Multimodal Learning

Danna Gurari
University of Colorado Boulder
Spring 2024

Review

- Last lecture
  - Motivation
  - Foundation Models
  - NLP: Textual Prompting
  - CV: Visual Prompting (In-context Few-shot Learning)
  - Parameter-Efficient Tuning Methods
  - Latex Tutorial
- Assignments (Canvas)
  - Problem set 4 due earlier today
  - Final project outline due in 3 weeks
- Questions?

Today’s Topics

- Multimodal applications
- Image captioning dataset challenges
- Image captioning algorithms
- Visual question answering dataset challenges
- Visual question answering algorithms
- Foundation models

Today’s Topics

- Multimodal applications
- Image captioning dataset challenges
- Image captioning algorithms
- Visual question answering dataset challenges
- Visual question answering algorithms
- Foundation models

Simultaneously Use 2+ Modalities

To date, most work focuses on the intersection of CV + NLP; e.g.,

Caption:
A bunch of small light brown mushrooms in a green field.

Answer Visual Question:
Q: Is it edible or poisonous?

A: Poisonous

Visual Assistance for People with Vision Loss; e.g.,

Seeing AI

Lookout by Google

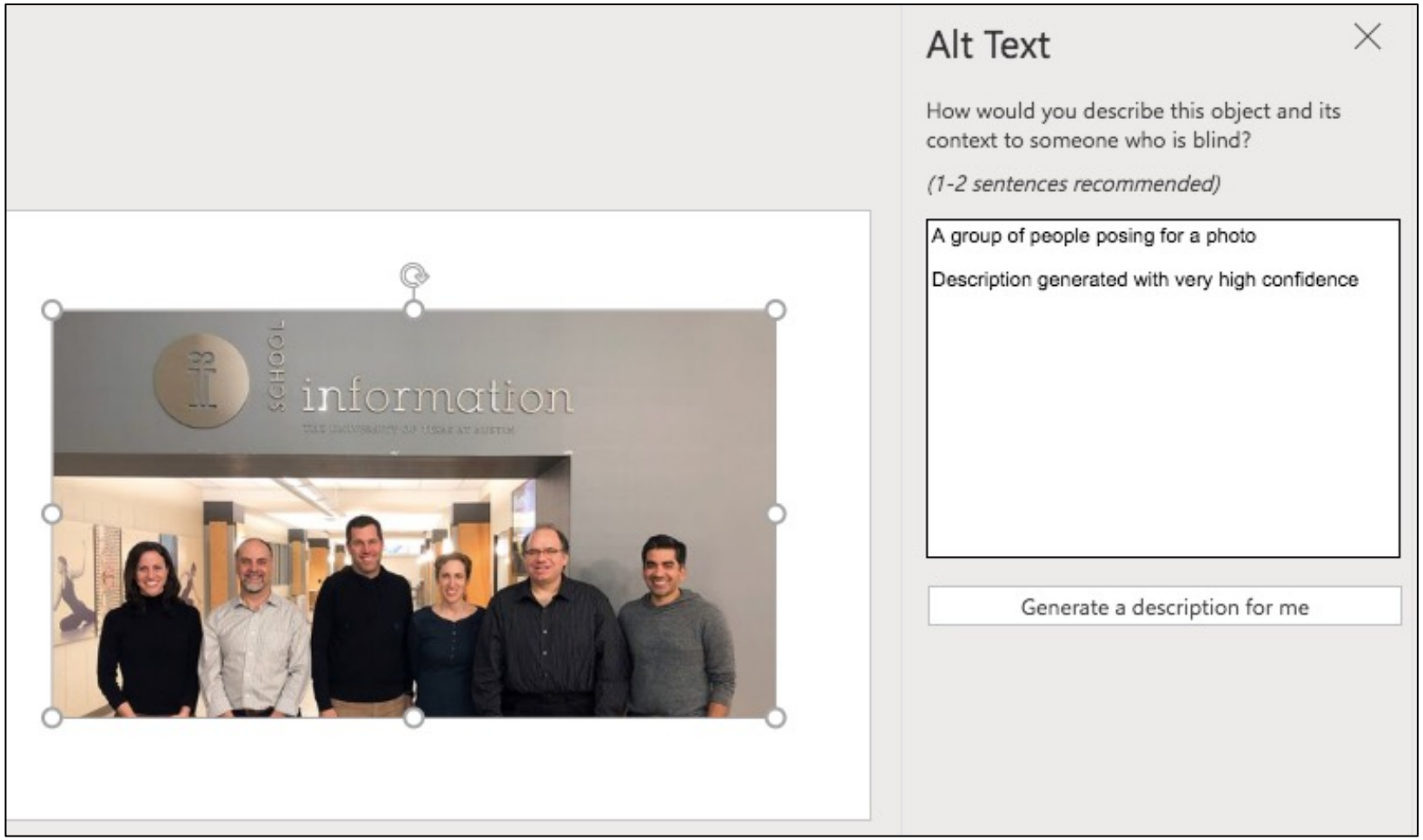
be my eyes

TapTapSee

ORCAM

envision

Visual Assistance for People with Vision Loss; e.g.,



e.g., Microsoft Power Point (Office 365 demo)

Visual Assistance for People with Vision Loss

https://www.youtube.com/watch?v=cUSeFnZGIzY

Describing and Responding to Images Posted to Social Media with “Personality”

Standard captioning output: A plate with a sandwich and salad on it.

Our model with different personality traits (215 possible traits, not all shown here):

Sweet	That is a lovely sandwich.
Dramatic	This sandwich looks so delicious! My goodness!
Anxious	I’m afraid this might make me sick if I eat it.
Sympathetic	I feel so bad for that carrot, about to be consumed.
Arrogant	I make better food than this
Optimistic	It will taste positively wonderful!
Money-minded	I would totally pay $100 for this plate.

Shuster et al. Engaging Image Captioning via Personality. 2019

Describing Products

Title: Stand Collar A-Line Dress

Fashion Caption: A pearly button accents the stand collar that gives this so-simple, yet so-chic A-line dress its retro flair

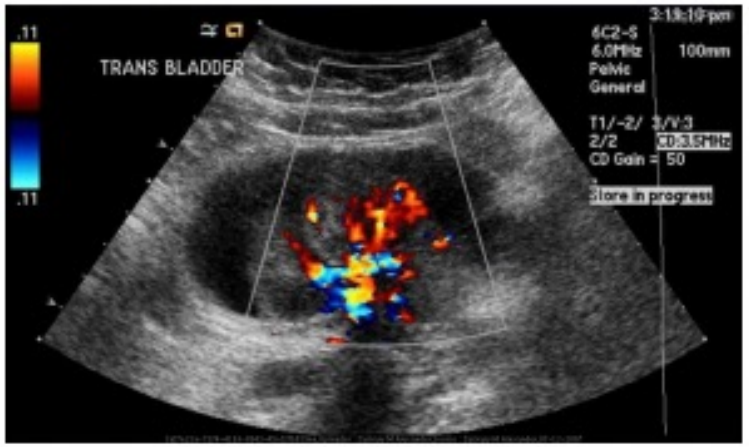
Color: Black and ivory

Meta: - 33" petite length (size 8P) - Hidden back-zip closure - Stand collar - Cap sleeves - Side-seam pockets – A-Lined - 63% polyester, 34% rayon, 3% spandex - Dry clean or hand wash, dry flat - Imported – Dress

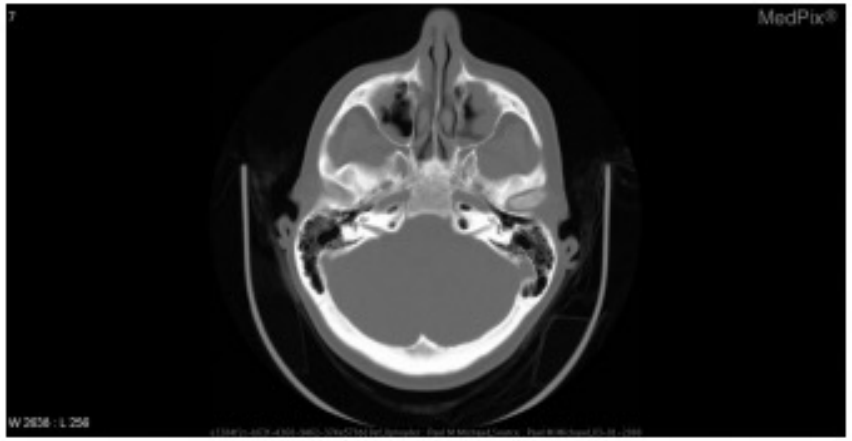
Image Caption: A person in a dress

Yang et al. Fashion Captioning: Towards Generating Accurate Descriptions with Semantic Rewards. ECCV 2020

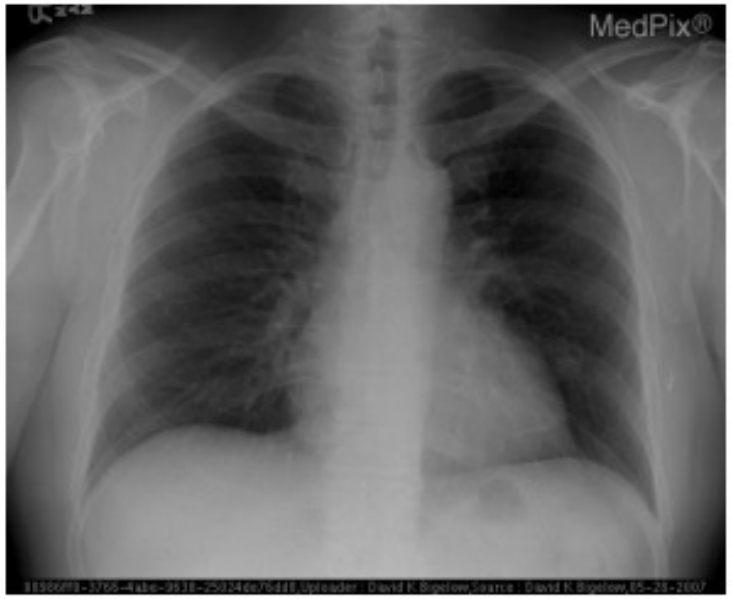
Medical VQA



(a) **Q**: what imaging method was used? **A**: us-d - doppler ultrasound



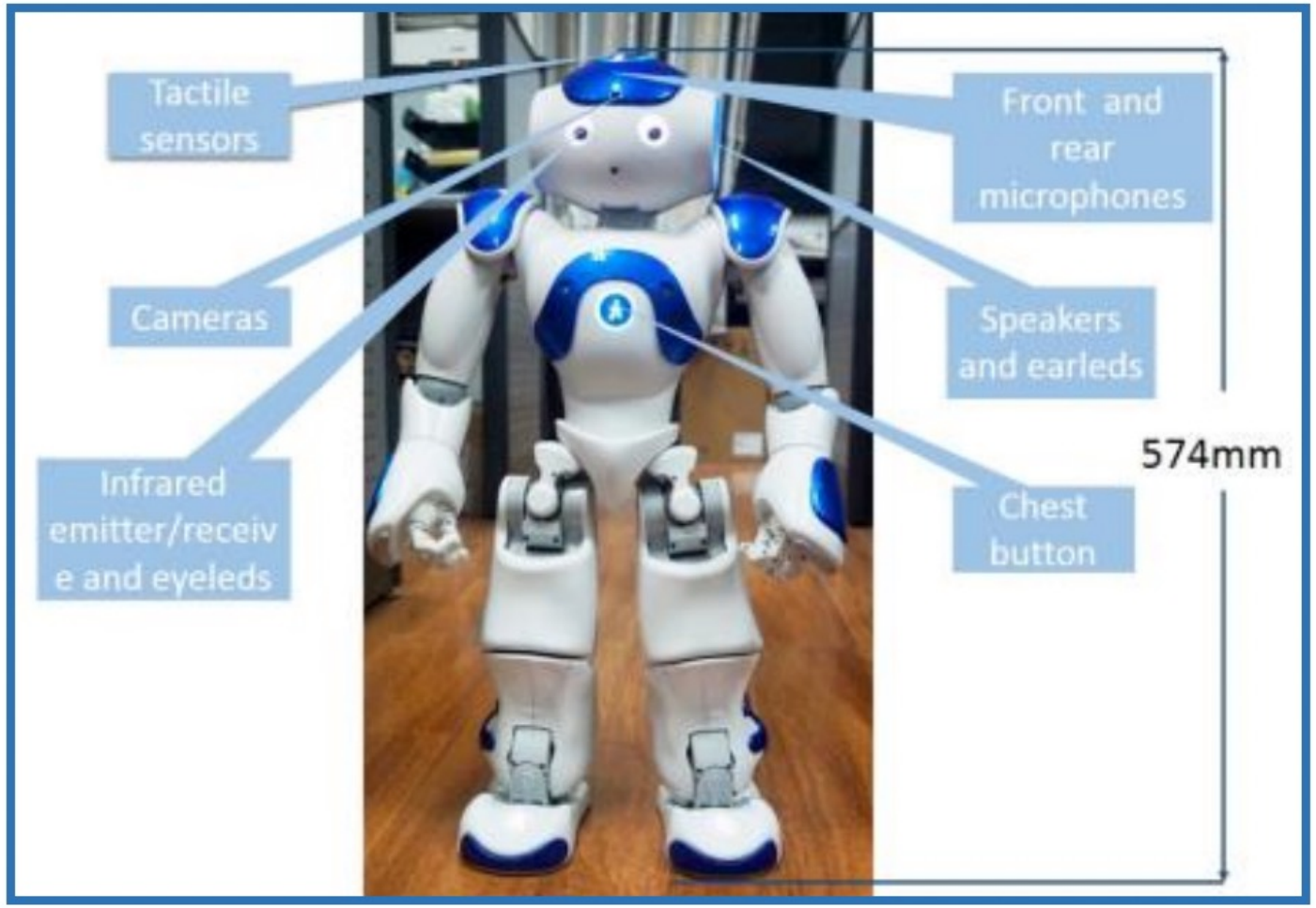
(b) **Q**: which plane is the image shown in? **A**: axial



(e) **Q**: what abnormality is seen in the image? **A**:nodular opacity on the left#metastastic melanoma

Abacha et al. VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019

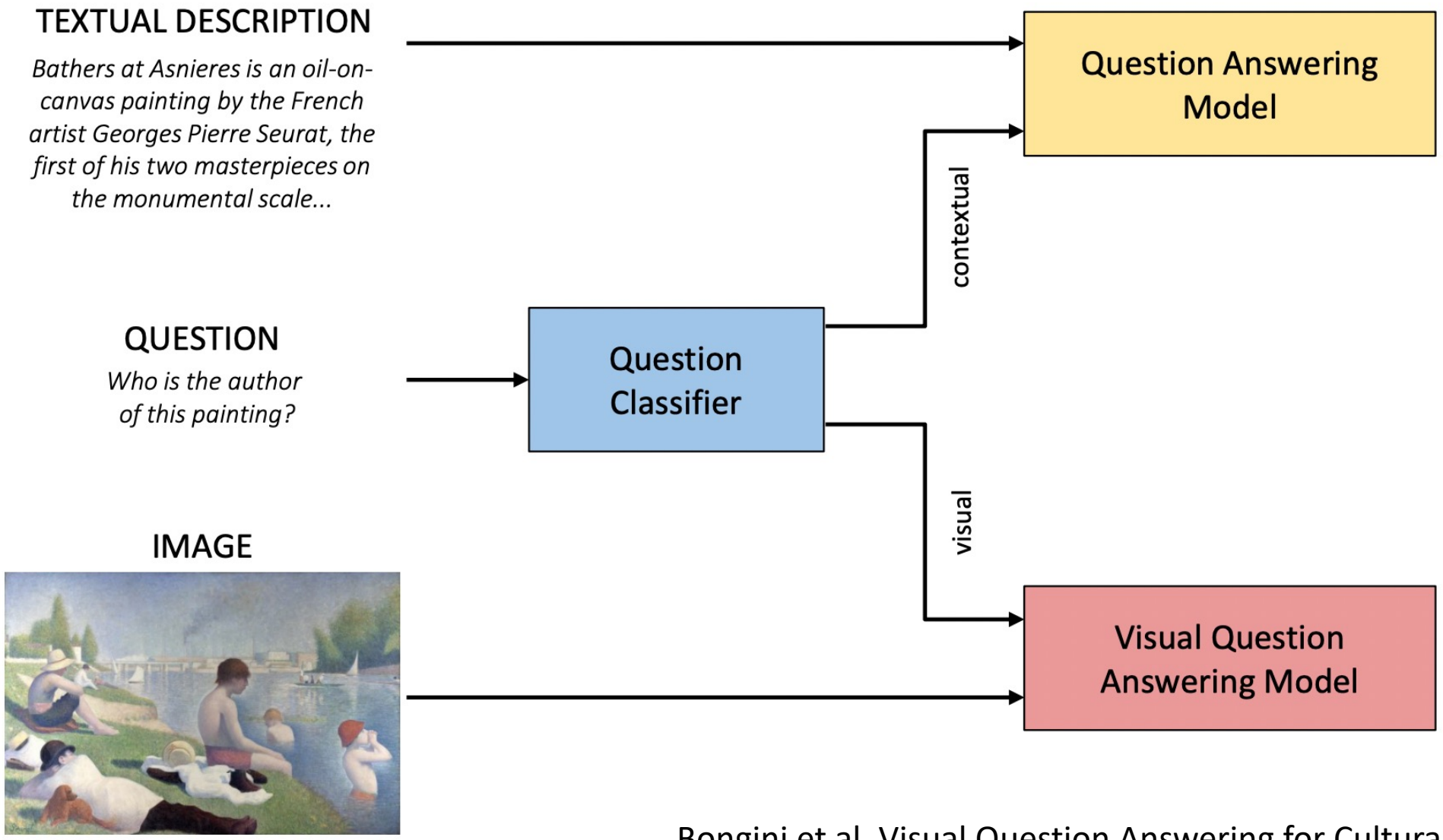
Education (e.g., for Preschoolers)



Answers questions about **quantity** and **colors** of detected objects

He et al. An Educational Robot System of Visual Question Answering for Preschoolers. 2017.

Audio Guide for Museums and Art Galleries



Bongini et al. Visual Question Answering for Cultural Heritage. 2020.

Today’s Topics

- Multimodal applications
- Image captioning dataset challenges
- Image captioning algorithms
- Visual question answering dataset challenges
- Visual question answering algorithms
- Foundation models

Sample of Existing Dataset Challenges

COCO

Woman on a horse jumping over a pole jump.

A glass bowl contains peeled tangerines and cut strawberries.

VizWiz

A person is holding a small container of cream upside down.

TextCaps

The billboard displays 'Welcome to Yakima The Palm Springs of Washington'.

Conceptual Captions

Cars are on the streets.

Small stand of trees, just visible in the distance in the previous photo.

Fashion Captioning

A decorative leather padlock on a compact bag with croc embossed leather.

CUB-200

This bird is blue with white on its chest and has a very short beak.

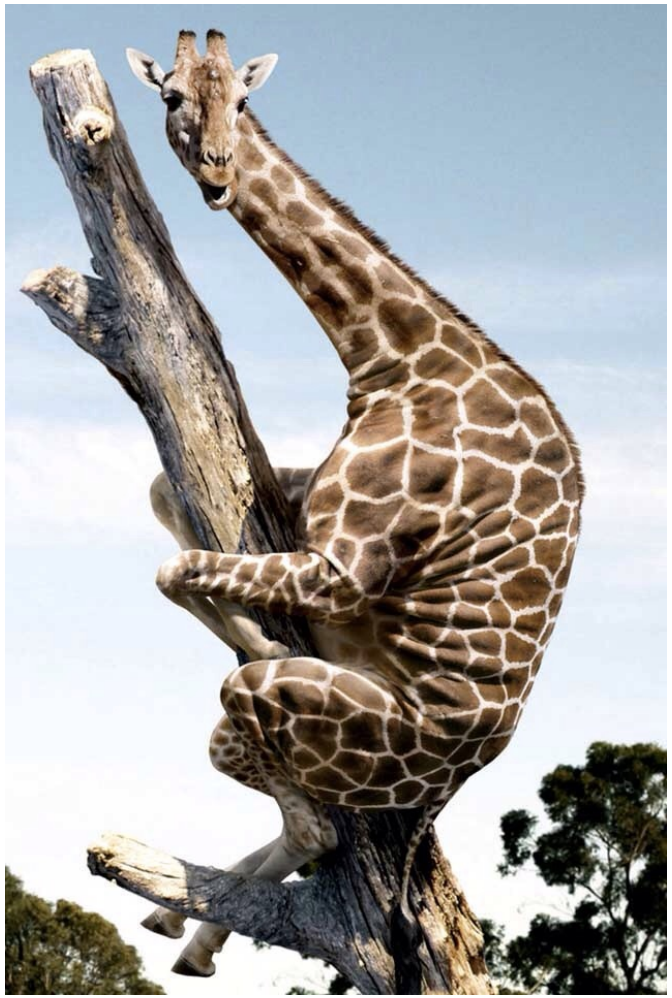
Stefanini et al. From Show to Tell: A Survey on Deep Learning-based Image Captioning. arXiv 2021

Sample of Existing Dataset Challenges

	Domain	Nb. Images	Nb. Caps (per Image)	Vocab Size	Nb. Words (per Cap.)
COCO [128]	Generic	132K	5	27K (10K)	10.5
Flickr30K [129]	Generic	31K	5	18K (7K)	12.4
Flickr8K [19]	Generic	8K	5	8K (3K)	10.9
CC3M [130]	Generic	3.3M	1	48K (25K)	10.3
CC12M [131]	Generic	12.4M	1	523K (163K)	20.0
SBU Captions [4]	Generic	1M	1	238K (46K)	12.1
VizWiz [132]	Assistive	70K	5	20K (8K)	13.0
CUB-200 [133]	Birds	12K	10	6K (2K)	15.2
Oxford-102 [133]	Flowers	8K	10	5K (2K)	14.1
Fashion Cap. [134]	Fashion	130K	1	17K (16K)	21.0
BreakingNews [135]	News	115K	1	85K (10K)	28.1
GoodNews [136]	News	466K	1	192K (54K)	18.2
TextCaps [137]	OCR	28K	5/6	44K (13K)	12.4
Loc. Narratives [138]	Generic	849K	1/5	16K (7K)	41.8

Stefanini et al. From Show to Tell: A Survey on Deep Learning-based Image Captioning. arXiv 2021

Class Task: How Would You Describe This Image?



Form: https://forms.gle/Nbue5HcdP9Dib8Co8

VLT2K

Guidelines and Examples:

Read these guidelines carefully. You must write exactly two sentences.

1. Describe the action being performed and mention the person performing the action and all objects involved in the action.
2. Describe any objects in the image that are not directly involved in the action.

A man is reading a newspaper. It is cloudy and there are skyscrapers in the background.

A boy is typing on a laptop. There is a brown bookshelf behind him and a bright window.

A man is talking on the telephone. There is a red lampshade and three red chairs in the background.

D. Elliott and F. Keller. Image description using visual dependency representations. EMNLP 2013.

Flickr8K and 30K

Guidelines:

- You must describe each of the following five images with one sentence.
- Please provide an accurate description of the activities, people, animals and objects you see depicted in the image
- Each description must be a single sentence under 100 characters. Try to be concise.
- Please pay attention to grammar and spelling.
- We will accept your results if you provide a good description for all five images, leaving nothing blank.

Examples of good and bad descriptions.

(1) The dog is wearing a red sombrero.
Very Good: This describes the two main objects concisely and accurately.

(2) White dog wearing a red hat.
Good: Incomplete sentences like this are fine.

(3) The white dog is wearing a pink collar.
Okay: This describes the dog, but it ignores the hat.

(4) The red hat is adorned with gold sequins.
Bad: This ignores the dog.

(5) The dog is angry because he is hungry.
Bad: This is speculation.

(6) The dog.
Very Bad: This could describe any image of any dog.

Hodosh, Young, and Hockenmaier. Framing image description as a ranking task: Data, models, and evaluation metrics. JAIR 2013

MSCOCO

Instructions:

- Describe all the important parts of the scene.
- Do not start the sentences with "There is".
- Do not describe unimportant details.
- Do not describe things that might have happened in the future or past.
- Do not describe what a person might say.
- Do not give people proper names.
- The sentence should contain at least 8 words.

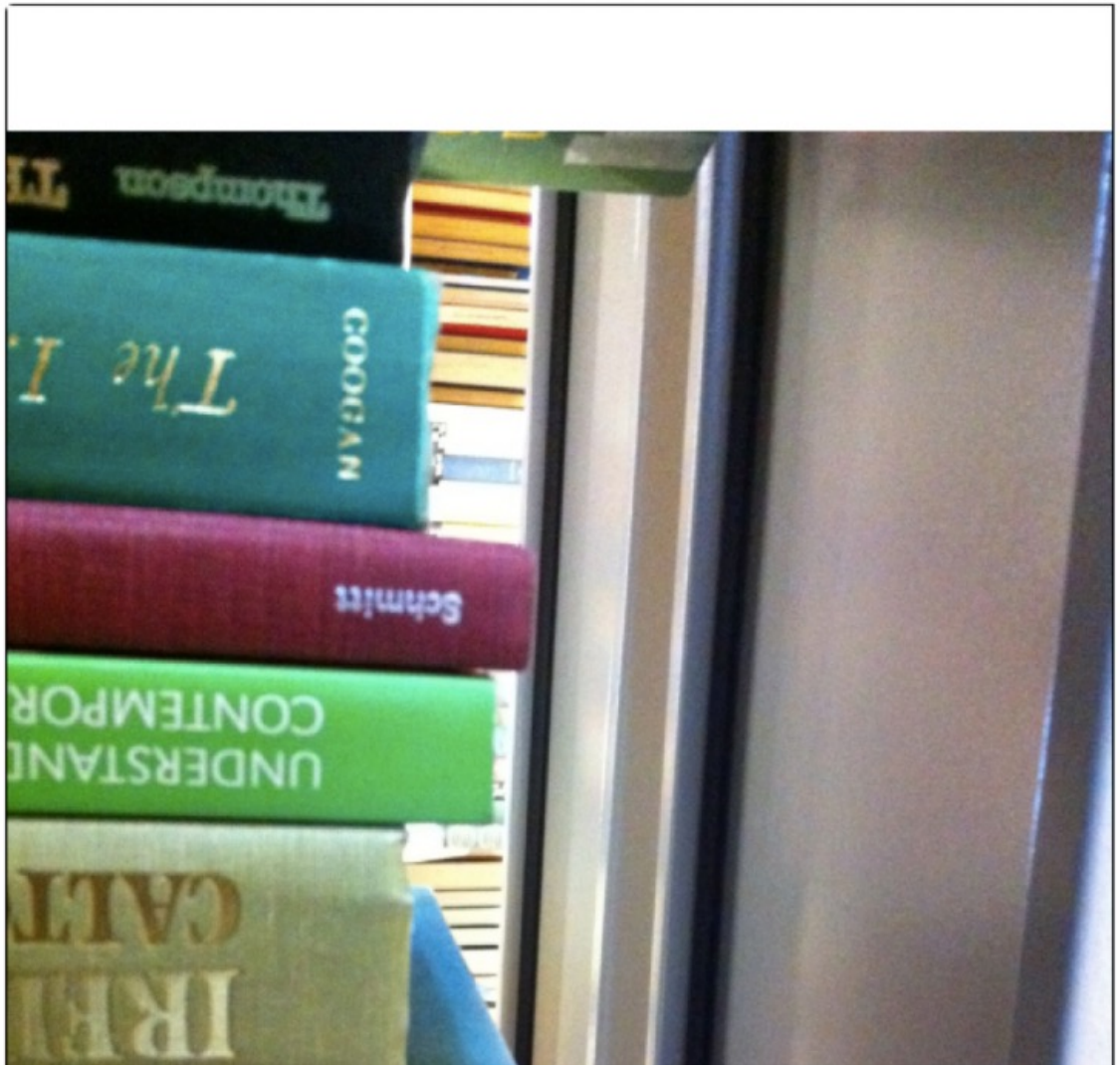
Please describe the image:

Enter description here

prev next

Chen et al. Microsoft COCO Captions: Data Collection and Evaluation Server. arXiv 2015.

VizWiz



Step 1: Please describe the image in one sentence.

- Describe all parts of the image that may be **important to a person who is blind.** *E.g., imagine how you would describe this image on the phone to a friend.*
- **DO NOT** speculate about what people in the image might be saying or thinking.
- **DO NOT** describe things that may have happened in the future or past.
- **DO NOT** use more than one sentence.
- If text is in the image, and is important, then you can summarize what it says. **DO NOT** use all the specific phrases that you see in the image as your description of the image.
- **DO NOT** describe the image quality issues. This is covered in Step 3. If the image quality issues make it **impossible to recognize the visual content** (e.g., image is totally black or white), then use the following description (you can copy-paste):

  Quality issues are too severe to recognize visual content. Copy to description

- Your description should contain at least **8 words**.

Type here. Do not start the description with:
- "There is/are ..."
- "This is / These are ..."
- "The/This image/picture ..."
- "It is/ It's ..."

Gurari et al. Captioning Images Taken by People Who Are Blind. ECCV 2020

Personality-Captions

215 personalities selected from this list: http://ideonomy.mit.edu/essays/traits.html

Comment on an Image

Description

In this task, you will be shown 5 images, and will write a comment about each image. The goal of this task is to write something about an image that someone else would find engaging.

STEP 1

With each new photo, you will be given a **personality trait** that you will try to emulate in your comment. For example, you might be given "**snarky**" or "**glamorous**". The personality describes ***YOU***, not the picture. It is *you* who is snarky or glamorous, not the contents of the image.

STEP 2

You will then be shown an image, for which you will write a comment *in the context of your given personality trait*. Please make sure your comment has at least **three words**. Note that these are *comments*, not captions.

E.g., you may be shown an image of a tree. If you are "**snarky**", you might write "What a boring tree, I bet it has bad wood;" or, if you were "**glamorous**", you might write "What an absolutely beautiful tree! I would put this in my living room it's so extravagent!"

Image

Your assigned personality is:

Adventurous

Reminder - please do not write anything that involves any level of discrimination, racism, sexism and offensive religious/politics comments, otherwise the submission will be rejected.

K. Shuster, S. Humeau, H. Hu, A. Bordes, and J. Weston. Engaging image captioning via personality. CVPR 2019.

How Would You Evaluate Captions from an Algorithm?

FEATURE NAME:	VALUE
Description	{ "tags": ["outdoor", "giraffe", "animal", "mammal", "standing", "field", "top", "branch", "bird", "eating", "head", "grazing", "neck", "water", "large", "man", "grassy", "tall", "group", "dirt", "zoo"], "captions": [{ "text": "a giraffe standing in the dirt", "confidence": 0.982929349 }] }

Evaluation: Human Judgments

Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
1	2	3	4	5	6

- The description accurately describes the image (Kulkarni et al., 2011; Li et al., 2011; Mitchell et al., 2012; Kuznetsova et al., 2012; Elliott & Keller, 2013; Hodosh et al., 2013).
- The description is grammatically correct (Yang et al., 2011; Mitchell et al., 2012; Kuznetsova et al., 2012; Elliott & Keller, 2013).
- The description has no incorrect information (Mitchell et al., 2012).
- The description is relevant for this image (Li et al., 2011; Yang et al., 2011).
- The description is creatively constructed (Li et al., 2011).
- The description is human-like (Mitchell et al., 2012).

