Doc4

Image Knowledge Graph– Scene Graphs

- https://arxiv.org/abs/2206.04863
- https://github.com/pzzhang/VinVL

VinVL has been trained for downstream Vision Language tasks and gives more precise object detection.

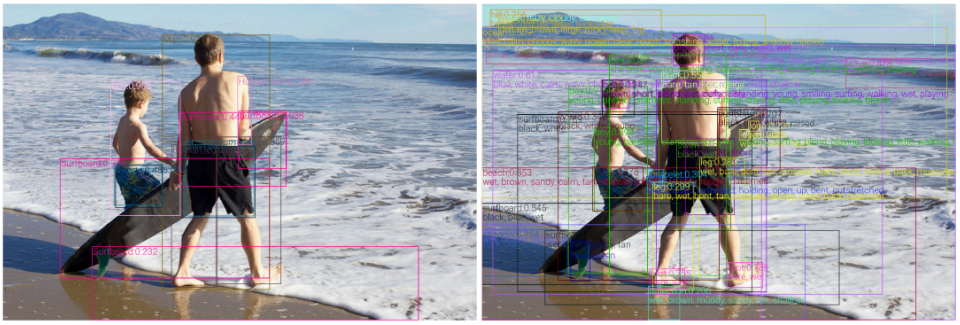
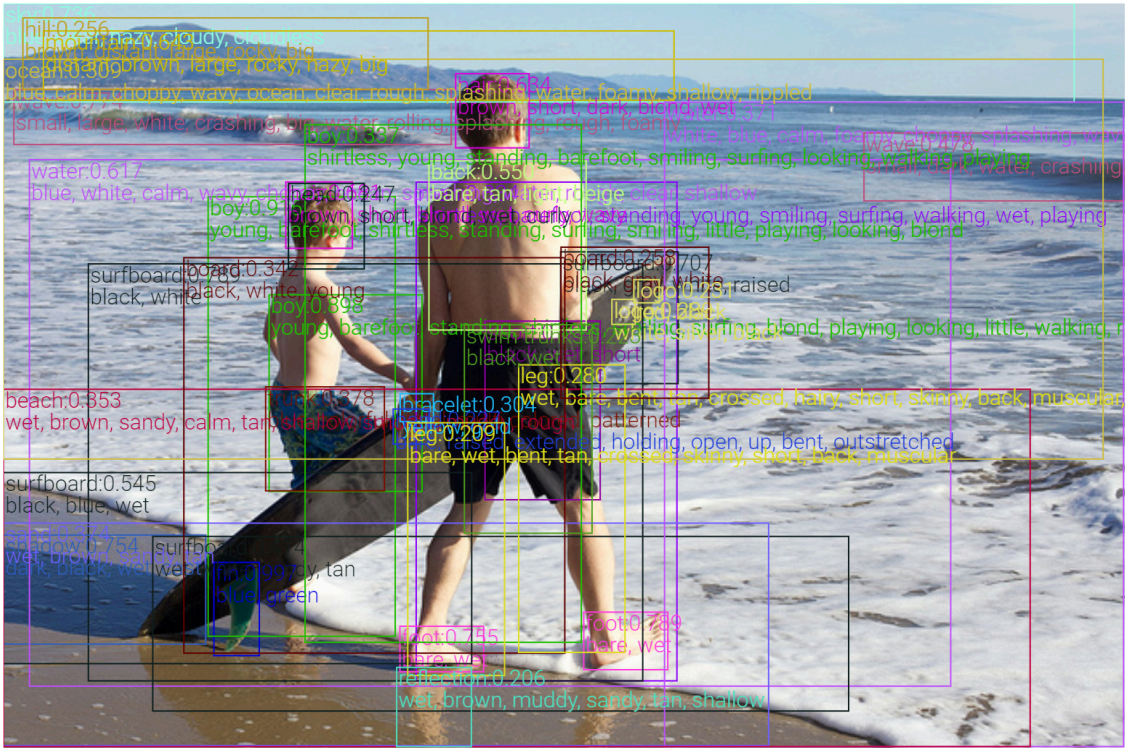


