**Goal**

Captioning
- Two people are in a wheelchair and one is holding a racket

Visual Question Answering
- How many people on wheelchairs? Two
- How many wheelchairs? One

**Task:** Given image, dialog history, follow-up question – predict free-form answer

**VisDial Dataset & Evaluation**

- **Evaluation:** Given Image, dialog history, question, 100 candidate answers – evaluate model on retrieval of ground-truth human response
- **Metrics:** mean reciprocal rank, recall@k, mean rank

>140k dialogs on COCO images; >1.4M dialog question-answers

**Encoder-Decoder Models**

Late Fusion (LF) Encoder: Naive embedding of image, history, question
Hierarchical Recurrent Encoder (HRE): Dialog-level recurrent block on top of QA-level recurrent block

Memory Network (MN) Encoder (Similar to Weston et al., 2014): Builds context vector from previous QA facts in memory

Generative (G) Decoder: RNN initialized with input encoding, predicts response word-by-word; Trained to maximize LL of ground truth response
Discriminative (D) Decoder: Dot product between input encoding and RNN encoding of 100 candidate answers + 100-way softmax

**Visual Chatbot**

Real-time visual chatbot hosted on CloudCV

demo.visualdialog.org

**Discriminative models work better than generative models**

MN works best in both generative and discriminative settings

Human performance, topic transition studies, dialog perplexities, etc. in paper