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. Automatic Description from Images: A Survey of Models, Datasets, and Measures. JAIR 2016.

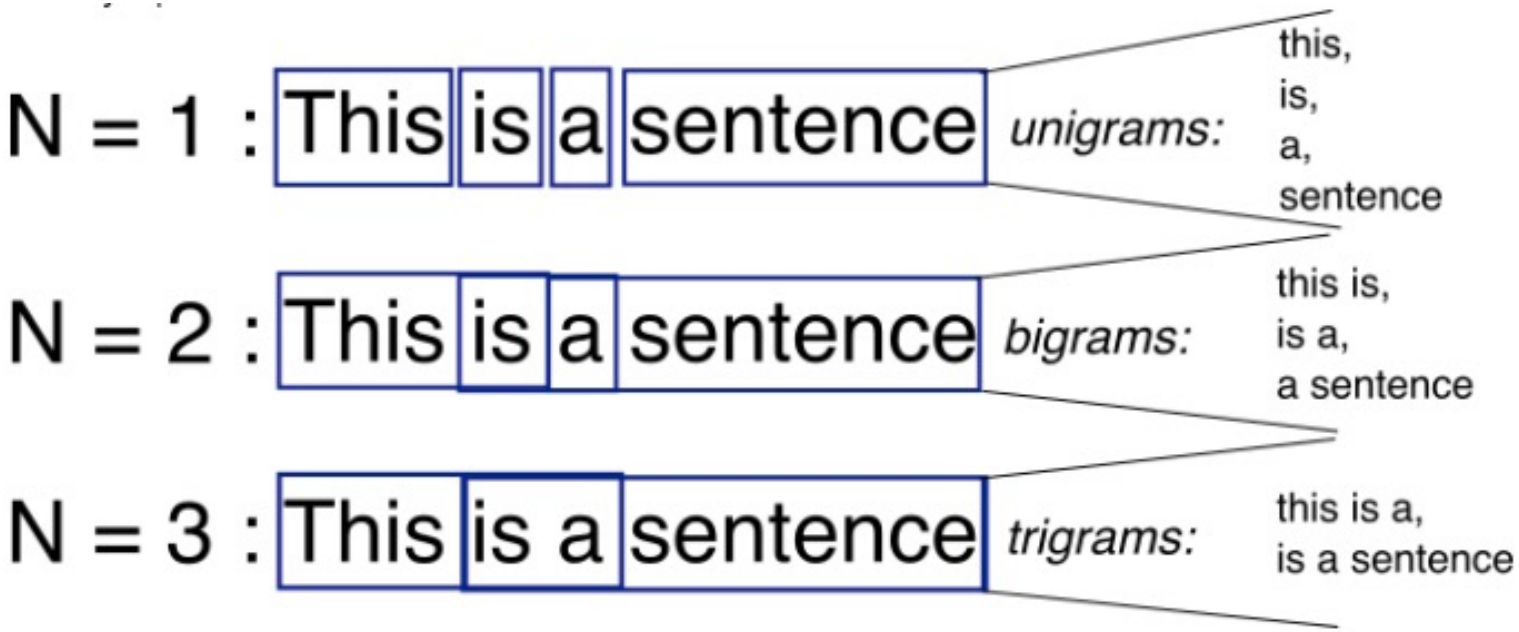
Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Idea: compute similarities of n-grams between a predicted caption and each ground truth caption



http://recognize-speech.com/language-model/n-gram-model/comparison

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Idea: measure similarity of a predicted caption to how most people describe an image based on *n*-grams unique to the image

A cow is standing in a field.

A cow with horns and long hair covering its face stands in a field.

A cow with hair over its eyes stands in a field.

This horned creature is getting his picture taken.

A furry animal with horns roams on the range.

R. Vedantam, C. L. Zitnick, and D. Parikh. CIDEr: Consensus-based Image Description Evaluation. CVPR 2015.

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

What content do most people describe in this image?

A cow is standing in a field.

A cow with horns and long hair covering its face stands in a field.

A cow with hair over its eyes stands in a field.

This horned creature is getting his picture taken.

A furry animal with horns roams on the range.

R. Vedantam, C. L. Zitnick, and D. Parikh. CIDEr: Consensus-based Image Description Evaluation. CVPR 2015.

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Do you think these two captions describe the same image?

(a) A young girl *standing on top of a* tennis court.
(b) A giraffe *standing on top of a* green field.

P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

Evaluation: Automated

- BLEU

Problem: n-gram methods scores these as very similar

- METEOR

(a) A young girl *standing on top of a* tennis court.
(b) A giraffe *standing on top of a* green field.

- Rouge
- CIDEr
- SPICE

P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Do you think these two captions describe the same image?

(c) A shiny metal pot filled with some diced veggies.
(d) The pan on the stove has chopped vegetables in it.

P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

Evaluation: Automated

- BLEU

Problem: n-gram methods scores these as very different

- METEOR

(c) A shiny metal pot filled with some diced veggies.
(d) The pan on the stove has chopped vegetables in it.

- Rouge
- CIDEr
- SPICE

P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

Evaluation: Automated

Idea: compare scene graph of prediction to scene graph of ground truth

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

"two women are sitting at a white table"

"two women sit at a table in a small store"

"two women sit across each other at a table smile for the photograph"

"two women sitting in a small store like business"

"two woman are sitting at a table"

P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

Evaluation: Automated

What is the meaningful semantic content in these captions?

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

"two women are sitting at a white table"

"two women sit at a table in a small store"

"two women sit across each other at a table smile for the photograph"

"two women sitting in a small store like business"

"two woman are sitting at a table"

P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Meaningful semantic content in these captions:

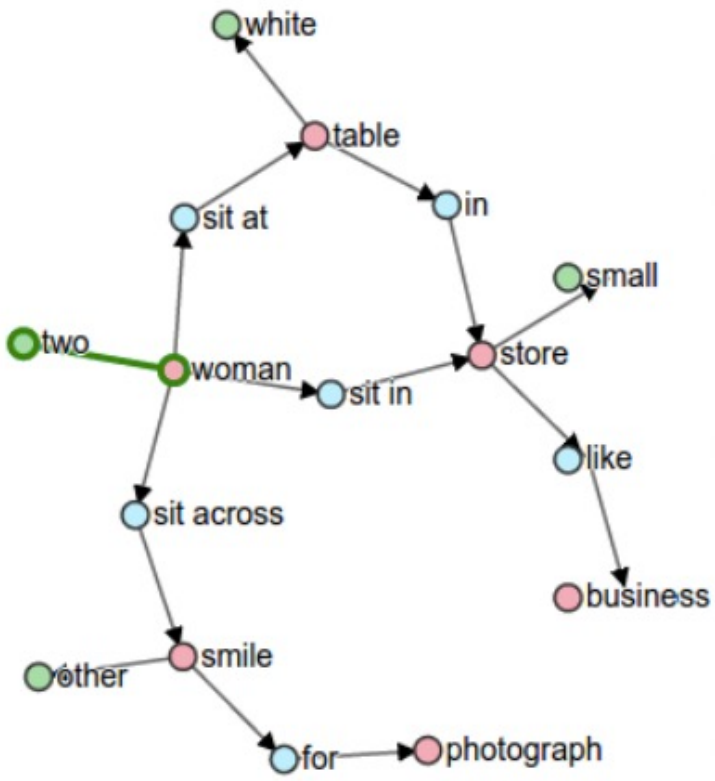
"two women are sitting at a white table"

"two women sit at a table in a small store"

"two women sit across each other at a table smile for the photograph"

"two women sitting in a small store like business"

"two woman are sitting at a table"

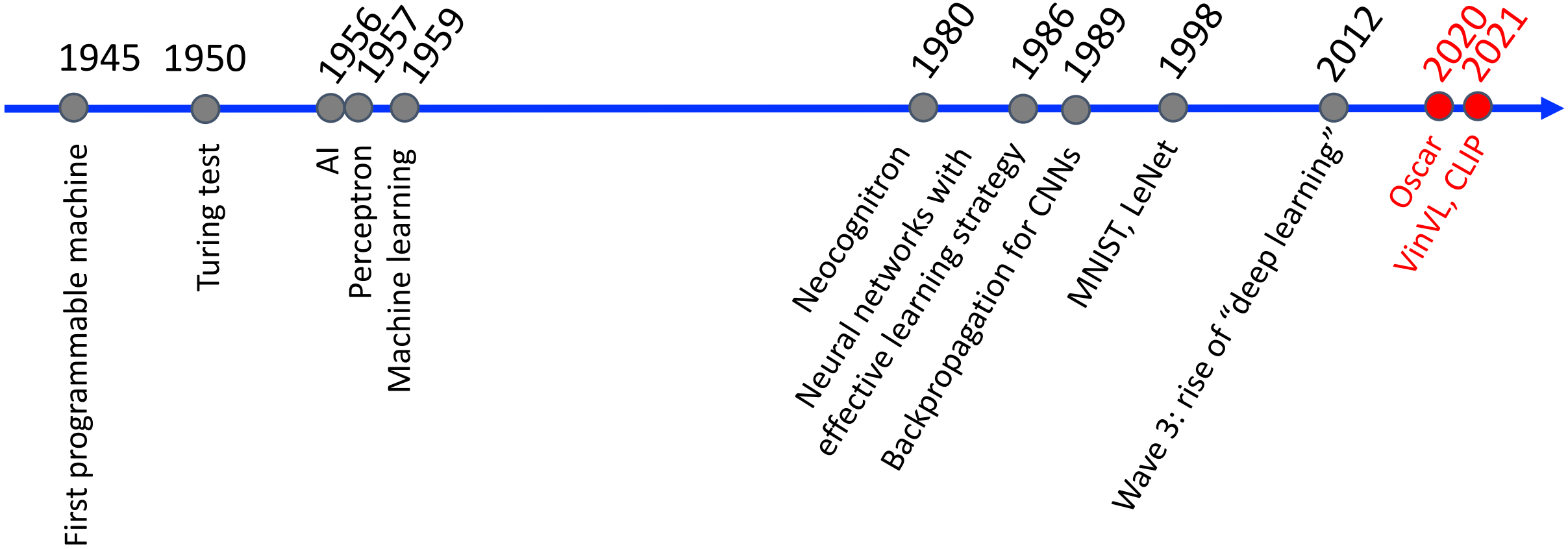


P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

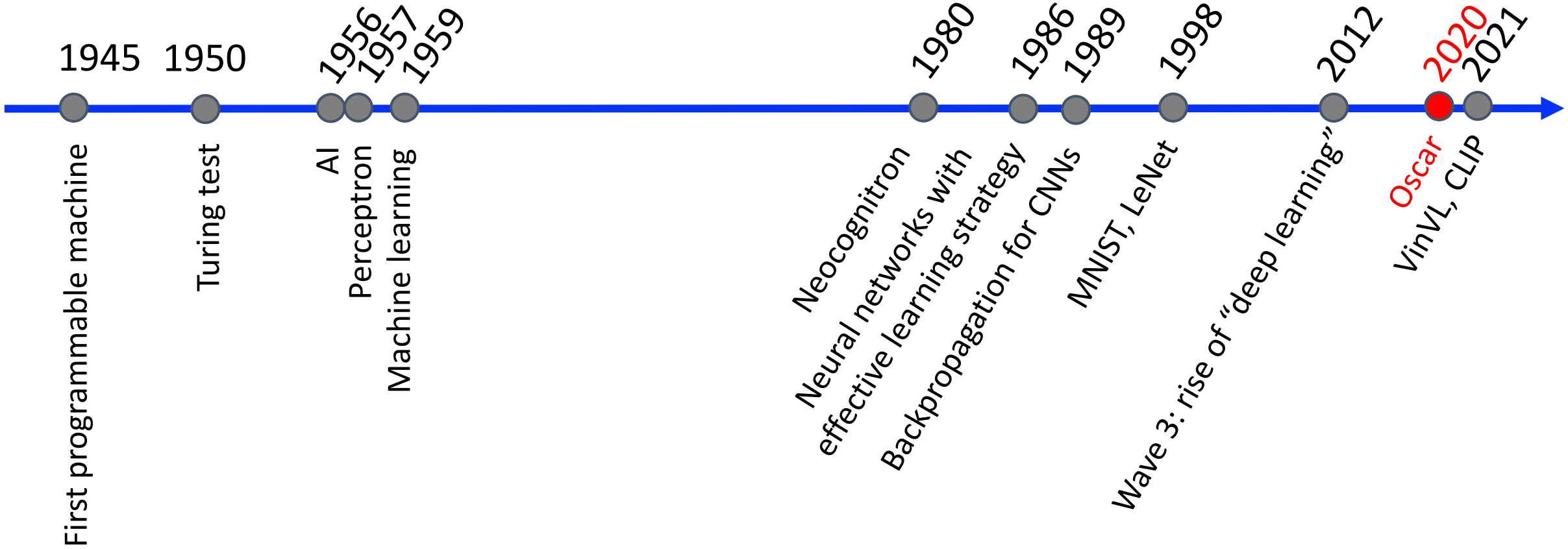
Today’s Topics

- Multimodal applications
- Image captioning dataset challenges
- Image captioning algorithms
- Visual question answering dataset challenges
- Visual question answering algorithms
- Foundation models

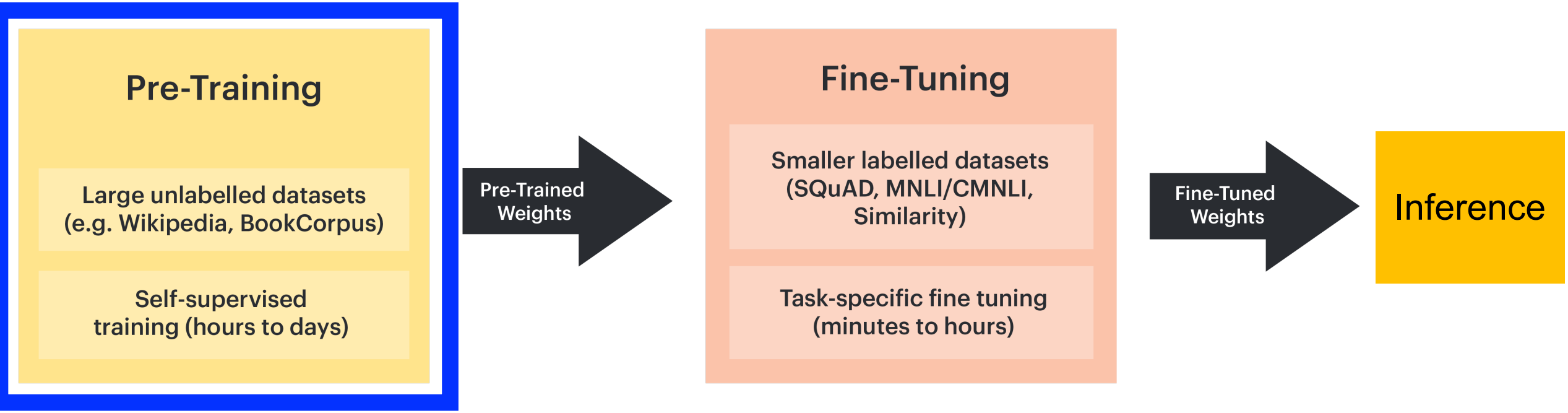
Historical Context



Historical Context



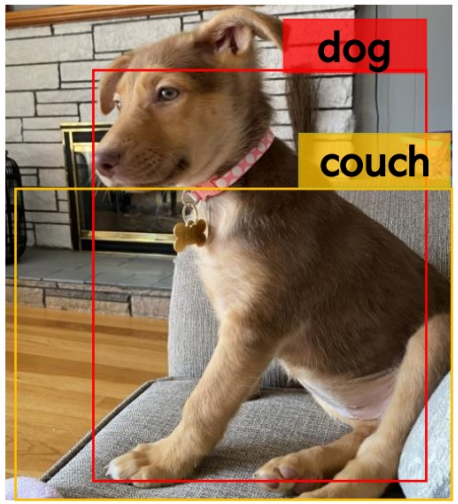
Oscar: Transformer Design



https://docs.graphcore.ai/projects/bert-training/en/latest/bert.html

Novelty: Adds **Explicit** Alignment Between Visual and Textual Concepts

- **Idea**: rather than have algorithm learn alignment between text and features describing image regions, align them explicitly
- **Motivating observations**: often, salient objects are mentioned in image descriptions and can be located by object detection algorithms

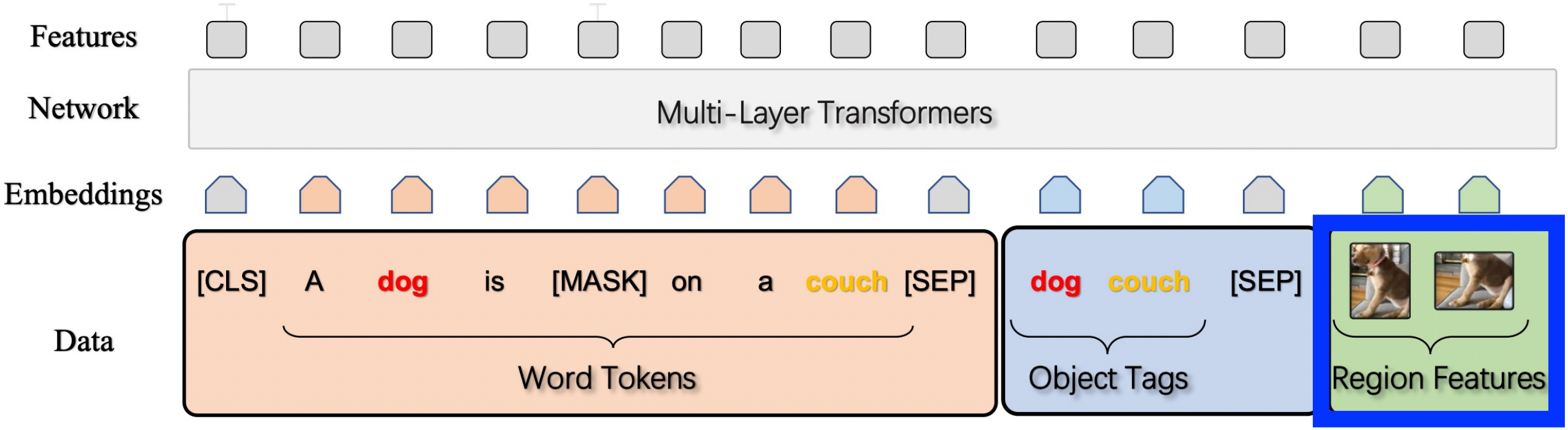


A dog is sitting on a couch

VS

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

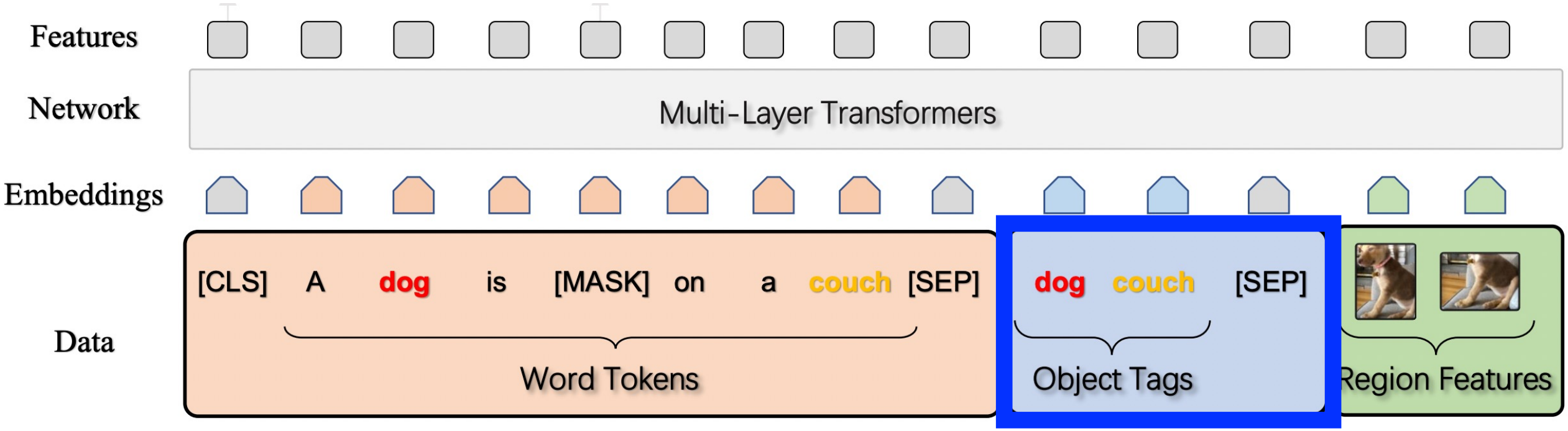
Oscar: Architecture



Each image is represented as description of objects detected with Faster R-CNN using features from Faster R-CNN

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

Oscar: Architecture

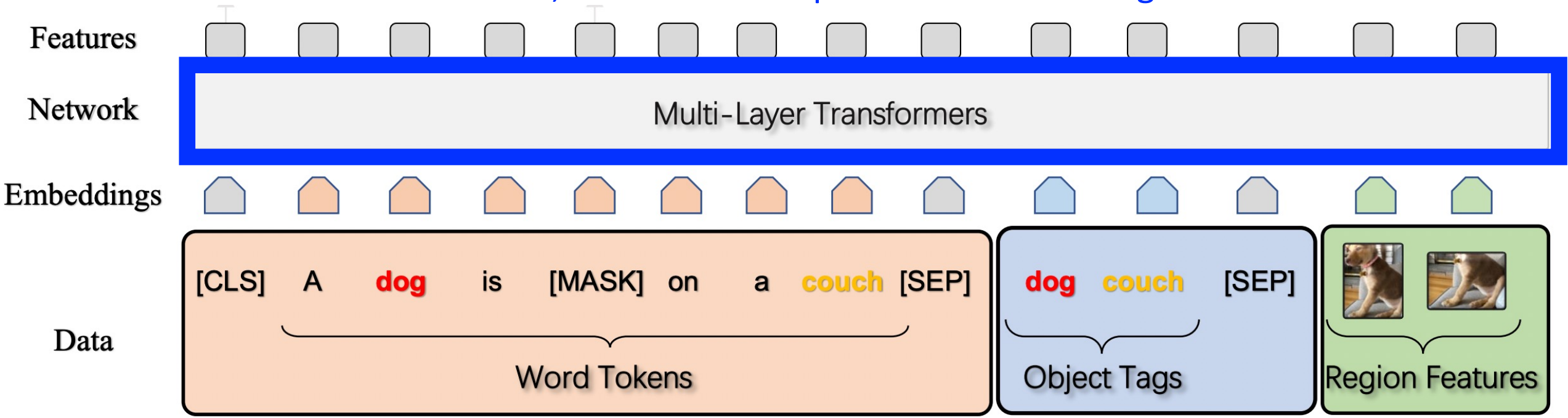


Novelty is to incorporate tags predicted by Faster R-CNN

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

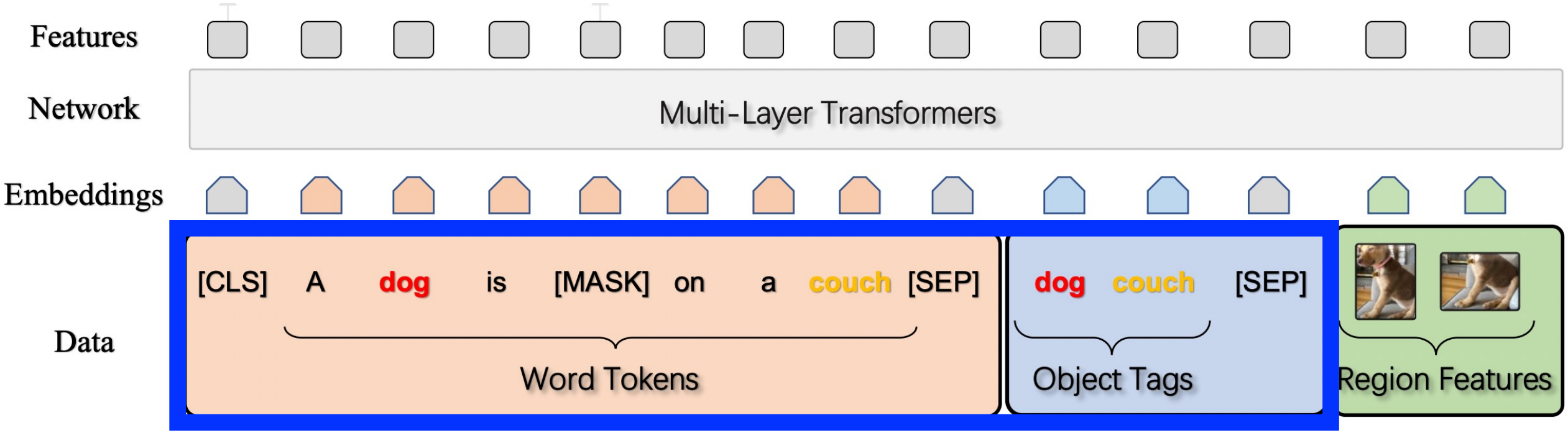
Oscar: Architecture

Uses BERT architecture, initialized with pretrained BERT weights



Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

Oscar: 2 Pretraining Tasks (Masked Token Loss and Contrastive Loss)

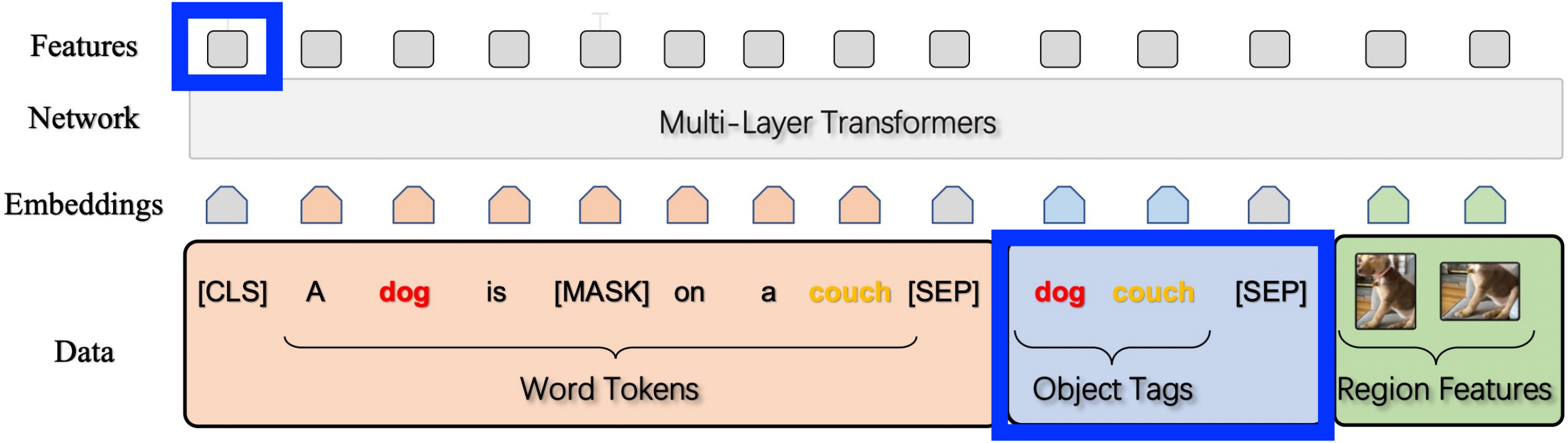


Like BERT, predict randomly masked tokens based on surrounding words, tags, and image information

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

Oscar: 2 Pretraining Tasks (Masked Token Loss and Contrastive Loss)

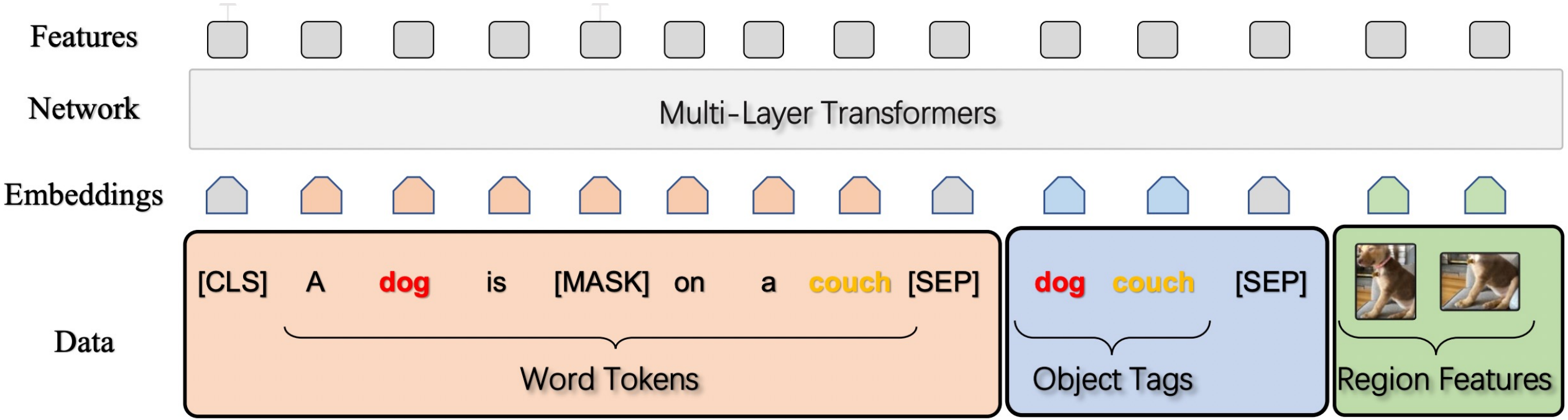
Fully-connected layer added to enable binary classification based on the fused vision-language token representation



Goal is to determine whether tags are original when 50% of tags are replaced with randomly selected tag sequence in the dataset

