Probing Contextual Language Models for Common Ground with Visual Representations


Gabriel Ilharco


Rowan Zellers


Ali Farhadi


Hannaneh Hajishirzi

UNIVERSITY of WASHINGTON

UWNLP

NAACL 2021

Motivation

How do **text representations** relate to the visual world?

Motivation

How do **text representations** relate to the visual world?

a dog is sleeping
on the floor

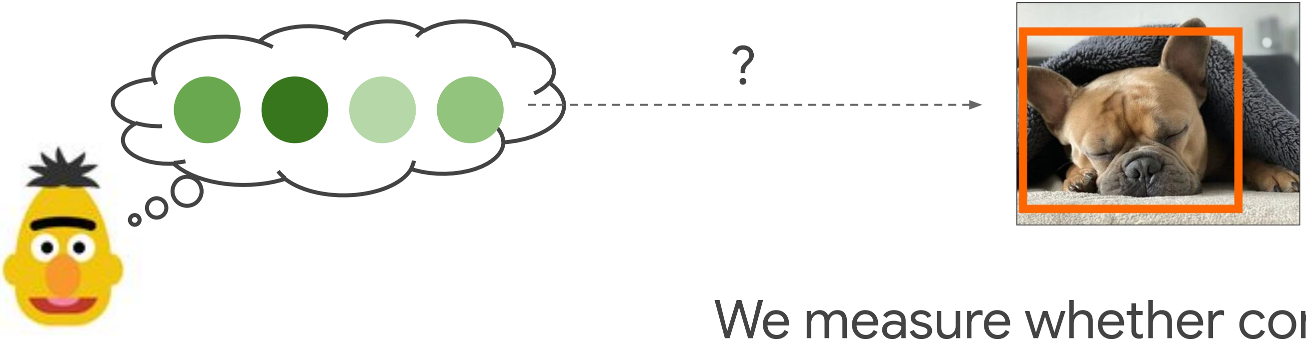
Motivation

How do **text representations** relate to the visual world?

a **dog** is sleeping
on the floor

Motivation

How do **text representations** relate to the visual world?

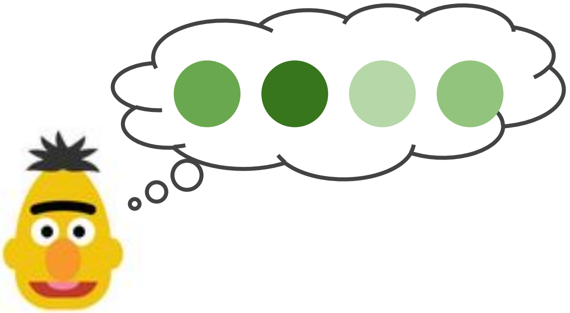


a **dog** is sleeping
on the floor

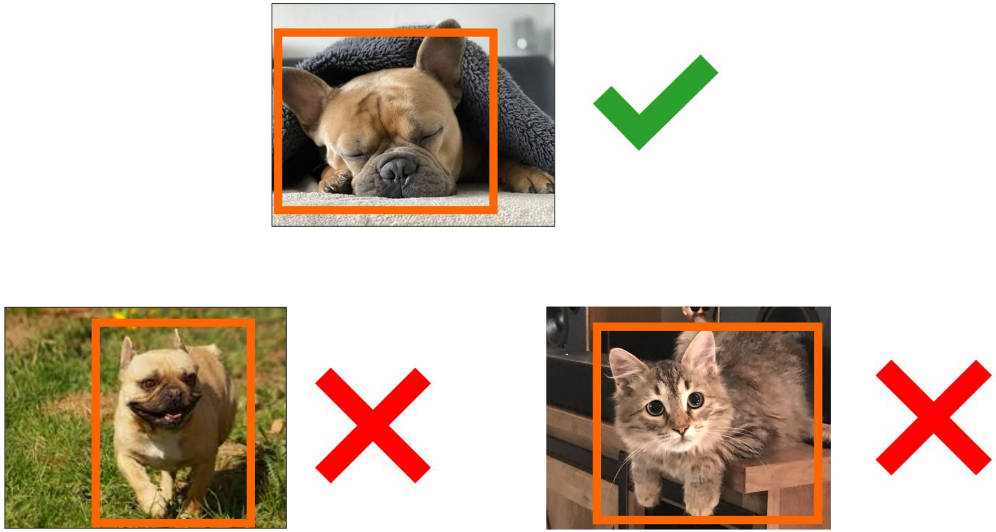
We measure whether contextual **text representations** of concrete **objects** are effective in finding aligned image patches

Motivation

Context in critical for this investigation



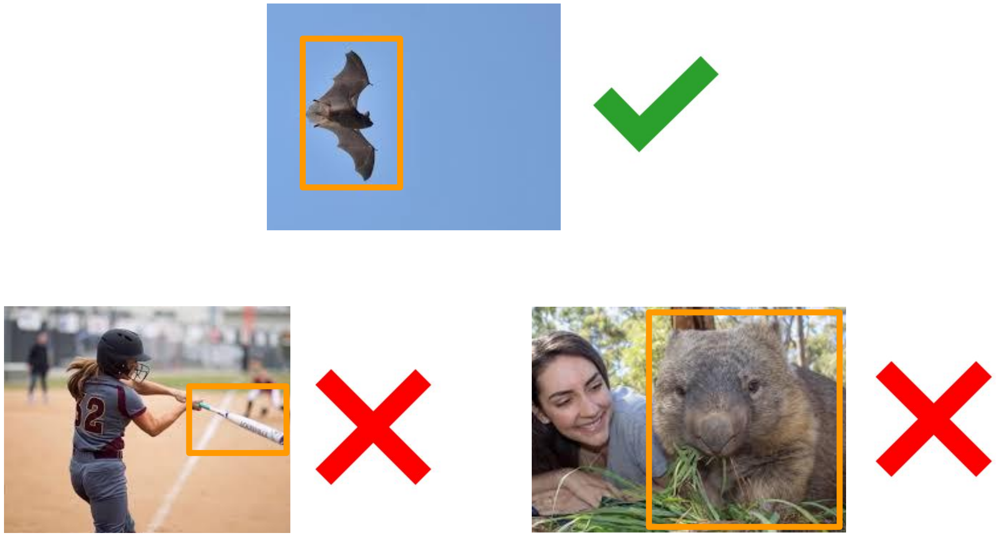
a **dog** is sleeping
on the floor



Motivation

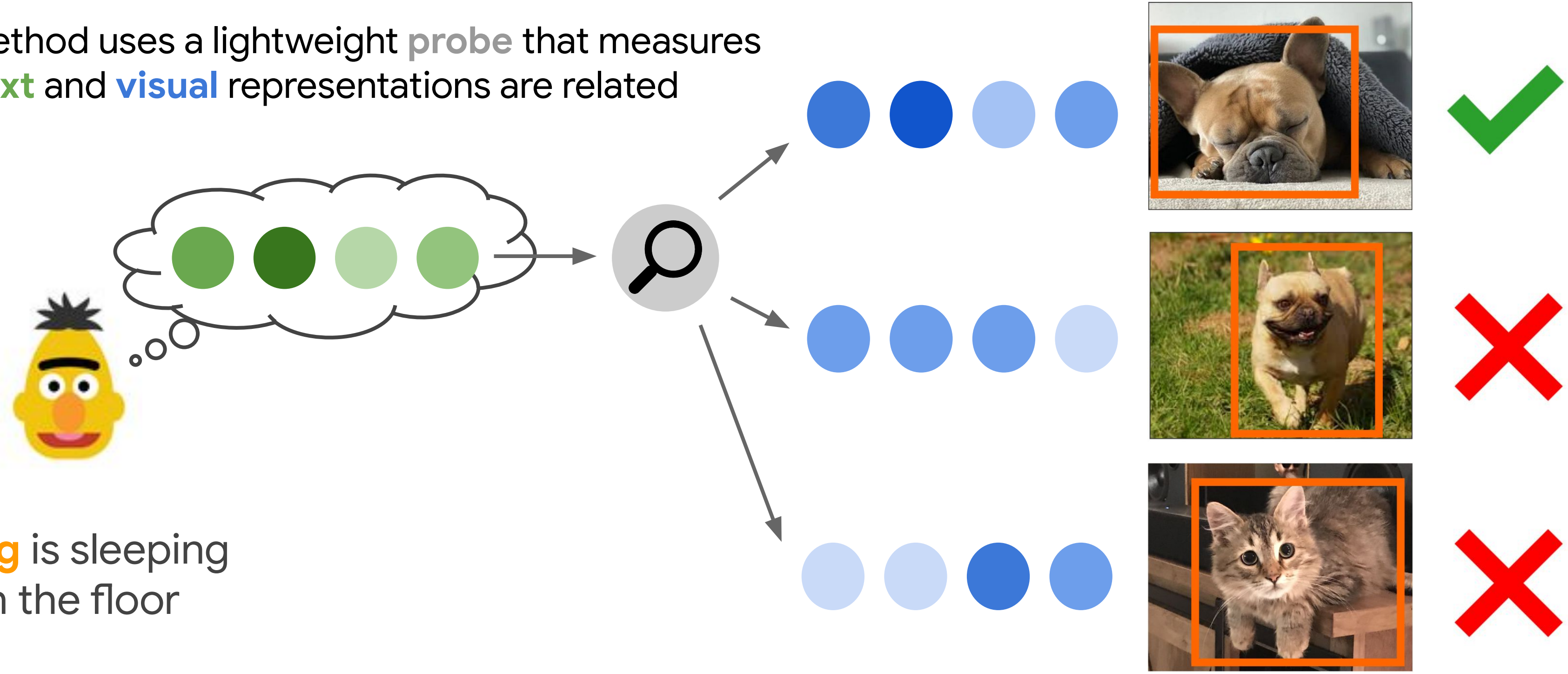
Context in critical for this investigation

