Image retrieval with mixed initiative and multimodal feedback

Nils Murrugarra-Llerena
Adriana Kovashka

Department of Computer Science
University of Pittsburgh

Motivation

I really like this shoe, how can I find it?

Draw a sketch and show to a friend.

(Yu et al, CVPR 2016)

OR

Describe it to a salesman, compare with another shoe.

(Kovashka et al, CVPR 2012)

OR

I am a salesman, can you answer some questions?
Is it more / less formal / pointy than this?

(Kovashka and Grauman, ICCV 2013)

Existing methods use these approaches in isolation.

Motivation

Is there an intelligent way to **combine** user and system **interactions** with **multimodal feedback**?

I really like this shoe, how can I find it?

User-initiated methods

exploration

Draw a sketch and show to a friend.

(Yu et al, CVPR 2016)

AND

Describe it to a salesman, compare with another shoe.

(Kovashka et al, CVPR 2012)

AND

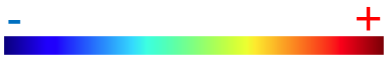
System-initiated methods

exploitation

I am a salesman, can you answer some questions?

Is it more / less formal / pointy than this?

(Kovashka and Grauman, ICCV 2013)



An agent adaptively chooses an interaction.

Key idea

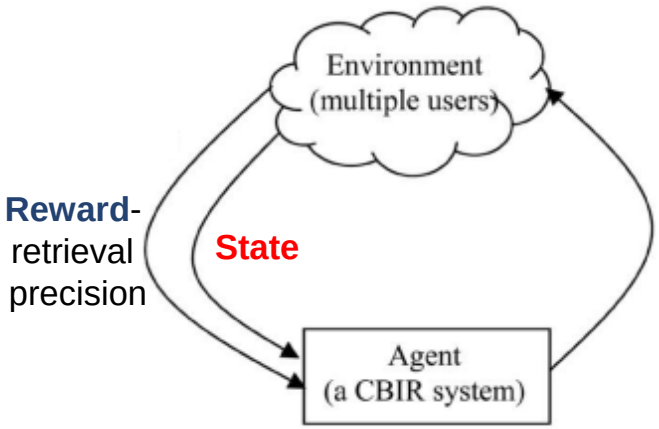
- We learn to **adaptively combine** different forms of **user feedback** (textual or visual) for interactive image retrieval.
- To select the best feedback, we train a reinforcement learning (**RL**) agent which finds the target image fast.

Image Search

Actions

Related work

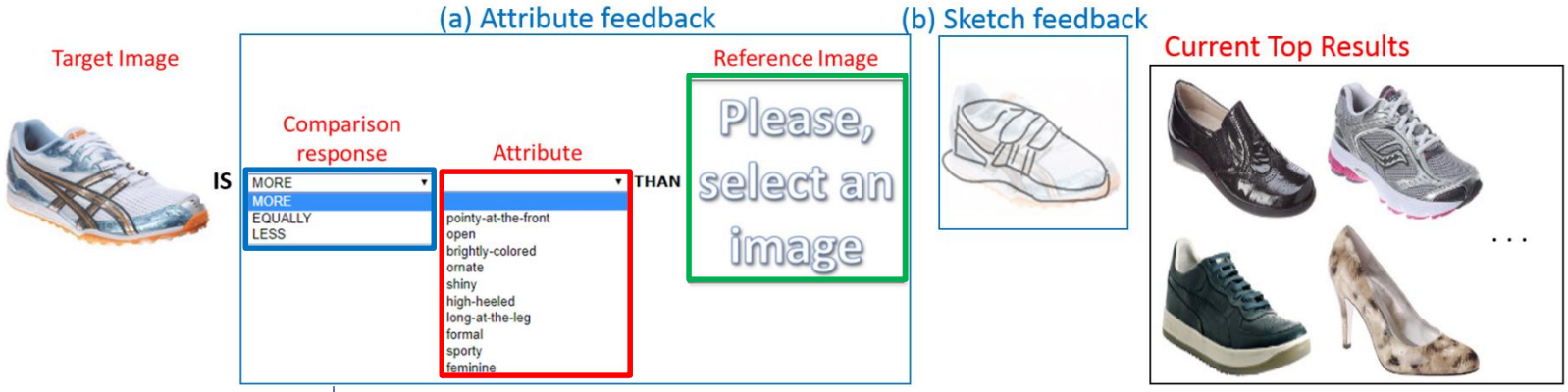
- Instead of improving sketch-based retrieval, we focus on **when to request** a sketch.
- Rather than retrieving a single result, users perform **iterative image retrieval**.
- The most similar work to ours is **Yin et al (TPAMI 2005)**, which also uses RL.



