When Language Meets Vision:
A Multimodal Perspective on the NLP World

Sapienza NLP Group
Bianca Scarlini
scarlini@di.uniroma1.it

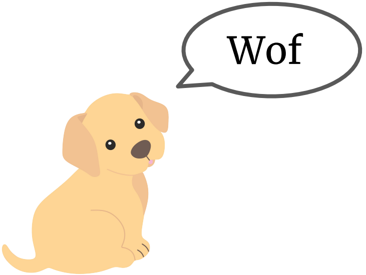
Reading Group @ Sapienza NLP

How do humans learn?

A dog is sitting on a couch with its toy.

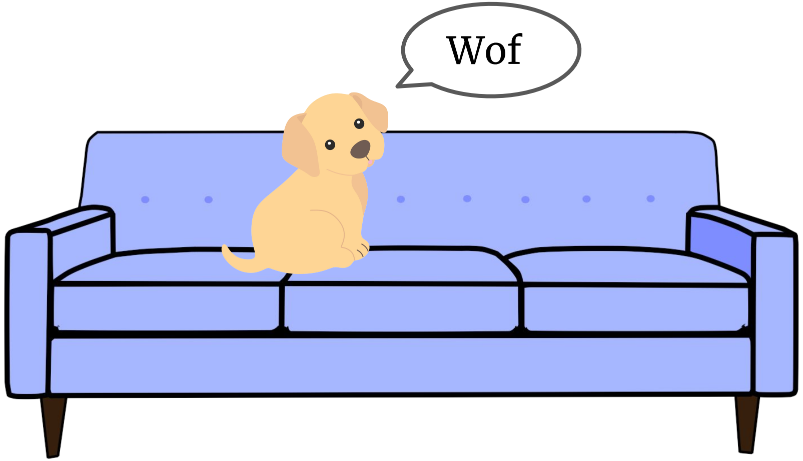
How do humans learn?

A **dog** is sitting on a couch with its toy.



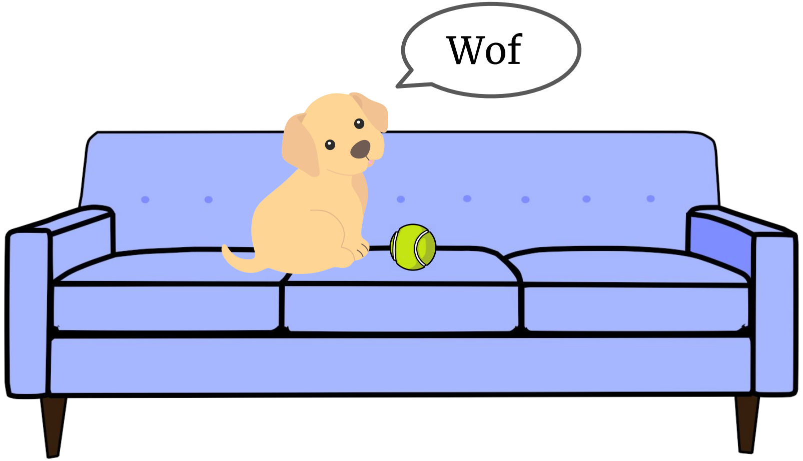
How do humans learn?

A dog is sitting on a **couch** with its toy.

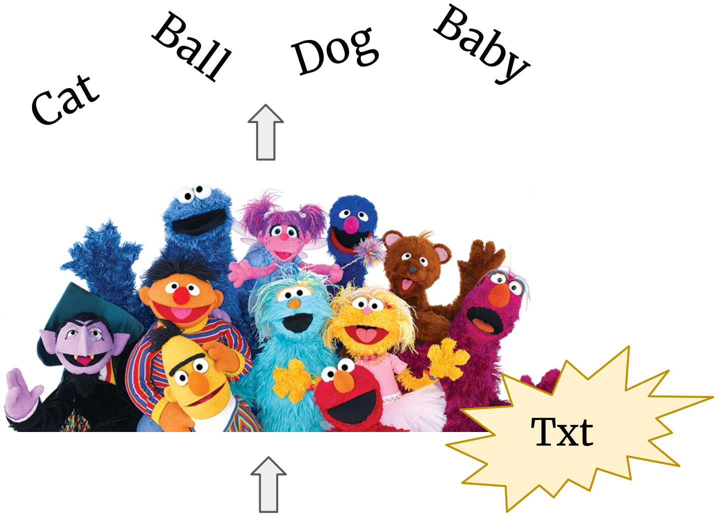


How do humans learn?

A dog is sitting on a couch with its **toy**.



How do text models learn?



A __ is sitting on a couch with its toy.

− Learn syntactic relations *Clark et al. 2019*

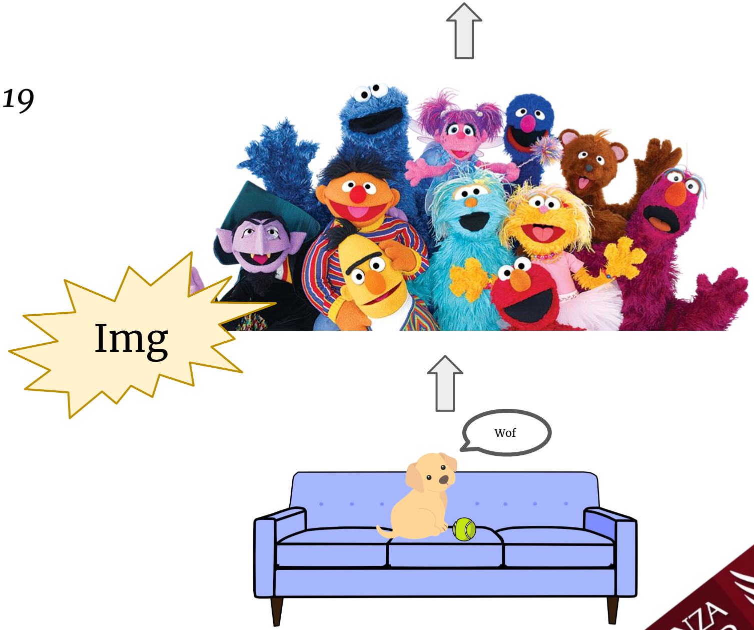
✖ Not grounded in the real world *Bender and Koller 2020*

How do visual models learn?

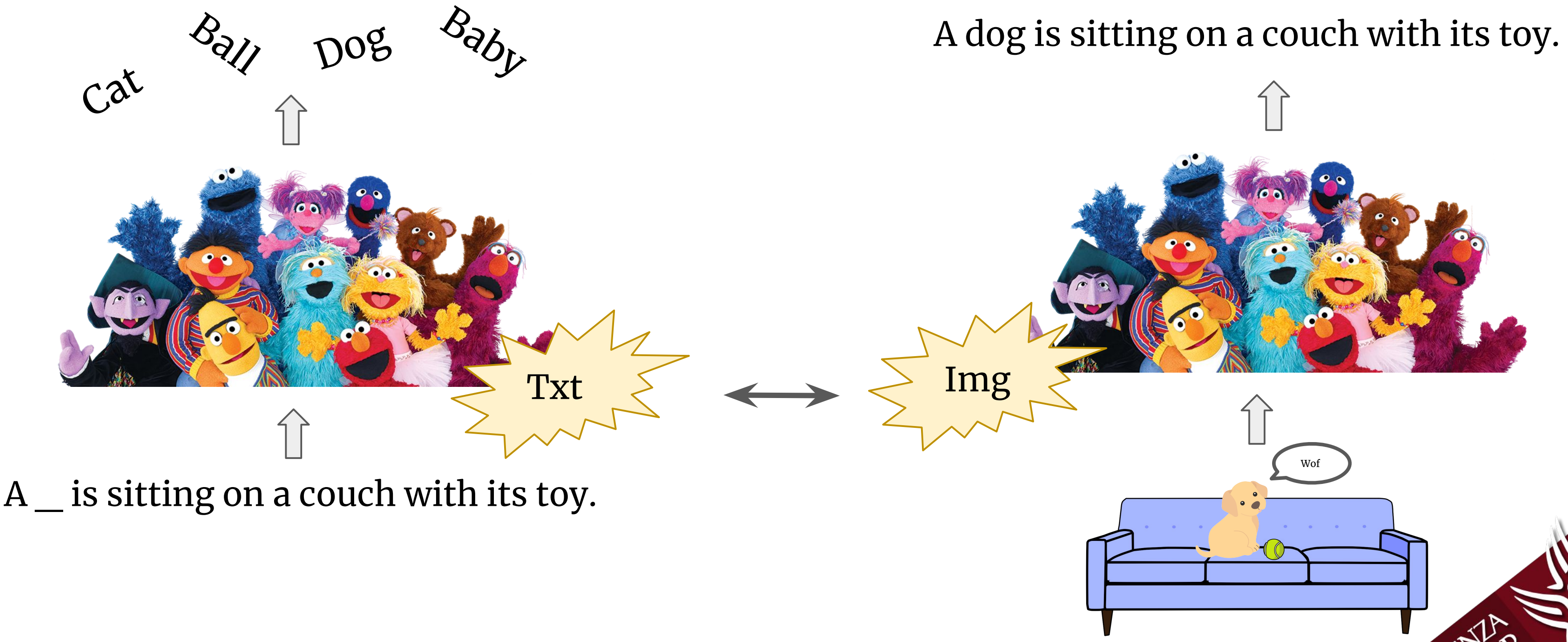
− Learn relations within objects in an image *Cadene et al. 2019*

✖ Need detailed semantics of the image for visual understanding *Johnson et al. 2015*

A dog is sitting on a couch with its toy.



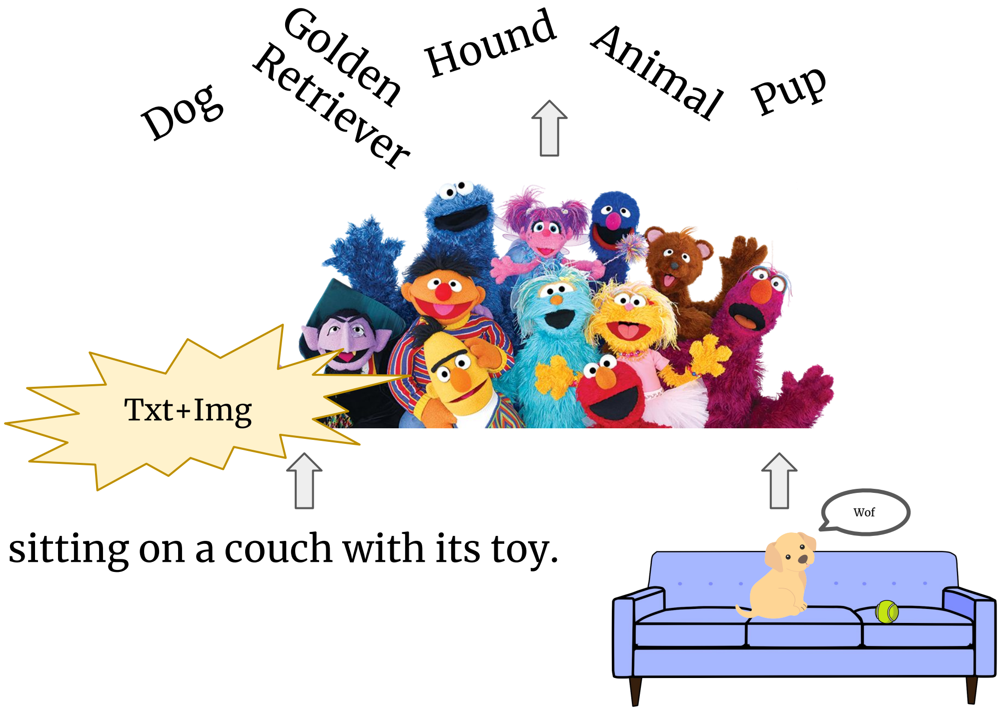
How do models learn?



A dog is sitting on a couch with its toy.

A __ is sitting on a couch with its toy.

How do models learn?