A bat flying in the sky



Overview

Our method uses a lightweight **probe** that measures how **text** and **visual** representations are related



a **dog** is sleeping
on the floor

Method - collecting representations

We find aligned representations of concrete **objects**

a **dog** is chasing
an orange frisbee

Method - collecting representations

We find aligned representations of concrete **objects**

a **dog** is chasing
an orange frisbee

Method - collecting representations

From image captioning datasets, we find aligned pairs of **instances**

using a trained object detector

a dog is chasing an
orange frisbee

Method

From image captioning datasets, we find aligned pairs of **instances** using a trained object detector

a dog is chasing an orange frisbee

a **dog** is chasing an orange frisbee

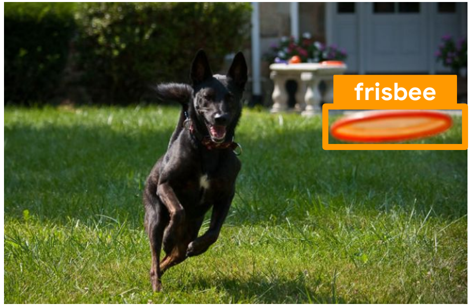
Method

From image captioning datasets, we find aligned pairs of **instances**

using a trained object detector

a dog is chasing an
orange frisbee

a **dog** is chasing an
orange frisbee

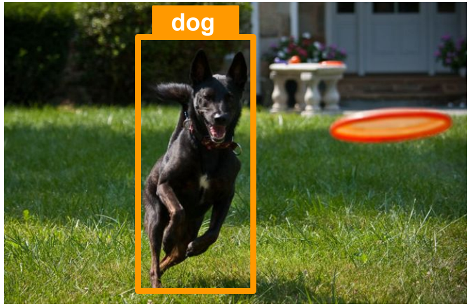


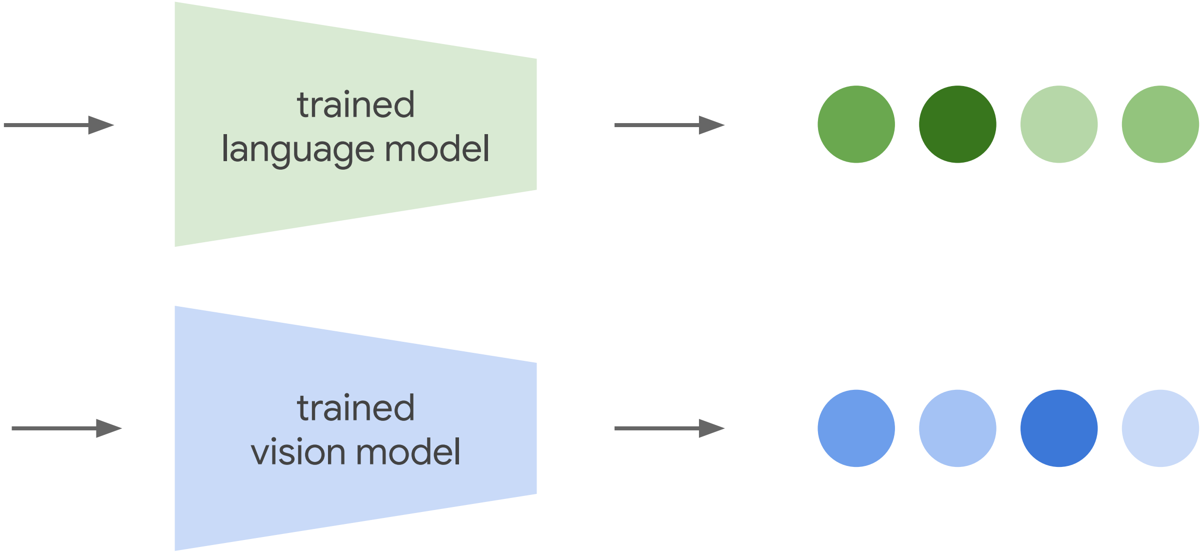
a dog is chasing an
orange **frisbee**

Method - Collecting Data

Text and **visual** representations are extracted by trained models

a **dog** is chasing an orange frisbee

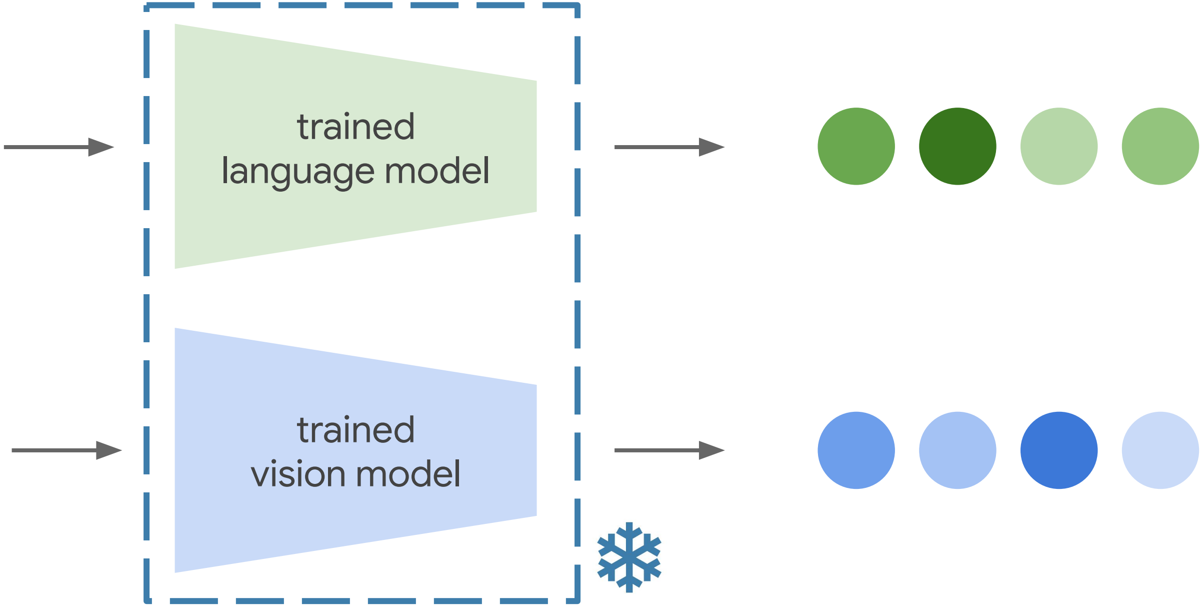




Method - Collecting Data

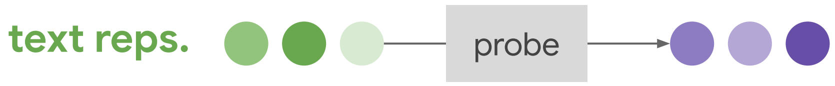
Text and **visual** representations are extracted by trained models

a **dog** is chasing an orange frisbee



Method - Inspecting Text Representations

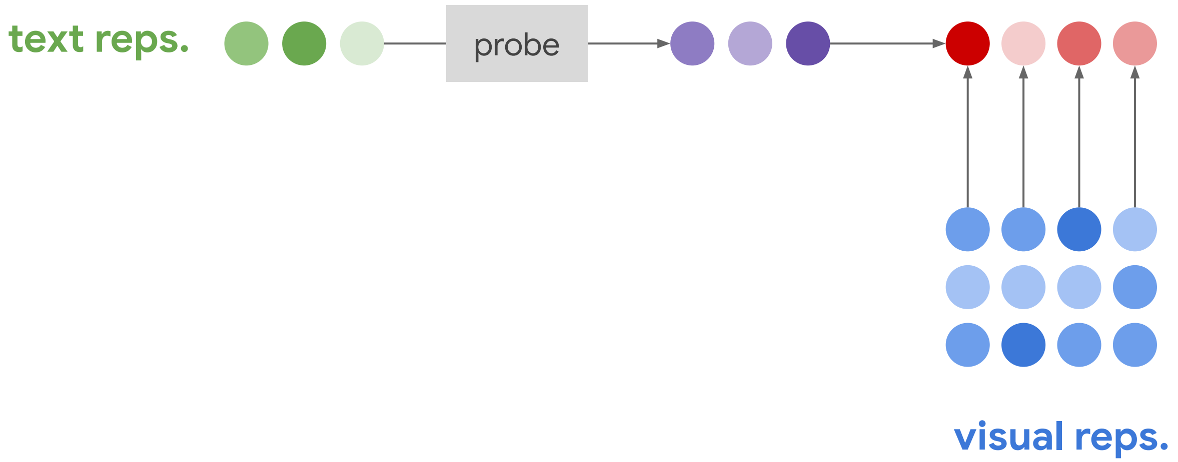
The **probe** maps **text representations** to the **visual domain**.



