

Multimodal CL and NLP: Combining Language and Vision for (Computational) Semantics

Carina Silberer

IMS, University of Stuttgart

CL Fall School 2024, Passau
16 - 20 September 2024

Outline

Introduction: Multimodal NLP

Basics: Multimodal Representations

NLP/CL: Linguistic Representations

Computer Vision: Visual Representations I

Appendix: CV Datasets

Visually Grounded Representations

Tasks and Applications in Multimodal NLP

Limitations of Models for NLU

Current Challenges

Multimodal (VL) Representations



<http://pixabay.com/photos/photo-collection-pictures-photos-382018>



Vision
images

Language
text



Computer Vision: Transformer-Based Approaches

ViT: Image recognition

[Dosovitskiy et al., 2021]

- ▶ End-to-end (only self-attention over pixels, no conv layers)
- ▶ The input image is split into fixed-size image patches
e.g., 16×16 or 32×32
- ▶ The patches are mapped to vector embeddings
- ▶ Positional embeddings and special *classification token* are added (cf. NLP)
- ⇒ Feed to encoder transformer (see BERT)

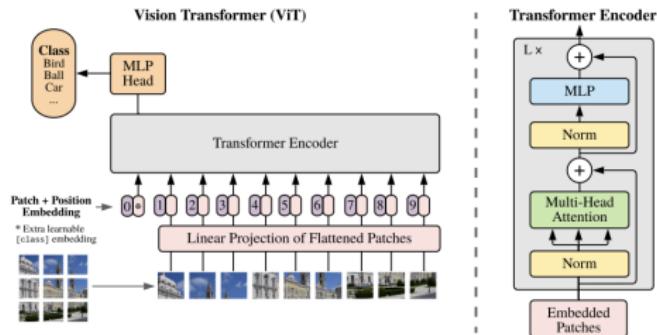


Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In addition, for classification, we add a learnable extra embedding, called [class] embedding.

Outline

Introduction: Multimodal NLP

Basics: Multimodal Representations

NLP/CL: Linguistic Representations

Computer Vision: Visual Representations I

Appendix: CV Datasets

Visually Grounded Representations

Tasks and Applications in Multimodal NLP

Limitations of Models for NLU

Current Challenges

APPENDIX

Deep NN Architectures for Object Detection

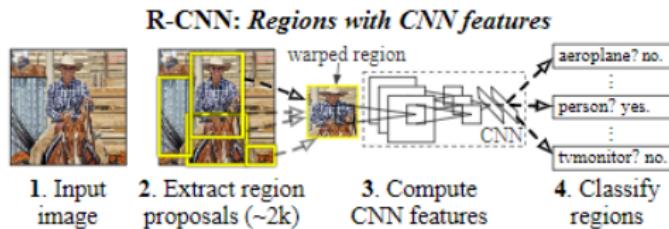
Well Known Architectures

- ▶ R-CNN
- ▶ Fast R-CNN
- ▶ Faster R-CNN
- ▶ Mask R-CNN

Blog: blog.paperspace.com/faster-r-cnn-explained-object-detection/

Deep NN Architectures for Object Detection

R-CNN: Three stage process

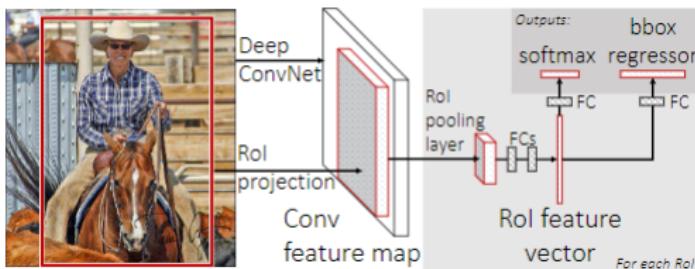


source: Girshick et al., 2014

- ▶ Extract possible objects using some method -> region proposals
- ▶ **Extract features with CNN**
- ▶ Use SVM for classification

