun

Model

How to construct the caption template?



A young woman standing next to a man holding a surfboard.

Object label: <person>, <person>, <surfboard>, <surfboard>

A young woman standing next to a man holding a surfboard.

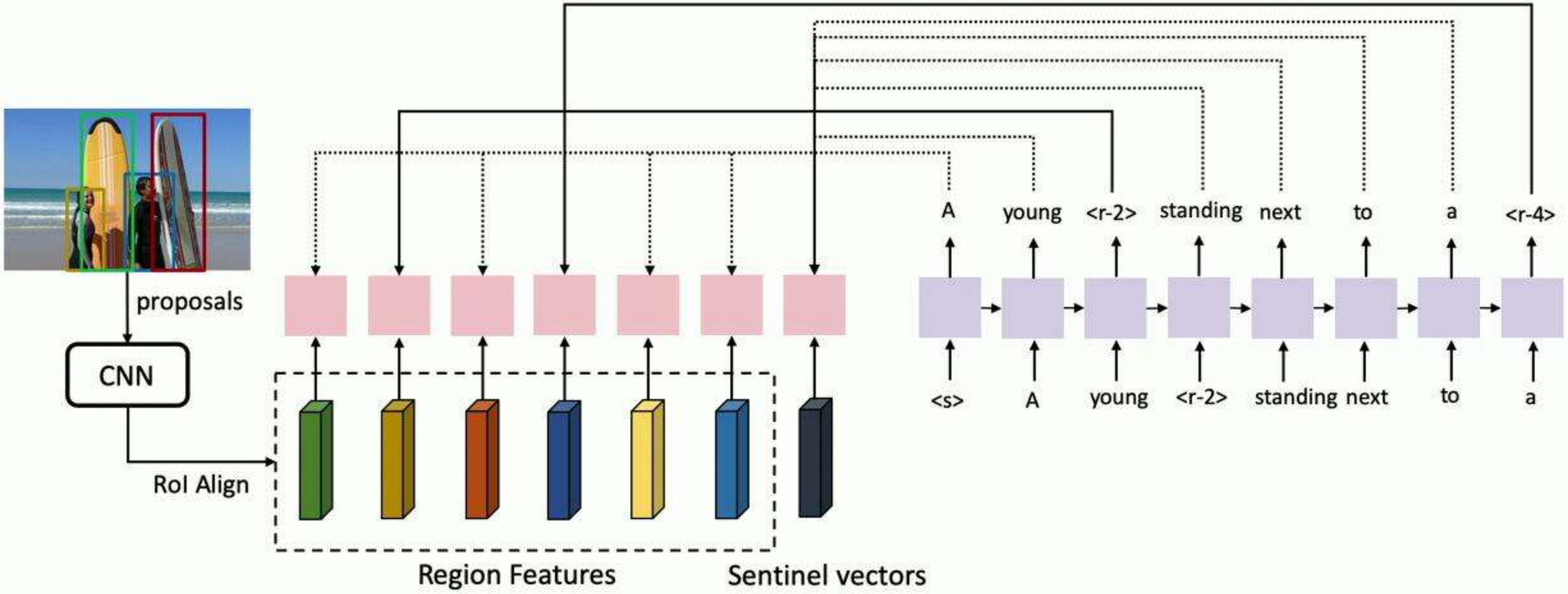
A young [redacted] standing next to a [redacted] holding a [redacted].

A [redacted] [redacted] [redacted] next to a [redacted] [redacted] a [redacted].

[redacted] noun [redacted] attribute [redacted] action

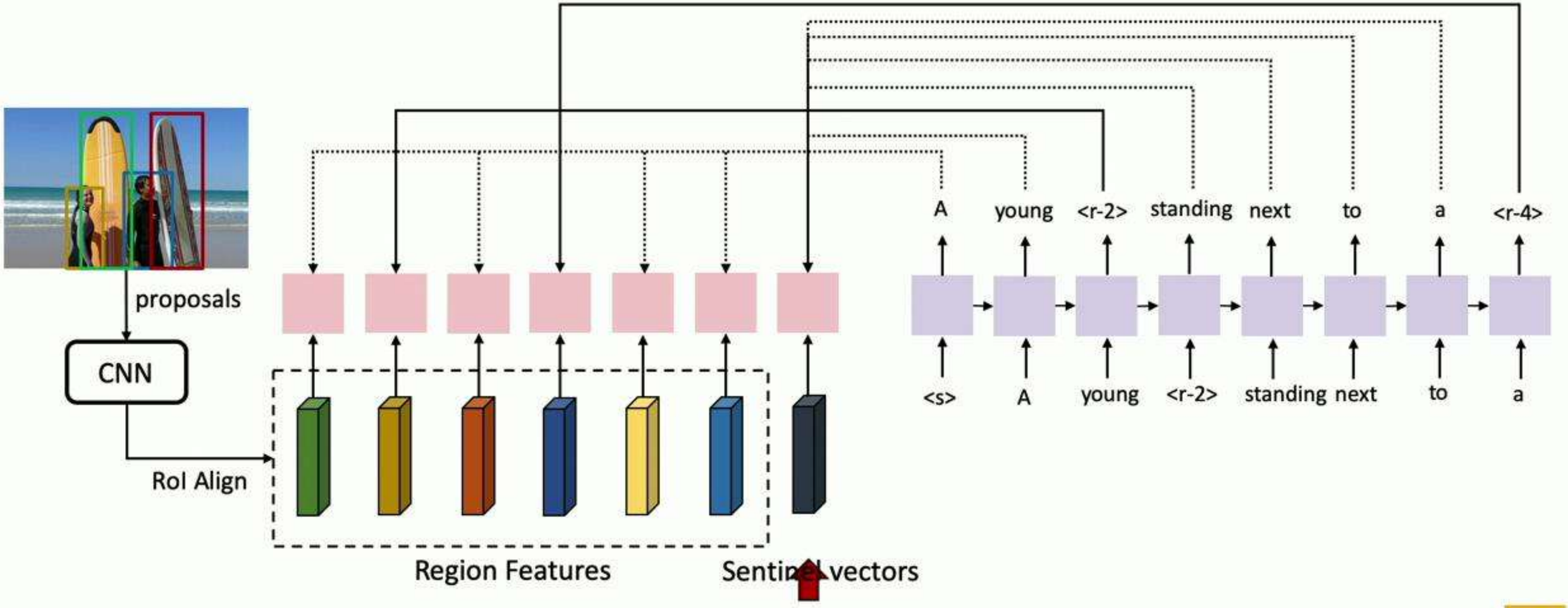
Model

Use pointer networks [Vinyals et.al 2015] with visual sentinel [Lu et.al 2017] to generate caption templates.



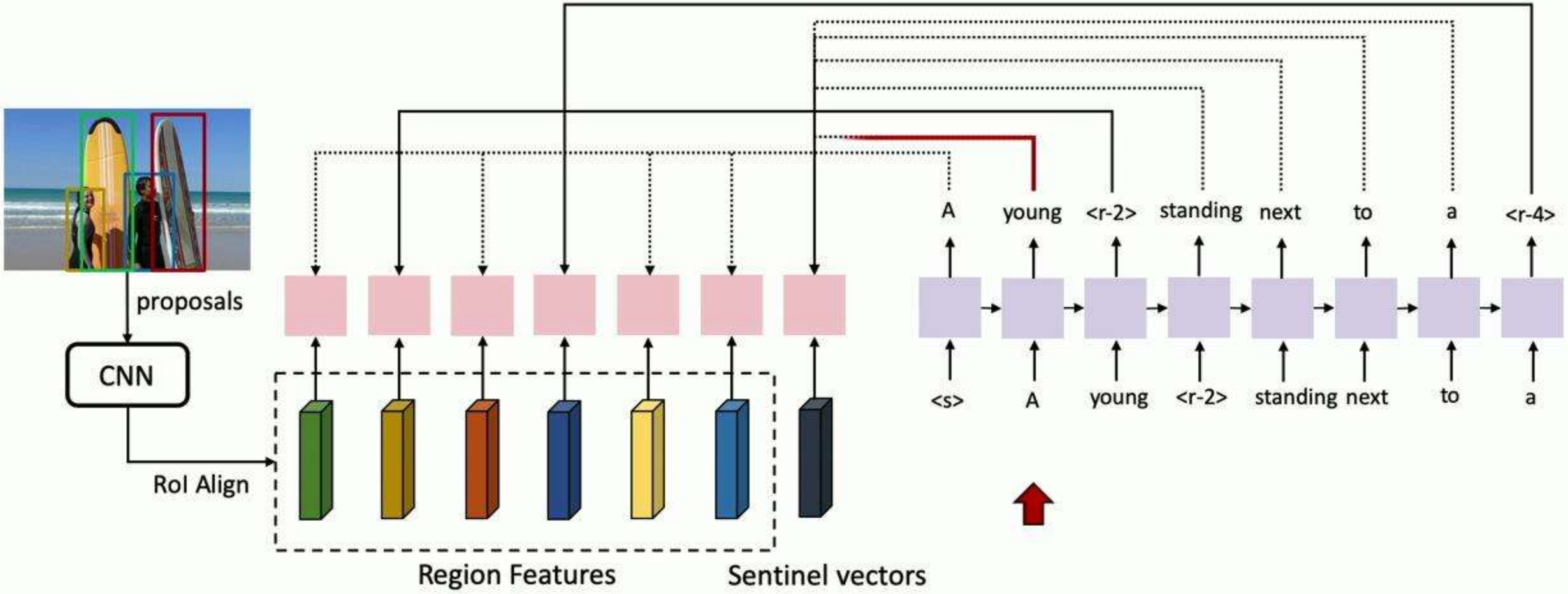
Model

Use pointer networks [Vinyals et.al 2015] with visual sentinel [Lu et.al 2017] to generate caption templates.



Model

Use pointer networks [Vinyals et.al 2015] with visual sentinel [Lu et.al 2017] to generate caption templates.

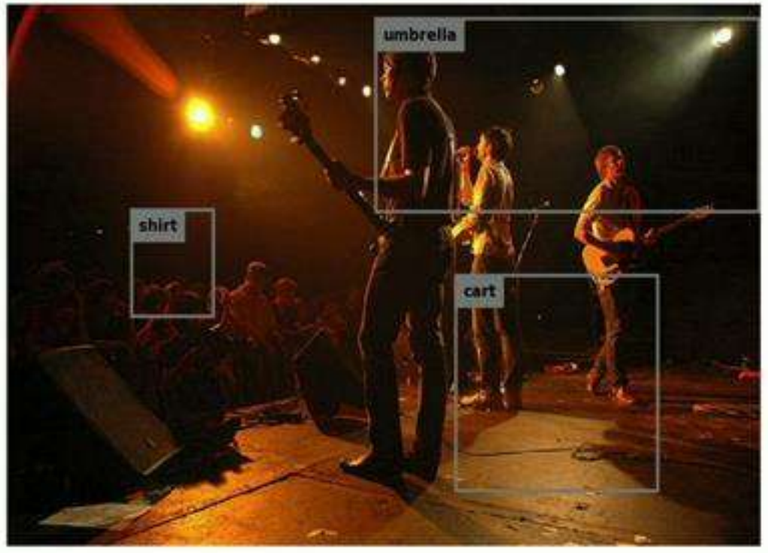


Model

How to handle inaccurate object detection?

Model

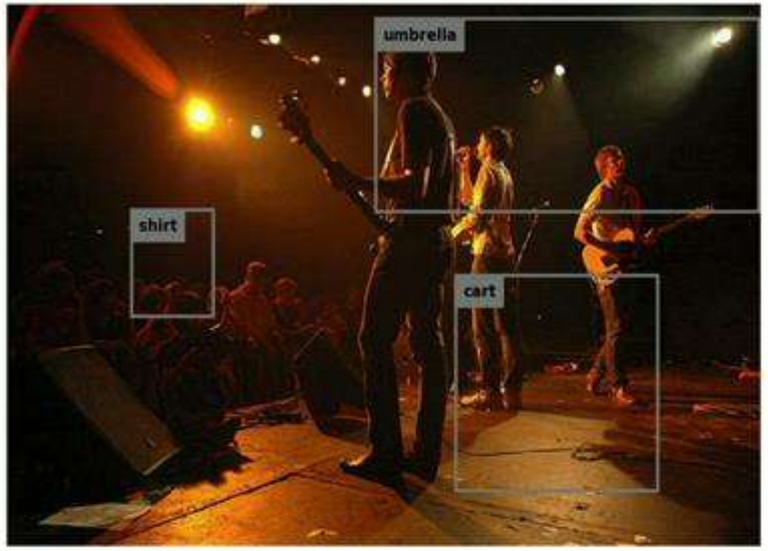
How to handle inaccurate object detection?



A man is playing guitar on the stage.

Model

How to handle inaccurate object detection?



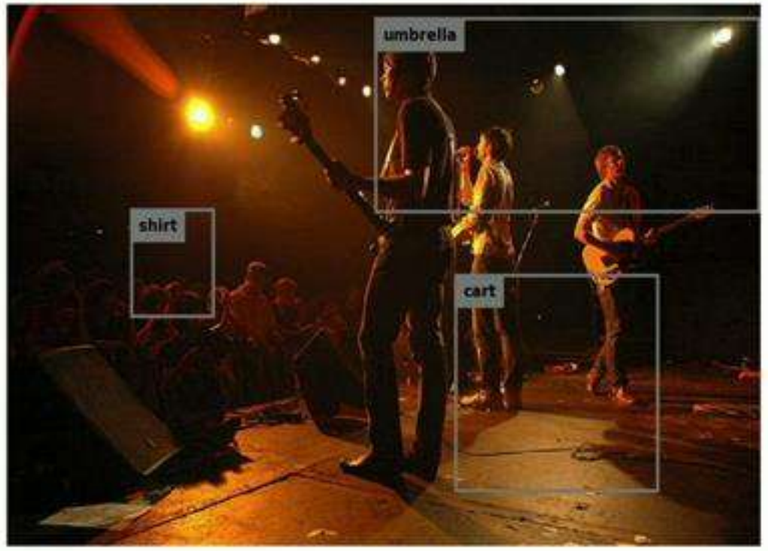
A man is playing guitar on the stage.

- Treat object labels as caption template.

  Template: *A man is playing guitar on the stage.*

Model

How to handle inaccurate object detection?



A man is playing guitar on the stage.

- Treat object labels as caption template.

  Template: *A man is playing guitar on the stage.*

- Dynamically identify the visual words.

  IoU(*detected bbox, GT bbox*) > 0.5 *and labels == words*

Model

Filling in the slots

Input: A young <region-2> standing next to a <region-3> holding a <region-5>.

person, person, surfboard (coarse names form the object detector)

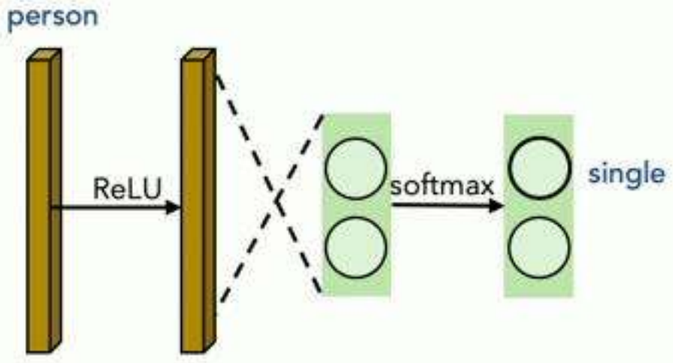
Model

Filling in the slots

Input: A young <region-2> standing next to a <region-3> holding a <region-5>.

person, person, surfboard (coarse names form the object detector)

1) Classify Plurality



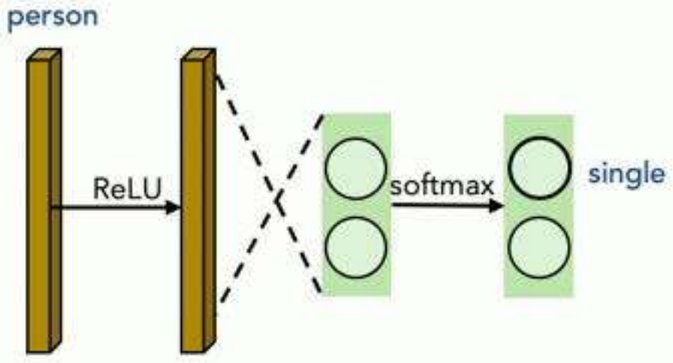
Model

Filling in the slots

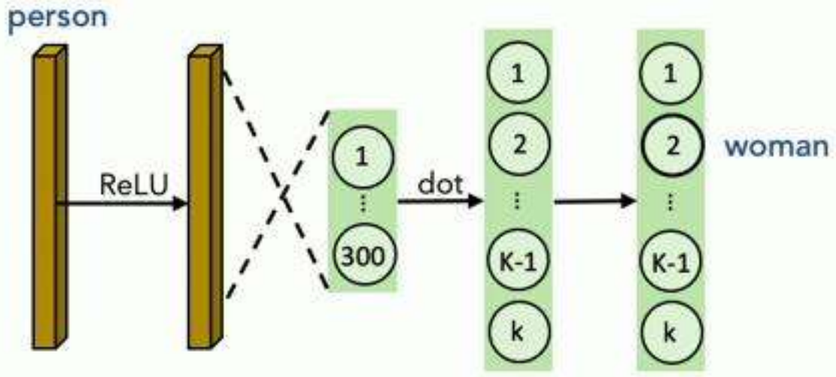
Input: A young <region-2> standing next to a <region-3> holding a <region-5>.

person, person, surfboard (coarse names form the object detector)

1) Classify Plurality



2) Determine Fine Grained Category



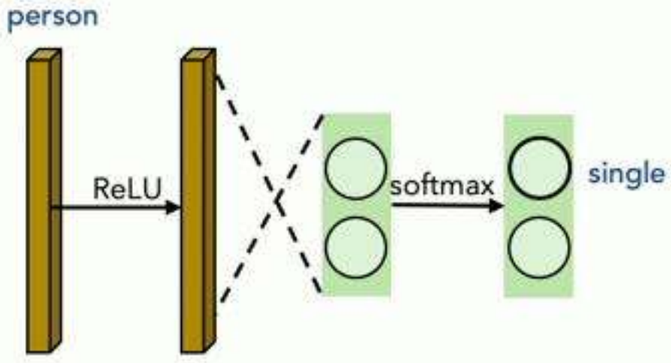
Model

Filling in the slots

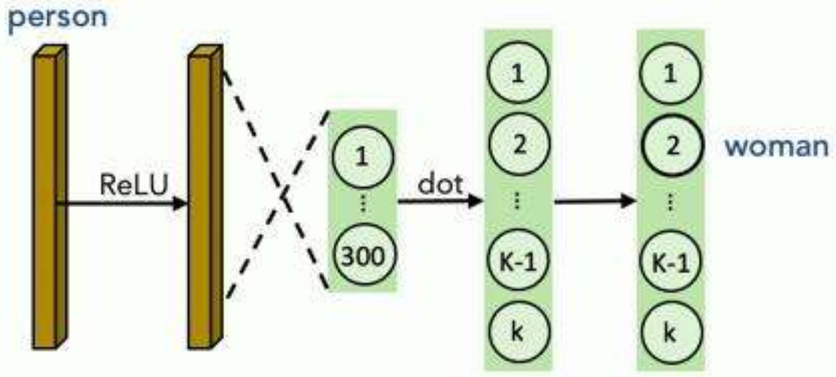
Input: A young <region-2> standing next to a <region-3> holding a <region-5>.

person, person, surfboard (coarse names form the object detector)

1) Classify Plurality



2) Determine Fine Grained Category



Output: A young woman standing next to a man holding a surfboard.

Objective

Minimize this cross-entropy loss

$$L(\boldsymbol{\theta}) = -\sum_{t=1}^{T} \log\left(p(y_t^* | \tilde{r}, y_{1:t-1}^*) p(\tilde{r} | \boldsymbol{y}_{1:t-1}^*) 1_{(y_t^* = y^{txt})} + p(b_t^*, s_t^* | \boldsymbol{r}_t, y_{1:t-1}^*) \left(\frac{1}{m} \sum_{i=1}^{m} p(r_t^i | y_{1:t-1}^*)\right) 1_{(y_t^* = y^{vis})} \right)$$

Objective

Minimize this cross-entropy loss

Template word prob.

$$L(\boldsymbol{\theta}) = -\sum_{t=1}^{T} \log\left(p(y_t^* | \tilde{r}, y_{1:t-1}^*) p(\tilde{r} | \boldsymbol{y}_{1:t-1}^*) 1_{(y_t^* = y^{txt})} + p(b_t^*, s_t^* | \boldsymbol{r}_t, y_{1:t-1}^*) \left(\frac{1}{m} \sum_{i=1}^{m} p(r_t^i | y_{1:t-1}^*)\right) 1_{(y_t^* = y^{vis})} \right)$$

Objective

Minimize this cross-entropy loss

Template word prob. ***Refinement prob.***

$$L(\boldsymbol{\theta}) = -\sum_{t=1}^{T} \log\left(p(y_t^* | \tilde{r}, y_{1:t-1}^*) p(\tilde{r} | \boldsymbol{y}_{1:t-1}^*) 1_{(y_t^* = y^{txt})} + p(b_t^*, s_t^* | \boldsymbol{r}_t, y_{1:t-1}^*) \left(\frac{1}{m} \sum_{i=1}^{m} p(r_t^i | y_{1:t-1}^*)\right) 1_{(y_t^* = y^{vis})} \right)$$

Objective

Minimize this cross-entropy loss

Template word prob. ***Refinement prob.*** ***target region prob.***

$$L(\boldsymbol{\theta}) = -\sum_{t=1}^{T} \log\left(p(y_t^*|\tilde{r}, y_{1:t-1}^*) p(\tilde{r}|\boldsymbol{y}_{1:t-1}^*) 1_{(y_t^*=y^{txt})} + p(b_t^*, s_t^*|\boldsymbol{r}_t, y_{1:t-1}^*) \left(\frac{1}{m}\sum_{i=1}^{m} p(r_t^i|y_{1:t-1}^*)\right) 1_{(y_t^*=y^{vis})} \right)$$

Objective

Minimize this cross-entropy loss

$$L(\boldsymbol{\theta}) = -\sum_{t=1}^{T} \log\Big(\underbrace{p(y_t^*|\tilde{r}, y_{1:t-1}^*)p(\tilde{r}|\boldsymbol{y}_{1:t-1}^*)1_{(y_t^*=y^{txt})}}_{\textit{Template word prob.}} + \underbrace{p(b_t^*, s_t^*|\boldsymbol{r}_t, y_{1:t-1}^*)}_{\textit{Refinement prob.}} \underbrace{(\frac{1}{m}\sum_{i=1}^{m} p(r_t^i|y_{1:t-1}^*))1_{(y_t^*=y^{vis})}}_{\textit{target region prob.}} \Big)$$

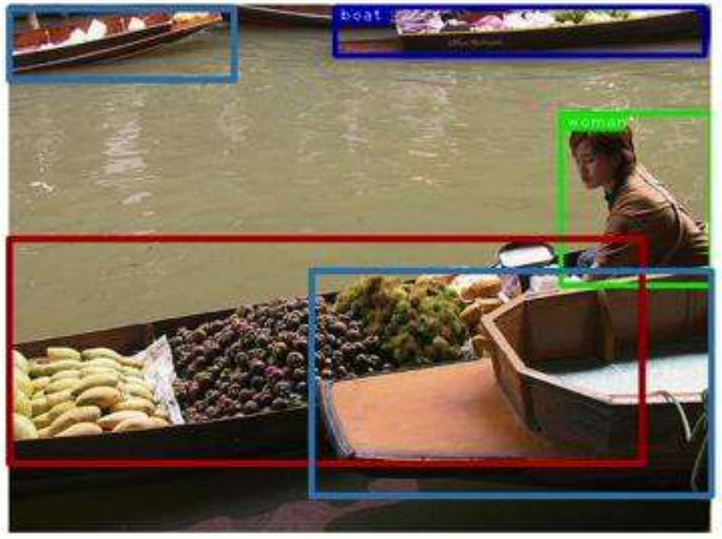
Co-reference when different kind of supervision exists.

Objective

Minimize this cross-entropy loss

$$L(\boldsymbol{\theta}) = -\sum_{t=1}^{T} \log\left(\underbrace{p(y_t^*|\tilde{r}, y_{1:t-1}^*)p(\tilde{r}|\boldsymbol{y}_{1:t-1}^*)1_{(y_t^*=y^{txt})}}_{\textbf{Template word prob.}} + \underbrace{p(b_t^*, s_t^*|\boldsymbol{r}_t, y_{1:t-1}^*)}_{\textbf{Refinement prob.}} \underbrace{\left(\frac{1}{m}\sum_{i=1}^{m} p(r_t^i|y_{1:t-1}^*)\right)1_{(y_t^*=y^{vis})}}_{\textbf{target region prob.}} \right)$$

Co-reference when different kind of supervision exists.



A young woman is sitting inside a boat.

Objective

Minimize this cross-entropy loss

$$L(\boldsymbol{\theta}) = -\sum_{t=1}^{T} \log\Big(\underbrace{p(y_t^*|\tilde{r}, y_{1:t-1}^*)p(\tilde{r}|\boldsymbol{y}_{1:t-1}^*)1_{(y_t^*=y^{txt})}}_{\textbf{\textit{Template word prob.}}} + \underbrace{p(b_t^*, s_t^*|\boldsymbol{r}_t, y_{1:t-1}^*)}_{\textbf{\textit{Refinement prob.}}} \underbrace{(\frac{1}{m}\sum_{i=1}^{m} p(r_t^i|y_{1:t-1}^*))1_{(y_t^*=y^{vis})}}_{\textbf{\textit{target region prob.}}} \Big)$$

Co-reference when different kind of supervision exists.

A young woman is sitting inside a boat.

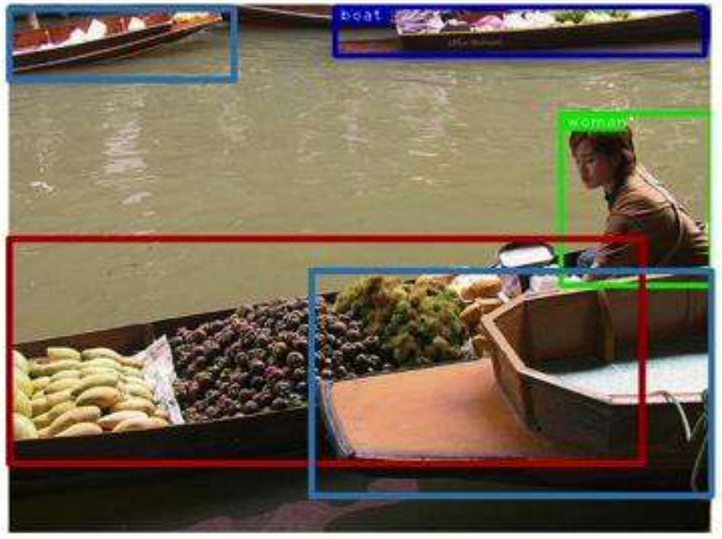
Problem: which boat is the caption referred to.

Objective

Minimize this cross-entropy loss

$$L(\boldsymbol{\theta}) = -\sum_{t=1}^{T} \log\left(\underbrace{p(y_t^*|\tilde{r}, y_{1:t-1}^*)p(\tilde{r}|\boldsymbol{y}_{1:t-1}^*)1_{(y_t^*=y^{txt})}}_{\textbf{Template word prob.}} + \underbrace{p(b_t^*, s_t^*|\boldsymbol{r}_t, y_{1:t-1}^*)}_{\textbf{Refinement prob.}} \underbrace{\left(\frac{1}{m}\sum_{i=1}^{m} p(r_t^i|y_{1:t-1}^*)\right)1_{(y_t^*=y^{vis})}}_{\textbf{target region prob.}} \right)$$

Co-reference when different kind of supervision exists.



A young woman is sitting inside a boat.

Problem: which boat is the caption referred to.

Solution: maximize the averaged target region prob.

Dataset

Datasets

- Flickr30k: 31,783 images, 5 captions per image, 275,555 annotated bounding boxes.
- COCO: 164,062 images, 5 captions per image.

Object category to words

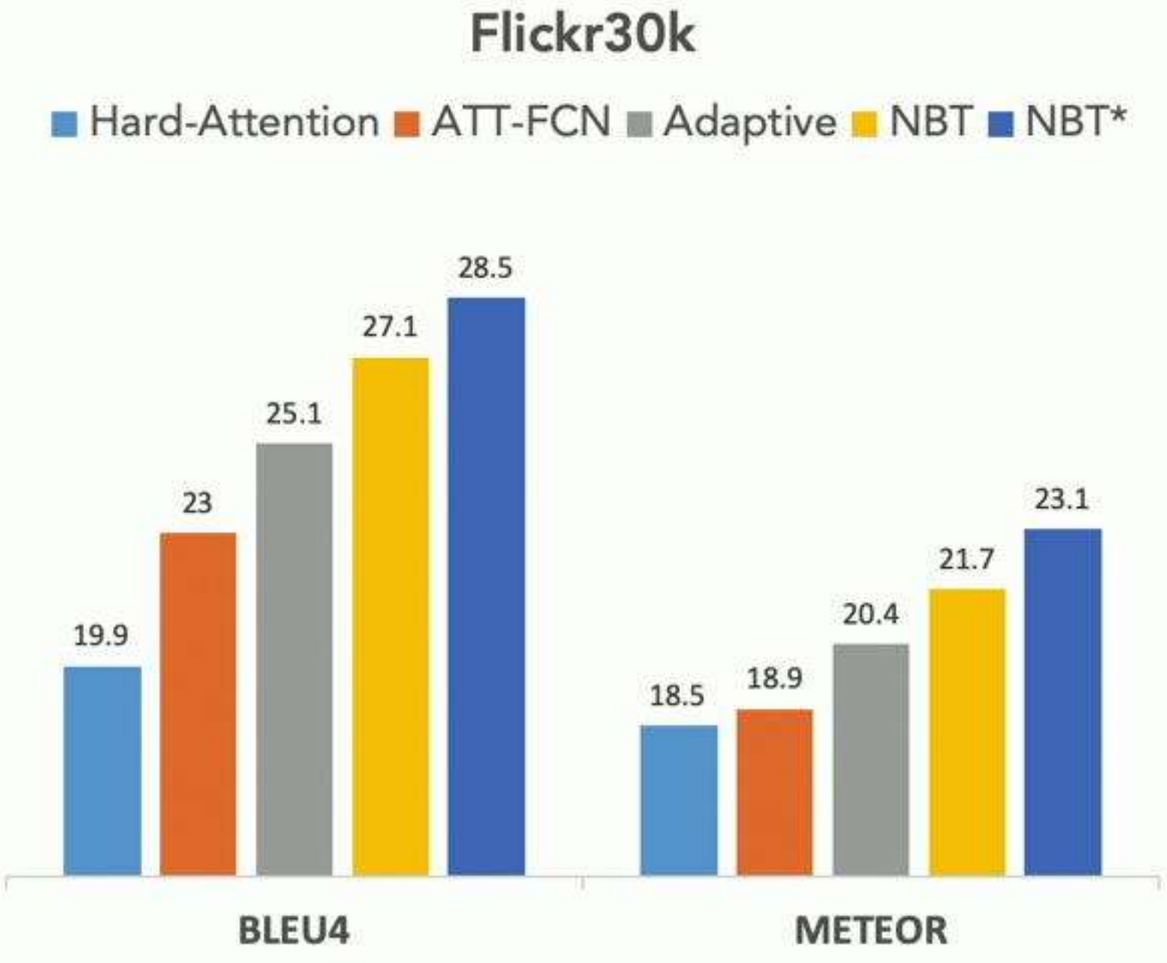
- For COCO dataset. (e.g., “person” mapping to [“child”, “baker”, …])

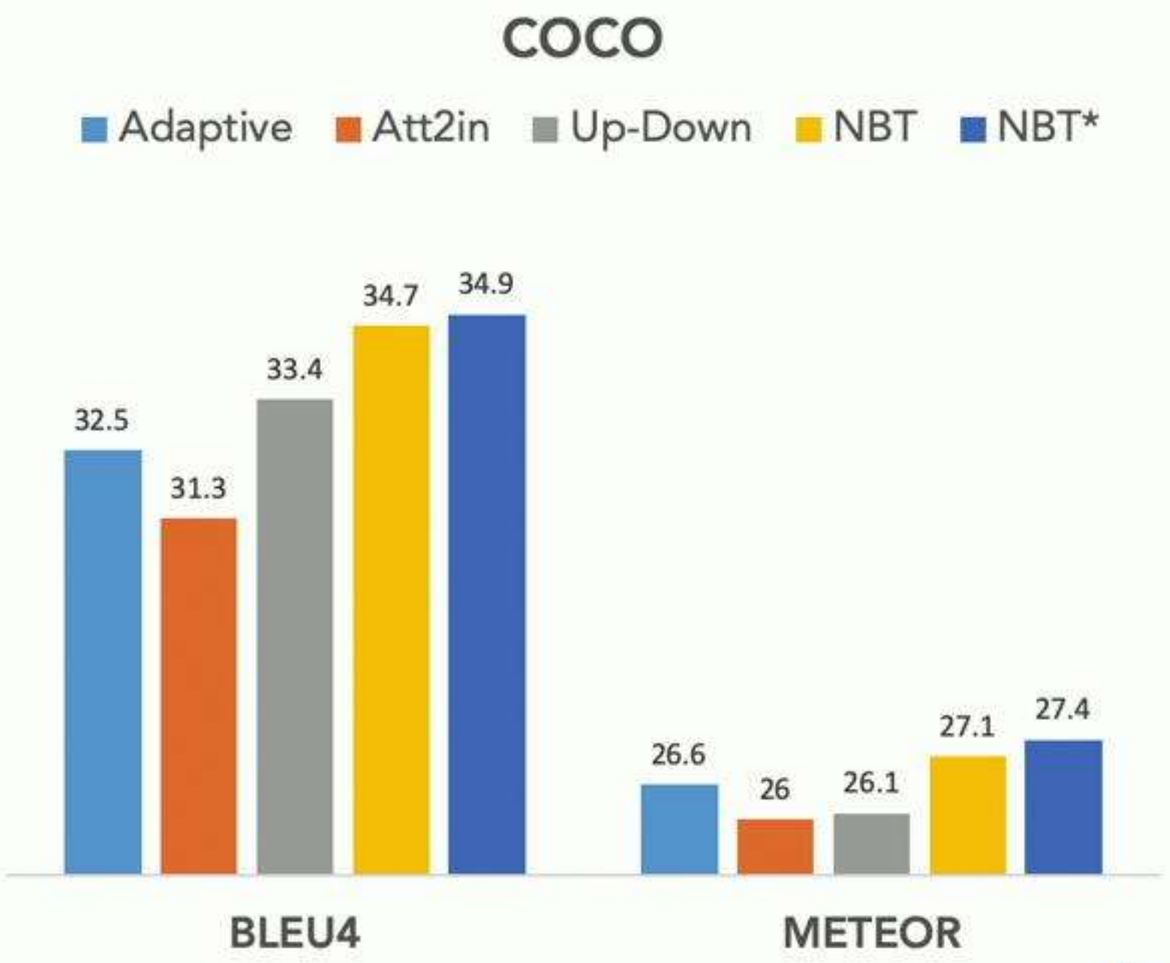
Caption pre-processing

- Caption truncation (if > 16 words)
- Building vocabulary (9,587 words for COCO, 6,864 words for Flickr30k)