Method - Inspecting Text Representations

The **probe** maps **text representations** to the **visual domain**.

We compute the **dot product** between **projected representations** and **visual representations**

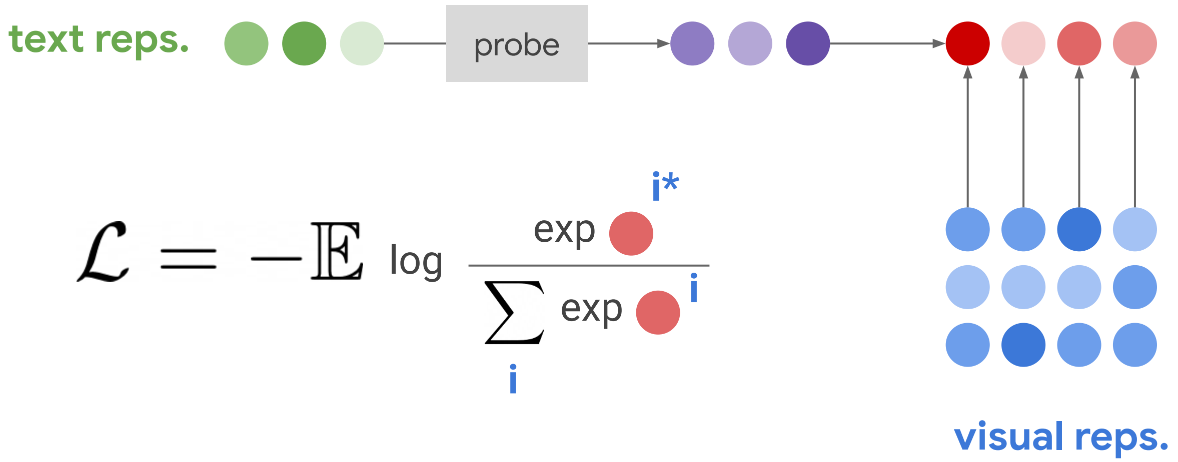


Method - Inspecting Text Representations

The probe maps **text representations** to the **visual domain.**

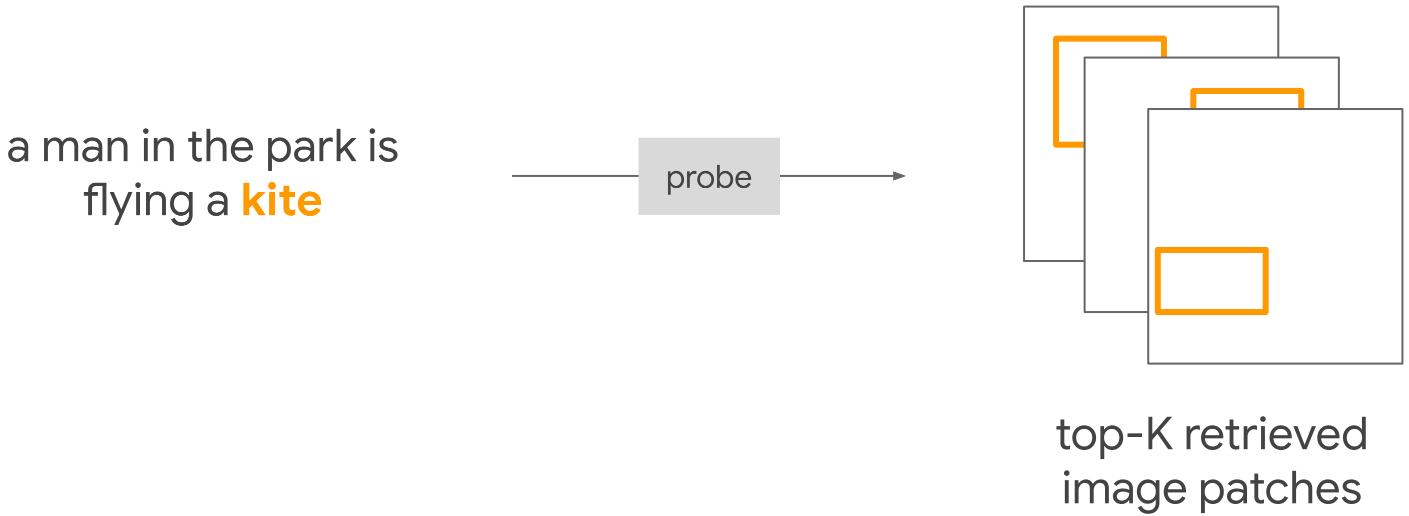
We compute the **dot product** between **projected representations** and **visual representations**

The probe is optimized via a **contrastive loss,** InfoNCE (Oord et al., 2018)



Method - Evaluation

We then evaluate by retrieving image patches of **unseen object categories**



Method - Evaluation

We then evaluate by retrieving image patches of **unseen object categories**.

We report two metrics:

- **Category Recall at K**:
    - how often an image patch of the correct object category was in the top-K

Method - Evaluation

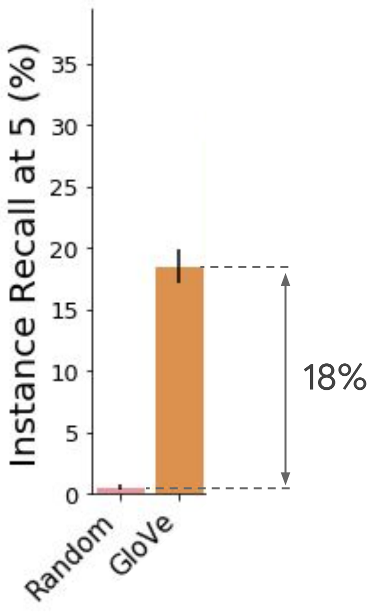
We then evaluate by retrieving image patches of **unseen object categories**.

We report two metrics:

- **Category Recall at K**:
    - how often an image patch of the correct object category was in the top-K
- **Instance Recall at K**:
    - how often the correct instance was in the top-K

Results

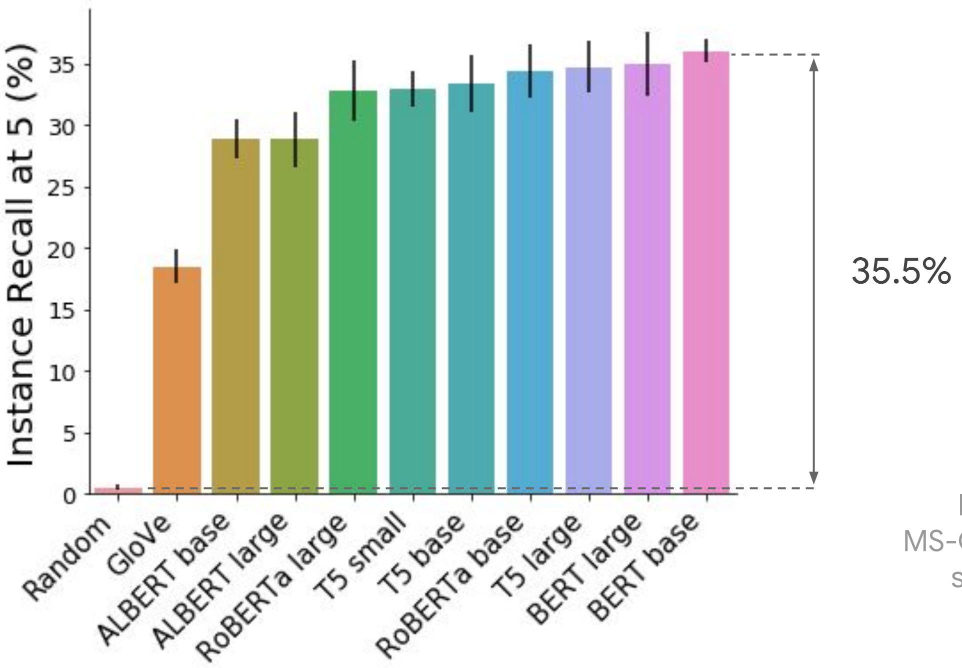
Language representations provide a strong signal for retrieval