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

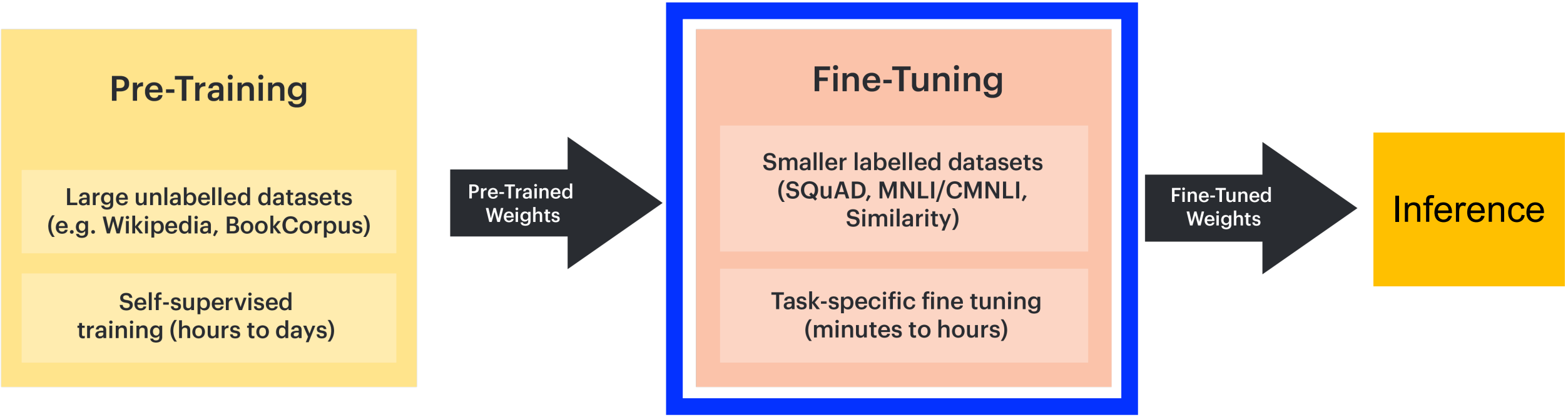
Oscar: 2 Pretraining Dataset



6.5 million text-tag-image triplets derived from existing V+L datasets

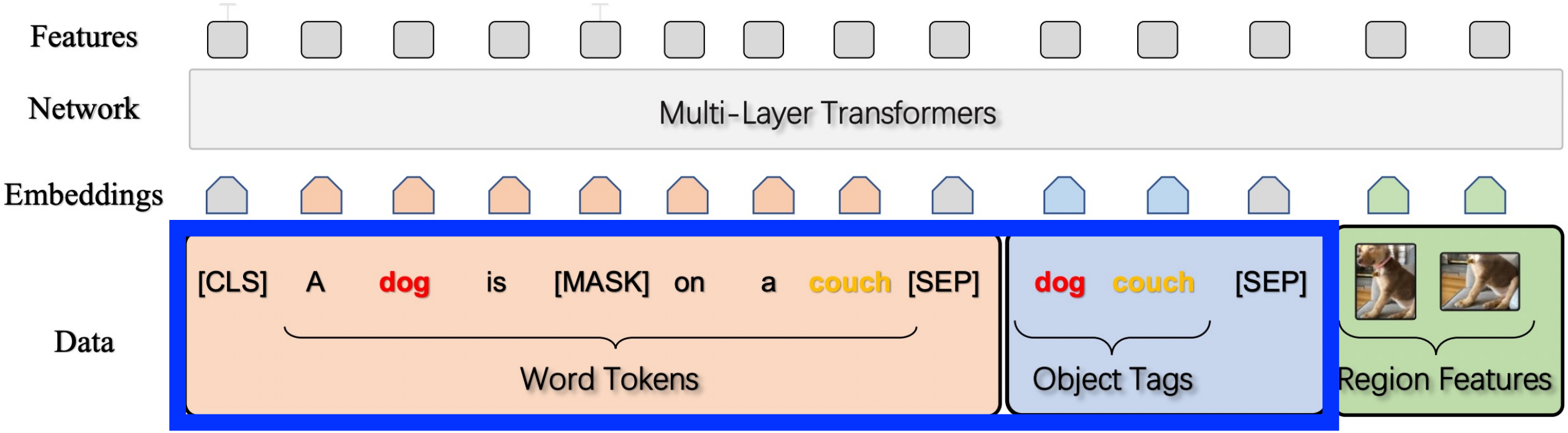
Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

Oscar: Transformer Design



https://docs.graphcore.ai/projects/bert-training/en/latest/bert.html

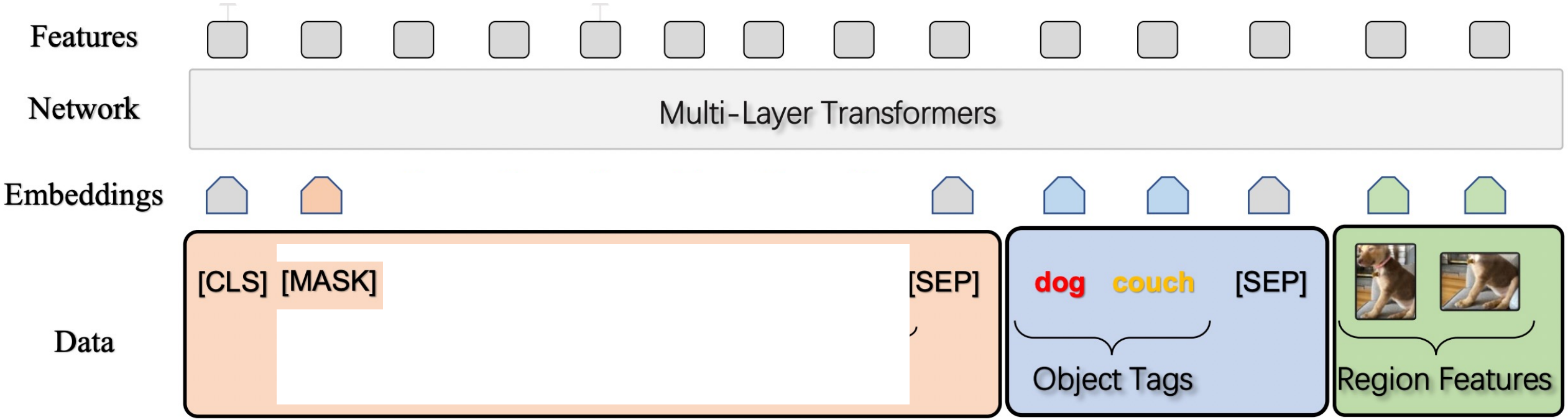
Oscar: 2 Fine-Tuning Task (Masked Token Loss)



Similar to pre-training, predict randomly masked tokens based on surrounding words, tags, and image information (on COCO dataset)

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

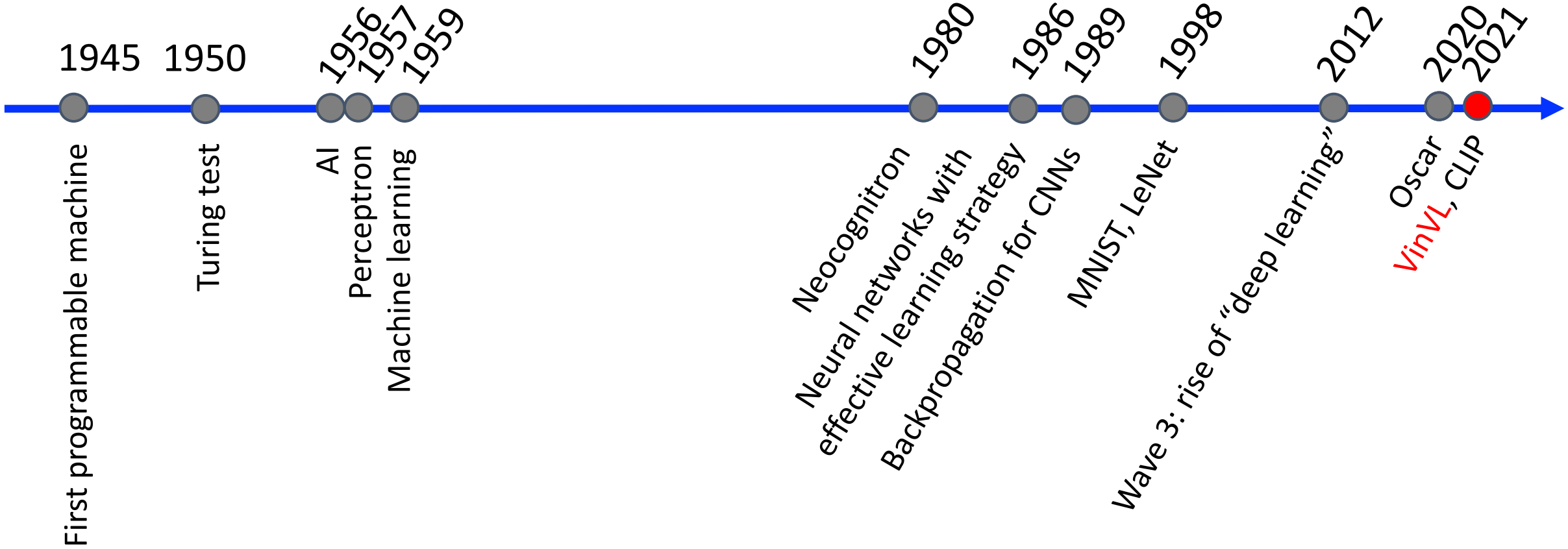
Oscar: Inference Time



Repeatedly predict a new [MASK] token, incorporating the predicted word into the sequence, until [STOP] is predicted.

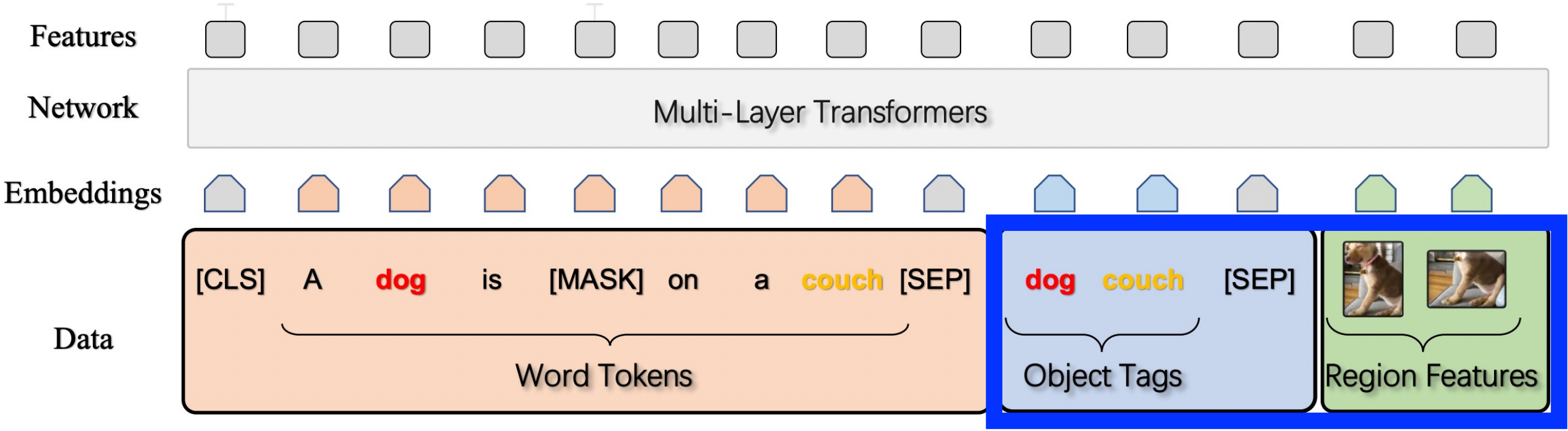
Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

Historical Context



Idea: Oscar + Improved Visual Representation

VinVL Architecture: Oscar + New Object Detector

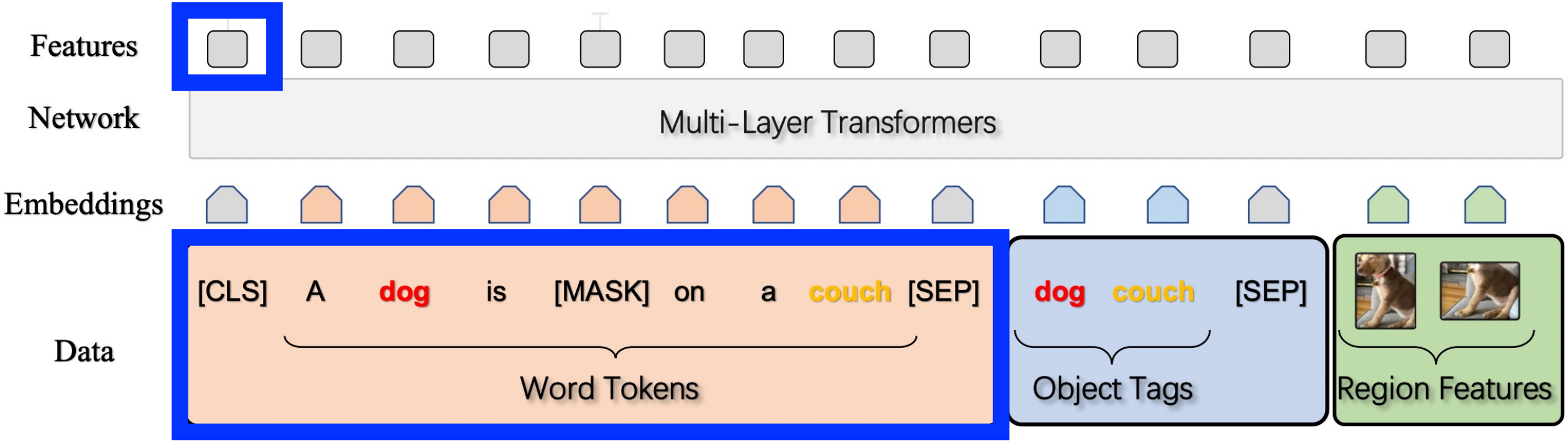


Improved object detector to predict more diverse categories and train larger models on larger datasets

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

VinVL: 2 Pretraining Tasks (Masked Token Loss and Contrastive Loss)

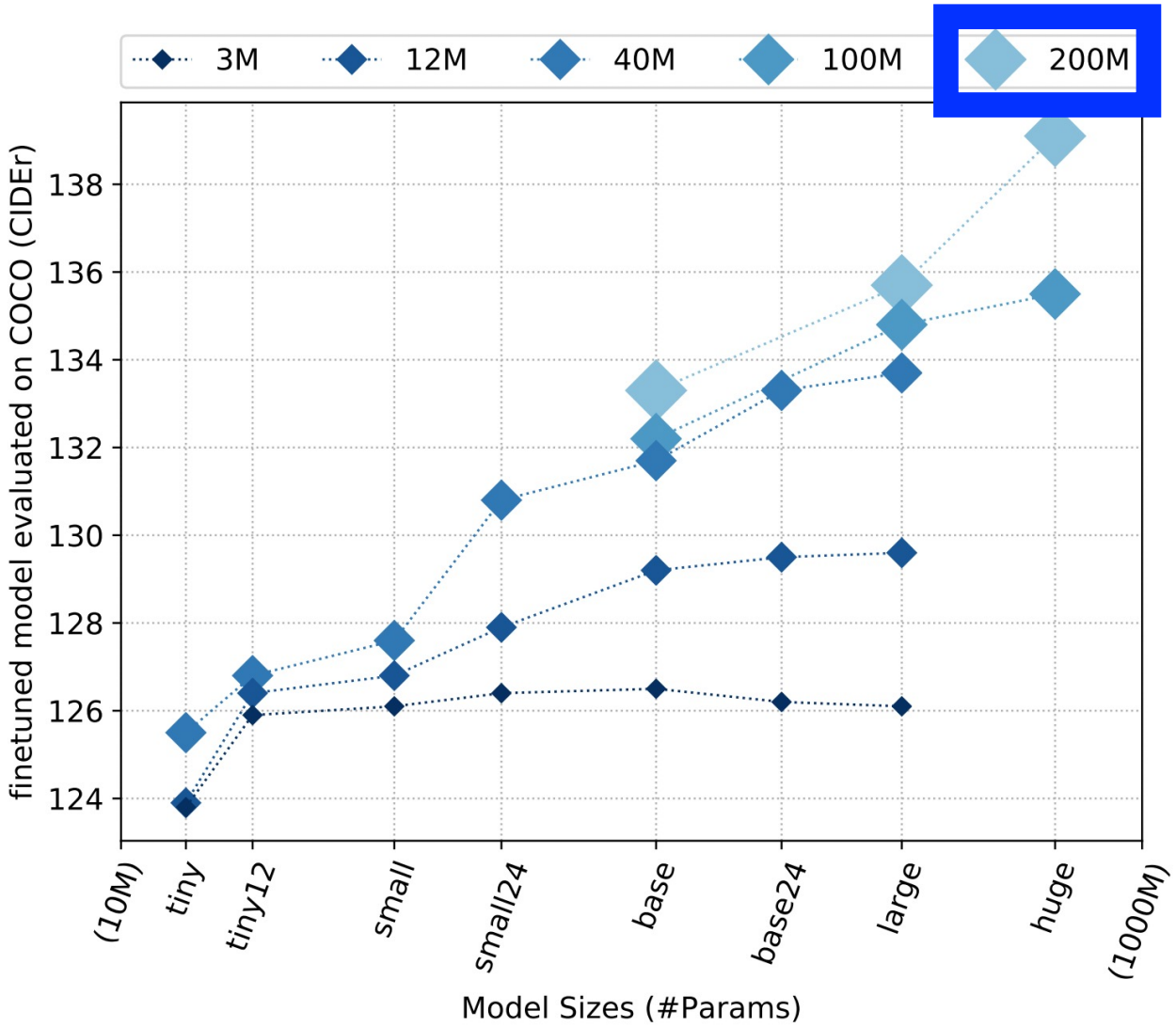
Fully-connected layer added to enable 3-way classification based on the fused vision-language token representation



Trained on 8.85 million text-image pairs to decide whether either captions or answers are corrupted (50% are not) for caption-tags-image triplets and question-answer-image triplets

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

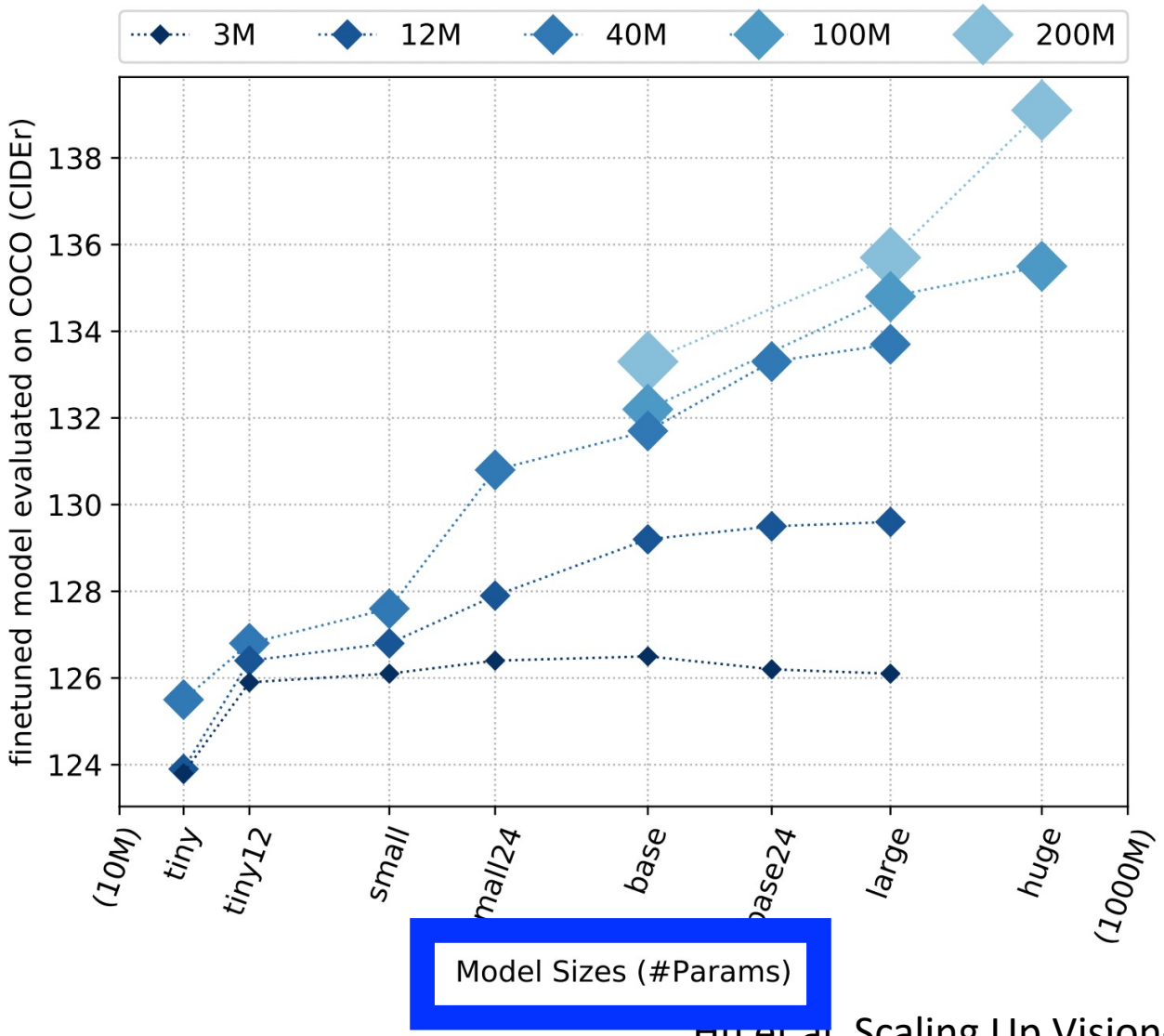
VinVL: Influence of Model and Dataset Sizes



200M images, each with 1 alt text description, collected from Internet

Hu et al. Scaling Up Vision-Language Pre-training for Image Captioning. CVPR 2022

VinVL: Influence of Model and Dataset Sizes

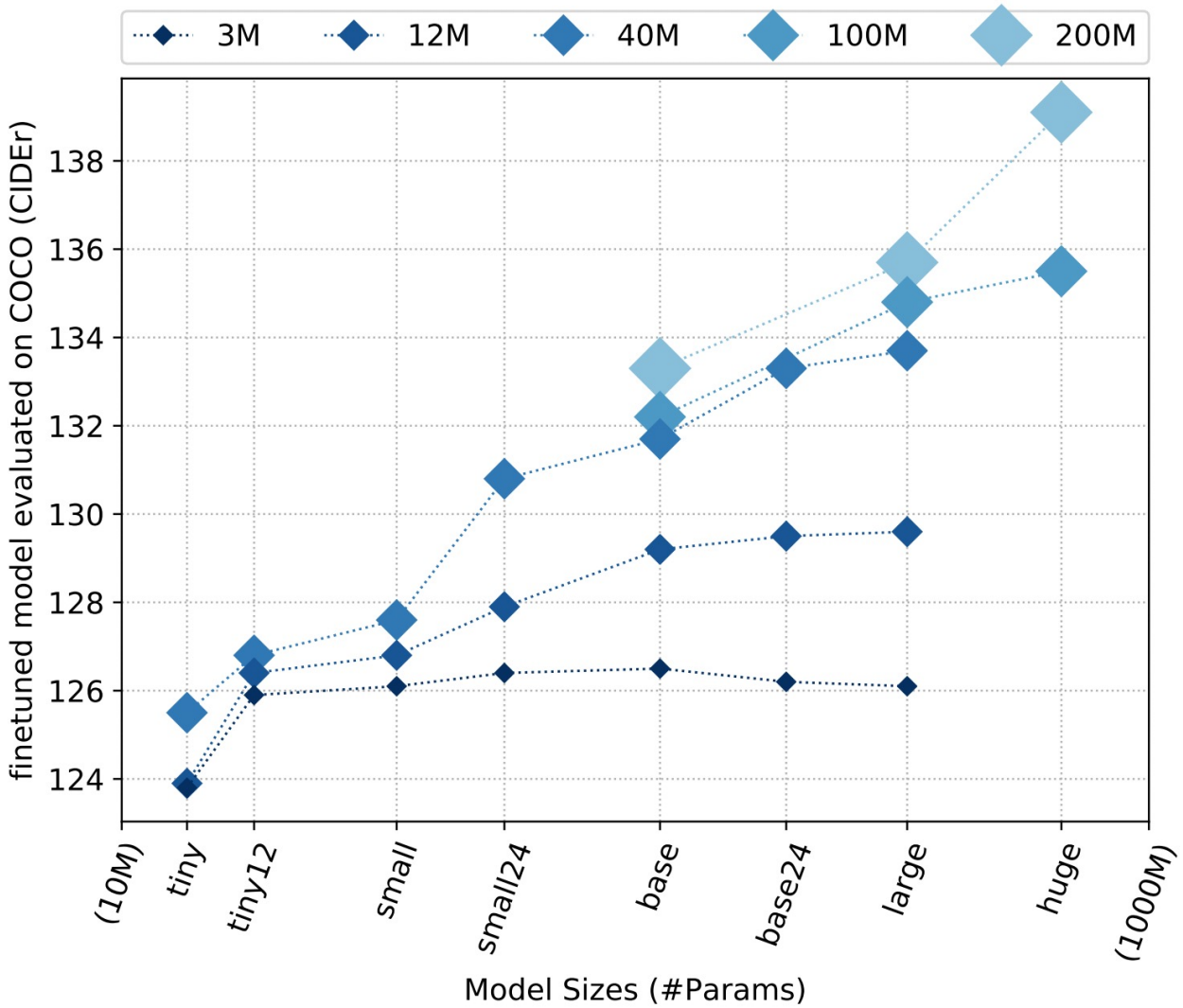


8 model sizes tested on COCO dataset

Model	Layers	Width	MLP	Heads	Param (M)
tiny	6	256	1024	4	13.4
tiny12	12	256	1024	4	18.1
small	12	384	1536	6	34.3
small24	24	384	1536	6	55.6
base	12	768	3072	12	111.7
base24	24	768	3072	12	196.7
large	24	1024	4096	16	338.3
huge	32	1280	5120	16	675.4

Hu et al. Scaling Up Vision-Language Pre-training for Image Captioning. CVPR 2022

VinVL: Influence of Model and Dataset Sizes

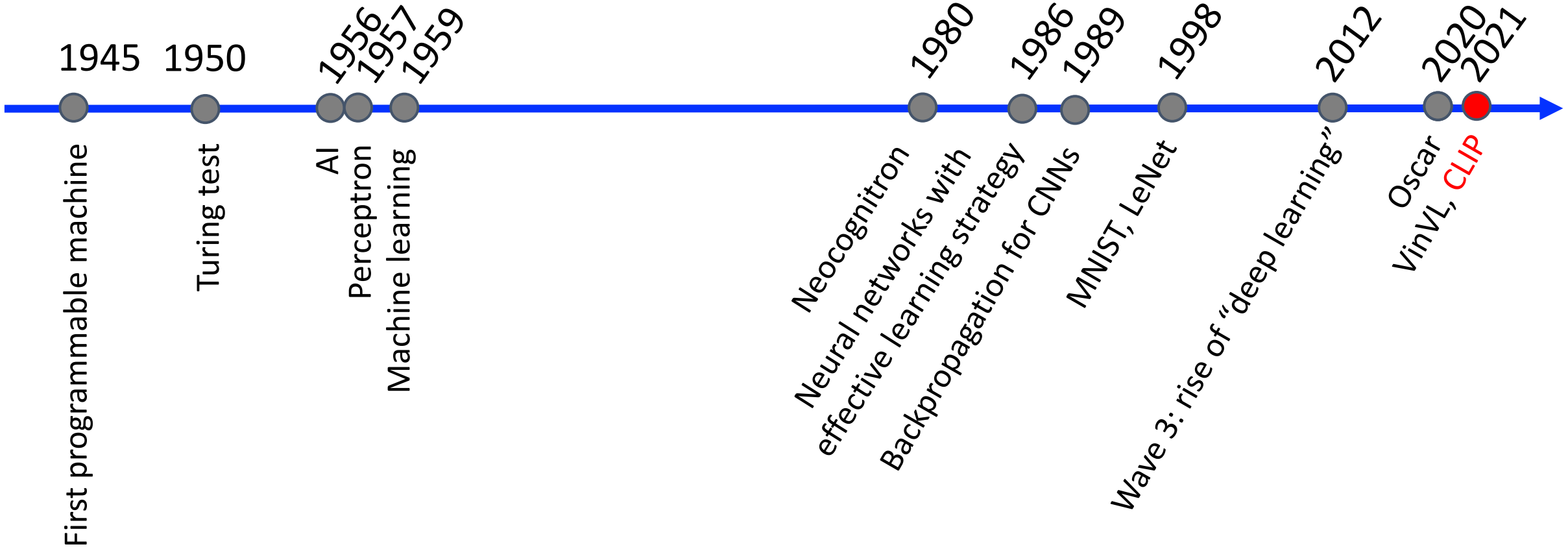


What trend(s) do you observe?

The trends of improved performance for large models and training datasets is generally observed for transformers

Hu et al. Scaling Up Vision-Language Pre-training for Image Captioning. CVPR 2022

Historical Context



Why CLIP?

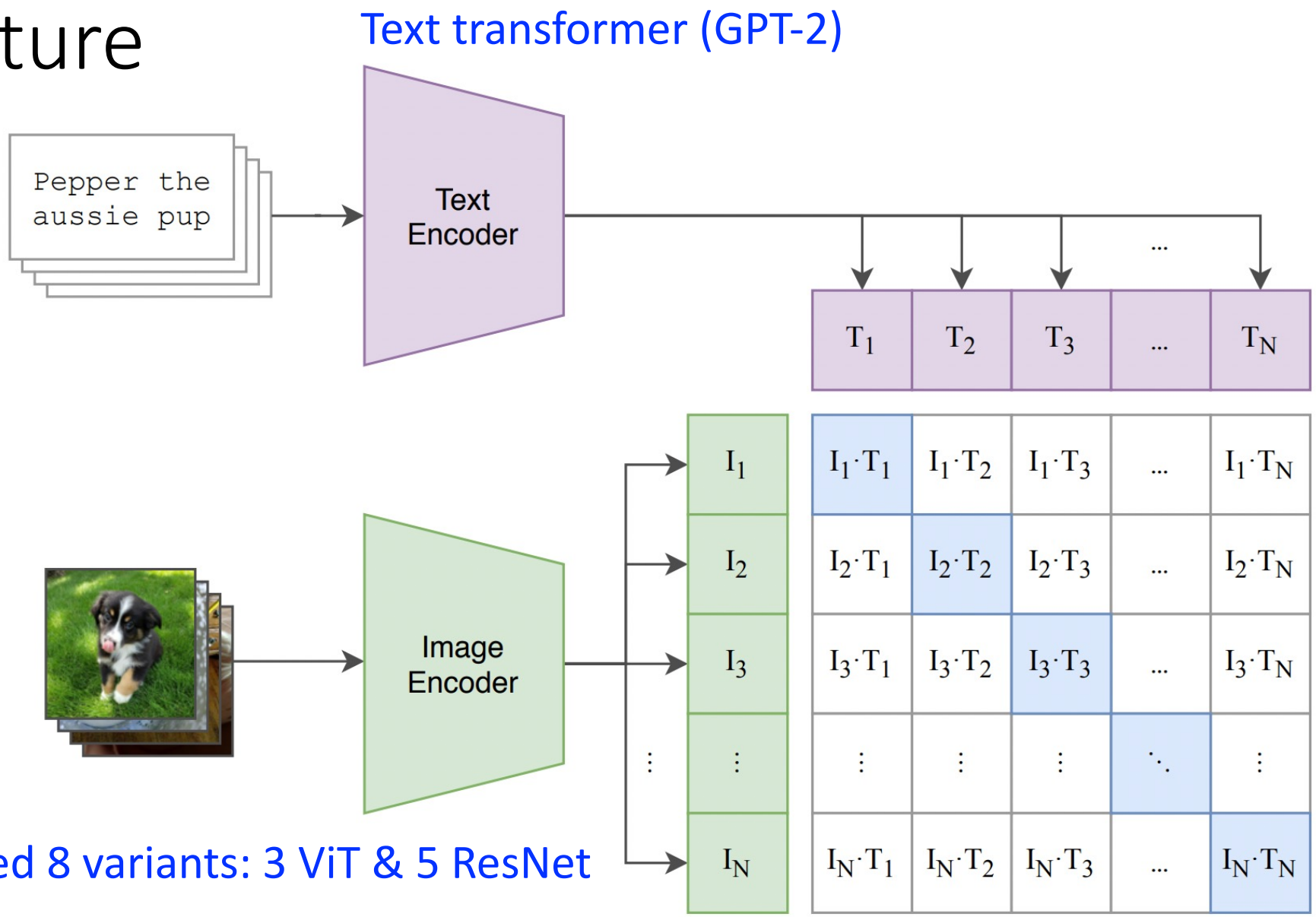
Named after the proposed technique: Contrastive Language Image Pre-training

Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.