Results

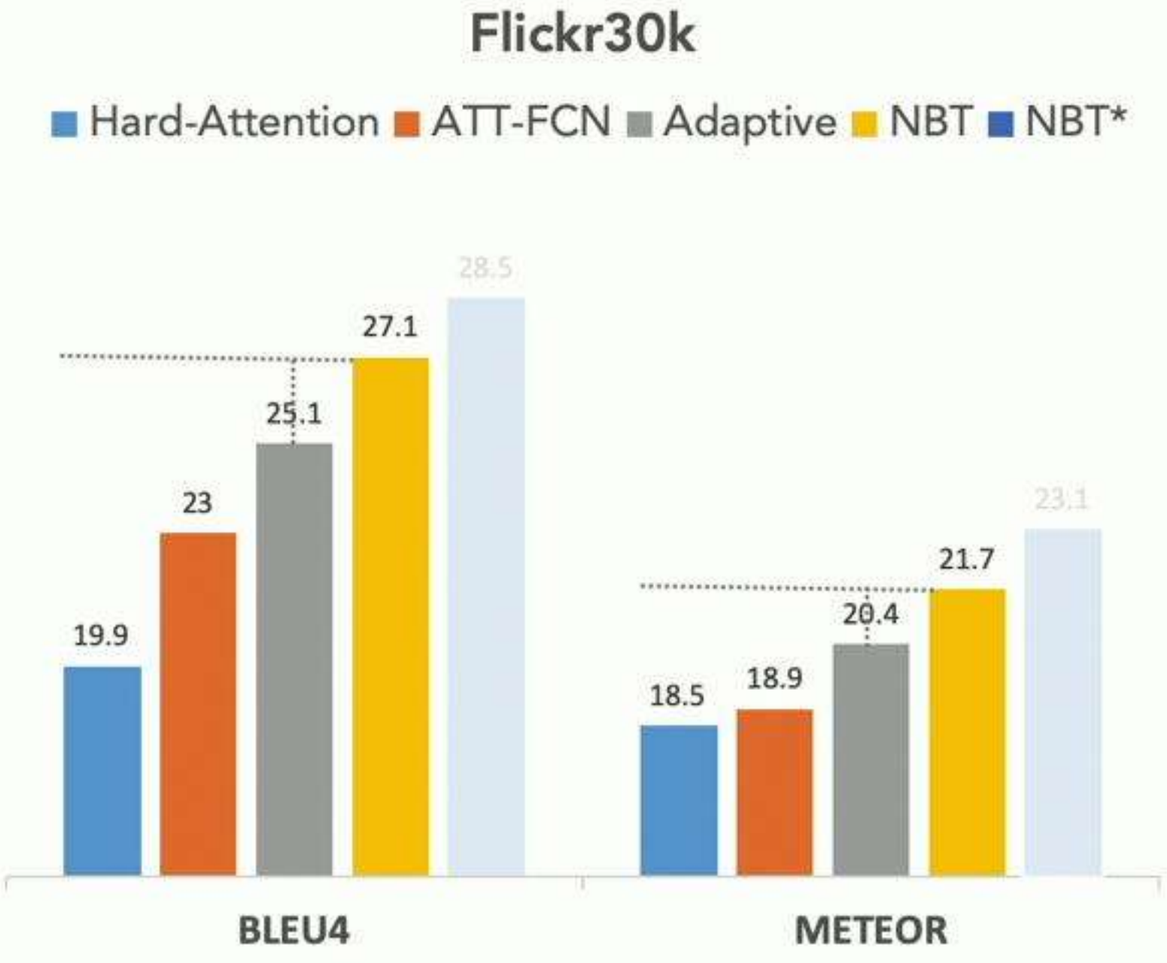
Standard Image Captioning

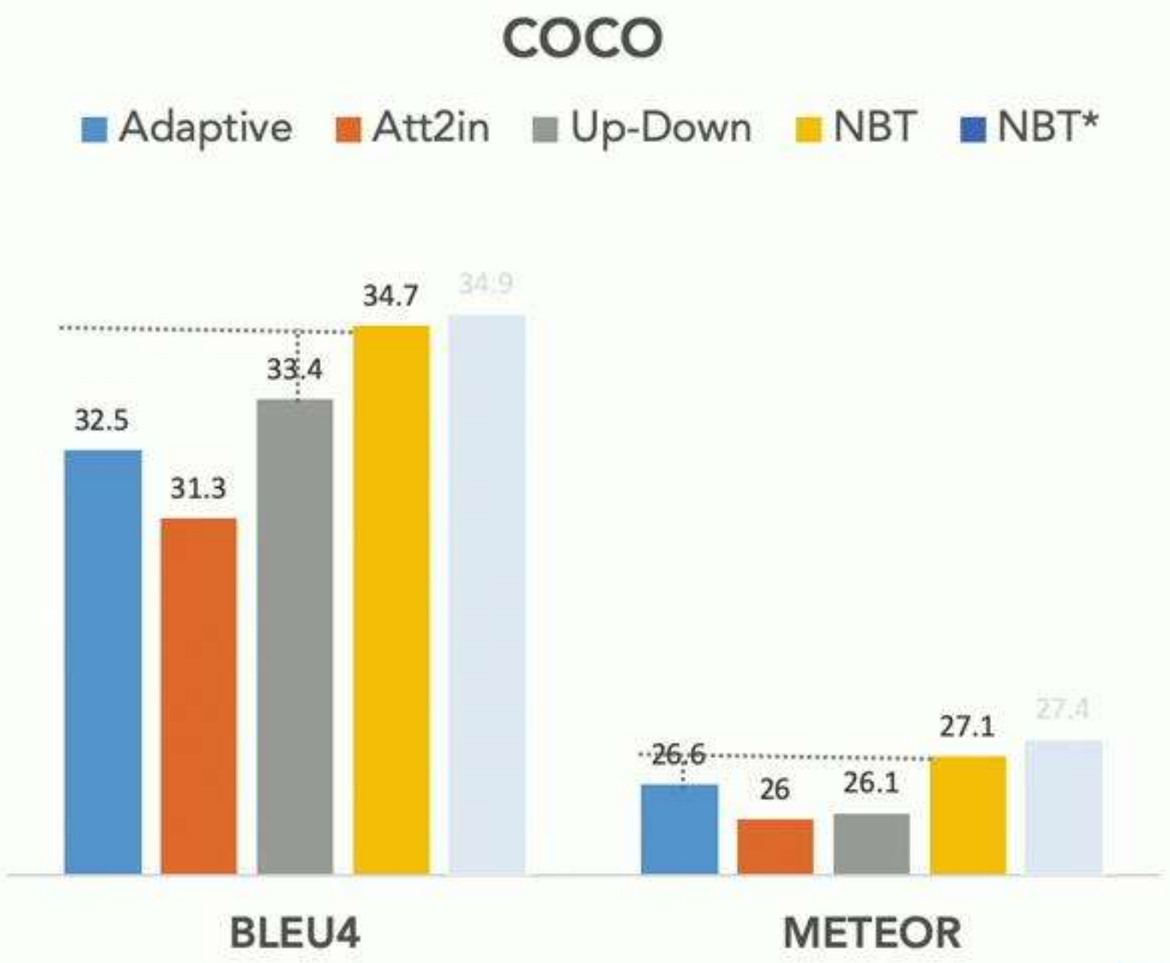




Results

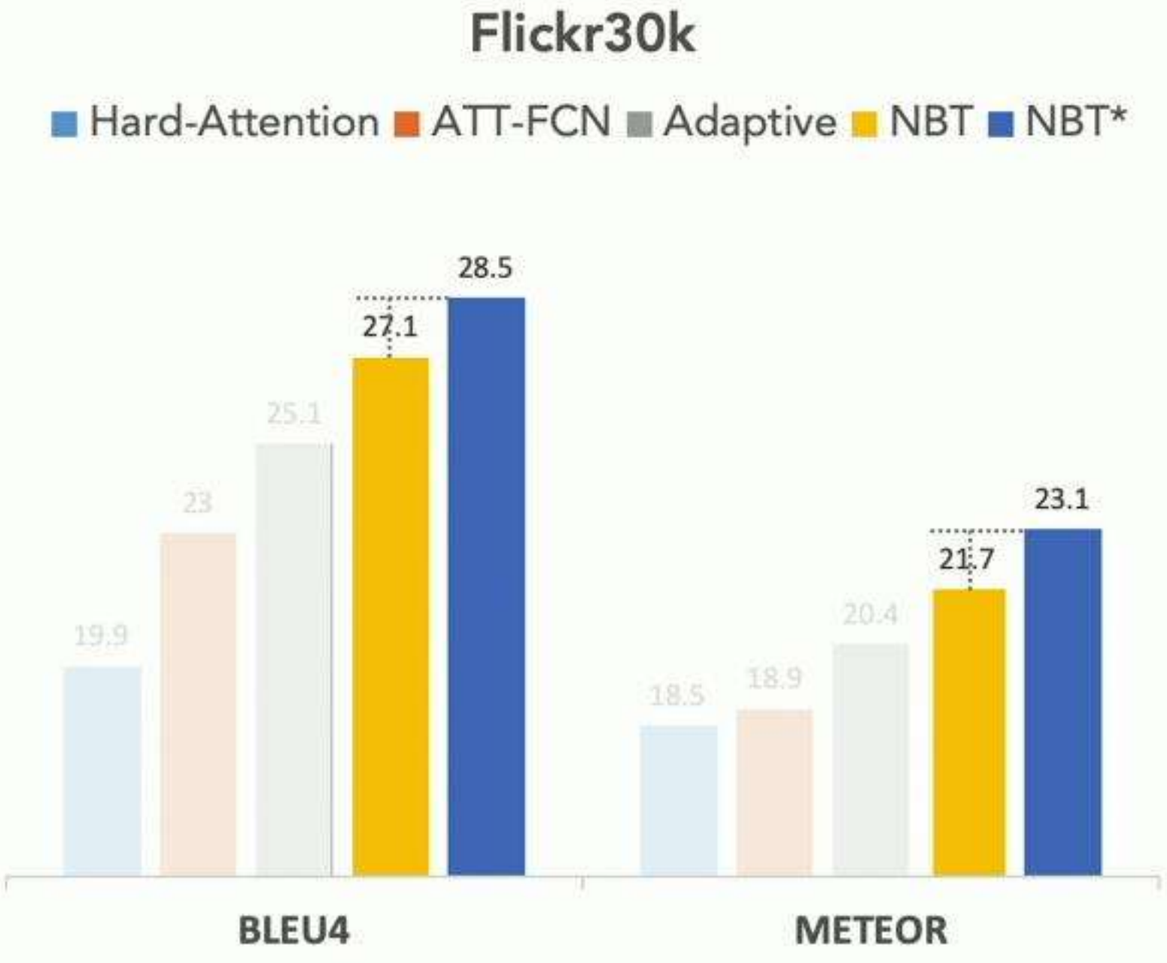
Standard Image Captioning

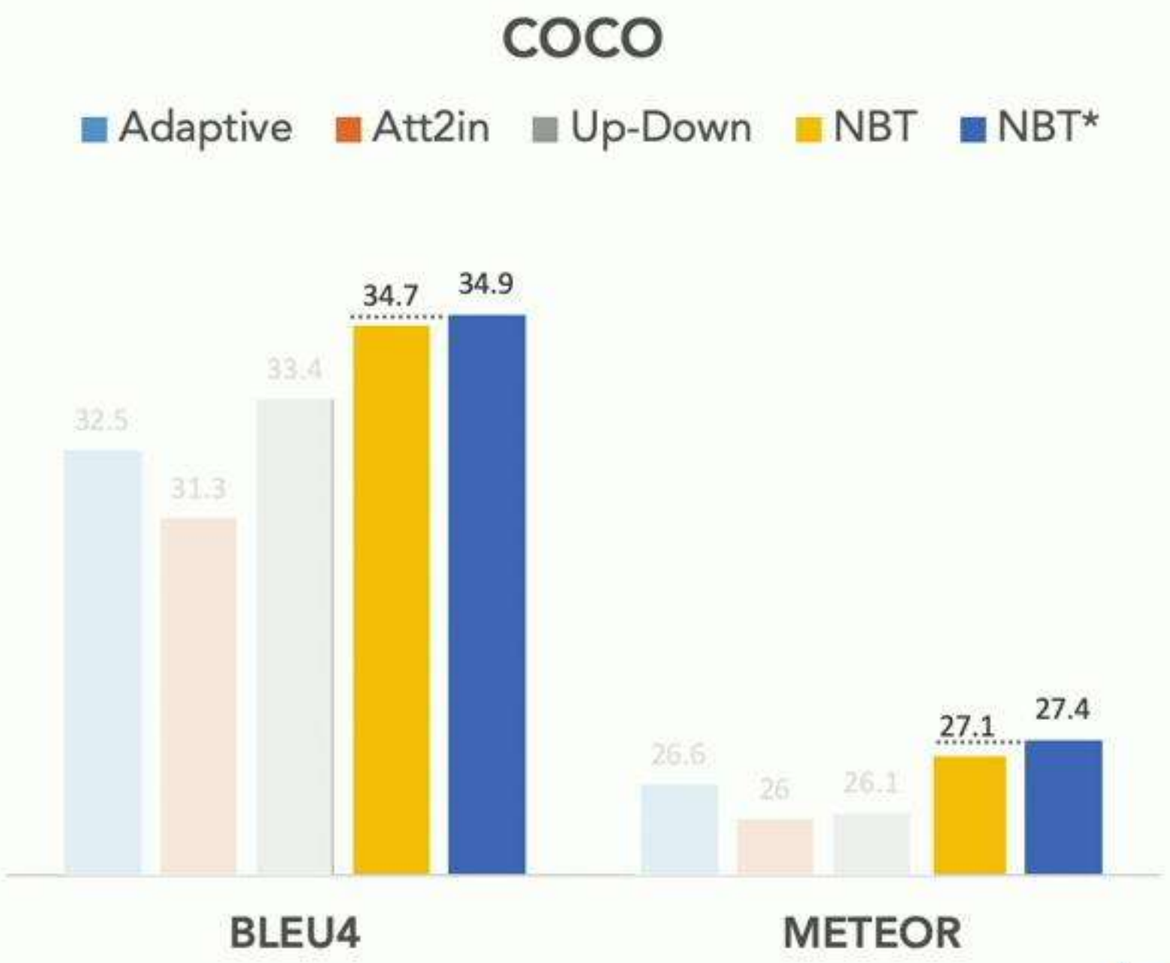




Results

Standard Image Captioning





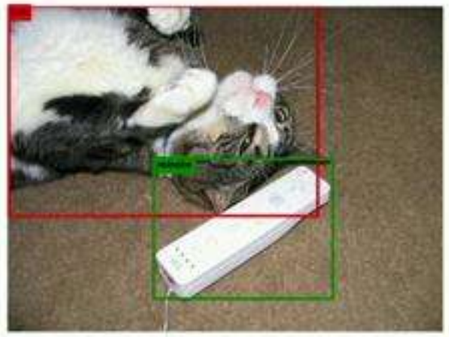
Results

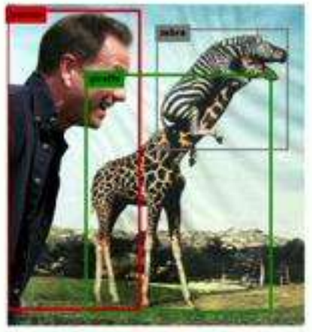
Robust Image Captioning

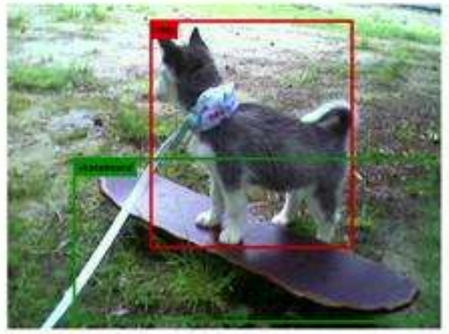
Results

Robust Image Captioning

- To evaluate image captioning for novel scene compositions.



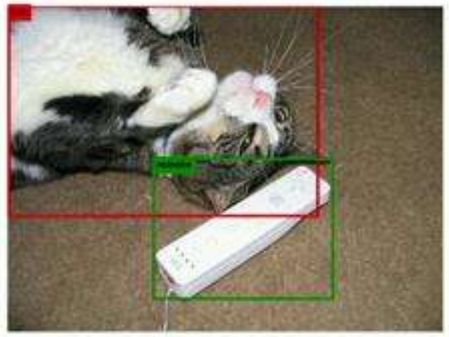
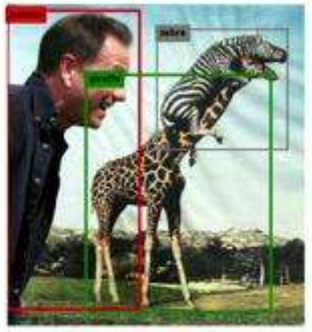
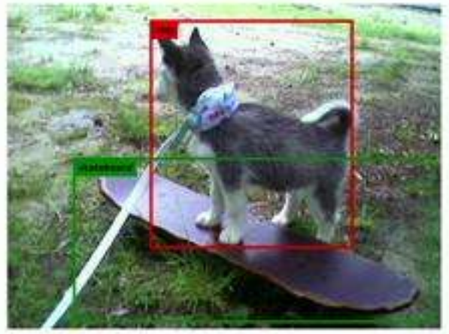




Results

Robust Image Captioning

- To evaluate image captioning for novel scene compositions.

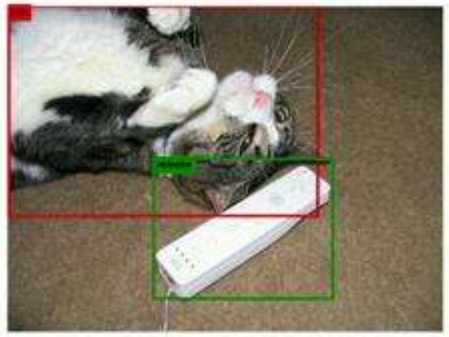
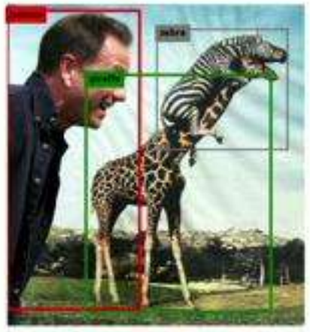
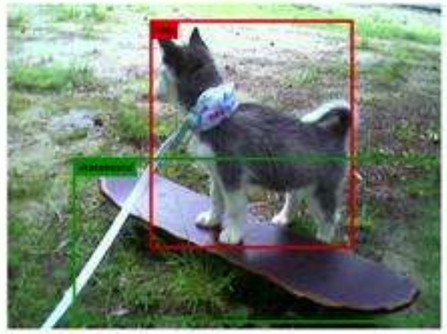
Robust-COCO split

- Distribution of co-occurring objects in train data is different from test data.
- Sufficient examples from each category in train set.
- Novel compositions (pairs) of categories in test set.

Results

Robust Image Captioning

- To evaluate image captioning for novel scene compositions.

Robust-COCO split

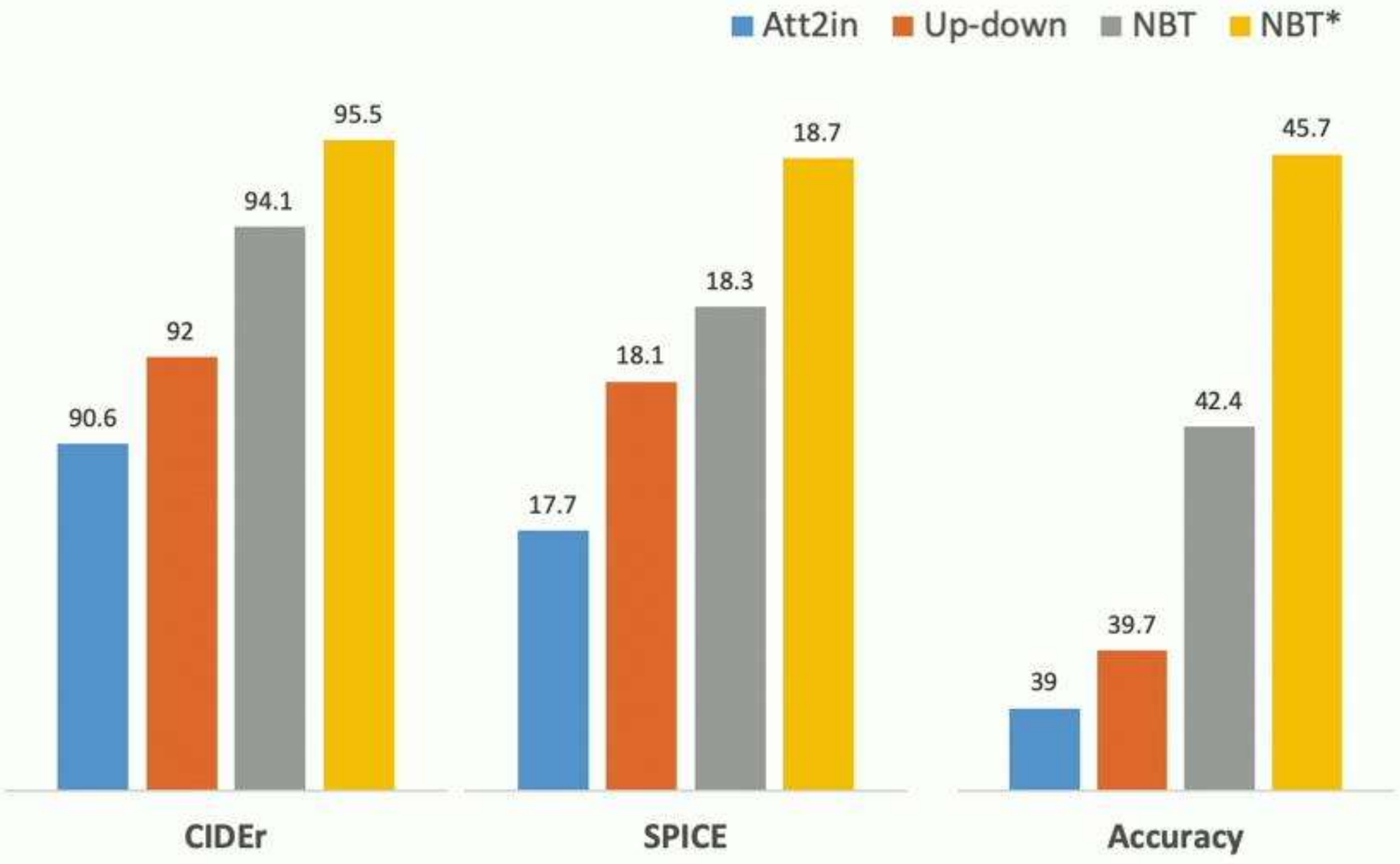
- Distribution of co-occurring objects in train data is different from test data.
- Sufficient examples from each category in train set.
- Novel compositions (pairs) of categories in test set.

Accuracy (new metric)

- Whether or not a generated caption includes the new object combination.
- 100% accuracy for at least one mention of the novel category pair.

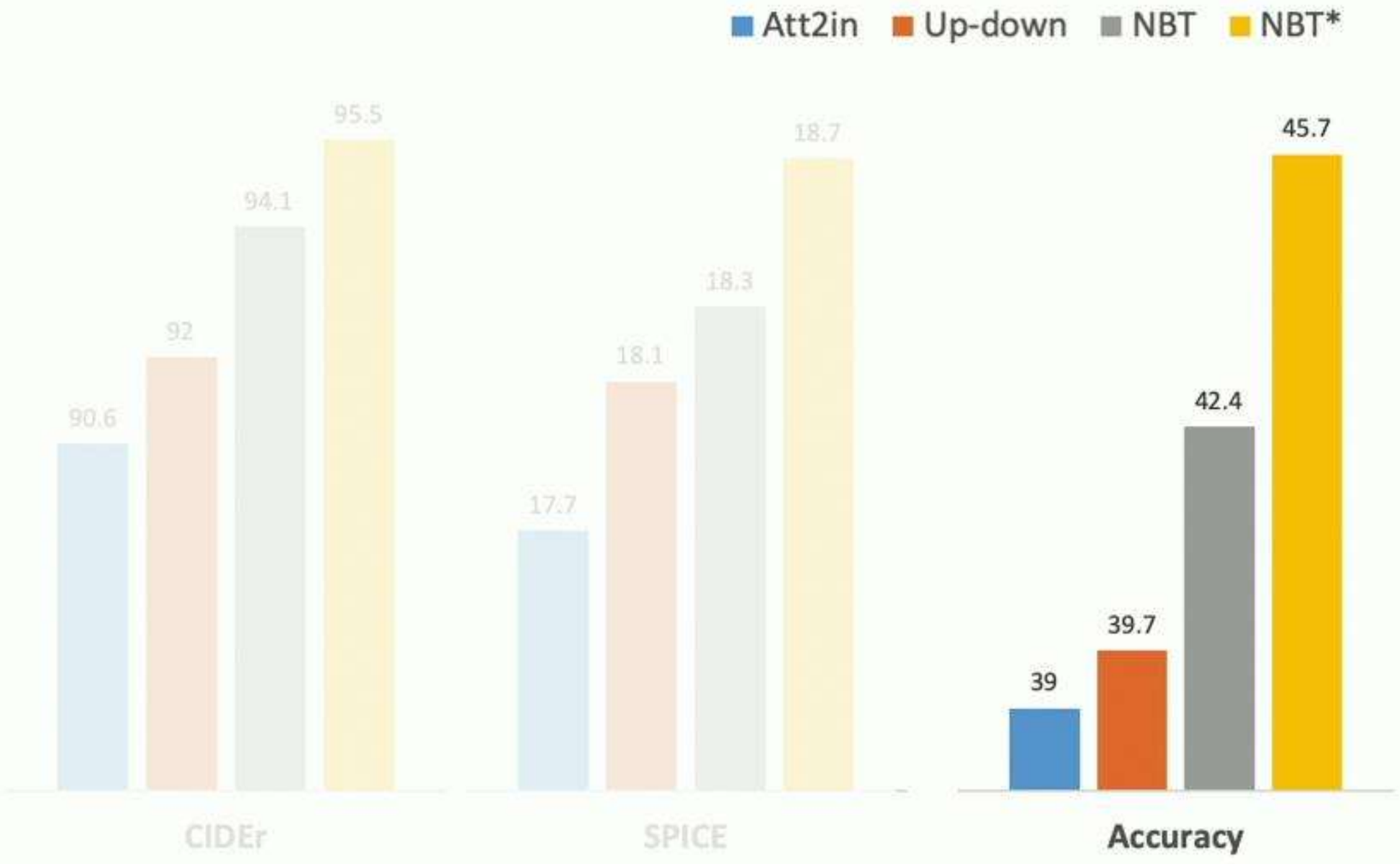
Results

Robust Image Captioning



Results

Robust Image Captioning



Results

Novel Object Captioning *[Hendricks et.al.]*

Results

Novel Object Captioning *[Hendricks et.al.]*

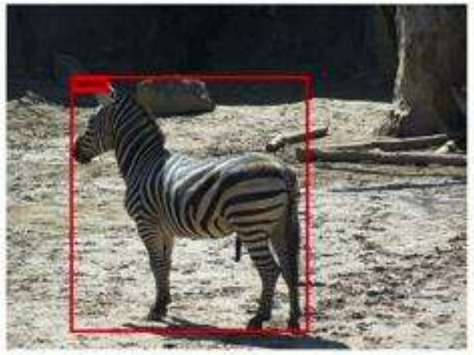
- Describe image with novel objects.

Results

Novel Object Captioning *[Hendricks et.al.]*

- Describe image with novel objects.
- Excludes all the pairs that contain at least one of the eight objects in COCO

("bottle", "bus", "couch", "microwave", "pizza", "racket", "suitcase", and "zebra")

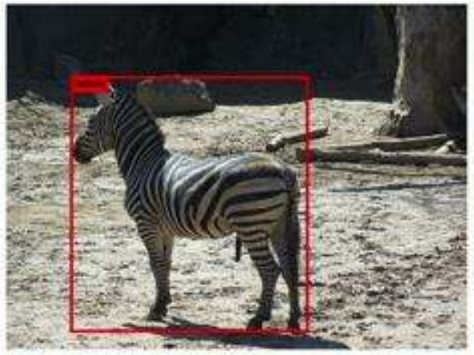


Results

Novel Object Captioning *[Hendricks et.al.]*

- Describe image with novel objects.
- Excludes all the pairs that contain at least one of the eight objects in COCO

("bottle", "bus", "couch", "microwave", "pizza", "racket", "suitcase", and "zebra")



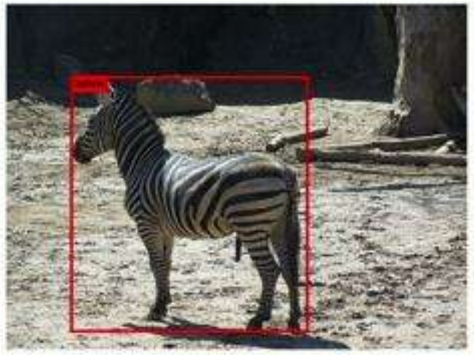
- Test set is split into in-domain and out-of-domain subsets.

Results

Novel Object Captioning *[Hendricks et.al.]*

- Describe image with novel objects.
- Excludes all the pairs that contain at least one of the eight objects in COCO

("bottle", "bus", "couch", "microwave", "pizza", "racket", "suitcase", and "zebra")

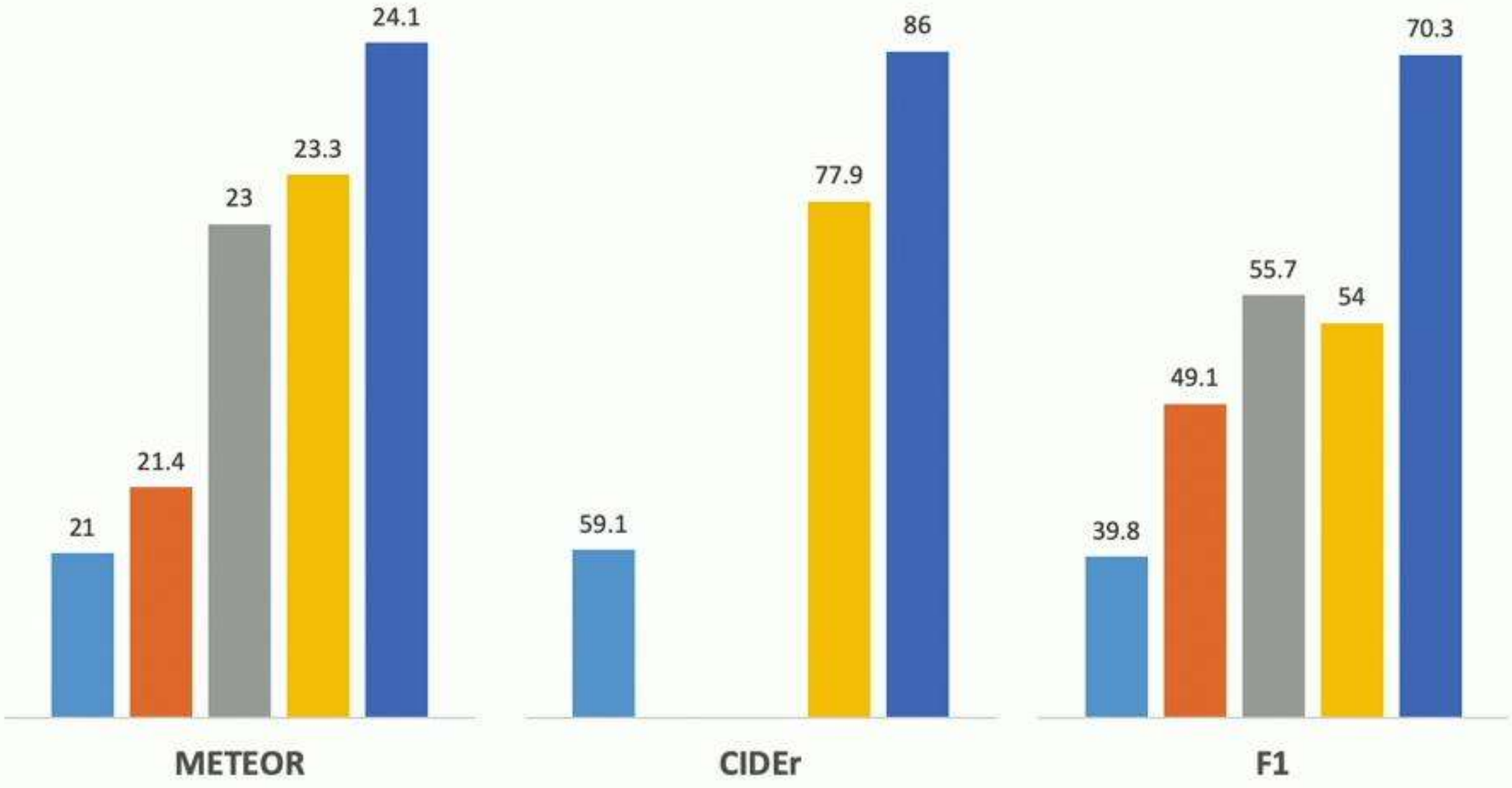


- Test set is split into in-domain and out-of-domain subsets.
- F1 score (metric)

  Checks if the excluded object is correctly mentioned in the generated caption.