Instance retrieval results on MS-COCO using 1000 test samples spanning 200 unseen object categories

Results

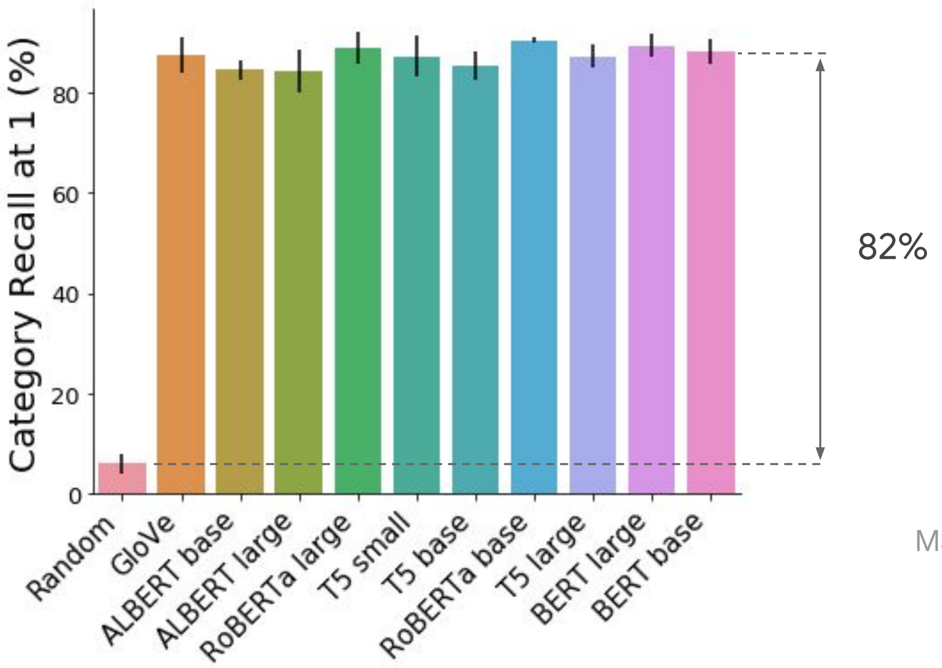
Language representations provide a strong signal for retrieval



Instance retrieval results on MS-COCO using 1000 test samples spanning 200 unseen object categories

Results

Language representations provide a strong signal for retrieval

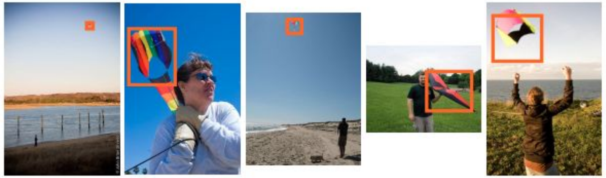


Category retrieval results on MS-COCO using 1000 test samples spanning 200 unseen object categories

Probing Contextual Language Models for Common Ground with Visual Representations. Ilharco et. al, 2021

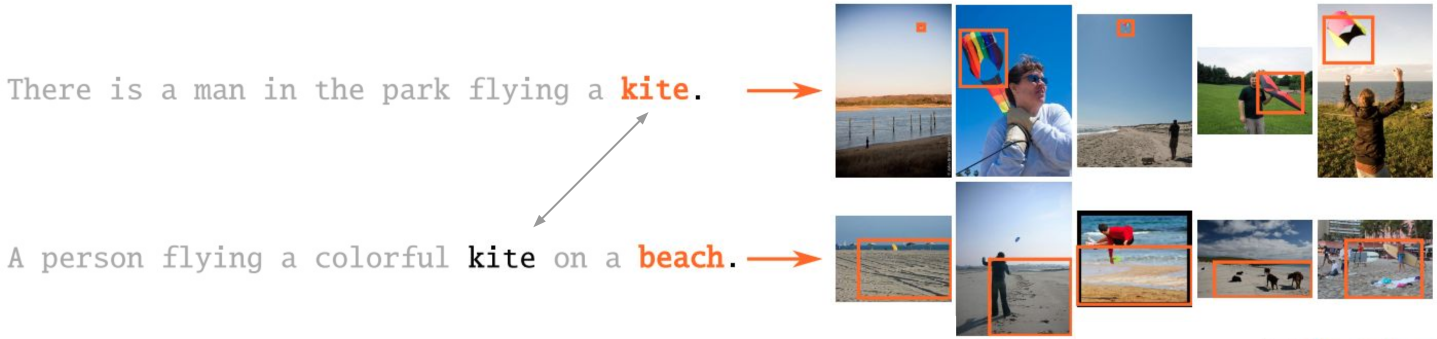
Results - Qualitative Results

There is a man in the park flying a **kite**. →



A person flying a colorful kite on a **beach**. →

Results - Qualitative Results



Results - Qualitative Results

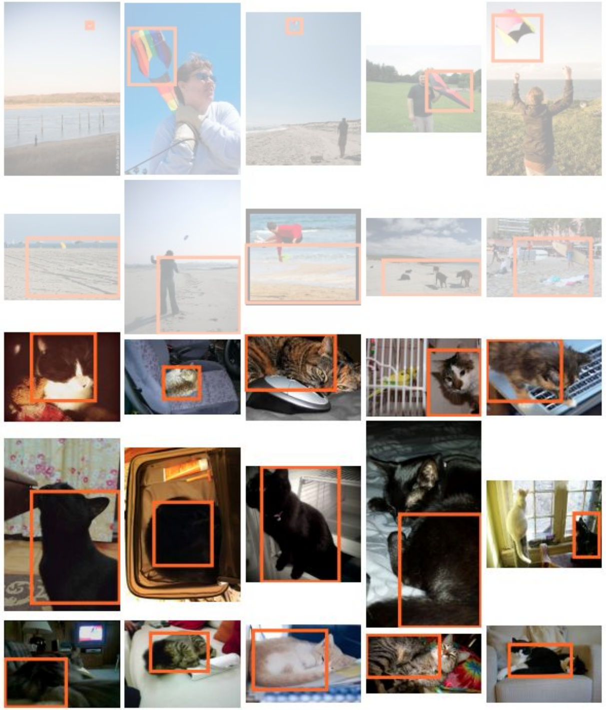
There is a man in the park flying a **kite**.

A person flying a colorful kite on a **beach**.

A **cat**.

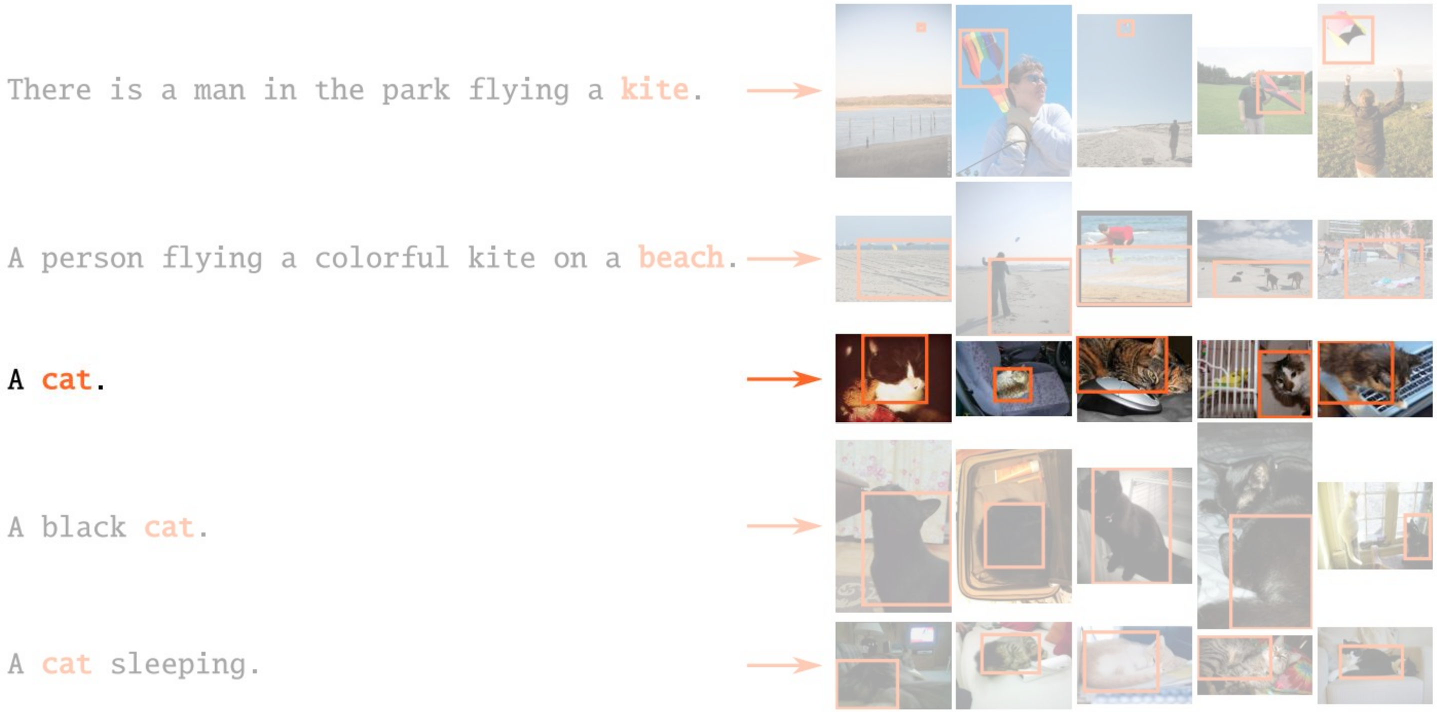
A black **cat**.

