Vision-Language Tasks: Image Captioning & Visual Question Answering

Danna Gurari
University of Colorado Boulder
Fall 2024

https://dannagurari.colorado.edu/course/recent-advances-in-computer-vision-fall-2024/

Review

- Last lecture: object tracking
  - Problem
  - Applications
  - Datasets
  - Evaluation metrics
  - Computer vision models
  - Discussion

- Assignments (Canvas)
  - Reading assignment was due earlier today
  - Reading assignment due Monday
  - Project outline due in one week

- Questions?

Today’s Topics

- Multimodal applications
- Image captioning dataset challenges
- Image captioning algorithms
- Visual question answering dataset challenges
- Discussion (chosen by YOU ☺)

Today’s Topics

- Multimodal applications
- Image captioning dataset challenges
- Image captioning algorithms
- Visual question answering dataset challenges
- Discussion (chosen by YOU ☺)

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

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

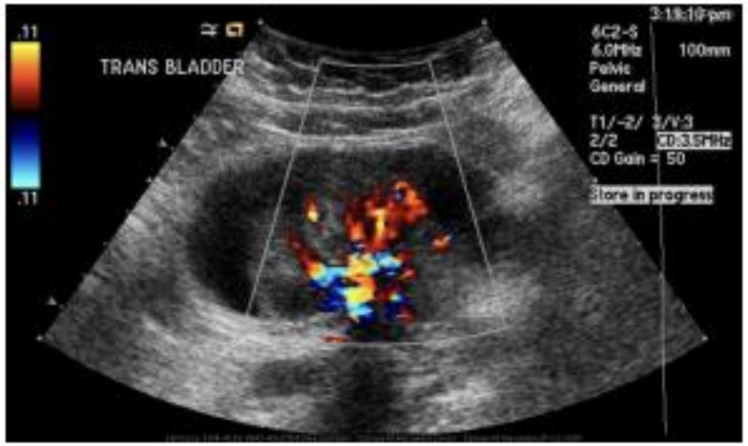
Color: Black and ivory

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

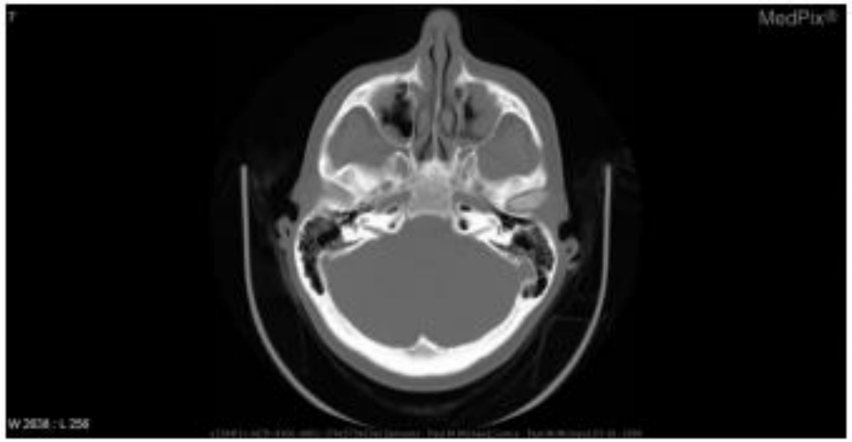
Image Caption: A person in a dress

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

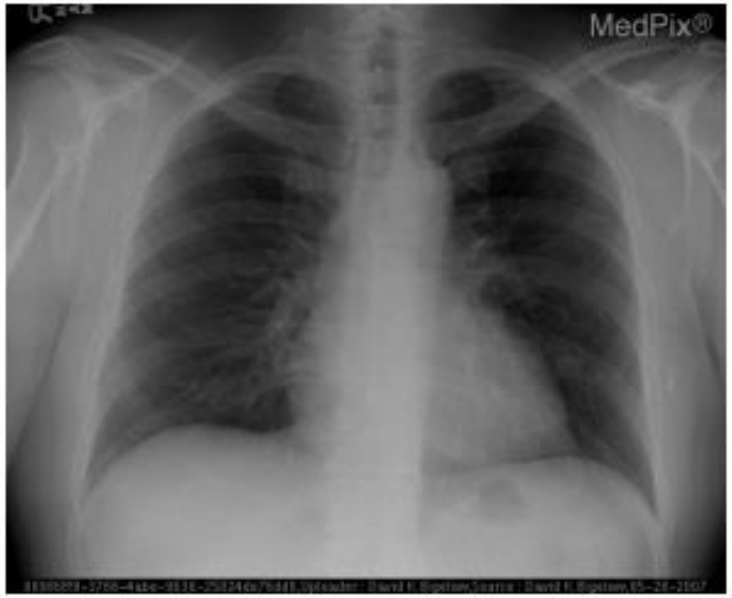
Medical VQA



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



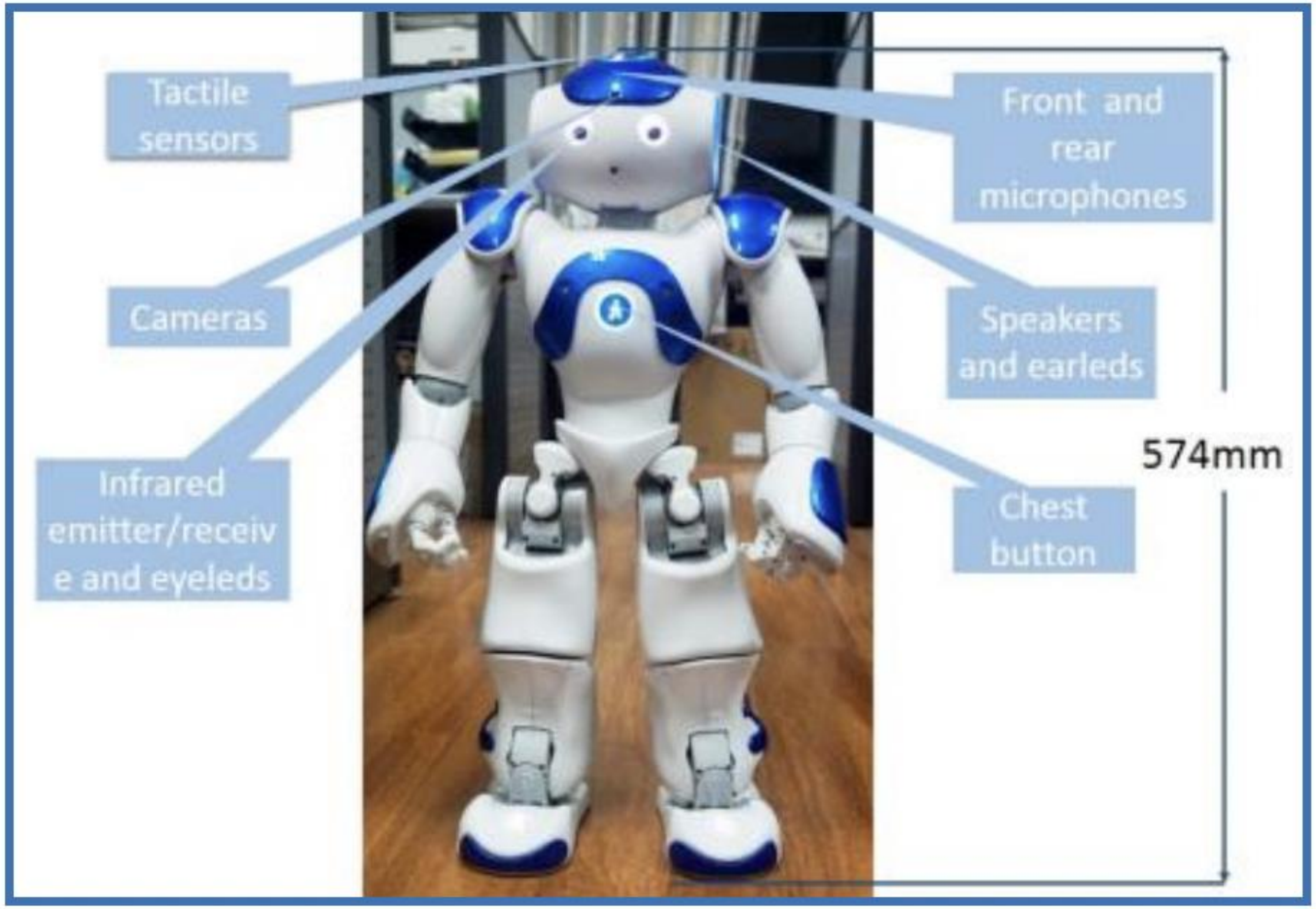
(b) **Q**: which plane is the image shown in? **A**: axial



(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

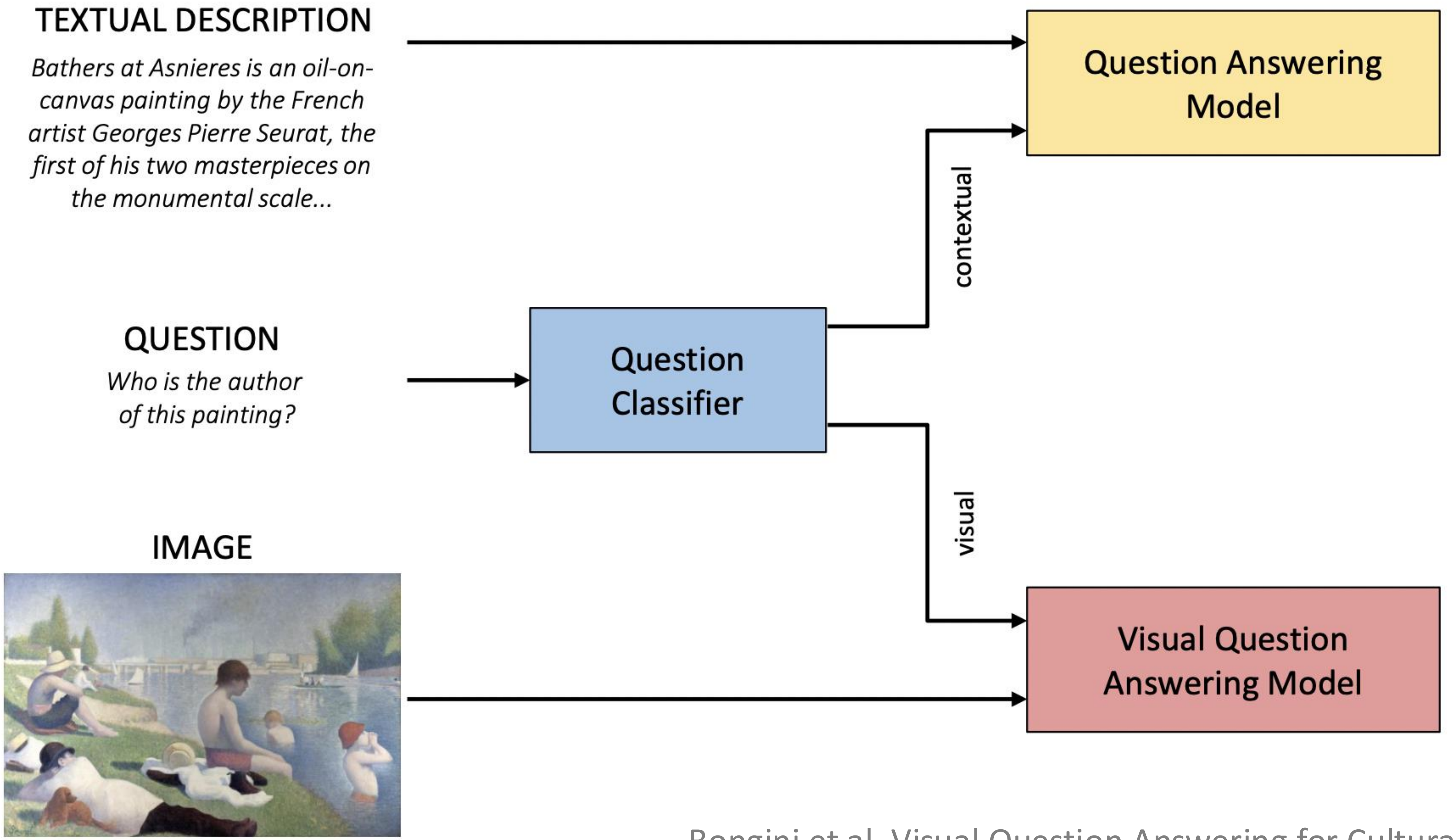
Education (e.g., for Preschoolers)