Actions – Relevance Feedback (RF) methods
- QVM
- FREE
- BI

- However, it does not allow users to **describe comparatively** how the results should change. Instead, each image property is defined as desirable or not.
- Unlike **Yin et al (TPAMI 2005)**, we consider both **textual** and **visual feedback**.

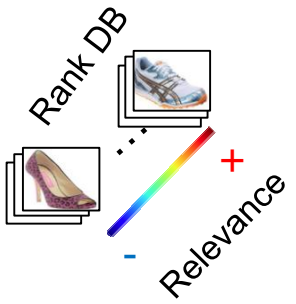
Approach - interactions



A1 **(WS)**: Free-form attribute feedback

(att?, ?)

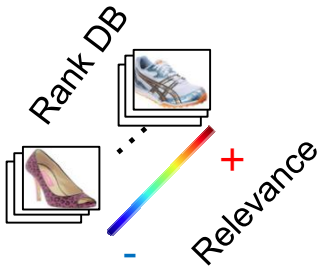
Choose (attribute, reference image) + response



A2 **(PRR)**: Suggested question

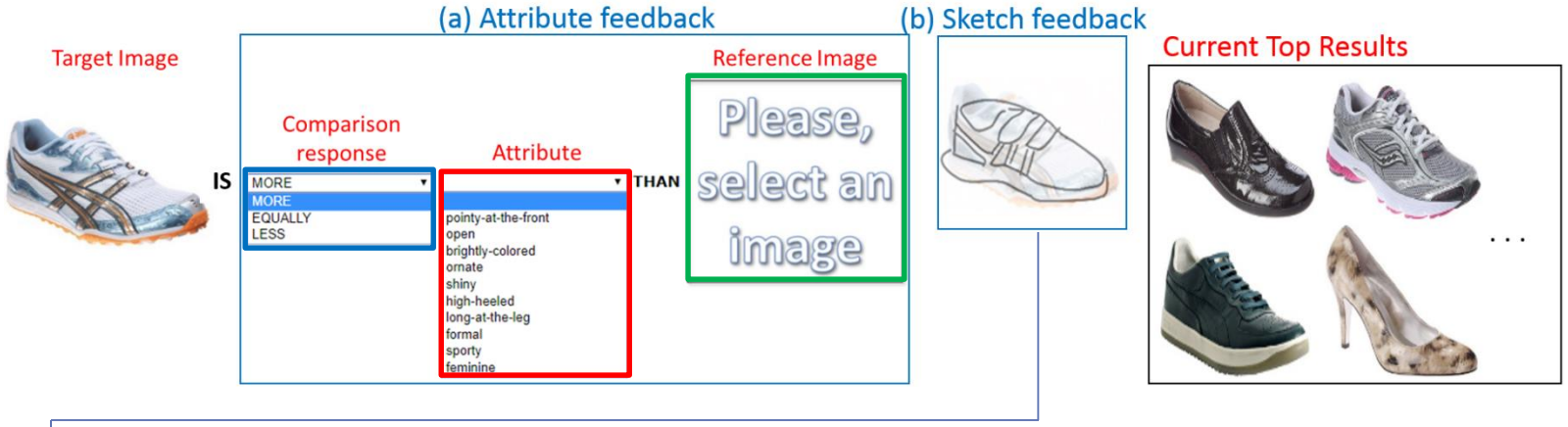
(pointy,)