Results

Novel Object Captioning

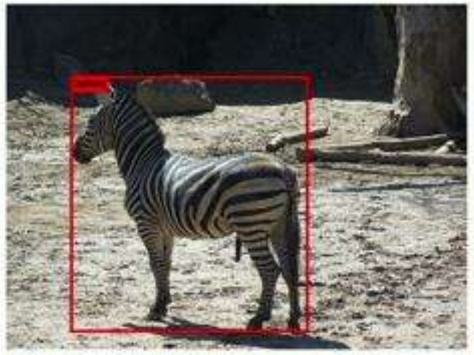


Results

Novel Object Captioning *[Hendricks et.al.]*

- Describe image with novel objects.
- Excludes all the pairs that contain at least one of the eight objects in COCO

("bottle", "bus", "couch", "microwave", "pizza", "racket", "suitcase", and "zebra")

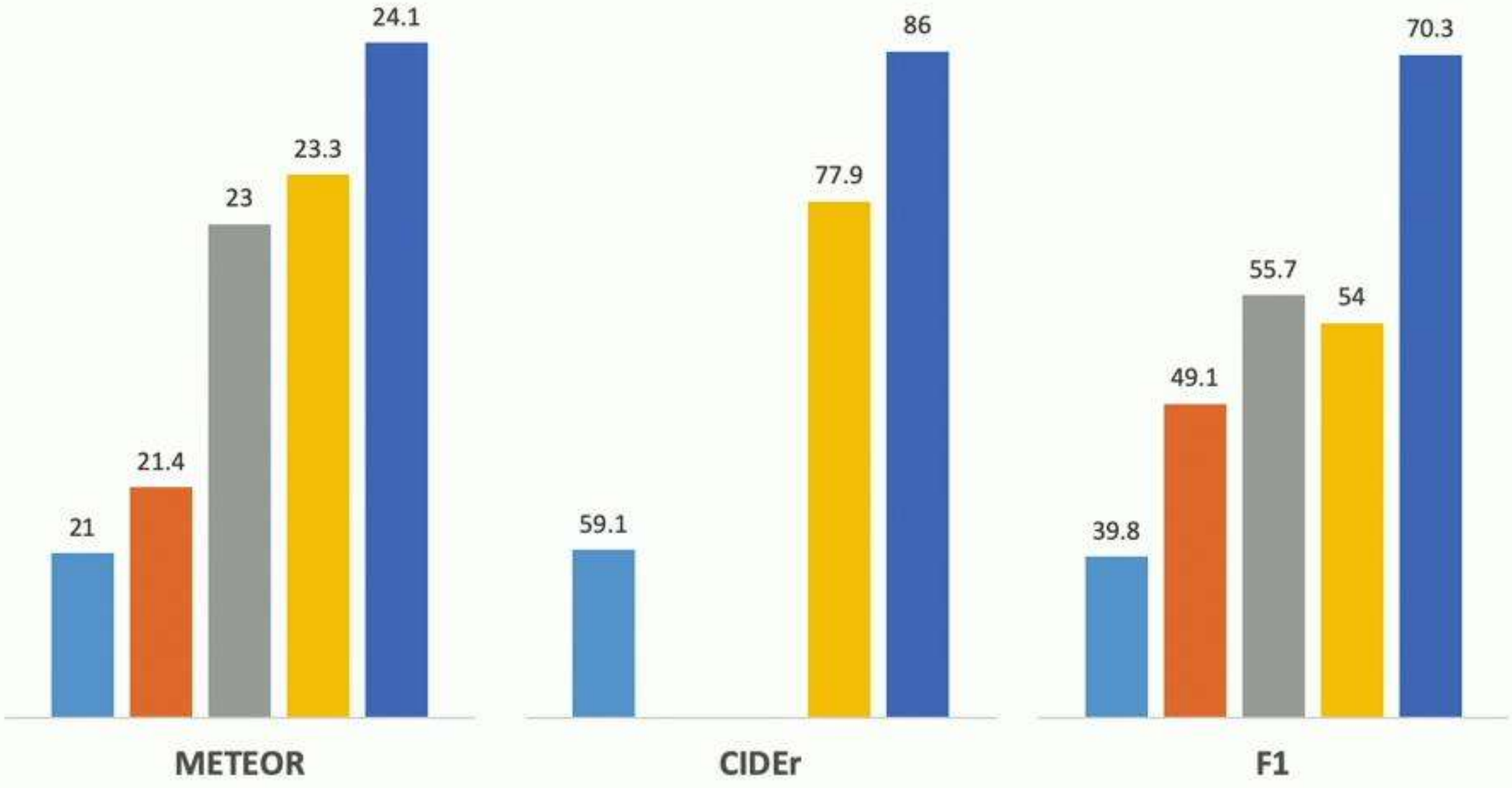


- Test set is split into in-domain and out-of-domain subsets.
- F1 score (metric)

  Checks if the excluded object is correctly mentioned in the generated caption.

Results

Novel Object Captioning



Results

Summary

Two-stage approach for image captioning

- generate hybrid template
- fills the slots with categories recognized by object detector

Results

Summary

Two-stage approach for image captioning

- generate hybrid template
- fills the slots with categories recognized by object detector

Limitation

Results

Summary

Two-stage approach for image captioning

- generate hybrid template
- fills the slots with categories recognized by object detector

Limitation

Only ground on noun words which found by object detector

- Extend to attributes, actions, OCR etc.

Results

Summary

Two-stage approach for image captioning

- generate hybrid template
- fills the slots with categories recognized by object detector

Limitation

Only ground on noun words which found by object detector

- Extend to attributes, actions, OCR etc.

Limited by the image captioning dataset.

- Learn from other data sources -- ViLBERT.

This Talk

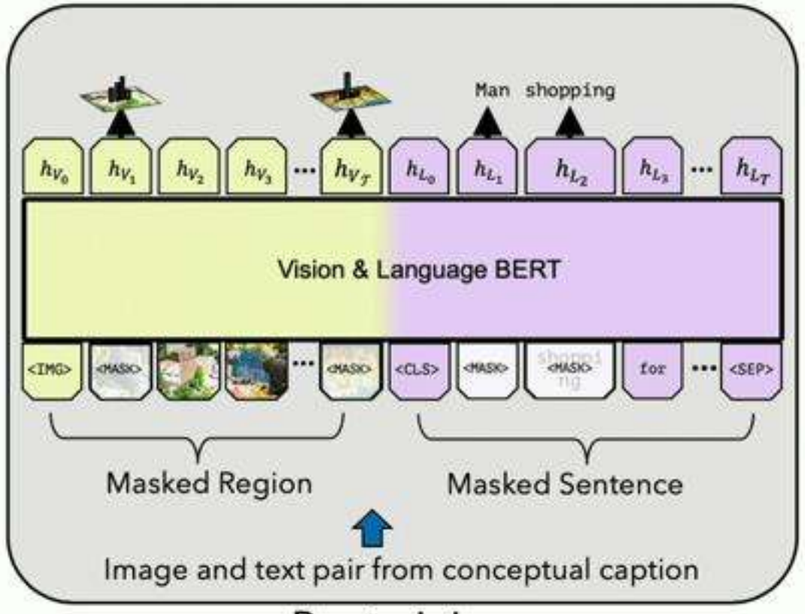
ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks

NeurIPS 2019

This Talk

ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks

NeurIPS 2019

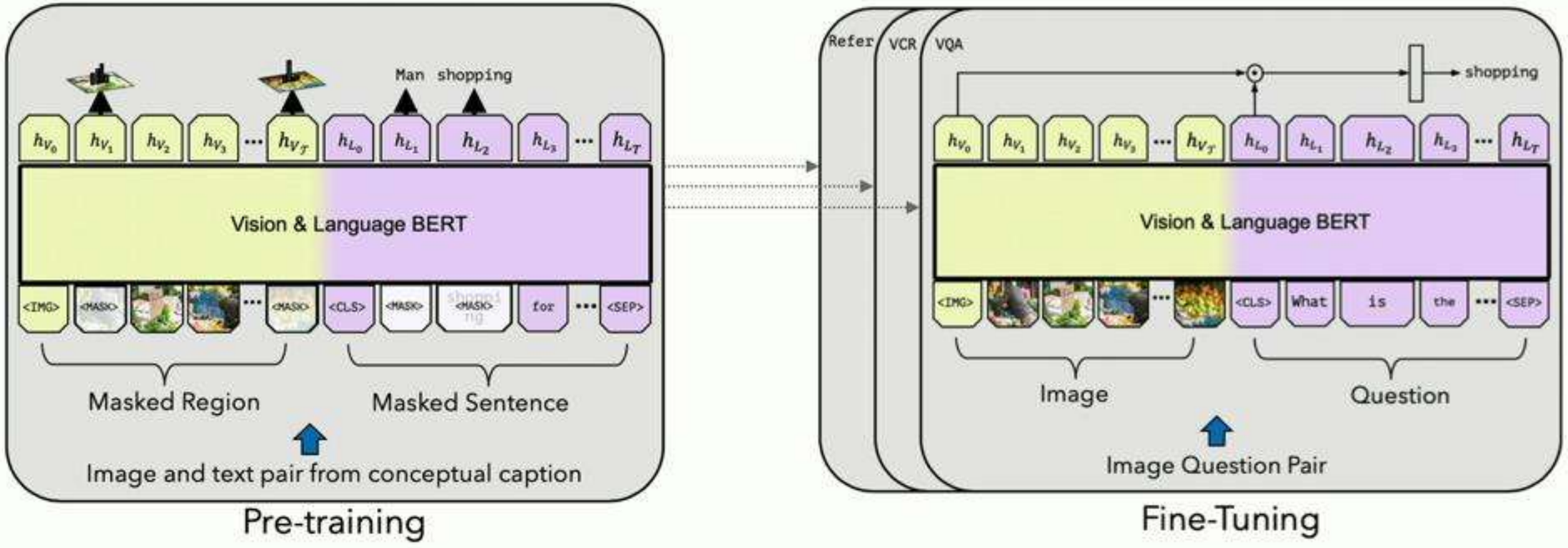


Pre-training

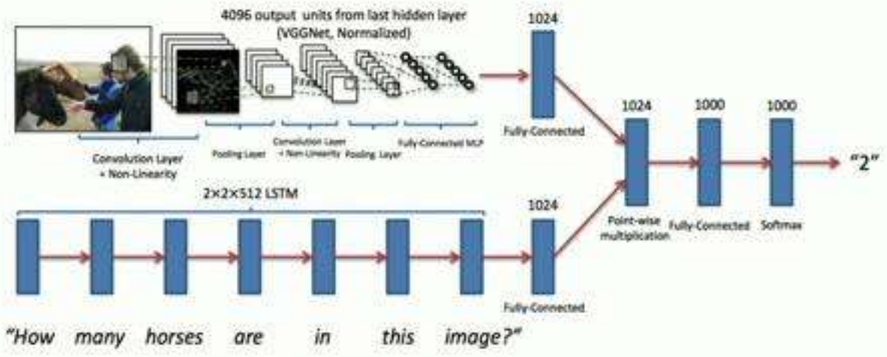
This Talk

ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks

NeurIPS 2019



Vision and Language



Visual Question Answering

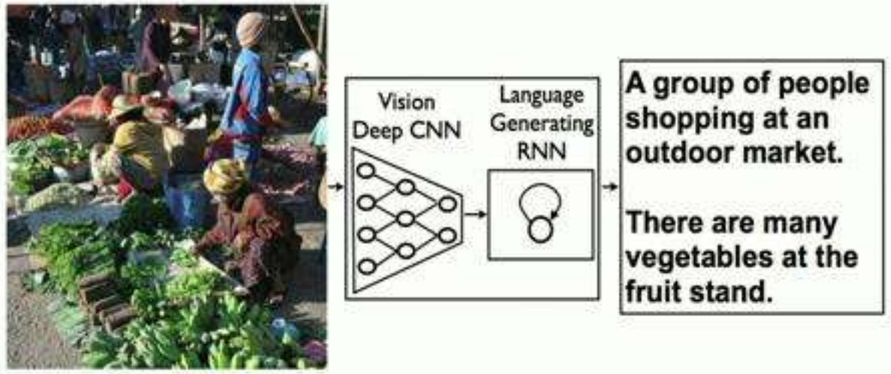
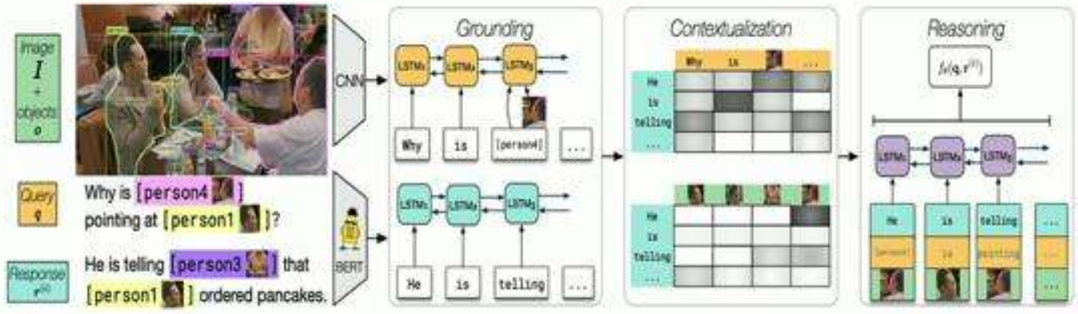
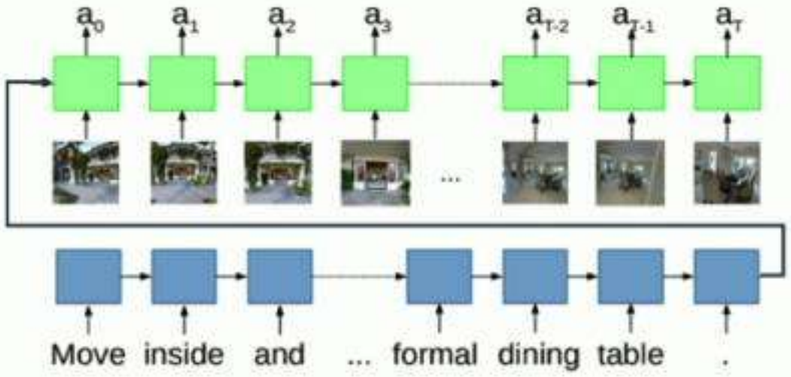


Image Captioning

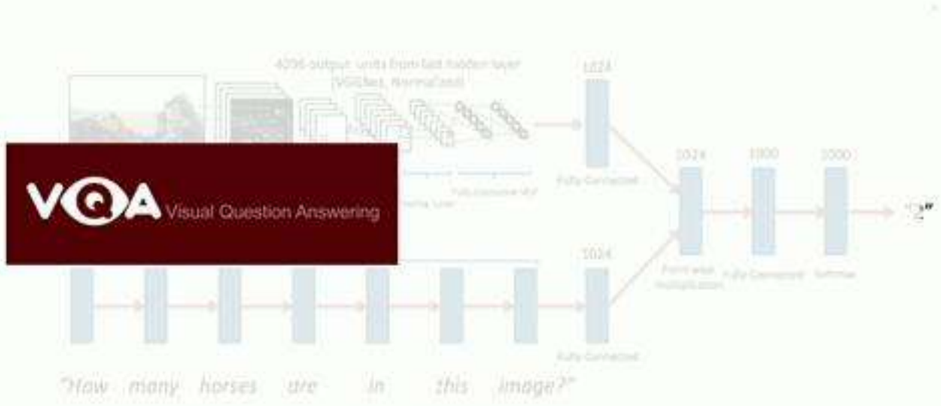


Visual Commonsense Reasoning



Vision and Language Navigation

Vision and Language



Visual Question Answering

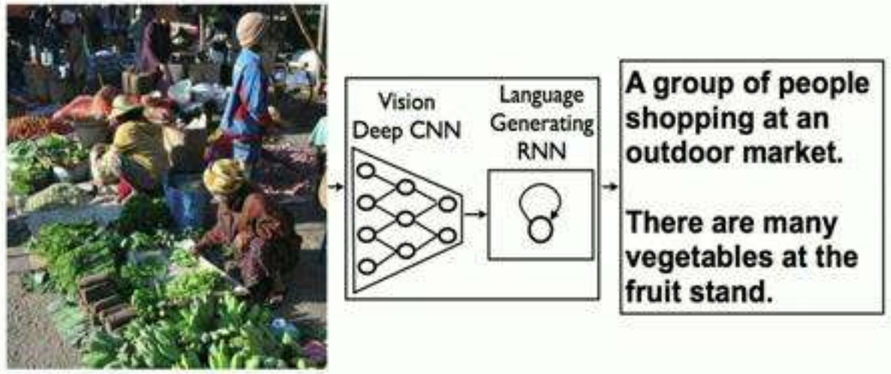
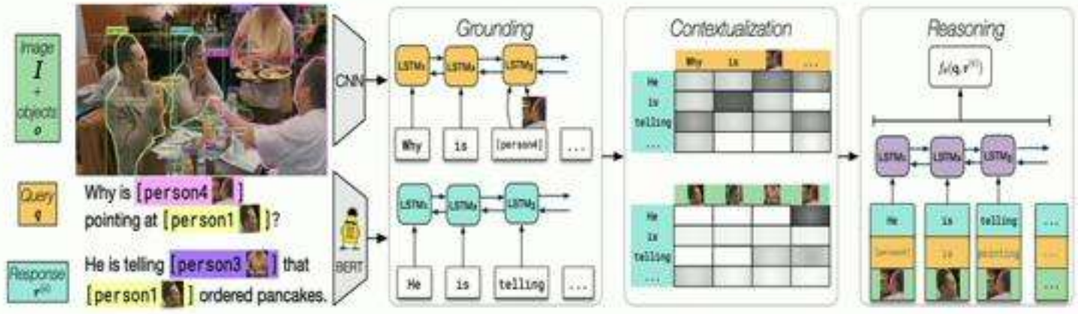
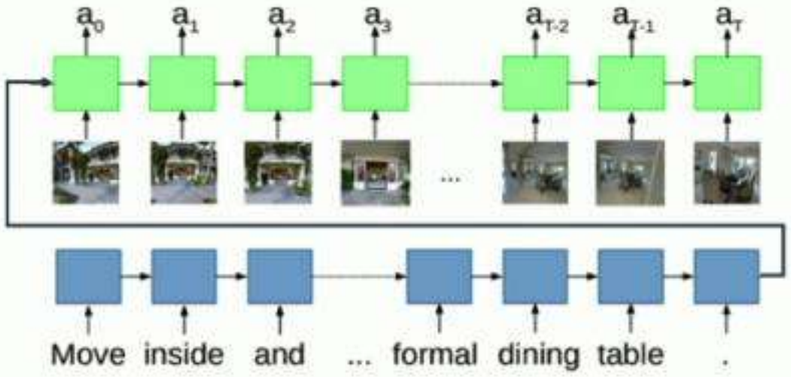


Image Captioning

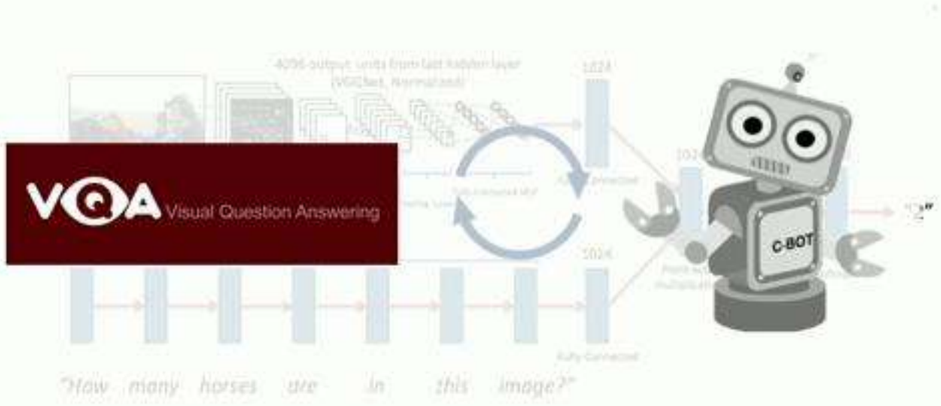


Visual Commonsense Reasoning



Vision and Language Navigation

Vision and Language



Visual Question Answering

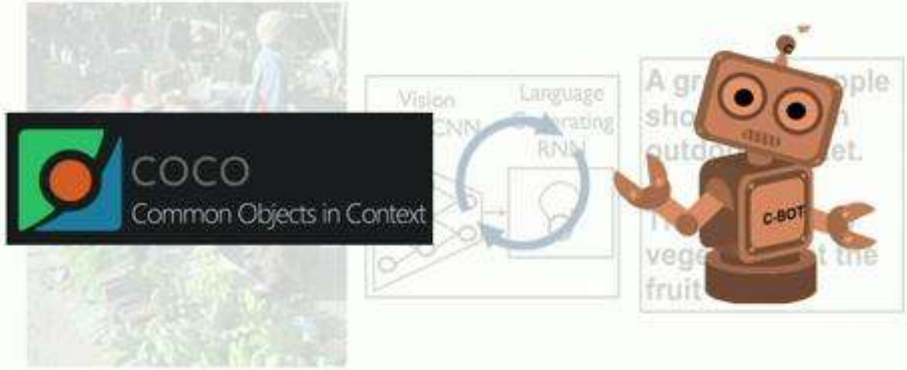
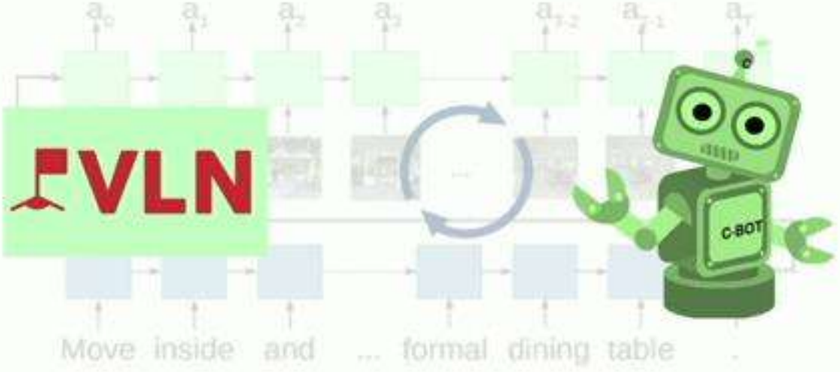


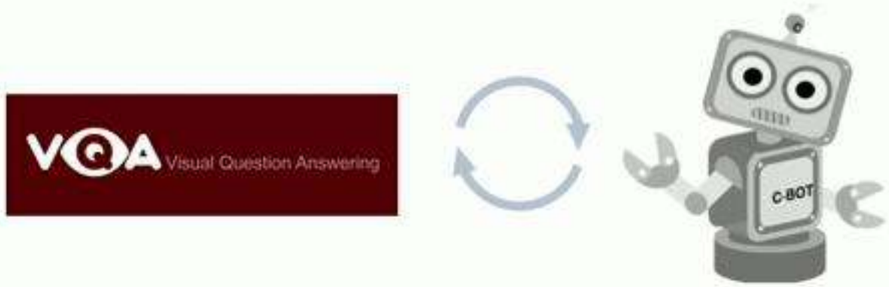
Image Captioning

Visual Commonsense Reasoning



Vision and Language Navigation

Visual Grounding



[Shen et.al 2018]

Visual Grounding

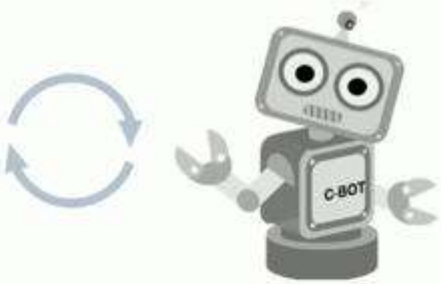
VQA Visual Question Answering

COCO Common Objects in Context

[Shen et.al 2018]

Visual Grounding

VQA Visual Question Answering



COCO Common Objects in Context

Q: What type of plant is this?

[Shen et.al 2018]

Visual Grounding

VQA Visual Question Answering

COCO Common Objects in Context

Q: What type of plant is this?

A: Banana

C: A bunch of red and yellow flowers on a branch.

[Shen et.al 2018]

Visual Grounding

VQA Visual Question Answering

COCO Common Objects in Context

Q: What type of plant is this?
A: Banana

C: A bunch of red and yellow flowers on a branch.

Common model for visual grounding and leverage them on a wide array of vision-and-language tasks

[Shen et.al 2018]

Pretrain-Transfer

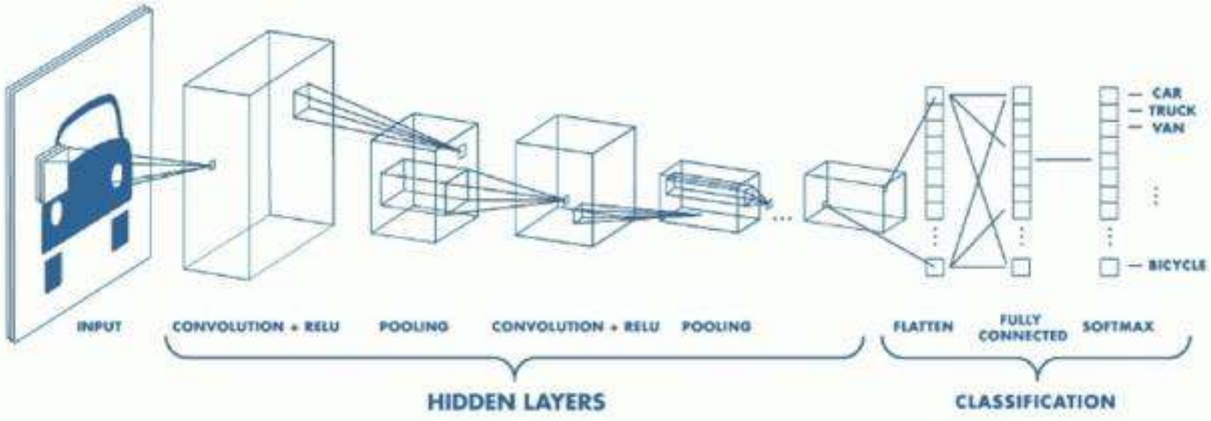
[Deng et.al 2009, Devlin 2018]

Pretrain-Transfer

IMAGENET

[Deng et.al 2009, Devlin 2018]

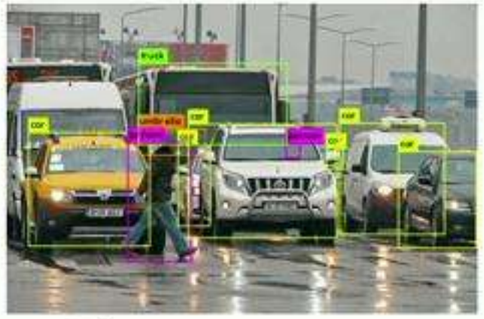
Pretrain-Transfer



[Deng et.al 2009, Devlin 2018]

Pretrain-Transfer

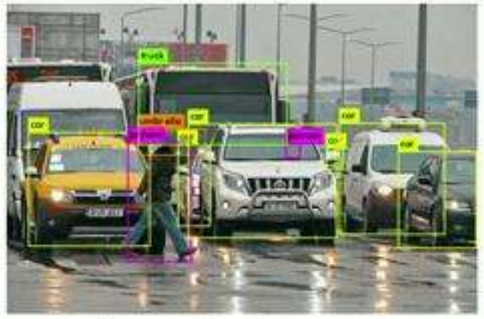
IMAGENET



Object Detection

[Deng et.al 2009, Devlin 2018]

Pretrain-Transfer



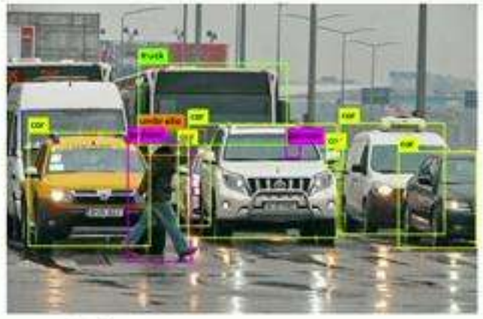
Object Detection

Semantic Segmentation

[Deng et.al 2009, Devlin 2018]

Pretrain-Transfer

IMAGENET



Object Detection

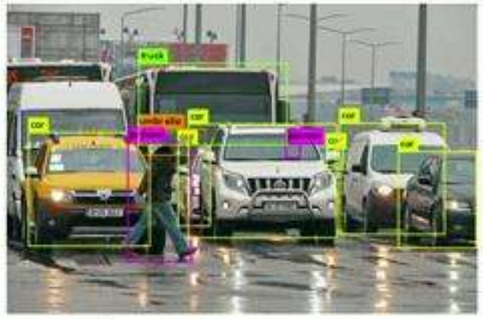
Semantic Segmentation

Pose Estimation

[Deng et.al 2009, Devlin 2018]

Pretrain-Transfer

IMAGENET



Object Detection

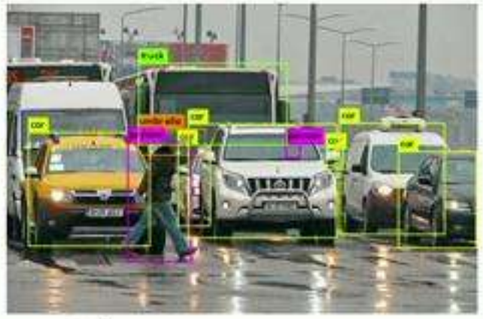
Semantic Segmentation

Pose Estimation

[Deng et.al 2009, Devlin 2018]

Pretrain-Transfer

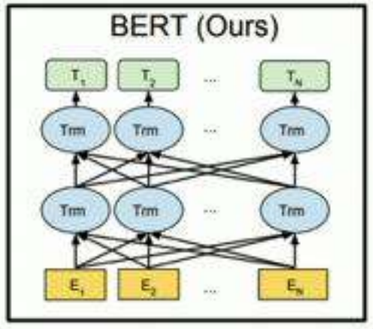
IMAGENET

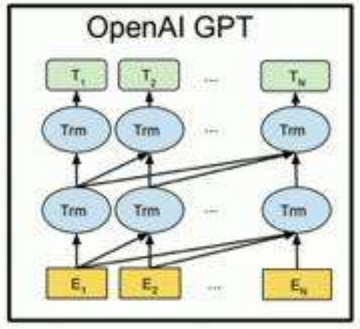


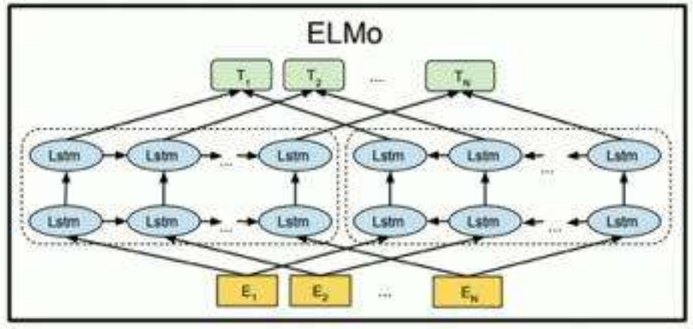
Object Detection

Semantic Segmentation

Pose Estimation



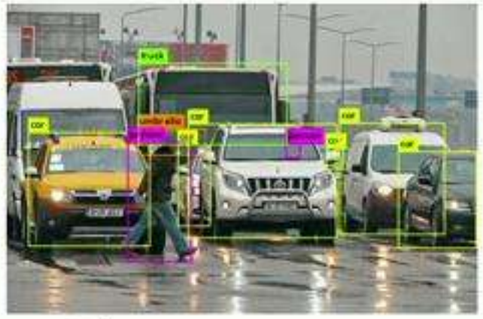




[Deng et.al 2009, Devlin 2018]

Pretrain-Transfer

IMAGENET



Object Detection

Semantic Segmentation

Pose Estimation

Passage Segment

...The European Parliament and the Council of the European Union have powers of amendment and veto during the legislative process...

Question

Which governing bodies have veto power?