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

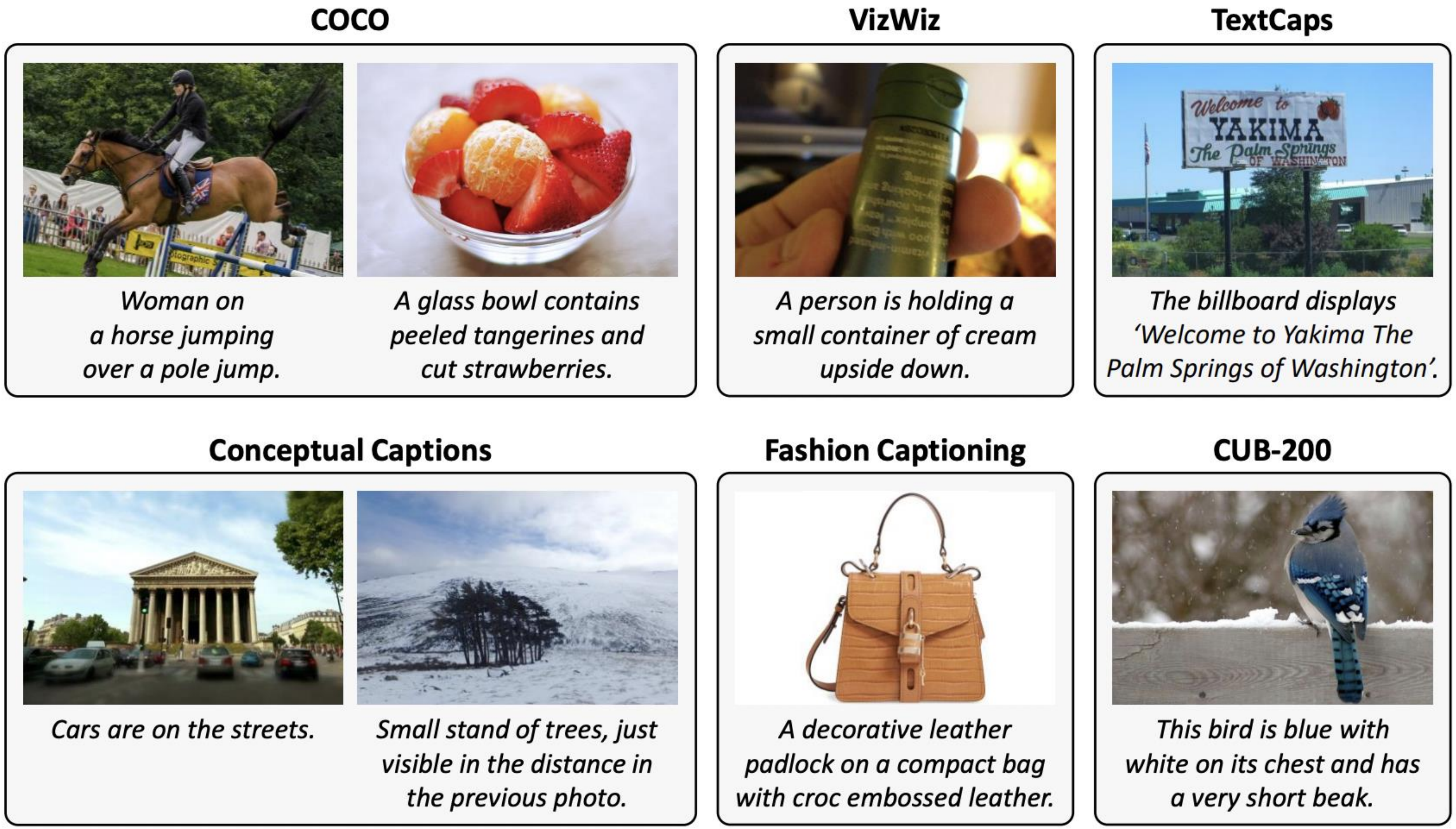


Bongini et al. Visual Question Answering for Cultural Heritage. 2020

Today's Topics

- Multimodal applications
- Image captioning dataset challenges
- Image captioning algorithms
- Visual question answering dataset challenges
- Discussion (chosen by YOU ☺)

Sample of Existing Dataset Challenges



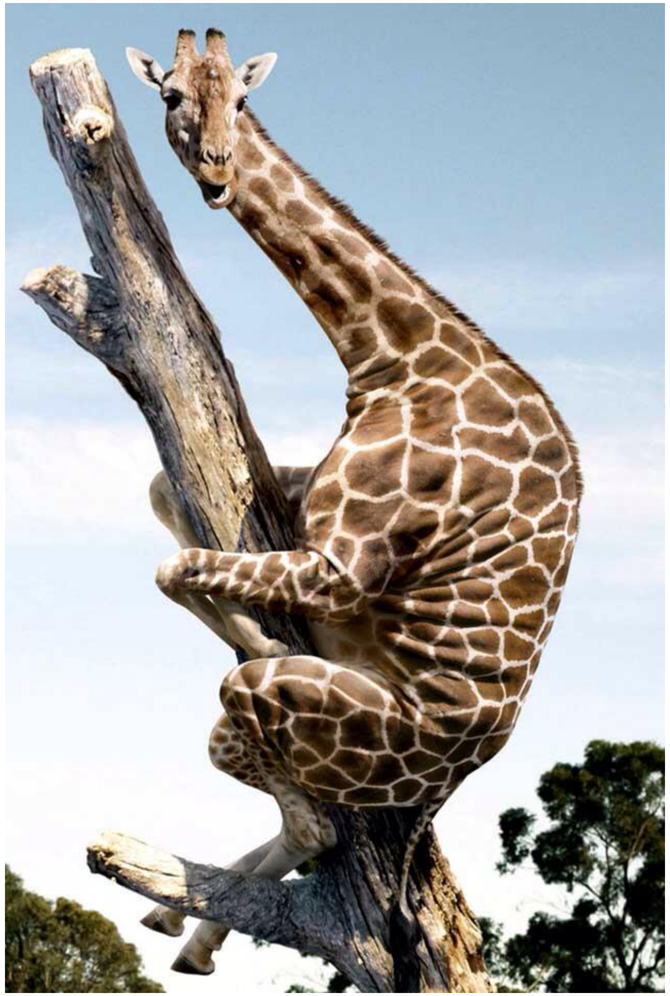
Stefanini et al. From Show to Tell: A Survey on Deep Learning-based Image Captioning. arXiv 2021

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?



Fill out Google form

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

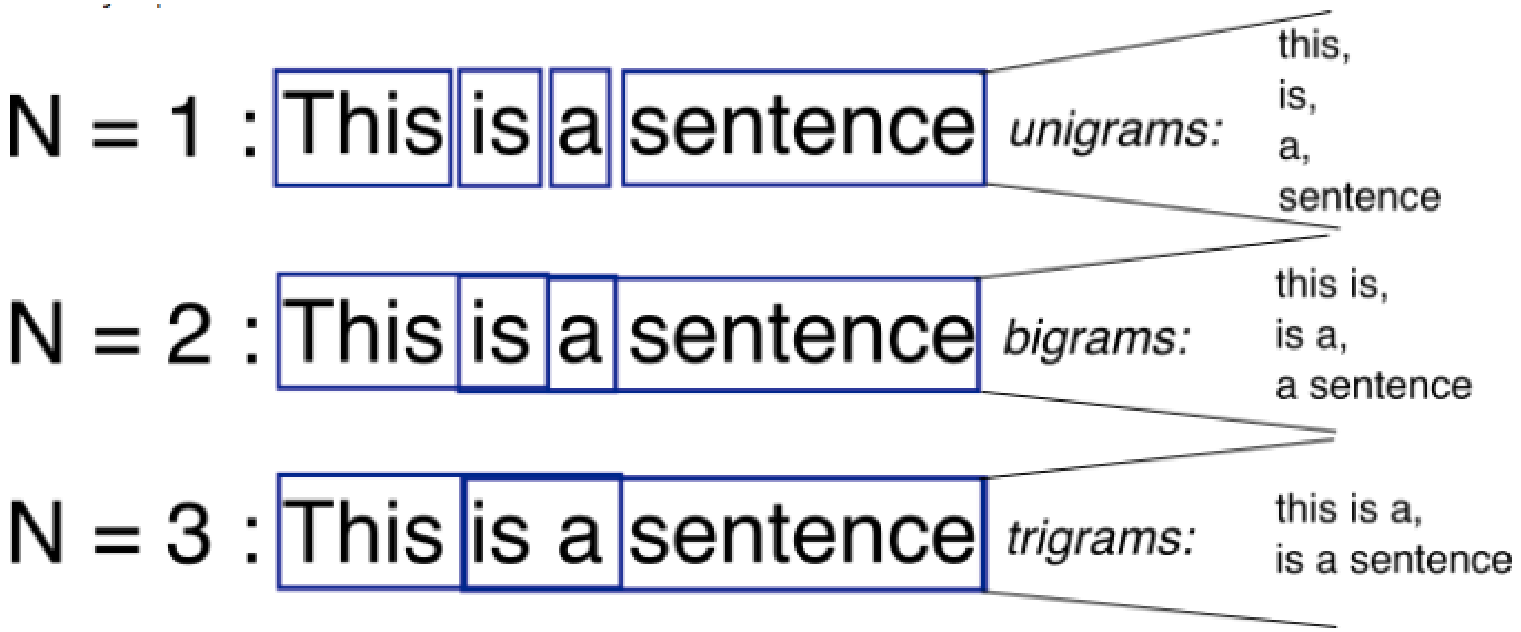
Evaluation: Automated

- 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

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
- METEOR
- Rouge
- CIDEr
- SPICE

Problem: n-gram methods scores these as very similar

(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

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.