Provided (attribute, reference image) + Choose response

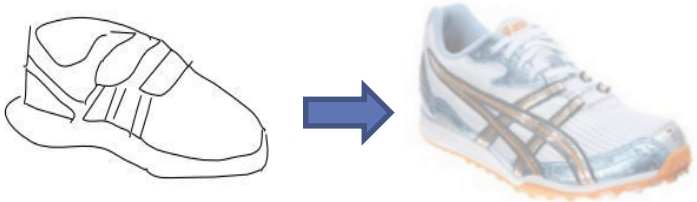


Approach - interactions

A3 **(SK_PRR)**: Sketch

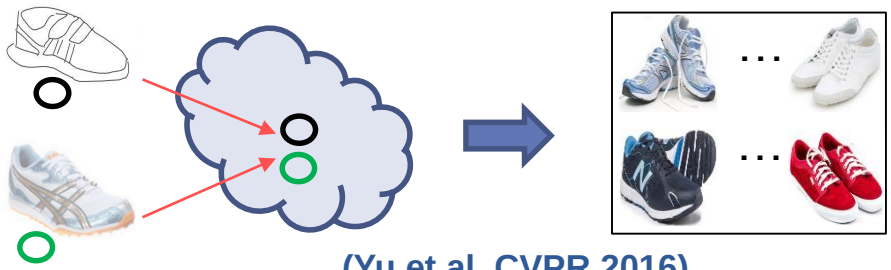


GAN

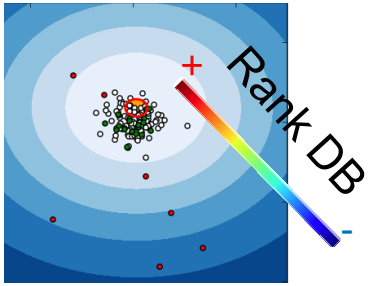


(Isola et al, CVPR 2017)

Image retrieval



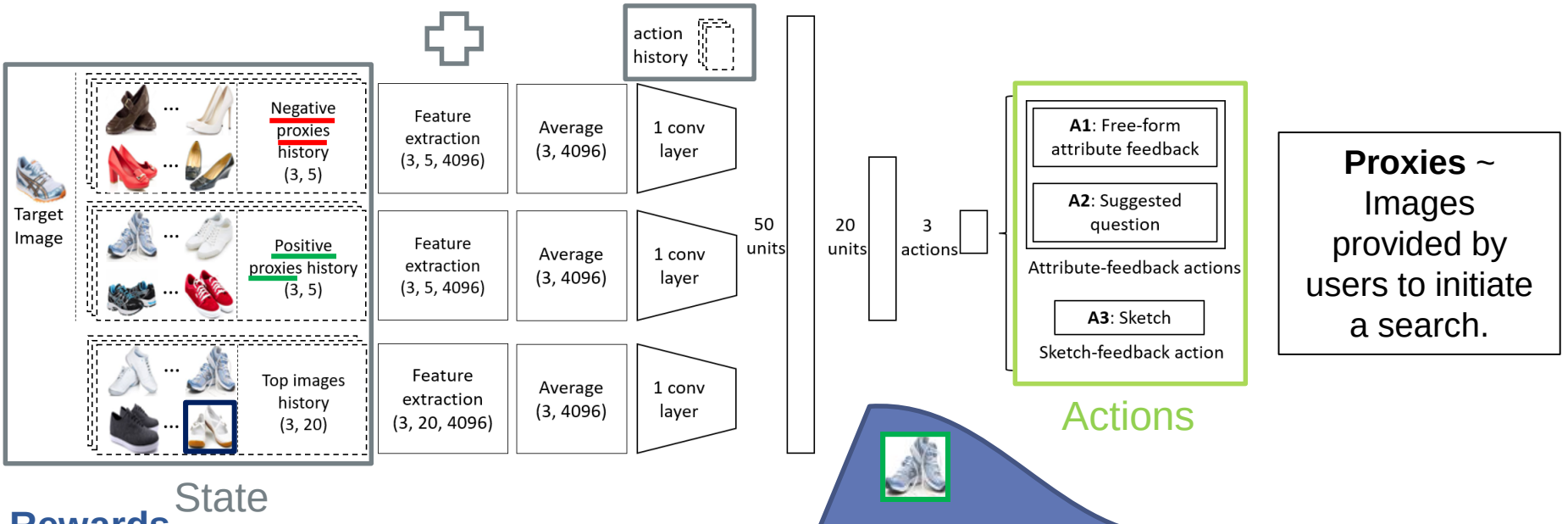
(Yu et al, CVPR 2016)



