Drill-down: Interactive Retrieval of Complex Scenes using Natural Language Queries
Fuwen Tan¹, Paola Cascante-Bonilla¹, Xiaoxiao Guo², Hui Wu², Song Feng², Vicente Ordonez³
¹University of Virginia, ²IBM Research AI

Retrieving an Image using a Single Query is HARD
Especially when the scene is complex

Goal: Retrieving an Image by Multiple Round Queries

Contributions
• Drill-down, an interactive image search approach with multiple round queries which leverages region captions as form of weak supervision during training;
• A compact representation, outperforming competing baseline methods by a significant margin;
• Experiments on a large-scale natural image dataset: Visual Genome, demonstrating superior performance of our model on both simulated and real user queries.

Evaluations on Simulated/Real Scenarios

Multiple Round Retrieval Examples

Evaluations on Region Captions

Human Evaluations

Interpretable & Compact Representation

Target Image

New Query

Target Image

Region Captions as Weak Supervision

Observation

ID query representations can NOT distinguish entities sharing the same feature space

Proposed: 2D query representations of fixed sizes

Model

State Vectors $X^t$

State Vectors $X'$

Cross Modal Similarity $s(X, X')$

Sentence Rep. $q$

State vec. 1

State vec. 2

Proposed: 2D query representations of fixed sizes

Region Features

Faster R-CNN

Linear Proj.

GRU

R@5

R@10

R-HRED

Human Evaluations

Queries

State Vectors

Attended Image Regions

Methods

Evaluation on Region Captions

Evaluations on Simulated/Real Scenarios

Datasets

Building Drill-down Dialogue Systems using Automatic Interactive Retrieval in Neural Network Models, Arash V. Farahat, Alejandro Carrazedo, Yoav Shalit, Iulian V. Serban, Yang Song, Stamatios Lassig, Victoria Rusu, and Matthew Johnson, ICLR 2018

Drill-down Interactive Image Retrieval, Shihang Sun, Hui Wu, Ti Ceng, Steven Heitz, Gerald Tesauro, and Rogerio Feris. NeurIPS 2018