CLIP Model: Novelty

- Train image analysis models with natural language supervision using the vast amounts of publicly available data on the Internet

CLIP Architecture



Tried 8 variants: 3 ViT & 5 ResNet

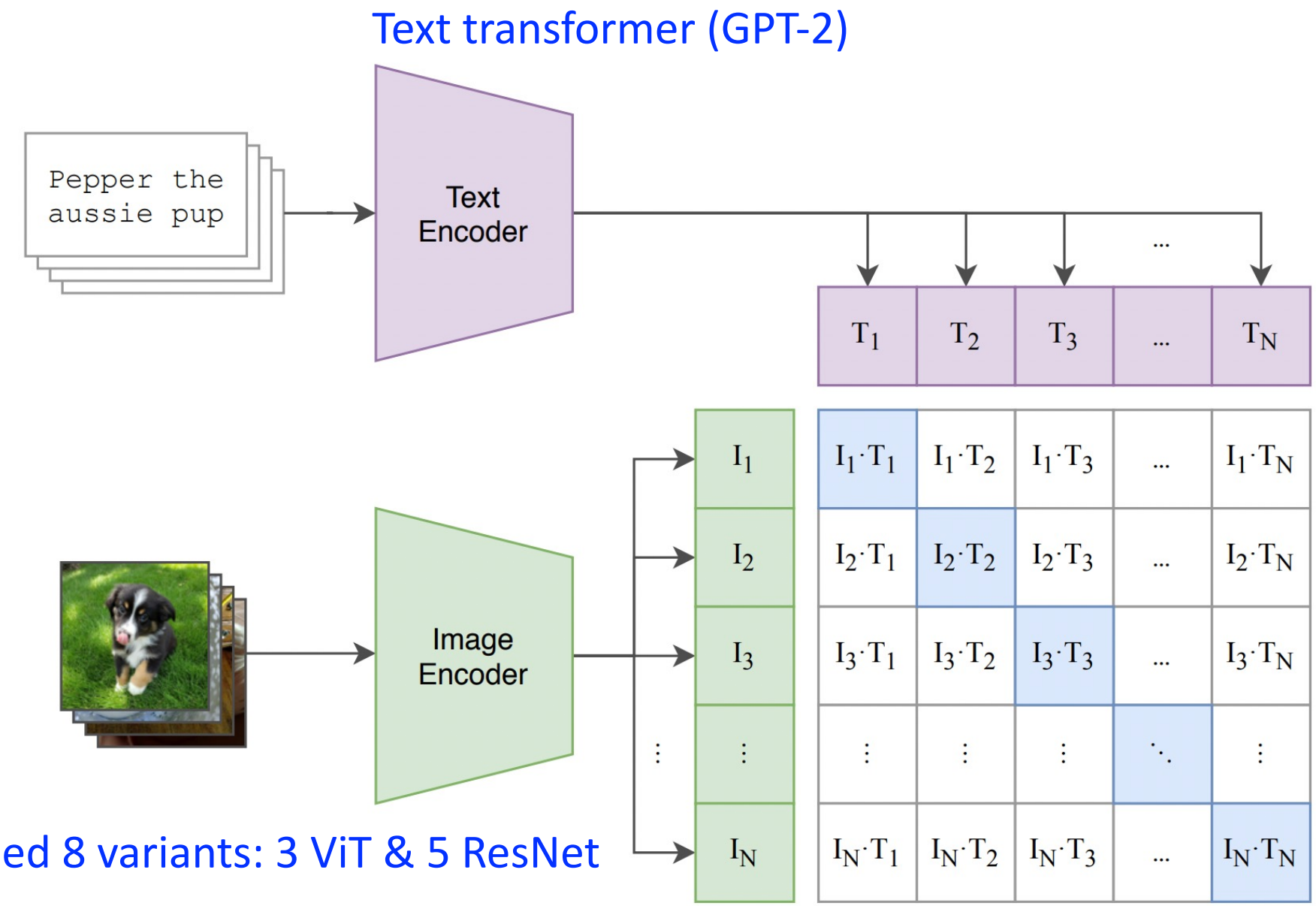
Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.

CLIP Training

Task: predict which image-text pairs match using 400 million image-text pairs from Internet containing any of 500,000 queries (e.g., words occurring 100+ times in English version of Wikipedia and all WordNet synonyms)

- Largest ResNet model took 18 days to train on 592 V100 GPUs and largest ViT took 12 days on 256 V100 GPUs

- Experiments run with largest ("best") ViT model

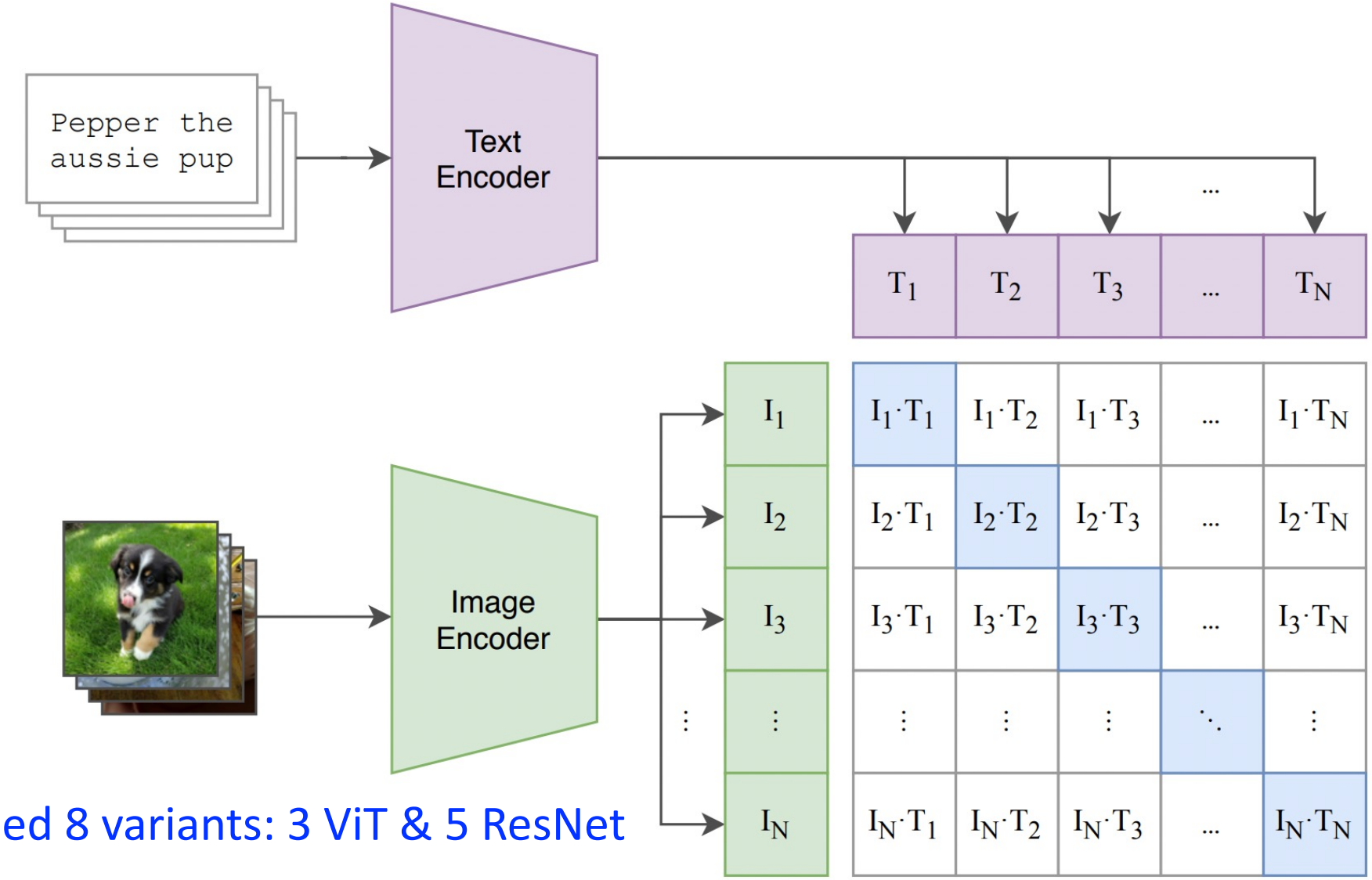


Tried 8 variants: 3 ViT & 5 ResNet

Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.

CLIP Training

Text transformer (GPT-2)



- Learns feature embeddings for image and text encoders that push correct image-text pairs together and incorrect image-text pairs apart.

- Learns nouns, verbs, adjectives, and more!

Tried 8 variants: 3 ViT & 5 ResNet

Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.

Zero-Shot Performance Evaluated on Over 30 Datasets

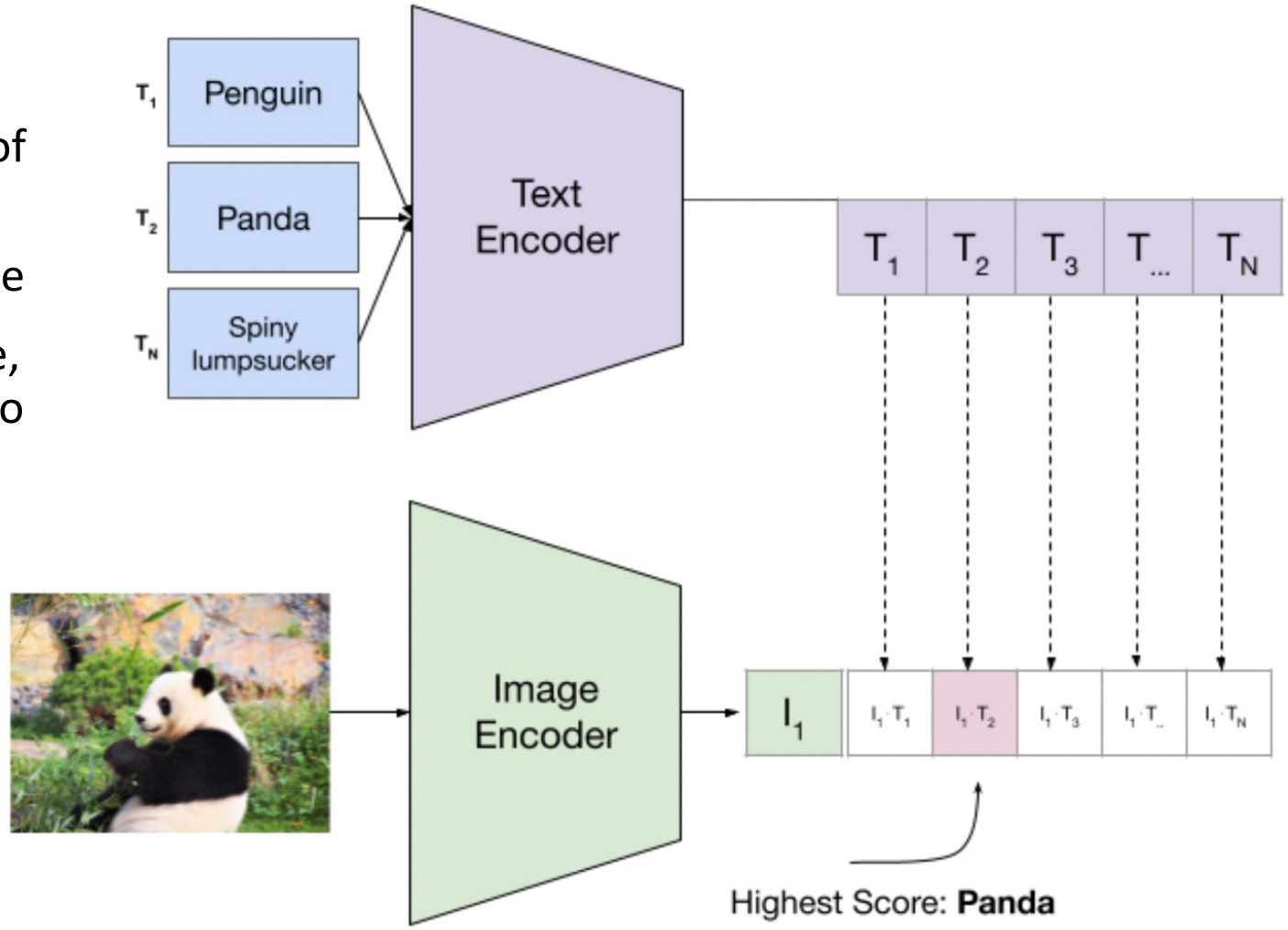
CLIP Inference

e.g., zero-shot classification:

1. Compute feature embedding for names of all classes in the dataset by its encoder

2. Compute feature embedding of the image

3. Compute cosine similarity of each (image, text) pair embedding followed by softmax to identify most probable match



https://towardsdatascience.com/understanding-zero-shot-learning-making-ml-more-human-4653ac35ccab

CLIP Inference

Prompts “engineered” to mimic that training data often had sentences (instead of words):

- classification: “A photo of a {label}”
- fine-grained classification: “A photo of a {label}, a type of pet/food/aircraft/etc”
- satellite image classification: “A satellite photo of a {label}”
- ensembles: “A photo of a big/small/etc {label}”

Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.

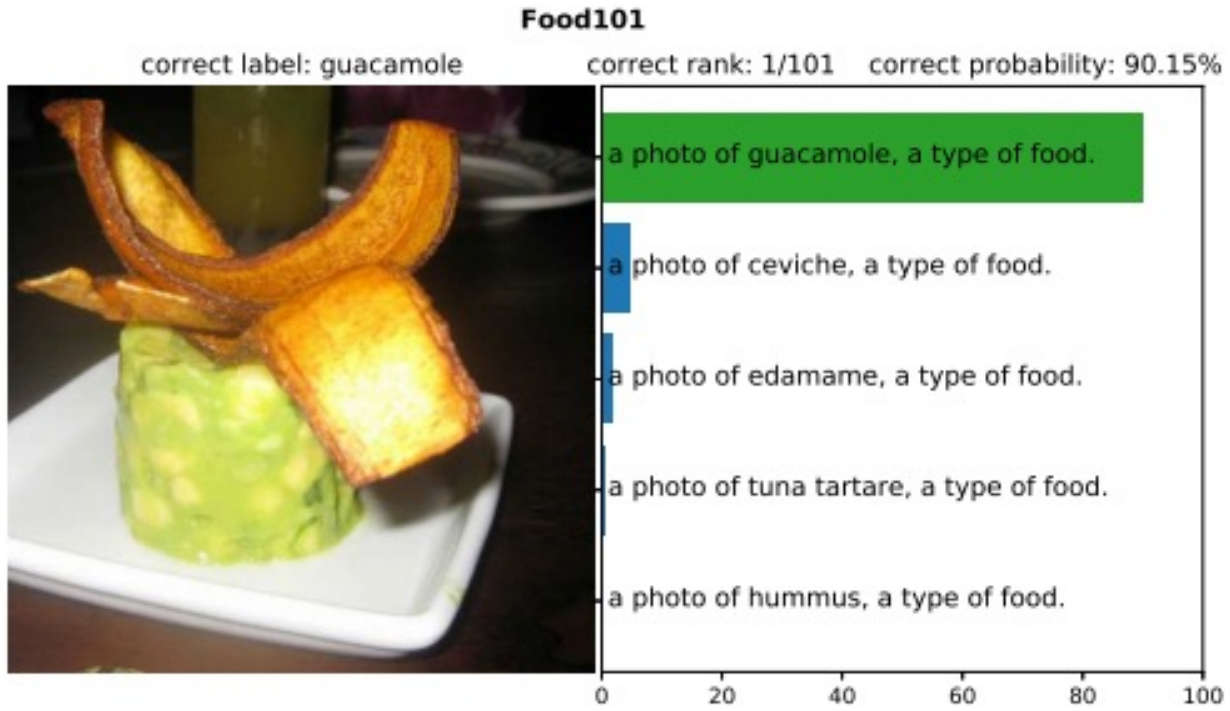
CLIP Evaluation

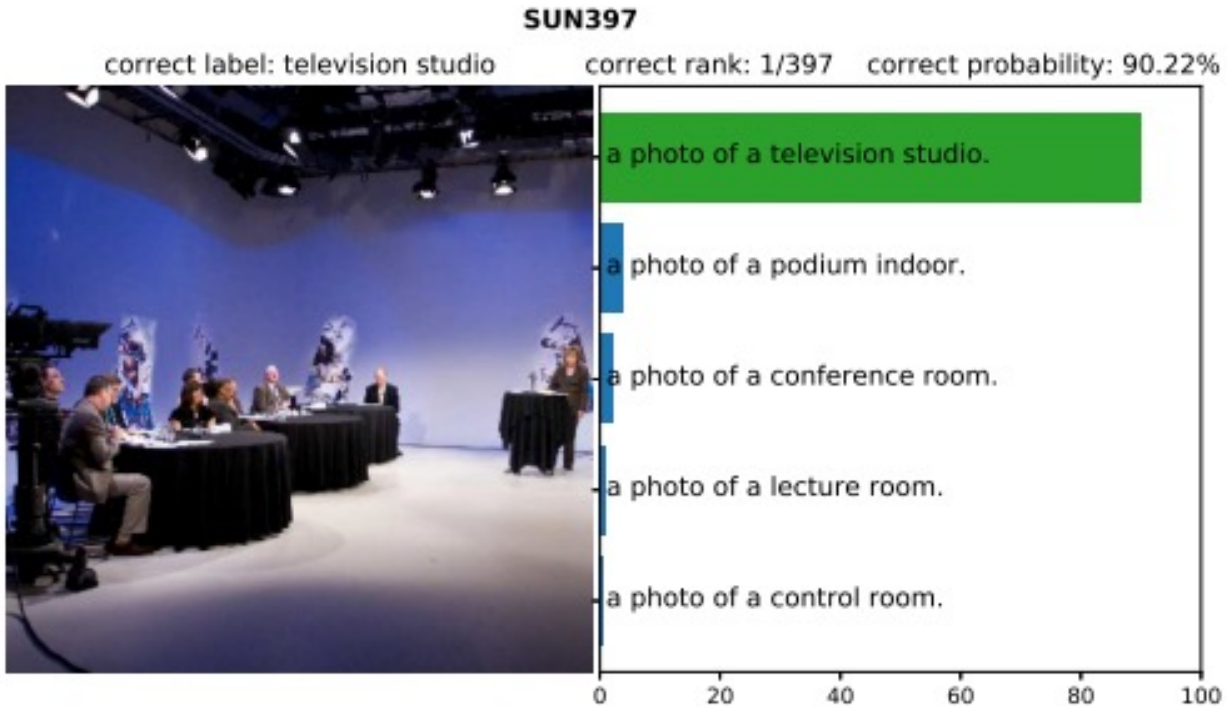
Subset of datasets shown here:

Classification evaluation spanned fine-grained classification (e.g., food, bird, aircraft, and car categories), distribution shifts for ImageNet categories (e.g., corrupted images), and more

Dataset	Classes	Train size	Test size	Evaluation metric
Food-101	102	75,750	25,250	accuracy
CIFAR-10	10	50,000	10,000	accuracy
CIFAR-100	100	50,000	10,000	accuracy
Birdsnap	500	42,283	2,149	accuracy
SUN397	397	19,850	19,850	accuracy
Stanford Cars	196	8,144	8,041	accuracy
FGVC Aircraft	100	6,667	3,333	mean per class
Pascal VOC 2007 Classification	20	5,011	4,952	11-point mAP
Describable Textures	47	3,760	1,880	accuracy
Oxford-IIIT Pets	37	3,680	3,669	mean per class
Caltech-101	102	3,060	6,085	mean-per-class
Oxford Flowers 102	102	2,040	6,149	mean per class
MNIST	10	60,000	10,000	accuracy
Facial Emotion Recognition 2013	8	32,140	3,574	accuracy
STL-10	10	1000	8000	accuracy
EuroSAT	10	10,000	5,000	accuracy
RESISC45	45	3,150	25,200	accuracy
GTSRB	43	26,640	12,630	accuracy
KITTI	4	6,770	711	accuracy
Country211	211	43,200	21,100	accuracy
PatchCamelyon	2	294,912	32,768	accuracy
UCF101	101	9,537	1,794	accuracy
Kinetics700	700	494,801	31,669	mean(top1, top5)
CLEVR Counts	8	2,000	500	accuracy
Hateful Memes	2	8,500	500	ROC AUC
Rendered SST2	2	7,792	1,821	accuracy
ImageNet	1000	1,281,167	50,000	accuracy

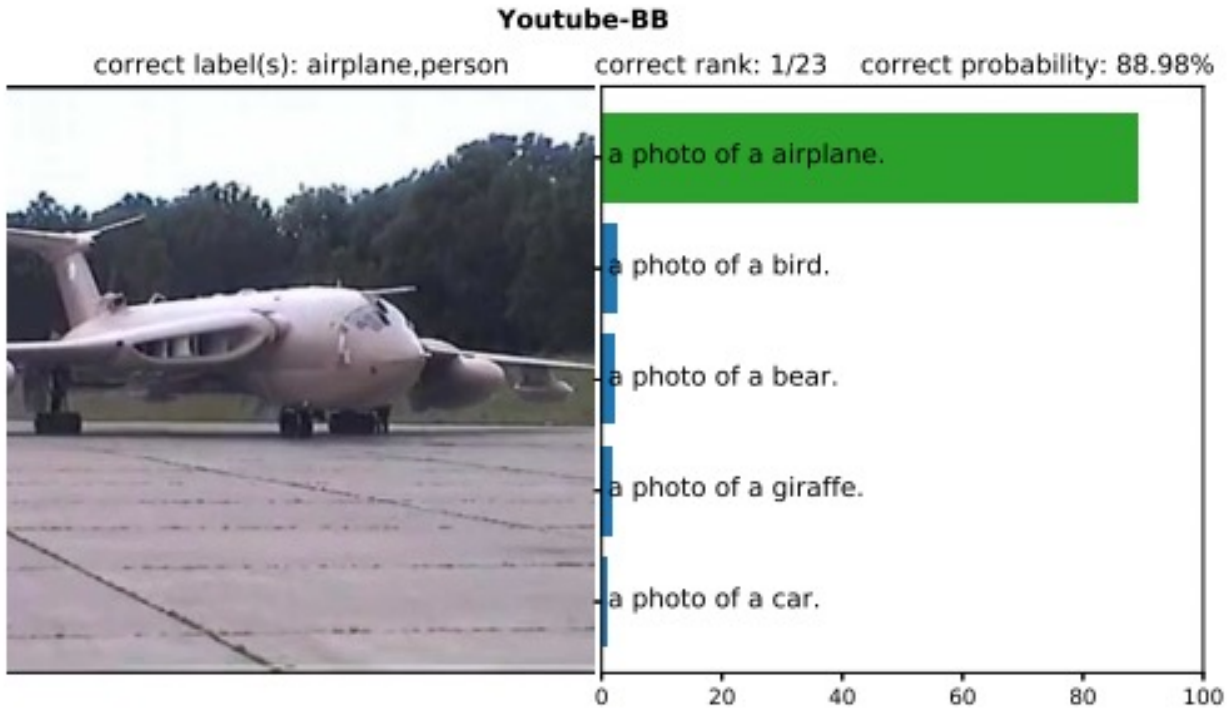
CLIP: Qualitative Results

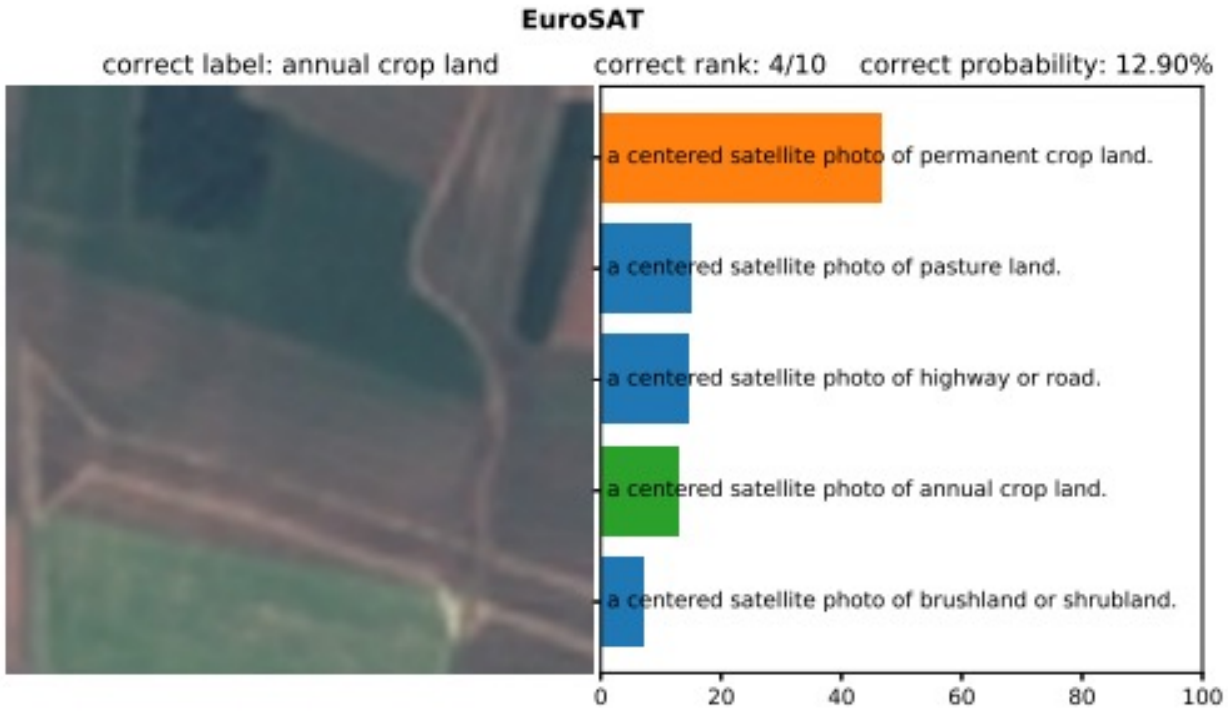




Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.

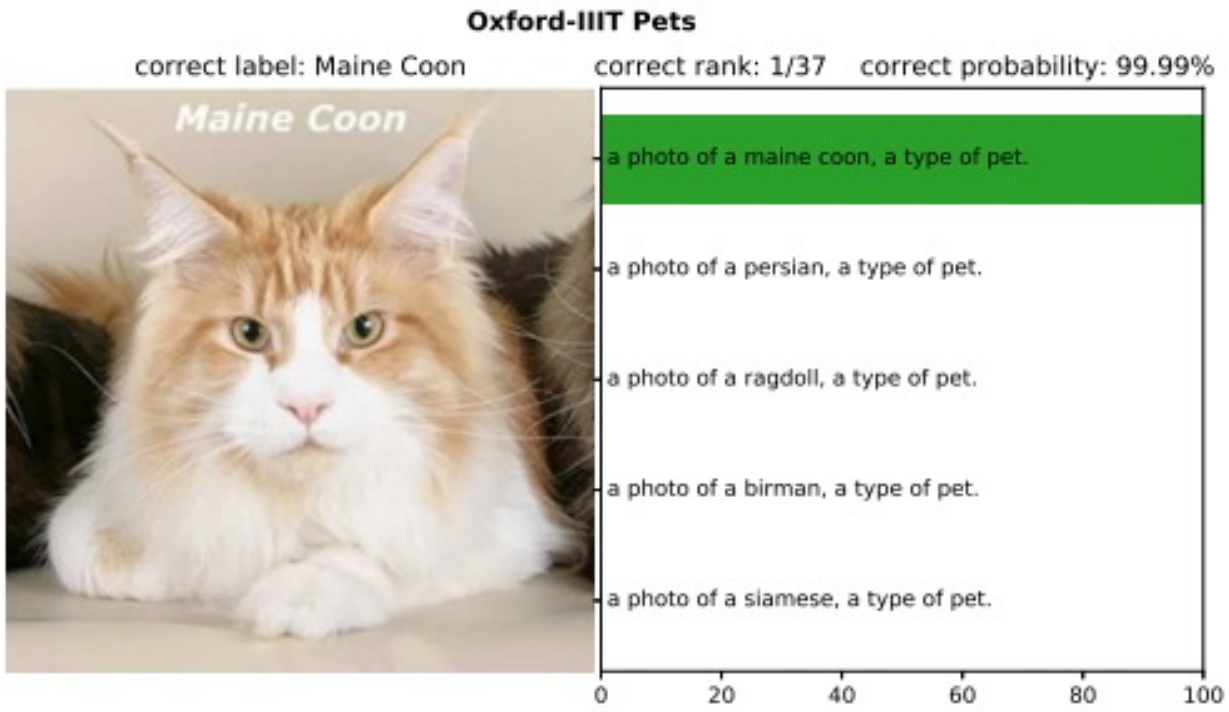
CLIP: Qualitative Results

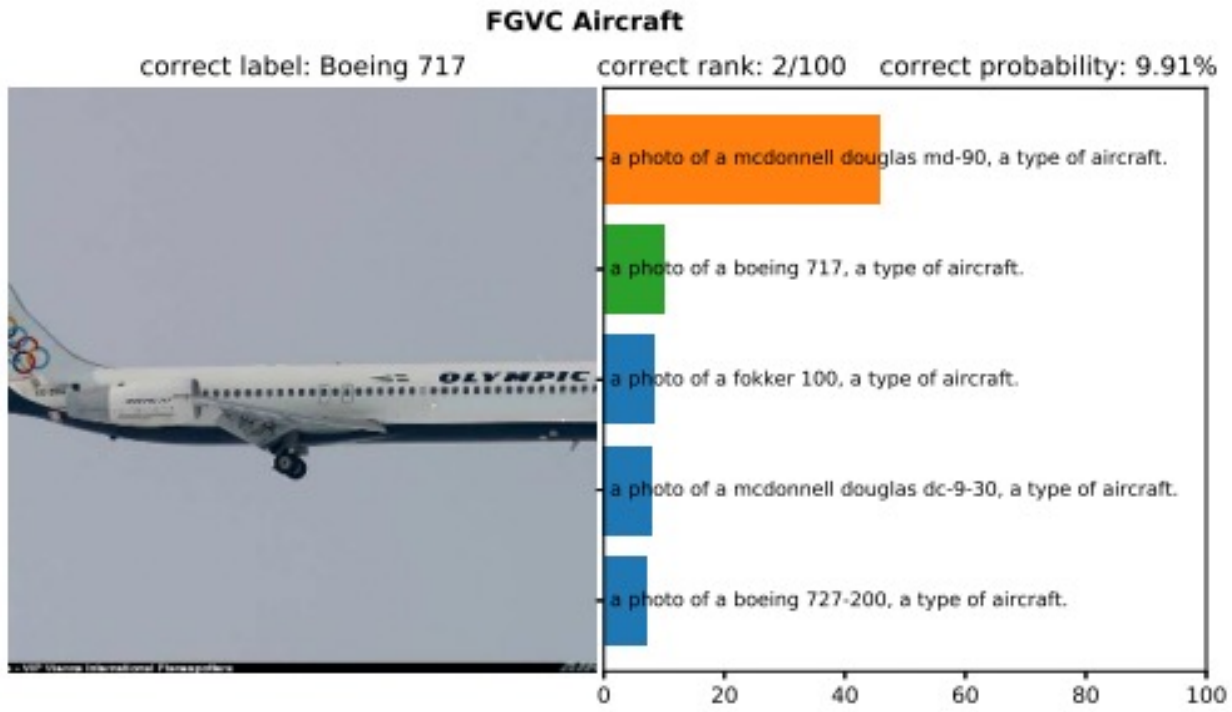




Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.