Question Answering

[Deng et.al 2009, Devlin 2018]

Pretrain-Transfer

IMAGENET

Object Detection

Semantic Segmentation

Pose Estimation

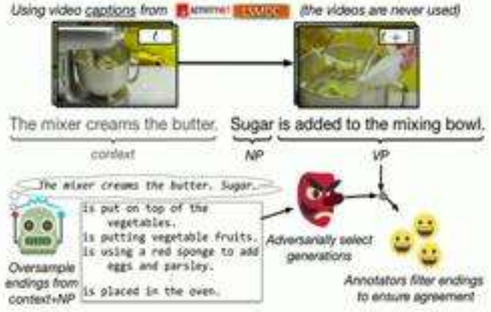
Passage Segment

...The European Parliament and the Council of the European Union have powers of amendment and veto during the legislative process...

Question

Which governing bodies have veto power?

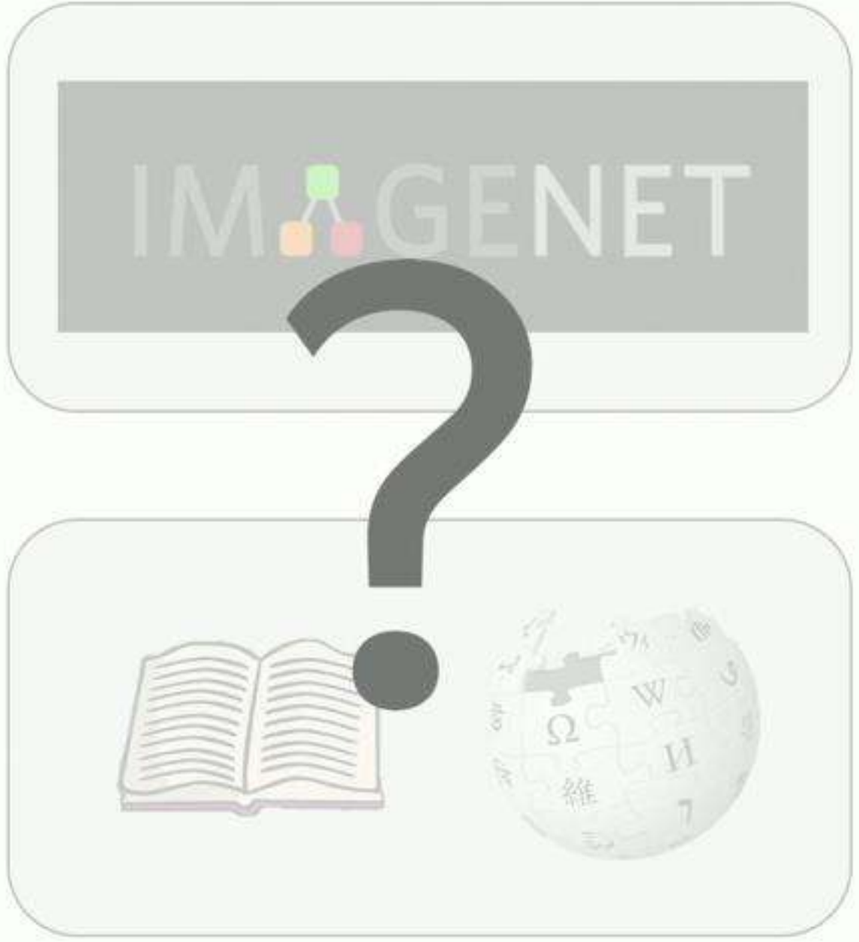
Question Answering

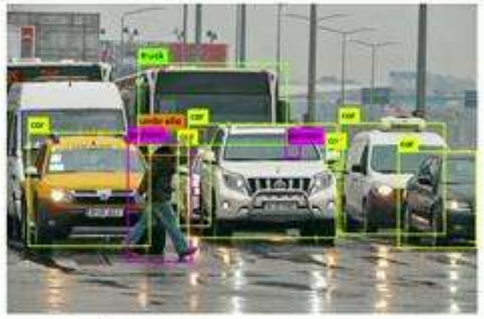


Commonsense Inference

[Deng et.al 2009, Devlin 2018]

Pretrain-Transfer





Object Detection

Semantic Segmentation

Pose Estimation

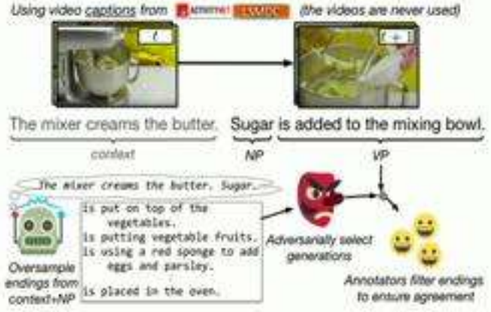
Passage Segment

...The European Parliament and the Council of the European Union have powers of amendment and veto during the legislative process...

Question

Which governing bodies have veto power?

Question Answering



Commonsense Inference

Sentiment	Tweets
Negative	@united is the worst. Nonrefundable First class tickets? Oh because when you select Global/FC their system auto selects economy w/upgrade.
	@united I will not be flying you again
Neutral	@VirginAmerica my drivers license is expired by a little over a month. Can I fly Friday morning using my expired license?
	@VirginAmerica any plans to start flying direct from DAL to LAS?
Positive	@VirginAmerica done! Thank you for the quick response, apparently faster than sitting on hold ;)
	@united I appreciate your efforts getting me home!

Sentiment Analysis

[Deng et.al 2009, Devlin 2018]

Pretrain-Transfer

Alt-text: Musician Justin Timberlake performs at the 2017 Pilgrimage Music & Cultural Festival on September 23, 2017 in Franklin, Tennessee.

Conceptual Captions: pop artist performs at the festival in a city.

Conceptual Caption Dataset

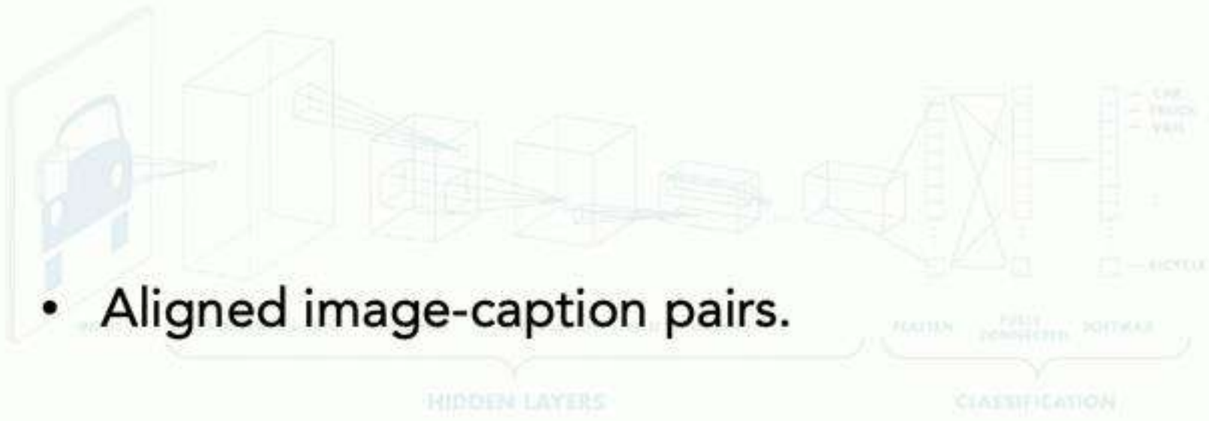
[Sharma et.al 2018]

Pretrain-Transfer

Alt-text: Musician Justin Timberlake performs at the 2017 Pilgrimage Music & Cultural Festival on September 23, 2017 in Franklin, Tennessee.

Conceptual Captions: pop artist performs at the festival in a city.

Conceptual Caption Dataset



- Aligned image-caption pairs.

[Sharma et.al 2018]

Pretrain-Transfer

Alt-text: Musician Justin Timberlake performs at the 2017 Pilgrimage Music & Cultural Festival on September 23, 2017 in Franklin, Tennessee.

Conceptual Captions: pop artist performs at the festival in a city.

Conceptual Caption Dataset

- Aligned image-caption pairs.
- Large Scale. (3.3 million)

[Sharma et.al 2018]

Pretrain-Transfer

Alt-text: Musician Justin Timberlake performs at the 2017 Pilgrimage Music & Cultural Festival on September 23, 2017 in Franklin, Tennessee.

Conceptual Captions: pop artist performs at the festival in a city.

Conceptual Caption Dataset

- Aligned image-caption pairs.
- Large Scale. (3.3 million)
- Automatically collected.

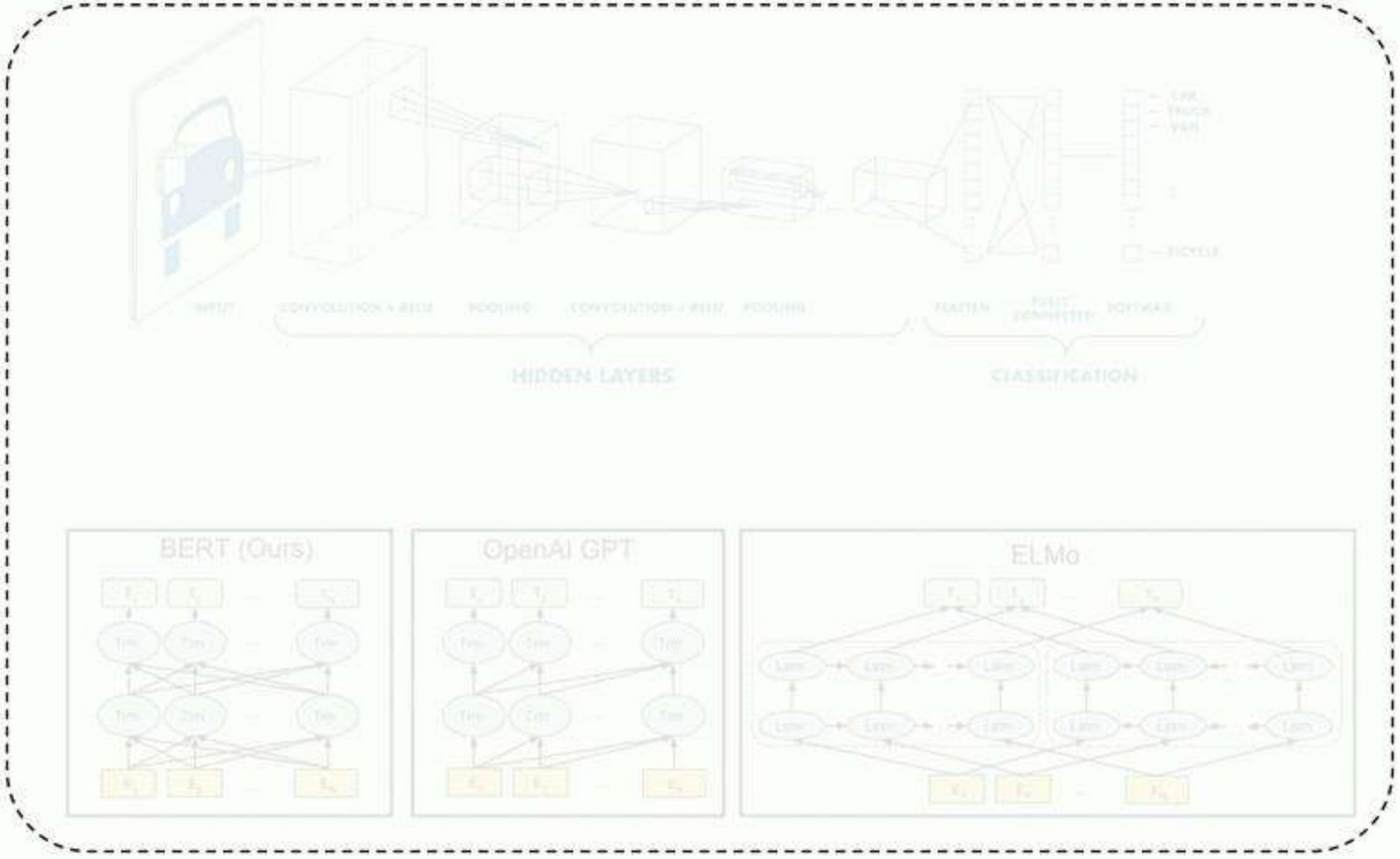
[Sharma et.al 2018]

BERT

Alt-text: Musician Justin Timberlake performs at the 2017 Pilgrimage Music & Cultural Festival on September 23, 2017 in Franklin, Tennessee.

Conceptual Captions: pop artist performs at the festival in a city.

Conceptual Caption Dataset



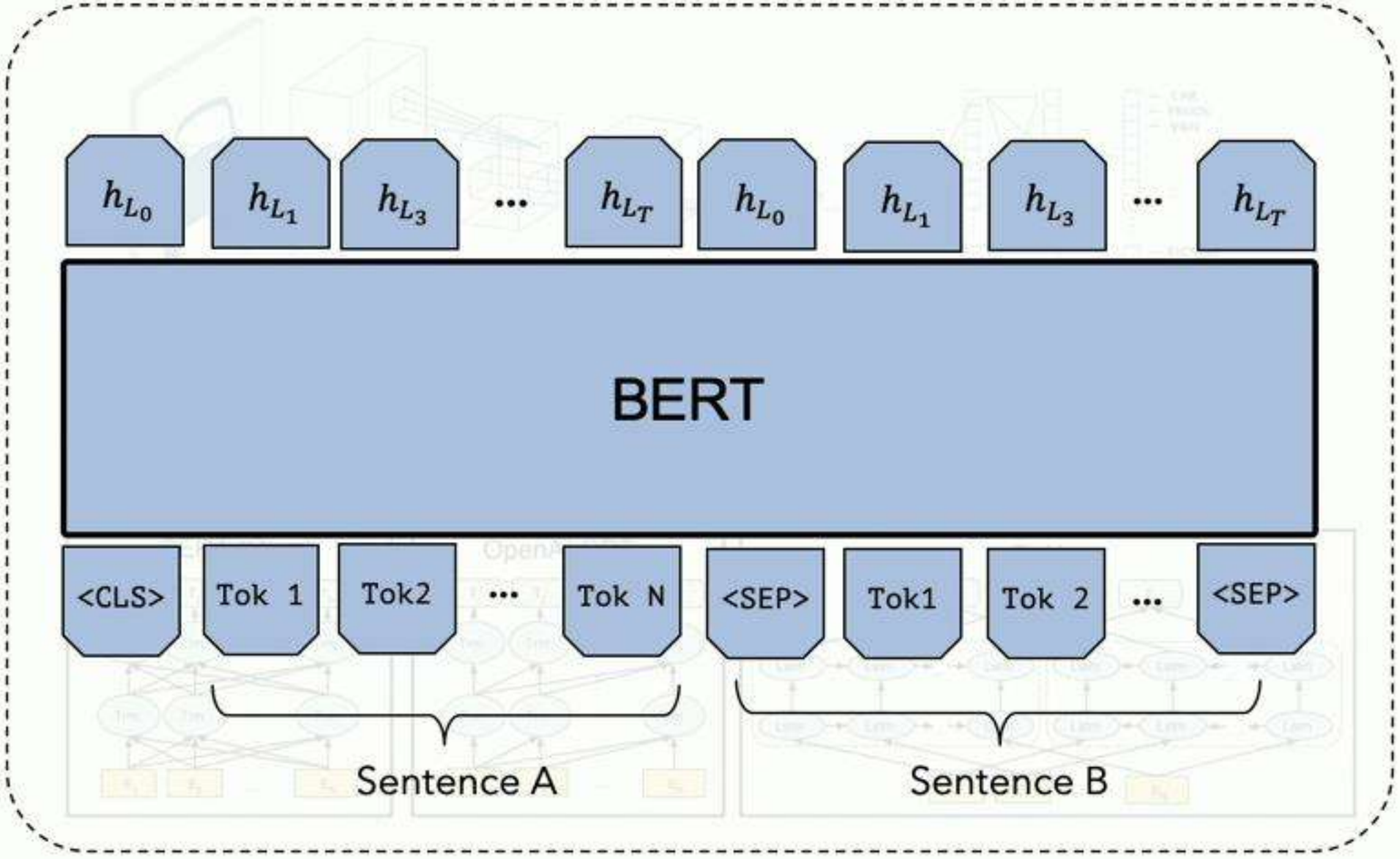
[Sharma et.al 2018, Devlin 2018]]

BERT

Alt-text: Musician Justin Timberlake performs at the 2017 Pilgrimage Music & Cultural Festival on September 23, 2017 in Franklin, Tennessee.

Conceptual Captions: pop artist performs at the festival in a city.

Conceptual Caption Dataset



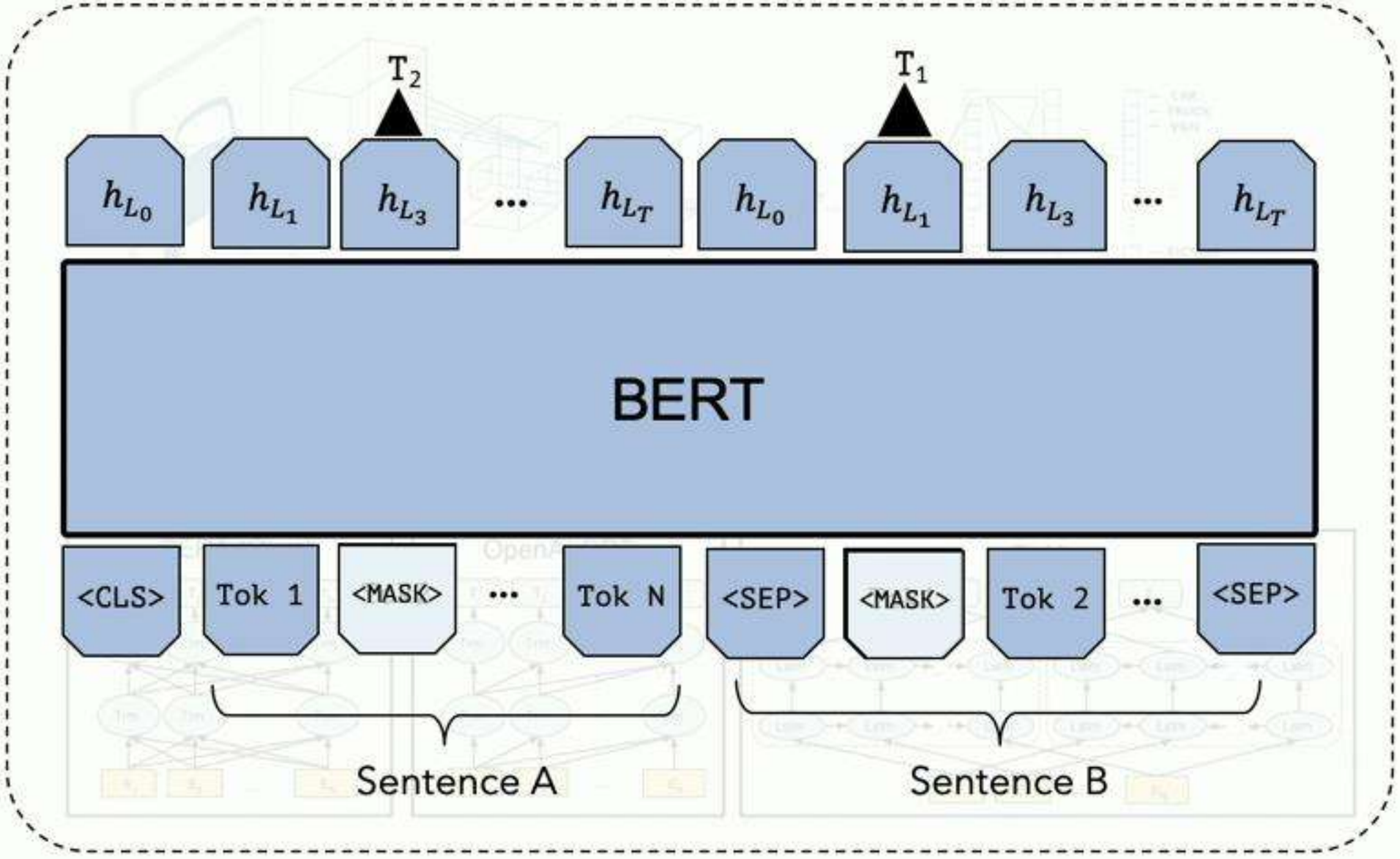
[Sharma et.al 2018, Devlin 2018]]

ViLBERT

Alt-text: Musician Justin Timberlake performs at the 2017 Pilgrimage Music & Cultural Festival on September 23, 2017 in Franklin, Tennessee.

Conceptual Captions: pop artist performs at the festival in a city.

Conceptual Caption Dataset



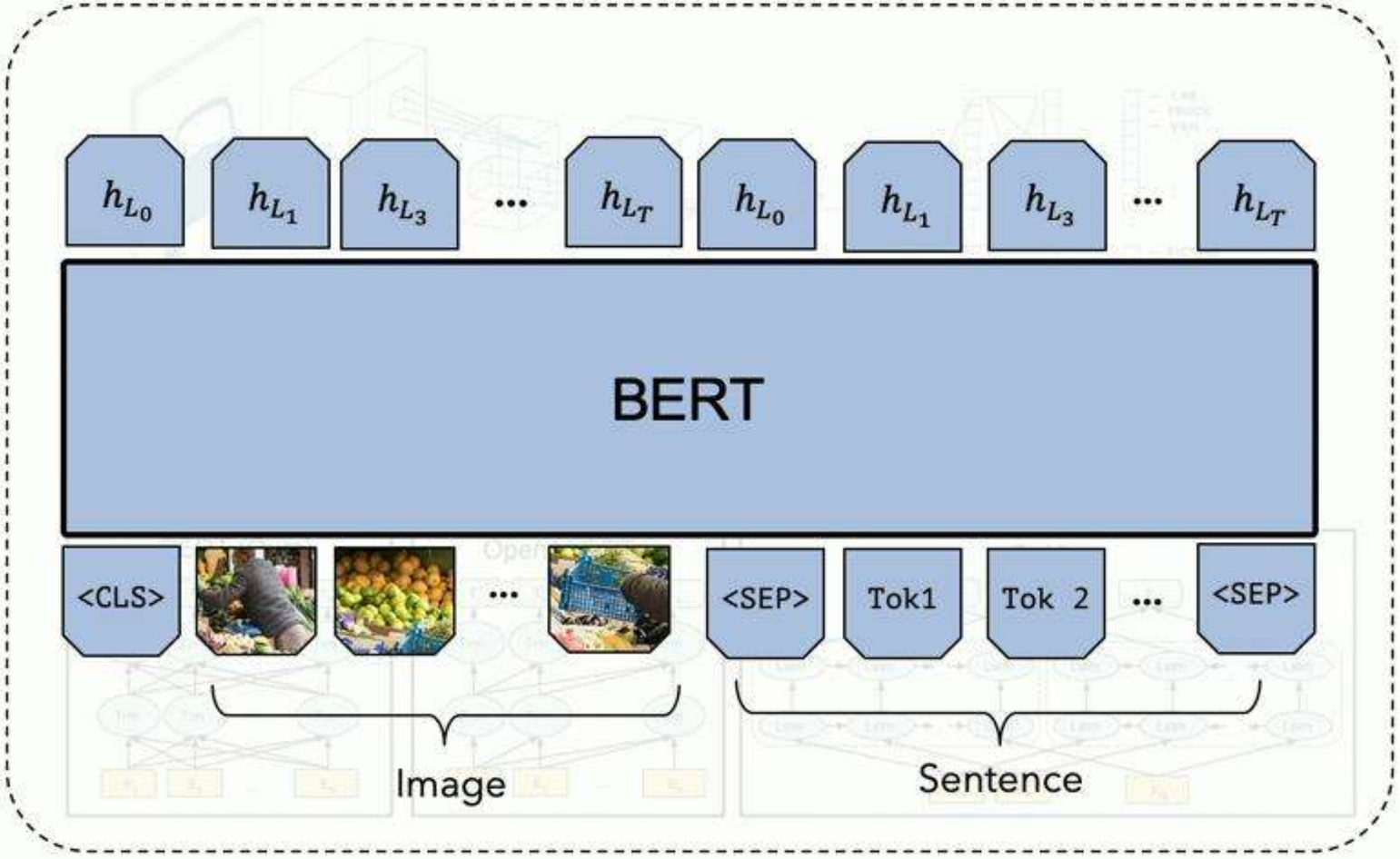
[Sharma et.al 2018, Devlin 2018]]

Single-Stream model

Alt-text: Musician Justin Timberlake performs at the 2017 Pilgrimage Music & Cultural Festival on September 23, 2017 in Franklin, Tennessee.

Conceptual Captions: pop artist performs at the festival in a city.

Conceptual Caption Dataset



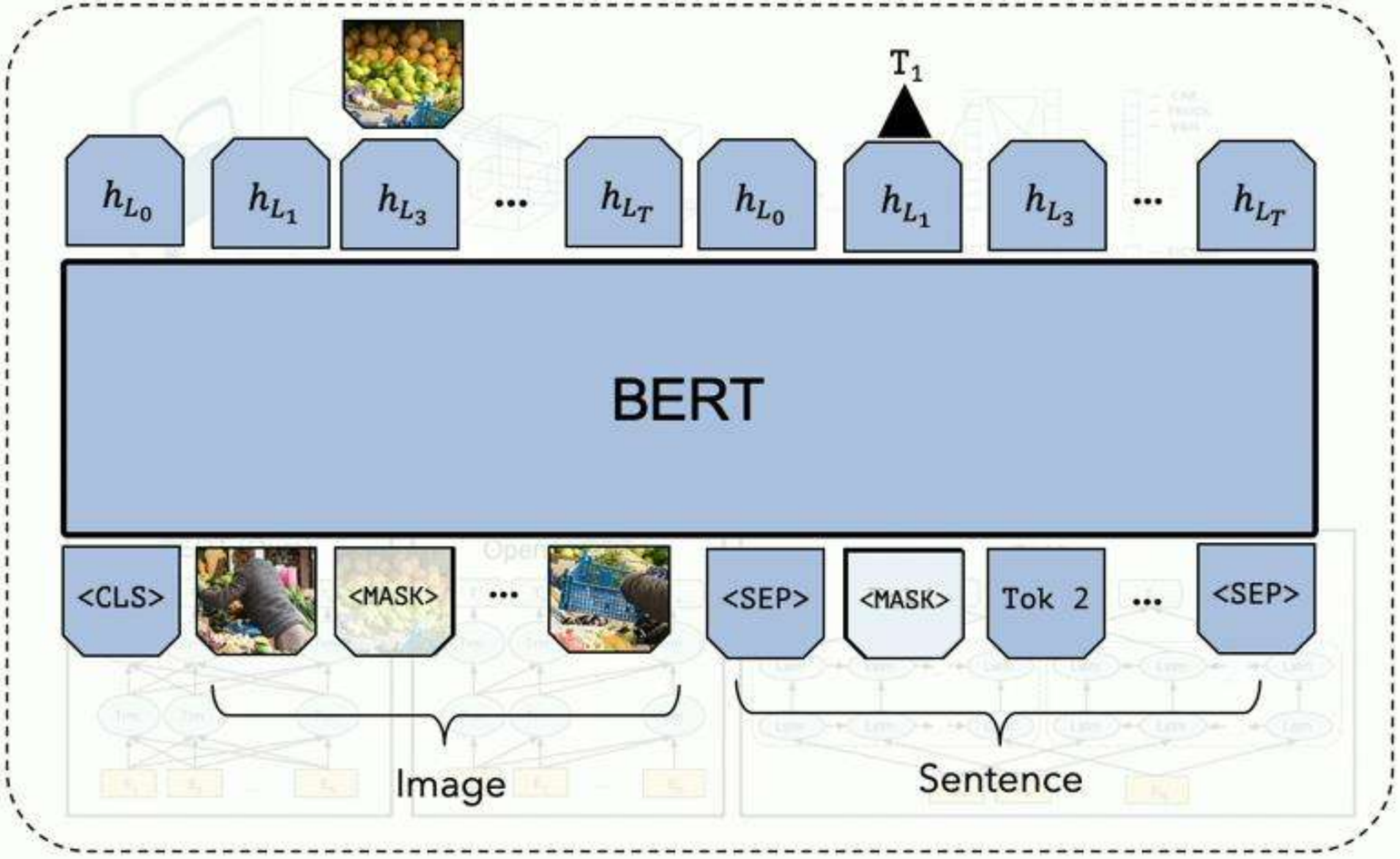
[Sharma et.al 2018, Devlin 2018]]

Single-Stream model

Alt-text: Musician Justin Timberlake performs at the 2017 Pilgrimage Music & Cultural Festival on September 23, 2017 in Franklin, Tennessee.

Conceptual Captions: pop artist performs at the festival in a city.

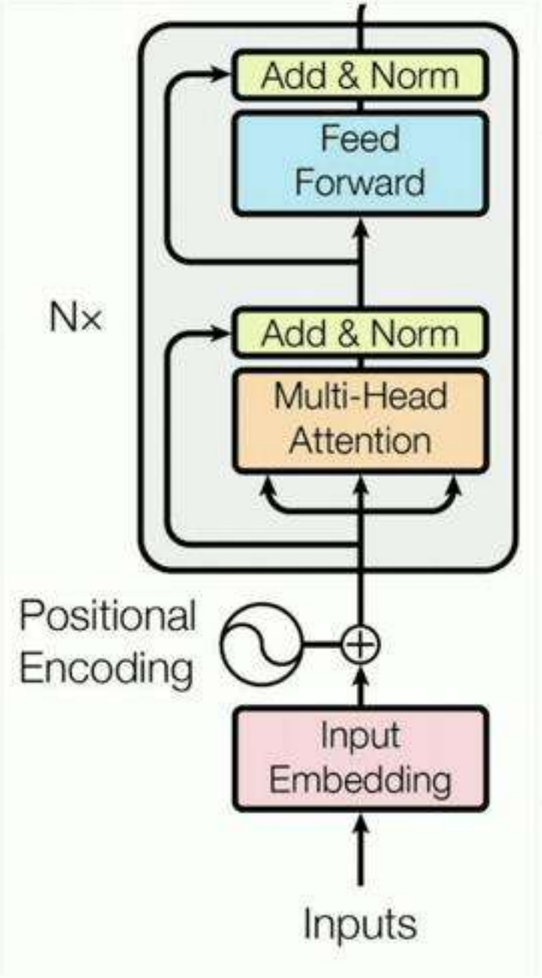
Conceptual Caption Dataset

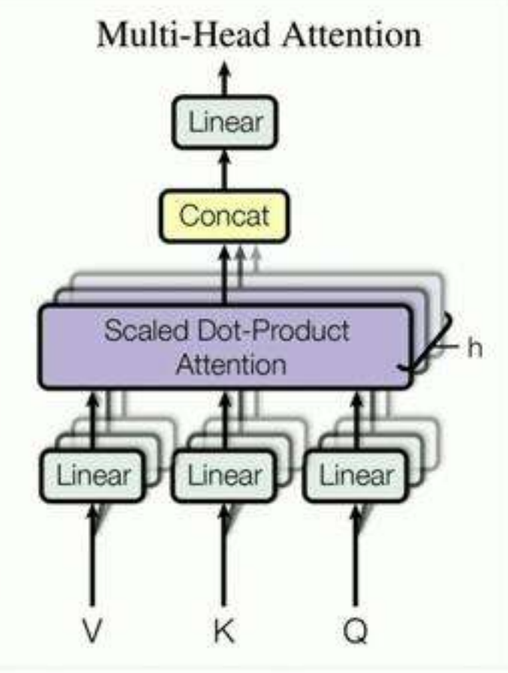


[Sharma et.al 2018, Devlin 2018]]

Transformer

Transformer encoder



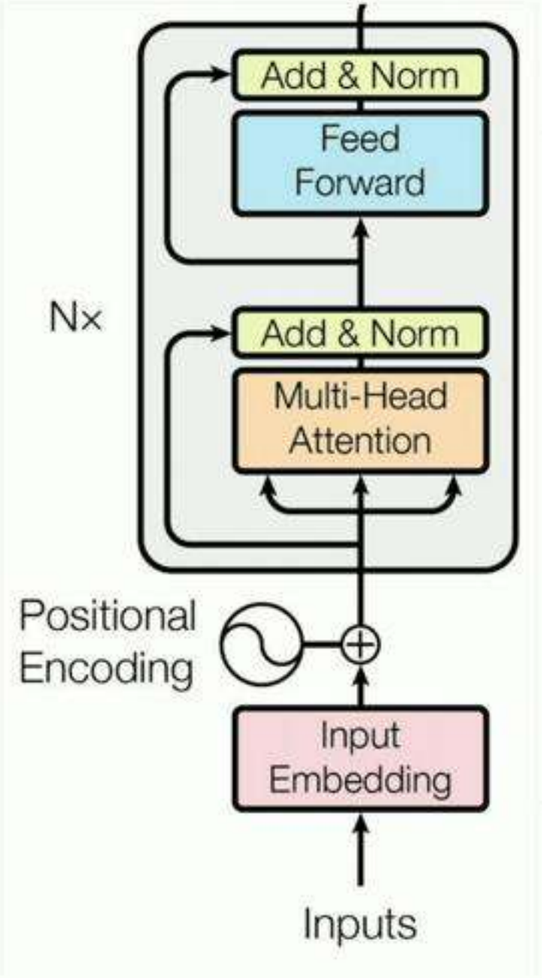


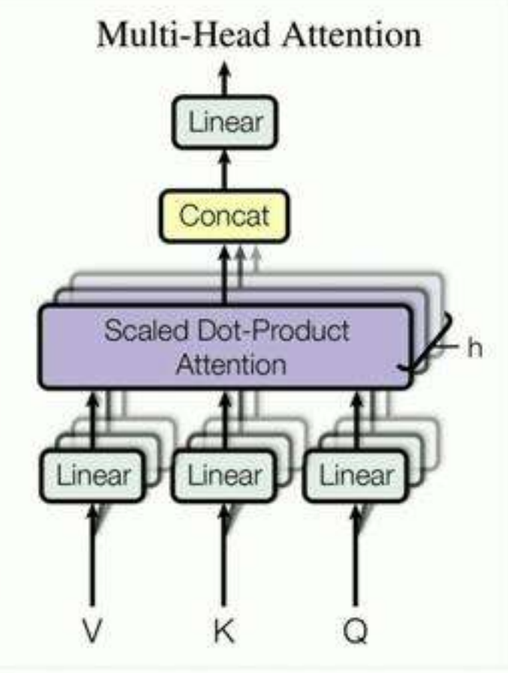
[Vaswani 2017]