Dog

Golden Retriever

Hound

Animal

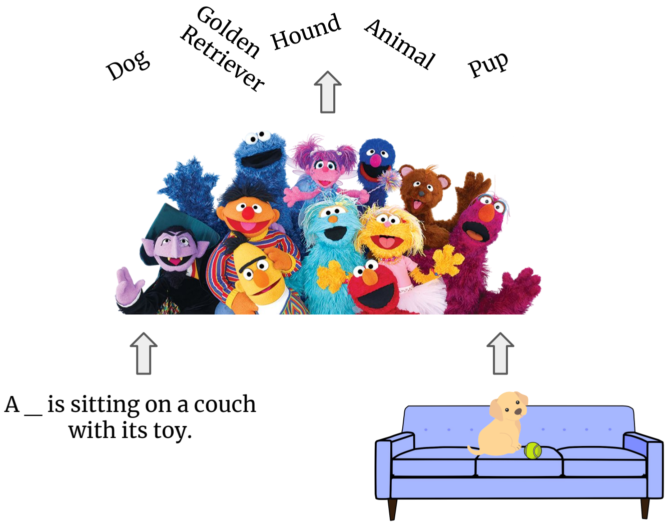
Pup

Txt+Img

A _ is sitting on a couch with its toy.

Cross-modal architectures

Single-stream architecture



Dog

Golden Retriever

Hound

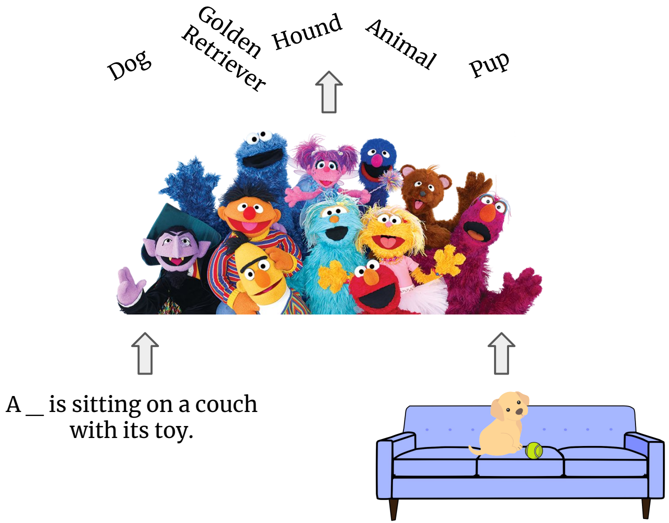
Animal

Pup

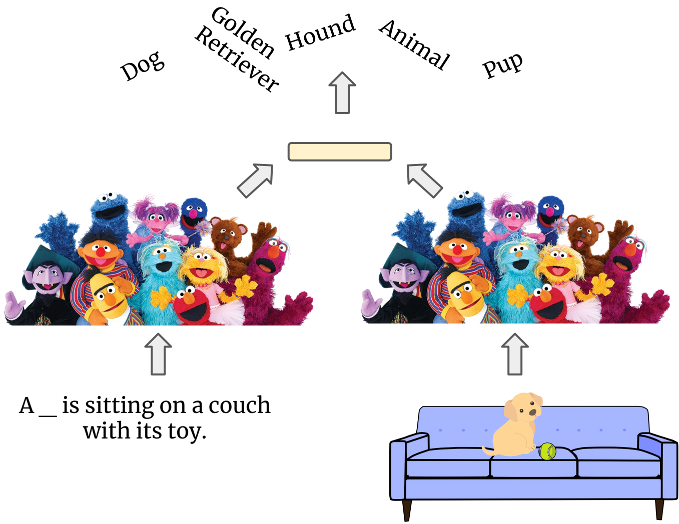
A __ is sitting on a couch with its toy.

Cross-modal architectures

Single-stream architecture

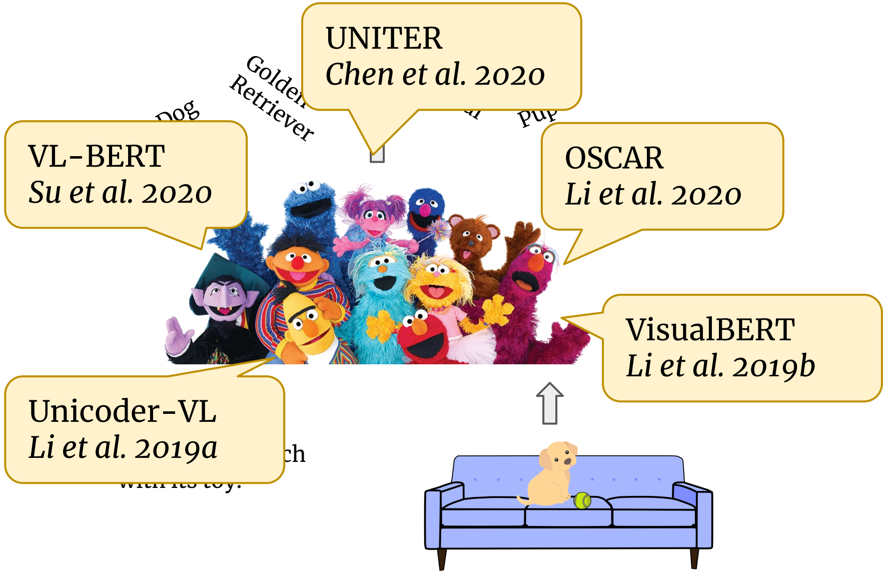


Two-stream architecture



Cross-modal architectures

Single-stream architecture



UNITER
Chen et al. 2020

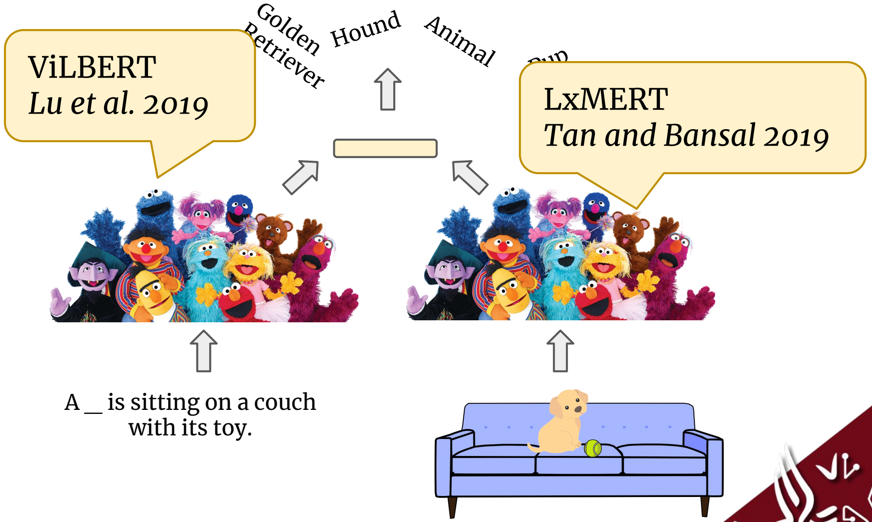
VL-BERT
Su et al. 2020

OSCAR
Li et al. 2020

VisualBERT
Li et al. 2019b

Unicoder-VL
Li et al. 2019a

Two-stream architecture



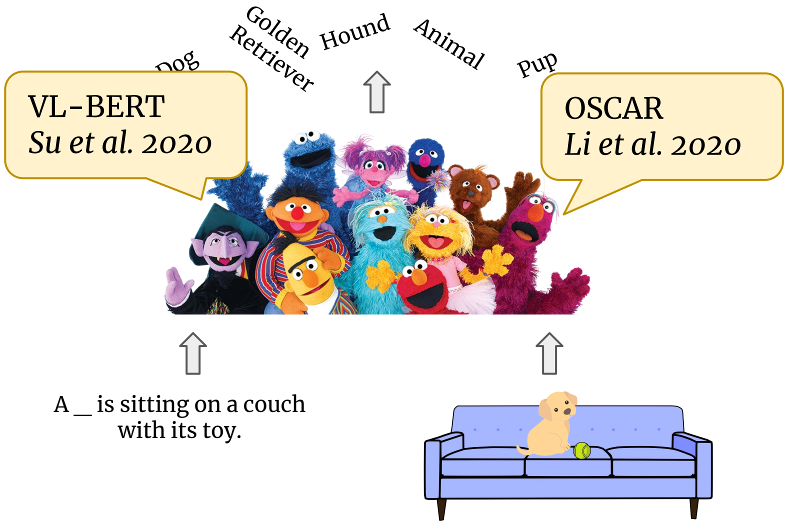
ViLBERT
Lu et al. 2019

LxMERT
Tan and Bansal 2019

A __ is sitting on a couch
with its toy.

Cross-modal architectures

Single-stream architecture

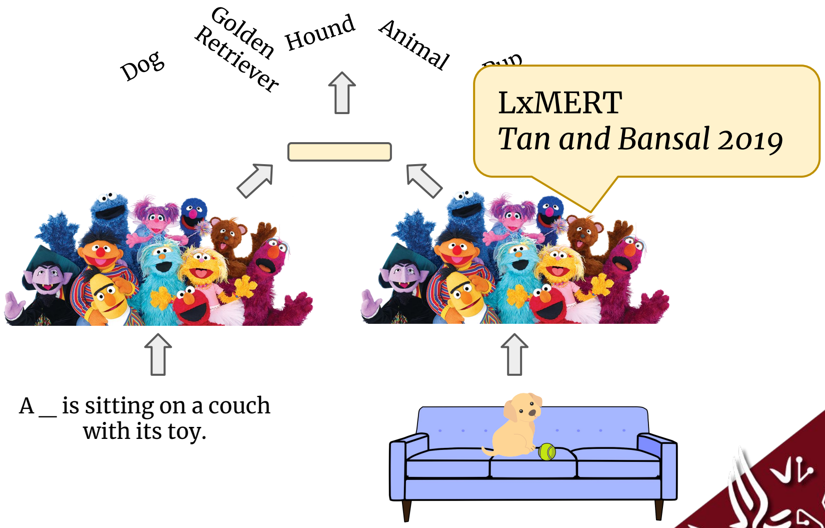


VL-BERT
Su et al. 2020

OSCAR
Li et al. 2020

A __ is sitting on a couch with its toy.

Two-stream architecture



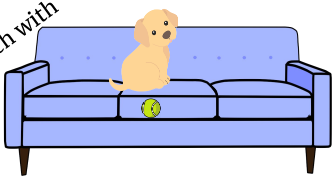
LxMERT
Tan and Bansal 2019

A __ is sitting on a couch with its toy.

VL-BERT *[Su et al. 2020]*

Visual and linguistic contents interact freely

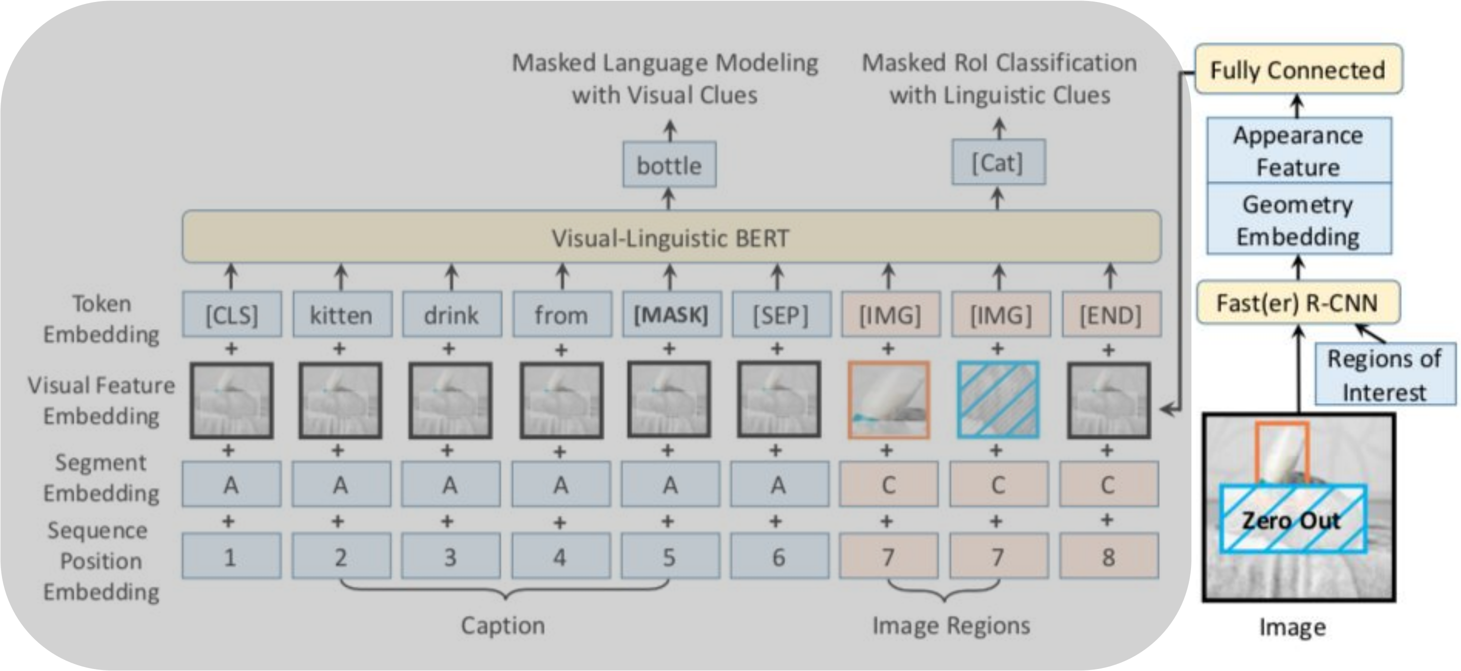
Pretrain on visual-linguistic and text-only data



A _ is sitting on a couch with its toy.