A **cat** sleeping.



Probing Contextual Language Models for Common Ground with Visual Representations. Ilharco et. al, 2021

Results - Qualitative Results



Probing Contextual Language Models for Common Ground with Visual Representations. Ilharco et. al, 2021

Results - Qualitative Results

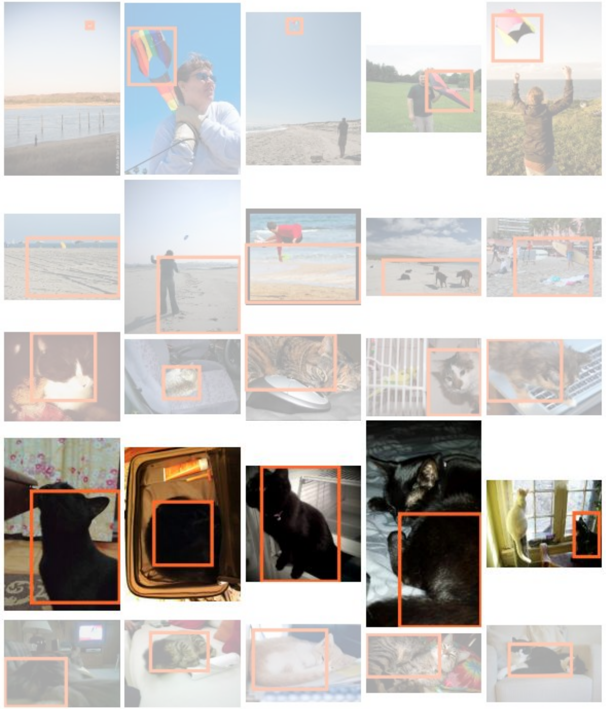
There is a man in the park flying a kite.

A person flying a colorful kite on a beach.

A cat.

A black cat.

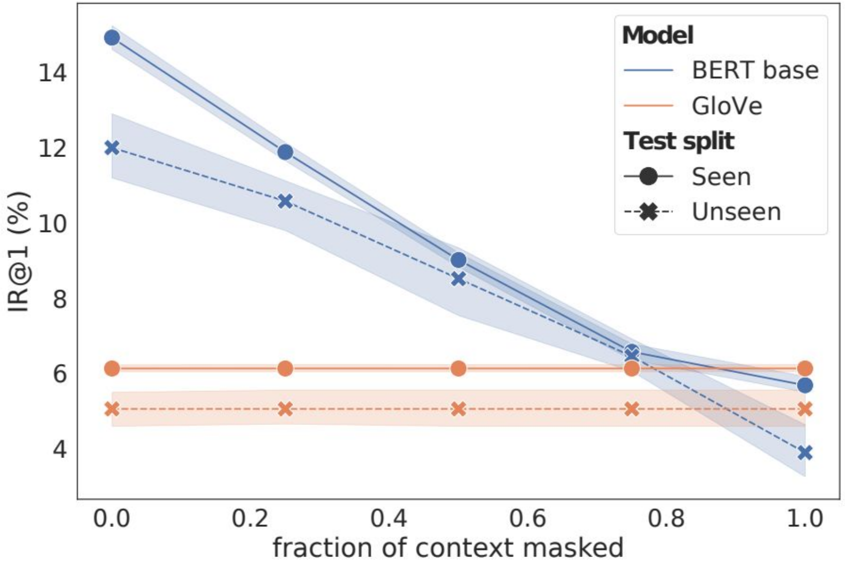
A cat sleeping.



Results - Qualitative Results

Results - Influence of context

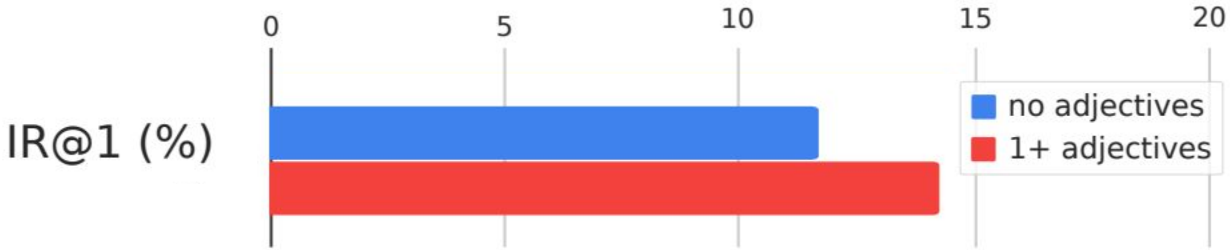
Performance of contextual models quickly degrades as context tokens are progressively masked out



Probing Contextual Language Models for Common Ground with Visual Representations. Ilharco et. al, 2021

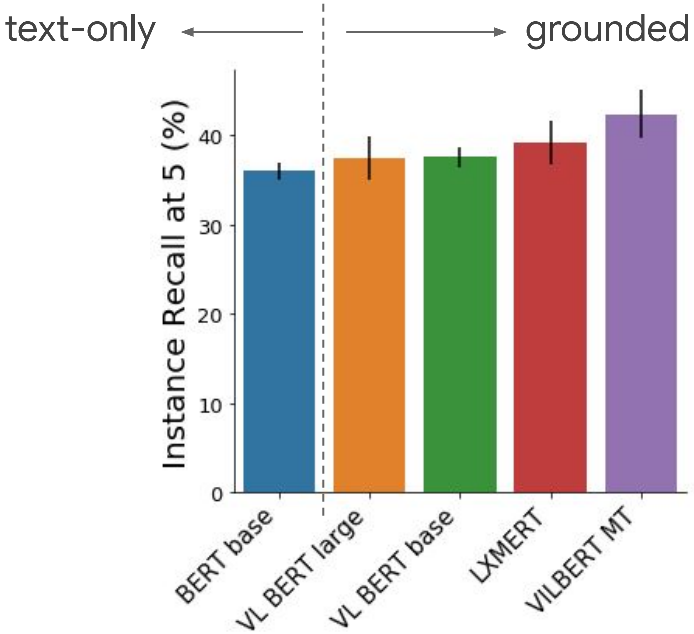
Results - Influence of context

More descriptive sentences lead to better retrieval:
performance increases when objects are accompanied by at least one adjective



Results - Grounded Models

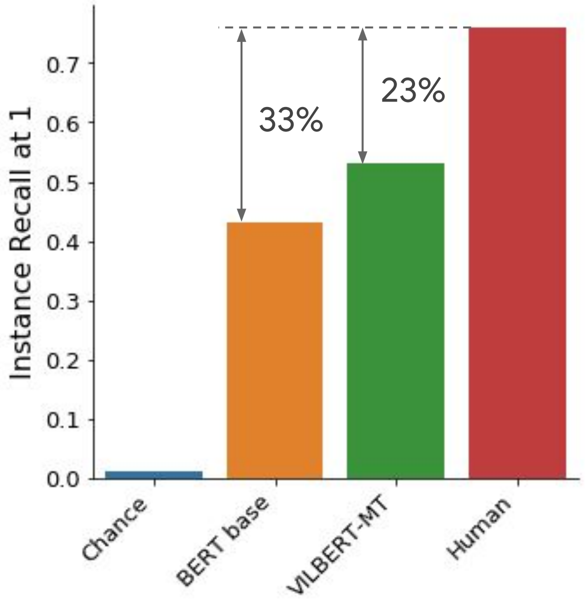
Grounded models slightly outperform text-only models