Rank with one-class SVM

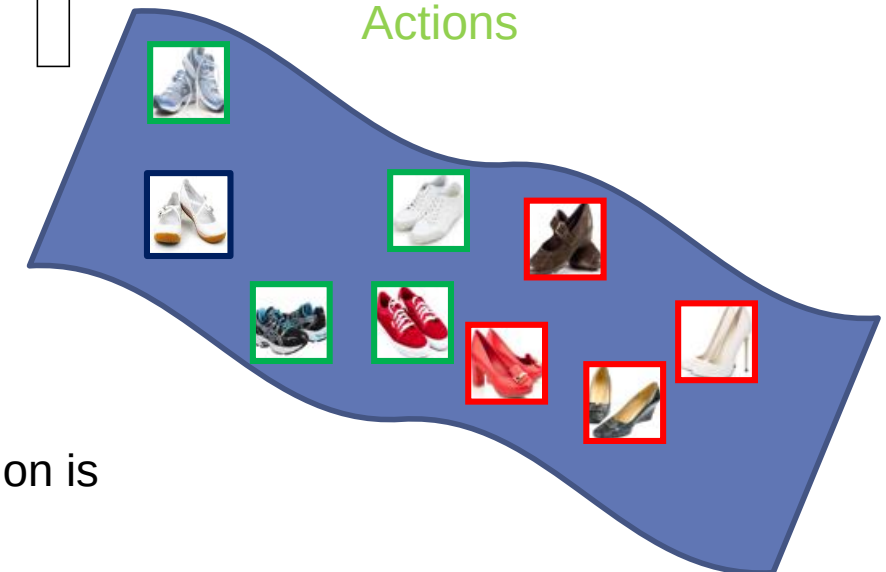
Approach – RL agent

We formulate reinforcement learning as a Markov decision process (MDP).