Add new visual features to BERT input embeddings

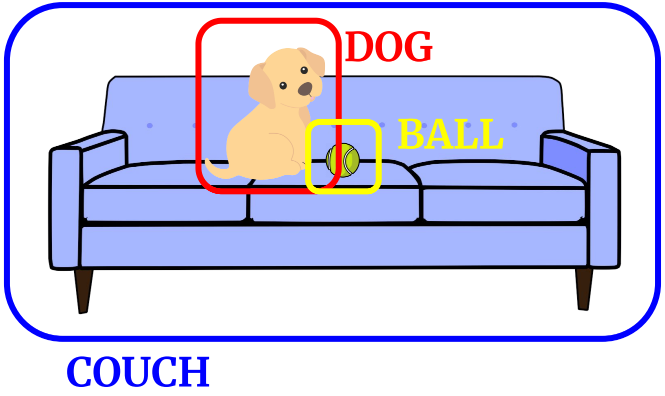
VL-BERT *[Su et al., 2020]*



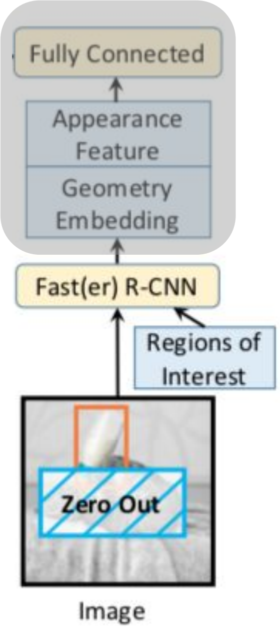
VL-BERT *[Su et al., 2020]*

Fast R-CNN *Girshick, 2015*

⇨ Object detection model



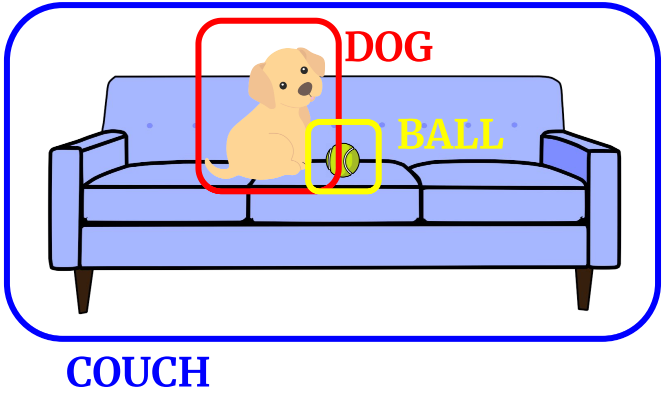
Region of Interests (RoI)



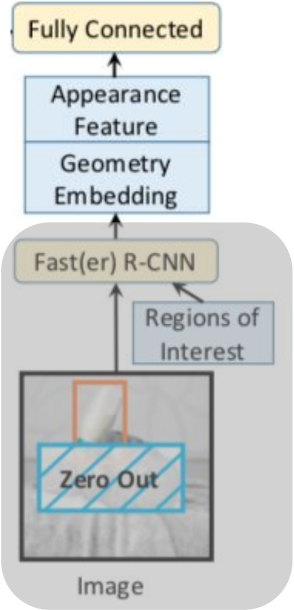
VL-BERT *[Su et al., 2020]*

Fast R-CNN *Girshick, 2015*

⇨ Object detection model



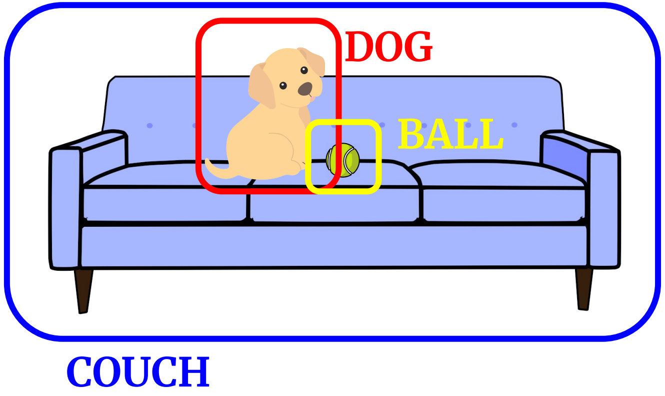
Appearance Feature
Feature vector prior to the output layer of RoI



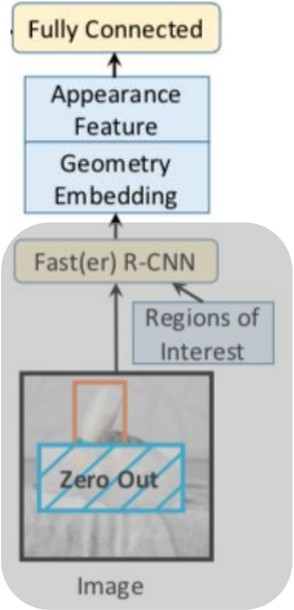
VL-BERT *[Su et al., 2020]*

Fast R-CNN *Girshick, 2015*

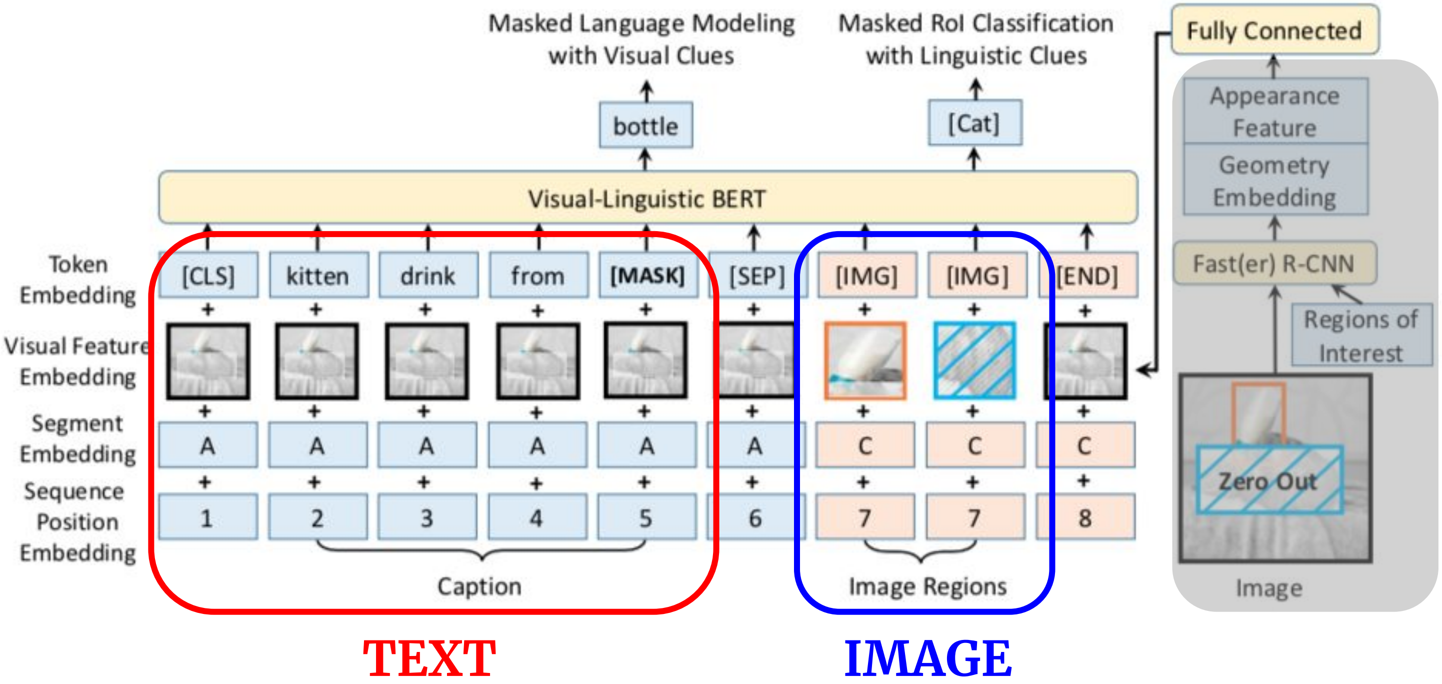
⇨ Object detection model



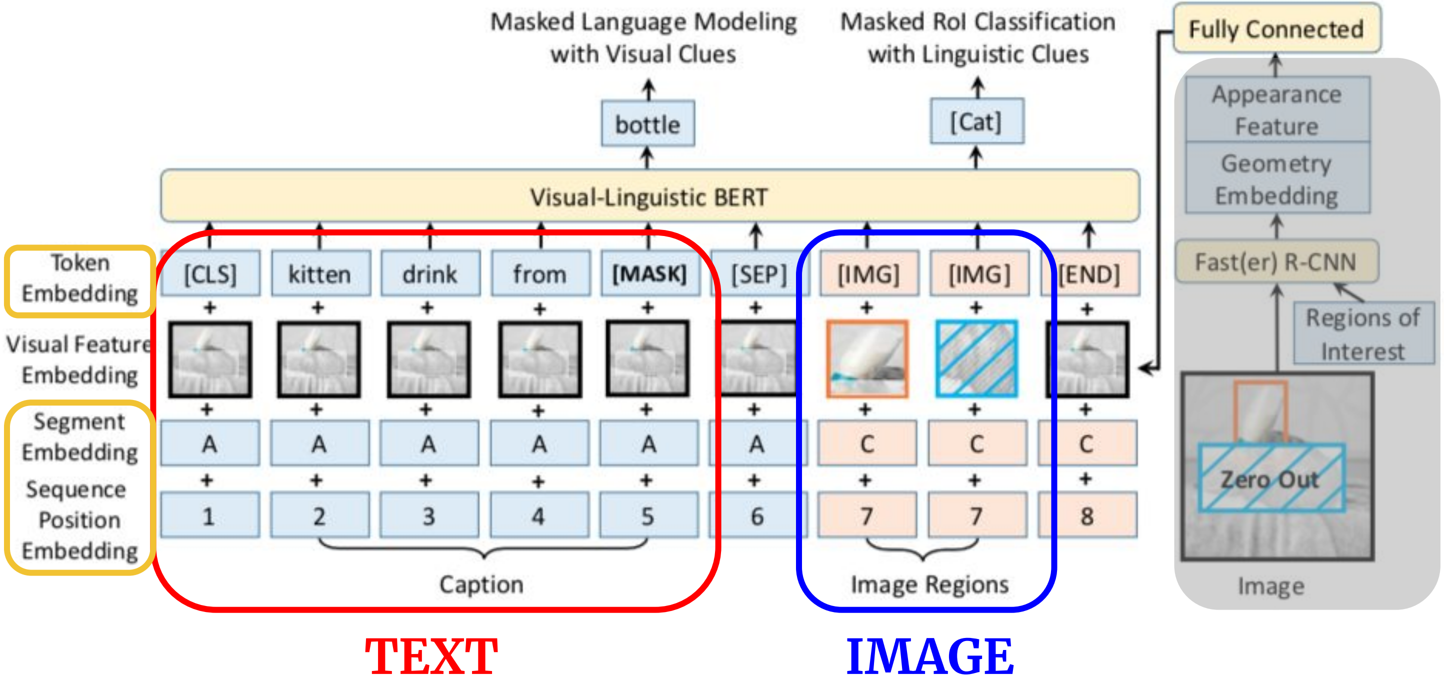
Geometry Embedding
Sine and cosine functions in different wavelengths applied to normalized coordinates of RoI