Transformer

Transformer encoder

- Multi-headed self attention
  - Model context



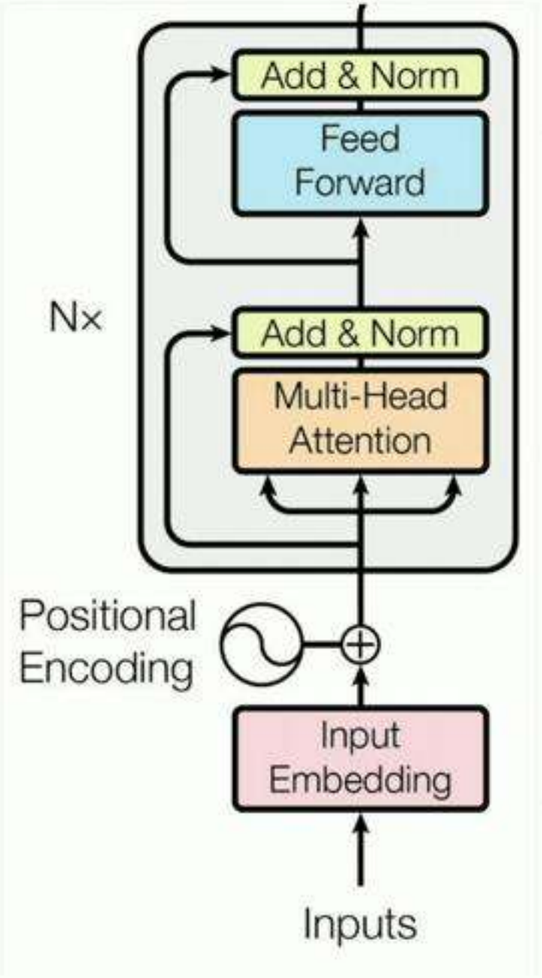


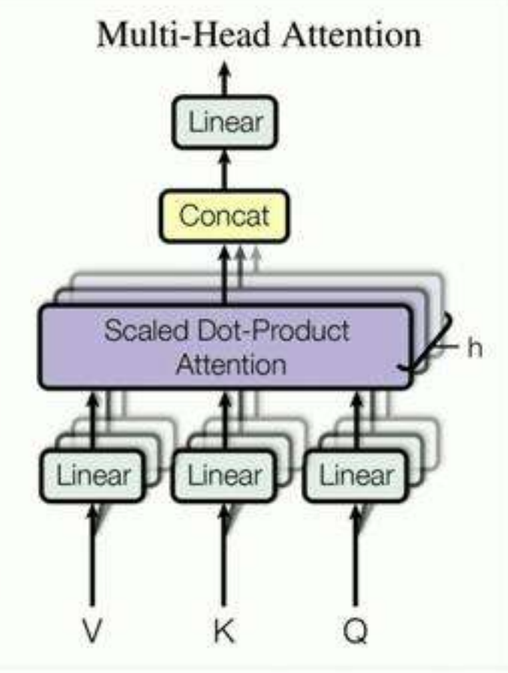
[Vaswani 2017]

Transformer

Transformer encoder

- Multi-headed self attention
  - Model context
- Feed-forward layers
  - Computes non-linear hierarchical features



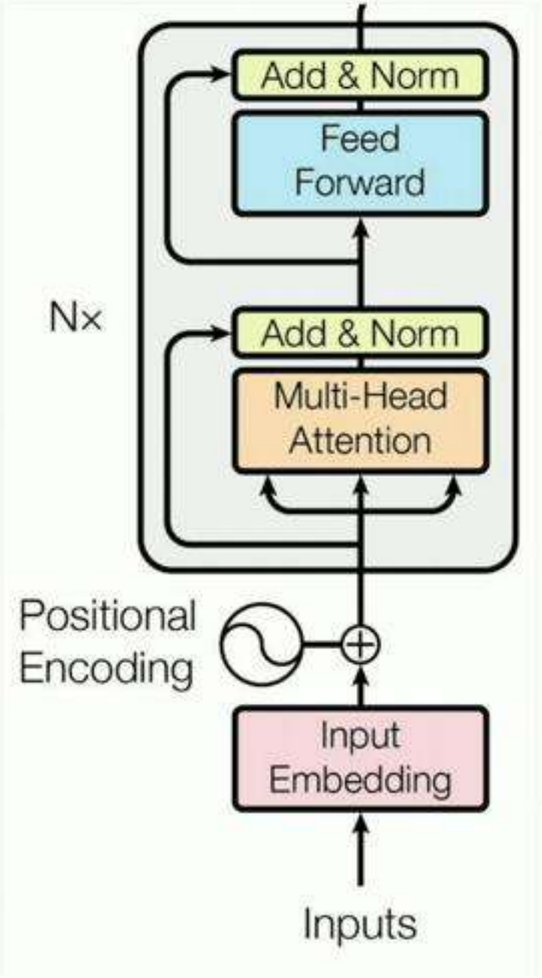


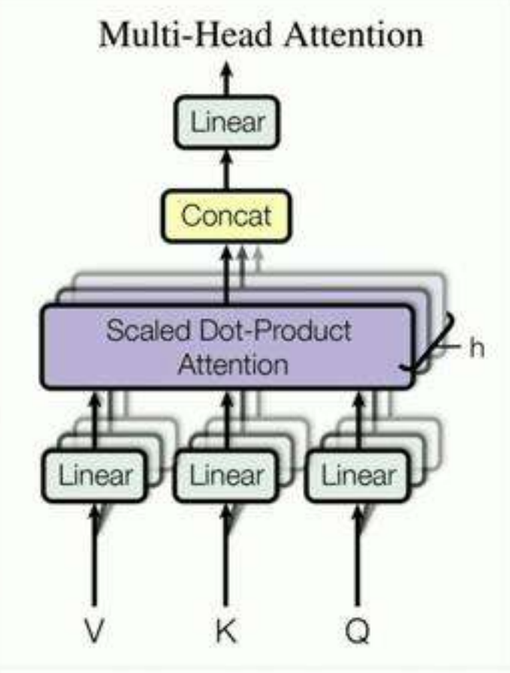
[Vaswani 2017]

Transformer

Transformer encoder

- Multi-headed self attention
  - Model context
- Feed-forward layers
  - Computes non-linear hierarchical features
- Layer norm and residuals
  - Makes training deep networks healthy



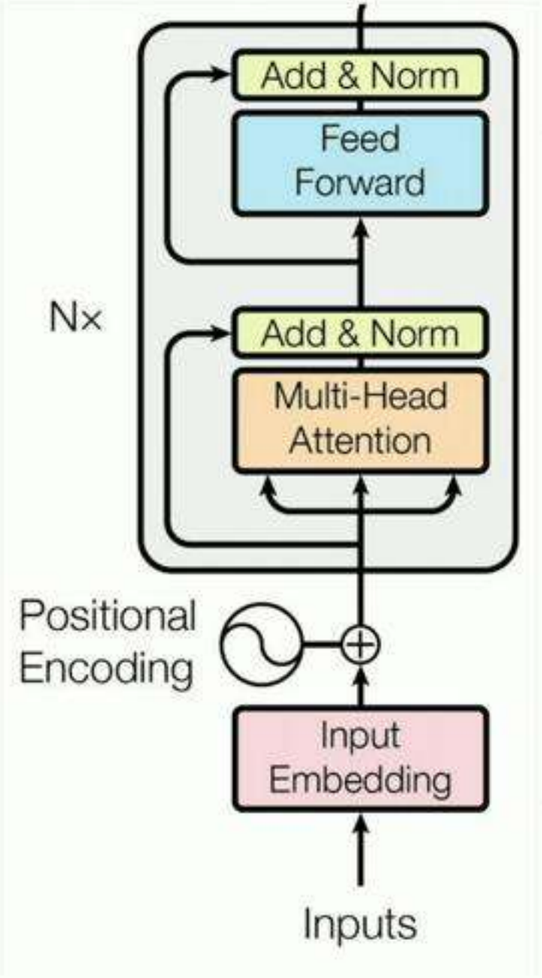


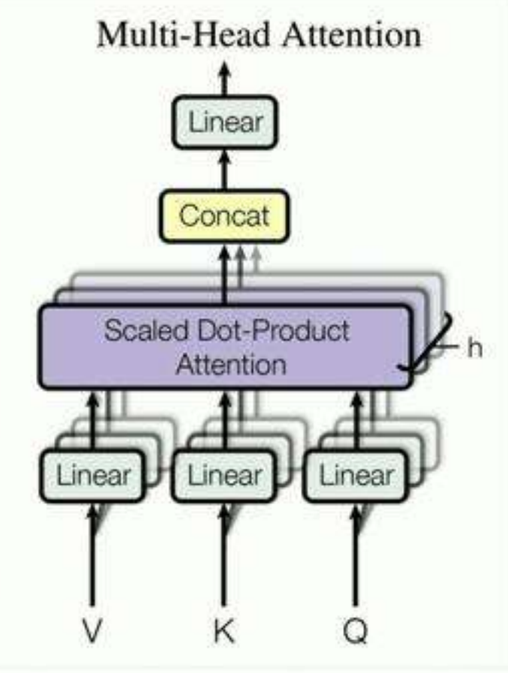
[Vaswani 2017]

Transformer

Transformer encoder

- Multi-headed self attention
  - Model context
- Feed-forward layers
  - Computes non-linear hierarchical features
- Layer norm and residuals
  - Makes training deep networks healthy
- Positional embeddings
  - Allows model to learn relative positioning





[Vaswani 2017]

ViLBERT

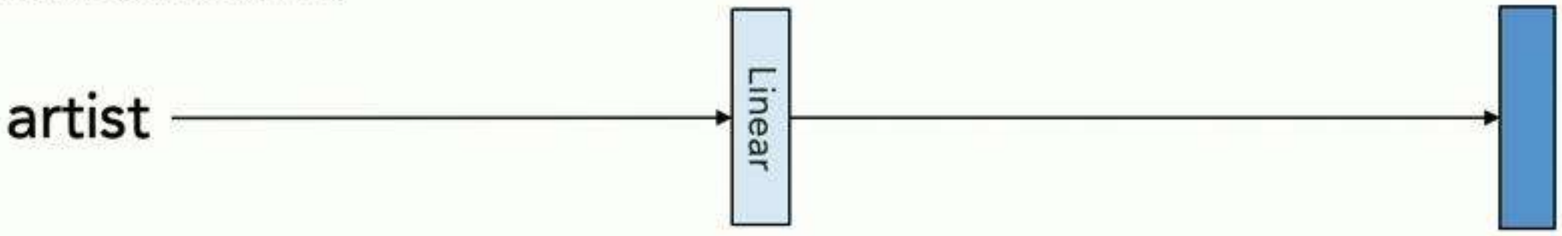
Problem: Different modalities may requires different level of abstractions.

[He 2015]

ViLBERT

Problem: Different modalities may requires different level of abstractions.

- Linguistic stream:

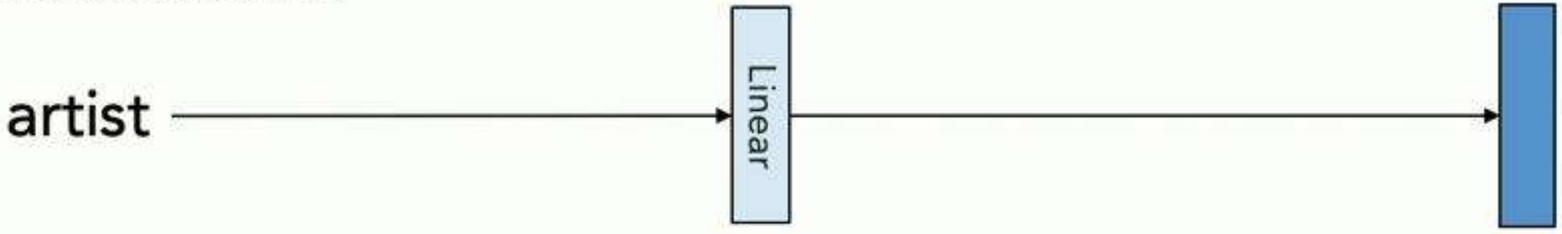


[He 2015]

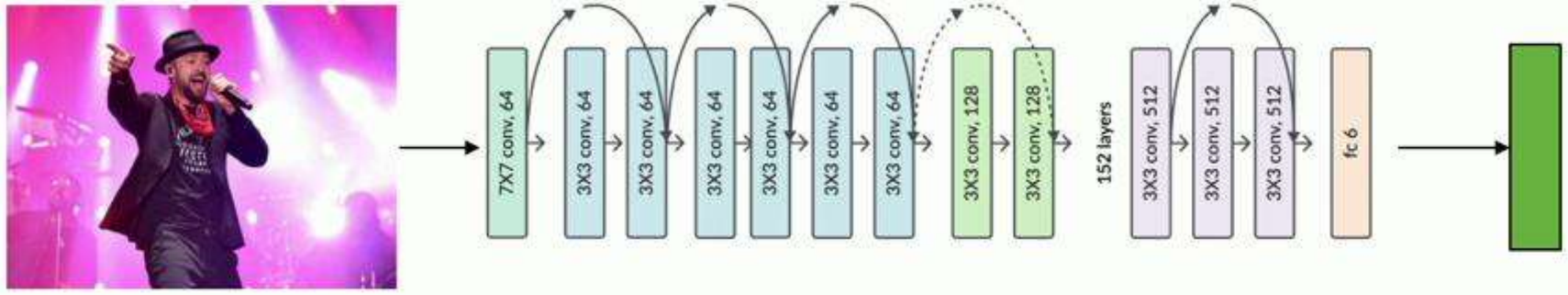
ViLBERT

Problem: Different modalities may requires different level of abstractions.

- Linguistic stream:



- Visual stream:



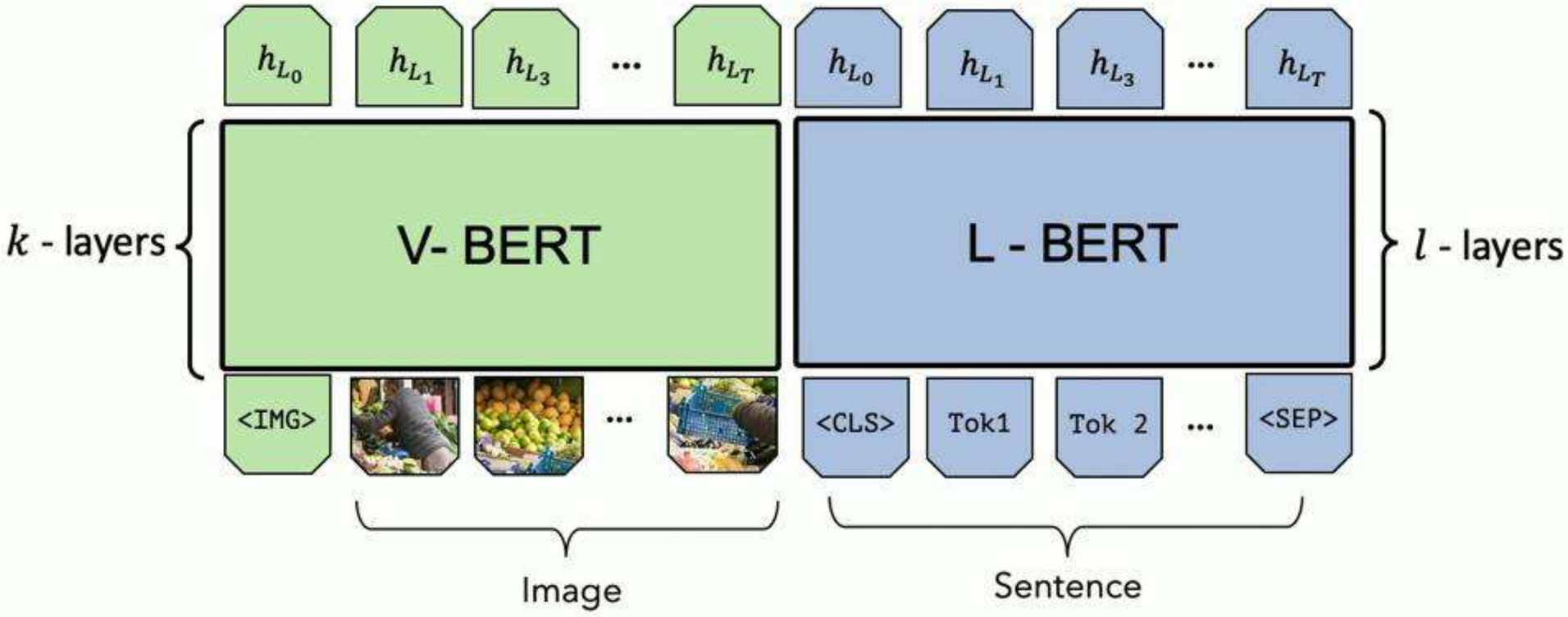
[He 2015]

ViLBERT

Solution: two-stream model which process visual and linguistic separately.

ViLBERT

Solution: two-stream model which process visual and linguistic separately.



ViLBERT

Problem: how to fuse two different modality?

ViLBERT

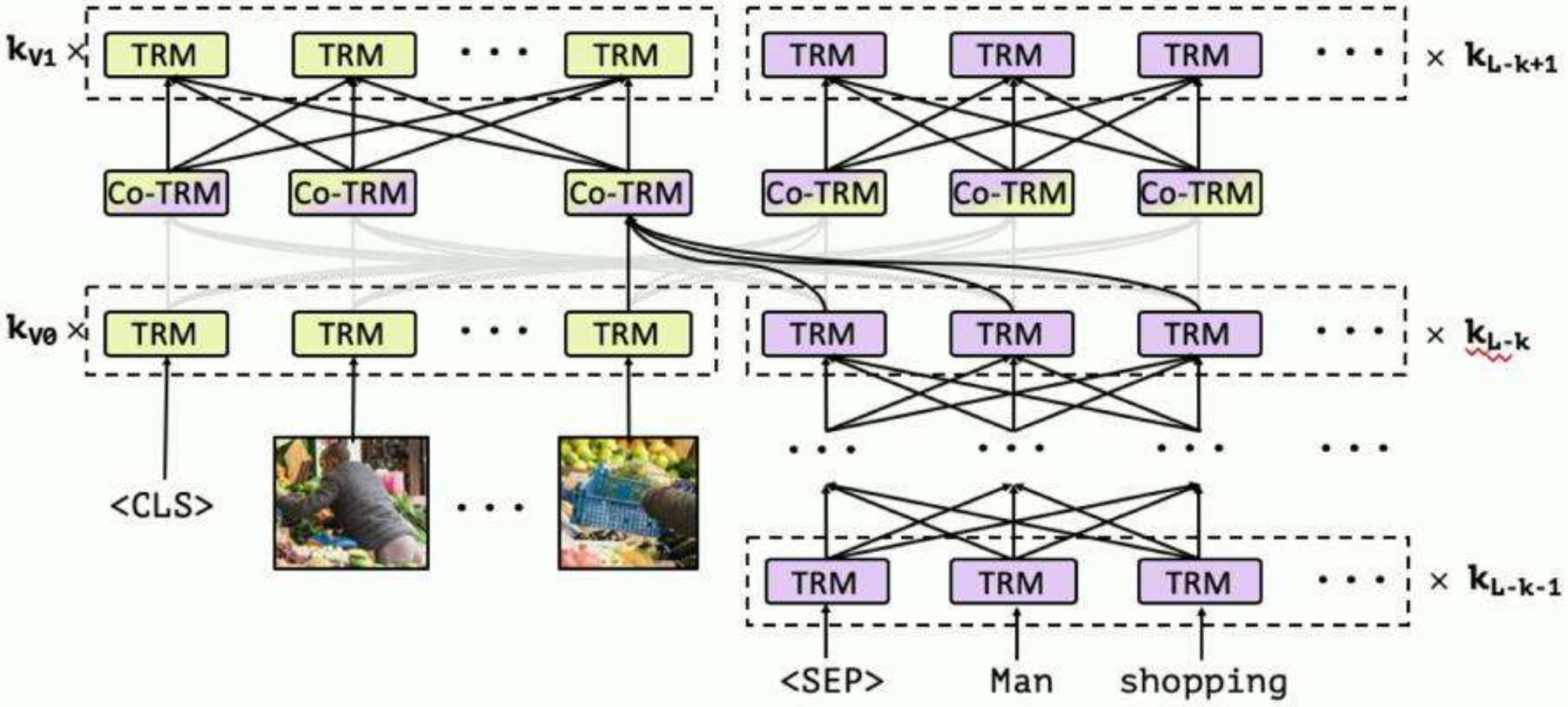
Problem: how to fuse two different modality?

Solution: use **co-attention** [Lu.et.al 2016] to fuse information between different source.

ViLBERT

Problem: how to fuse two different modality?

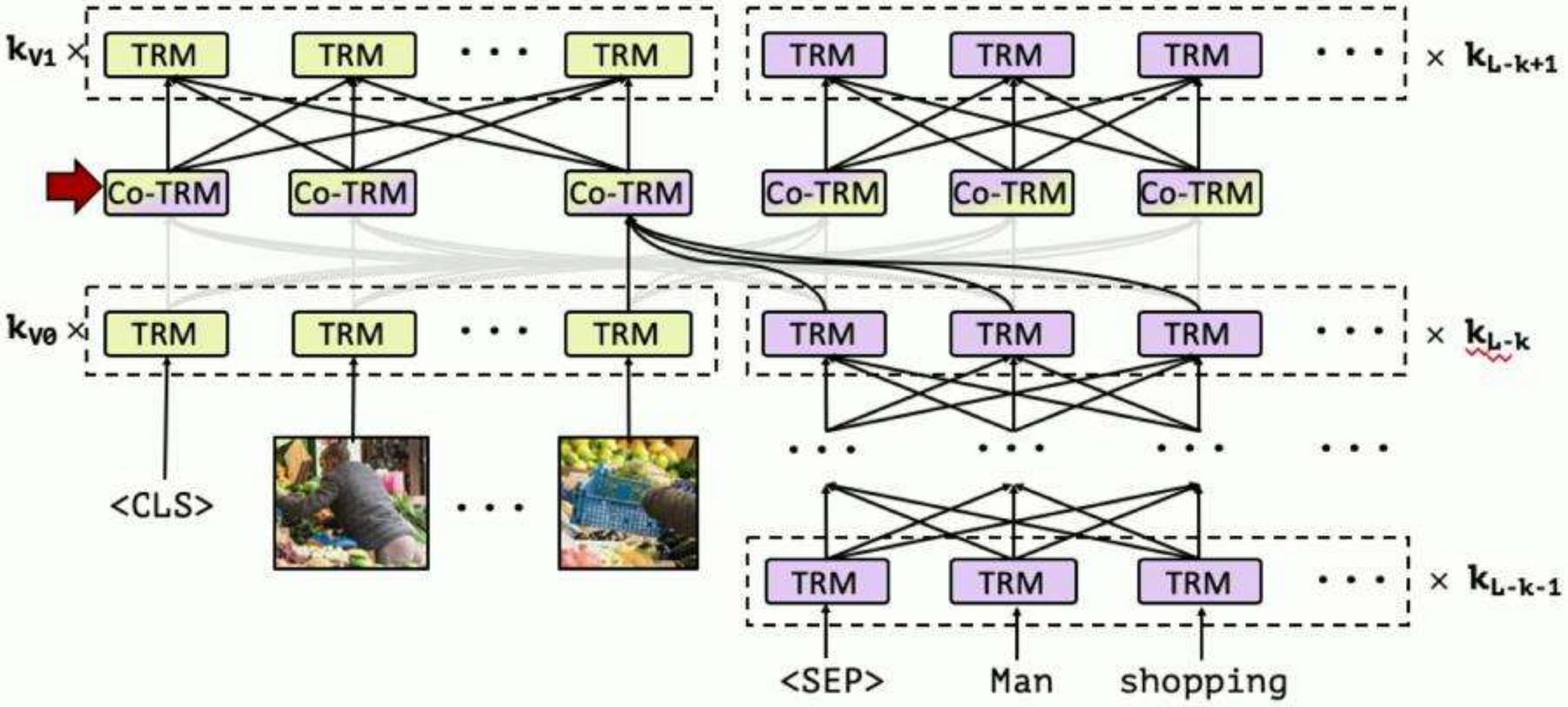
Solution: use **co-attention** [Lu.et.al 2016] to fuse information between different source.



ViLBERT

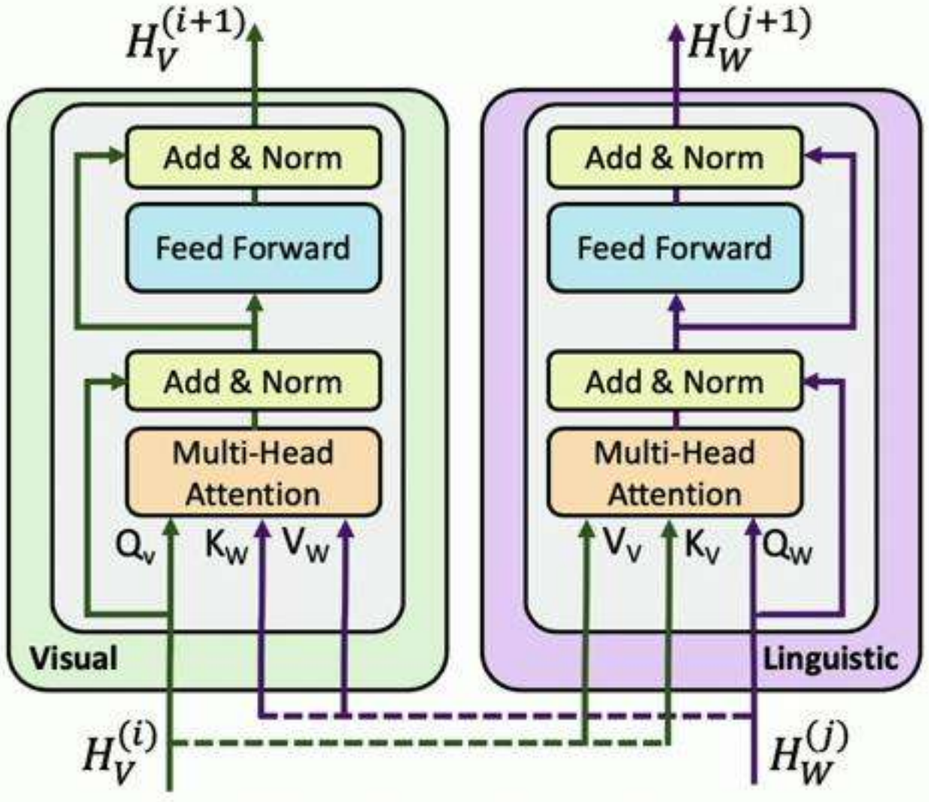
Problem: how to fuse two different modality?

Solution: use **co-attention** [Lu.et.al 2016] to fuse information between different source.



ViLBERT

Co-Attentional Transformer

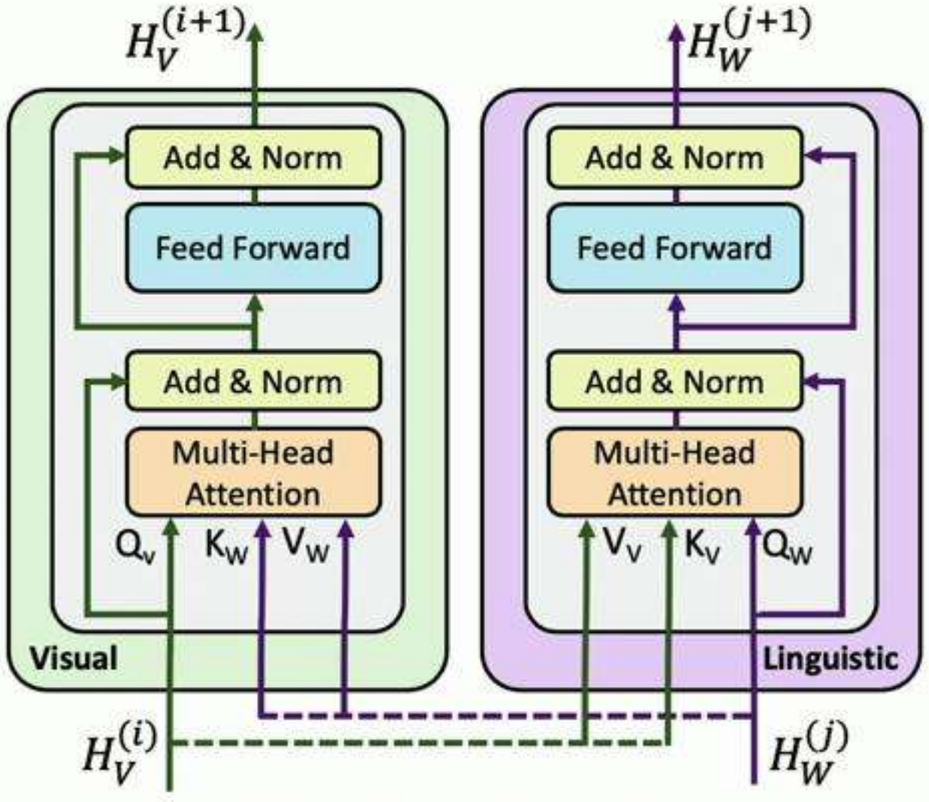


Co-Attentional Transformer

ViLBERT

Co-Attentional Transformer

- Transformer encoder with query from another modality.

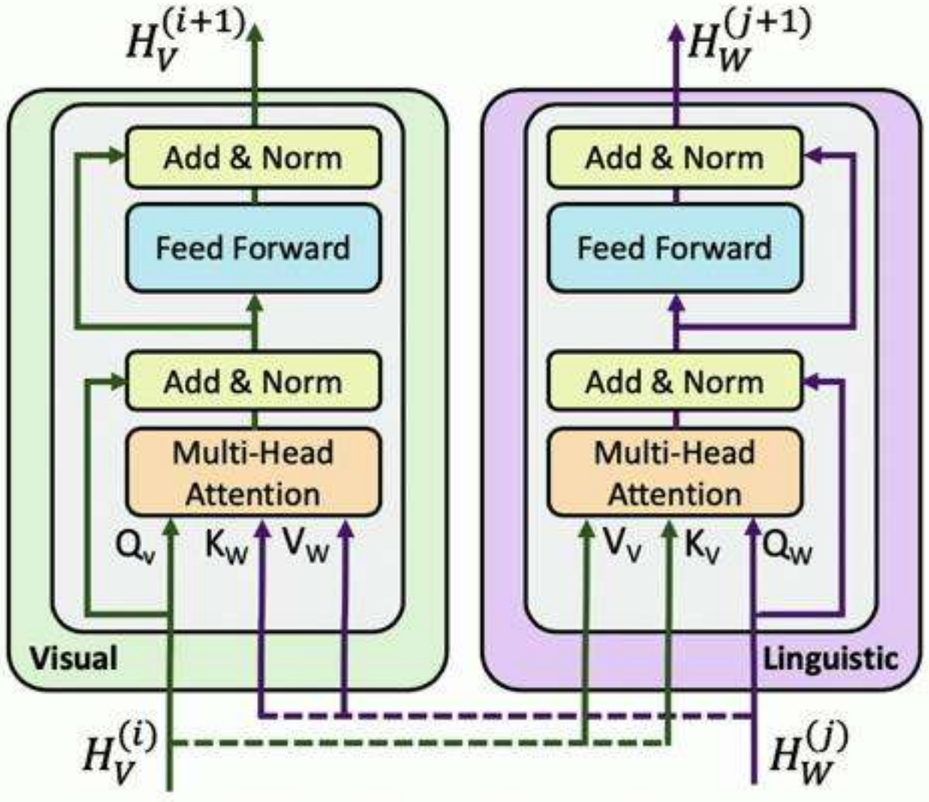


Co-Attentional Transformer

ViLBERT

Co-Attentional Transformer

- Transformer encoder with query from another modality.
- Aggregate information with residual add operation.



Co-Attentional Transformer

Pre-training Objective

Pre-training Objective

Masked multi-modal modelling

Pre-training Objective

Masked multi-modal modelling

- Follows masked LM in BERT.

Pre-training Objective

Masked multi-modal modelling

- Follows masked LM in BERT.
- 15% of the words or image regions to predict.

Pre-training Objective

Masked multi-modal modelling

- Follows masked LM in BERT.
- 15% of the words or image regions to predict.
- Linguistic stream:
  - 80% of the time, replace with *[MASK]*.
  - 10% of the time, replace random word.
  - 10% of the time, keep same.

Pre-training Objective

Masked multi-modal modelling

- Follows masked LM in BERT.
- 15% of the words or image regions to predict.
- Linguistic stream:
  - 80% of the time, replace with *[MASK]*.
  - 10% of the time, replace random word.
  - 10% of the time, keep same.
- Visual stream:
  - 80% of the time, replace with zero vector.

