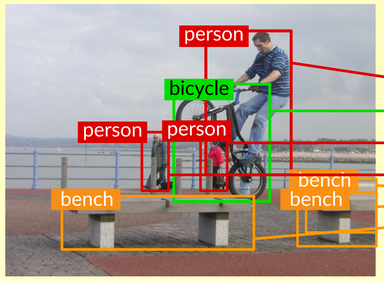



# Object Counts!

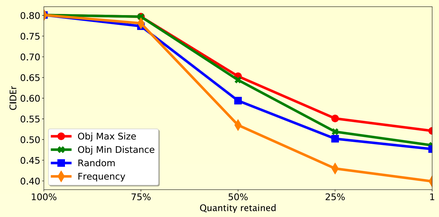
# Bringing Explicit Detections Back into Image Captioning

## What can we extract from explicit detections for image captioning?

| | Person | Bicycle | Train | Banana | Bench | Spoon | CIDEr score |
|---|---|---|---|---|---|---|---|
| Frequency | 3 | 1 | 0 | 0 | 3 | 0 | 0.807 |
| Normalized Frequency | 0.43 | 0.14 | 0 | 0 | 0.43 | 0 | 0.762 |
| Presence | 1 | 1 | 0 | 0 | 1 | 0 | 0.751 |
| Most central instance | 0.09 | 0.15 | 0 | 0 | 0.81 | 0 | 0.759 |
| Size of biggest instance | 0.18 | 0.27 | 0 | 0 | 0.20 | 0 | 0.793 |
| Size \|\| centrality | 0.18 0.27 0 0 0.20 0 | | | 0.09 0.15 0 0 0.81 0 | | | 0.799 |
| Frequency \|\| centrality | 3 1 0 0 3 0 | | | 0.09 0.15 0 0 0.81 0 | | | 0.830 |
| Frequency \|\| size | 3 1 0 0 3 0 | | | 0.18 0.27 0 0 0.20 0 | | | 0.836 |
| Freq. \|\| centrality \|\| size | 3 1 0 0 3 0 | | | 0.09 0.15 ... 0 | | 0.18 0.27 ... 0 | 0.849 |
| Full spatial info | *person* x y w h a x y w h a x y w h a ... | | | | | | 0.870 |
| | *bicycle* x y w h a 0 0 0 0 0 0 0 0 0 ... | | | | | | |
| | *train* 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 ... | | | | | | |
| | *bench* x y w h a x y w h a x y w h a ... | | | | | | |
| | ... | | | | | | |

(x, y): coordinates of object center, (w, h): width/height of object bounding box, (a): area of object segment

What happens when information is removed? Mask N% of object frequencies with different heuristics:

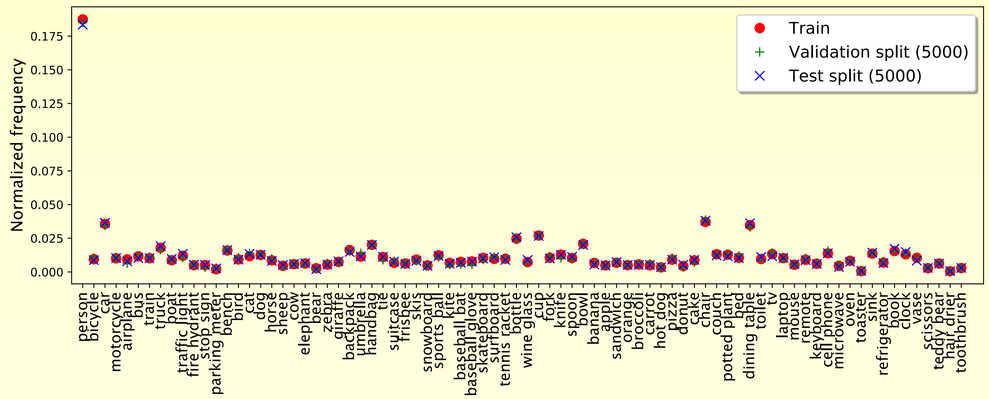



Summary:

- Importance: Frequency > Size > Centrality
- All three features are complementary
- A more frequent category is not necessarily more important
- Larger objects are more important than smaller objects
- Objects near the image center > away from the center (but size is more important)

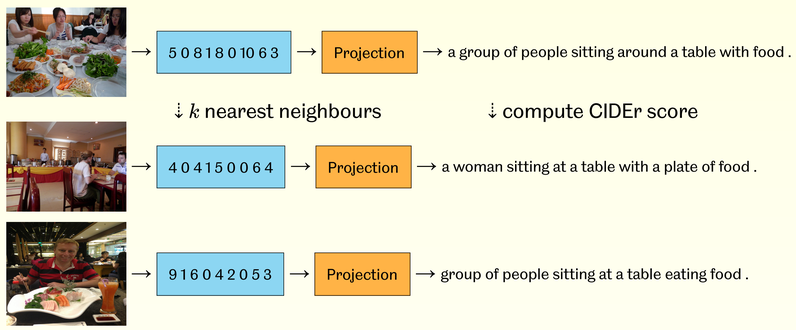
## Why does bag of objects work so well?

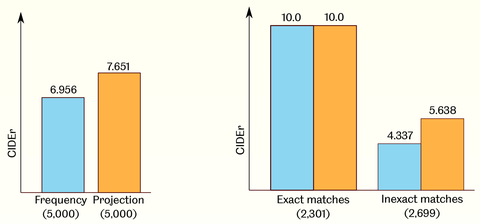
Object distribution in train vs. test is almost identical! A retrieval machine maybe?




## Is neural image captioning merely a multimodal retrieval machine?

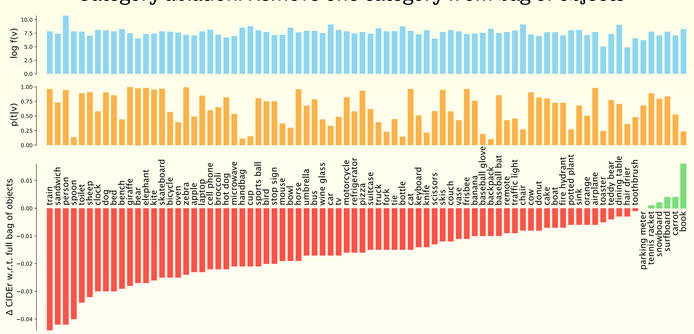
If yes, then the model should generate similar captions for the nearest neighbour(s) of an image!




## Which categories are more important?

Category ablation: Remove one category from bag of objects




Rank correlation between ΔCIDEr versus:

- frequency of object being annotated
  - $f(v_c)$: Kendal's $\tau = 0.093$, Spearman's $\rho = 0.137$
- probability of object being mentioned given that it is depicted
  - $p(t_c|v_c)$: Kendal's $\tau = 0.153$, Spearman's $\rho = 0.227$

## Example output

| | |
|---|---|
| Objects: | *person*, *clock* |
| Frequency: | a large clock tower with a large clock on it . |
| Centrality: | a clock tower with a large clock on it 's face . |
| Size: | a man standing in front of a clock tower . |
| All 3 feats: | a clock tower with people standing in the middle of the water . |
| *(x,y):* | a large clock tower with a clock on the front . |
| *(w,h):* | a clock on a pole in front of a building |
| *(a):* | a large clock tower with people walking around it |
| *(x,y,w,h,a):* | a group of people standing around a clock tower . |
| CNN: | a large building with a clock tower in the middle of it. |
| *person* removed: | a clock tower with a weather vane on top of it . |

Josiah Wang, Pranava Madhyastha, Lucia Specia

The University Of Sheffield.