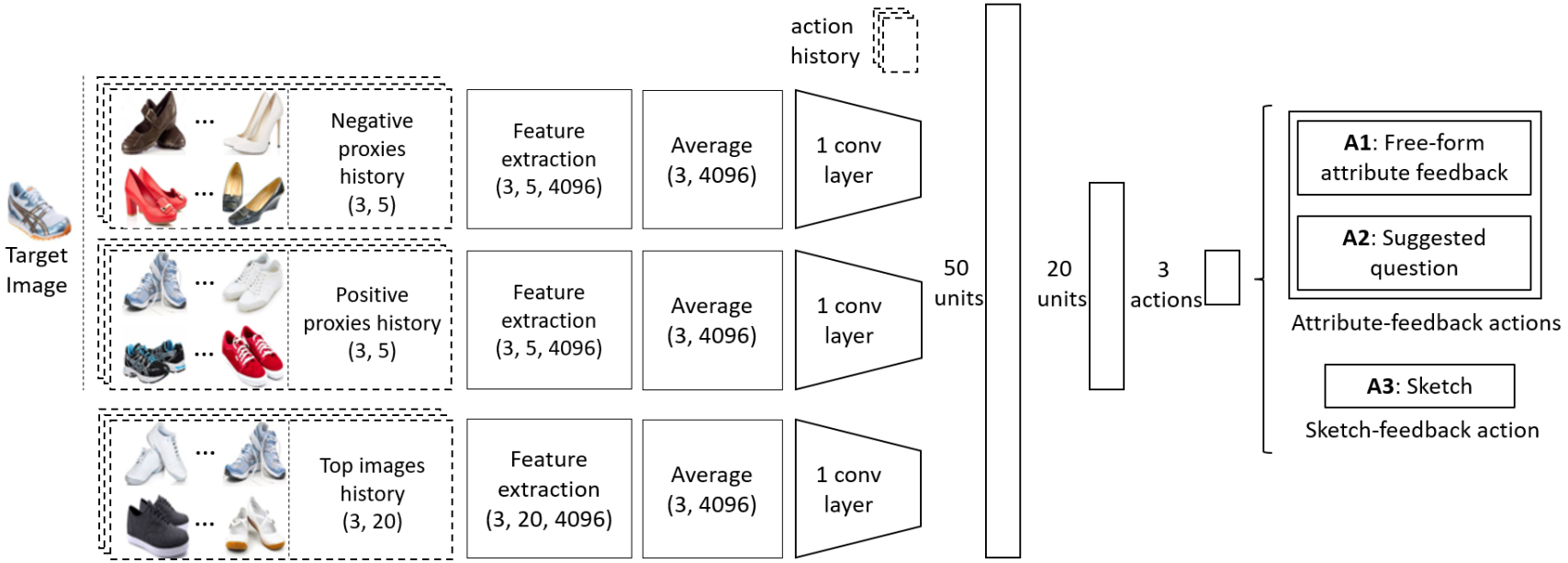
Rewards

- Distance to positive proxies should decrease.
- Distance to negative proxies should increase.
- Assign a negative reward if sketch action is queried more than once.



Approach – learning

Q-learning network

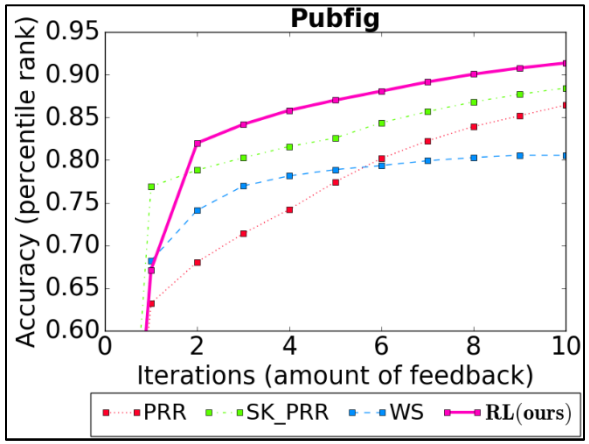


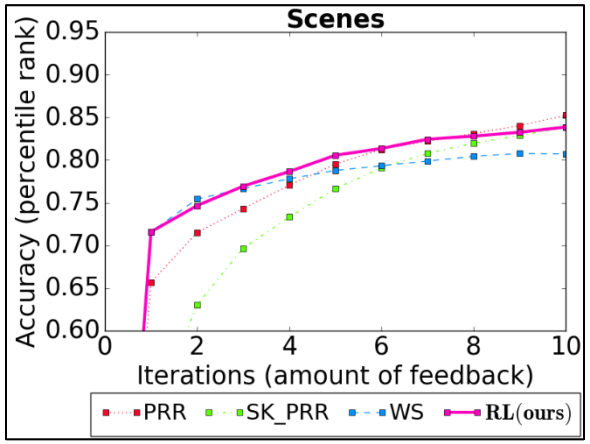
- Neural network with **convolutions** ⇨ capture image features and ordering.
- **Replay-memory** ⇨ collect many instances as the agent is running; useful to remove short-term correlation.
- **Random actions** w/ probability decreasing from 1 to 0.1 ⇨ exploration/exploitation.

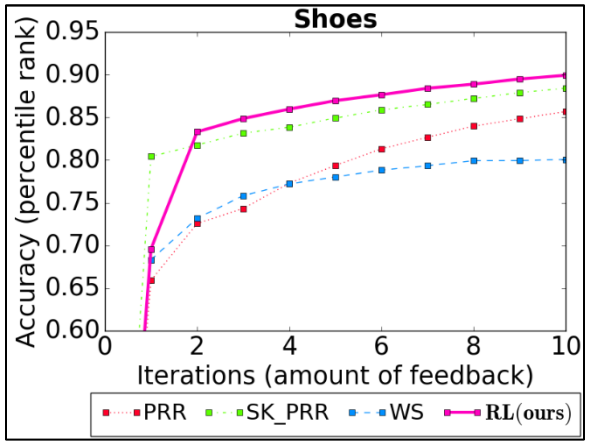
Evaluation – simulated users

Baselines

- *Whittle Search (**WS**):* Users select a (reference image, attribute) and compare target and reference for the chosen attribute.
- *Pivot round-robin (**PRR**):* The machine provides a (reference image, attribute) pair and users select a response.
- *Sketch retrieval + Pivot round robin (**SK_PRR**):* Users provide a sketch, then the machine follows the pivot round-robin strategy.