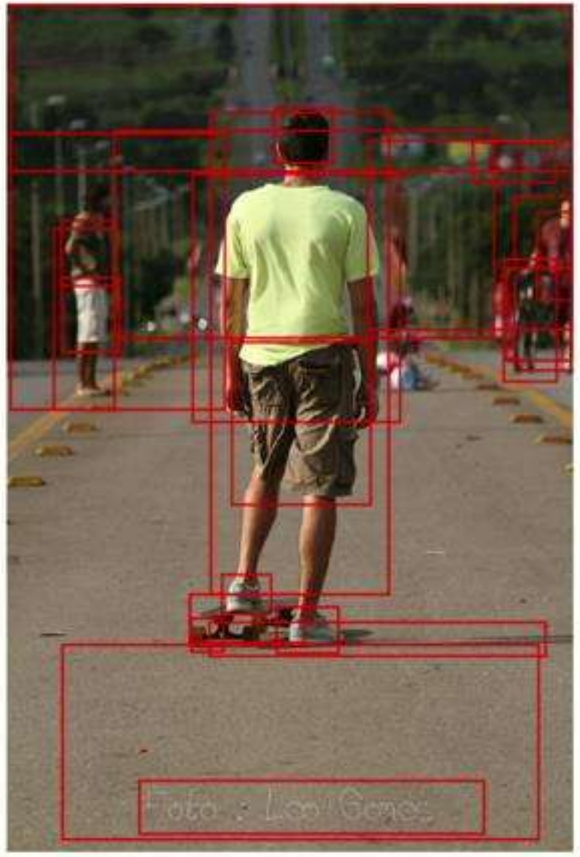
Pre-training Objective

Masked multi-modal modelling

- Follows masked LM in BERT.
- 15% of the words or image regions to predict.
- Linguistic stream:
  - 80% of the time, replace with *[MASK]*.
  - 10% of the time, replace random word.
  - 10% of the time, keep same.
- Visual stream:
  - 80% of the time, replace with zero vector.

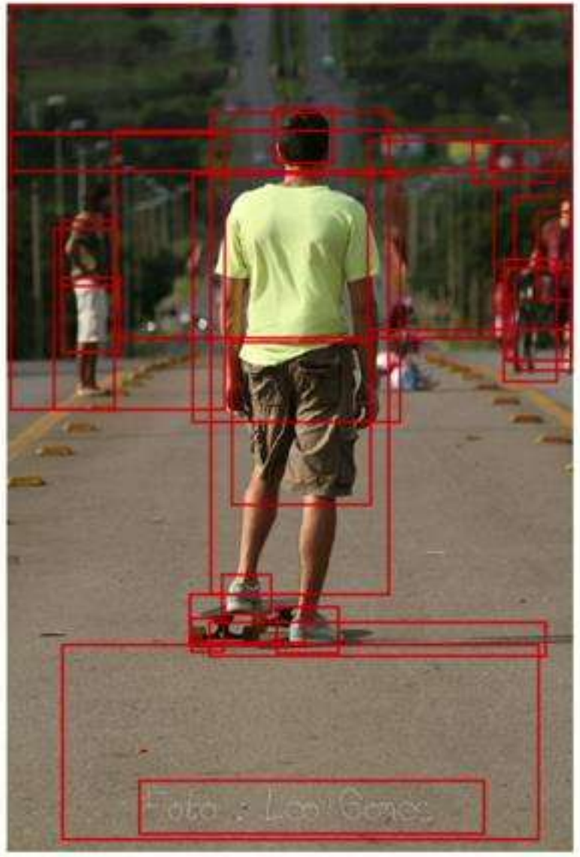
Multi-modal alignment prediction

Image Representation



[Anderson *et.al. 2018*]

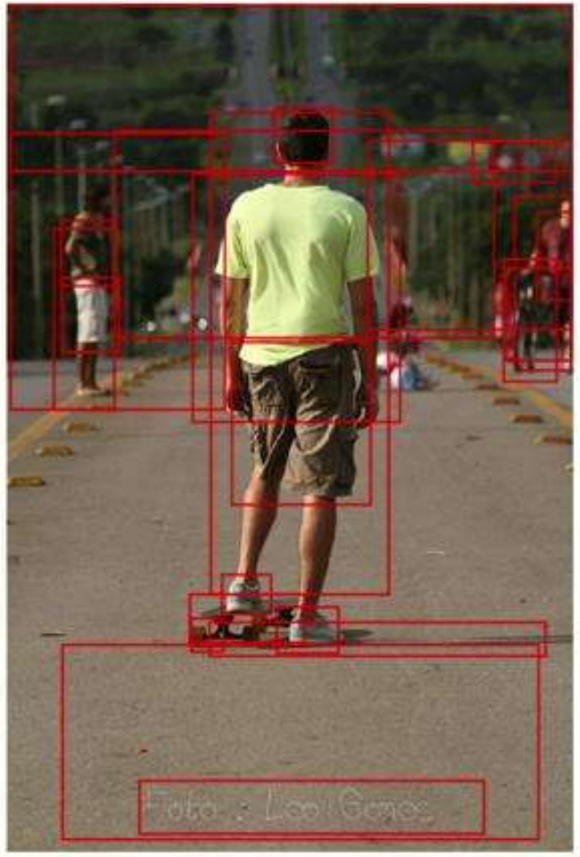
Image Representation



- Faster R-CNN with Res101 backbone.

[Anderson *et.al. 2018*]

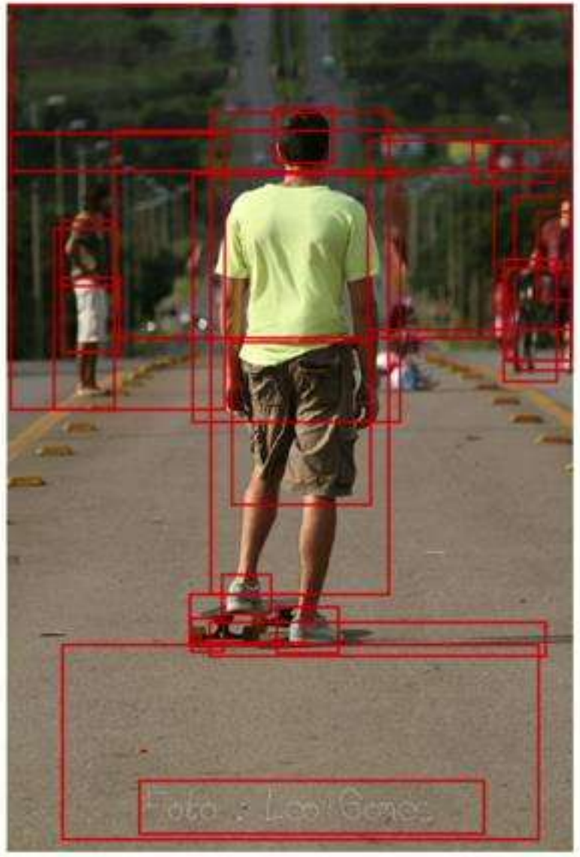
Image Representation



- Faster R-CNN with Res101 backbone.
- Trained on Visual Genome dataset.

[Anderson *et.al. 2018*]

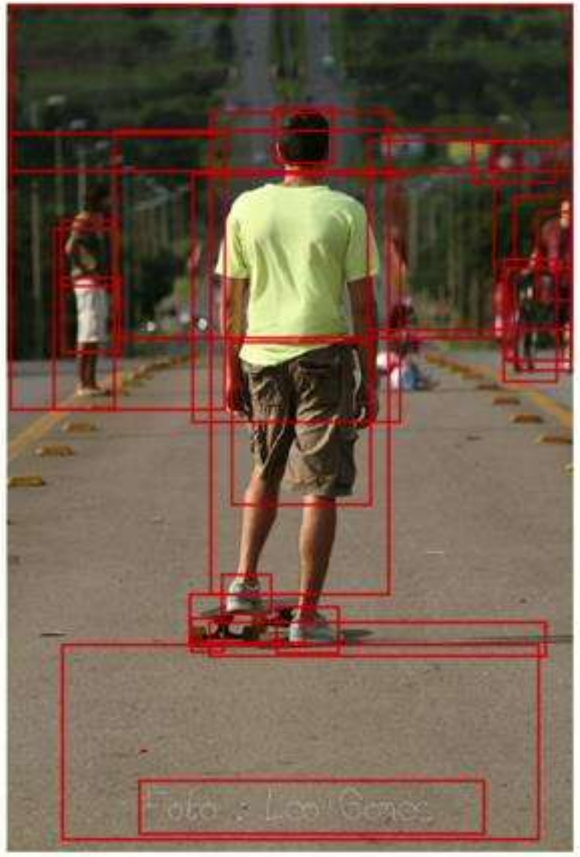
Image Representation



- Faster R-CNN with Res101 backbone.
- Trained on Visual Genome dataset.
- 1600 detection classes.

[Anderson *et.al. 2018*]

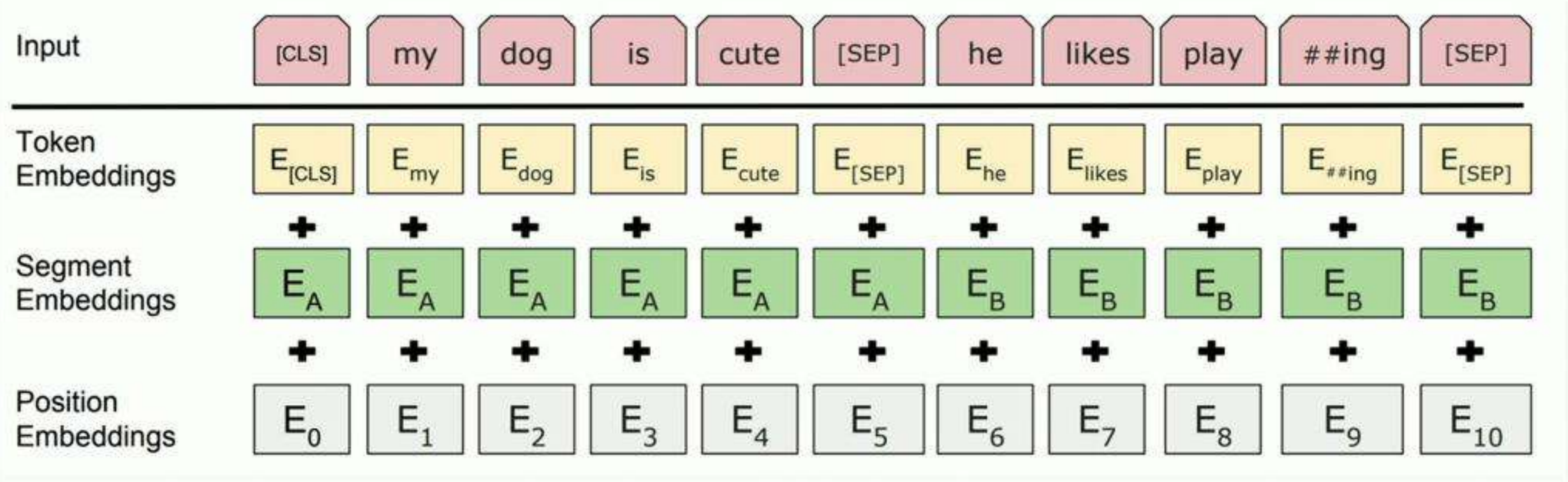
Image Representation



- Faster R-CNN with Res101 backbone.
- Trained on Visual Genome dataset.
- 1600 detection classes.
- Sum of region embeddings + location embeddings

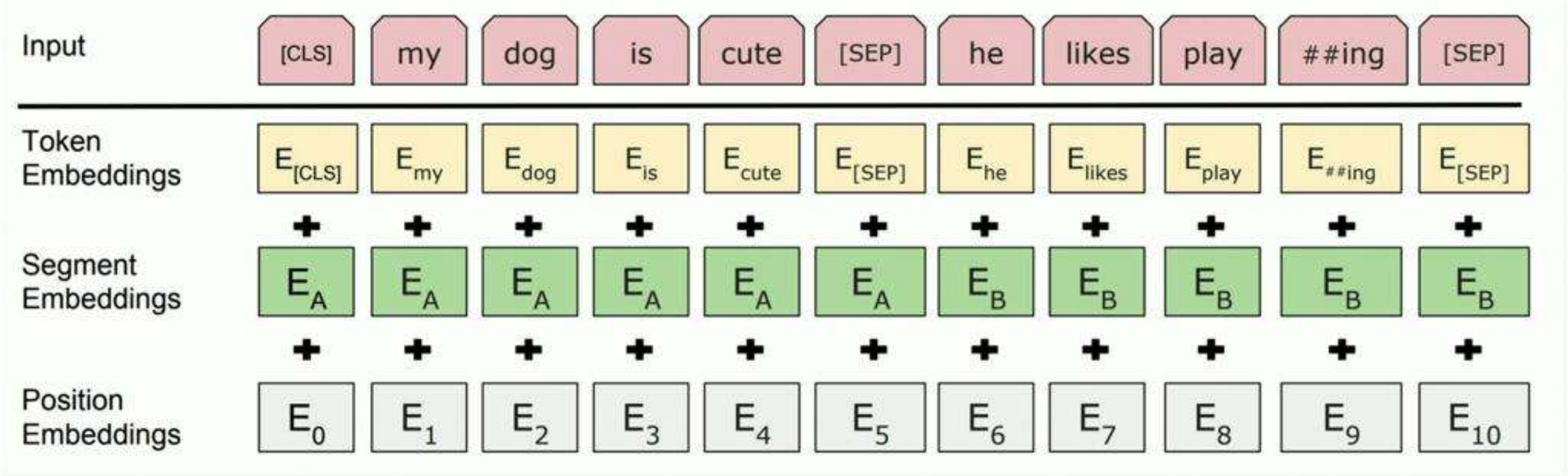
[Anderson *et.al. 2018*]

Text Representation



[Devlin 2018]

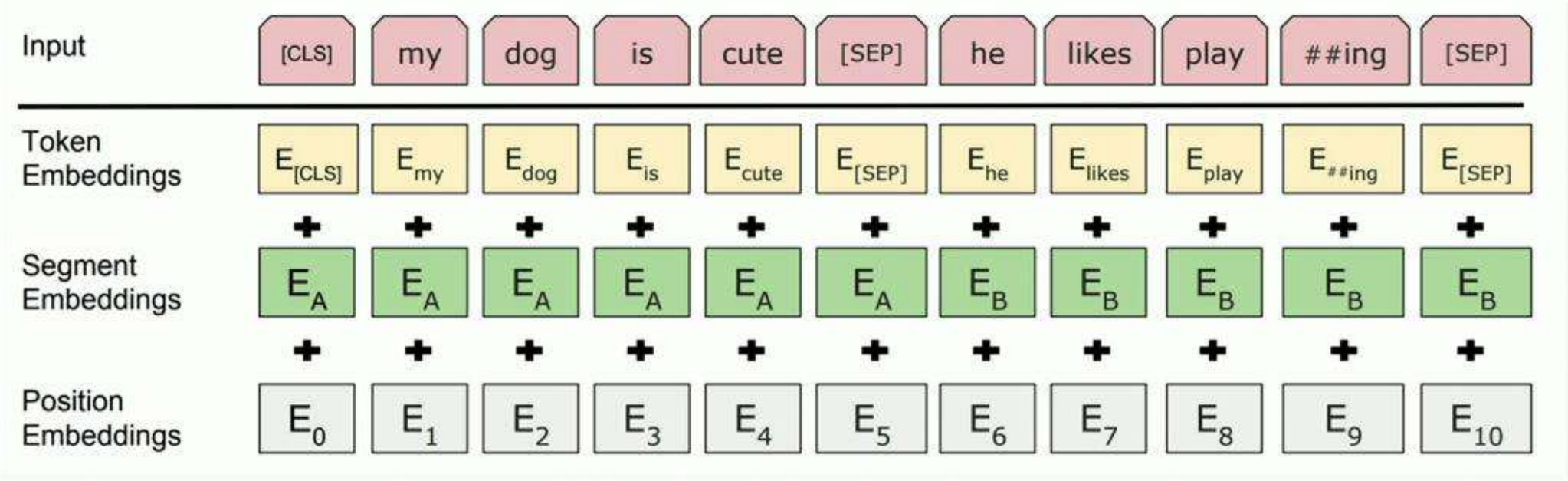
Text Representation



- Use 30,000 WordPiece vocabulary on input.

[Devlin 2018]

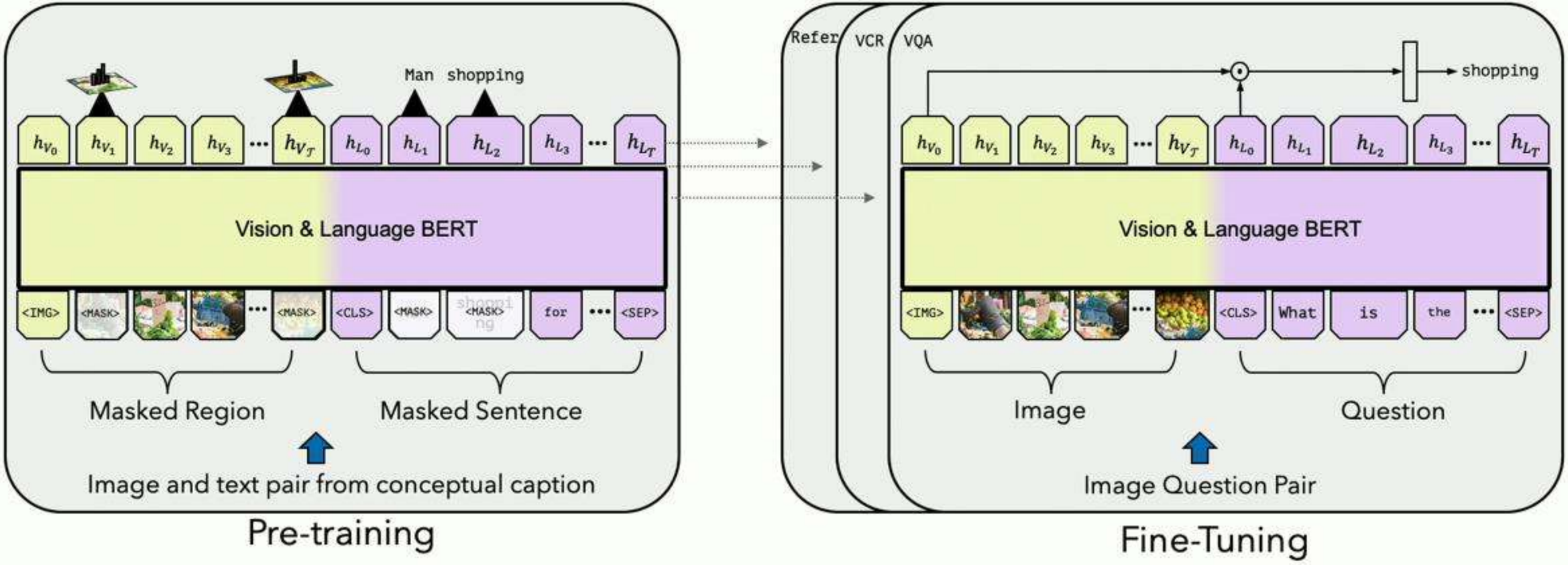
Text Representation



- Use 30,000 WordPiece vocabulary on input.
- Each token is sum of three embeddings

[Devlin 2018]

Fine-tuning Procedure



Tasks

[Antol 2015, zeller 2018, Yu 2016, Plummer 2015]

Tasks

VQA

[Antol 2015, zeller 2018, Yu 2016, Plummer 2015]

Tasks

VQA

VCR Q→A VCR QA→R

[Antol 2015, zeller 2018, Yu 2016, Plummer 2015]

Tasks

VQA

VCR Q→A VCR QA→R

Referring Expressions

[Antol 2015, zeller 2018, Yu 2016, Plummer 2015]

Tasks

VQA

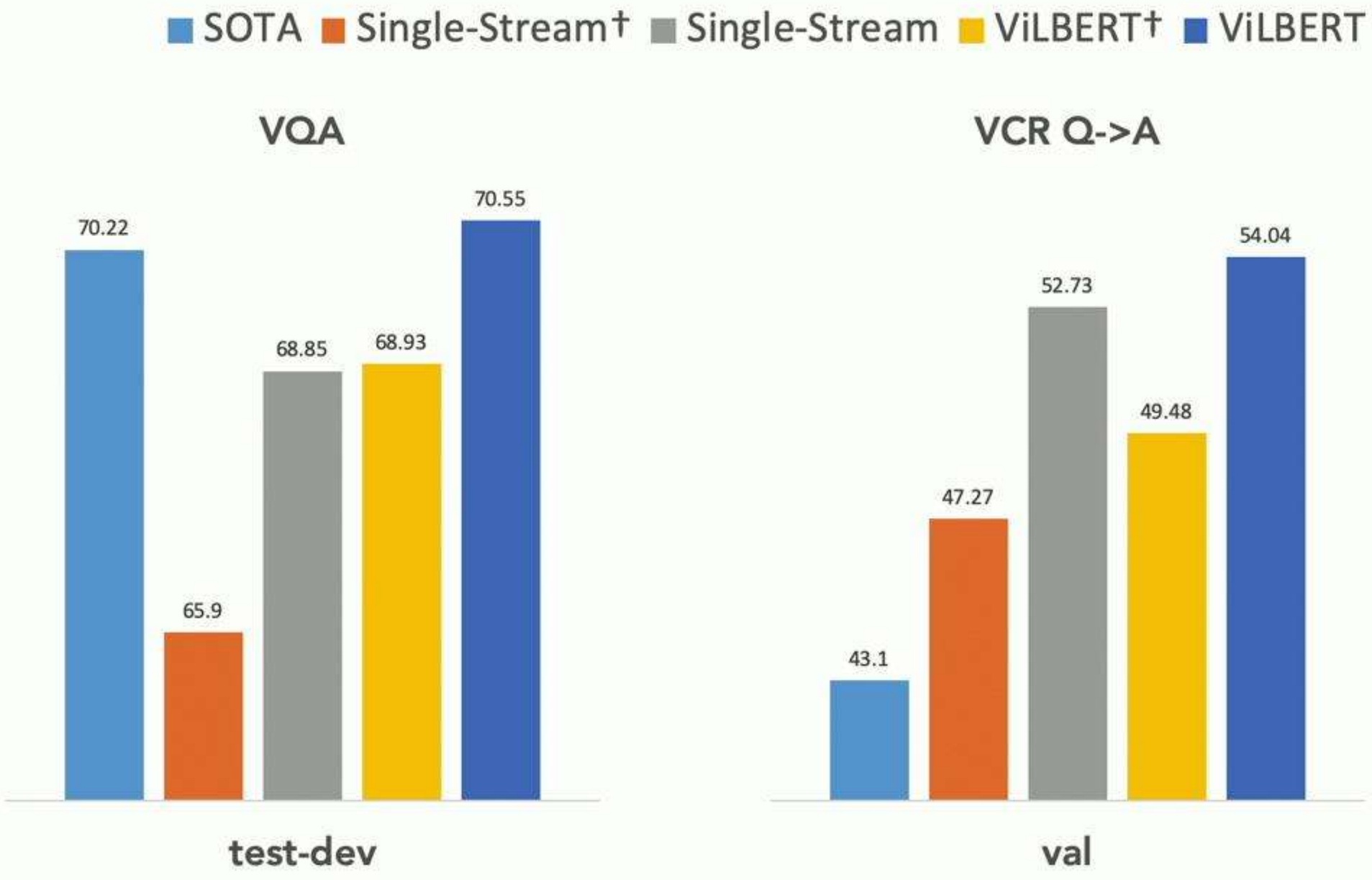
VCR Q→A VCR QA→R

Referring Expressions

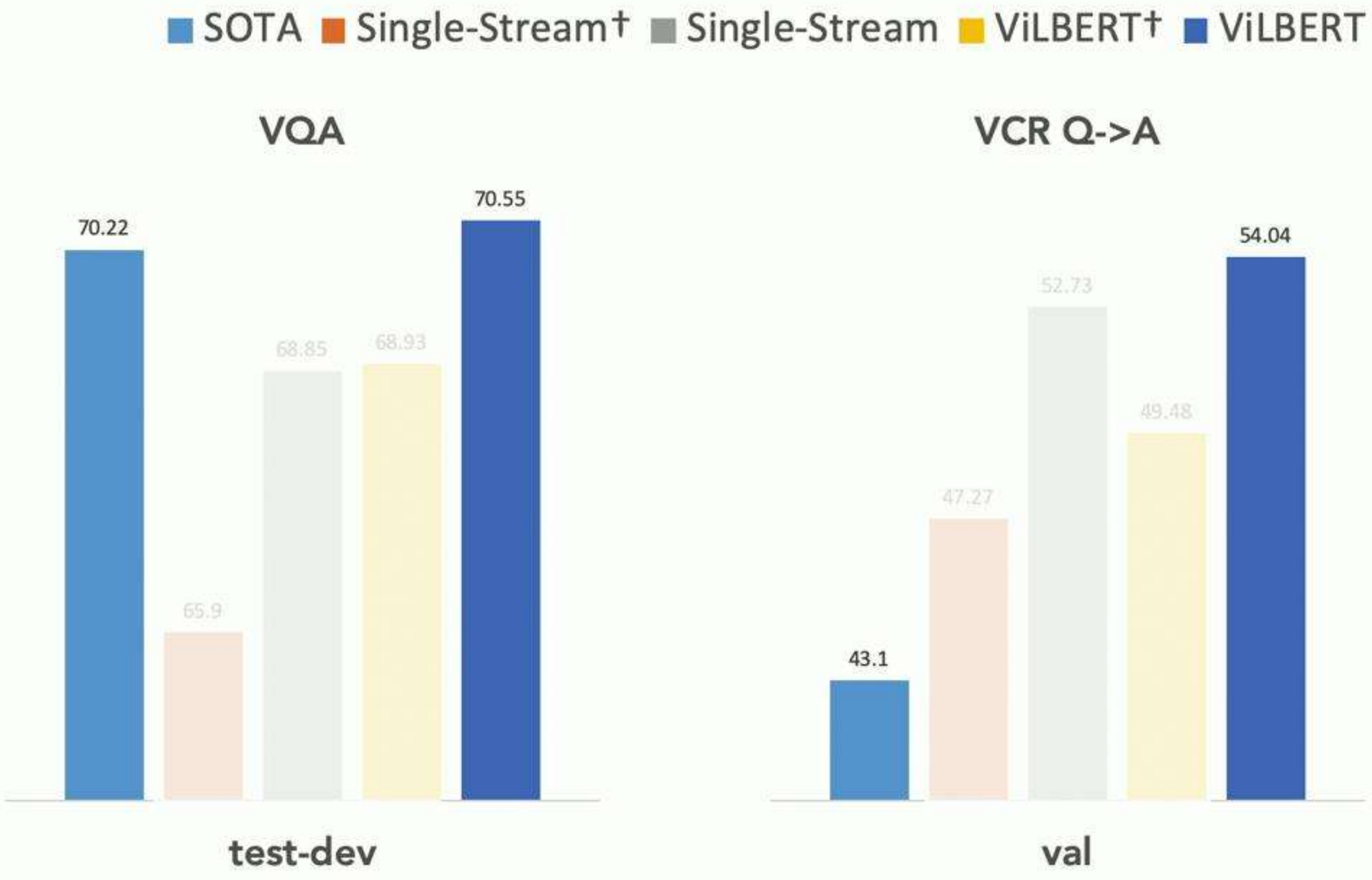
Caption-Based Image Retrieval

[Antol 2015, zeller 2018, Yu 2016, Plummer 2015]

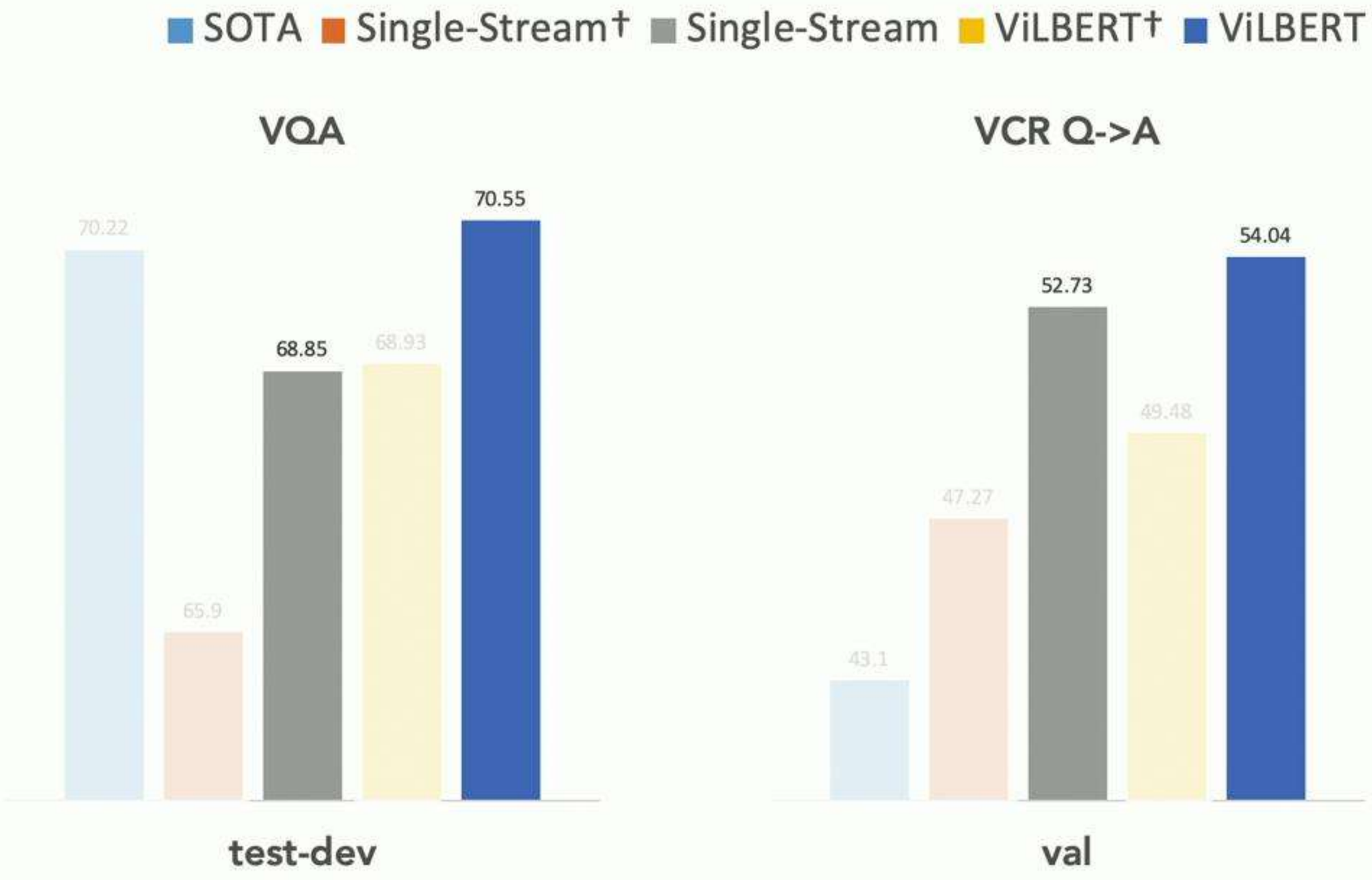
Results


SOTA
Single-Stream†
Single-Stream
ViLBERT†
ViLBERT
VQA
70.22
65.9
68.85
68.93
70.55
test-dev
VCR Q->A
43.1
47.27
52.73
49.48
54.04
val

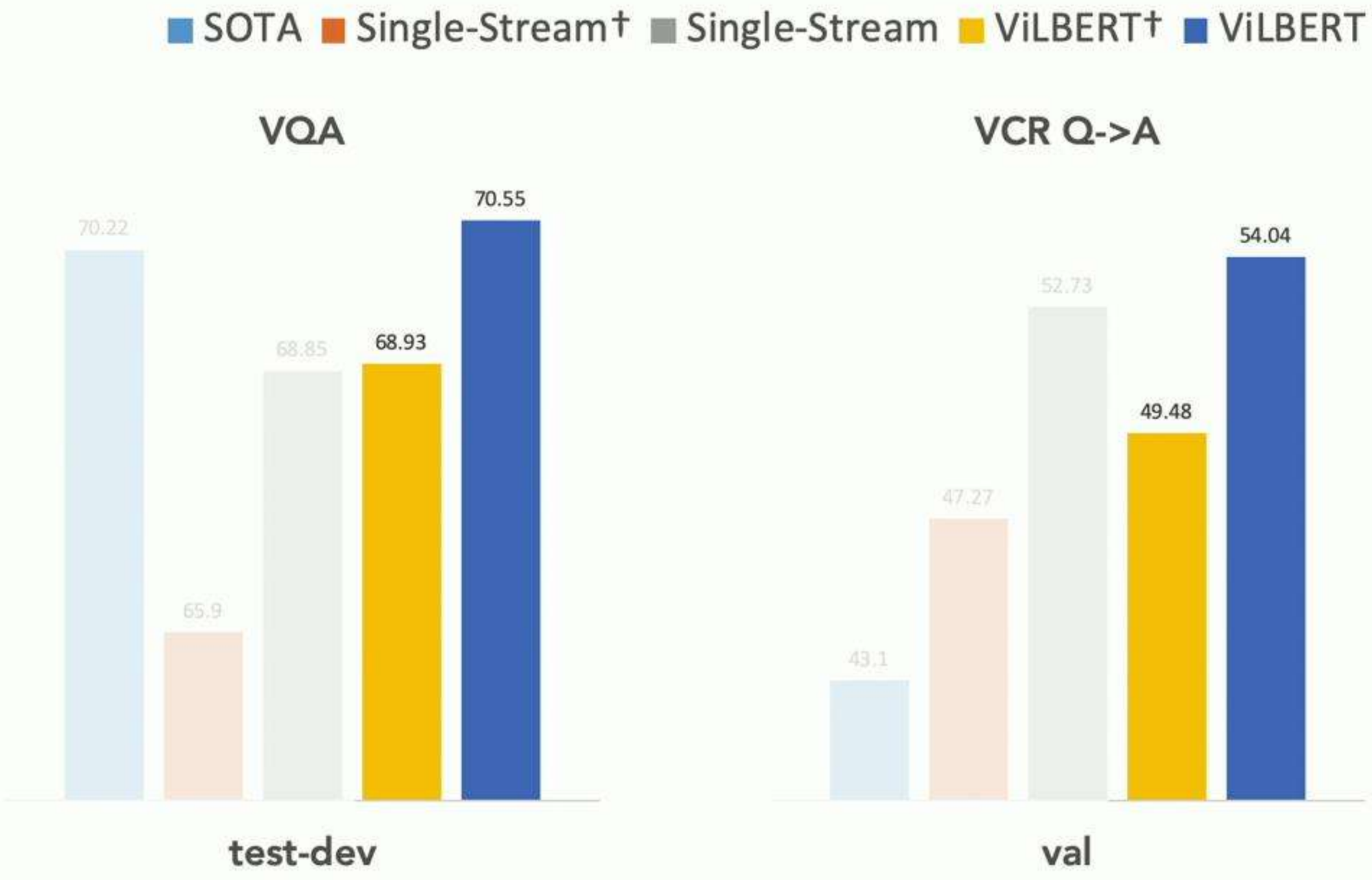
Results


SOTA
Single-Stream†
Single-Stream
ViLBERT†
ViLBERT
VQA
70.22
65.9
68.85
68.93
70.55
test-dev
VCR Q->A
43.1
47.27
52.73
49.48
54.04
val