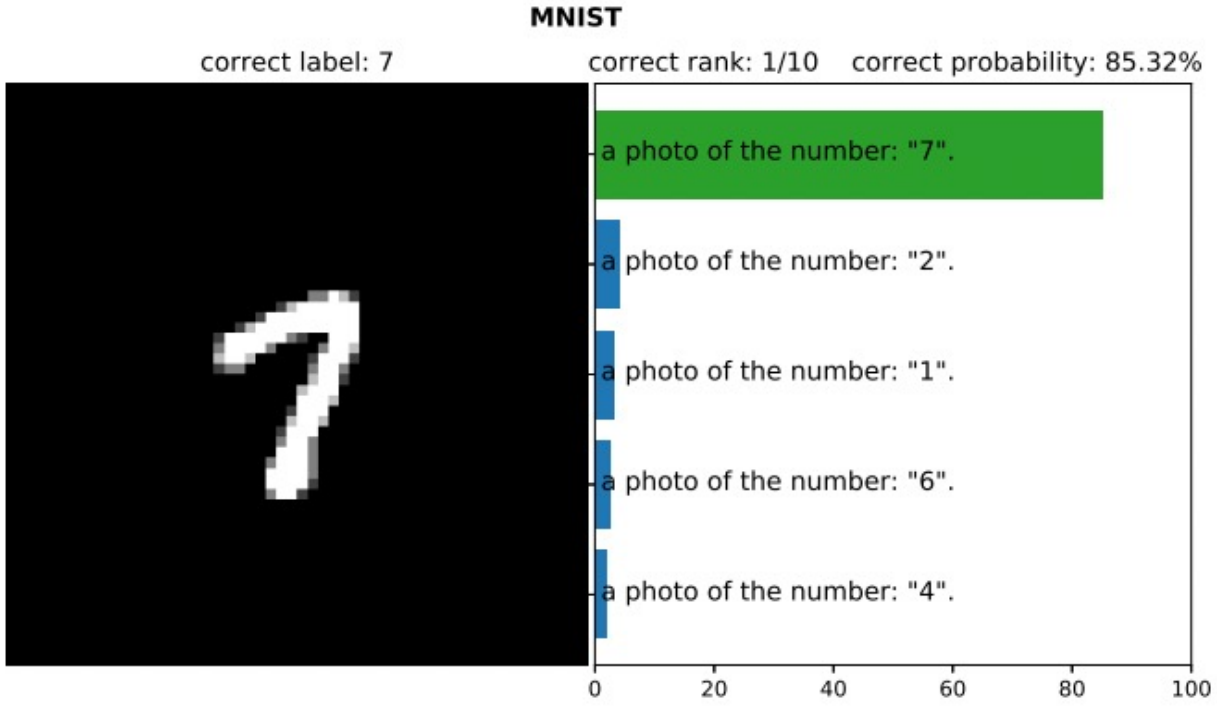
CLIP: Qualitative Results

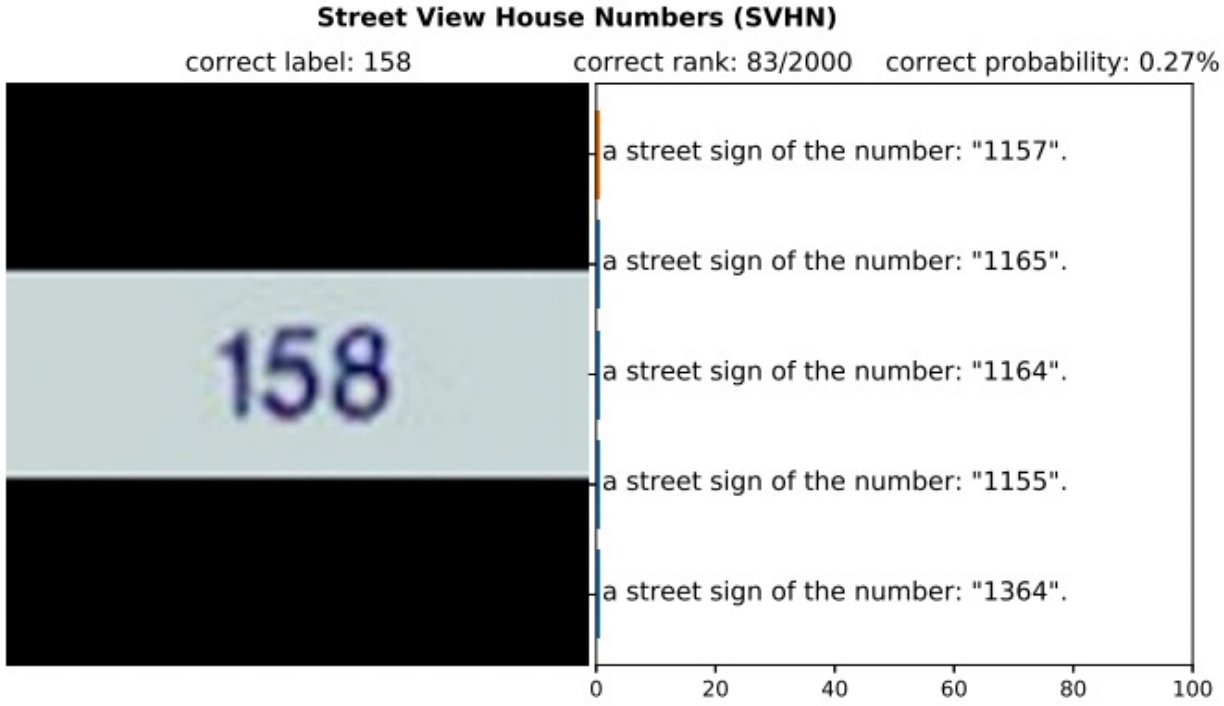




Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.

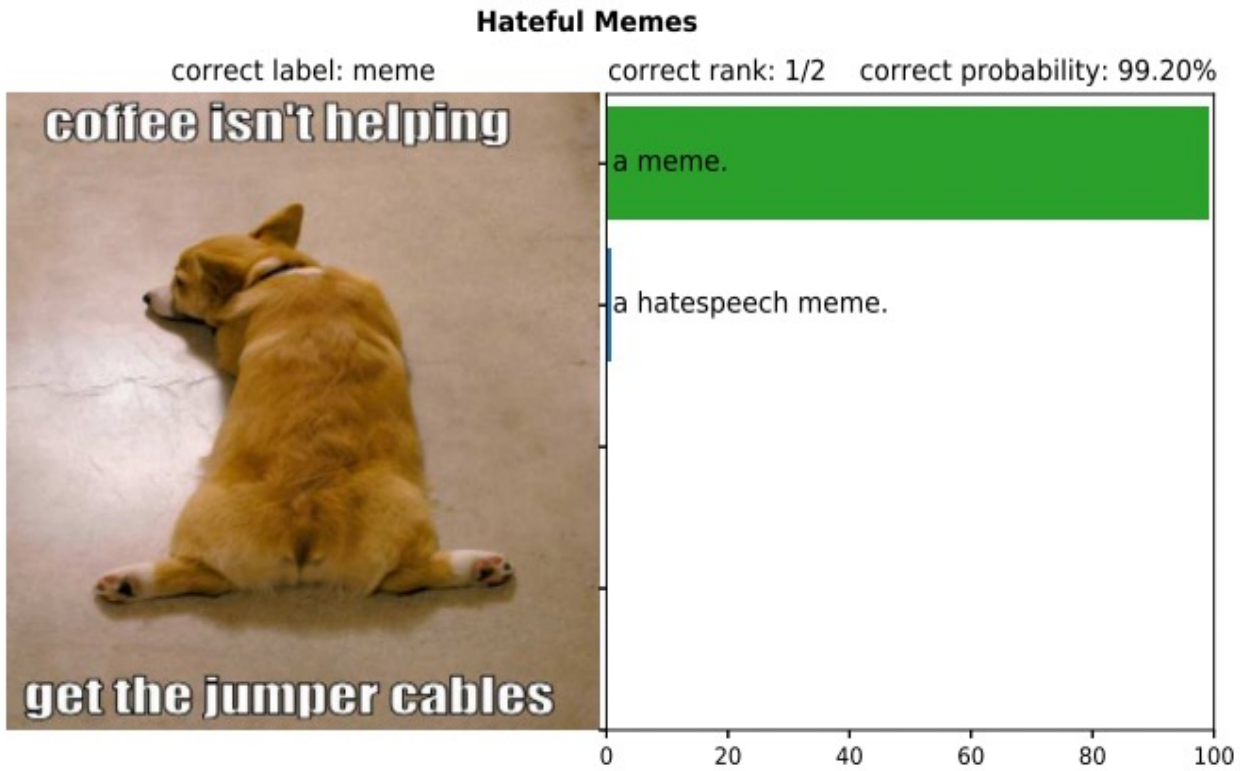
CLIP: Qualitative Results

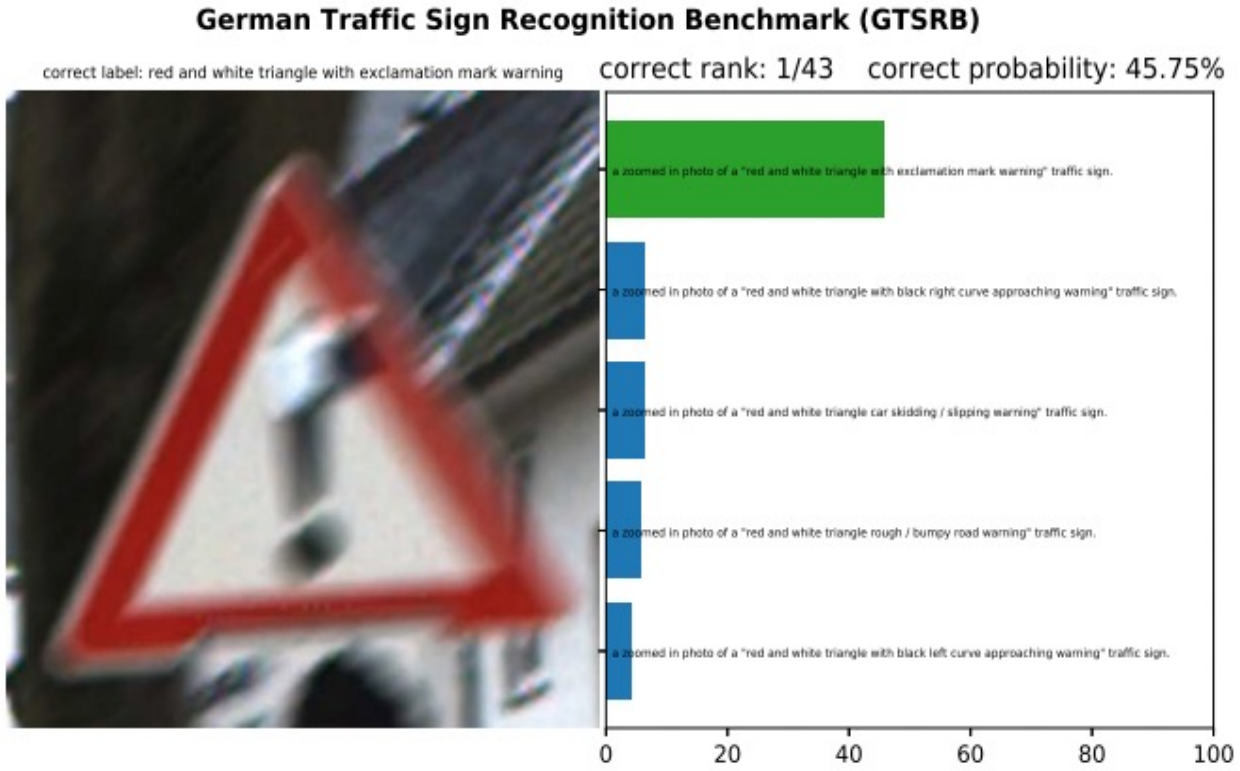




Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.

CLIP: Qualitative Results



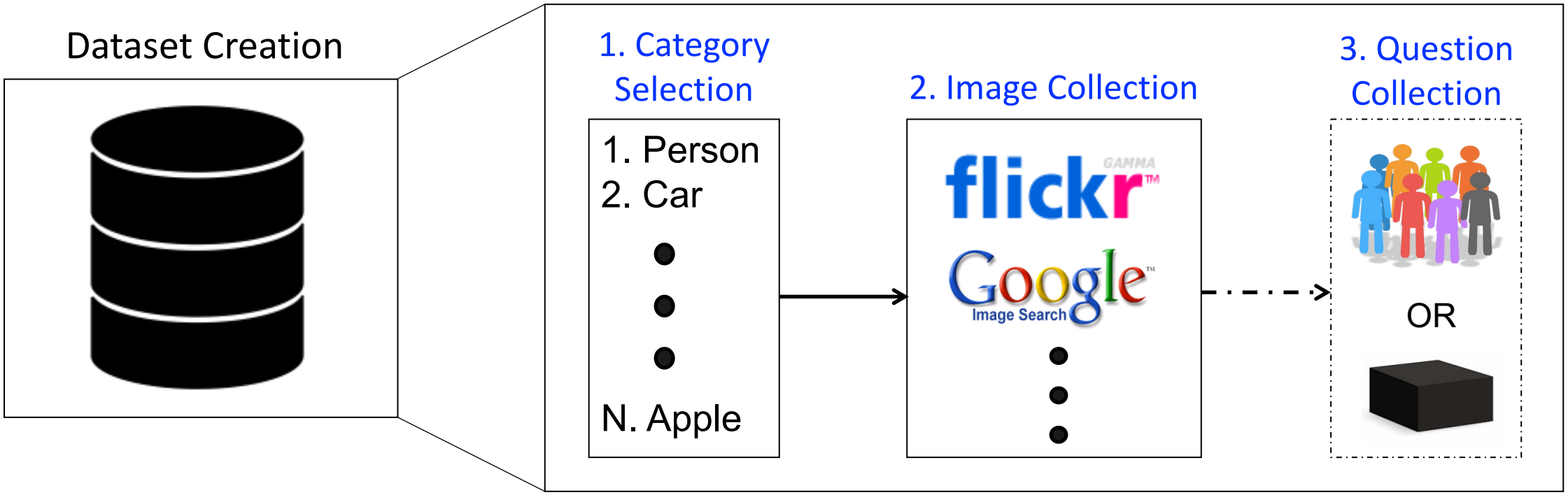


Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.

Today’s Topics

- Multimodal applications
- Image captioning dataset challenges
- Image captioning algorithms
- Visual question answering dataset challenges
- Visual question answering algorithms
- Foundation models

Status Quo (Approach to Create 14+ Datasets)



Constrained set of concepts

Pre-qualified images (quality, privacy)

Contrived Questions

e.g., Question Generation

Stump a smart robot! Ask a question about this scene that a human can answer, but a smart robot probably can't!

Updated instructions: Please read carefully

Hide Show

We have built a smart robot. It understands a lot about scenes. It can recognize and name all the objects, it knows where the objects are, it can recognize the scene type (e.g., kitchen, beach), people's expressions and poses, and properties of objects (e.g., the color of objects, their texture). Your task is to stump this smart robot! In particular, it already knows answers to some questions about this scene. We will tell you what these questions are.

Ask a question about this scene that this SMART robot probably can not answer, but any human can easily answer while looking at the scene in the image. IMPORTANT: The question should be about this scene. That is, the human should need the image to be able to answer the question -- the human should not be able to answer the question without looking at the image.

Your work will get rejected if you do not follow the instructions below:

- Do not ask questions that are similar to the ones listed below each image. As mentioned, the robot already knows the answers to those questions for the scene in this image. Please ask about something different.
- Do not repeat questions. Do not ask the same questions or the same questions with minor variations over and over again across images. Think of a new question each time specific to the scene in each image.
- Each question should be a single question. Do not ask questions that have multiple parts or multiple sub-questions in them.
- Do not ask generic questions that can be asked of many other scenes. Ask questions specific to the scene in each image.

Below is a list of questions the smart robot can already answer. Please ask a different question about this scene that a human can answer "if" looking at the scene in the image (and not otherwise), but would stump this smart robot:

Q1: What is unusual about this mustache? (The robot already knows the answer to this question.)

Q2: What is her facial expression? (The robot already knows the answer to this question.)

Q3: Write your question, different from the questions above, here to stump this smart robot.

Agrawal et al. VQA: Visual Question Answering. CVPR 2015.

e.g., Answer Generation

Help Us Answer Questions About Images!

Updated instructions: Please read carefully

Hide Show

Please answer some questions about images **with brief answers**. Your answers should be how most other people would answer the questions. If the question doesn't make sense, please try your best to answer it and indicate via the buttons that you are unsure of your response.

If you don't follow the following instructions, your work will be rejected.

Your work will get rejected if you do not follow the instructions below:

- Answer the question based on what is going on in the scene depicted in the image.
- Your answer should be a brief phrase (not a complete sentence).
  - "It is a kitchen." -> "kitchen"
- For yes/no questions, please just say yes/no.
  - "You bet it is!" -> "yes"
- For numerical answers, please use digits.
  - "Ten." -> "10"
- If you need to speculate (e.g., "What just happened?"), provide an answer that most people would agree on.
- If you don't know the answer (e.g., specific dog breed), provide your best guess.
- Respond matter-of-factly and avoid using conversational language or inserting your opinion.

10 answers collected from 10 crowdworkers

Please answer the question using as few words as possible:

Q1: What is unusual about this mustache?

A1: Write your answer here.

Do you think you were able to answer the question correctly?
(Clicking an option will take you to the next question.)

no maybe yes **Page 1/2**

Agrawal et al. VQA: Visual Question Answering. CVPR 2015.

Mainstream VQA Challenge (held for 6 years)

VQA Visual Question Answering

VirginiaTech Invent the Future

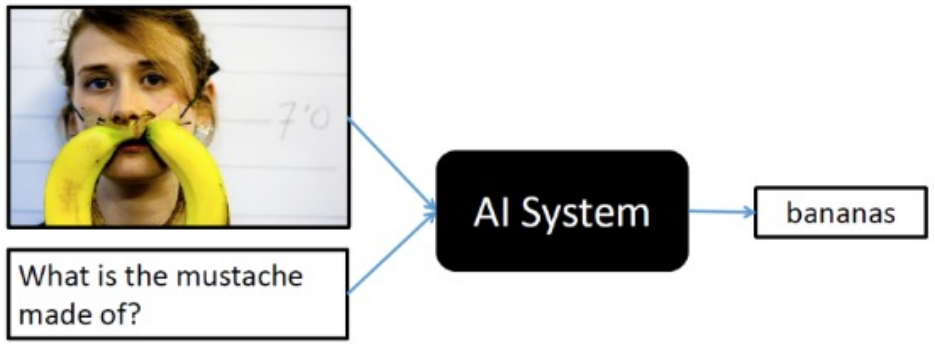
Georgia Tech

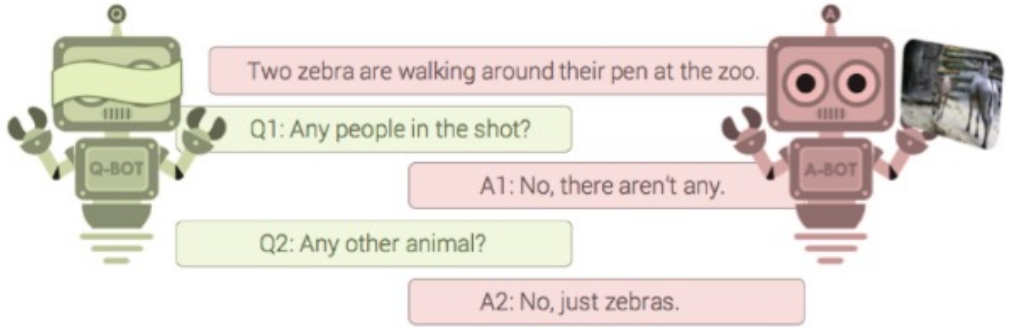
Home | People | Code | Demo | Download | Evaluation | Challenge | Browse | Visualize | Workshop | Sponsors | Terms | External

Visual Question Answering and Dialog Workshop

Location: **Seaside Ballroom B, Long Beach Convention & Entertainment Center**

at CVPR 2019, June 17, Long Beach, California, USA



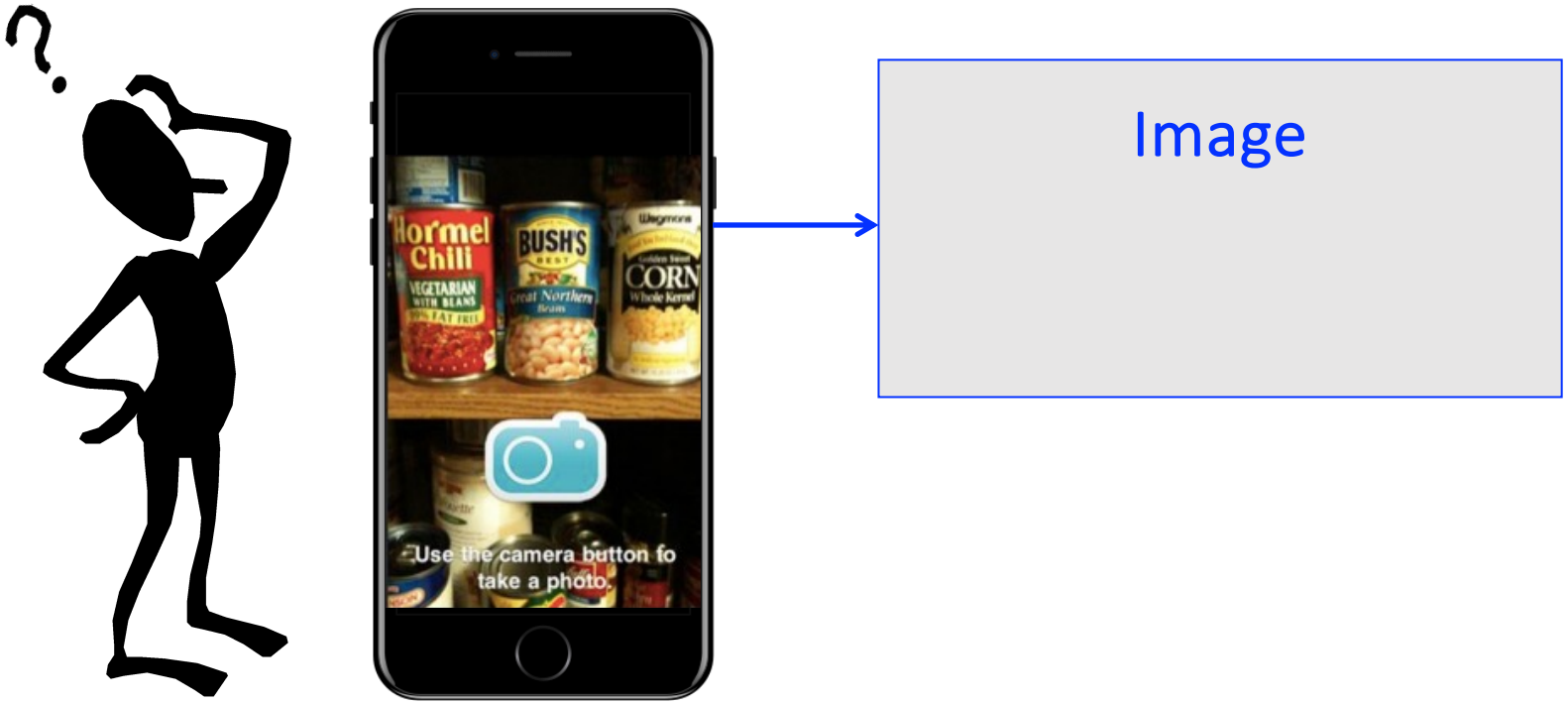


https://visualqa.org/workshop.html

VizWiz: Authentic Use Case

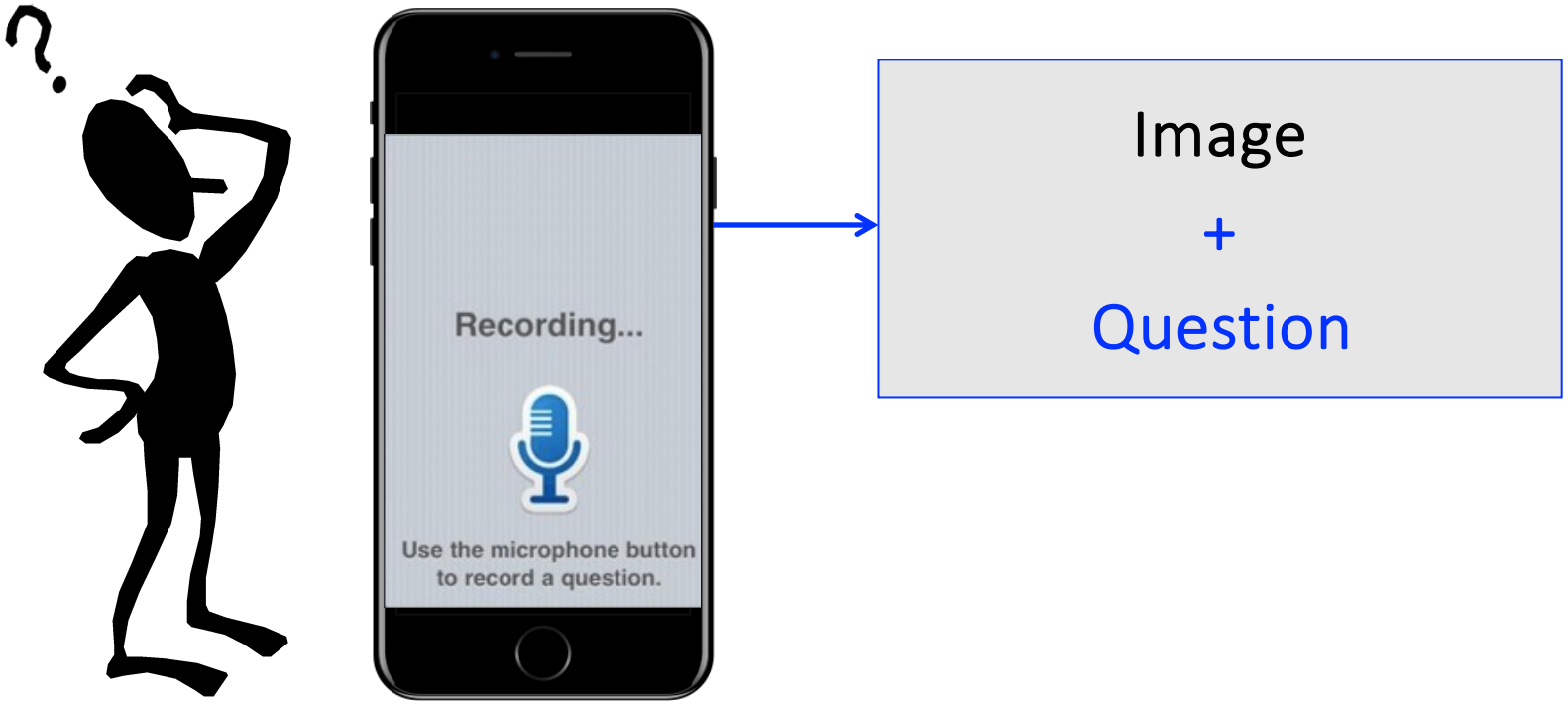
Jeffrey Bigham et al. VizWiz: Nearly Real-time Answers to Visual Questions. UIST 2010

VizWiz: Authentic Use Case



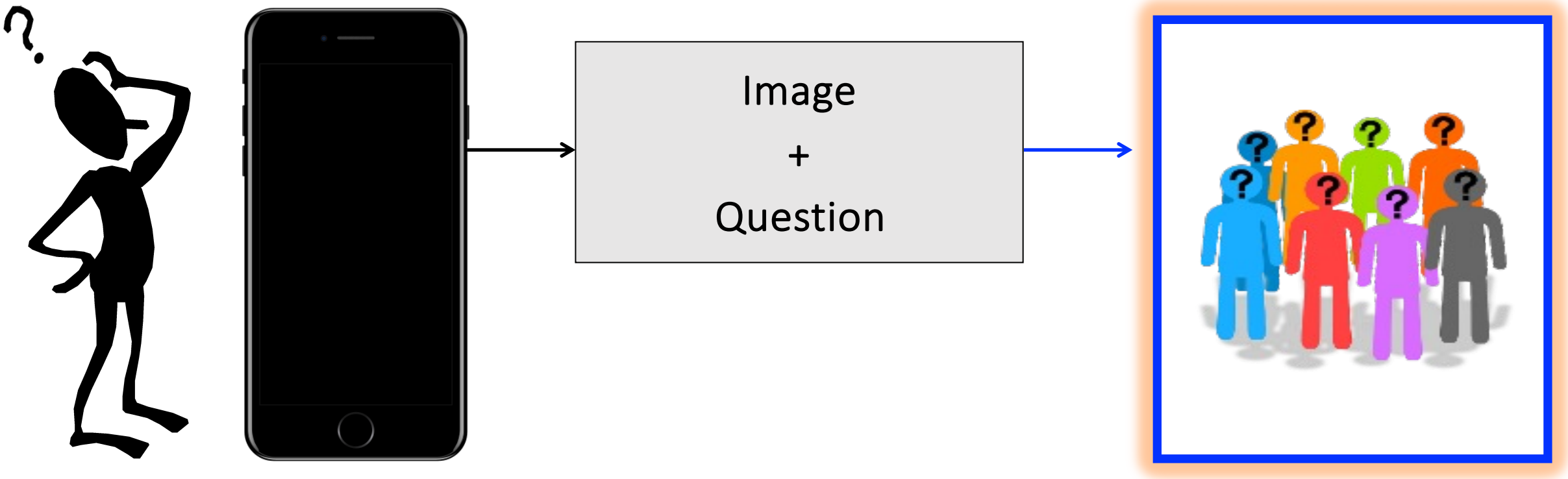
Jeffrey Bigham et al. VizWiz: Nearly Real-time Answers to Visual Questions. UIST 2010