Figure 1: Predictions from an X152-FPN model trained on OpenImages (Left) and our X152-C4 model trained on four public object detection datasets (Right). Our model contains much richer semantics, such as richer visual concepts and attribute information, and the detected bounding boxes cover nearly all semantically meaningful regions. Compared with those from the common object classes in typical OD models (Left), the rich and diverse region features from our model (Right) are crucial for vision-language tasks. For concepts detected by both models, e.g., "boy", attributes from our model offer richer information, e.g., "young barefoot shirtless standing surfing smiling little playing looking blond boy". There are object concepts that are detected by our model but not by the OpenImages model, including fin, wave, foot, shadow, sky, hair, mountain, water, (bare, tan, light, beige) back, (blue, colorful, floral, multi colored, patterned) trunk, sand, beach, ocean, (yellow, gold) bracelet, logo, hill, head, (black, wet) swim trunks, black, wet swim trunks. Compared to the R101-C4 model of [2], our model produces more accurate object-attribute detection results and better visual features for VL applications; see Appendix A for the full pictures and predictions from [2].



- For using VinVL for downstream tasks- https://github.com/microsoft/Oscar
- https://www.microsoft.com/en-us/research/blog/vinvl-advancing-the-state-of-the-art-for-vision-language-models/#:~:text=VinVL%3A%20A%20generic%20object%2Dattribute,concepts%20in%20the%20text%20modality.

AzCopy is a command-line utility by Microsoft designed for copying data to and from Azure Storage. It allows you to transfer files between your local storage and Azure Blob, File, and Table storage quickly and efficiently.

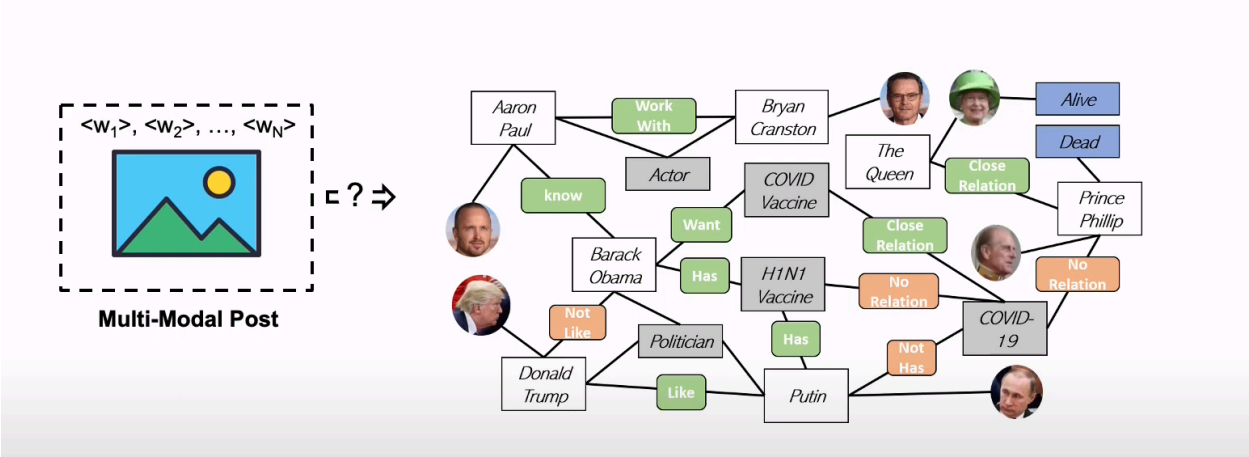
- https://www.mdpi.com/2227-7390/11/19/4174

based on the accuracy of the image recognition of objects, we can explore the similarity between instances of the same species to establish the triples of similarity between instances. Finally, we construct an image knowledge graph through these triples; at the same time, we set a certain threshold to be used

to determine the same kinds of species similarity: the similarity to the threshold of the entity to judge for the same species