Probing Contextual Language Models for Common Ground with Visual Representations. Ilharco et. al, 2021

Results - Human Experiments

All models substantially underperform humans



Takeaways

Takeaways

- We introduce a method for measuring similarities between text and visual representations

Takeaways

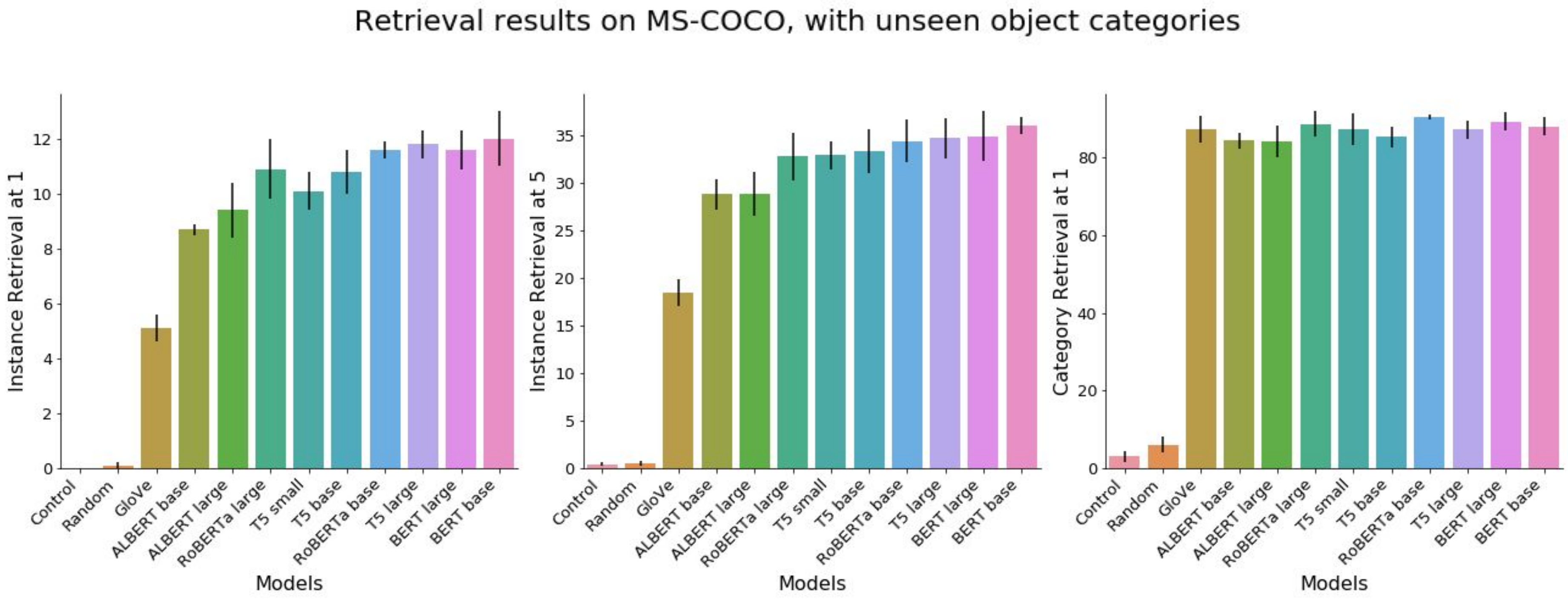
- We introduce a method for measuring similarities between text and visual representations
- Contextual language representations are useful in finding aligned image patches
    - We explore how results are affected by variables such as context and explicit grounding during training

Takeaways

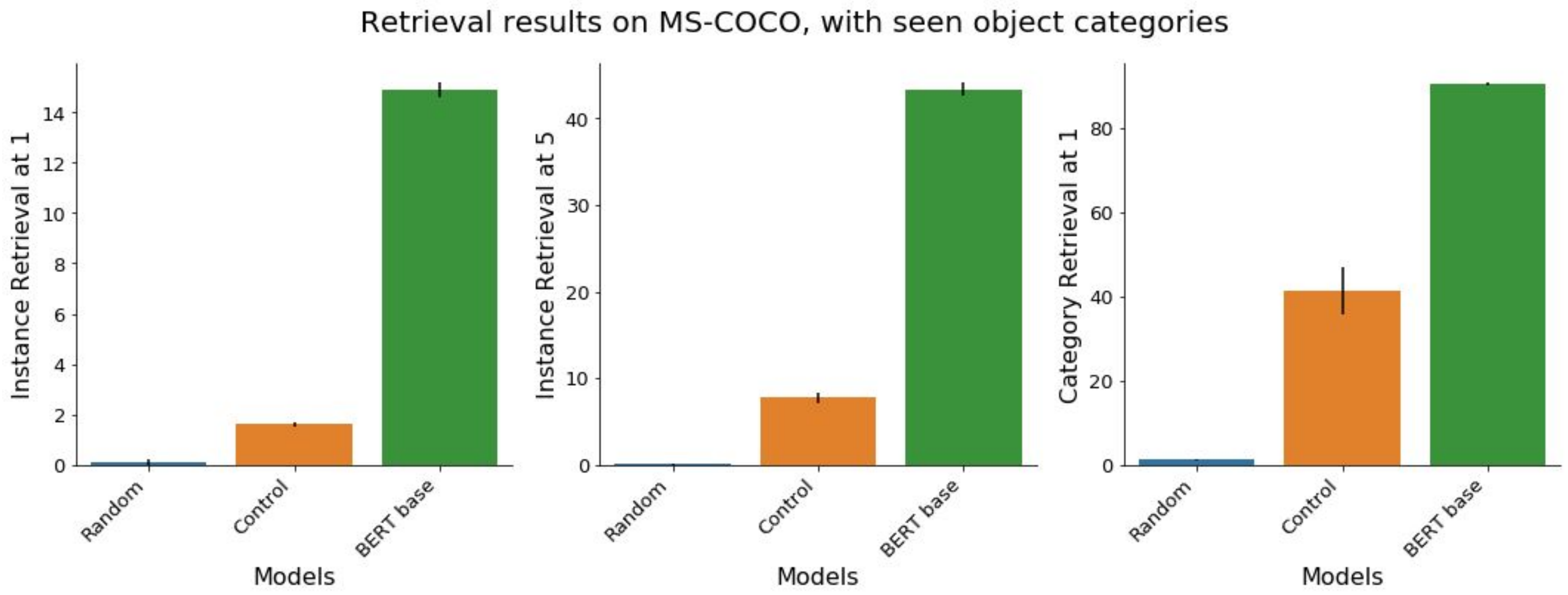
- We introduce a method for measuring similarities between text and visual representations
- Contextual language representations are useful in finding aligned image patches
  - We explore how results are affected by variables such as context and explicit grounding during training
- All studied models significantly underperform humans, showing much room for future progress

Thank you!

Results - Control



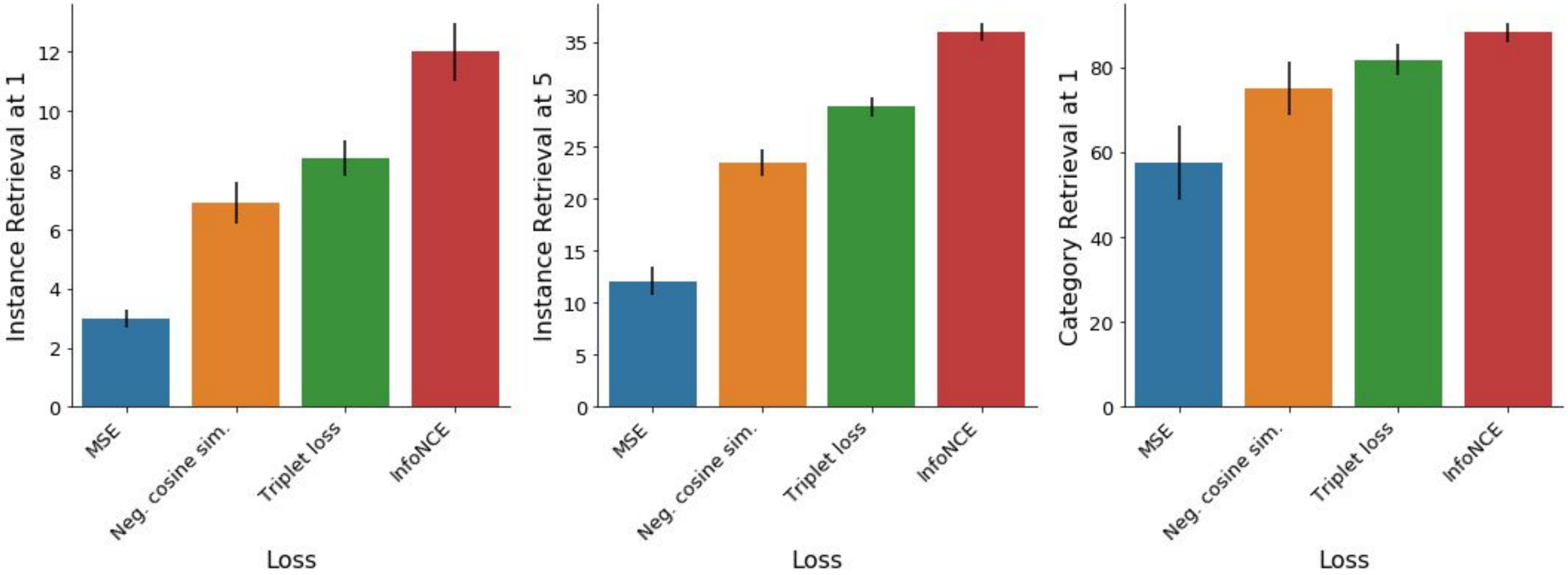
Results - Seen object categories



Probing Contextual Language Models for Common Ground with Visual Representations. Ilharco et. al, 2021