VizWiz: Authentic Use Case



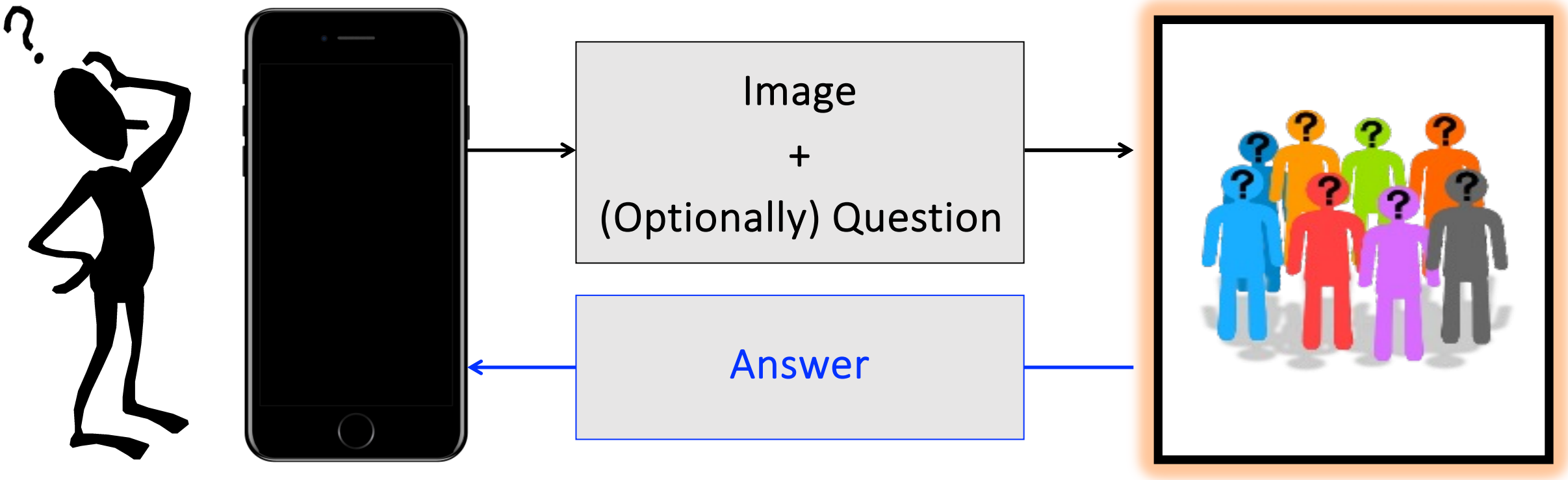
Jeffrey Bigham et al. VizWiz: Nearly Real-time Answers to Visual Questions. UIST 2010

VizWiz: Authentic Use Case



Jeffrey Bigham et al. VizWiz: Nearly Real-time Answers to Visual Questions. UIST 2010

VizWiz: Authentic Use Case



Jeffrey Bigham et al. VizWiz: Nearly Real-time Answers to Visual Questions. UIST 2010

VizWiz: Authentic Use Case

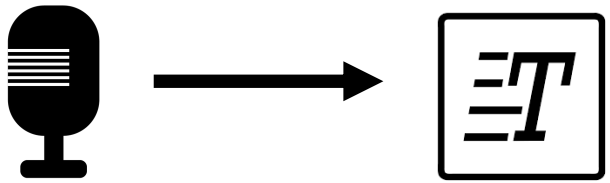
Users agreed to share 44,799 (62%) of requests for dataset creation

Jeffrey Bigham et al. VizWiz: Nearly Real-time Answers to Visual Questions. UIST 2010

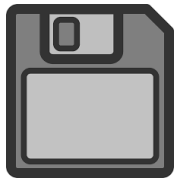
VizWiz: Authentic Use Case

Anonymization

1. Transcribe questions (removes voice)



2. Re-save images (removes metadata)

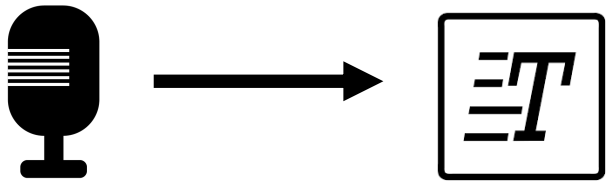


Gurari et al. CVPR 2018

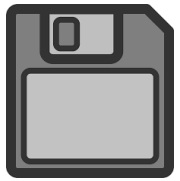
VizWiz: Authentic Use Case

Anonymization

1. Transcribe questions

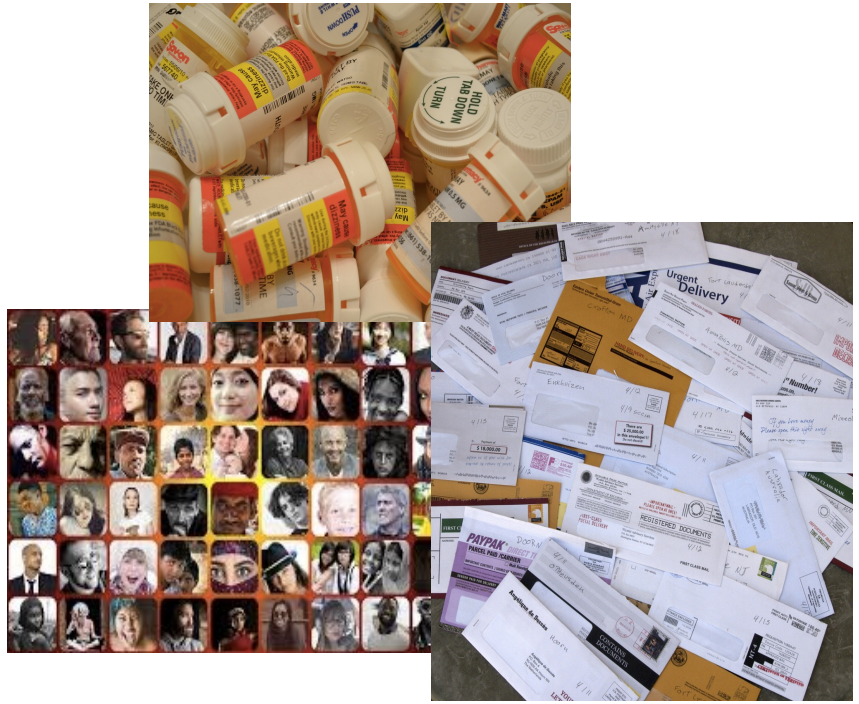


2. Re-save images



In-House Filtering

(personally identifying information)

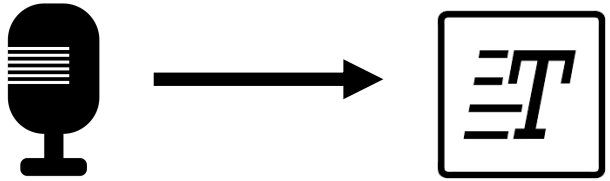


Gurari et al. CVPR 2018

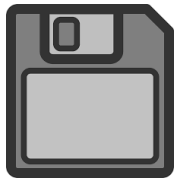
VizWiz: Authentic Use Case

Anonymization

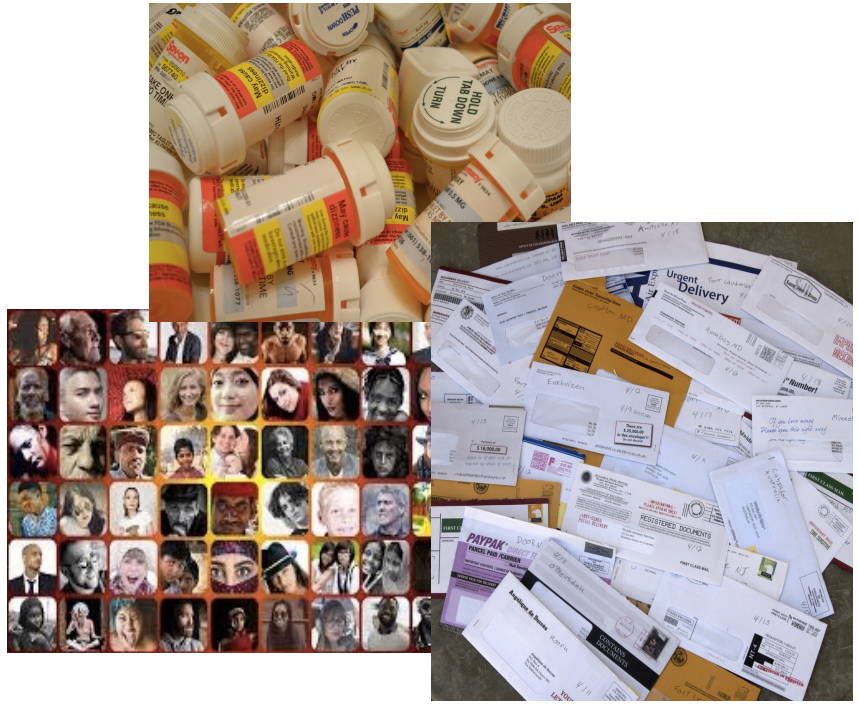
1. Transcribe questions



2. Re-save images



In-House Filtering



Data Labeling

(high quality answers)

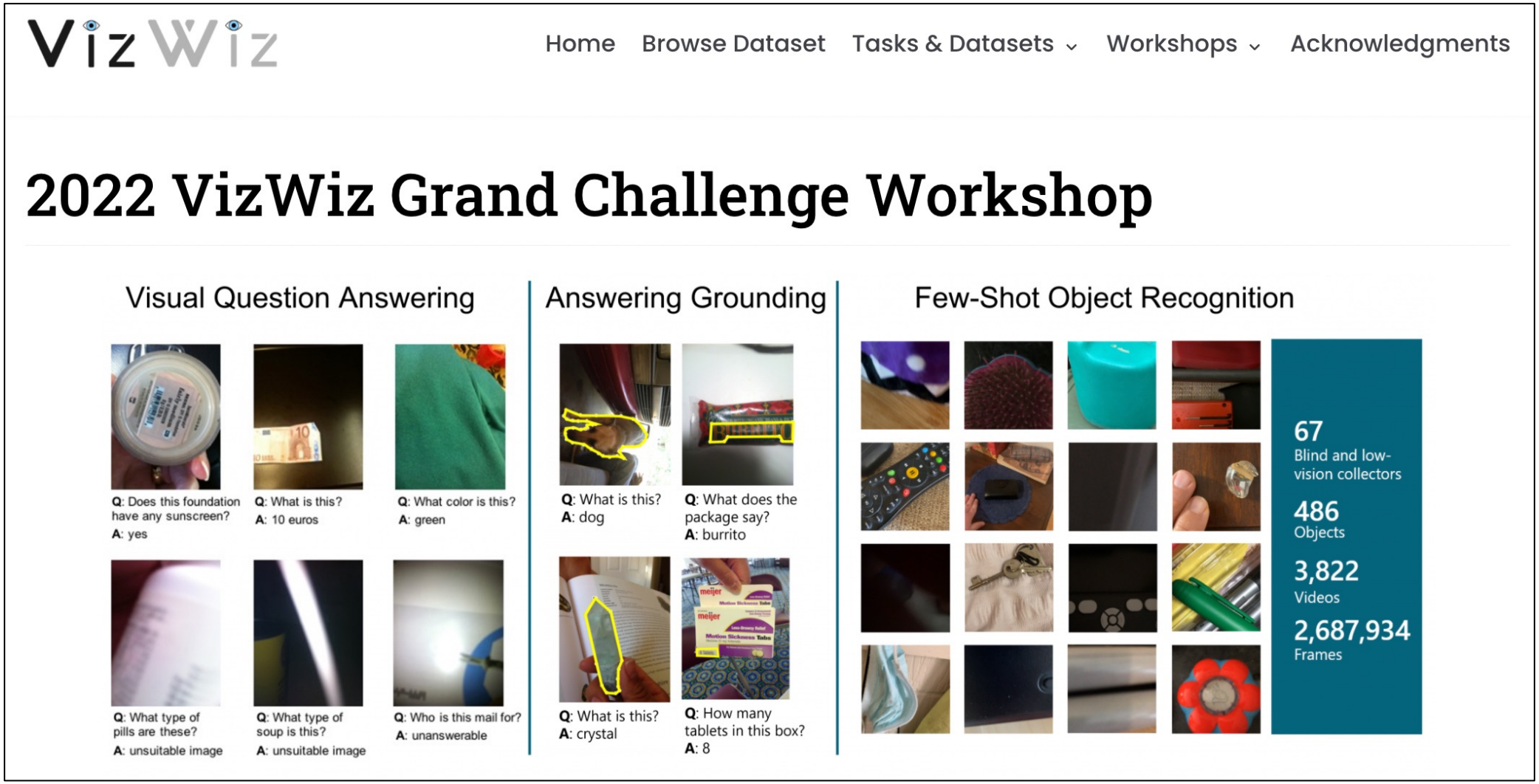
Gurari et al. CVPR 2018

VizWiz: Authentic Use Case

VQA: 32,842 image/question pairs → 328,420 answers

Gurari et al. CVPR 2018

VizWiz-VQA Grand Challenge (5th year in 2023)

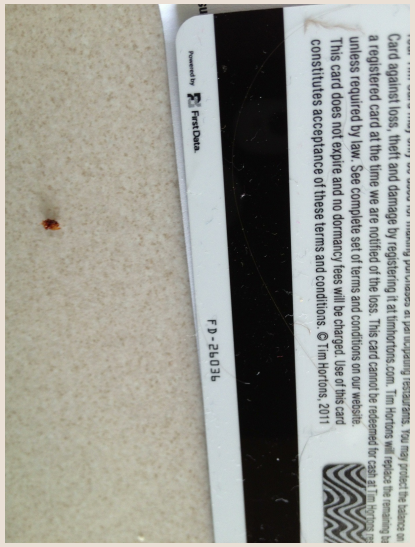


https://vizwiz.org

Key Difference of Real-World Use Case from Status Quo: VQs Can Be Unanswerable!

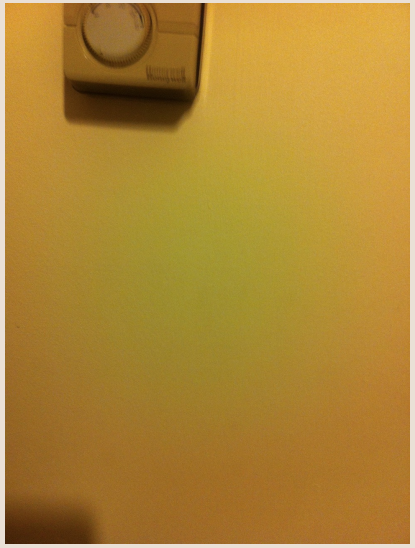
Q: What is the expiration date?

A: unanswerable



Q: What is this a gift card for?

A: unanswerable

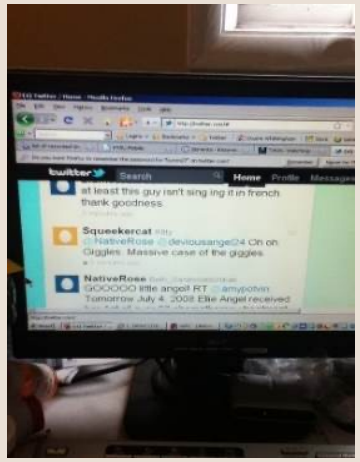


Q: What temperature is the dial set to?

A: unanswerable

[Gurari et al. CVPR 2018]

Class Task: Answer Visual Question (https://forms.gle/A58M1gm23QMckZBq6)



Is my monitor on?

(1)

Hi there can you please tell me what flavor this is?

(2)

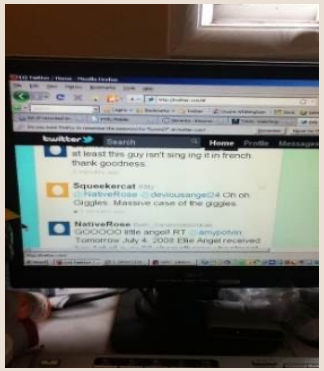
Does this picture look scary?

(3)

Which side of the room is the toilet on?

(4)

Crowdsourced Answers



Is my monitor on?

(1) yes
(2) yes
(3) yes
(4) yes
(5) yes
(6) yes
(7) yes
(8) yes
(9) yes
(10) yes

Hi there can you please tell me what flavor this is?

(1) sweet pepper
(2) sweet pepper
(3) sweet pepper
(4) sweet pepper
(5) sweet pepper
(6) sweet pepper
(7) sweet pepper
(8) sweet pepper
(9) sweet pepper
(10) sweet pepper

Does this picture look scary?

(1) yes
(2) no
(3) no
(4) yes
(5) no
(6) yes
(7) yes
(8) no
(9) no
(10) no

Which side of the room is the toilet on?

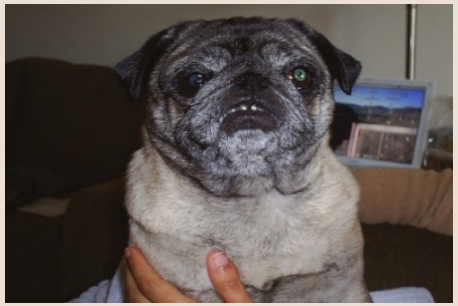
(1) right
(2) left
(3) right
(4) right
(5) right
(6) right
(7) right side
(8) right
(9) center
(10) right

Evaluating Automated Predictions: Basic Equation

$$\text{accuracy} = \min\left(\frac{\#\ \text{humans that provided that answer}}{3}, 1\right)$$

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. VQA: Visual Question Answering. CVPR 2015.

Evaluating Automated Predictions: Example



Does this picture look scary?

(1) yes
(2) no
(3) no
(4) yes
(5) no
(6) yes
(7) yes
(8) no
(9) no
(10) no

What is the accuracy of an algorithm prediction of

- "yes"?
- "no"?
- "maybe"?

$$\text{accuracy} = \min\left(\frac{\text{\# humans that provided that answer}}{3}, 1\right)$$

Evaluating Automated Predictions: Example

Which side of the room is the toilet on?

(1) right
(2) left
(3) right
(4) right
(5) right
(6) right
(7) right side
(8) right
(9) center
(10) right

What is the accuracy of an algorithm prediction of

- "right"?
- "left"?
- "right side"?
- "center"?
- "bottom"?

$$\text{accuracy} = \min\left(\frac{\#\text{ humans that provided that answer}}{3}, 1\right)$$

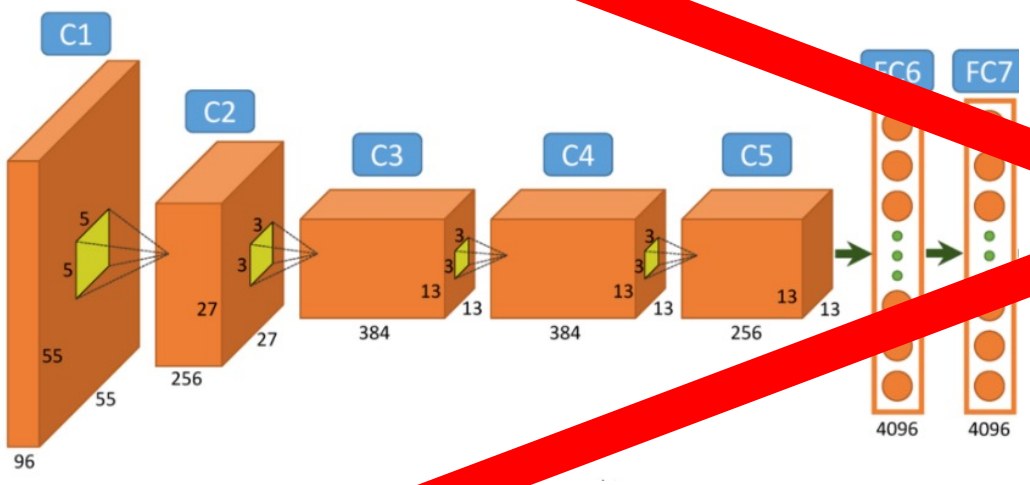
Implementation detail: to enable fair comparison to humans, compute 10 rounds of comparing a prediction with each possible set of 9 human-supplied answers

Today’s Topics

- Multimodal applications
- Image captioning dataset challenges
- Image captioning algorithms
- Visual question answering dataset challenges
- Visual question answering algorithms
- Foundation models

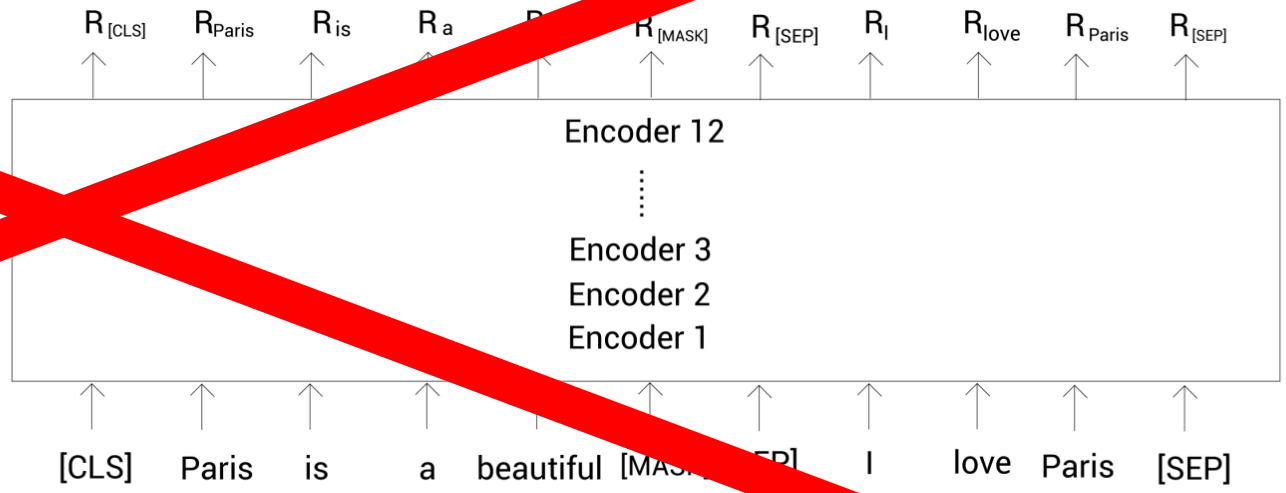
Key Idea: Multimodal Representation Rather Than Single Modality Representations

e.g., visi[illegible] representation with AlexNet



https://www.resea[illegible]te.net/figure/Architecture-of-Alexnet-From-left-to-righ[illegible]t-to-output-five-convolutional-layers_fig2_312303454

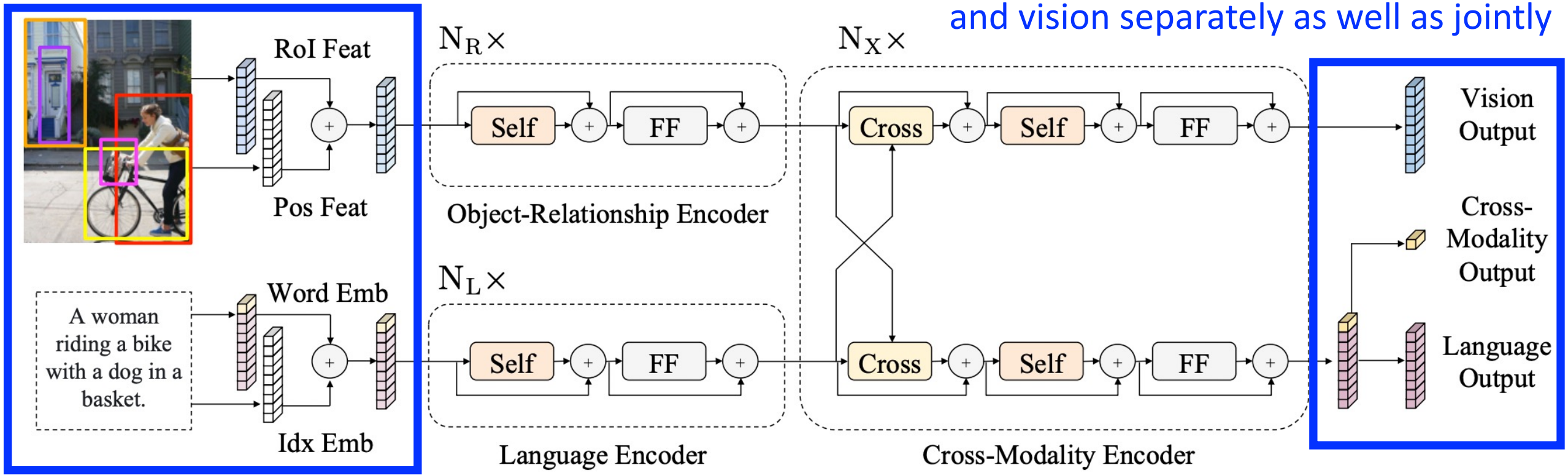
e.g., language representation wi[illegible] BERT



https://static.packt-cdn.com/downloads/9781838821[illegible]olorImages.pdf

LXMERT: Learning Cross-Modality Encoder Representations from Transformers

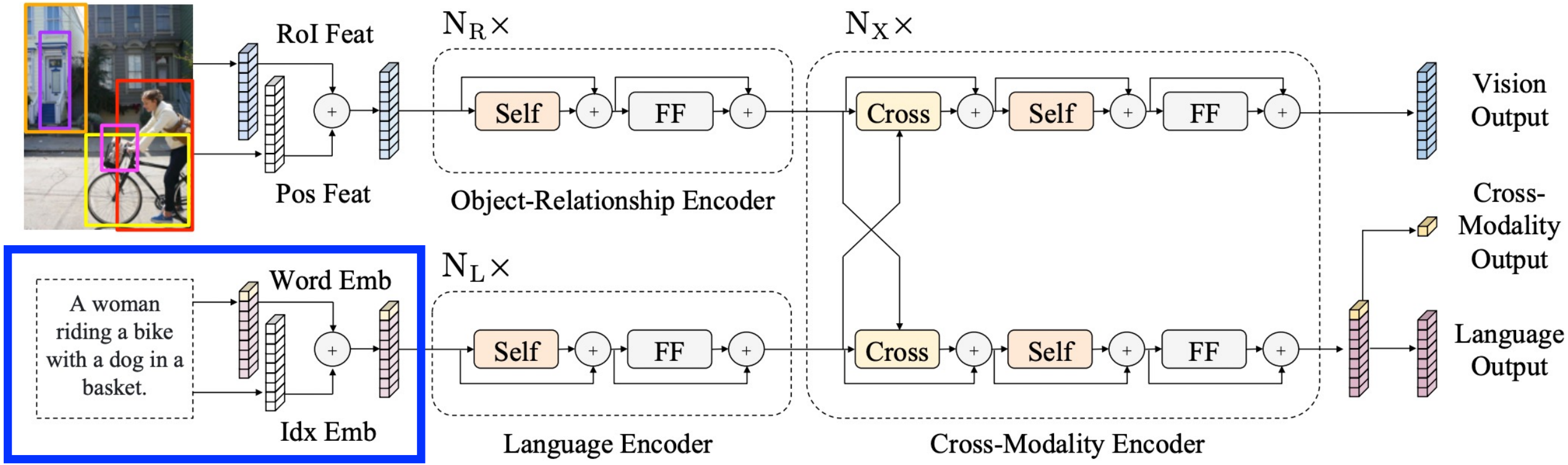
Generates representations for image and vision separately as well as jointly



Pretrains using language and vision input

Tan and Bansal, EMNLP 2019.

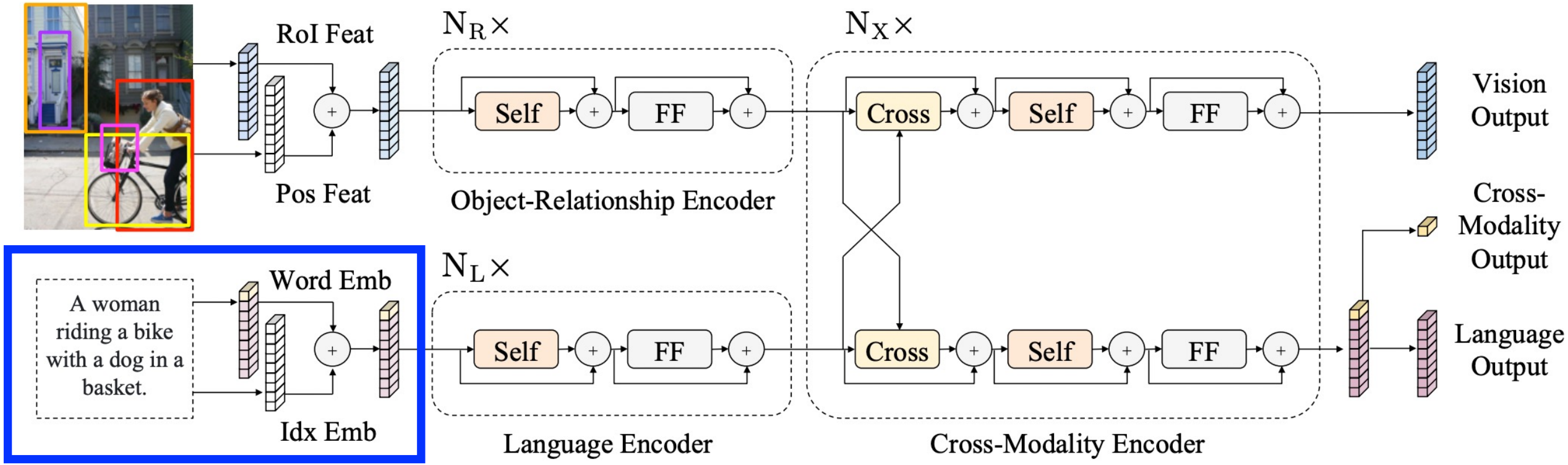
LXMERT: Language Input



[CLS] is added to the start of the sequence

Tan and Bansal, EMNLP 2019.

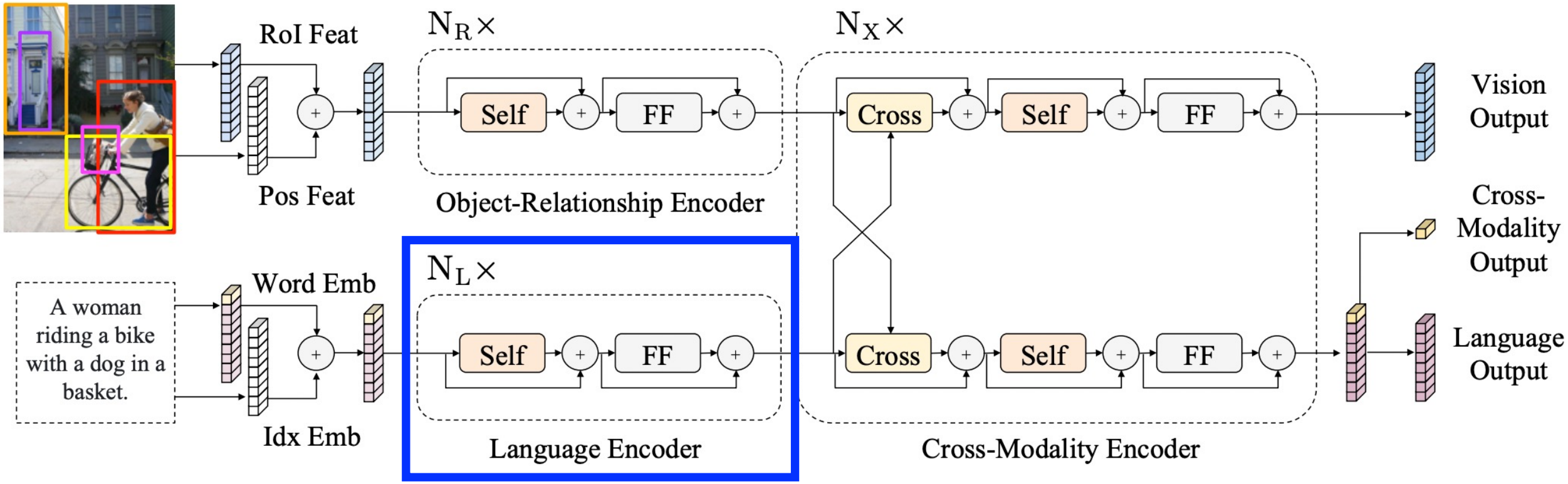
LXMERT: Language Input



Each word is represented as sum of its word embedding and position encoding

Tan and Bansal, EMNLP 2019.

