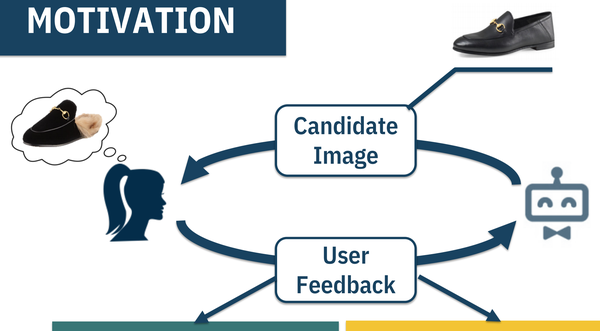
Dialog-based Interactive Image Retrieval

Xiaoxiao Guo* xiaoxiao.guo@ibm.com
Hui Wu* wuhu@us.ibm.com
Yu Cheng chengyu@us.ibm.com
Gerald Tesauro gtesauro@us.ibm.com
Steve J Rennie srennie@gmail.com
Rogerio S Feris rsferis@us.ibm.com

MOTIVATION



Relevance Feedback: Positive

Attribute Feedback: more open

Dialog Feedback: Unlike the provided image, the one I want has an open back design with suede texture.

we introduce a new approach to interactive image search that enables users to provide feedback via **natural language**, allowing for more natural and effective interaction.

CONTRIBUTIONS

New vision/NLP task for interactive image search, where the dialog agent learns to interact with a human user, and the user gives feedback in natural language.

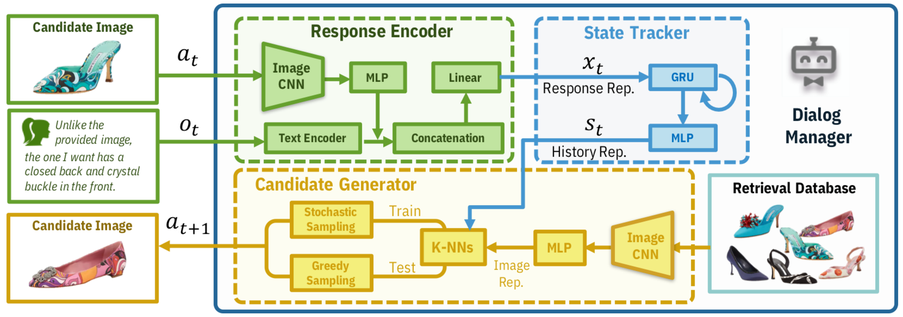
A deep dialog manager architecture: the network is trained end-to-end based on an efficient policy optimization strategy.

Novel vision task (relative image captioning), where the generated captions describe the salient visual differences between two images, and **a new dataset**, which supports further research on this task.

[1] A Kovashka, D Parikh, and K Grauman. Whittlesearch: Image search with relative attribute feedback. In CVPR, 2012.
[2] S J Rennie, E Marcheret, Y Mroueh, J Ross, and V Goel. Self-critical sequence training for image captioning. In CVPR, 2017.
[3] X Guo*, H Wu*, Y Cheng, G Tesaur, S J Rennie, and R S Feris. Dialog-based Interactive Image Retrieval. arXiv preprint, 2018

APPROACH

The Learning Framework

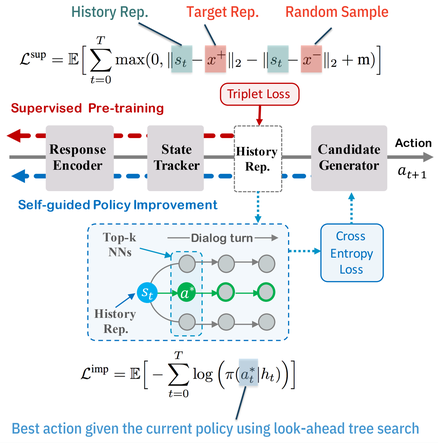


Response Encoder embeds the information from the current dialog turn to a visual-semantic representation;

State Tracker: receives the response representation and combines it with the history information;

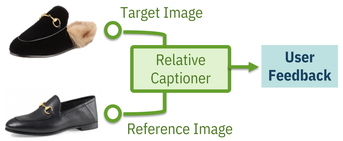
Candidate Generator: samples an image to return to the user based on distances between the history representation to each database image.

Policy Learning



Best action given the current policy using look-ahead tree search

User Simulator



Dataset (~10k pairs)

AMT User : *"unlike what you showed, the one I want is bolder with cow pattern and more ridged sole"*

User Simulator: *"unlike what you showed, the one I want has a print with a strap"*

- User simulator enables efficient exploration of the retrieval dialogs
- Relative Captioner model: show-attend-tell with image feature fusion

* These two authors contributed equally.

RESULTS AND EVALUATION

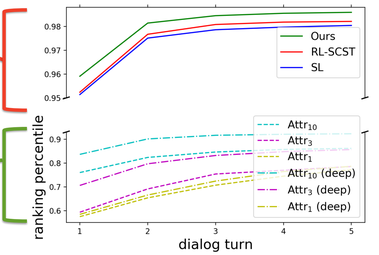
Policy Learning Results

SL: supervised learning where the agent is trained only using triplet loss;

RL-SCST: policy learning using Self-Critical Sequence Training after pre-training using SL. [2]

Effectiveness of Natural Language Feedback

Attr$_n$ and **Attr$_n$(deep)**: dialog managers trained with relative attribute feedback [1] . A rule based feedback generator concatenates respective attribute words with "more" or "less". n denotes the number of attributes used in each feedback, such as *"more shiny and less sporty"*.



- RL based methods resulted in improved retrieval ranking percentile than triplet loss.
- Dialog-based feedback is more effective than attribute feedback using a limited vocabulary.

USER DIALOGS

- Dialog-based feedback is more natural compared to selecting attributes from a pre-defined list.
- Coarse to fine feedback as dialog progresses

Project Website: www.spacewu.com/posts/fashion-retrieval/