- METEOR
- 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

Problem: n-gram methods scores these as very different

(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

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:

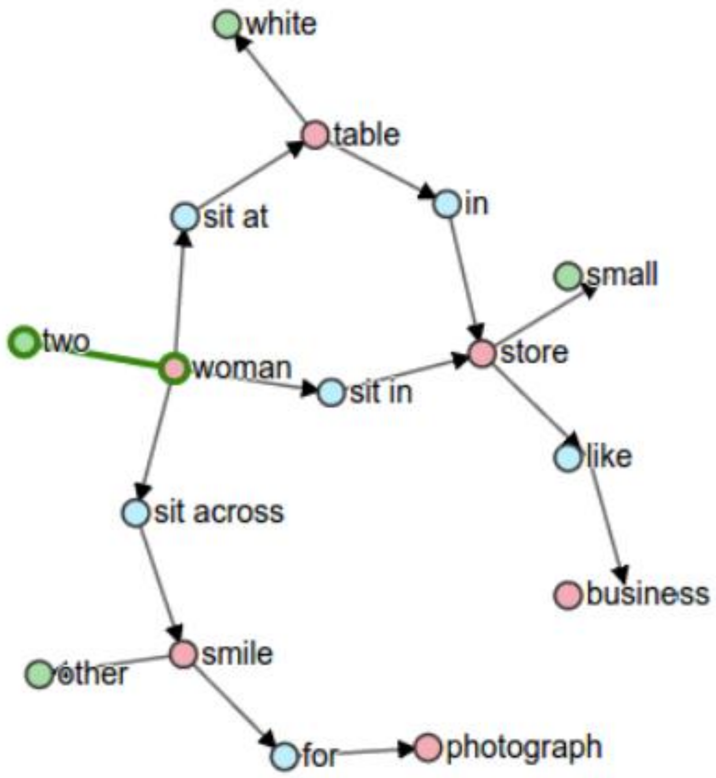
"two women are sitting at a white table"

"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"

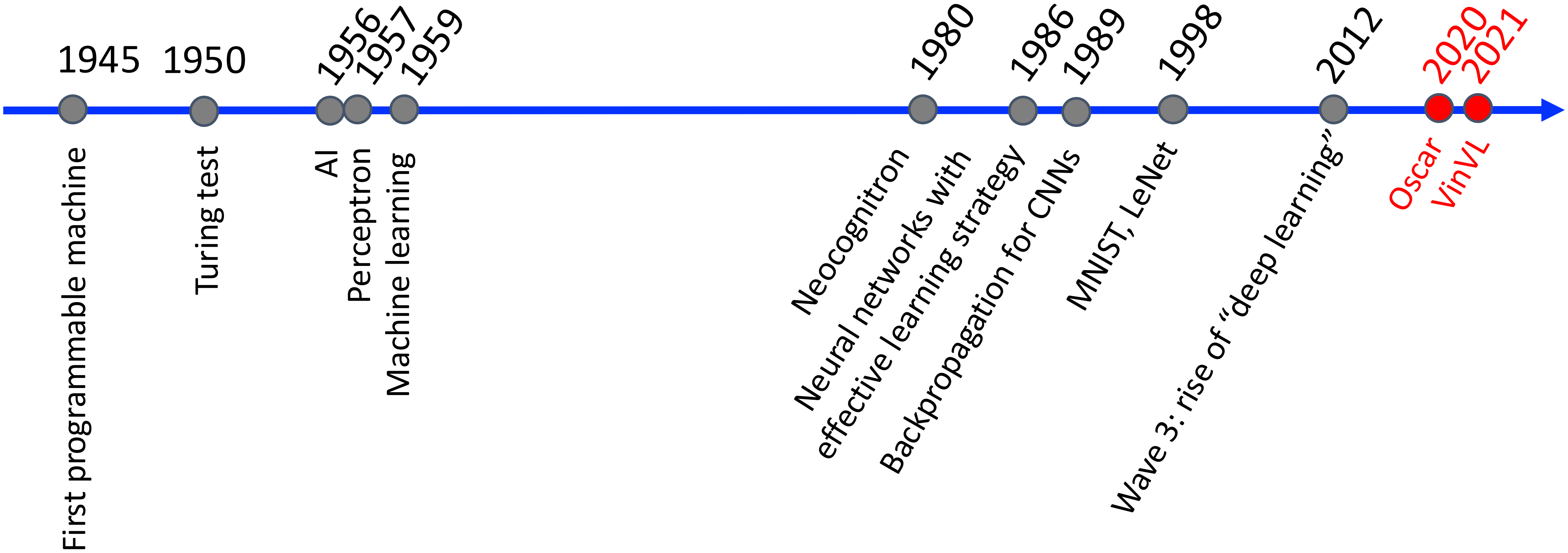


P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016

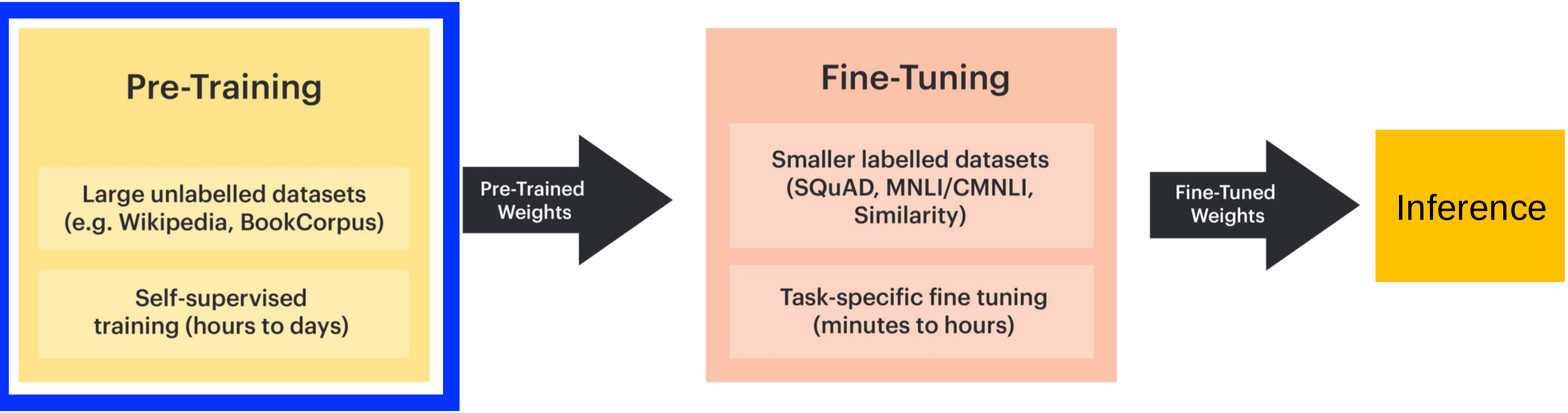
Today’s Topics

- Multimodal applications
- Image captioning dataset challenges
- Image captioning algorithms
- Visual question answering dataset challenges
- Discussion (chosen by YOU ☺)

Historical Context



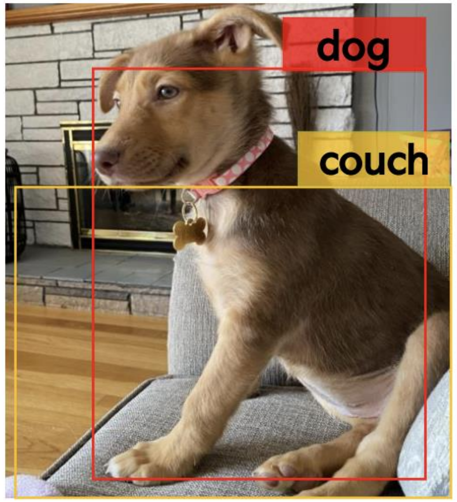
Oscar: Transformer Design



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

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

- **Idea**: explicitly learn alignment between text and features
- **Motivating observations**: often, salient objects are mentioned in image descriptions and can be located by object detection algorithms

