Context-aware Captions from Context-agnostic Supervision

Midterm Report

Introduction:

We'll be implementing this paper which doesn't have any public code available to the best of our knowledge. The paper attempts to solve the problem of generating context aware captions (captions that describe differences between images or visual concepts) by introducing a novel inference technique using only generic context-agnostic training data (captions that describe a concept or an image in isolation) .

Approach:

- The data provided while training the model were only the 'image', 'class of the image' and 'captions'; No paired 'images' and 'context-aware captions' were provided.
- The architecture of the model builds upon the architecture from the popular paper Show, Attend and Tell which is attention based model for captioning images. The main changes to the original architecture was that the decoder was conditioned on the class of the image (during Justification).
- The inference technique was the main novelty, which conditions the captions of the target image on the distractor class, and then uses Emitter-Suppressor beam search on this conditioned probability distribution.
- The model trade-offs between linguistic adequacy of the sentence, and discriminativeness.

Present Progress:

- Major chuck of our work till now was to understand the original paper and a cascade of other papers which it builds upon.
- Gotten hold of the unformatted dataset(image, class and labels), formatted and cleaned it, hence to increase the efficiency during training.
- Written code for Show, Attend and Tell, using only soft Attention which is yet to be trained.

Future work:

- Change the original Show, Attend and Tell architecture to incorporate the Justification and Discrimination tasks.
- Implement Emitter-Suppressor beam search inference technique.
- Tune the hyper-parameters and do the training on CUB dataset.

160050092, 160050098, 1600500104