Deep NN Architectures for Object Detection

Fast R-CNN

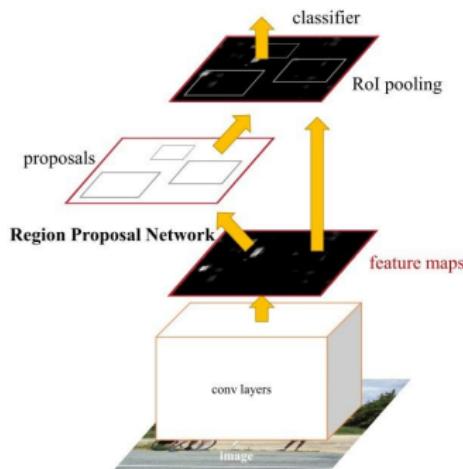


source: *Girshick, 2015*

- ▶ Extract features with CNN
- ▶ **Convolutional layer computations are shared**
- ▶ **Output layers: softmax for classification; FC to predict bounding boxes**

Deep NN Architectures for Object Detection

Faster R-CNN

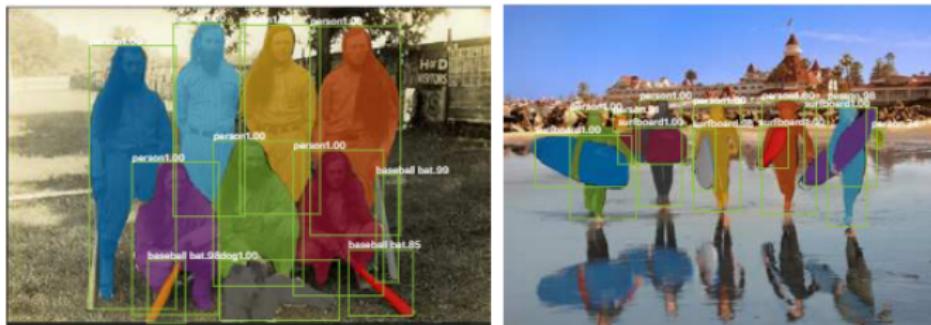


source: *Ren et al., 2015*

- ▶ Generates Region Proposals for Objects
- ▶ Convolutional layer computations are shared
- ▶ Output layers: softmax for classification; FC to predict bounding boxes

Deep NN for Instance Segmentation

Mask R-CNN



source: *He et al., 2017*

Outline

Introduction: Multimodal NLP

Basics: Multimodal Representations

NLP/CL: Linguistic Representations

Computer Vision: Visual Representations I

Appendix: CV Datasets

Visually Grounded Representations

Tasks and Applications in Multimodal NLP

Limitations of Models for NLU

Current Challenges

Computer Vision Datasets (also L&V datasets)

Computer Vision Datasets (also L&V datasets)

- ▶ **ImageNet** <http://www.image-net.org/>
- ▶ **Conceptual Captions** [?, ?]
- ▶ **MS COCO** [?]
Human-elicited captions for “Common Objects in Context”
<https://cocodataset.org/>
- ▶ **Visual Genome** [?]
Human-elicited descriptions of image regions
<https://visualgenome.org/>
- ▶ **SBU Captions** [?]
- ▶ **OpenImages** <storage.googleapis.com/openimages/web/index.html>
- ▶ **PASCAL** <http://host.robots.ox.ac.uk/pascal/VOC/>
- ▶ **Scenes (indoor**) <http://web.mit.edu/torralba/www/indoor.html>
SUN scenes <http://groups.csail.mit.edu/vision/SUN/>
- ▶ **Attributes**
of celebrities (faces) <http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html>
of animals <https://cvml.ist.ac.at/AwA2>
- ▶ **Pets** <http://www.robots.ox.ac.uk/~vgg/data/pets>

Computer Vision Datasets (also L&V datasets)

ImageNet <http://www.image-net.org/>

- ▶ localisation, detection, image classification, segmentation, ...
- ▶ Standard dataset for pre-training models



ImageNet is an image database organized according to the [WordNet](#) hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. Currently we have an average of over five hundred images per node. We hope ImageNet will become a useful resource for researchers, educators, students and all of you who share our passion for pictures.