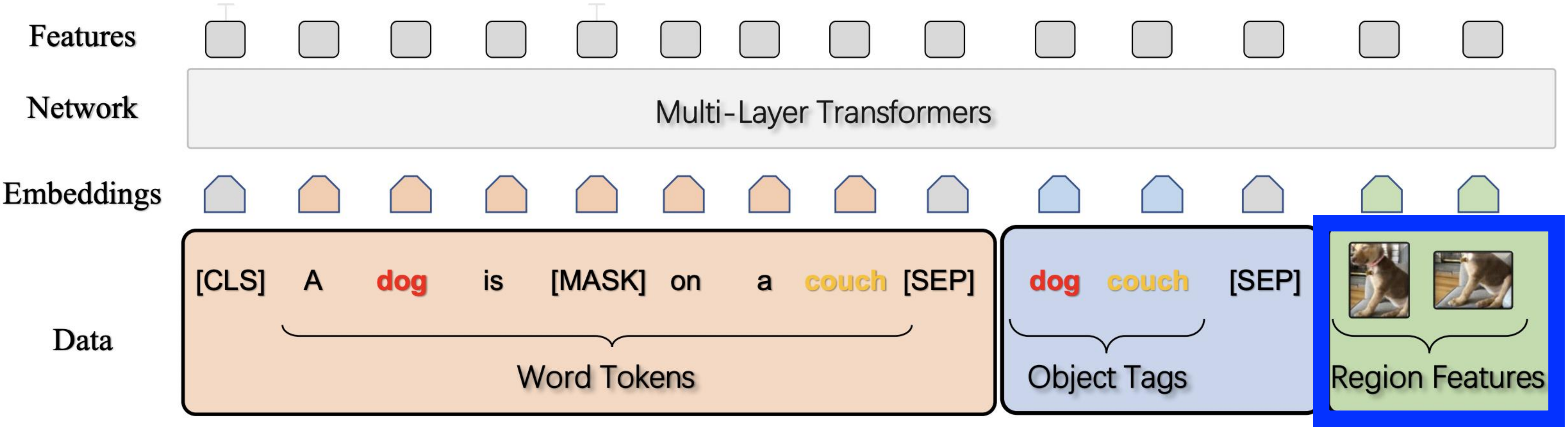


A dog is sitting on a couch

VS

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

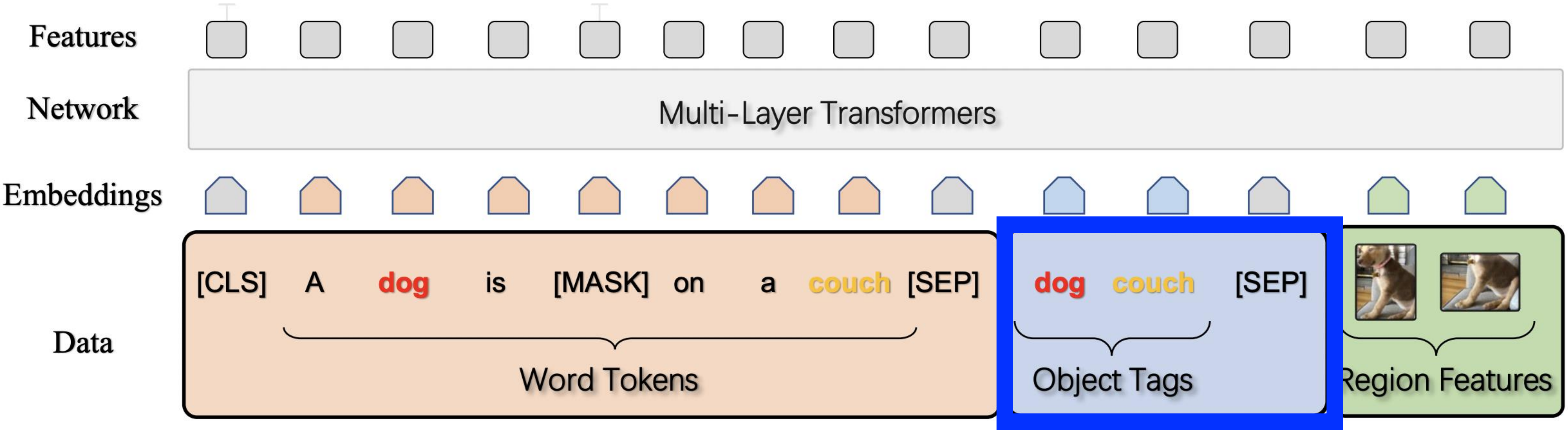
Oscar: Architecture



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

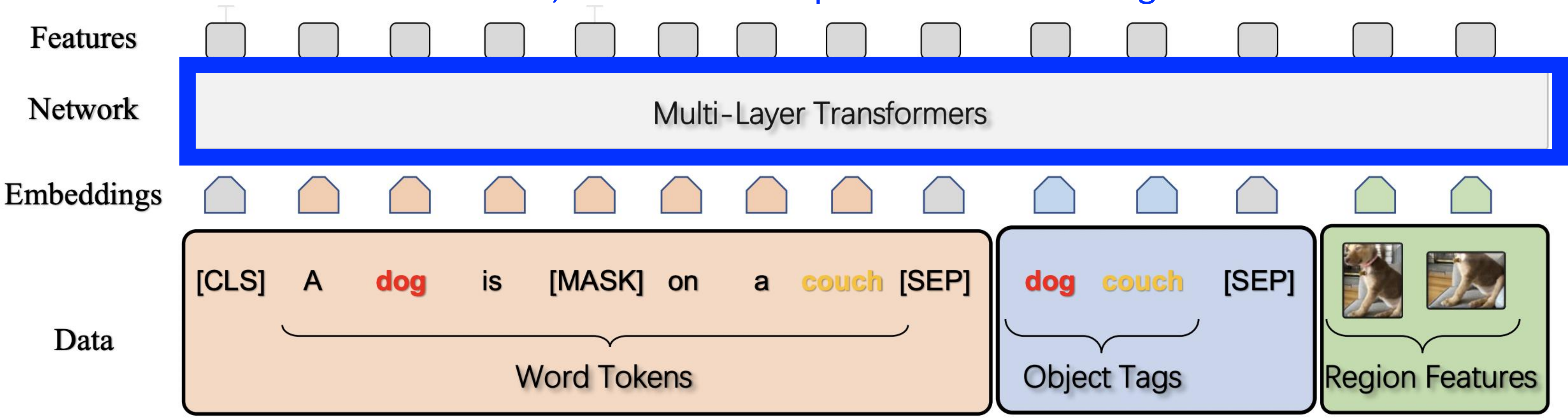


Novelty is to incorporate tags predicted by Faster R-CNN

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

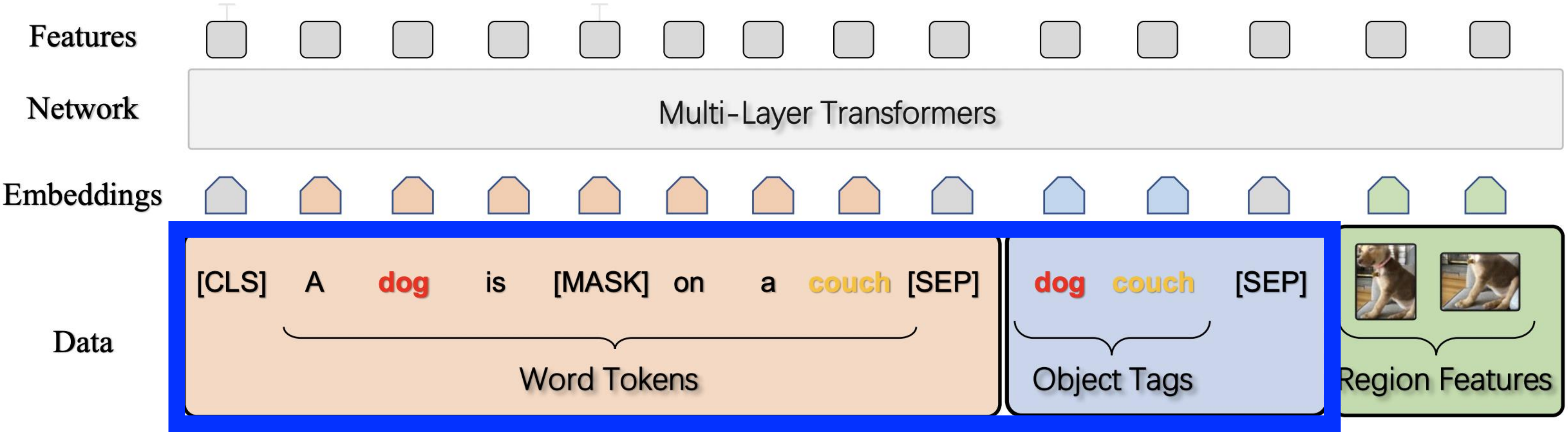
Oscar: Architecture

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)

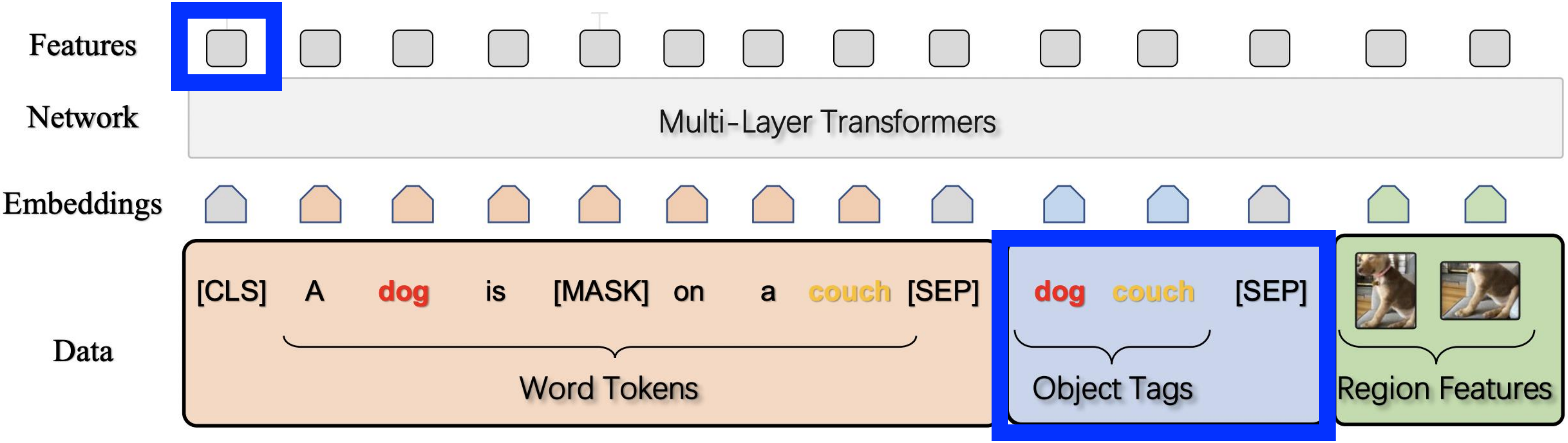


Like BERT, predict randomly masked tokens based on surrounding words, tags, and image information

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)

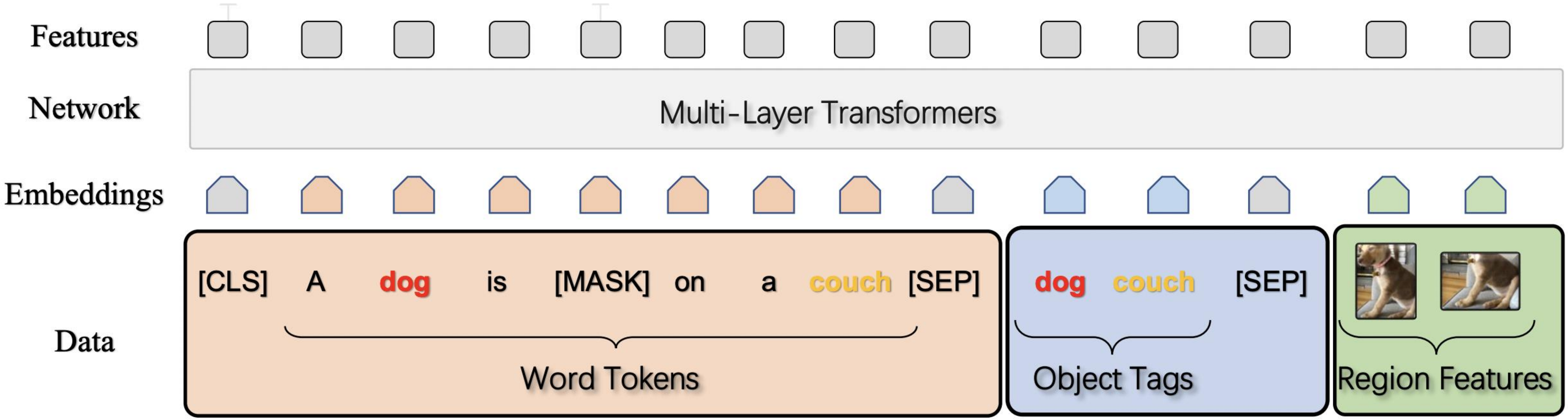
Fully-connected layer added to enable binary classification based on the fused vision-language token representation