Multimodal Knowledge Graphs

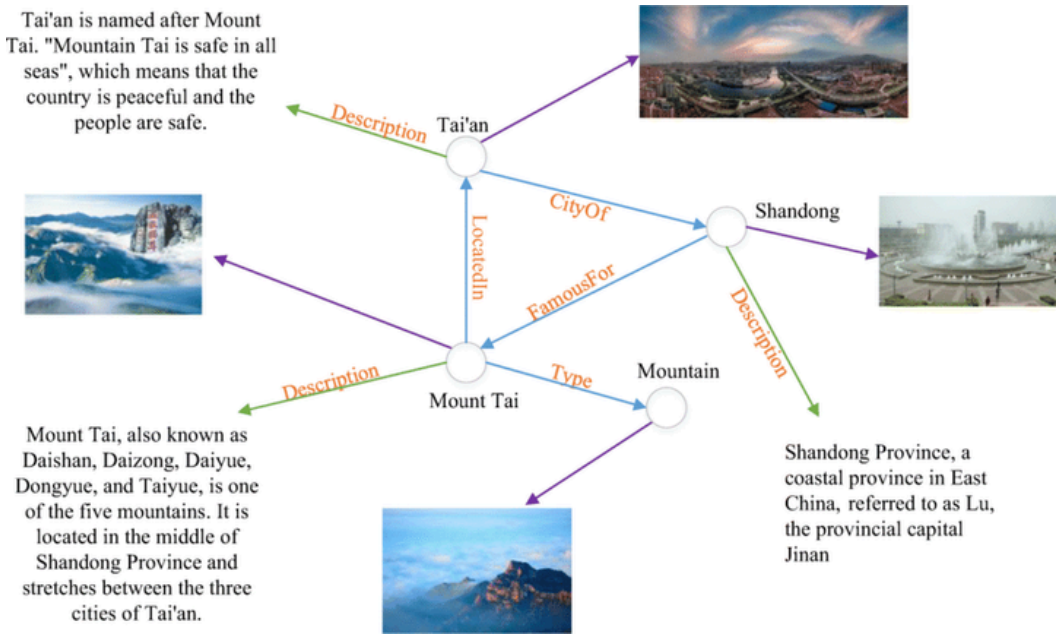
- Now that we have made Text corpus knowledge graph, we would move on to use both image and text to create a knowledge graph



https://www.frontiersin.org/journals/neurorobotics/articles/10.3389/fnbot.2019.00093/full

- “Semantic Graph” for textual paragraphs (Sorokin and Gurevych, 2017), and “Scene Graph” for visual images
- the generation of knowledge graphs (KGs) is decomposed into two phases: (1) detecting the entities (or objects) as nodes, and (2) extracting relations between entities as edges.
- process of predicting unknown relations based on the current (incomplete) KG shares some commonalities with knowledge graph completion (KGC)

- to understand and reason about the context of an image we need not only information about objects within the scene, but also about relations between these objects. Therefore, extracting the relations between objects (e.g., in/on/under, support, etc.) yields a better scene understanding compared to just recognizing objects and their individual properties (Elliott and de Vries, 2015).
- Scene graphs are a way of representing the context of an image in a structured way to improve the performance of tasks such as visual question answering or image retrieval. Existing scene graph generators usually extend an object detection framework that first detects bounding boxes for objects, then extracts visual features and classifies objects inside bounding boxes, and finally predicts relations between objects in a parallel manner



Resources:

- https://www.sciencedirect.com/science/article/abs/pii/S2214579623000138

The conventional knowledge graph (CKG) refers to the text knowledge graph as a form of structured human knowledge, which is a semantic network constructed by triples with text as the object.