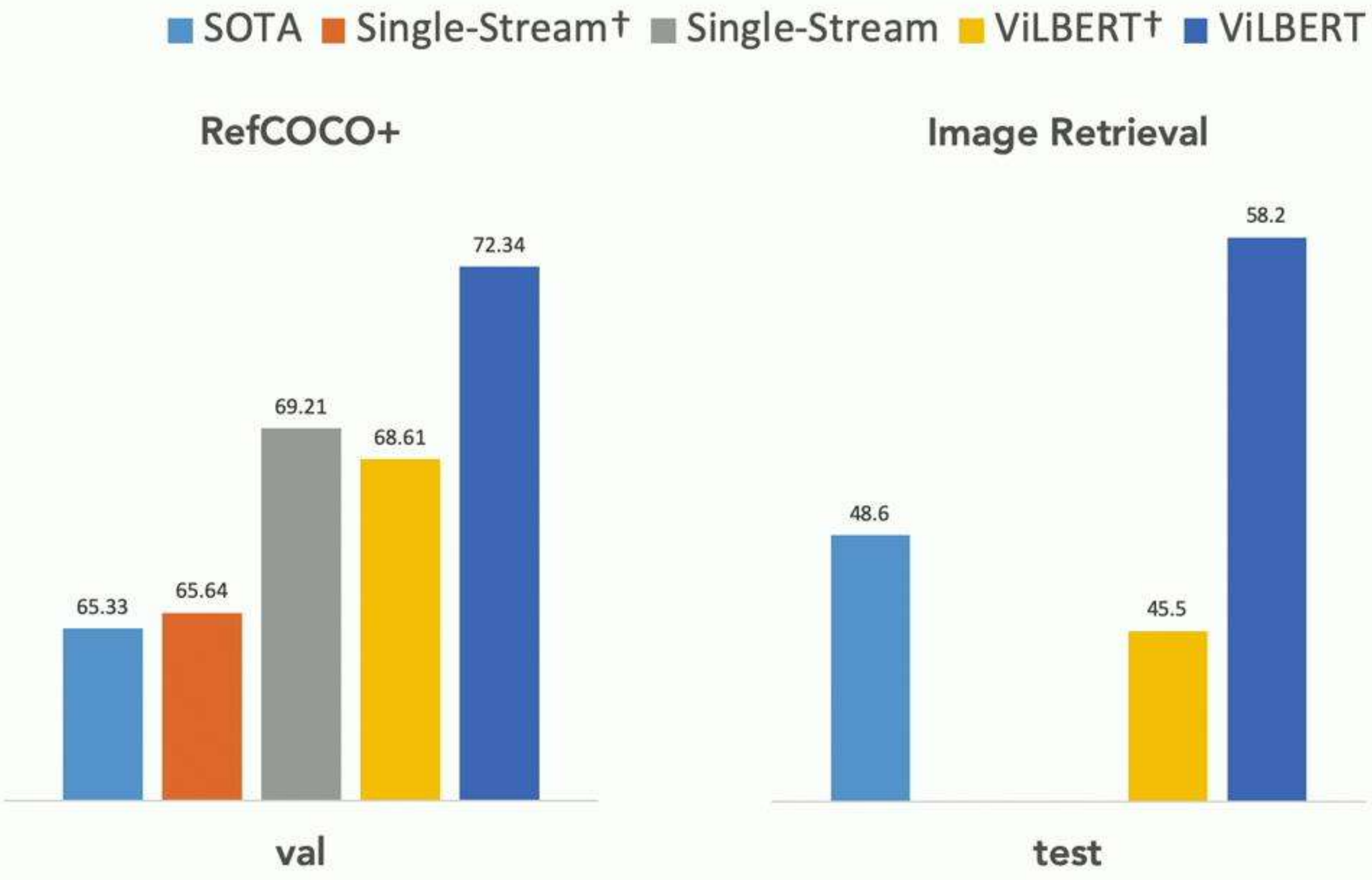
Results


SOTA
Single-Stream†
Single-Stream
ViLBERT†
ViLBERT
VQA
70.22
65.9
68.85
68.93
70.55
test-dev
VCR Q->A
43.1
47.27
52.73
49.48
54.04
val

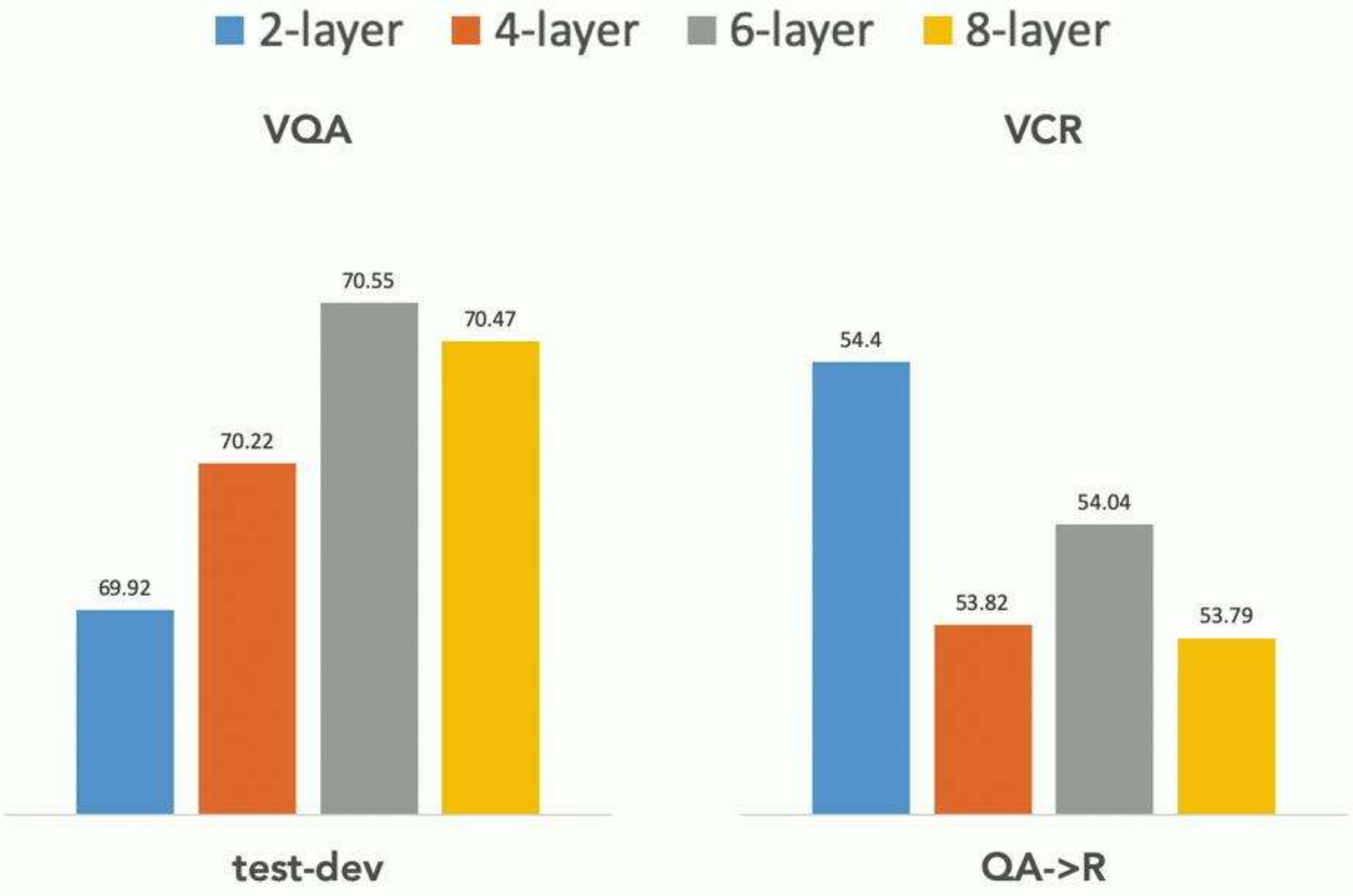
Results


SOTA
Single-Stream†
Single-Stream
ViLBERT†
ViLBERT
VQA
70.22
65.9
68.85
68.93
70.55
test-dev
VCR Q->A
43.1
47.27
52.73
49.48
54.04
val

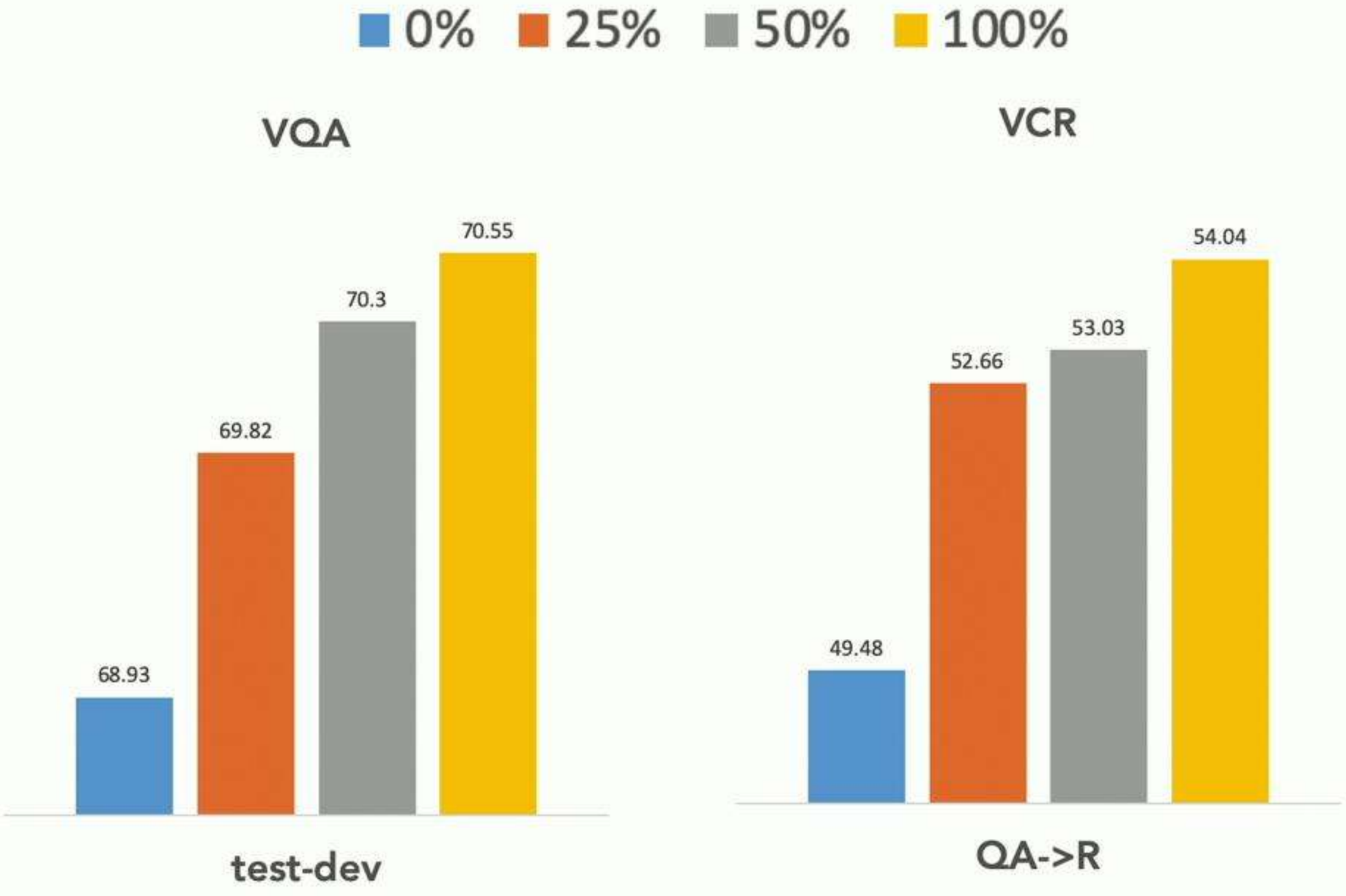
Results


SOTA
Single-Stream†
Single-Stream
ViLBERT†
ViLBERT
RefCOCO+
65.33
65.64
69.21
68.61
72.34
val
Image Retrieval
48.6
45.5
58.2
test

Ablations - Depth



Ablations - Pretraining



Concurrent Work

	Method	Architecture	Visual Token	Pre-train Datasets	Pre-train Tasks	Downstream Tasks
Published Works	VideoBERT (Sun et al., 2019b)	single cross-modal Transformer	video frame	Cooking312K (Sun et al., 2019b)	1) sentence-image alignment 2) masked language modeling 3) masked visual-words prediction	1) zero-shot action classification 2) video captioning
Works Under Review / Just Got Accepted	CBT (Sun et al., 2019a)	two single-modal Transformer (vision & language respectively) + one cross-modal Transformer	video frame	Cooking312K (Sun et al., 2019b)	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature regression	1) action anticipation 2) video captioning
	ViLBERT (Lu et al., 2019)	one single-modal Transformer (language) + one cross-modal Transformer (with restricted attention pattern)	image RoI	Conceptual Captions (Sharma et al., 2018)	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature classification	1) visual question answering 2) visual commonsense reasoning 3) grounding referring expressions 4) image retrieval 5) zero-shot image retrieval
	B2T2 (Alberti et al., 2019)	single cross-modal Transformer	image RoI	Conceptual Captions (Sharma et al., 2018)	1) sentence-image alignment 2) masked language modeling	1) visual commonsense reasoning
	LXMERT (Hao Tan, 2019)	two single-modal Transformer (vision & language respectively) + one cross-modal Transformer	image RoI	‡ COCO Caption + VG Caption + VG QA + VQA + GQA	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature classification 4) masked visual-feature regression 5) visual question answering	1) visual question answering 2) natural language visual reasoning
Works in Progress	VisualBERT (Li et al., 2019b)	single cross-modal Transformer	image RoI	COCO Caption (Chen et al., 2015)	1) sentence-image alignment 2) masked language modeling	1) visual question answering 2) visual commonsense reasoning 3) natural language visual reasoning 4) grounding phrases
	Unicoder-VL (Li et al., 2019a)	single cross-modal Transformer	image RoI	Conceptual Captions (Sharma et al., 2018)	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature classification	1) image-text retrieval 2) zero-shot image-text retrieval
	Our VL-BERT	single cross-modal Transformer	image RoI	Conceptual Captions (Sharma et al., 2018)	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature classification	1) visual question answering 2) visual commonsense reasoning 3) grounding referring expressions

‡ LXMERT is pre-trained on COCO Caption (Chen et al., 2015), VG Caption (Krishna et al., 2017), VG QA (Zhu et al., 2016), VQA (Antol et al., 2015) and GQA (Hudson & Manning, 2019).

Image credit: VL-BERT

Concurrent Work

	Method	Architecture	Visual Token	Pre-train Datasets	Pre-train Tasks	Downstream Tasks
Published Works	VideoBERT (Sun et al., 2019b)	single cross-modal Transformer	video frame	Cooking312K (Sun et al., 2019b)	1) sentence-image alignment 2) masked language modeling 3) masked visual-words prediction	1) zero-shot action classification 2) video captioning
Works Under Review / Just Got Accepted	CBT (Sun et al., 2019a)	two single-modal Transformer (vision & language respectively) + one cross-modal Transformer	video frame	Cooking312K (Sun et al., 2019b)	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature regression	1) action anticipation 2) video captioning
	ViLBERT (Lu et al., 2019)	one single-modal Transformer (language) + one cross-modal Transformer (with restricted attention pattern)	image RoI	Conceptual Captions (Sharma et al., 2018)	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature classification	1) visual question answering 2) visual commonsense reasoning 3) grounding referring expressions 4) image retrieval 5) zero-shot image retrieval
	B2T2 (Alberti et al., 2019)	single cross-modal Transformer	image RoI	Conceptual Captions (Sharma et al., 2018)	1) sentence-image alignment 2) masked language modeling	1) visual commonsense reasoning
	LXMERT (Hao Tan, 2019)	two single-modal Transformer (vision & language respectively) + one cross-modal Transformer	image RoI	‡ COCO Caption + VG Caption + VG QA + VQA + GQA	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature classification 4) masked visual-feature regression 5) visual question answering	1) visual question answering 2) natural language visual reasoning
Works in Progress	VisualBERT (Li et al., 2019b)	single cross-modal Transformer	image RoI	COCO Caption (Chen et al., 2015)	1) sentence-image alignment 2) masked language modeling	1) visual question answering 2) visual commonsense reasoning 3) natural language visual reasoning 4) grounding phrases
	Unicoder-VL (Li et al., 2019a)	single cross-modal Transformer	image RoI	Conceptual Captions (Sharma et al., 2018)	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature classification	1) image-text retrieval 2) zero-shot image-text retrieval
	Our VL-BERT	single cross-modal Transformer	image RoI	Conceptual Captions (Sharma et al., 2018)	1) sentence-image alignment 2) masked language modeling 3) masked visual-feature classification	1) visual question answering 2) visual commonsense reasoning 3) grounding referring expressions

‡ LXMERT is pre-trained on COCO Caption (Chen et al., 2015), VG Caption (Krishna et al., 2017), VG QA (Zhu et al., 2016), VQA (Antol et al., 2015) and GQA (Hudson & Manning, 2019).

Image credit: VL-BERT

Results

Summary

Results

Summary

Task-agnostic visiolinguistic representations pretraining for visual grounding

- Introduce *pretrain-transfer* to vision and language tasks.
- Achieve SOTA on multiple vision and language tasks.

Results

Summary

Task-agnostic visiolinguistic representations pretraining for visual grounding

- Introduce *pretrain-transfer* to vision and language tasks.
- Achieve SOTA on multiple vision and language tasks.

Limitations

The model can still learn inconsistent grounding by task specific finetuning.

- Training multiple vision and language task together.

This Talk

Recent & Future Work

Results

Summary

Task-agnostic visiolinguistic representations pretraining for visual grounding

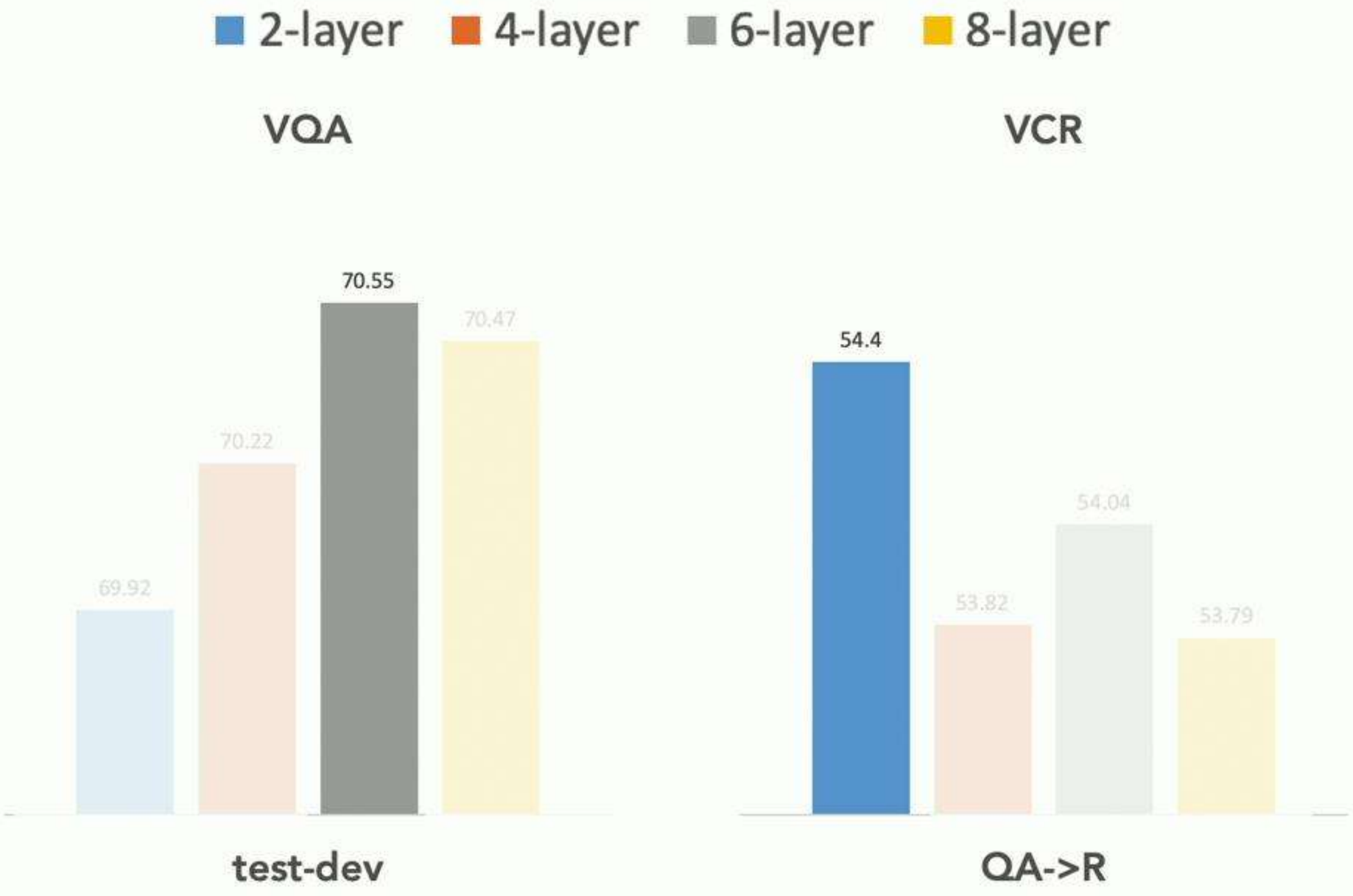
- Introduce *pretrain-transfer* to vision and language tasks.
- Achieve SOTA on multiple vision and language tasks.

Limitations

The model can still learn inconsistent grounding by task specific finetuning.

- Training multiple vision and language task together.

Ablations - Depth



This Talk

Recent & Future Work

Vision And Language Multi-Task Learning

Vision And Language Multi-Task Learning

ViLBERT Problems:

- Inconsistent grounding by task specific finetuning.
- Four V&L tasks.
- Model is huge, overfitting.

Vision And Language Multi-Task Learning

ViLBERT Problems:

- Inconsistent grounding by task specific finetuning.
- Four V&L tasks.
- Model is huge, overfitting.

What we want:

- More tasks.
- Consistent Grounding.
- Explore the limit of the model.

Vision And Language Multi-Task Learning

ViLBERT Problems:

- Inconsistent grounding by task specific finetuning.
- Four V&L tasks.
- Model is huge, overfitting.

What we want:

- More tasks.
- Consistent Grounding.
- Explore the limit of the model.

One Model for ALL V&L:

Vision And Language Multi-Task Learning

ViLBERT Problems:

- Inconsistent grounding by task specific finetuning.
- Four V&L tasks.
- Model is huge, overfitting.

What we want:

- More tasks.
- Consistent Grounding.
- Explore the limit of the model.

One Model for ALL V&L:

VQA
- VQA
- Genome QA
- GQA

Image Description
- Caption based Retrieval (COCO)
- Caption based Retrieval (COCO)

Referring Expression
- Ref COCO
- Ref COCO+
- Ref COCOg
- Visual 7w
- GuessWhat

V&L Verification
- NLVR2
- Visual Entailment

Vision And Language Multi-Task Learning

One Model for ALL V&L:

VQA
- VQA
- Genome QA
- GQA

Image Description
- Caption based Retrieval (COCO)
- Caption based Retrieval (COCO)

Referring Expression
- Ref COCO
- Ref COCO+
- Ref COCOg
- Visual 7w
- GuessWhat

V&L Verification
- NLVR2
- Visual Entailment

Vision And Language Multi-Task Learning

One Model for ALL V&L:

VQA
- VQA
- Genome QA
- GQA

Image Description
- Caption based Retrieval (COCO)
- Caption based Retrieval (COCO)

Referring Expression
- Ref COCO
- Ref COCO+
- Ref COCOg
- Visual 7w
- GuessWhat

V&L Verification
- NLVR2
- Visual Entailment

- Benchmark ALL vision and language understanding tasks with ViLBERT.

Vision And Language Multi-Task Learning

One Model for ALL V&L:

VQA
- VQA
- Genome QA
- GQA

Image Description
- Caption based Retrieval (COCO)
- Caption based Retrieval (COCO)

Referring Expression
- Ref COCO
- Ref COCO+
- Ref COCOg
- Visual 7w
- GuessWhat

V&L Verification
- NLVR2
- Visual Entailment

- Benchmark ALL vision and language understanding tasks with ViLBERT.
- Study inter-connections within task group and between task group.

Vision And Language Multi-Task Learning

One Model for ALL V&L:

VQA
- VQA
- Genome QA
- GQA

Image Description
- Caption based Retrieval (COCO)
- Caption based Retrieval (COCO)

Referring Expression
- Ref COCO
- Ref COCO+
- Ref COCOg
- Visual 7w
- GuessWhat

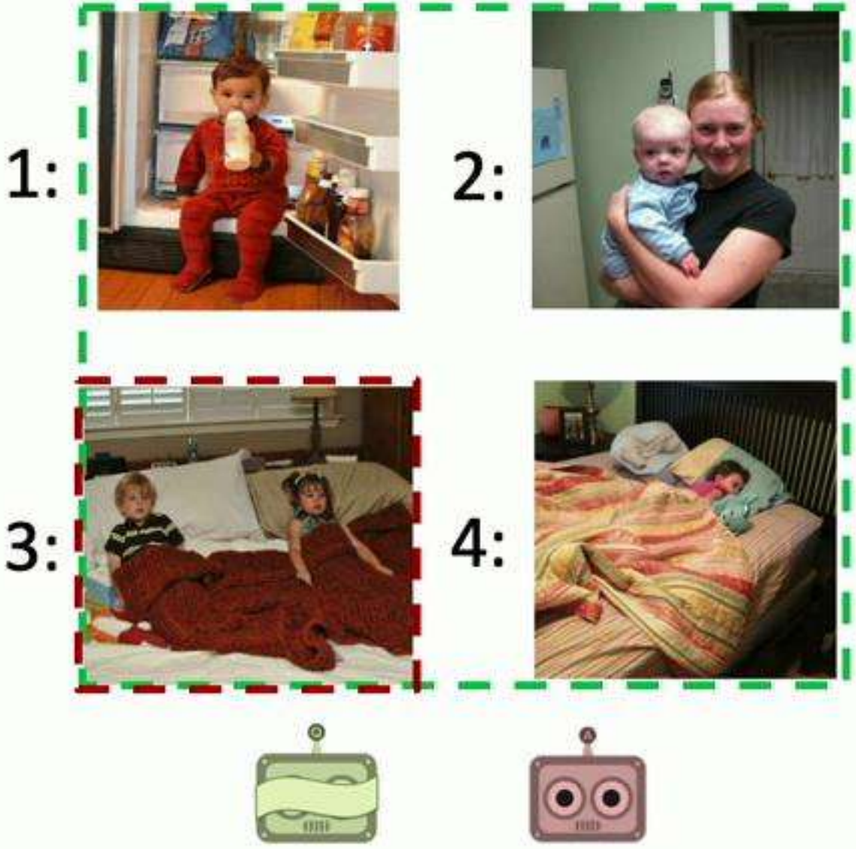
V&L Verification
- NLVR2
- Visual Entailment

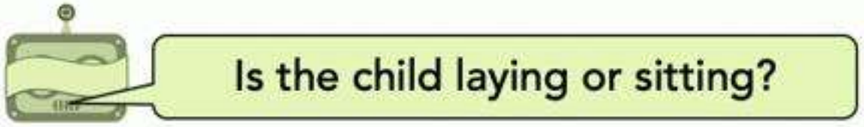
- Benchmark ALL vision and language understanding tasks with ViLBERT.
- Study inter-connections within task group and between task group.
- Explainable AI : Use other task outputs to provide explanations.