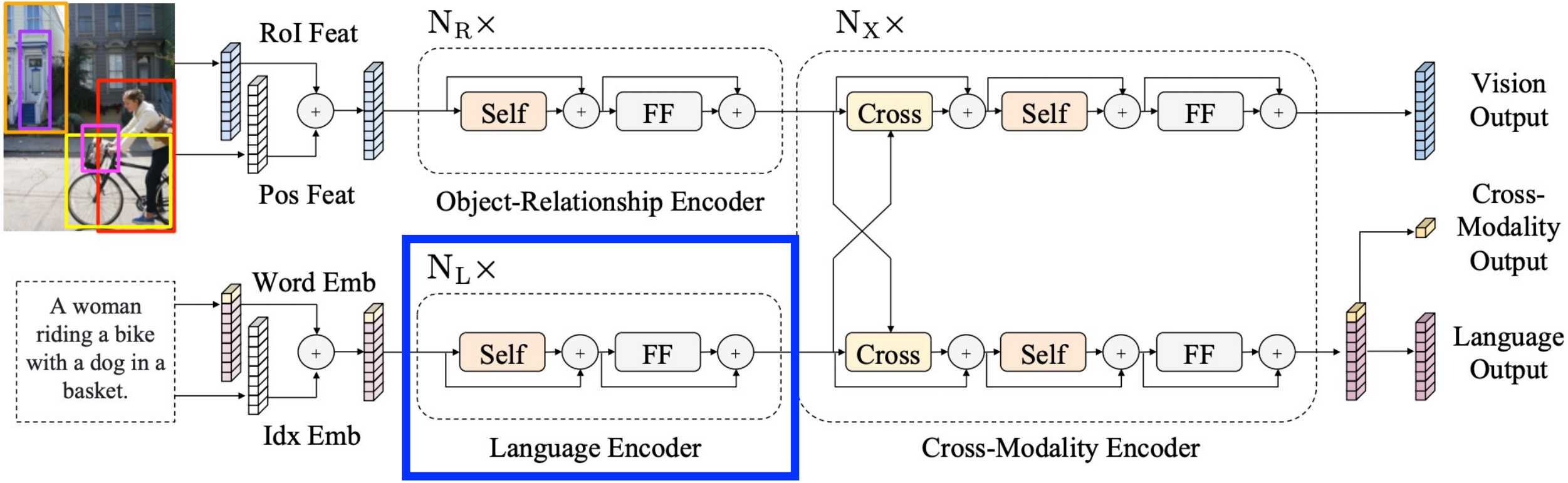
LXMERT: Language Input



Transformer encoder (i.e., BERT);
what does its output represent?

Tan and Bansal, EMNLP 2019.

LXMERT: Language Input

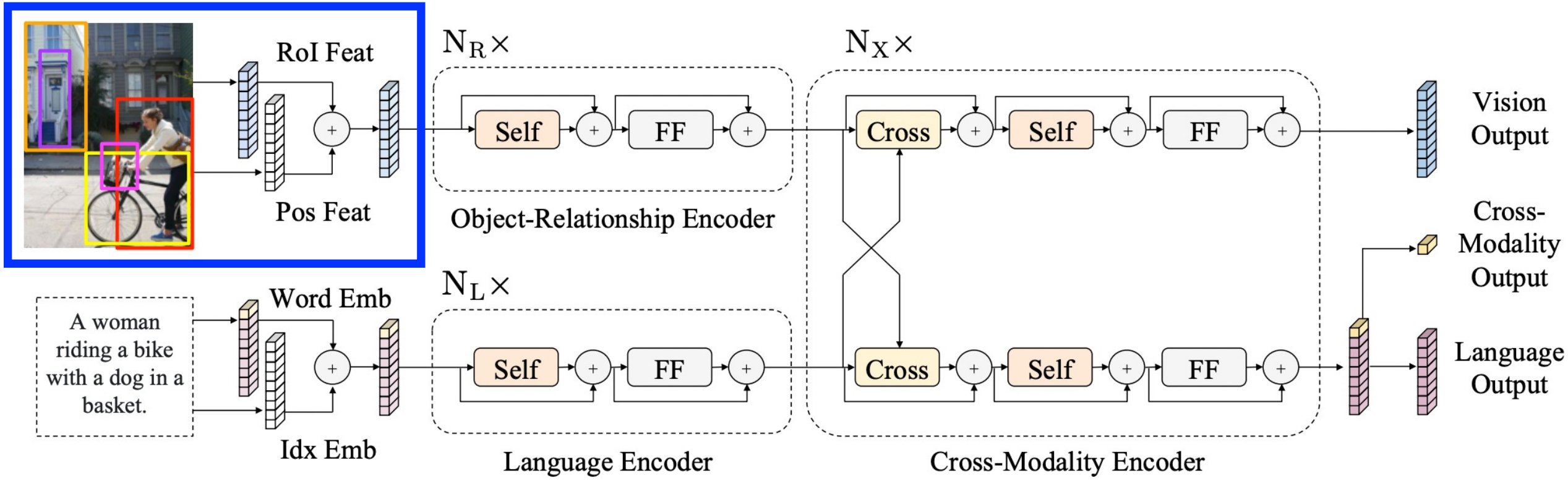


Transformer encoder (i.e., BERT); represents words with their relationships to all words

Tan and Bansal, EMNLP 2019.

LXMERT: Vision Input

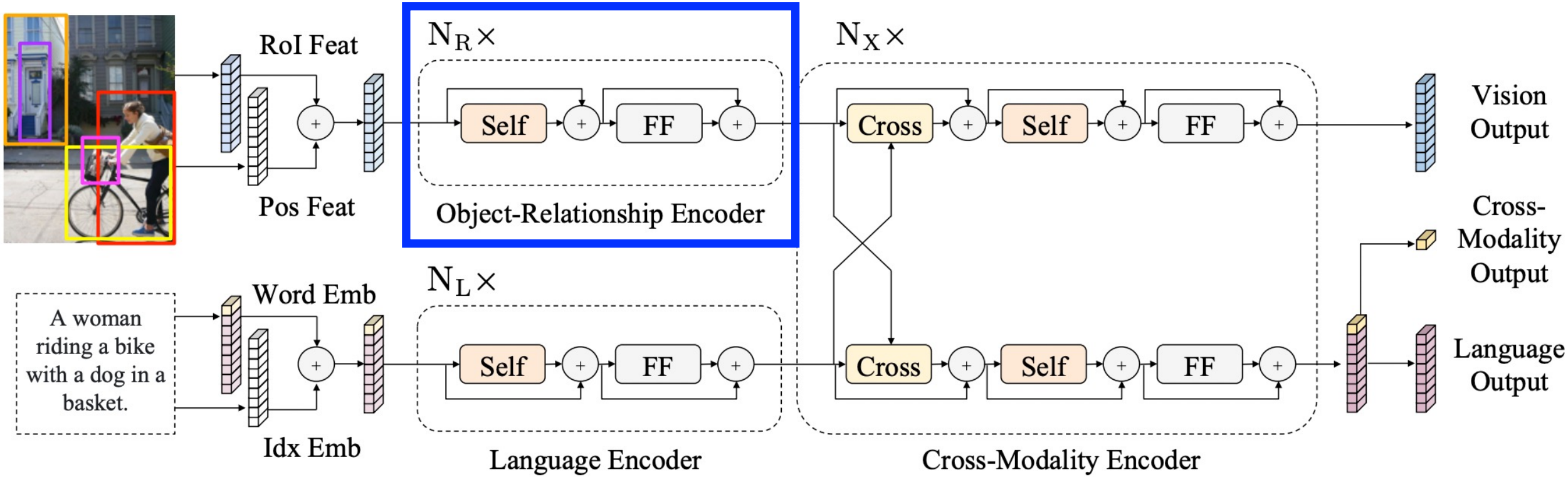
Each image is represented as a description of *m* objects detected with Faster R-CNN using features from Faster R-CNN and position encodings



Tan and Bansal, EMNLP 2019.

LXMERT: Architecture

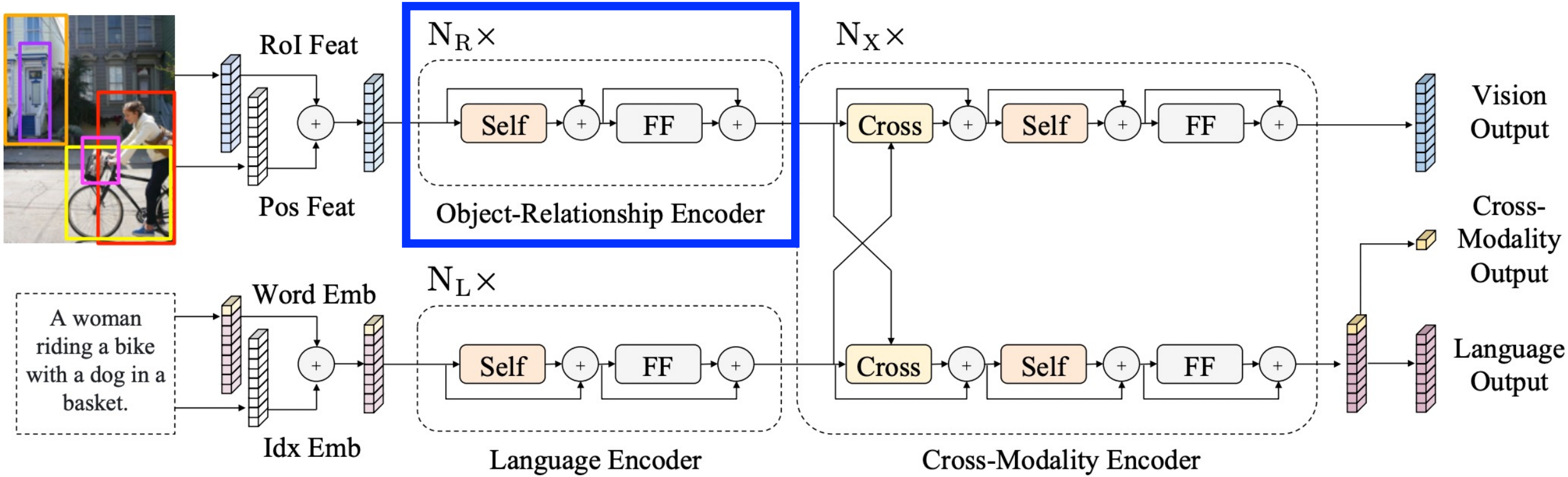
Transformer encoder (i.e., BERT);
what does its output represent?



Tan and Bansal, EMNLP 2019.

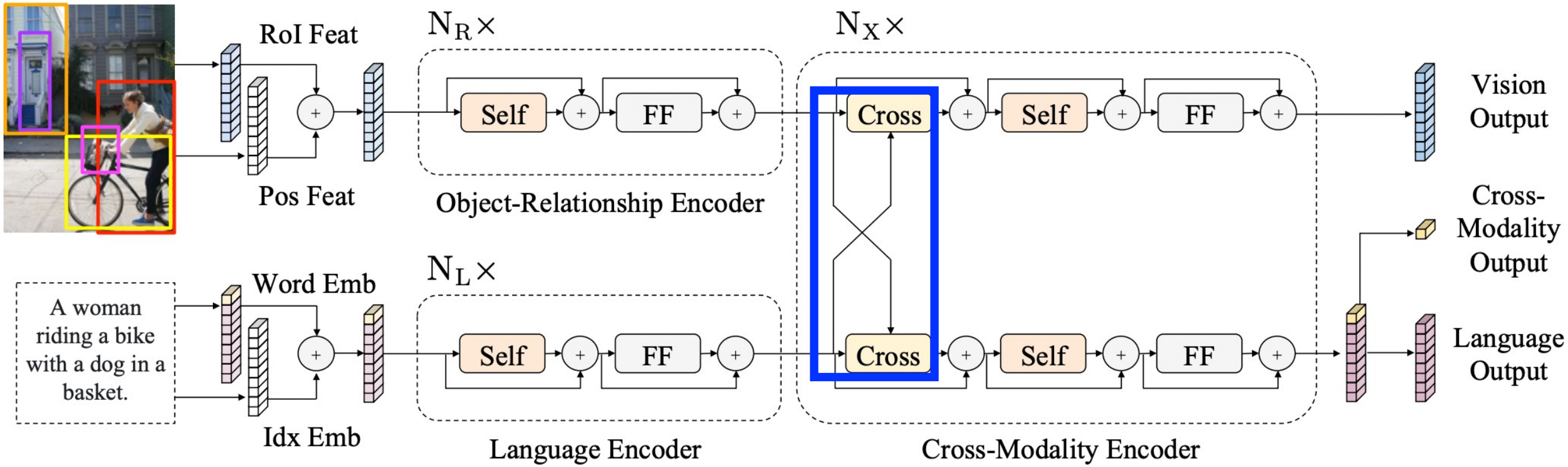
LXMERT: Architecture

Transformer encoder (i.e., BERT); represents objects with their relationships to all objects



Tan and Bansal, EMNLP 2019.

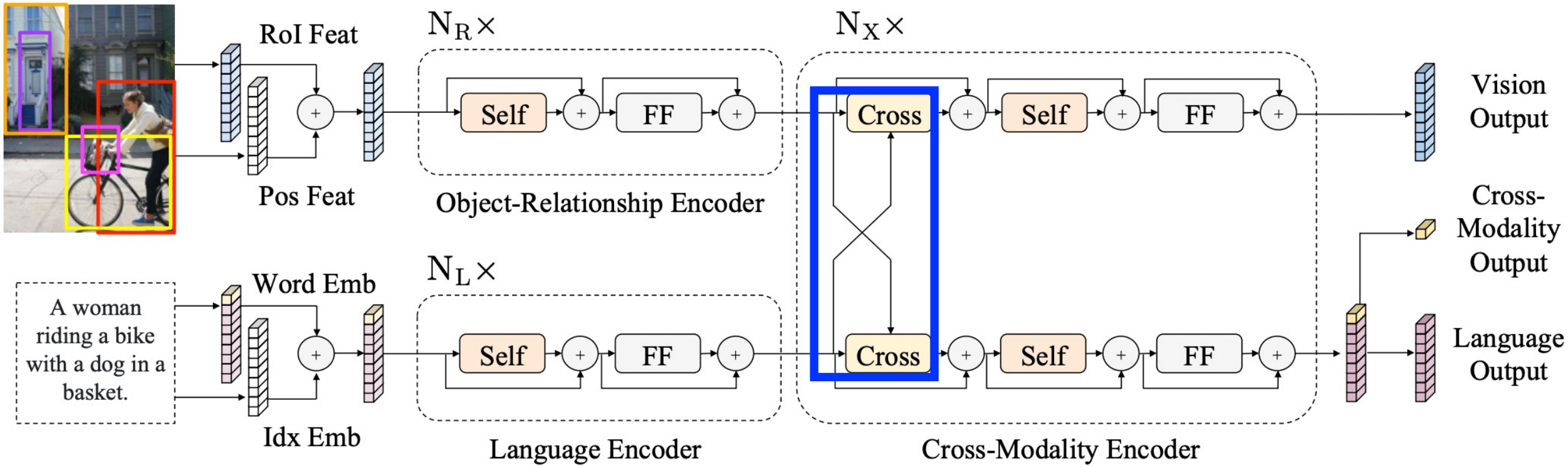
LXMERT: Architecture



Learns cross-modality representations
by aligning entities in the two modalities

Tan and Bansal, EMNLP 2019.

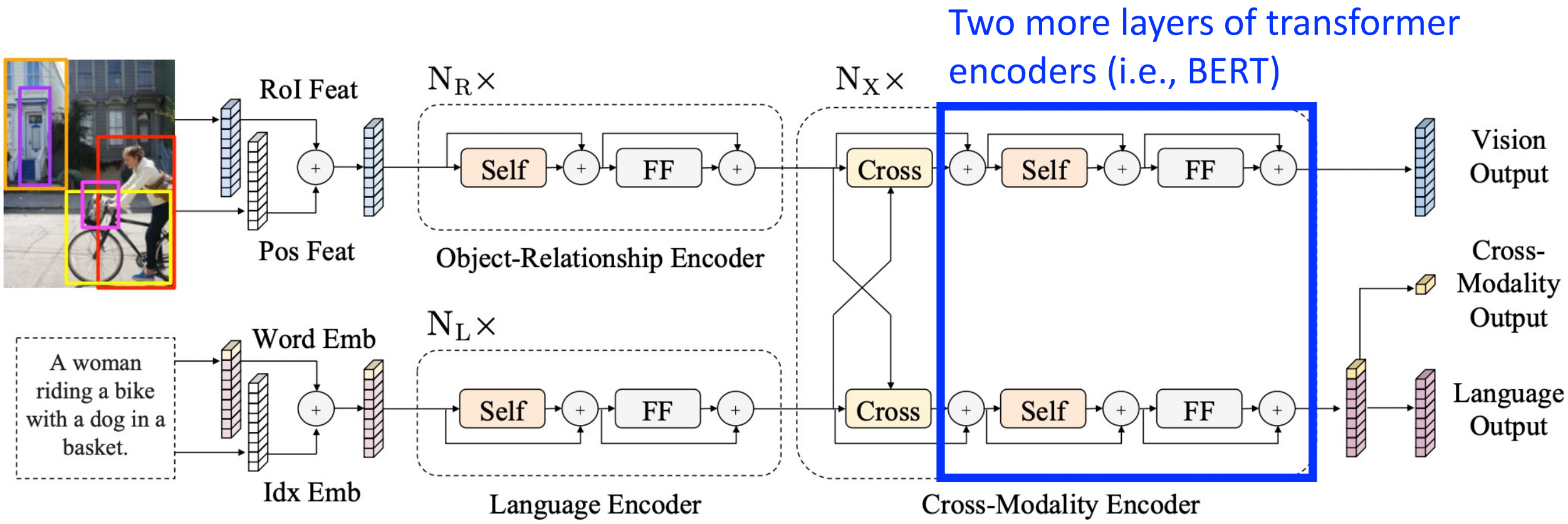
LXMERT: Architecture



Two cross-attention layers are functions of the "query" with "keys" and "values" from the opposite modalities' features

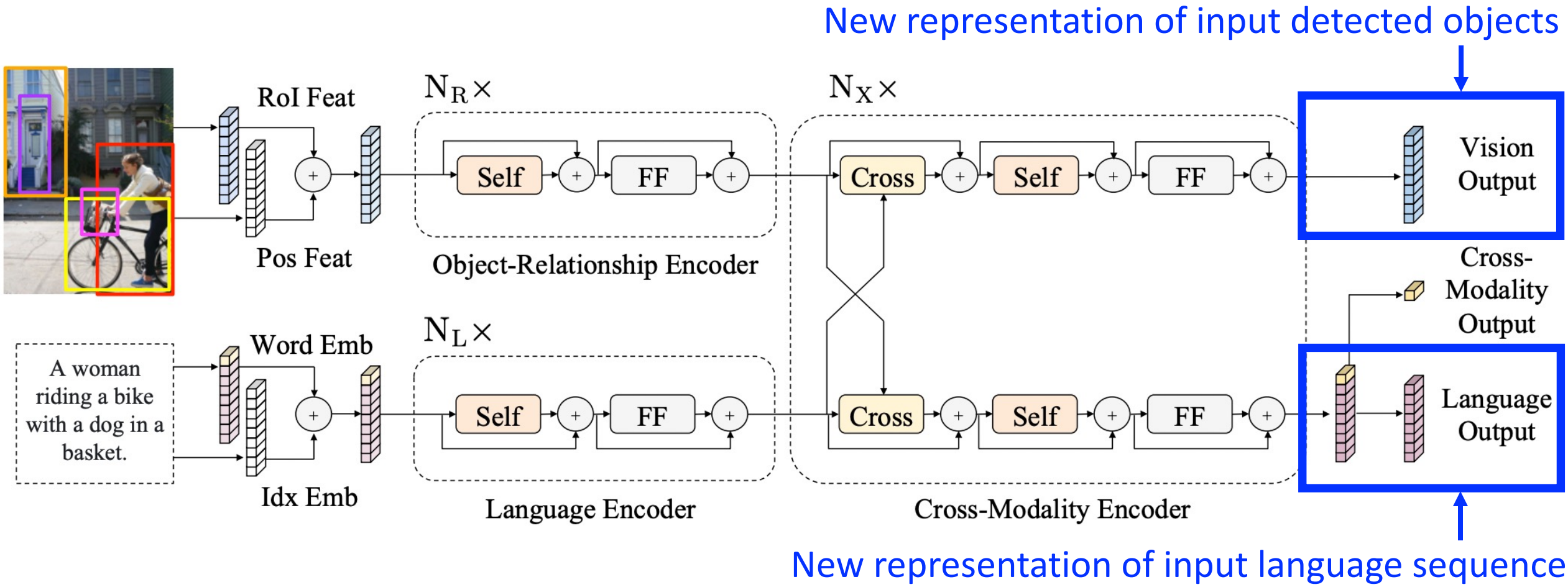
Tan and Bansal, EMNLP 2019.

LXMERT: Architecture



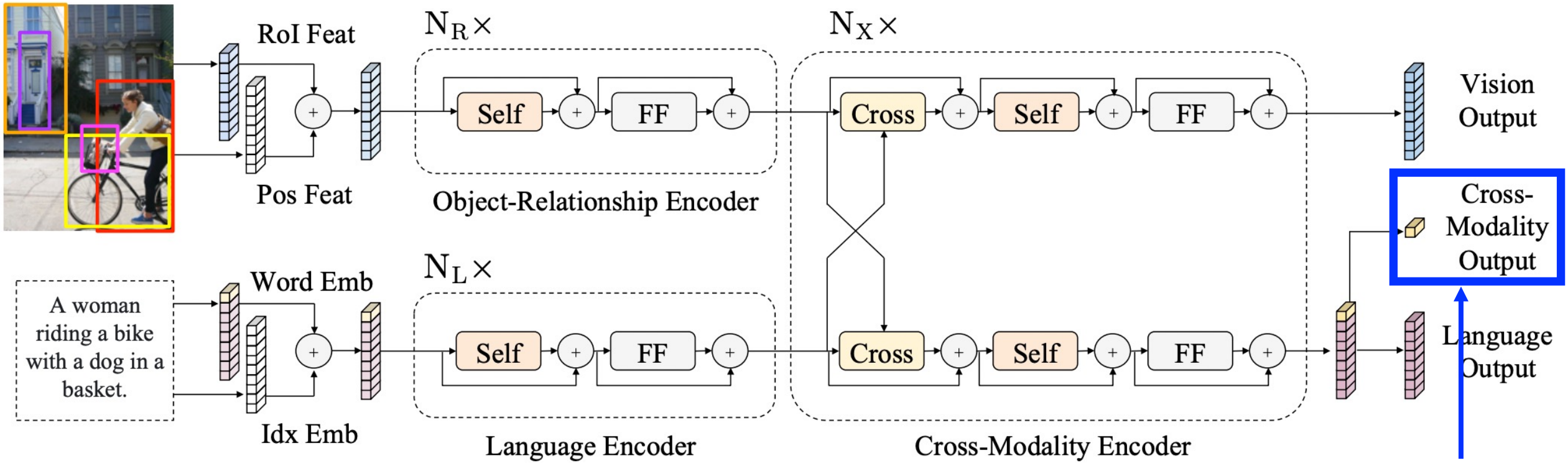
Tan and Bansal, EMNLP 2019.

LXMERT: Output



Tan and Bansal, EMNLP 2019.

LXMERT: Output

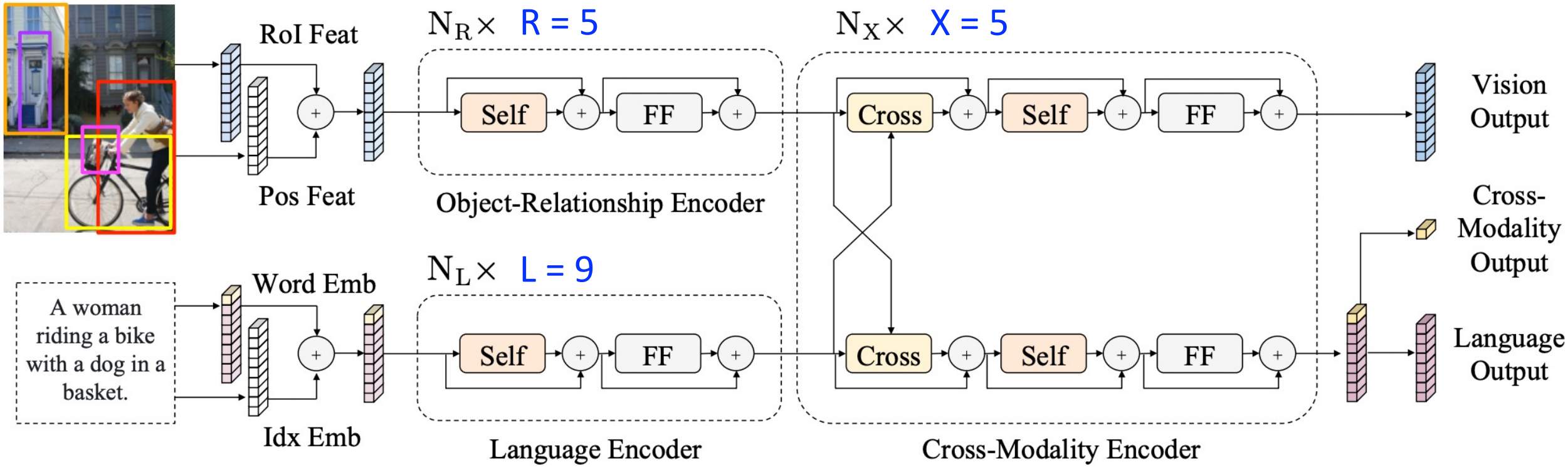


Cross-modality representation is the [CLS] token appended at the start of the sentence

Tan and Bansal, EMNLP 2019.

LXMERT: Implementation Details

Pretrained Faster R-CNN can locate 1,600 categories
and only 36 object detections are kept per image



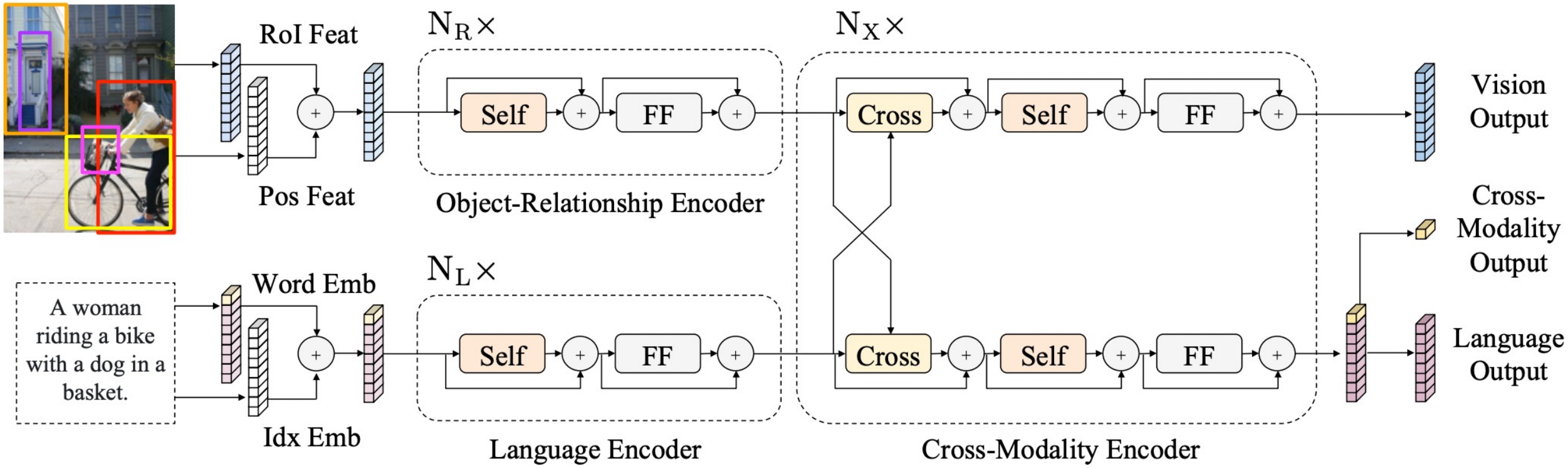
Number of layers mimics the size of BERT base of 12 layers;

i.e., (5+9)/2 + 5

Tan and Bansal, EMNLP 2019.

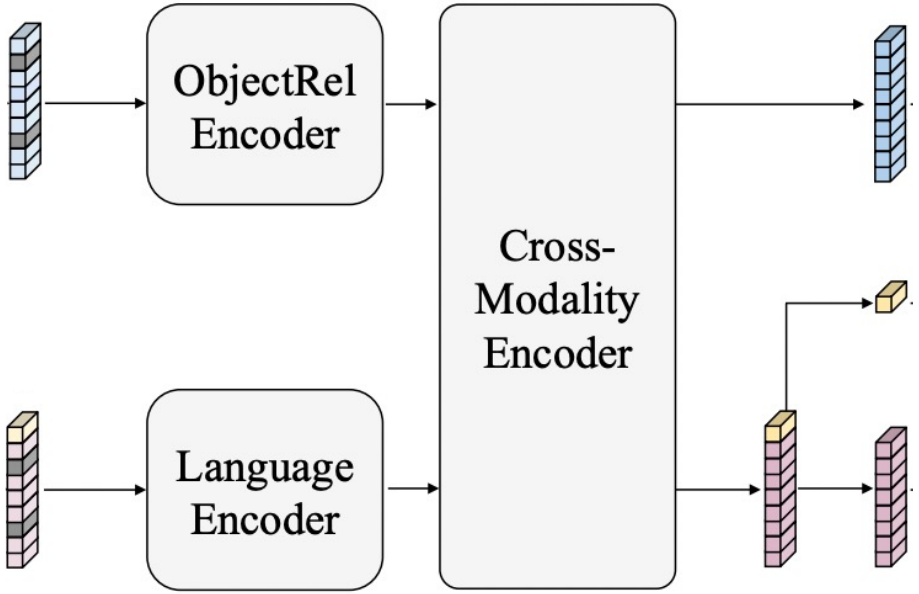
LXMERT: Architecture

What might be strengths and limitations of the resulting feature representations based on the architecture used?



