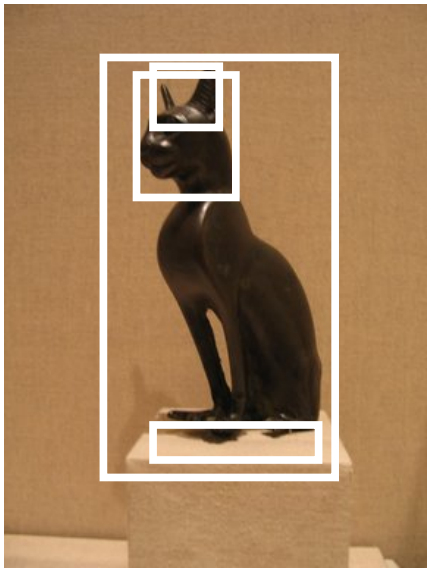
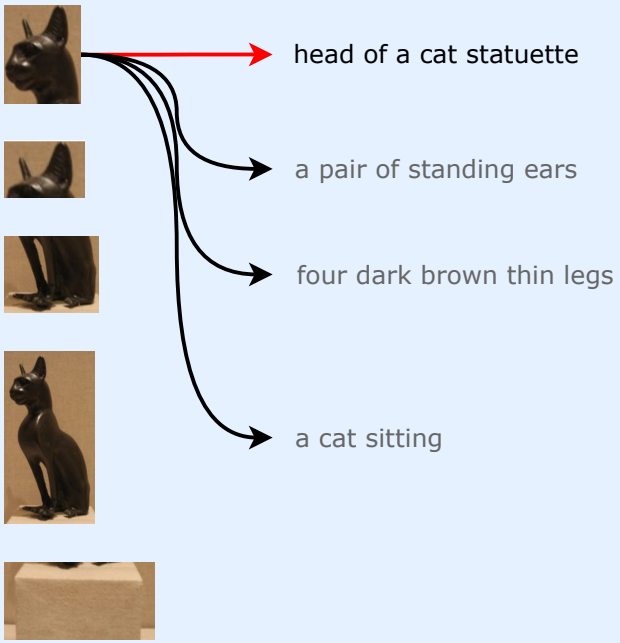


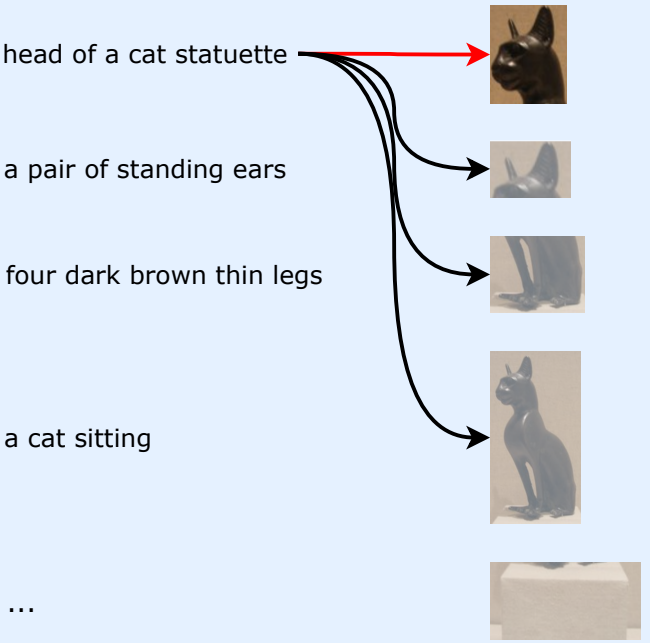
head of a cat statuette,
a pair of standing ears,
four dark brown thin legs, a cat sitting

1

attend to noun phrases with respect to each image feature



attend to image features with respect to each noun phrase



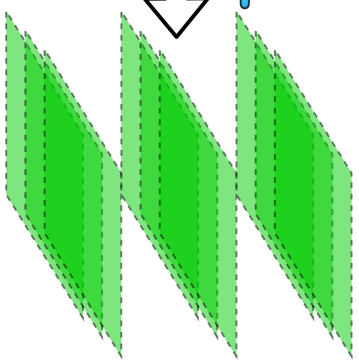
2

Calculate mutual attention between these noun phrases and the image fragments obtained from *faster-rcnn*

3

align targets

4



ResNet, Fully-connected layers, GRUs

5

representations of each phrase and its corresponding image fragment now in a common space

Calculate similarity between image fragement and phrases

6

Ranked recall

Stacked Cross Attention