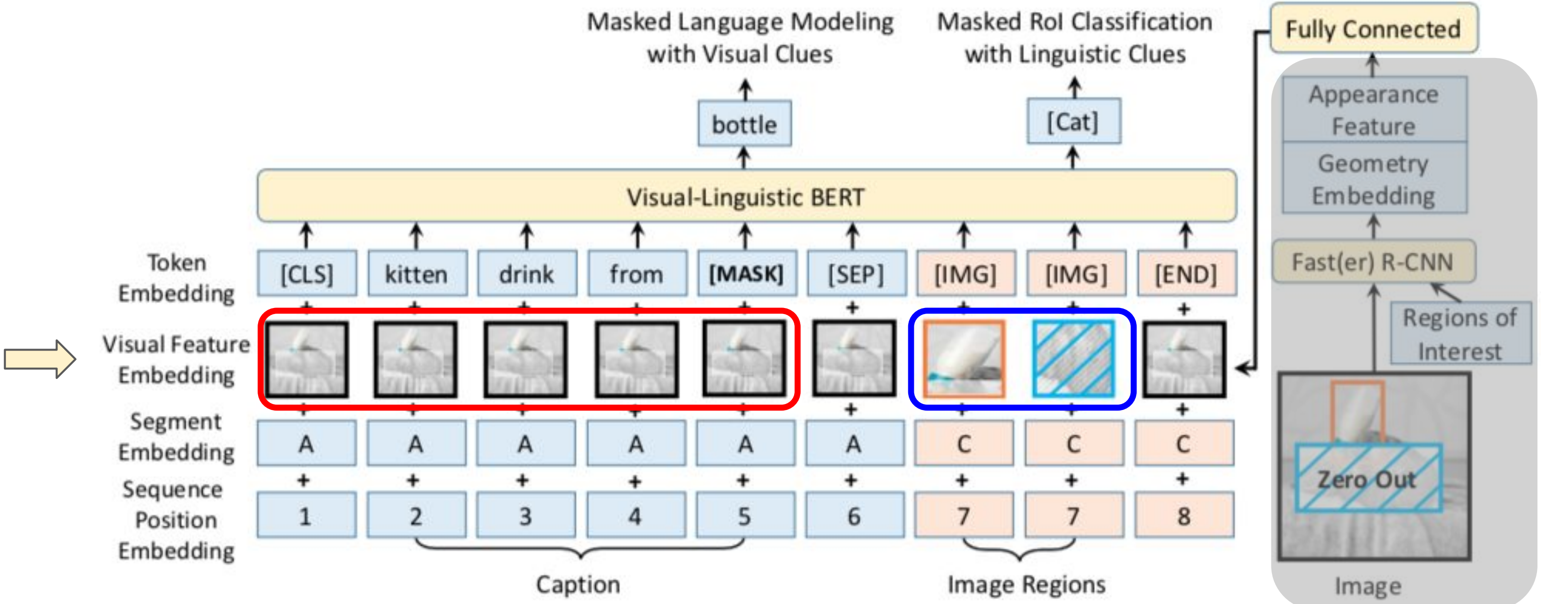
VL-BERT *[Su et al. 2020]*



VL-BERT *[Su et al. 2020]*



VL-BERT *[Su et al. 2020]*



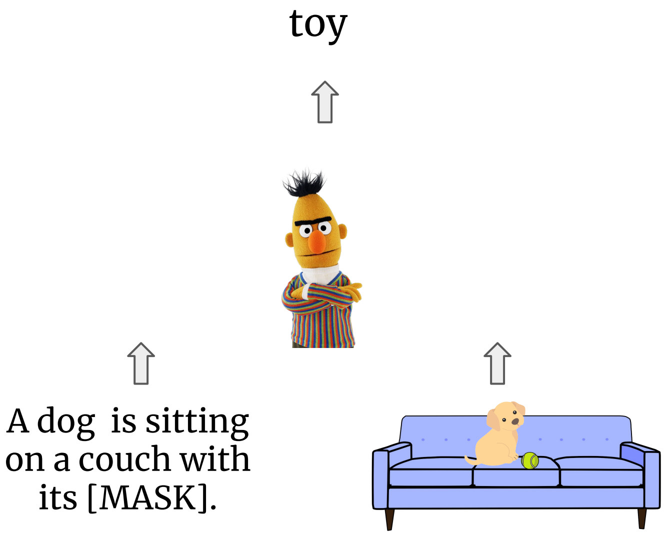
Whole Image

One per RoI

VL-BERT *[Su et al. 2020]*

Conceptual Captions
Sharma et al., 2018

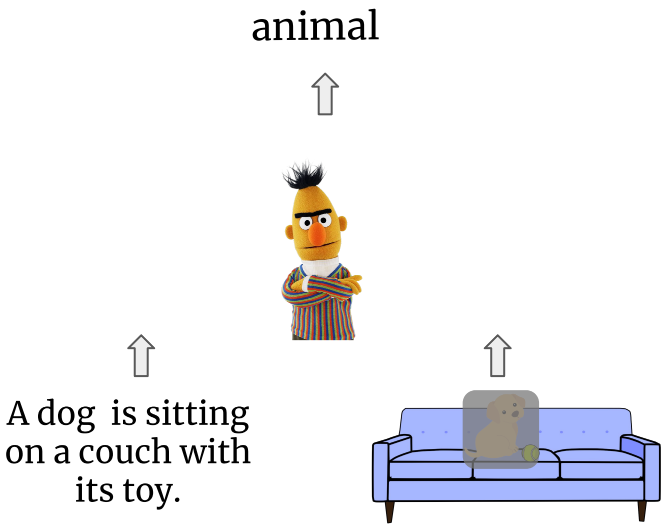
Masked Language Modeling with Visual Clues



toy

A dog is sitting on a couch with its [MASK].

Masked RoI Classification with Linguistic Clues



animal

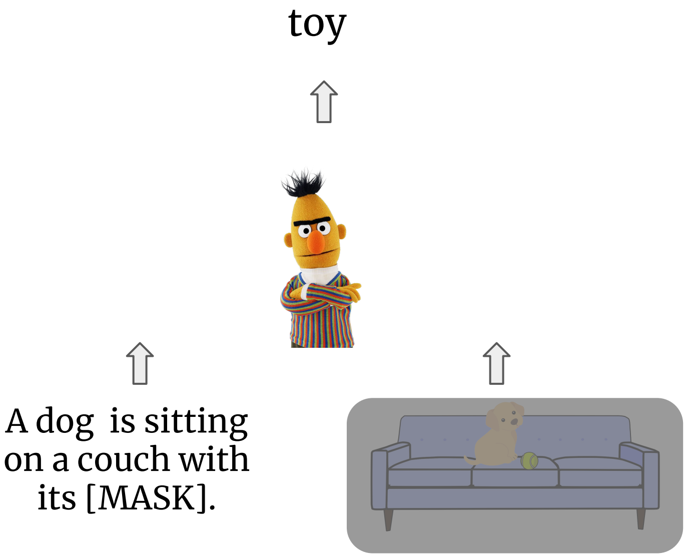
A dog is sitting on a couch with its toy.

VL-BERT *[Su et al. 2020]*

Book Captions
Zhu et al., 2015

English Wikipedia

Masked Language Modeling



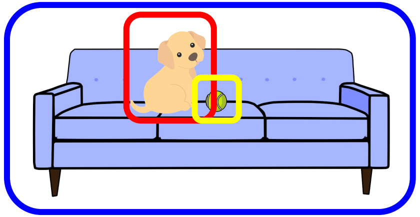
toy

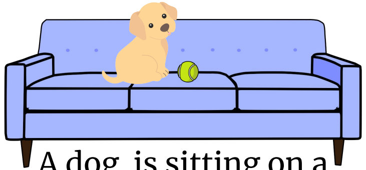
A dog is sitting on a couch with its [MASK].

OSCAR *[Li et al., 2020]*

Single-stream architecture

Uses object tags in an image as anchor points



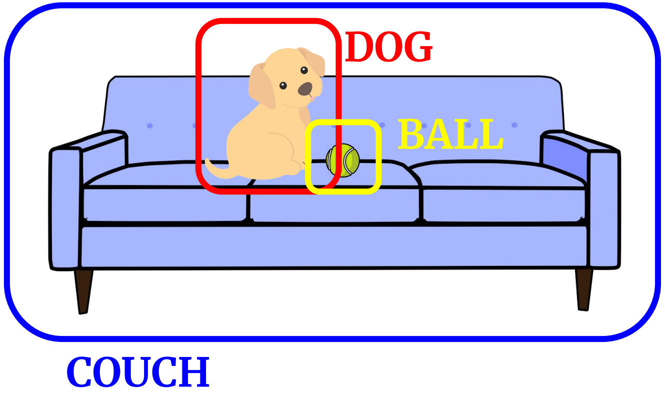


A dog is sitting on a couch with its toy.

Ease the learning of image-text alignment

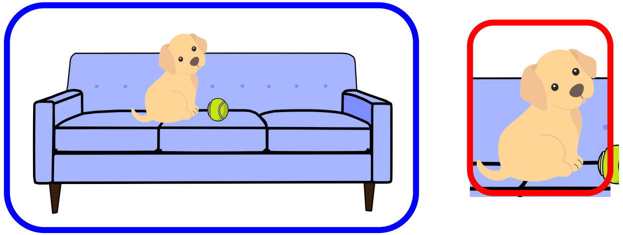
OSCAR *[Li et al., 2020]*

A **dog** is sitting on a **couch**.



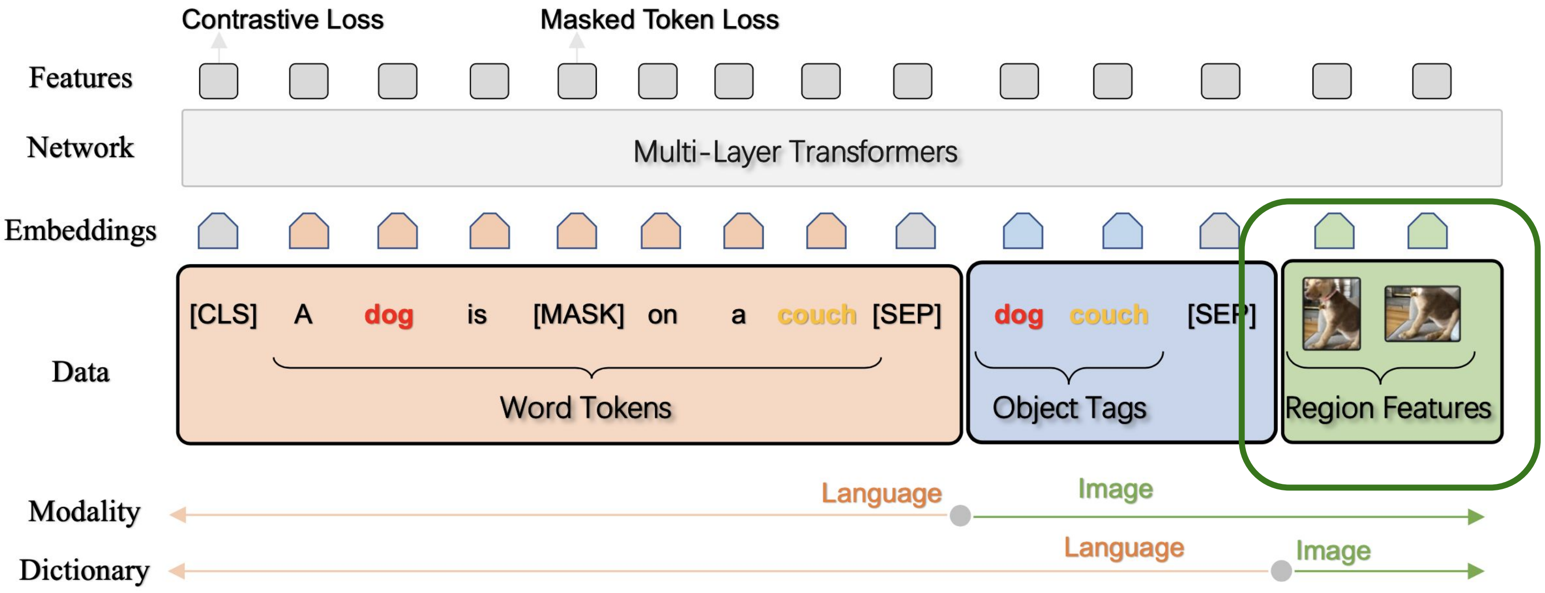
A dog is sitting on a couch.