[Click here](#) to learn more about ImageNet, [Click here](#) to join the ImageNet mailing list.



What do these images have in common? [Find out!](#)

[Research updates on improving ImageNet data](#)

© 2016 Stanford Vision Lab, Stanford University, Princeton University support@image-net.org Copyright infringement

Conceptual Captions

[?, ?]

- ▶ Automatically processed alt-text into clean image captions
- ▶ Used for training visual-language transformer models (Week 7)



Alt-text: A Pakistani worker helps to clear the debris from the Taj Mahal Hotel November 7, 2005 in Balakot, Pakistan.

Conceptual Captions: a worker helps to clear the debris.



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

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

Link to ConceptualCaptions

Computer Vision Datasets (also L&V datasets)

Visual Genome

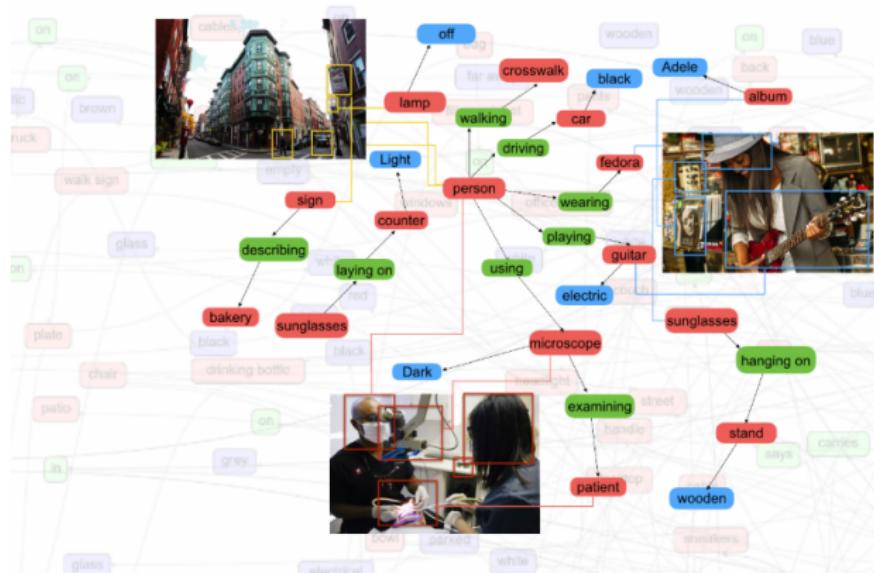
- ▶ region description, Q&A, detection, attributes ...

visualgenome.org



About Download Data Analysis Paper Explore Internal

Explore



Visual Genome is a dataset, a knowledge base, an ongoing connect structured image con language.

Explore our data:
throwing frisbee, helping, angry

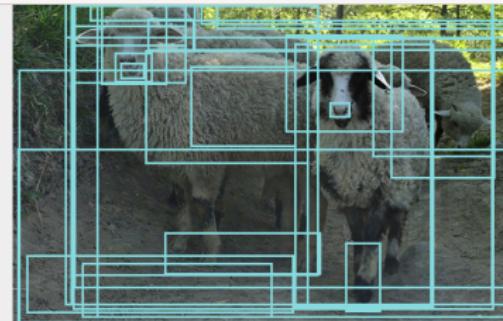
108,077 Images
5.4 Million Region Descriptions
1.7 Million Visual Question Answ
3.8 Million Object Instances
2.8 Million Attributes
2.3 Million Relationships
Everything Mapped to Word2vec

Computer Vision Datasets (also L&V datasets)