Task: decide if 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

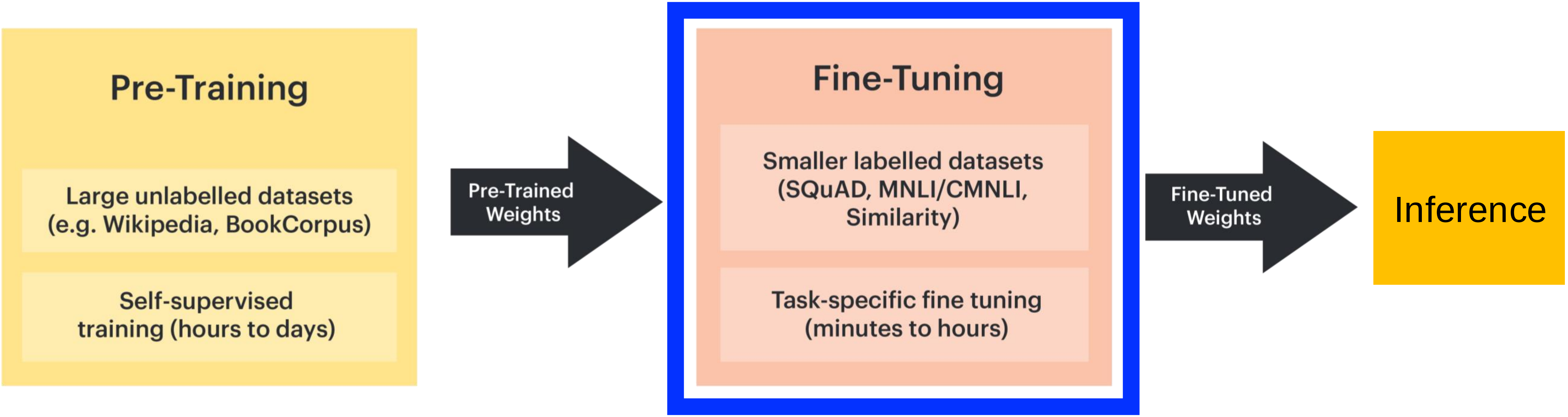
Oscar: 2 Pretraining Dataset



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

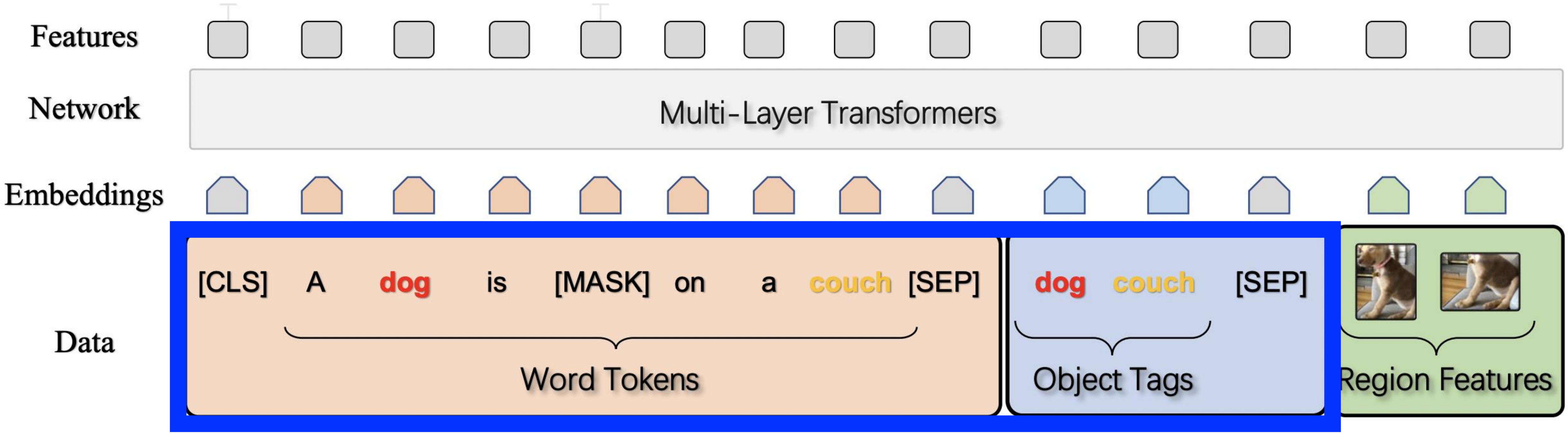
Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020

Oscar: Transformer Design



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

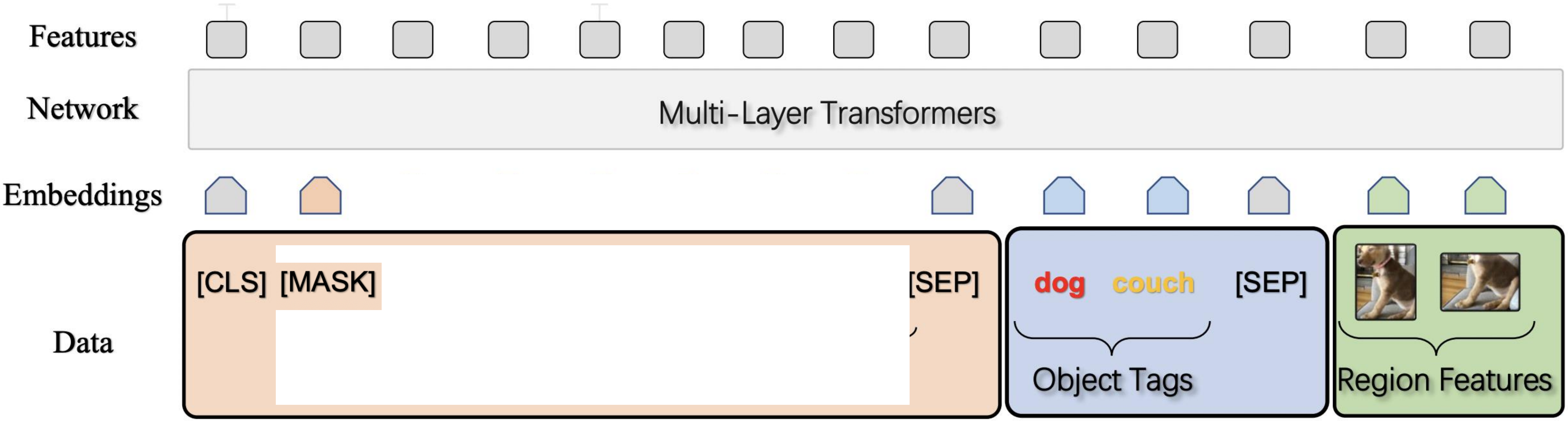
Oscar: 2 Fine-Tuning Task (Masked Token Loss)



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

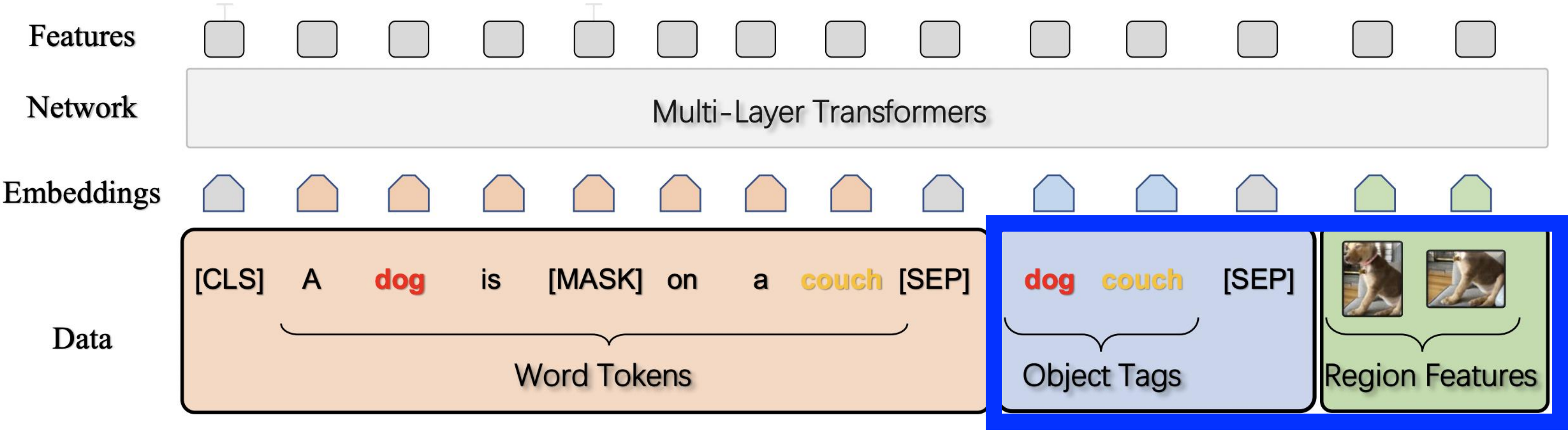


Repeatedly predict a new [MASK] token, incorporating the predicted word into the sequence, until [STOP] is predicted.

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

Idea: Oscar + Improved Object Detector

VinVL Architecture: Oscar + New Object Detector

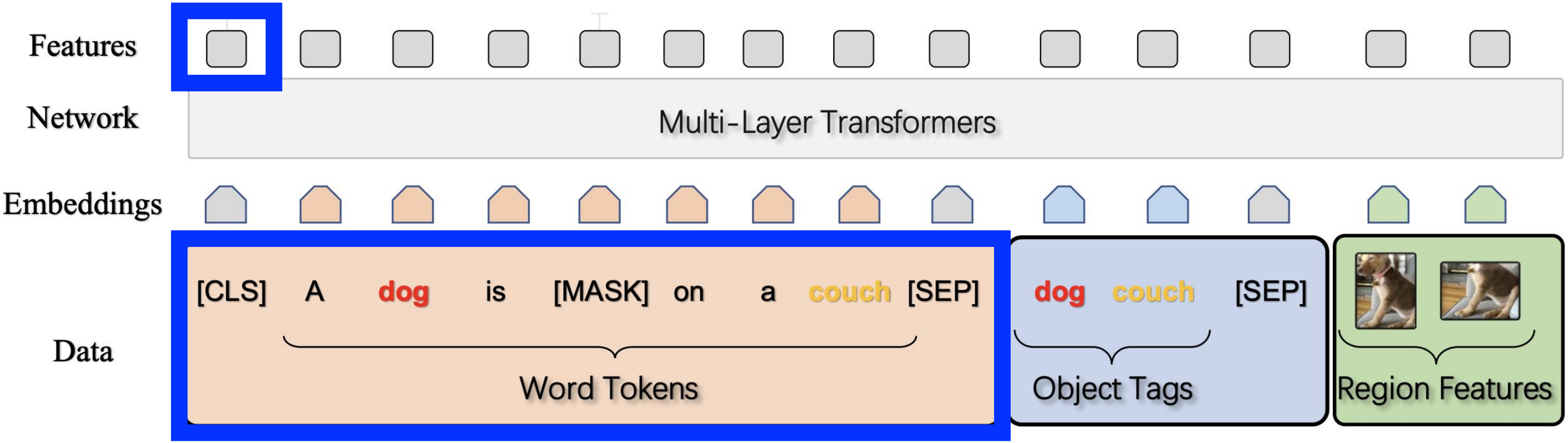


Improved object detector to predict more diverse categories and train larger models on larger datasets

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)

