Motivation & Target

- Motivation: Improving image captioning via leveraging unlabeled images from massive unpaired web sources

- Challenge:
  1. Cross-modal data - Unlike previous studies working on single-modal data (e.g., image, text, or graph), image caption generation is a cross-modal task on the intersection of image and text;
  2. Complex task - Image caption generation is a complex task that has to generate new content rather than simple classification or prediction task studied in previous work.

- Target: Image Captioning with Very Few Labels

Model

- Overall Framework

  -- 1. Pretrained Word Embedding for NLP Information

  -- 2. Big Dropout Rate as Semantics Augmentations

Experiment

- Performance Comparison with SOTAs

- Ablation Study

- Case Show

Table 1: Performances of different model variants with/without TAN and with pretrained word embedding, non-graph augmentations, and cross-modal training. (a) Overall setting (a) and (b) ablation setting.

**Note:** ROUGE-L and SPICE of Visual are not shown due to missing values in the original work.

**Summary:**

- **GT**: A herd of cattle laying on top of a sandy beach.
- **VSU**: A cattle a a a a.
- **SGAE**: A cow is a a a a.
- **M^2T**: A herd of sheep standing a a a a a.
- **SGCL**: A group of cattle grazing in the beach in front of the water.
Look Twice as Much as You Say: Scene Graph Contrastive Learning for Self-Supervised Image Caption Generation

Chunhui Zhang\textsuperscript{1}, Chao Huang\textsuperscript{2}, Youhuan Li\textsuperscript{3}, Xiangliang Zhang\textsuperscript{4}, Yanfang Ye\textsuperscript{4}, and Chuxu Zhang\textsuperscript{1}

Brandeis University\textsuperscript{1}, University of Hong Kong\textsuperscript{2}, Hunan University\textsuperscript{3}, University of Notre Dame\textsuperscript{4}