Visual Genome

visualgenome.org/VGViz/explore?query=sheep%20eats

a group of white sheep	sheep is white	group OF sheep
a black and white sheep	sheep is black	face OF sheep
the face of a white sheep	face is white	sheep looking down at ground
a sheep looking at the ground	trees is green	trees are in sunlight
green trees in the sunlight	terrain is sandy	sheep are in shade
a group of sheep under the shade	sheep is walking	sheep has ears
sandy terrain	ground is dirt	sheep has wool
a sheep's ear	sheep is looking down	sheep has hooves
		sheep are walking on ground
		trees IN



Question Answers

- When was the photo taken?
- What color are the trees?
- What is in the background?
- What time of day was this taken?
- What is hanging off the sheep head?

- Daytime.
- Green.
- The trees.
- In the daytime.
- His ears.

Regions

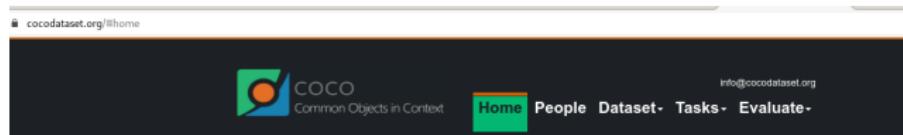
Attributes

Relationships

Computer Vision Datasets (also L&V datasets)

MS COCO

- ▶ localisation, segmentation, detection, image description
- ▶ many 3rd-party extensions (see website)



The screenshot shows the COCO dataset homepage. At the top, there's a navigation bar with links for Home, People, Dataset, Tasks, and Evaluate. Below the navigation bar, there's a section titled "News" which contains a bulleted list of recent announcements. To the left of the news, there's a "What is COCO?" section with a list of features and some sample images. To the right of the news, there's a "Collaborators" section listing names and institutions, and a "Sponsors" section with logos for CVDF, Microsoft, Facebook, and Mighty AI.

News

- We are pleased to announce the COCO 2020 Detection, Keypoint, Panoptic, and DensePose Challenges.
- The new rules and awards for this year challenges encourage innovative methods.
- Results to be announced at the Joint COCO and LVIS Recognition ECCV workshop.

What is COCO?



COCO is a large-scale object detection, segmentation, and captioning dataset. COCO has several features:

- ✓ Object segmentation
- ✓ Recognition in context
- ✓ Superpixel stuff segmentation
- ✓ 330K images (>200K labeled)
- ✓ 1.5 million object instances
- ✓ 80 object categories
- ✓ 91 stuff categories
- ✓ 5 captions per image
- ✓ 250,000 people with keypoints

Collaborators

Tsung-Yi Lin Google Brain
Genevieve Patterson MSR, Trash TV
Matteo R. Ronchi Caltech
Yin Cui Google
Michael Malie TTI-Chicago
Serge Belongie Cornell Tech
Lubomir Bourdev WaveOne, Inc.
Ross Girshick FAIR
James Hays Georgia Tech
Pietro Perona Caltech
Deva Ramanan CMU
Larry Zitnick FAIR
Piotr Dollar FAIR

Sponsors



Research Paper

Download the paper that describes the Microsoft COCO dataset.



Download
paper here

Computer Vision Datasets (also L&V datasets)