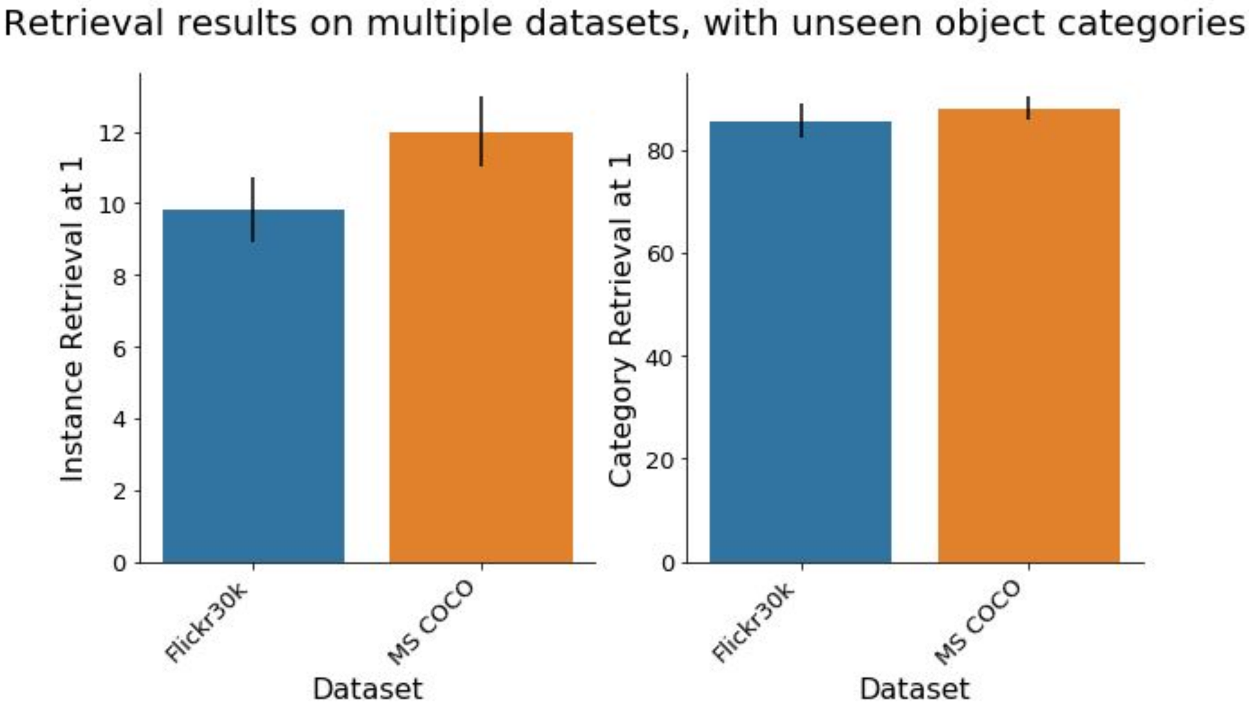
Results - Loss ablations

Retrieval results on MS-COCO, with unseen object categories

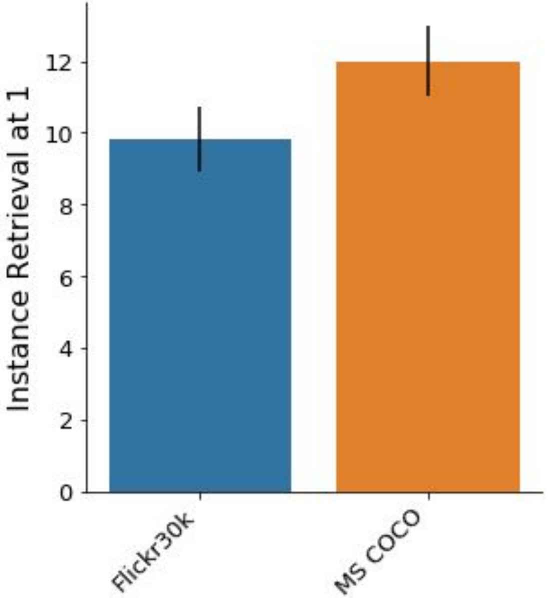


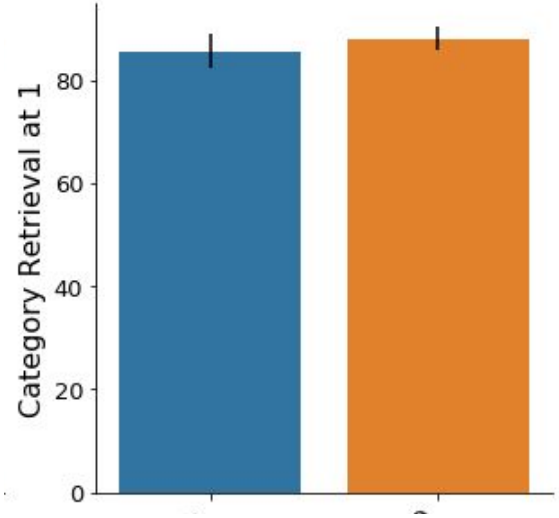
Probing Contextual Language Models for Common Ground with Visual Representations. Ilharco et. al, 2021