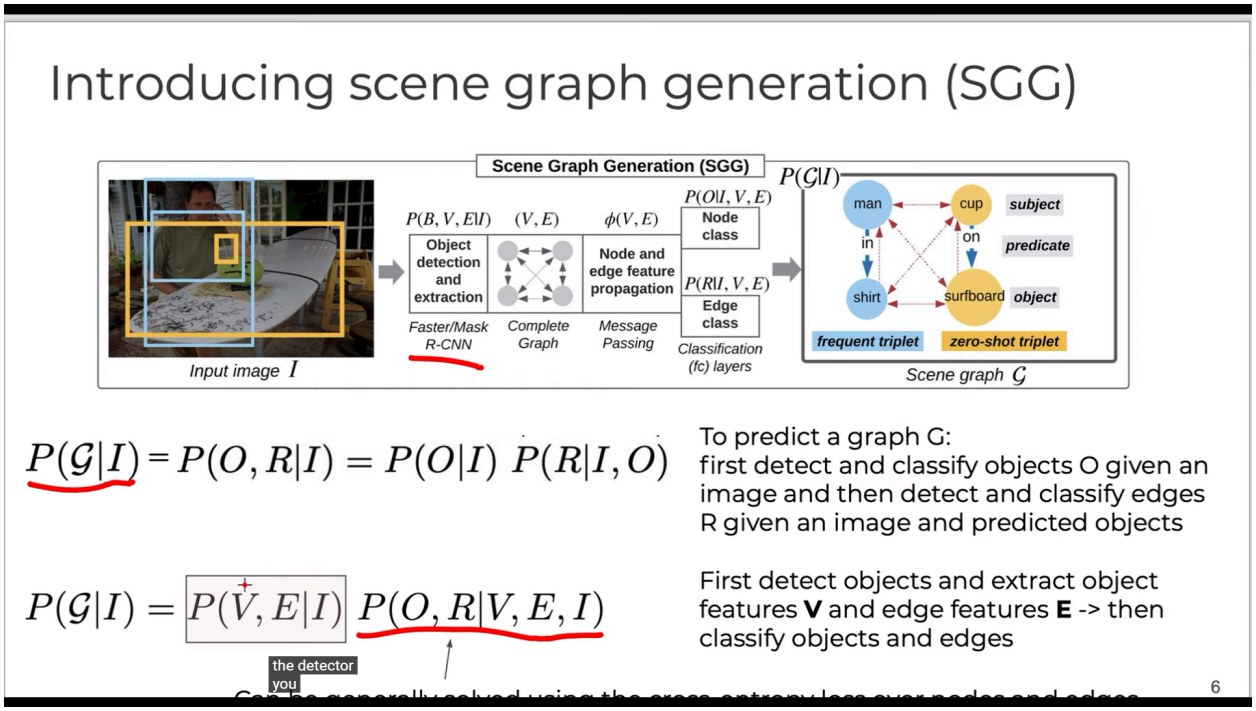
- https://github.com/heathersherry/Knowledge-Graph-Tutorials-and-Papers/blob/master/topics/Knowledge%20Graph%20Enhanced%20Machine%20Learning.md

(They have used CNNs for object detection and then created their Knowledge Graph)

Plan:

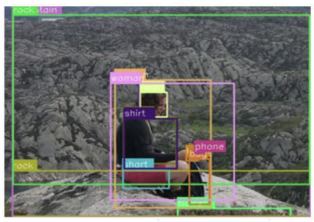
- ek text ka KG
- Image KG
- Multimodal KG
- uske baad embeddings

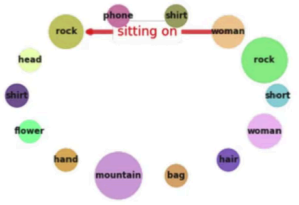
https://github.com/bknyaz/sgg/blob/master/Scene_Graph_Predictions_GQA.ipynb



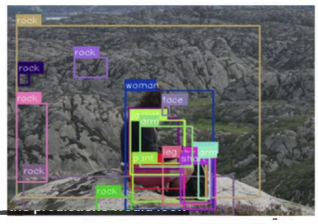
Example of predictions

Ground truth (annotations)





Predictions



like
for some image

7

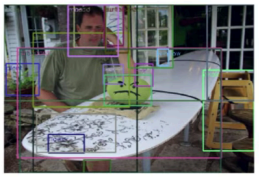
Evaluation

Standard recall metric (dominated by frequent

$$\mathrm{R@K} = |\mathrm{Top}_K \cap \mathrm{GT}|/|\mathrm{GT}|$$

Top-K triplets retrieved by the mode

Ground truth triplets



GT triplets:
1. cup on surfboard
2. surfboard with flower
3. table made of surfboard
4. head of man
5. **man in shirt**
6. **chair behind surfboard**
7. plant near door

Graph Constraint

keep only top-1 prediction between 'man' and 'shirt' (same for other pairs of objects)

Top model predictions	Score
1. man wearing shirt	0.10
2. cup on table	0.08
3. man has head	0.07
4. **chair behind surfboard**	0.04
5. sufrboard near man	0.03
6. man near surfboard	0.02
...	
100. plant near house	0.01

Recall@100 = 1/7 ≈ 14%

$$\mathrm{R@K}_{\mathrm{ZS}} = |\mathrm{Top}_K \cap \mathrm{GT}|/|\mathrm{GT}|$$

Top-K triplets retrieved by th mode

perspective
but it's it's not so popular so

Ground truth **zero-shot** triplets

8

https://github.com/KaihuaTang/Scene-Graph-Benchmark.pytorch