MS COCO

| cocodataset.org/#explore



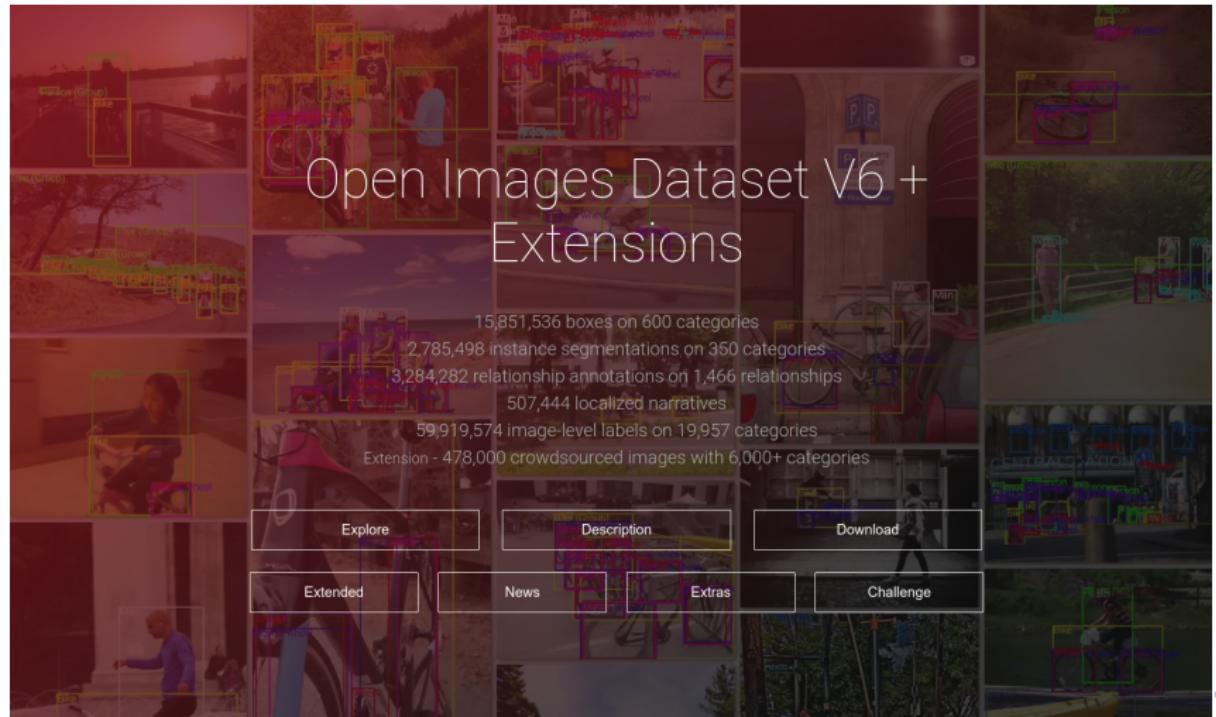
a man and two boys herd 5 adult sheep into a truck.
sheep herded together in pen by adults on farm.
the men are herding the sheep into the truck.
there are men herding sheep into a crate
a vintage photo of some sheep being herded



Computer Vision Datasets (also L&V datasets)