Results - Data ablations



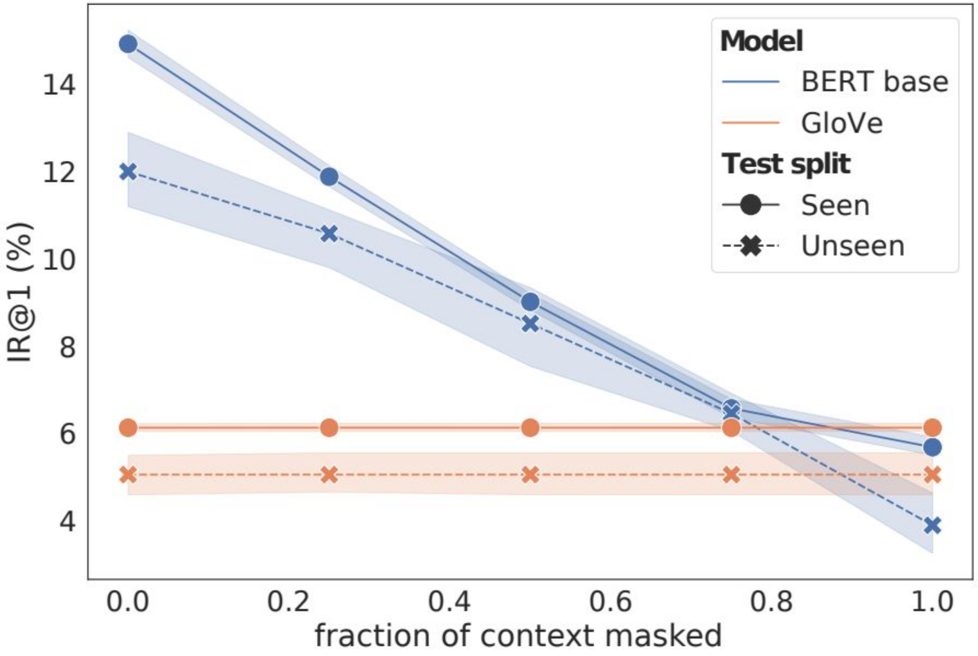
Results - Data ablations





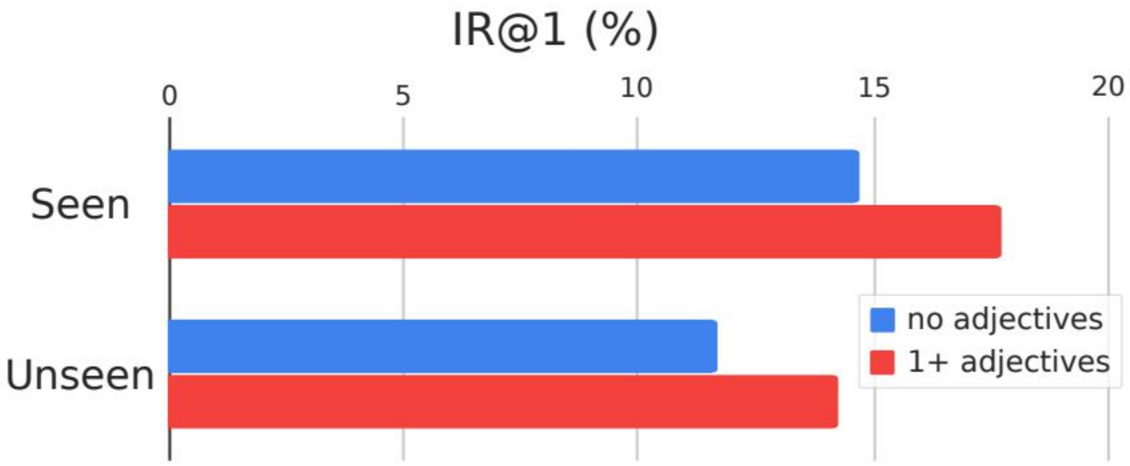
Probing Contextual Language Models for Common Ground with Visual Representations. Ilharco et. al, 2021

Results - Influence of context

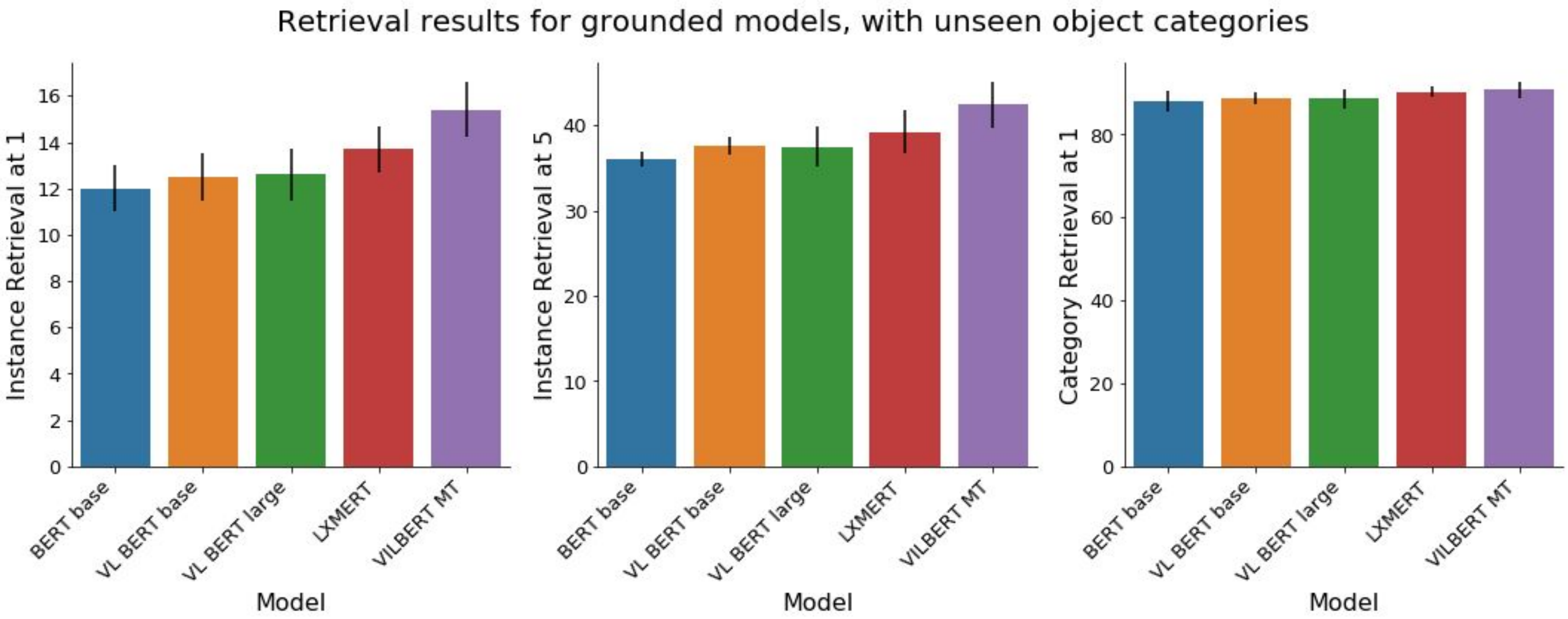


Probing Contextual Language Models for Common Ground with Visual Representations. Ilharco et. al, 2021

Results - Influence of context

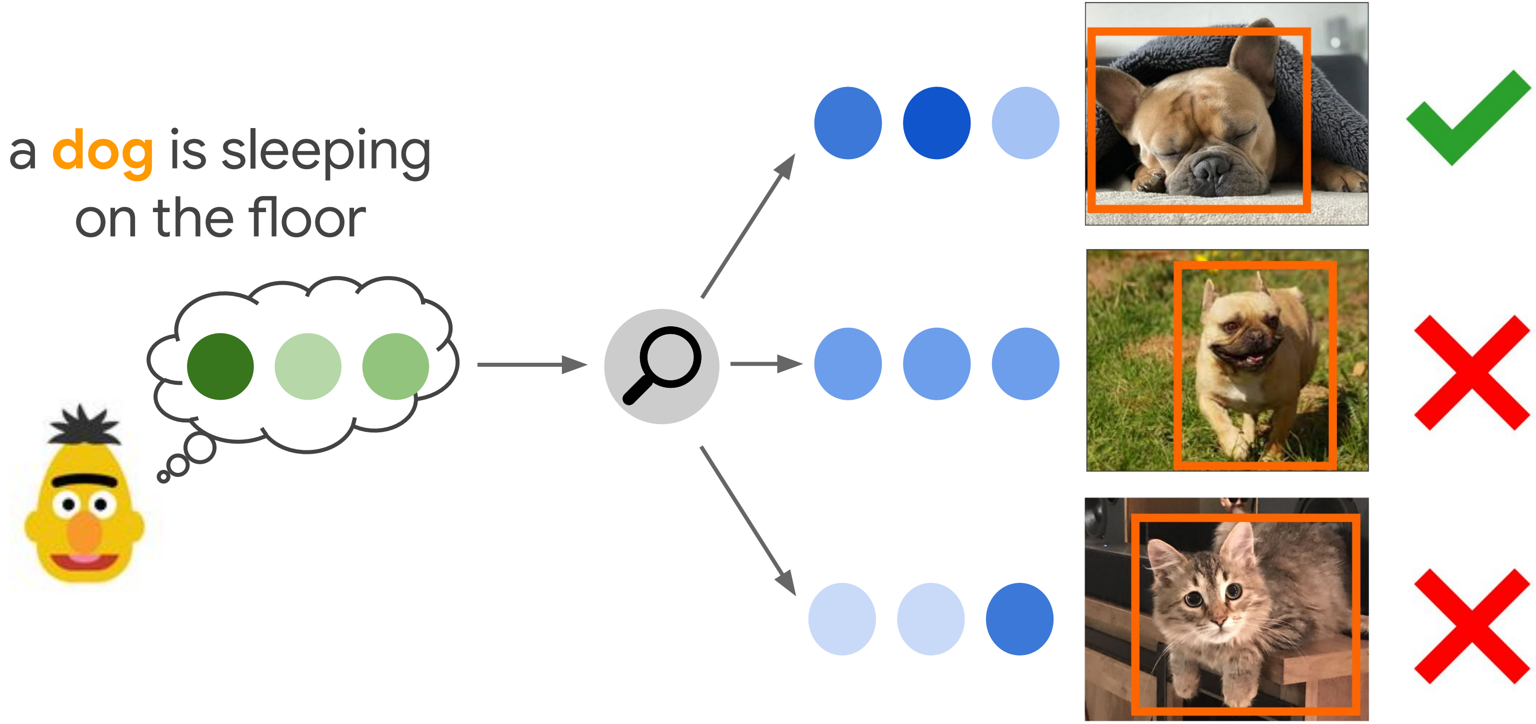


Results - Grounded Models



Probing Contextual Language Models for Common Ground with Visual Representations. Ilharco et. al, 2021

Results - Grounded Models



Probing Contextual Language Models for Common Ground with Visual Representations. Ilharco et. al, 2021