Highlights
- What: Generate context-aware captions without context-aware supervision
- Why: Humans are able to account for context to "say the right thing" Collecting captions across all relevant contexts is expensive Can we generate captions without context-aware supervision?
- How: Modify inference (beam search) to account for context explicitly

Models
- Why is it: a black-throated blue warbler and not a black and white warbler?

Task 2: Discriminative Justification
- Why: Generating captions across all relevant contexts is expensive
- What: Can we generate captions without context?

Image Captioning
- Lots of recent progress [Vinyals et al., Karpathy and Li, Fang et al.]

Task 1: Discriminative Image Captioning
- An airplane flying in the sky

Task 2: Discriminative Justification
- Why do you think this image is a black-throated blue warbler, and not a mourning warbler?

Justification Results
- Validation:

Datasets and Evaluation
- CUB-Justify for Justification Evaluation
- Built on top of the CUB dataset [Wah et al.] Target distractor chosen among 25 hyper-classes Overall 3161 target, distractor image triplets
- Metrics: CIDEr-D

COCO Human Studies
- Target distractor image sets: easy confusion, hard confusion
- Easy confusion: random 1500 1 nearest neighbor in FC7 Hard confusion: 1 nearest neighbors with most similar generated captions
- Human studies (Guess Which?):

Conclusion
- Novel inference procedure to generate context-aware discriminative language from context agnostic supervision
- Two novel tasks: discriminative image captioning and justification
- New dataset CUB-Justify with ground truth justifications for evaluation with 3161 image, target distractor classes
- Results on discriminative image captioning as well as justification show improvement over generative and speaker-listener approaches.