OpenImages

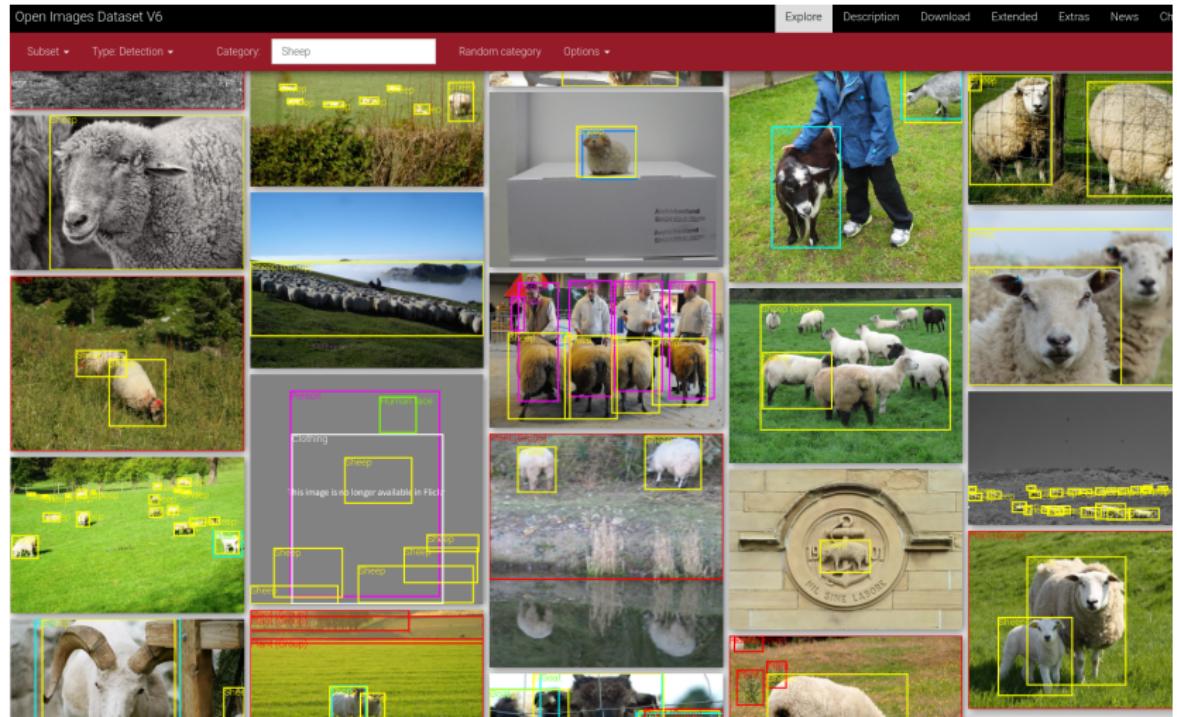
- **detection**, segmentation, relationships, "localised narratives"



Computer Vision Datasets (also L&V datasets)

OpenImages

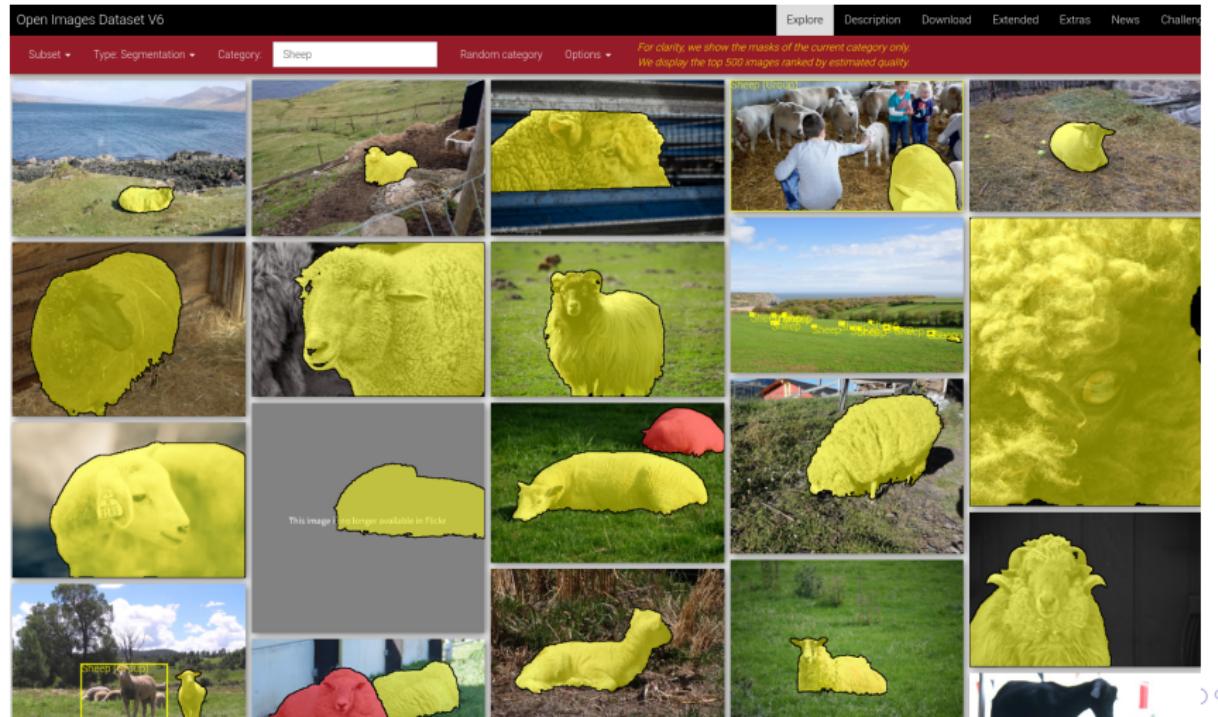
- detection, **segmentation**, relationships, "localised narratives"