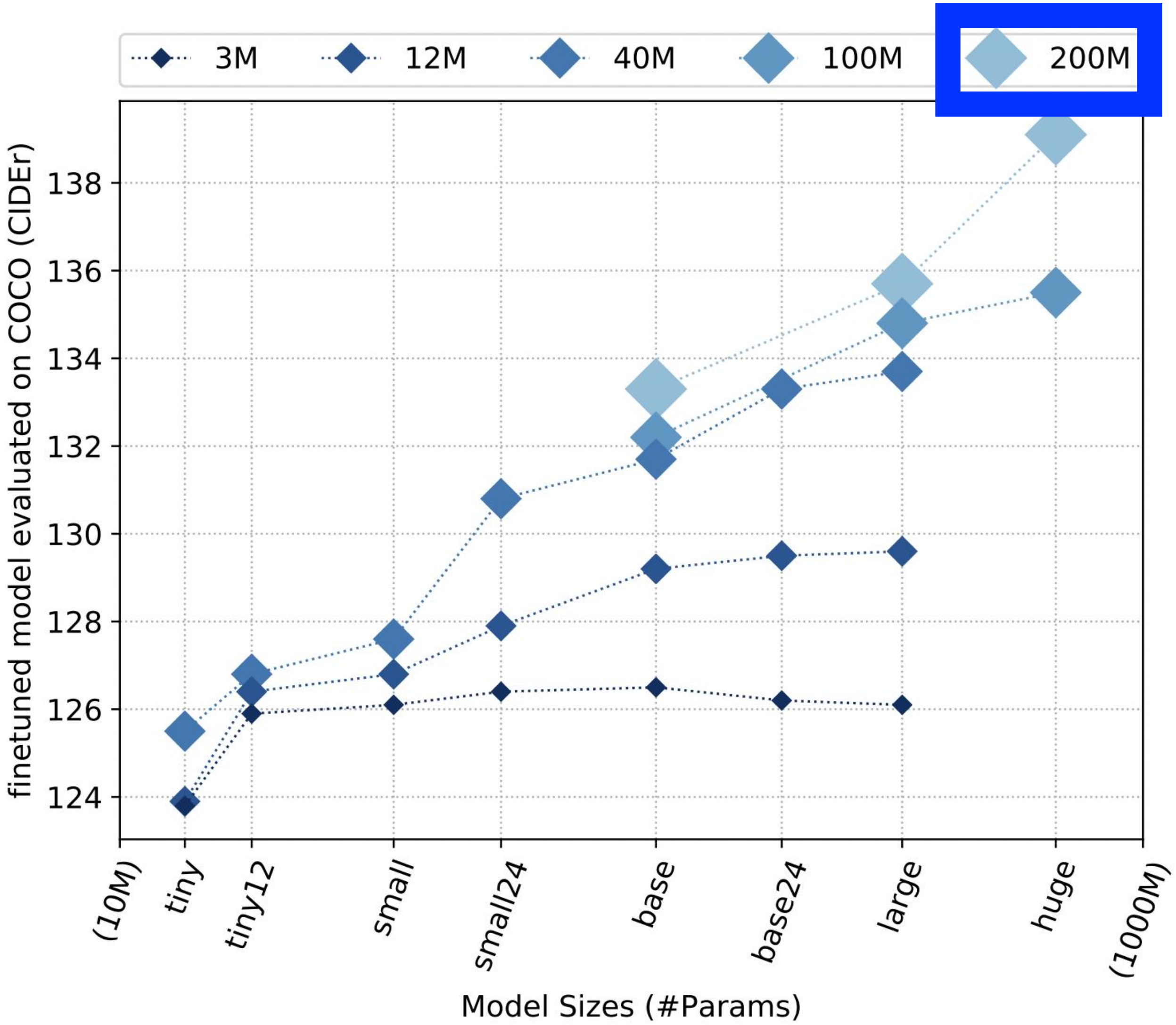
Fully-connected layer added to enable 3-way classification based on the fused vision-language token representation



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

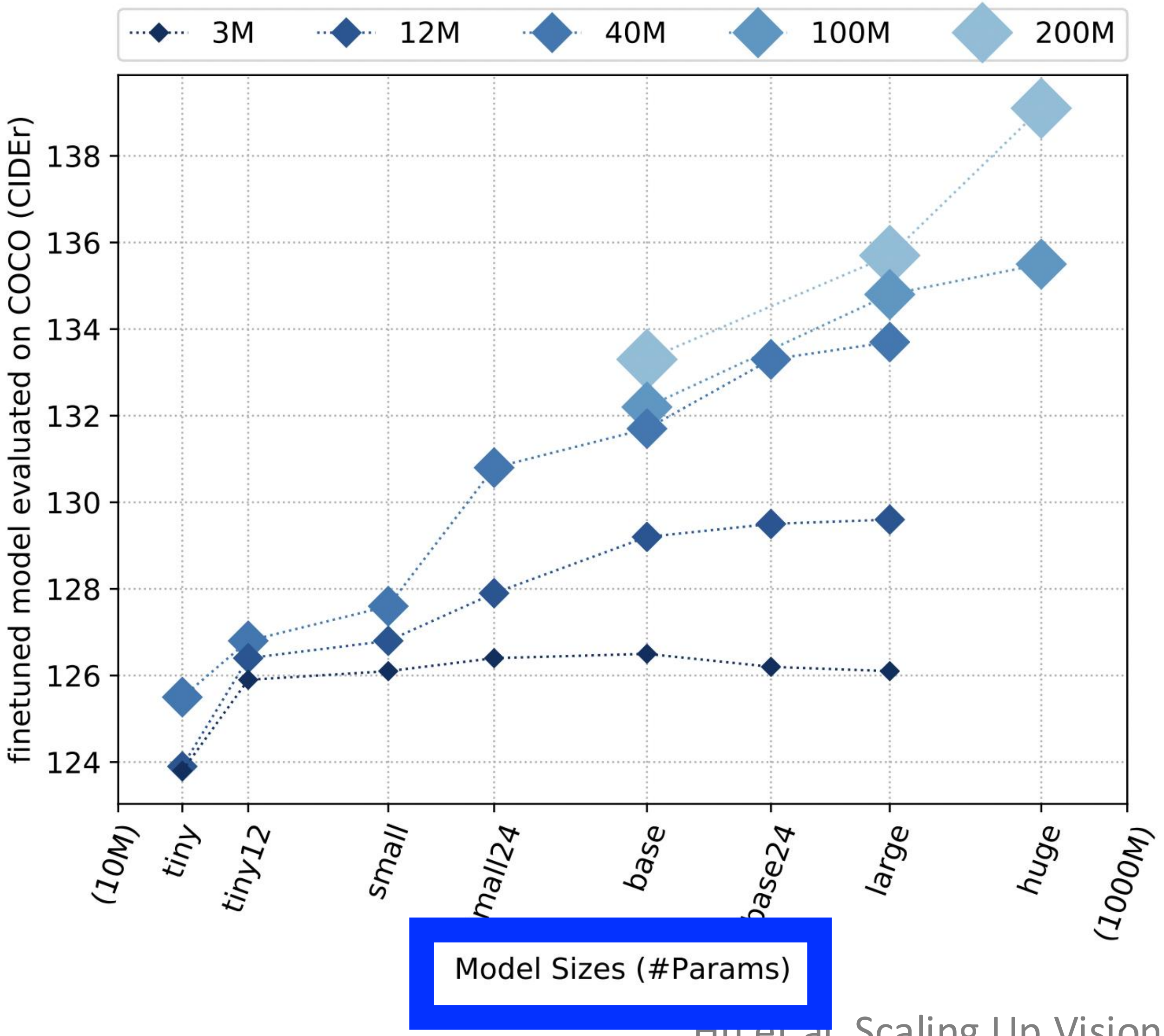
VinVL: Influence of Model and Dataset Sizes



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

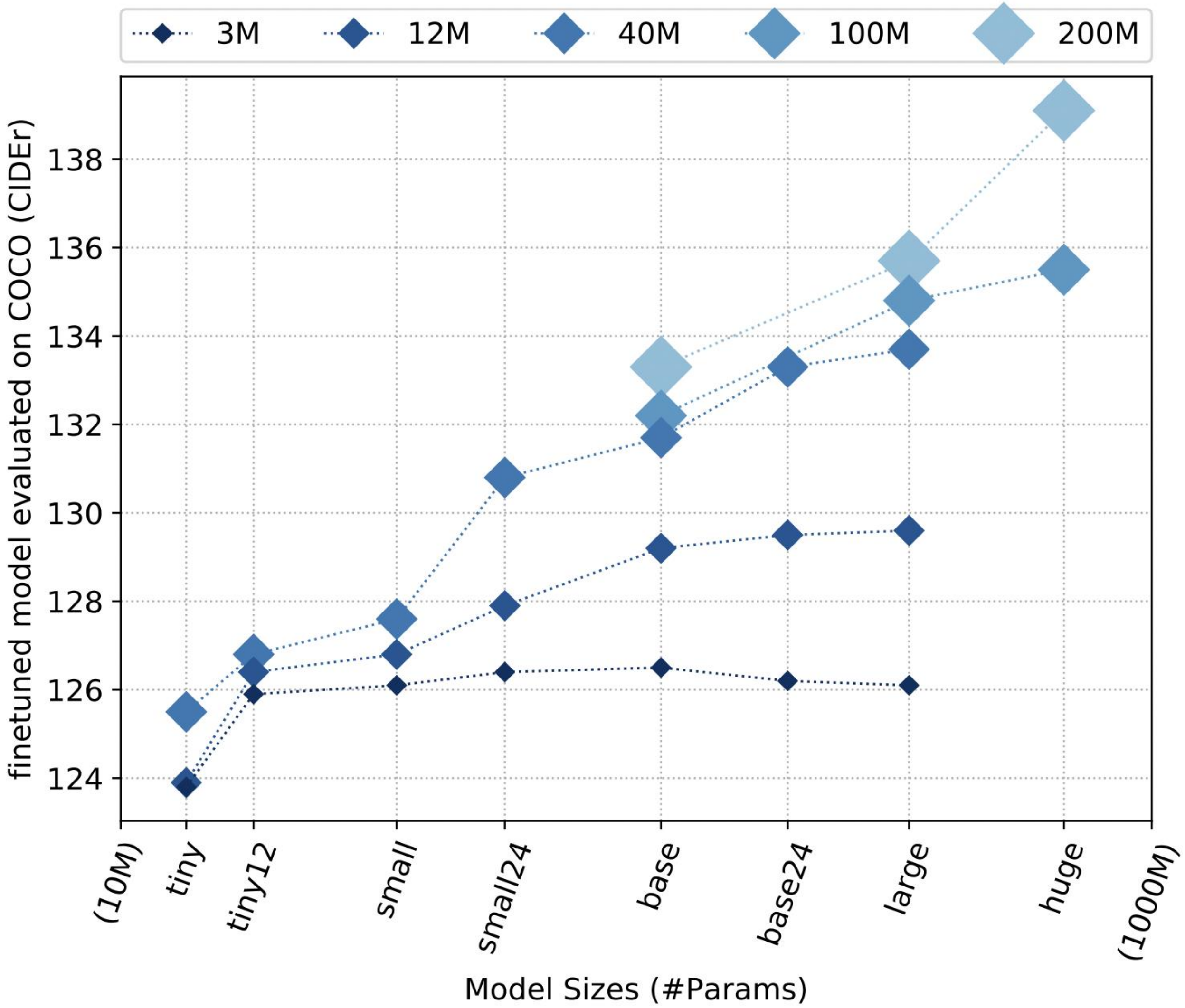


8 model sizes tested on COCO dataset

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?