dog
couch



OSCAR *[Li et al., 2020]*

Faster R-CNN, *Ren et al., 2015*

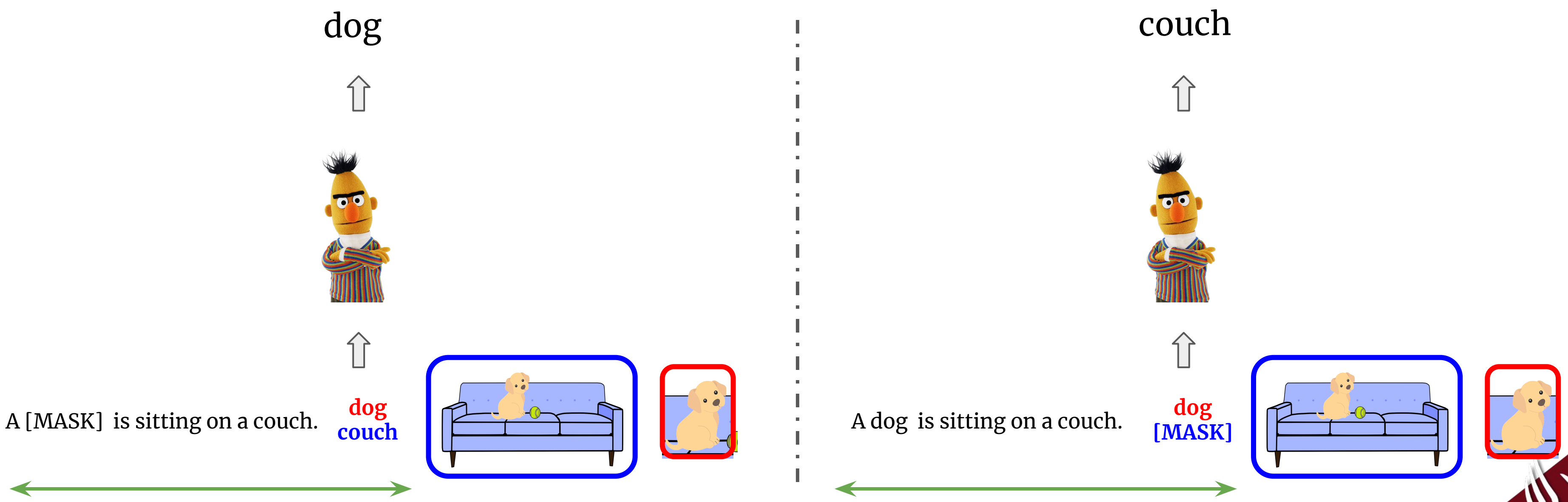


OSCAR *[Li et al., 2020]*

Conceptual Captions
Sharma et al., 2018

COCO
Lin et al., 2014
etc.

Masked Token Loss



dog

A [MASK] is sitting on a couch.

dog
couch

couch

A dog is sitting on a couch.

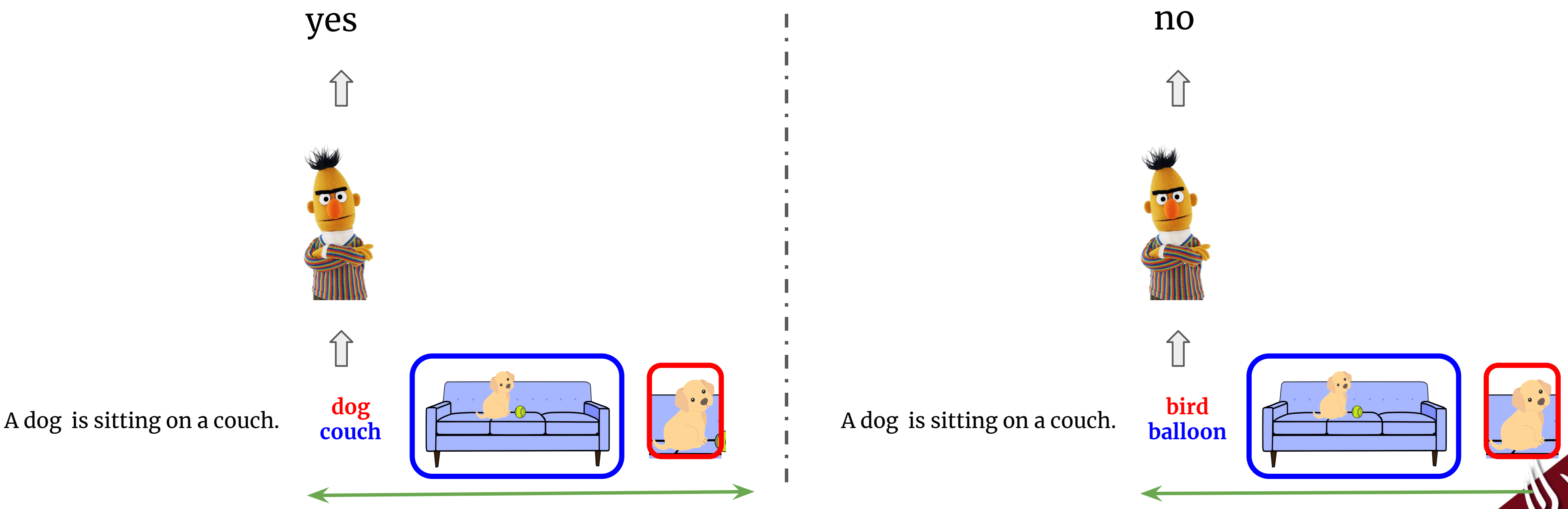
dog
[MASK]

OSCAR *[Li et al., 2020]*

Conceptual Captions
Sharma et al., 2018

COCO
Lin et al., 2014
etc.

Contrastive loss



yes

A dog is sitting on a couch.

dog
couch

no

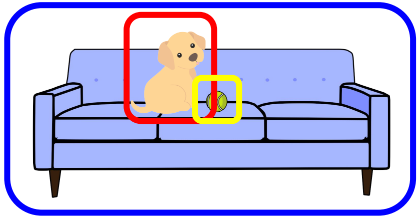
A dog is sitting on a couch.

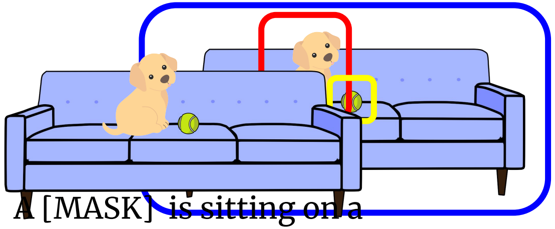
bird
balloon

LxMERT *[Tan and Bansal, 2020]*

Double-stream architecture

Builds both intra-modality and cross-modality relations



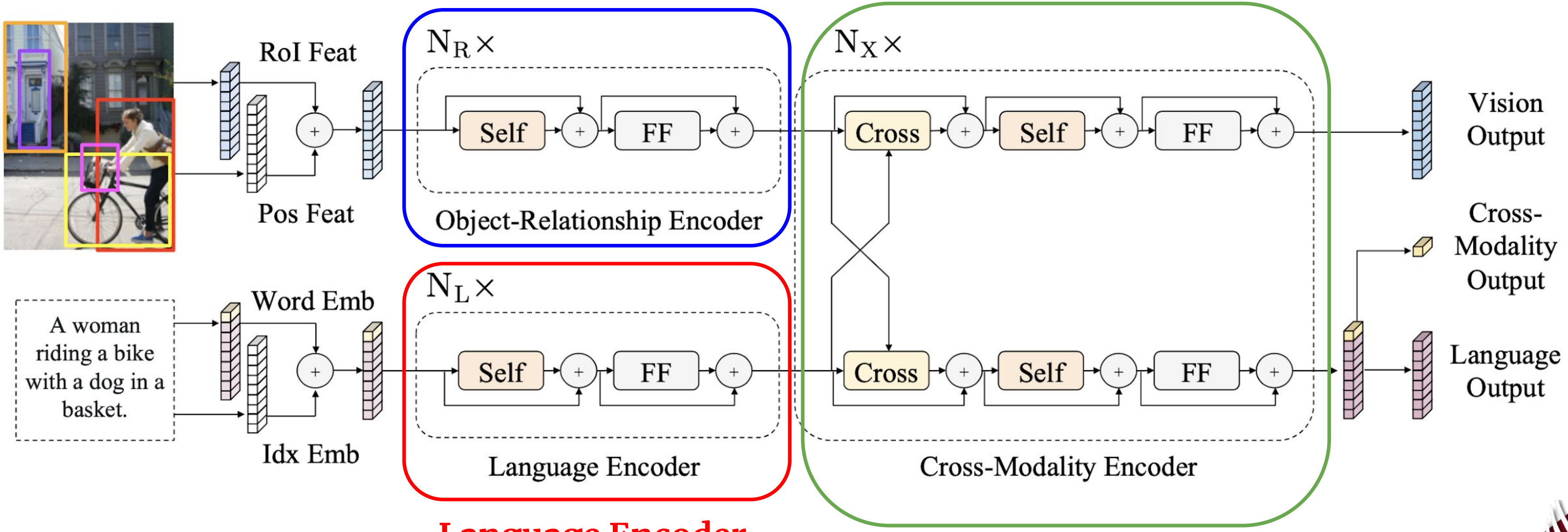


A [MASK] is sitting on a couch with its toy.