Dialog without Dialog

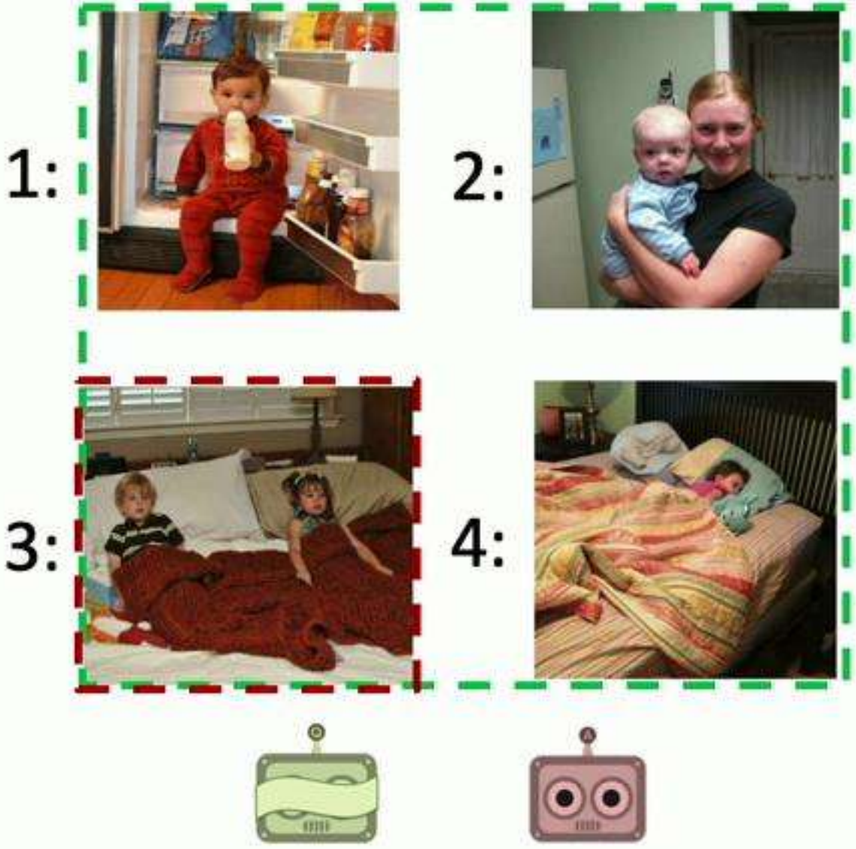
Dialog without Dialog

Dialog without Dialog

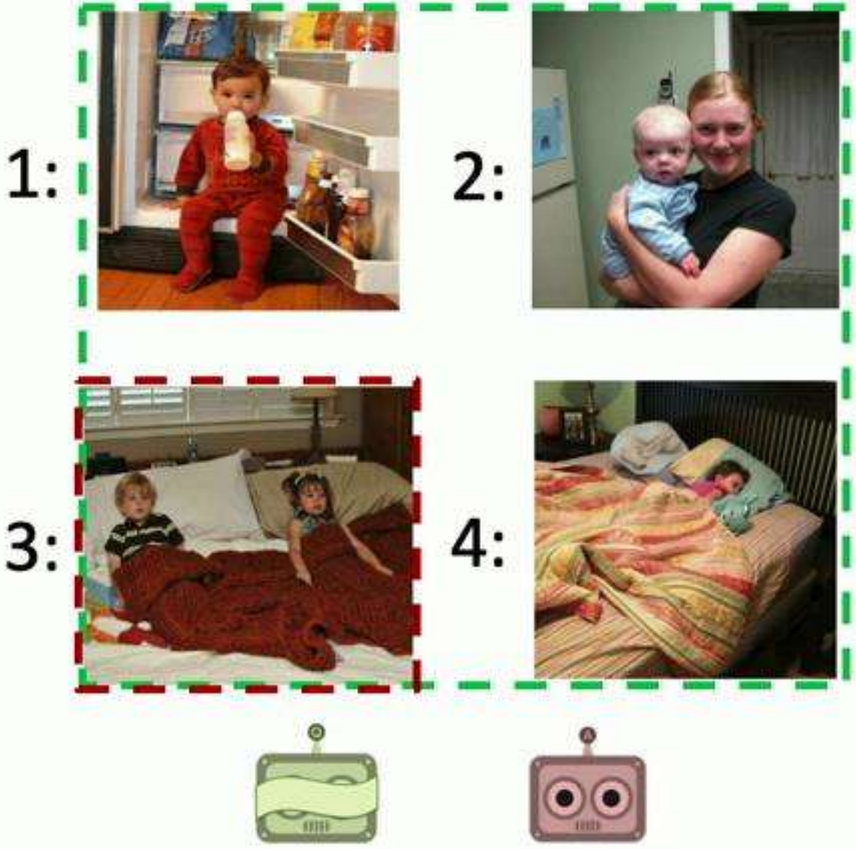


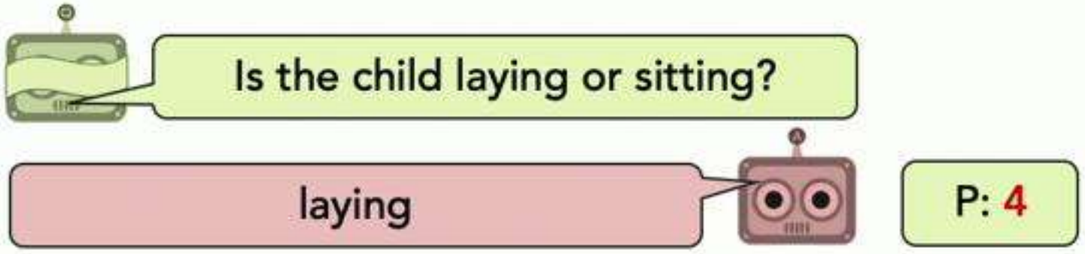


Dialog without Dialog

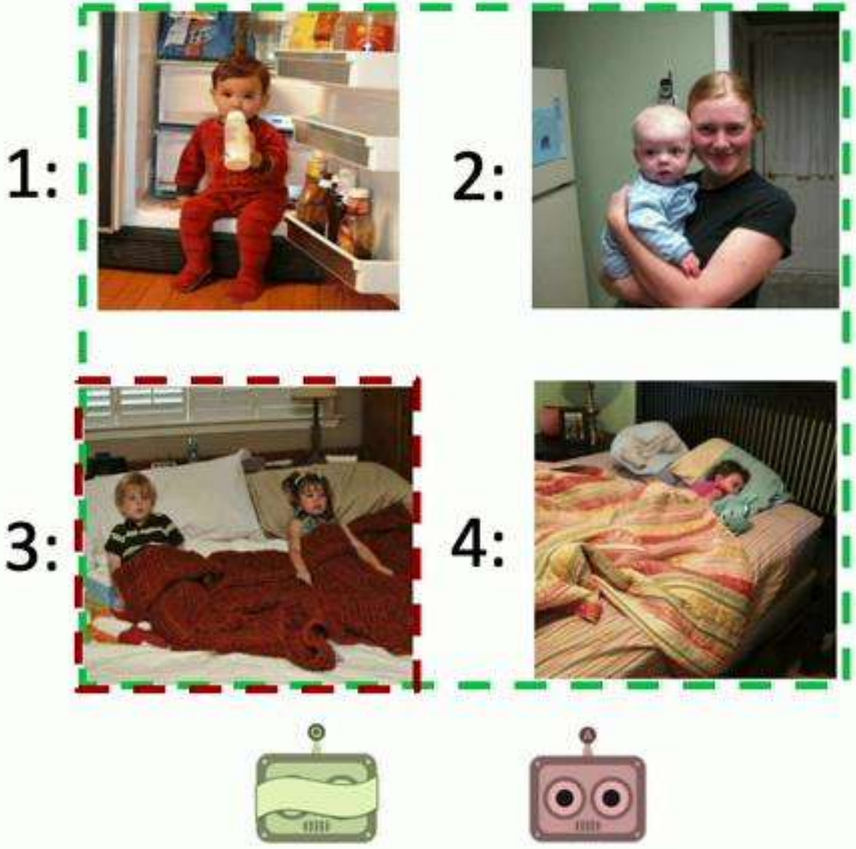


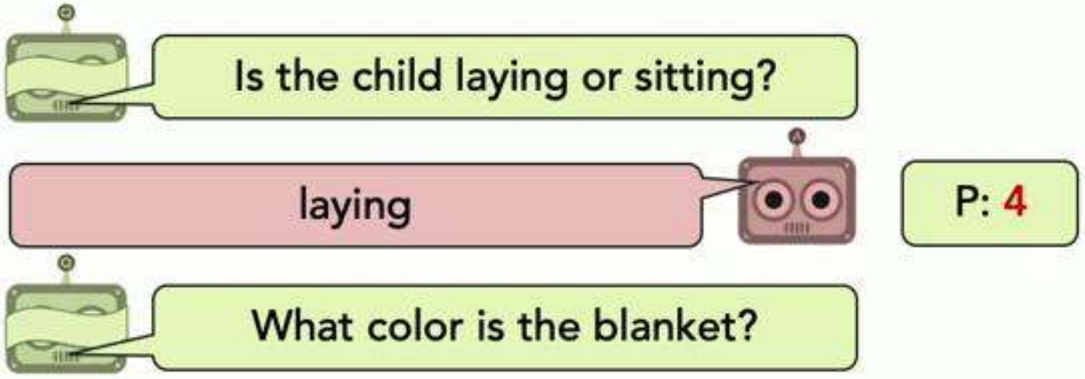
Dialog without Dialog



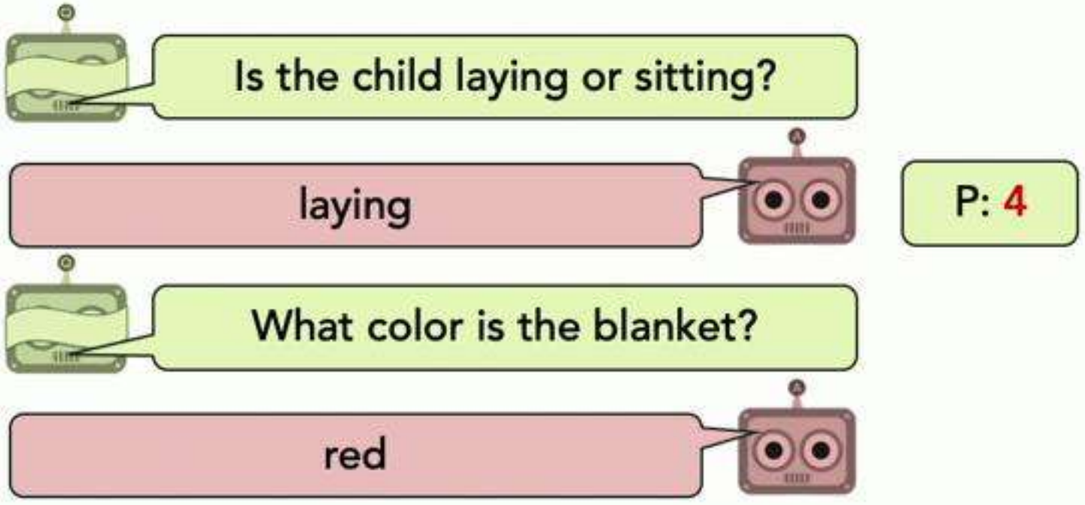


Dialog without Dialog

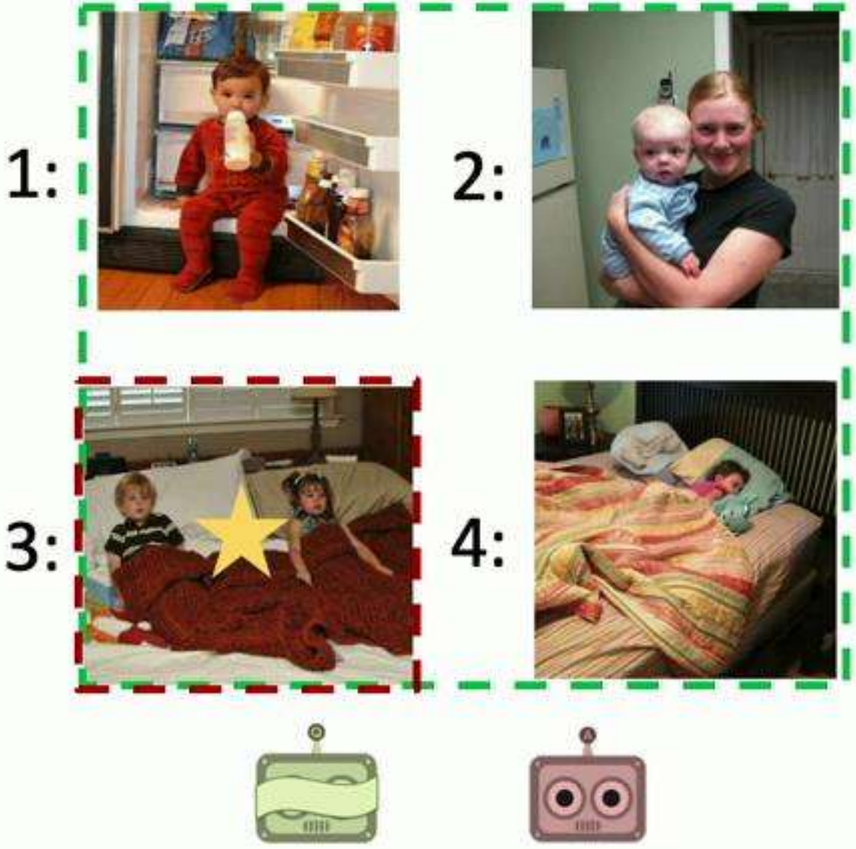


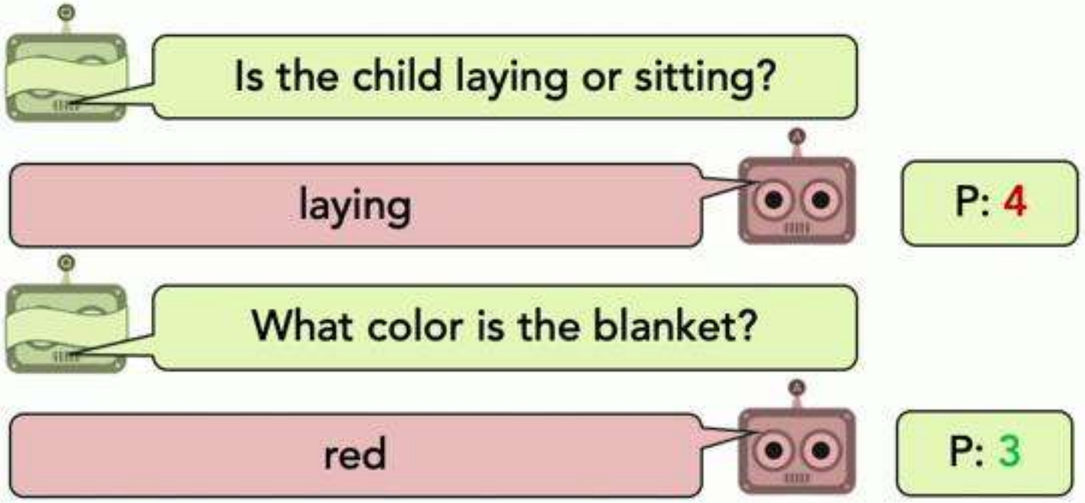


Dialog without Dialog

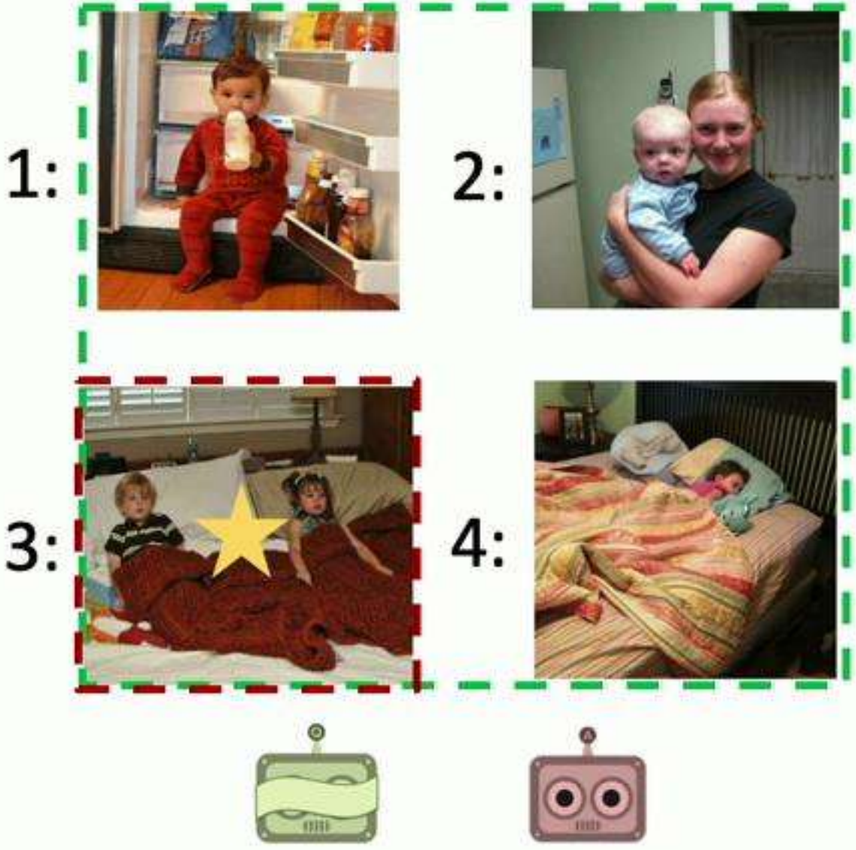


Dialog without Dialog

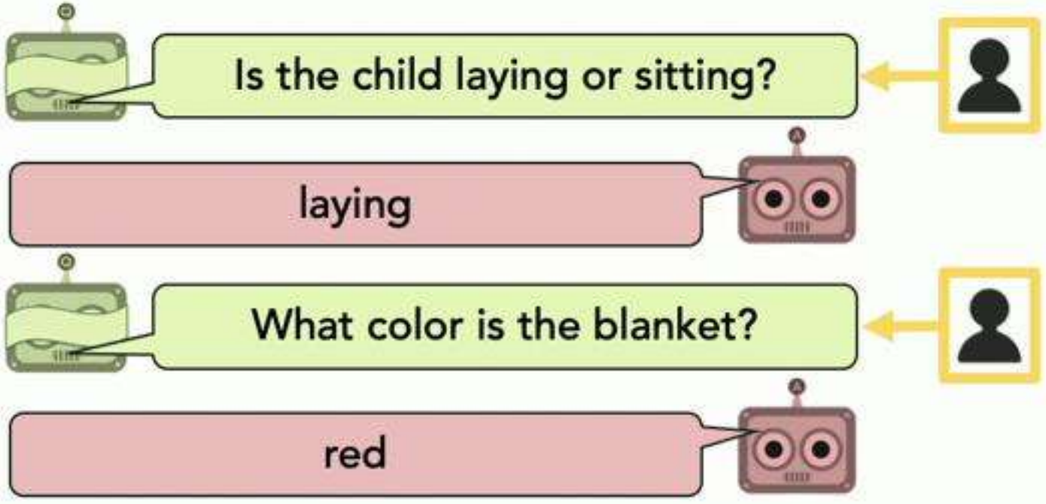




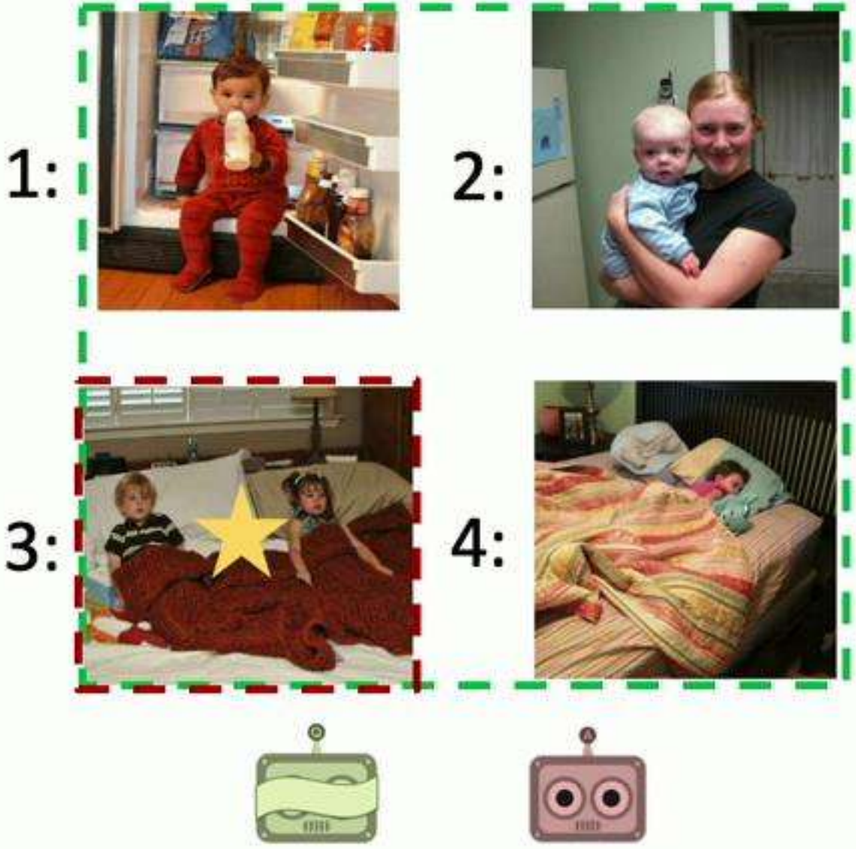
Dialog without Dialog



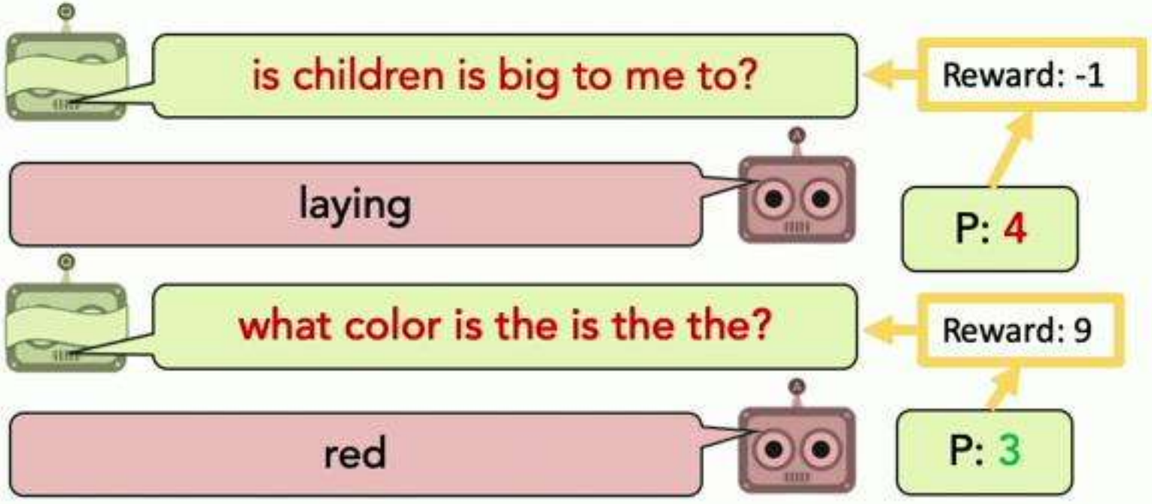
How does Q-Bot know what to say?



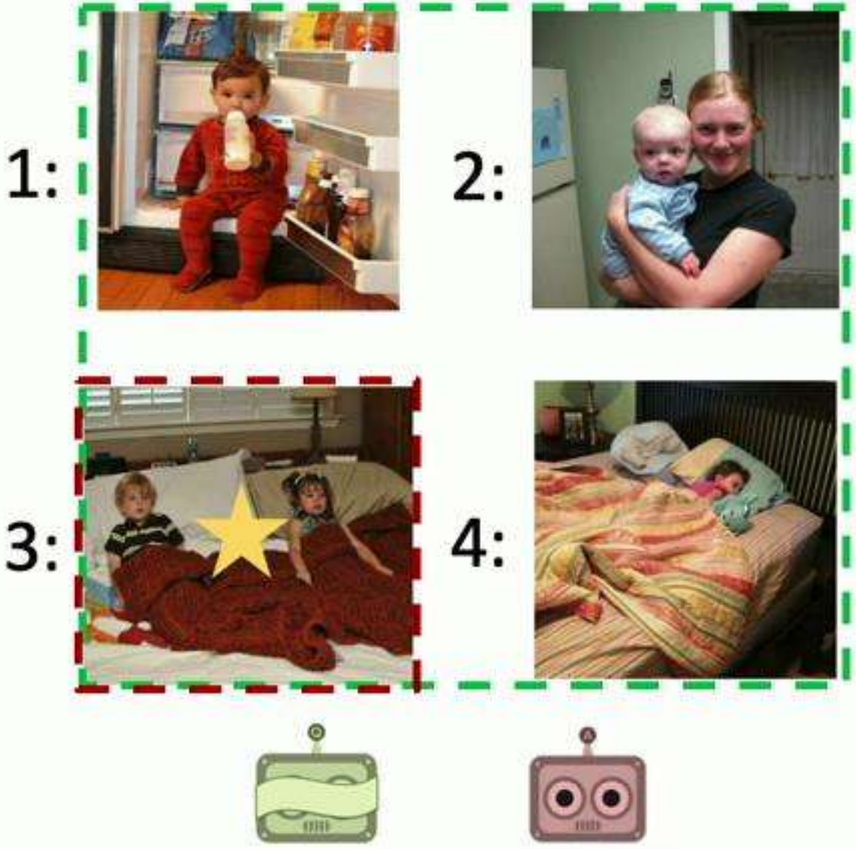
Dialog without Dialog



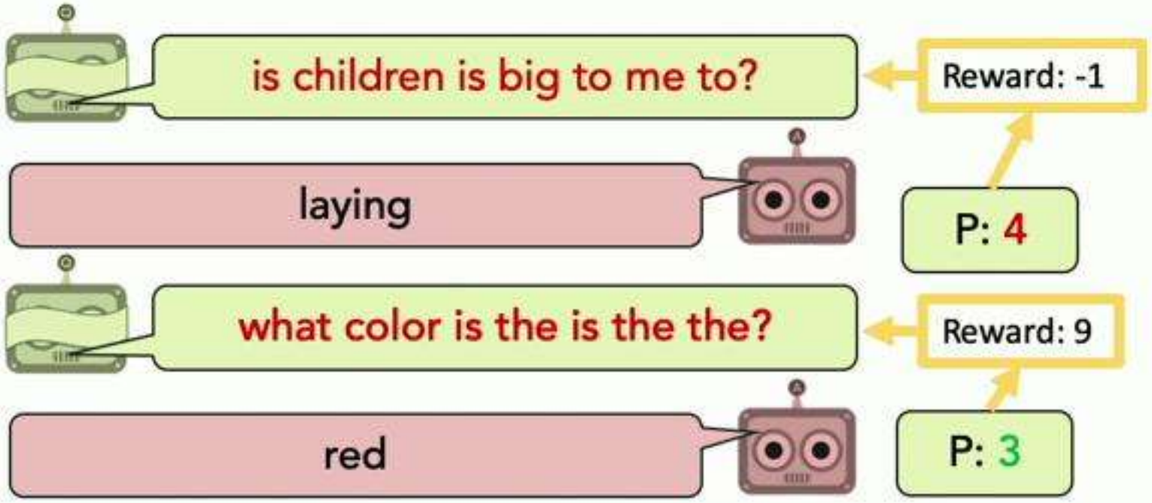
How does Q-Bot know what to say?



Dialog without Dialog



How does Q-Bot know what to say?



Goals (with sparse reward):

1. Language fluency
2. Task performance

Ongoing Work

Language Fluency from VQA

Who is wearing glasses?

man

woman

Is the umbrella upside down?

yes

no

Learn how to ask from VQA

Ongoing Work

Language Fluency from VQA

Who is wearing glasses?

man

woman

Is the umbrella upside down?

yes

no

Learn how to ask from VQA

Key Insight: LM with discrete latent action space

Interpolation

1. how many beds ?

10.where is he looking ?

Ongoing Work

Language Fluency from VQA

Who is wearing glasses?

man

woman

Is the umbrella upside down?

yes

no

Learn how to ask from VQA

Key Insight: LM with discrete latent action space

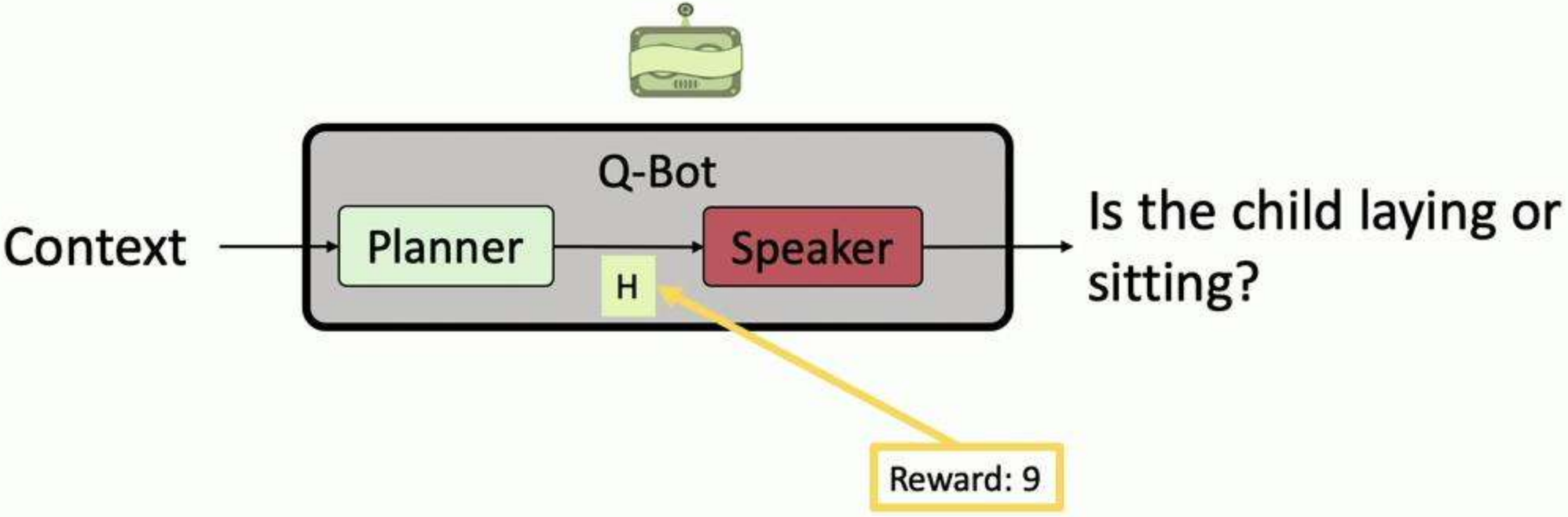
Interpolation

1. how many beds ?
2. how many cats ?
3. how many dogs ?
4. where is the dog ?
5. where is the man ?
6. where is this man ?
7. where is this woman ?
8. where is this ?
9. where is he ?
10. where is he looking ?

Ongoing Work

Task Performance via Fine-Tuning without Forgetting Language

Key Insight: Reinforce discrete latent action space.



Ongoing Work

Initial Results

Ongoing Work

Initial Results

target

Ongoing Work

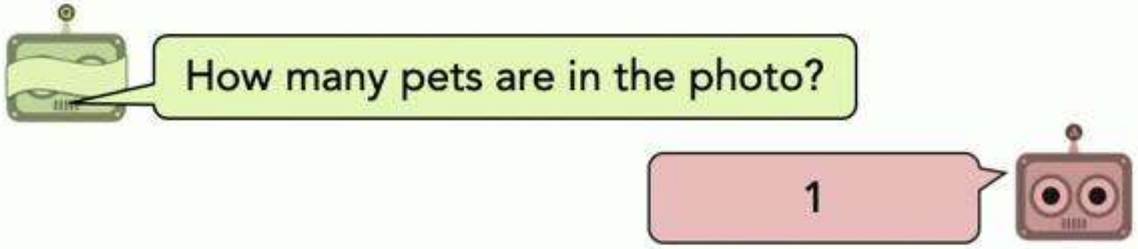
Initial Results

target

Ongoing Work

Initial Results

target



Ongoing Work

Initial Results

target

How many pets are in the photo?

1

P: 1

Ongoing Work

Initial Results

target

How many pets are in the photo?

1

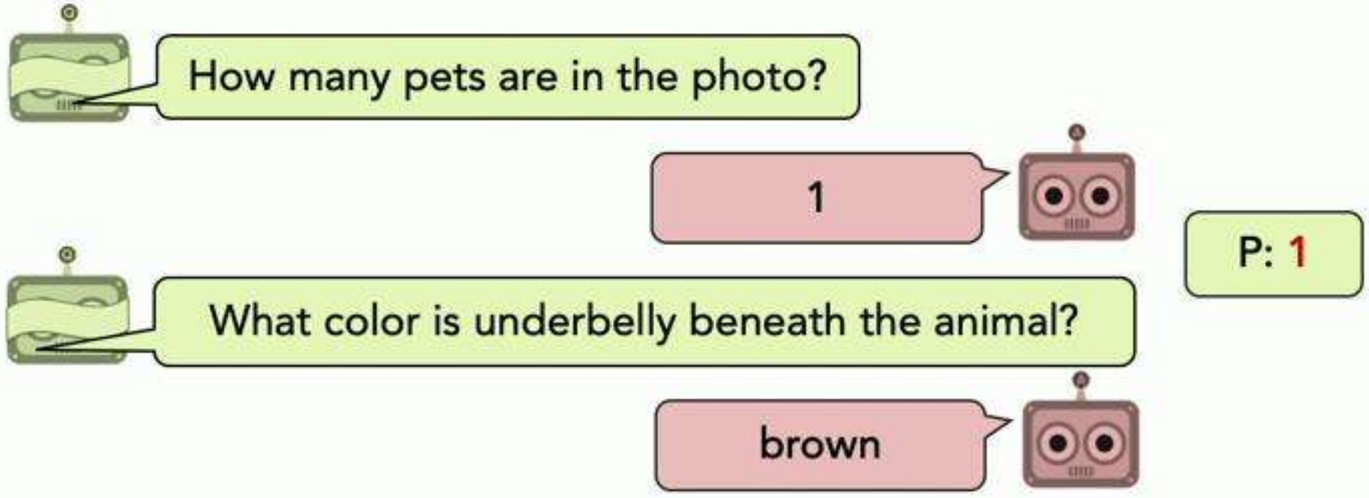
What color is underbelly beneath the animal?

P: 1

Ongoing Work

Initial Results

target



Ongoing Work

Initial Results

target

How many pets are in the photo?

1

P: 1

What color is underbelly beneath the animal?

brown

P: 1

Ongoing Work

Initial Results

target

How many pets are in the photo?

1

P: 1

What color is underbelly beneath the animal?

brown

P: 1

What color is the zebra's fur ?

Ongoing Work

Initial Results

target

How many pets are in the photo?

1

P: 1

What color is underbelly beneath the animal?

brown

P: 1

What color is the zebra's fur ?

Black and white

Ongoing Work

Initial Results

target

How many pets are in the photo?

1

P: 1

What color is underbelly beneath the animal?

brown

P: 1

What color is the zebra's fur ?

Black and white

P: 2

Ongoing Work

Initial Results

target

How many pets are in the photo?

1

P: 1

What color is underbelly beneath the animal?

brown

P: 1

What color is the zebra's fur ?

Black and white

P: 2

Is the animal looking at the camera?

Future Interest

Multi-modality representation learning in vision, language, sound and action.

Future Interest

Multi-modality representation learning in vision, language, sound and action.

- Environment around us is not un-modal

Future Interest

Multi-modality representation learning in vision, language, sound and action.

- Environment around us is not un-modal

Table:

Future Interest

Multi-modality representation learning in vision, language, sound and action.

- Environment around us is not un-modal

Table: 

Future Interest

Multi-modality representation learning in vision, language, sound and action.

- Environment around us is not un-modal

Table:  

Future Interest

Multi-modality representation learning in vision, language, sound and action.

- Environment around us is not un-modal

Table:  

Providing a level surface on which objects may be placed

Future Interest

Multi-modality representation learning in vision, language, sound and action.

- Environment around us is not un-modal

Table:  

Providing a level surface on which objects may be placed

- Learn a representation that connect all the modalities.
  - More complete understanding of the environment.

Future Interest

Emerging of goal orientated dialog with natural language.

- Dialog is the most natural way to communicate.

Future Interest

Emerging of goal orientated dialog with natural language.

- Dialog is the most natural way to communicate.
  - Impossible to collect dialog data for all the dialog tasks.

Future Interest

Emerging of goal orientated dialog with natural language.

- Dialog is the most natural way to communicate.
  - Impossible to collect dialog data for all the dialog tasks.
- Emerging of language (machine code) is not useful.

Future Interest

Emerging of goal orientated dialog with natural language.

- Dialog is the most natural way to communicate.
  - Impossible to collect dialog data for all the dialog tasks.
- Emerging of language (machine code) is not useful.
  - People can not understand them

Future Interest

Emerging of goal orientated dialog with natural language.

- Dialog is the most natural way to communicate.
  - Impossible to collect dialog data for all the dialog tasks.
- Emerging of language (machine code) is not useful.
  - People can not understand them
- Emerging of strategy (language that human can understand)

Future Interest

Common sense abstraction and causal reasoning.

- Intrinsic "motivation" behind the "scene".

Future Interest

Common sense abstraction and causal reasoning.

- Intrinsic "motivation" behind the "scene".
- Emerging of the commonsense and reasoning of the "motivation"

Future Interest

Common sense abstraction and causal reasoning.

- Intrinsic "motivation" behind the "scene".
- Emerging of the commonsense and reasoning of the "motivation"
- More human-like and interpretable agent which helps people in the daily life.

End

QUESTIONS?