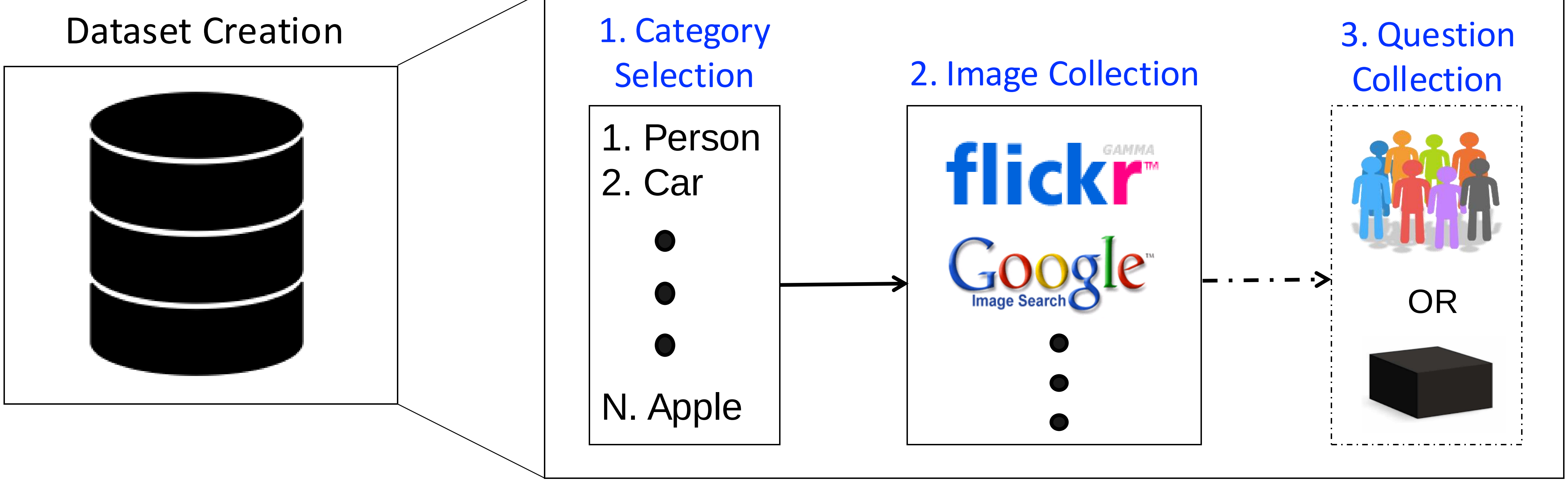
The trends of improved performance for large models and training datasets is generally observed for transformers

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

Today’s Topics

- Multimodal applications
- Image captioning dataset challenges
- Image captioning algorithms
- Visual question answering dataset challenges
- Discussion (chosen by YOU ☺)

Status Quo (Approach to Create 14+ Datasets)



Constrained set of concepts

Pre-qualified images (quality, privacy)

Contrived Questions

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.

e.g., Question Generation

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.

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.

e.g., Answer Generation

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?

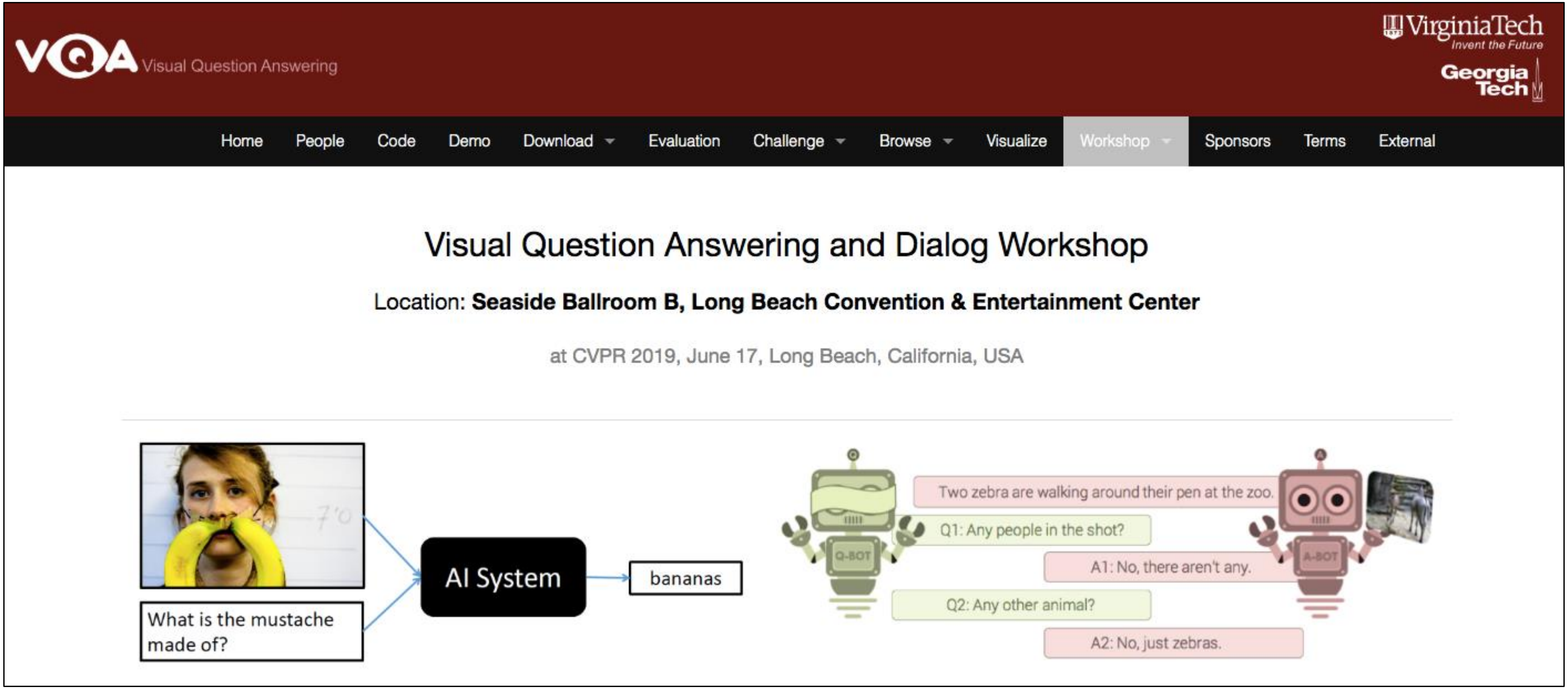
A1: Write your answer here.

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)