Five diverse pre-training tasks

LxMERT *[Tan and Bansal, 2020]*

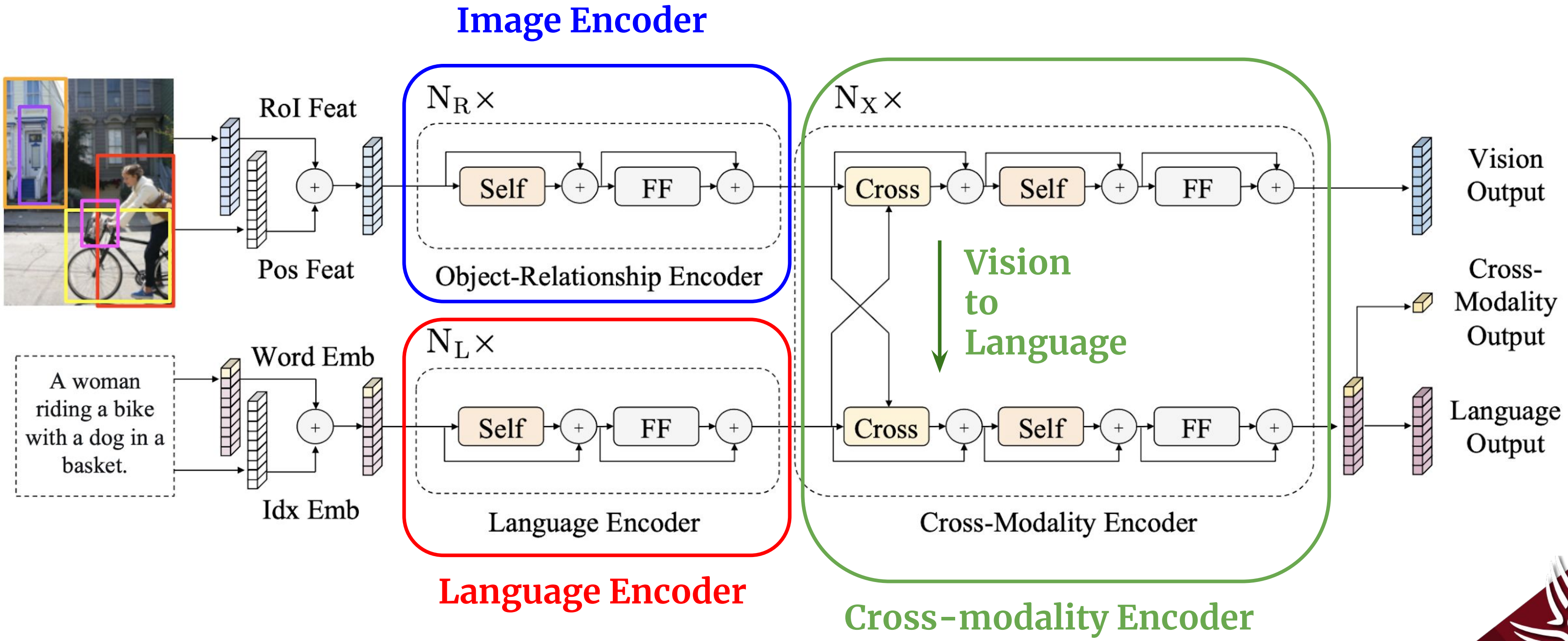
Image Encoder



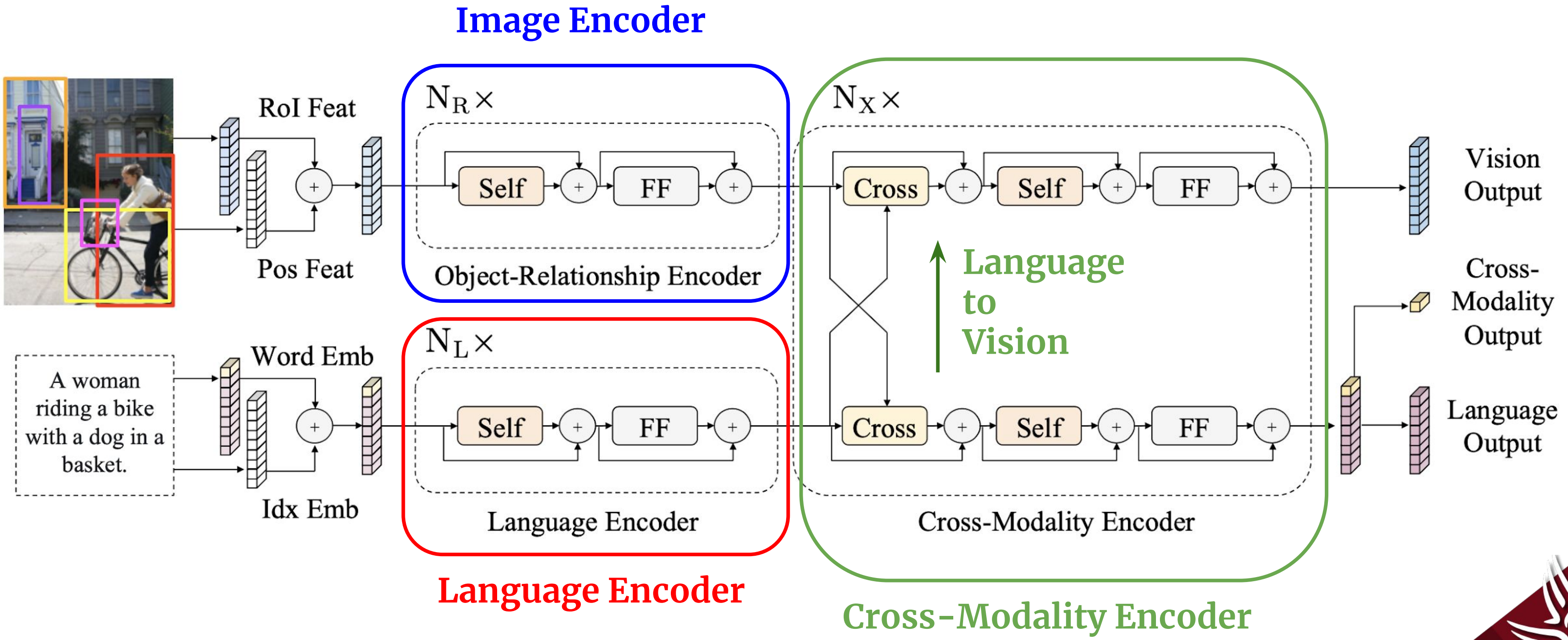
Language Encoder

Cross-modality Encoder

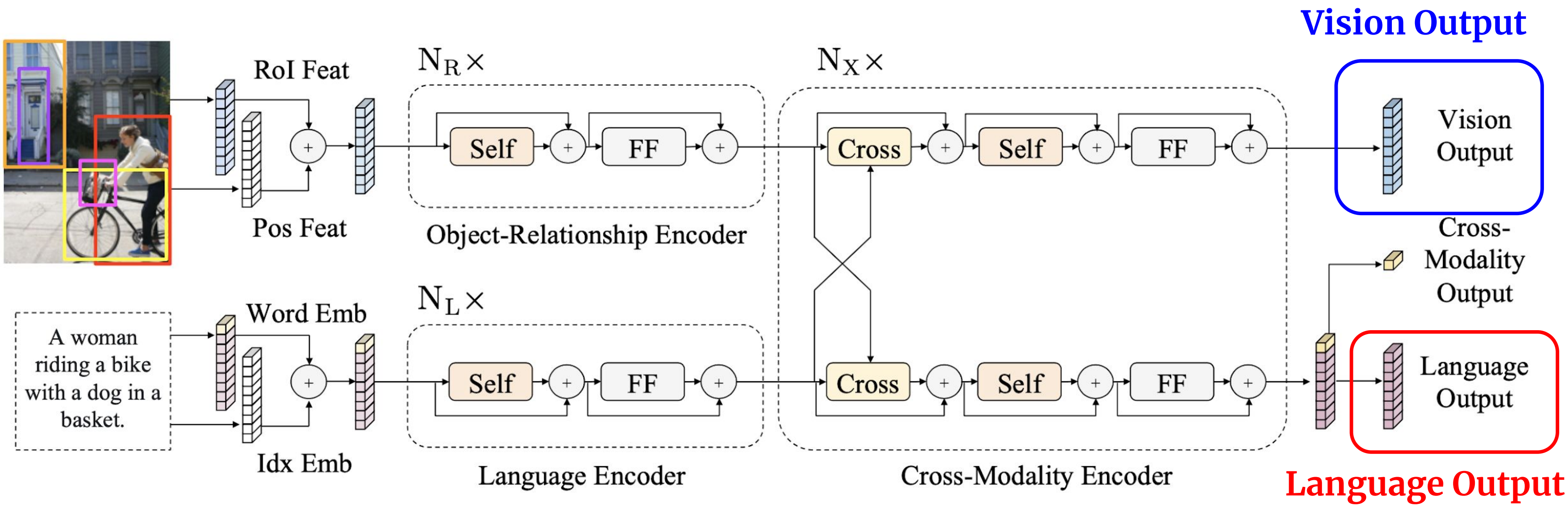
LxMERT *[Tan and Bansal, 2020]*



LxMERT *[Tan and Bansal, 2020]*

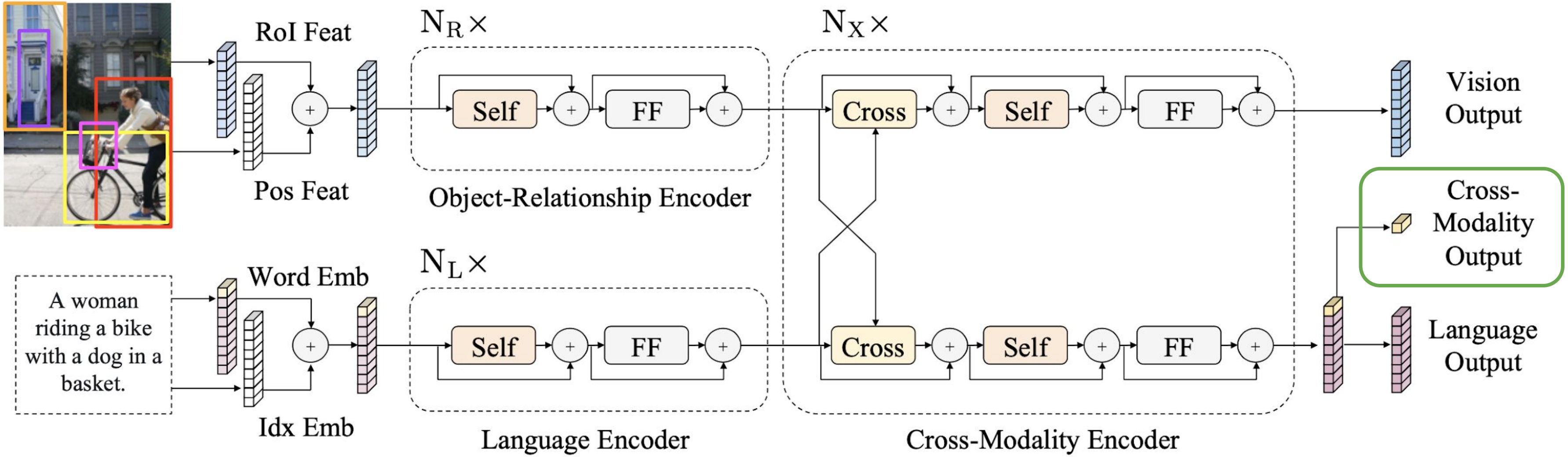


LxMERT *[Tan and Bansal, 2020]*



LxMERT *[Tan and Bansal, 2020]*

Cross-Modality Output
[CLS]



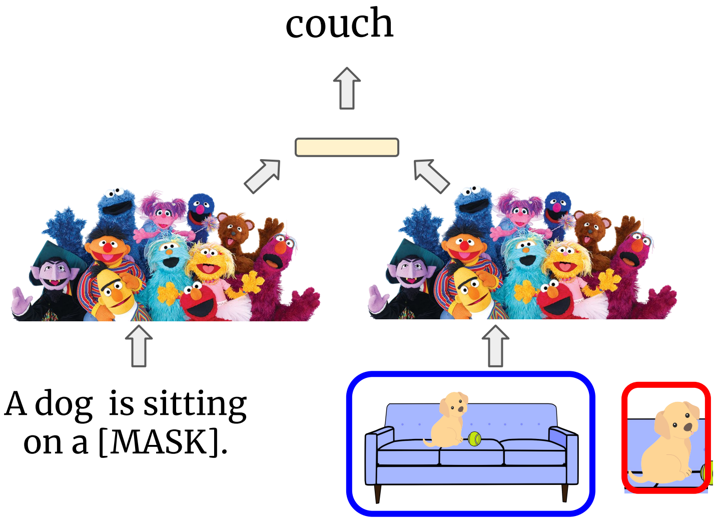
LxMERT *[Tan and Bansal, 2020]*

MS COCO
Lin et al., 2014

Visual Genome
Krishna et al., 2017

etc.

Masked Language Modeling with Visual Clues



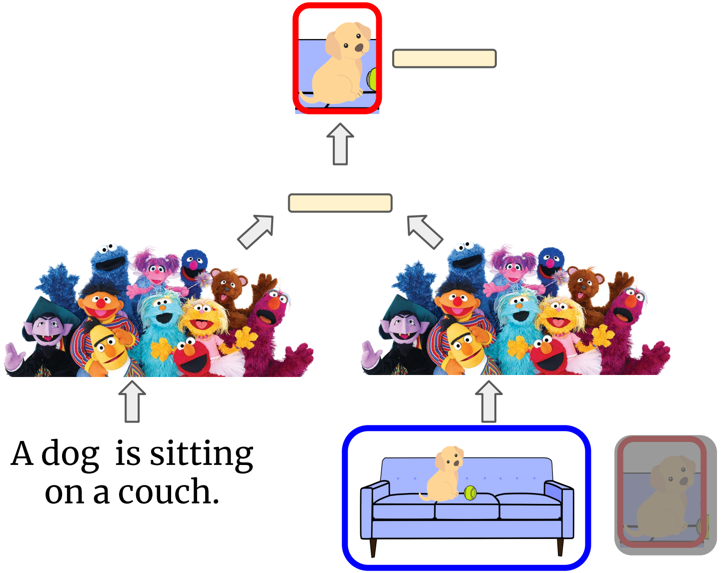
LxMERT *[Tan and Bansal, 2020]*

MS COCO
Lin et al., 2014

Visual Genome
Krishna et al., 2017

etc.