Computer Vision Datasets (also L&V datasets)

OpenImages

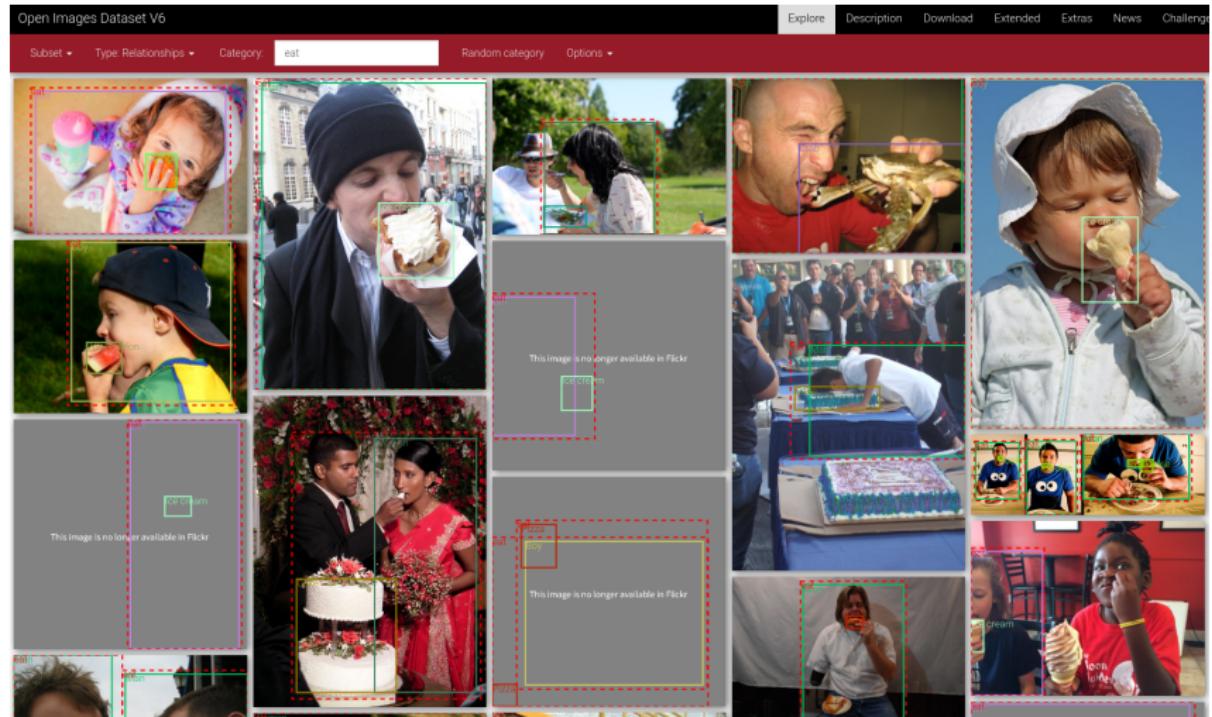
- detection, segmentation, **relationships**, "localised narratives"



Computer Vision Datasets (also L&V datasets)

OpenImages

- detection, segmentation, relationships, "localised narratives"



References I



Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021).
An image is worth 16x16 words: Transformers for image recognition at scale.
In *International Conference on Learning Representations*.