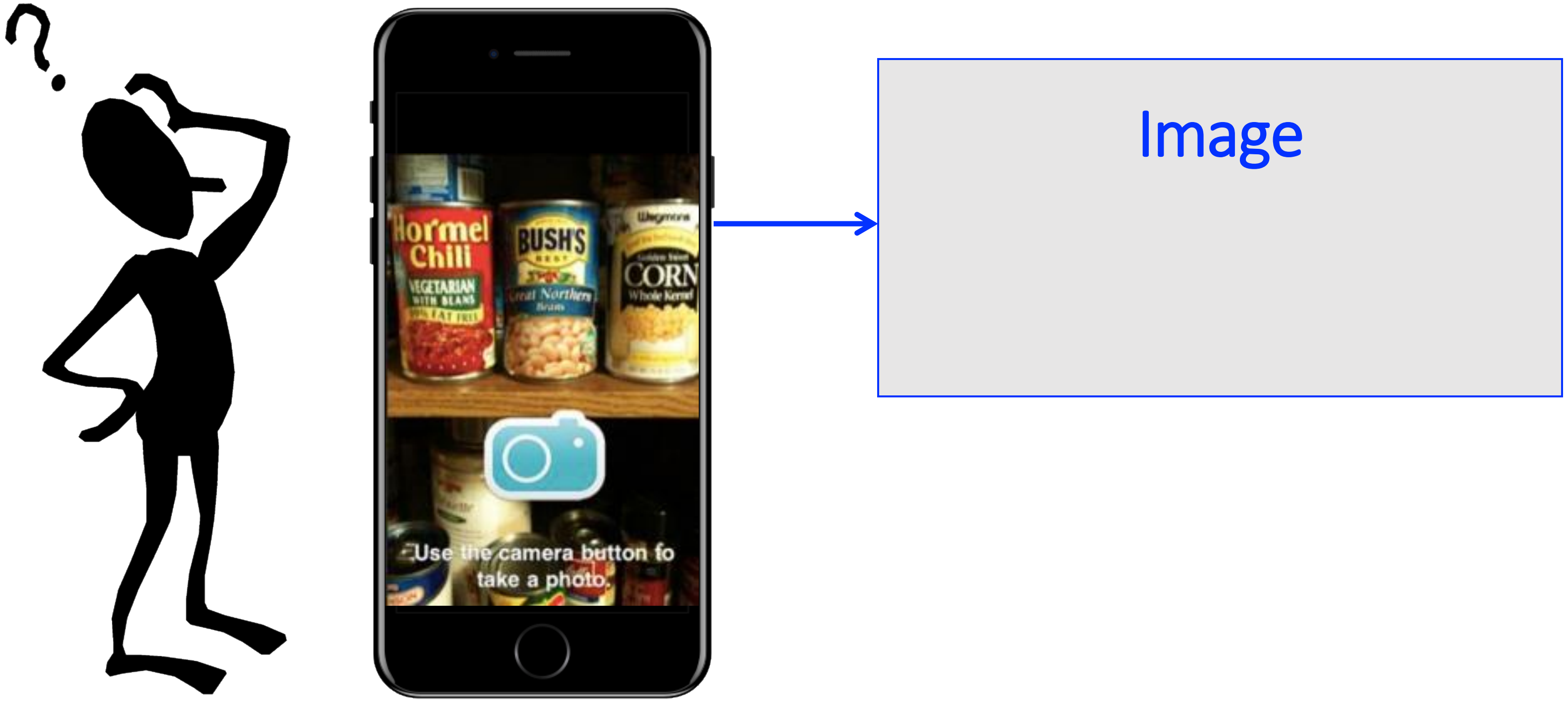


https://visualqa.org/workshop.html

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

Image

+

Question

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

VizWiz: Authentic Use Case

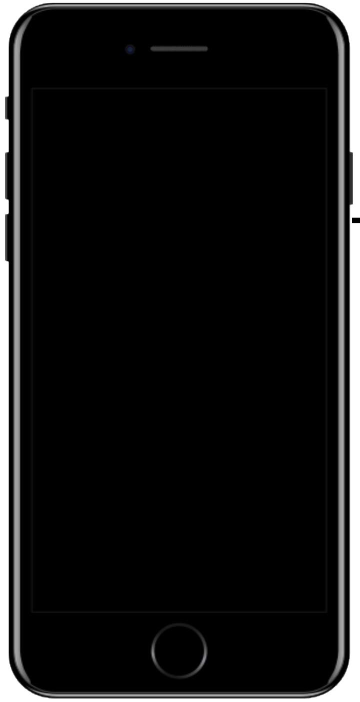
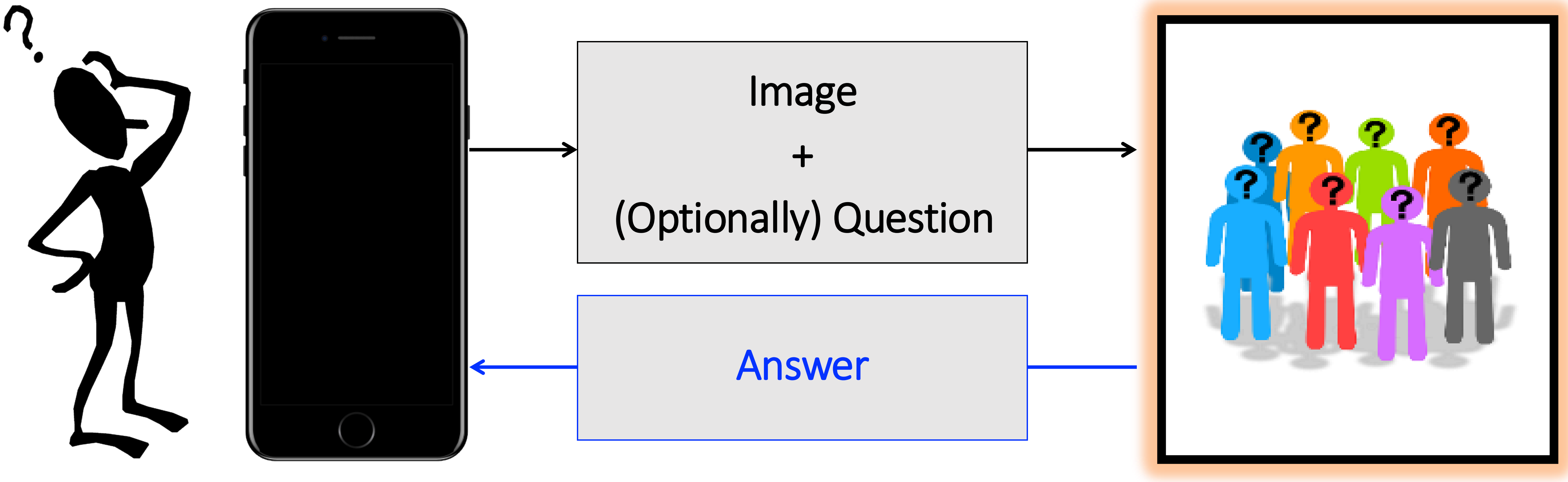


Image
+
Question

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

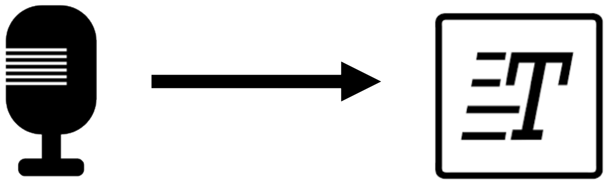
Users agreed to share 44,799 (62%) of requests for dataset creation

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

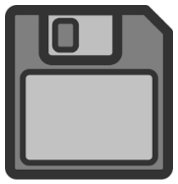
VizWiz: Authentic Use Case

Anonymization

1. Transcribe questions (removes voice)



2. Re-save images (removes metadata)

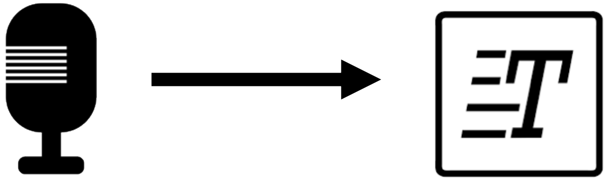


Gurari et al. CVPR 2018

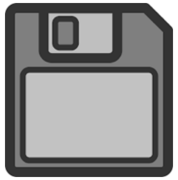
VizWiz: Authentic Use Case

Anonymization

1. Transcribe questions

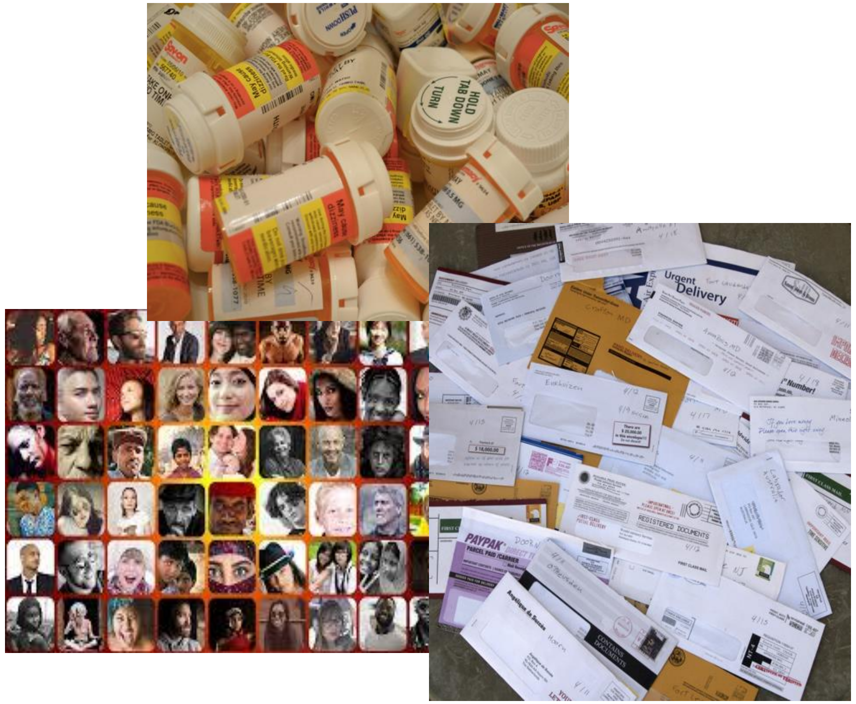


2. Re-save images



In-House Filtering

(personally identifying information)

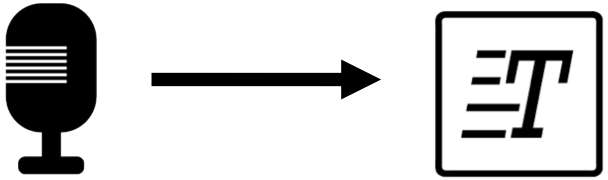


Gurari et al. CVPR 2018

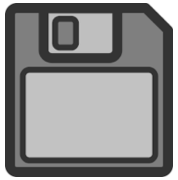
VizWiz: Authentic Use Case

Anonymization

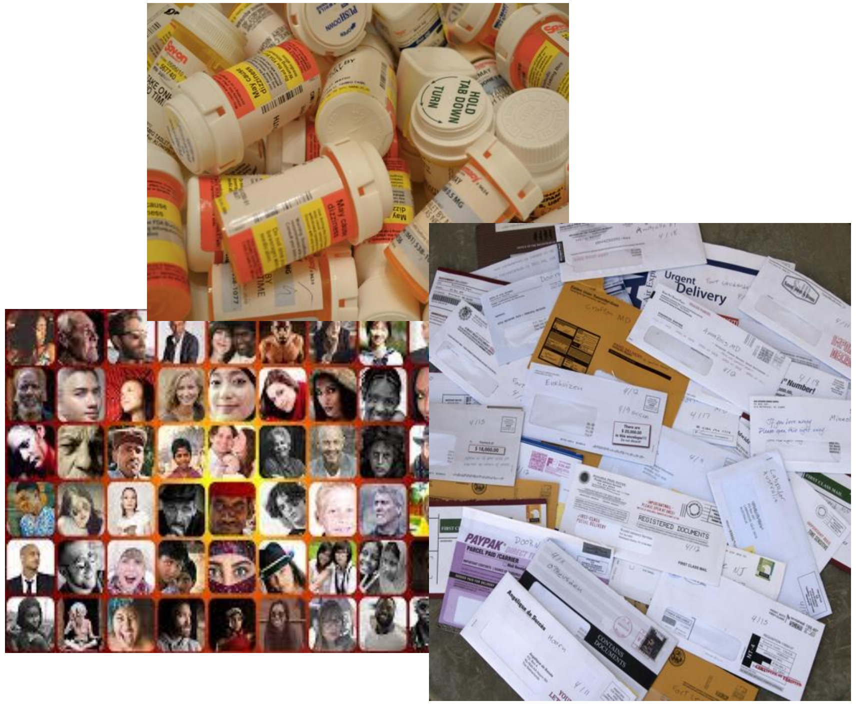
1. Transcribe questions



2. Re-save images



In-House Filtering



Data Labeling

(high quality answers)

amazon mechanicalturk

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 (6th year in 2024)

VizWiz   Home   Browse Dataset   Tasks & Datasets ⌄   Workshops ⌄   Acknowledgments

2024 VizWiz Grand Challenge Workshop

Overview

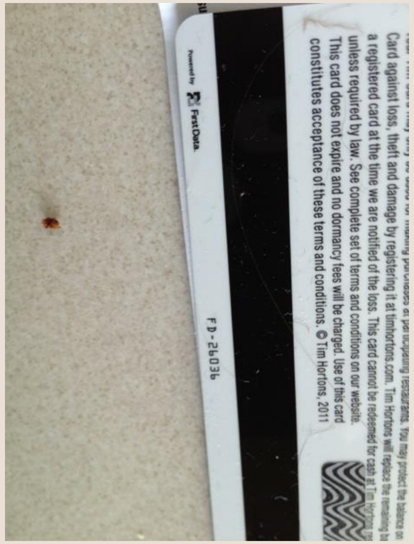
Our goal for this workshop is to educate researchers about the technological needs of people with vision impairments while empowering researchers to improve algorithms to meet these needs. A key component of this event will be to **track progress on six dataset challenges**, where the tasks are to answer visual questions, ground answers, recognize visual questions with multiple answer groundings, recognize objects in few-shot learning scenarios, locate objects in few-shot learning scenarios, and classify images in a zero-shot setting. The second key component of this event will be a discussion about current research and application issues, including invited speakers from both academia and industry who will share their experiences in building today's state-of-the-art assistive technologies as well as designing next-generation tools.

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?

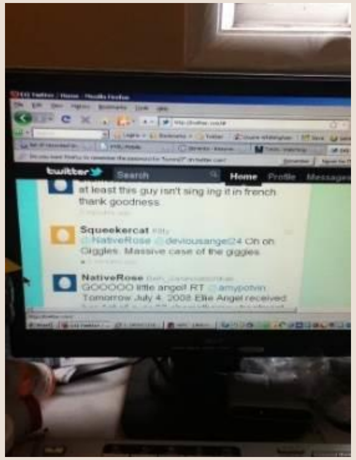
A: unanswerable

Q: What temperature is the dial set to?

A: unanswerable

[Gurari et al. CVPR 2018]

Class Task: Answer Visual Question



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?

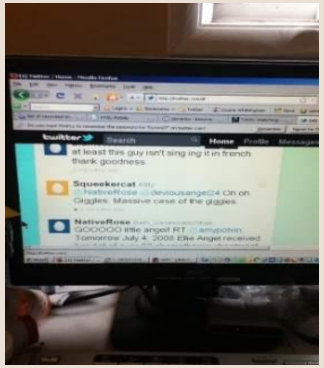
(4)

Fill out Google form

Class Task: Answer Visual Question

Fill out Google form

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

Discussion of models to come in next lecture

Today’s Topics

- Multimodal applications
- Image captioning dataset challenges
- Image captioning algorithms
- Visual question answering dataset challenges
- Discussion (chosen by YOU ☺)

Today’s Topics

- Multimodal applications
- Image captioning dataset challenges
- Image captioning algorithms
- Visual question answering dataset challenges
- Discussion (chosen by YOU ☺)

The End