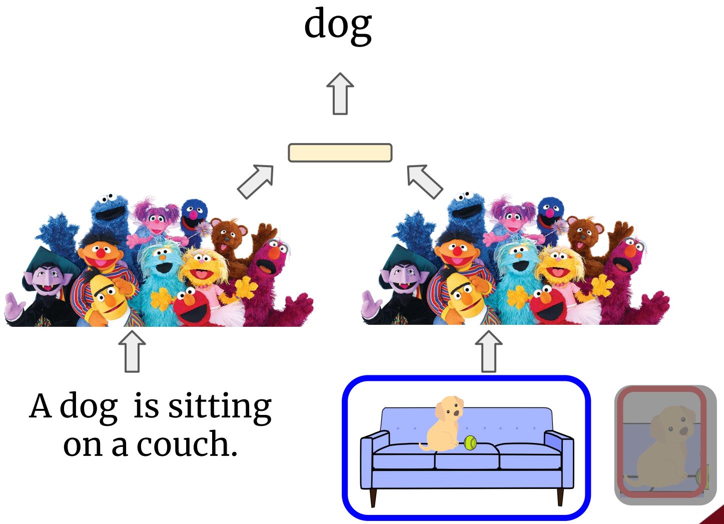
RoI Feature Regression



A dog is sitting on a couch.

Detected-Label Classification

dog



A dog is sitting on a couch.

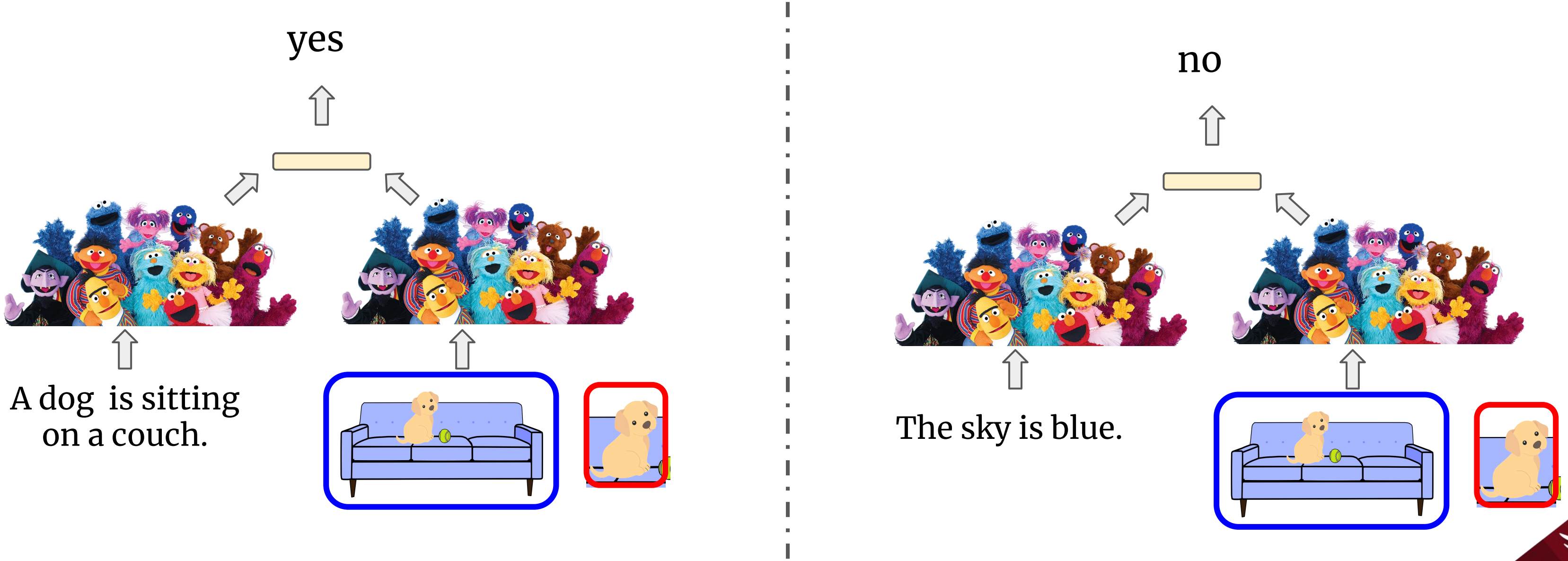
LxMERT *[Tan and Bansal, 2020]*

MS COCO
Lin et al., 2014

Visual Genome
Krishna et al., 2017

etc.

Cross-Modality Matching



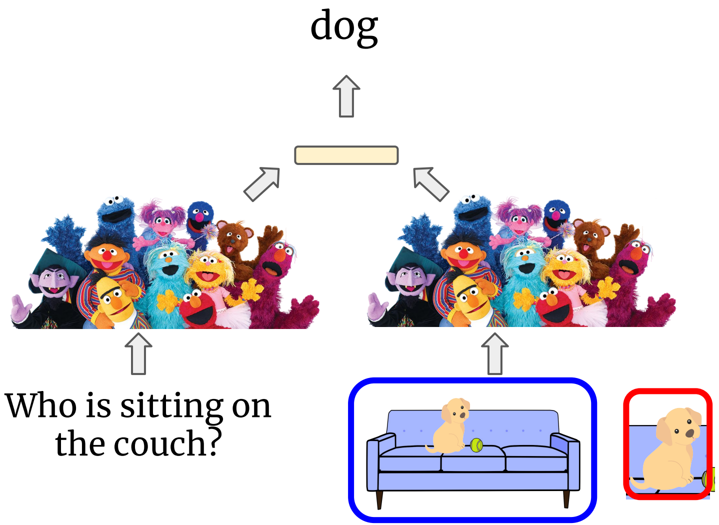
LxMERT *[Tan and Bansal, 2020]*

MS COCO
Lin et al., 2014

Visual Genome
Krishna et al., 2017

etc.

Image Question Answering



Experimental Setup

Image Text Retrieval

Visual Question Answering

Visual Commonsense Reasoning

Grounding Referring Expression

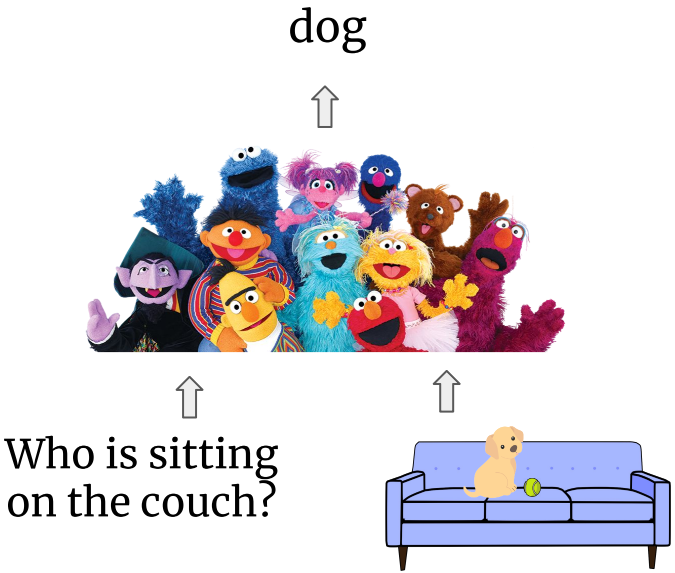
Image Captioning

Natural Language Visual Reasoning for Real

... and many more

Visual Question Answering

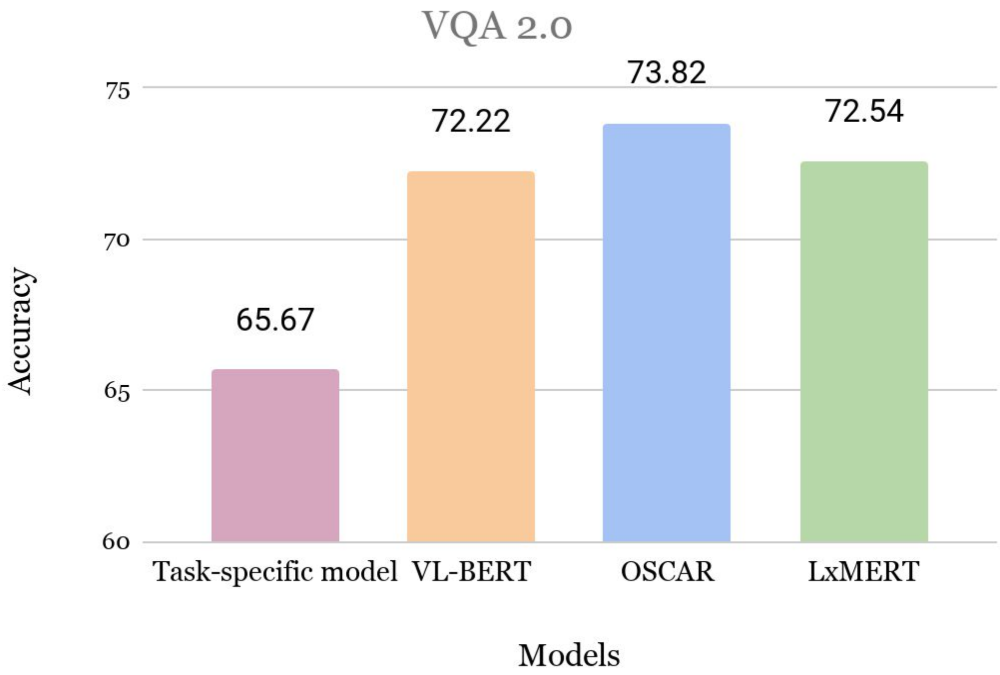
VQA v2.0
Goyal et al., 2017



dog

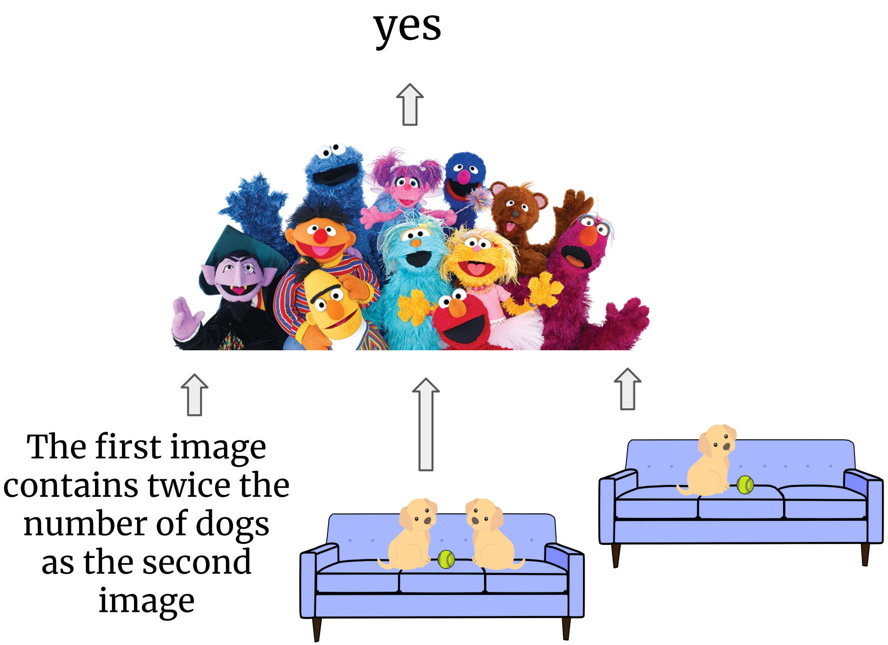
Who is sitting on the couch?

