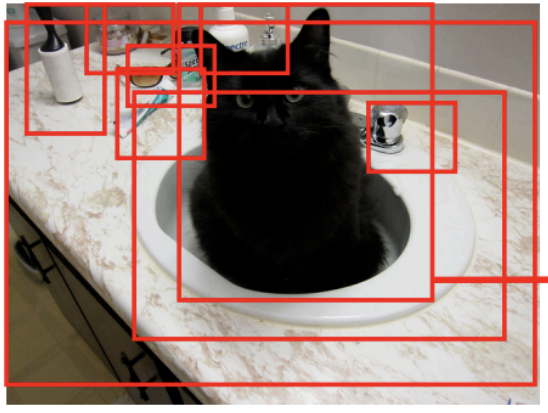

Bottom-Up Attention

Sentence T: A cat is sitting in the bathroom sink.

v_1

Attended sentence vector a_1^t

A cat sitting bathroom sink

Stage 1: Attend to words

v_2

A cat is sitting bathroom sink.

A cat is sitting in the bathroom sink.

A cat is sitting in the bathroom sink.

Stage 2:
Attend to the important image regions given a_i^t

Similarity $R(v_i, a_i^t)$

Pooling

Similarity $S(I, T)$