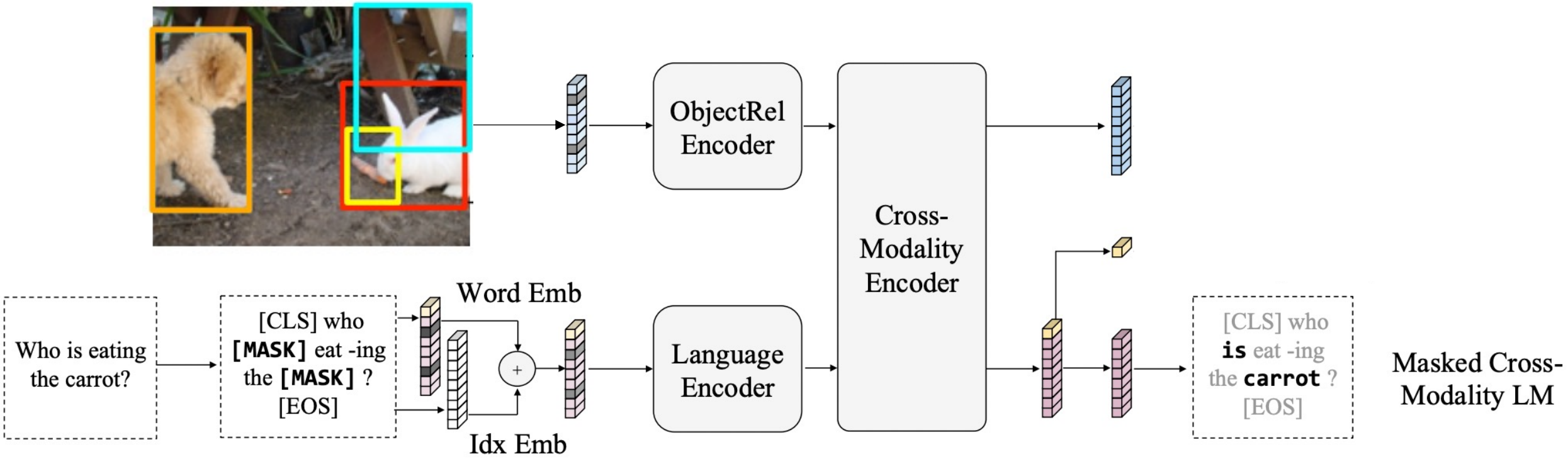
Tan and Bansal, EMNLP 2019.

LXMERT: Summary of Architecture



Tan and Bansal, EMNLP 2019.

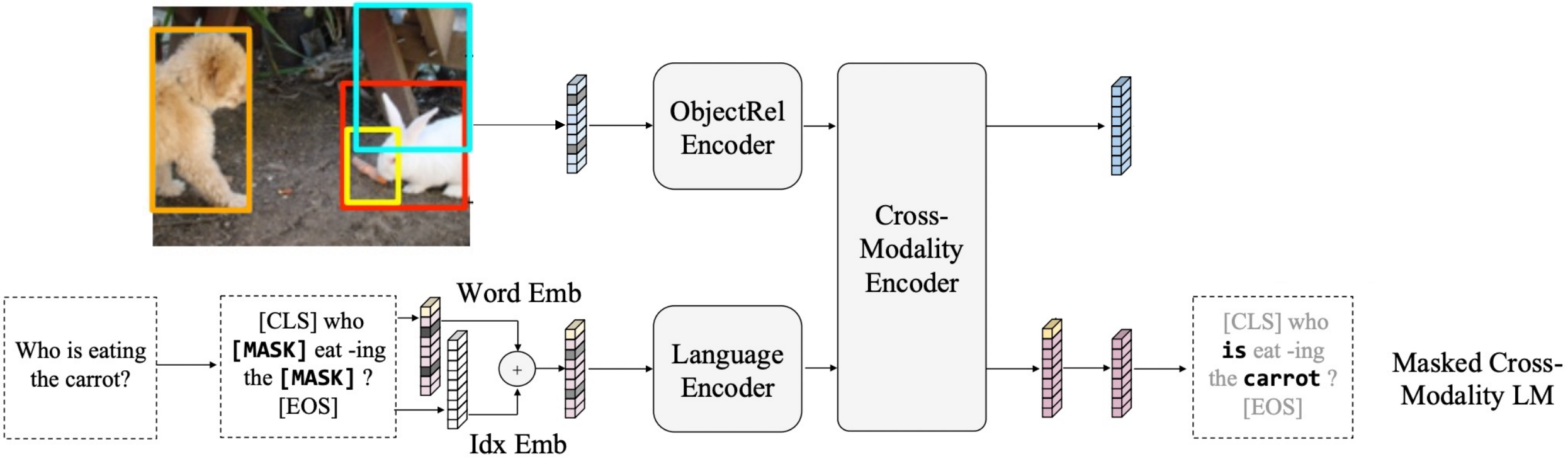
LXMERT: Pretraining Task 1 (Language)



Task used for BERT: mask 15% of input words and then predict them

Tan and Bansal, EMNLP 2019.

LXMERT: Pretraining Task 1 (Language)

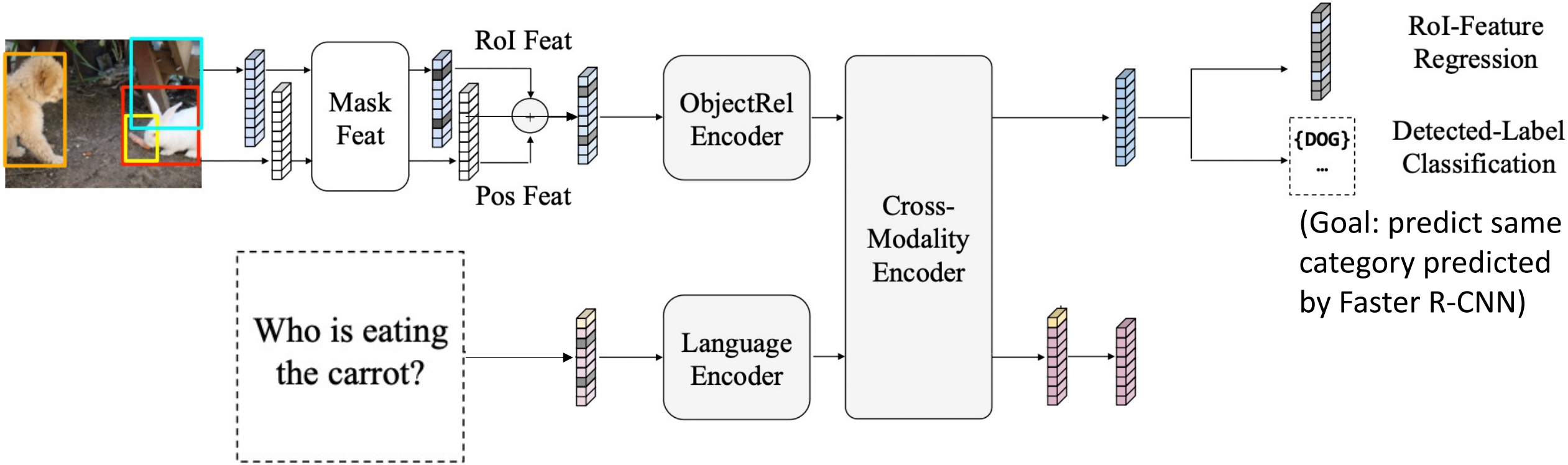


Unlike BERT, vision modality can resolve language ambiguity; e.g., shows what is being eaten

Tan and Bansal, EMNLP 2019.

LXMERT: Pretraining Tasks 2 & 3 (Vision)

Mask 15% of input objects and then predict their original feature values and categories

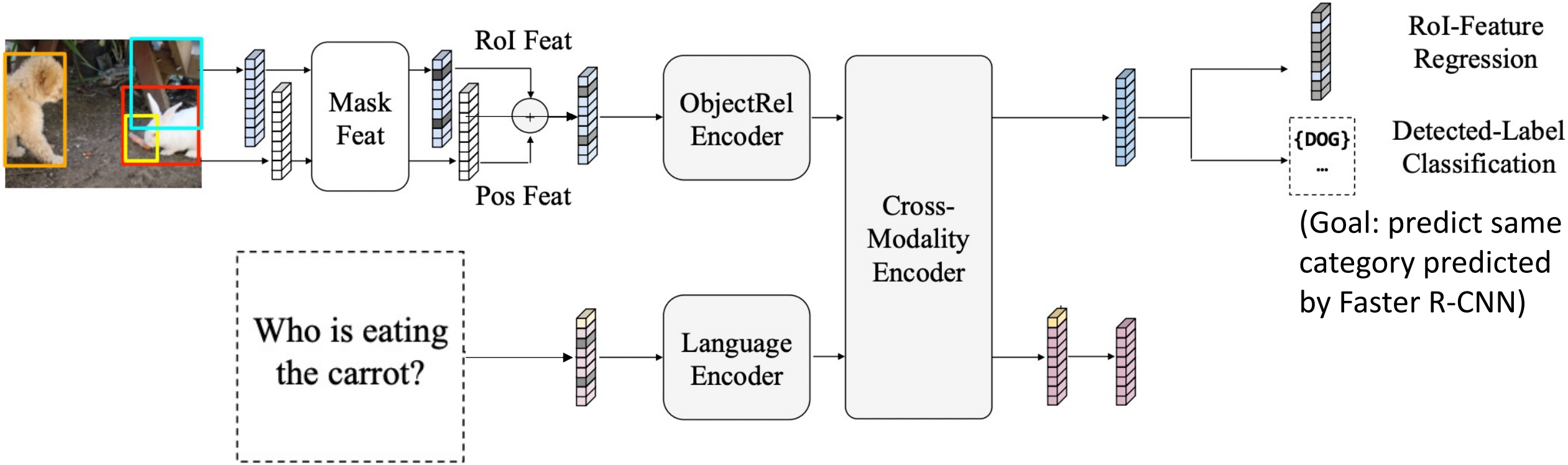


(Goal: predict same category predicted by Faster R-CNN)

Tan and Bansal, EMNLP 2019.

LXMERT: Pretraining Tasks 2 & 3 (Vision)

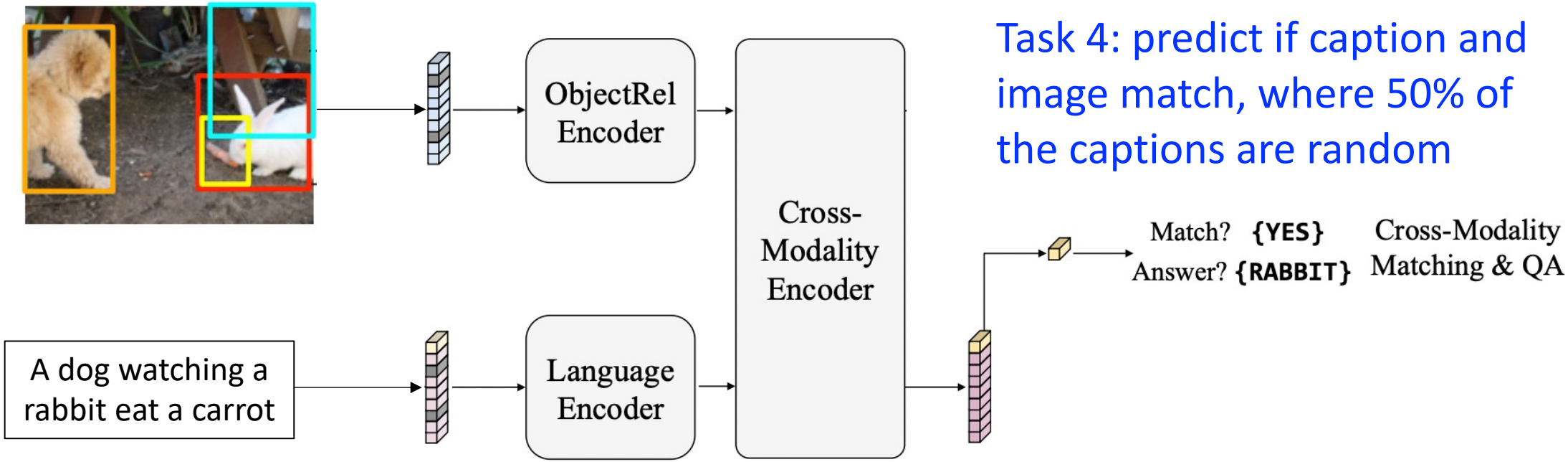
Knowledge about other objects and the language should help predict masked objects



(Goal: predict same category predicted by Faster R-CNN)

Tan and Bansal, EMNLP 2019.

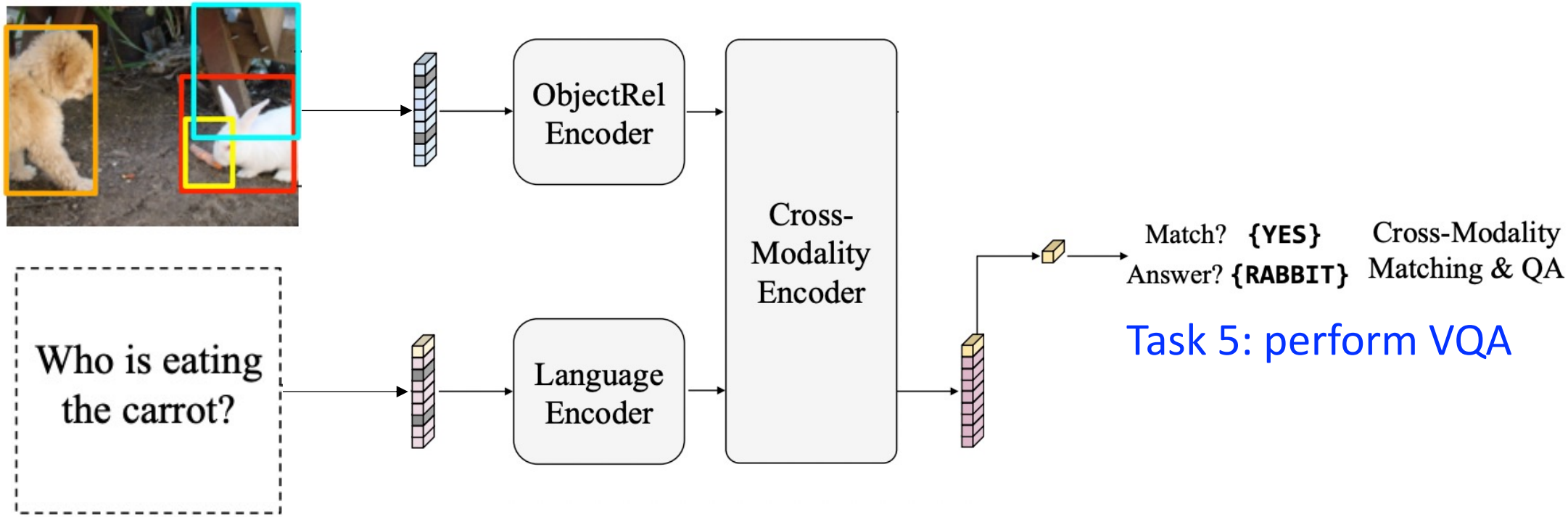
LXMERT: Pretraining Tasks 4 & 5 (Both Modalities)



Task 4: predict if caption and image match, where 50% of the captions are random

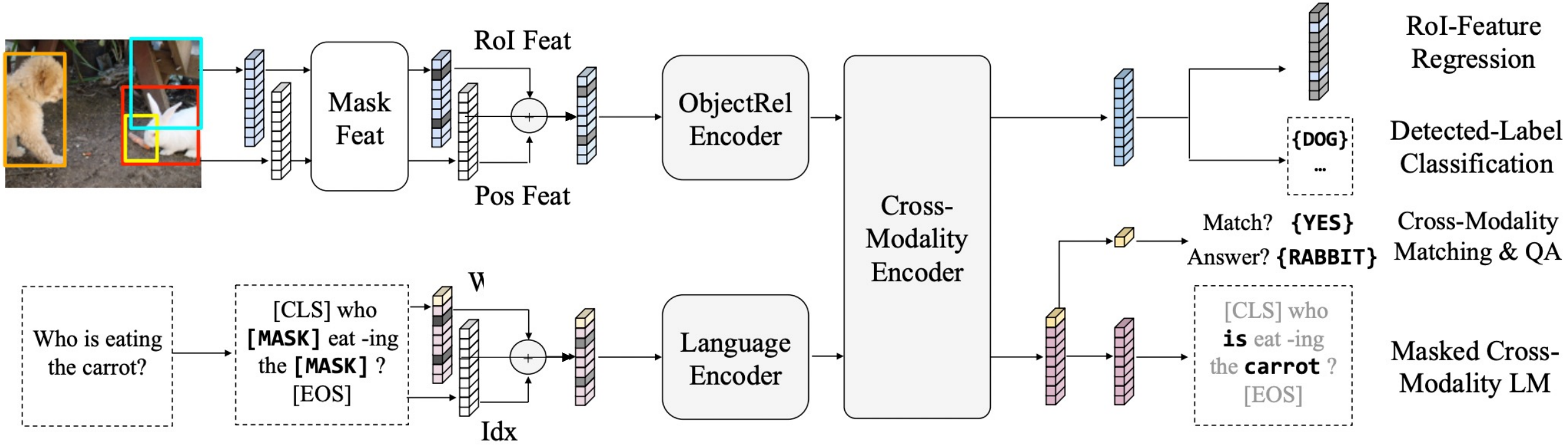
Tan and Bansal, EMNLP 2019.

LXMERT: Pretraining Tasks 4 & 5 (Both Modalities)



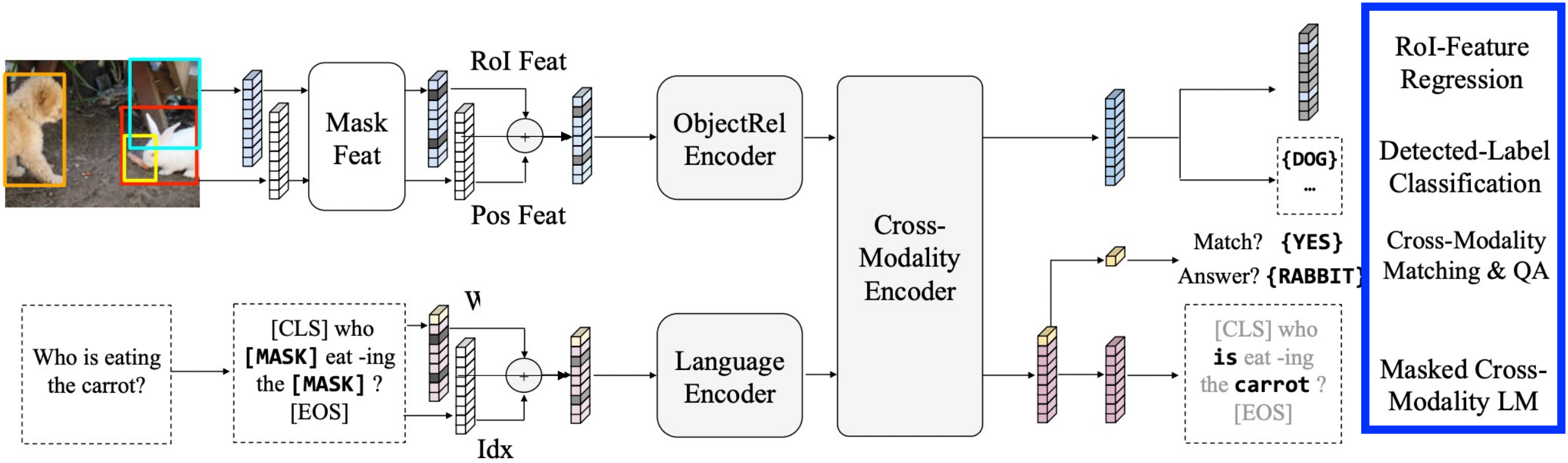
Tan and Bansal, EMNLP 2019.

LXMERT: 5 Pretraining Tasks



Tan and Bansal, EMNLP 2019.

LXMERT: All Pretraining Task Losses Are Summed During Training



What might be strengths and limitations of the resulting feature representations based on the type of pretraining tasks used?

Tan and Bansal, EMNLP 2019.

LXMERT: Training Data for Pretraining

Image Split	Images	Sentences (or Questions)					
		COCO-Cap	VG-Cap	VQA	GQA	VG-QA	All
MS COCO - VG	72K	361K	-	387K	-	-	0.75M
MS COCO ∩ VG	51K	256K	2.54M	271K	515K	724K	4.30M
VG - MS COCO	57K	-	2.85M	-	556K	718K	4.13M

All images are from two image sets, MS COCO and Visual Genome, which were collected by scraping images from the photo-sharing website Flickr

(Visual Genome includes the MS COCO images)

Tan and Bansal, EMNLP 2019.

LXMERT: Training Data for Pretraining

Image Split	Images	Sentences (or Questions)					
		COCO-Cap	VG-Cap	VQA	GQA	VG-QA	All
MS COCO - VG	72K	361K	-	387K	-	-	0.75M
MS COCO ∩ VG	51K	256K	2.54M	271K	515K	724K	4.30M
VG - MS COCO	57K	-	2.85M	-	556K	718K	4.13M

Language annotations came from 2 image captioning and 3 VQA datasets, authored by crowdworkers paid to create captions, questions, and answers

Tan and Bansal, EMNLP 2019.

LXMERT: Training Data for Pretraining

Image Split	Images	Sentences (or Questions)					
		COCO-Cap	VG-Cap	VQA	GQA	VG-QA	All
MS COCO - VG	72K	361K	-	387K	-	-	0.75M
MS COCO ∩ VG	51K	256K	2.54M	271K	515K	724K	4.30M
VG - MS COCO	57K	-	2.85M	-	556K	718K	4.13M
All	180K	617K	5.39M	658K	1.07M	1.44M	9.18M

A total of 9.18M image-sentence pairs are included for 180,000 images
(questions in VQA datasets are used for the image-sentence pairs)

Tan and Bansal, EMNLP 2019.

LXMERT: Training Data for Pretraining

Image Split	Images	Sentences (or Questions)					
		COCO-Cap	VG-Cap	VQA	GQA	VG-QA	All
MS COCO - VG	72K	361K	-	387K	-	-	0.75M
MS COCO ∩ VG	51K	256K	2.54M	271K	515K	724K	4.30M
VG - MS COCO	57K	-	2.85M	-	556K	718K	4.13M
All	180K	617K	5.39M	658K	1.07M	1.44M	9.18M

What might be strengths and limitations of the resulting feature representations based on the type of training data that is used?

Tan and Bansal, EMNLP 2019.

LXMERT: Fine-Tuning Experimental Results

Method	VQA			
	Binary	Number	Other	**Accu**
Human	-	-	-	-
Image Only	-	-	-	-
Language Only	66.8	31.8	27.6	44.3
State-of-the-Art	85.8	53.7	60.7	70.4
LXMERT	**88.2**	**54.2**	**63.1**	**72.5**

Achieved the best performance, with stronger gains over prior work for questions that lead to "binary" and "other" answers

Tan and Bansal, EMNLP 2019.

LXMERT: Fine-Tuning Experimental Results

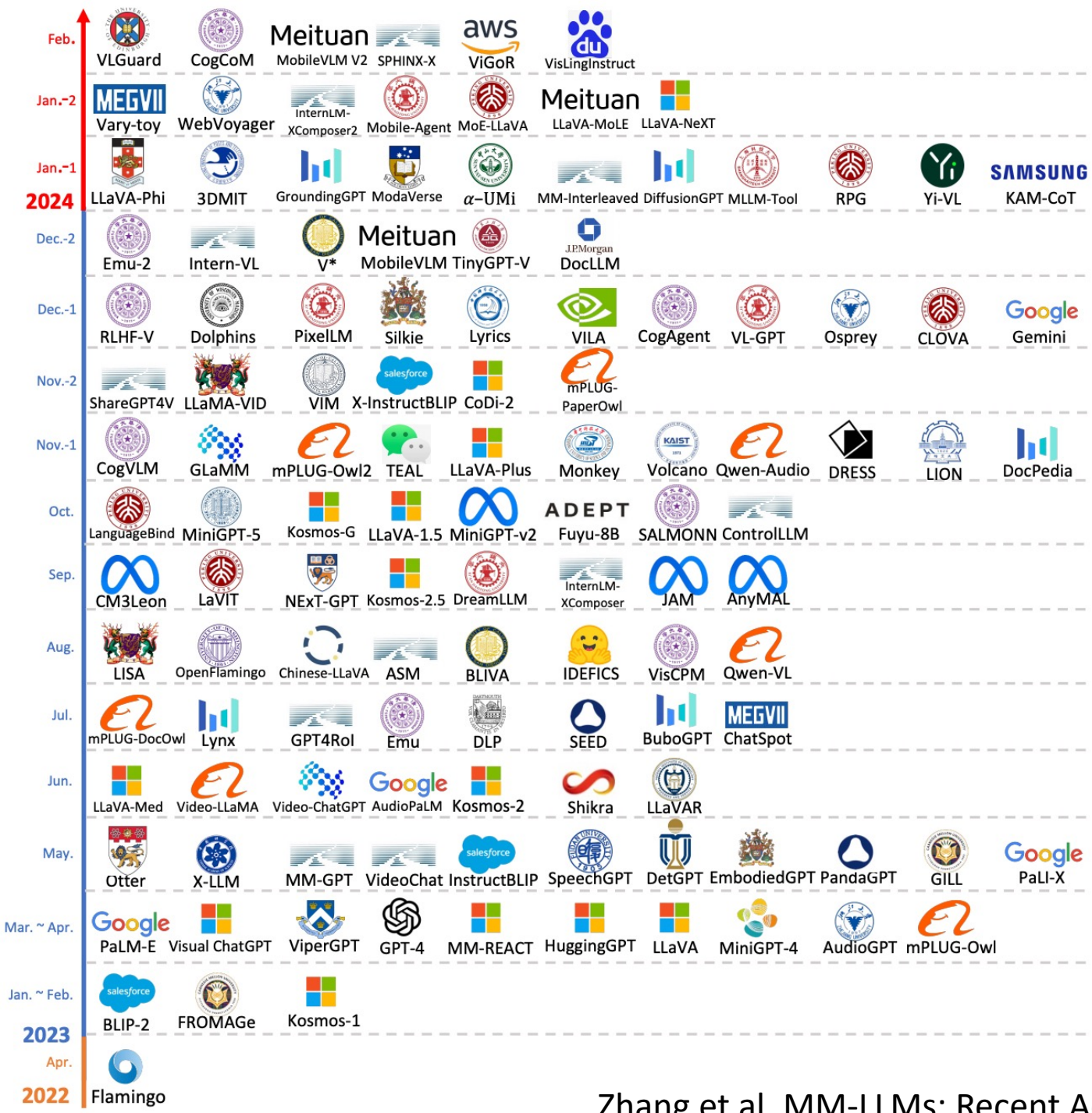
Method	VQA				GQA			NLVR^2	
	Binary	Number	Other	**Accu**	Binary	Open	**Accu**	Cons	**Accu**
Human	-	-	-	-	91.2	87.4	89.3	-	96.3
Image Only	-	-	-	-	36.1	1.74	17.8	7.40	51.9
Language Only	66.8	31.8	27.6	44.3	61.9	22.7	41.1	4.20	51.1
State-of-the-Art	85.8	53.7	60.7	70.4	76.0	40.4	57.1	12.0	53.5
LXMERT	**88.2**	**54.2**	**63.1**	**72.5**	**77.8**	**45.0**	**60.3**	**42.1**	**76.2**

The representations also led to the best performance for an additional VQA dataset and a visual reasoning task (i.e., does statement describe two images or not)

Tan and Bansal, EMNLP 2019.

Today’s Topics

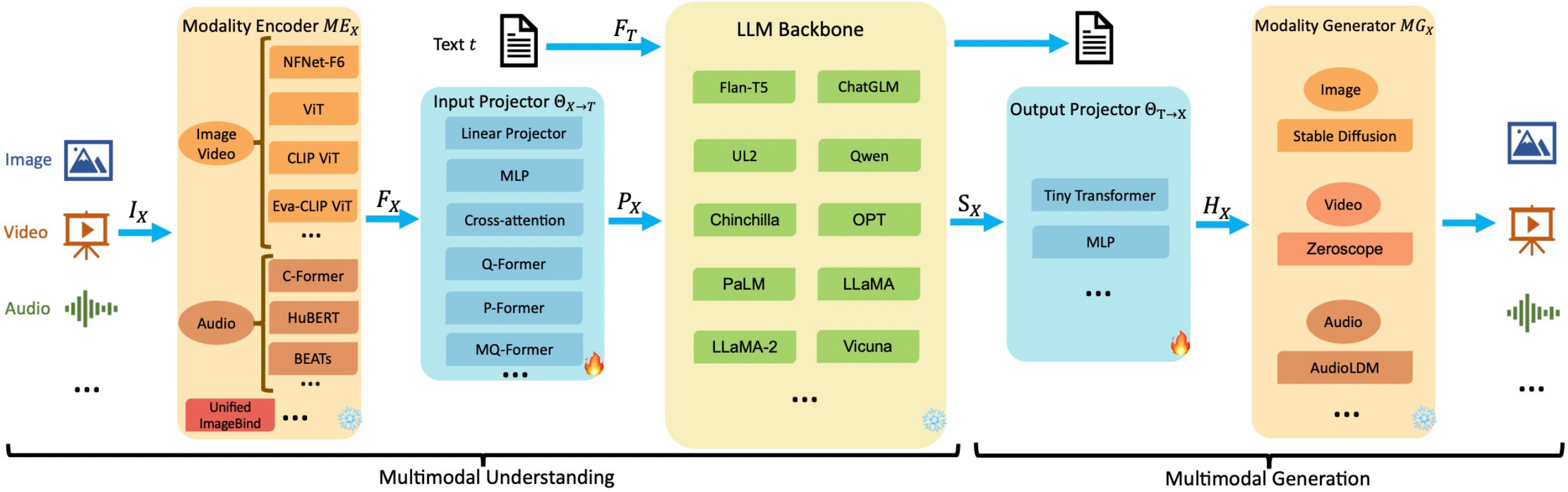
- Multimodal applications
- Image captioning dataset challenges
- Image captioning algorithms
- Visual question answering dataset challenges
- Visual question answering algorithms
- Foundation models



Contributors from academia & industry!

Zhang et al. MM-LLMs: Recent Advances in MultiModal Large Language Models. arXiv 2024.

Popular Variants of Architectures



Many convert visual inputs to text descriptions that LLMs understand (i.e., prompts) or finetune the vision encoder to align with a frozen pre-trained LLM

Zhang et al. MM-LLMs: Recent Advances in MultiModal Large Language Models. arXiv 2024.

Famous Models in Industry (Proprietary)

- Next frontier: models that are multimodal from design
- OpenAI's GPT-4(Vision): vision + language model
  - Rresults: https://openai.com/contributions/gpt-4v
- Google's Gemini: vision + language model
  - Testing results: https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf

Prompt Types: Textual vs Visual

- When Might One Choose A Visual Prompt Versus a Textual Prompt?
  - Greater equity for different languages as non-English languages often are poorly supported if at all
  - Empowering people appropriately based on their (dis)abilities: e.g., blind and deaf users

Today’s Topics

- Multimodal applications
- Image captioning dataset challenges
- Image captioning algorithms
- Visual question answering dataset challenges
- Visual question answering algorithms
- Foundation models

The End