Visual Question Answering



Natural Language Visual Reasoning for Real

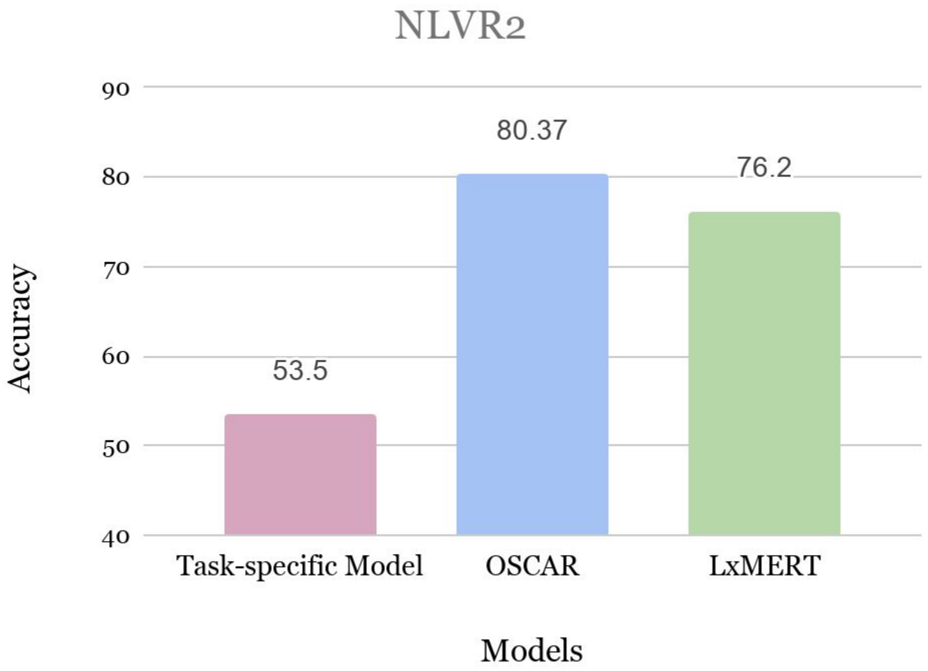
NLVR^2
Suhr et al., 2019



yes

The first image contains twice the number of dogs as the second image

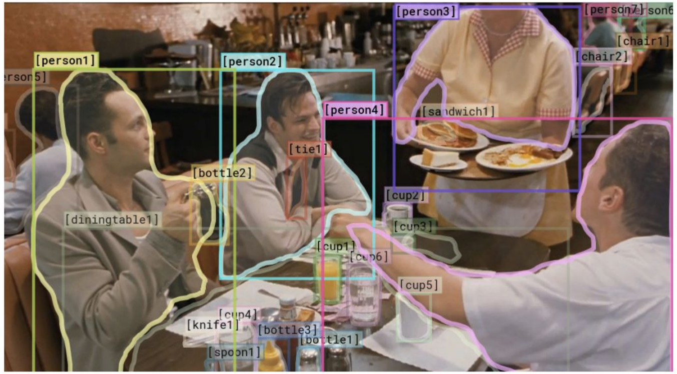
Natural Language Visual Reasoning for Real



Visual Commonsense Reasoning

VCR
Zellers et al., 2019

Image



Questions

Why is [person4] pointing at [person1]?

Answers

a) He is telling [person3] that [person1] ordered the pancakes.
b) He just told a joke.
c) He is feeling accusatory towards [person1].
d) He is giving [person1] directions.

I chose **a)** because...

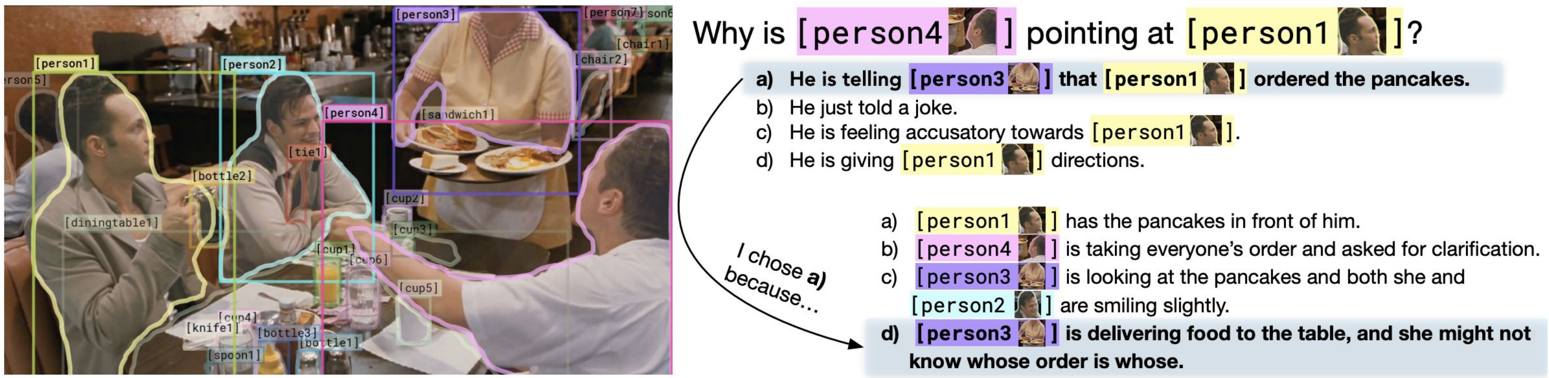
Rationales

a) [person1] has the pancakes in front of him.
b) [person4] is taking everyone's order and asked for clarification.
c) [person3] is looking at the pancakes and both she and [person2] are smiling slightly.
d) [person3] is delivering food to the table, and she might not know whose order is whose.

Visual Commonsense Reasoning

VCR
Zellers et al., 2019

⇨ Given the image and the question, return the correct answer
Q → A



Why is [person4] pointing at [person1]?

a) He is telling [person3] that [person1] ordered the pancakes.
b) He just told a joke.
c) He is feeling accusatory towards [person1].
d) He is giving [person1] directions.

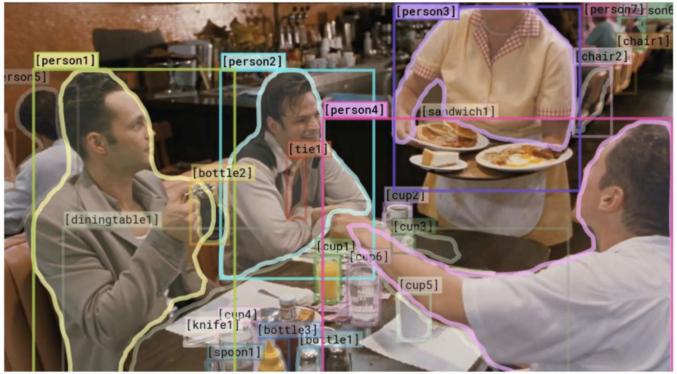
I chose **a)** because...

a) [person1] has the pancakes in front of him.
b) [person4] is taking everyone's order and asked for clarification.
c) [person3] is looking at the pancakes and both she and [person2] are smiling slightly.
d) [person3] is delivering food to the table, and she might not know whose order is whose.

Visual Commonsense Reasoning

VCR
Zellers et al., 2019

⇨ Given the image, the question and the answer return the correct rationale
QA → R



Why is [person4] pointing at [person1]?

a) He is telling [person3] that [person1] ordered the pancakes.
b) He just told a joke.
c) He is feeling accusatory towards [person1].
d) He is giving [person1] directions.

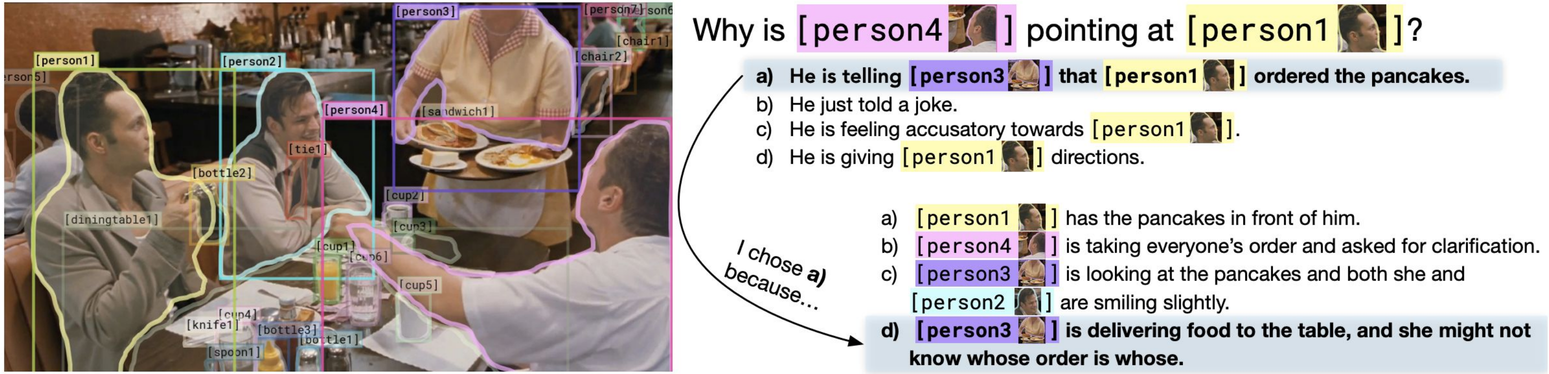
I chose **a)** because...

a) [person1] has the pancakes in front of him.
b) [person4] is taking everyone's order and asked for clarification.
c) [person3] is looking at the pancakes and both she and [person2] are smiling slightly.
d) [person3] is delivering food to the table, and she might not know whose order is whose.

Visual Commonsense Reasoning

VCR
Zellers et al., 2019

➡ Given the image and the question return the correct answer and rationale
Q → AR



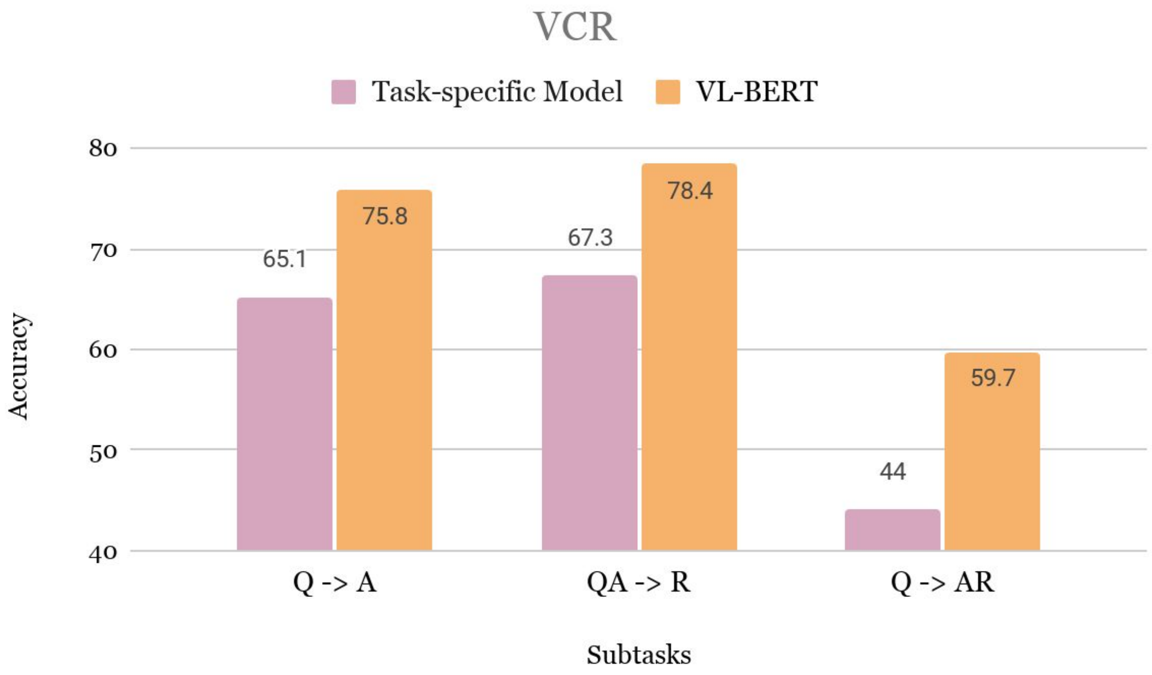
Why is [person4] pointing at [person1]?

a) He is telling [person3] that [person1] ordered the pancakes.
b) He just told a joke.
c) He is feeling accusatory towards [person1].
d) He is giving [person1] directions.

I chose **a)** because...

a) [person1] has the pancakes in front of him.
b) [person4] is taking everyone's order and asked for clarification.
c) [person3] is looking at the pancakes and both she and [person2] are smiling slightly.
d) [person3] is delivering food to the table, and she might not know whose order is whose.

Visual Commonsense Reasoning



Conclusions

☆ Vision-and-language models gained much interest in the last couple of years

Straightforward techniques to incorporate visual features in contextualized language models ☆

☆ Vision-and-language models raised the bar for the state-of-the-art in many vision-and language tasks

Many directions that are still worth to be explored!

Thanks for your attention!

Any questions? Feel free to ask