Compositional Image Captioning

A white cat sitting on a laptop computer
A white dog running along a beach
A big brown dog sitting on a couch

White things
Dogs

Compositional Image Captioning

Syntactic Granularity & Planning Approaches

• Syntactic tags: From chunking to part-of-speech to dependency labels to CCG tags (plus a synthetic IDLE tag for each word)

• Planning approaches: Standard Decoder
Sequential Decoder
Interleave Decoder
Multi-Task Decoder

Syntactic Planning

• Idea: Introduce a sentence planning step where the model plans the syntactic structure of the caption

• Hypothesis: a model assigns higher probability to the unseen phrase “white dog” if it explicitly expects to generate JJ NN

• Goal: caption images that have unseen compositions of known concepts

Nicklaus et al. (2019) showed that RNN-based captioning models do not compositionally generalise due to the text decoder

Syntax Awareness

Average Recall@5 of unseen concepts by BUTD (Anderson et al., 2018)

Directly mapping an image onto words is sub-optimal
Breaking bi-gram sequences with IDLE is useful
Syntactic planning improves compositional image captioning

Results & Examples

Average Recall@5 of unseen concepts

Interleaving POS tags with words is model-agnostic
Transformers do not compositionally generalise

BUTD
BUTR
BUTRweight
M2

Standard
9.5
15.0
14.9
10.6

Interleave POS
11.8
12.0
16.4
13.2

BUTD there is a woman that is on the floor
BUTD+POS a woman riding a bike on a wooden floor
BUTD a woman with a child sitting on a bench
BUTD+POS a girl that is standing on a skateboard