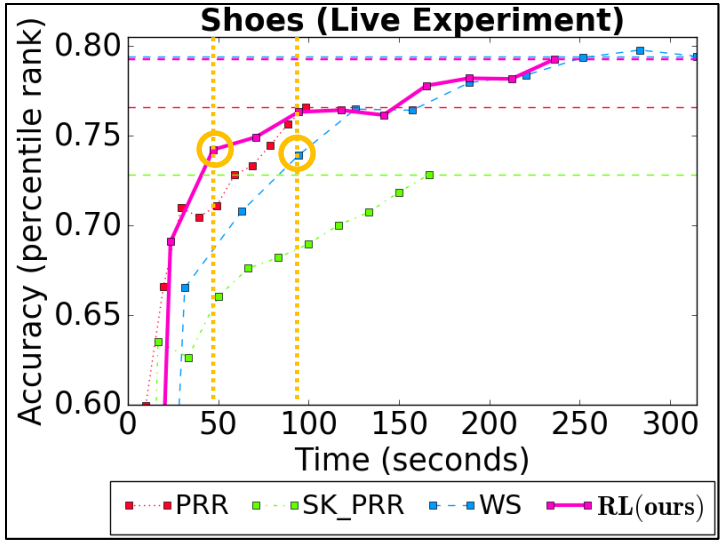




Evaluation – live users

We recruit university and Amazon MTurk participants to conduct **100 searches**.

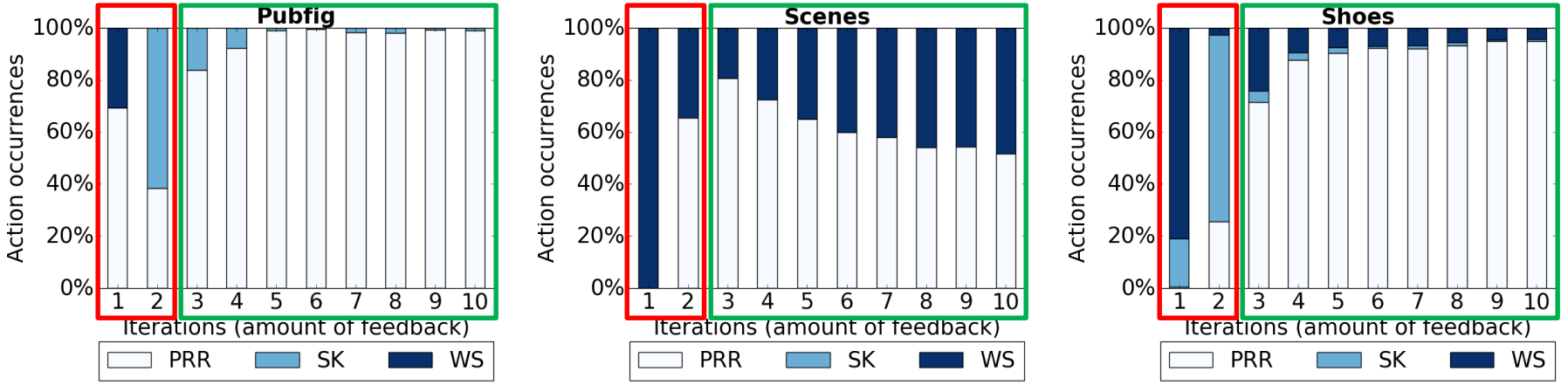
- Users provide sketch and attribute feedback. Users provide high-quality data.
- Our RL agent queries the next action using a REST API.
- We verify the benefit of our adaptive feedback strategy in this realistic scenario.



Qualitative results

To understand our **mixed-initiative RL agent**, we count the predicted action per iteration. We observe that:

- The SK and WS actions (sketch and free-form attribute feedback) are mainly performed in iterations 1 and 2, because they are *exploration-like* actions.
- After iteration 3, PRR (system-chosen attribute questions) is most common.



Our RL agent learned to prioritize **human-initiated** feedback **early on**, and complement it with **machine-initiated** feedback in **later** iterations.

Qualitative results

From our **sketch-to-photo generated images**,

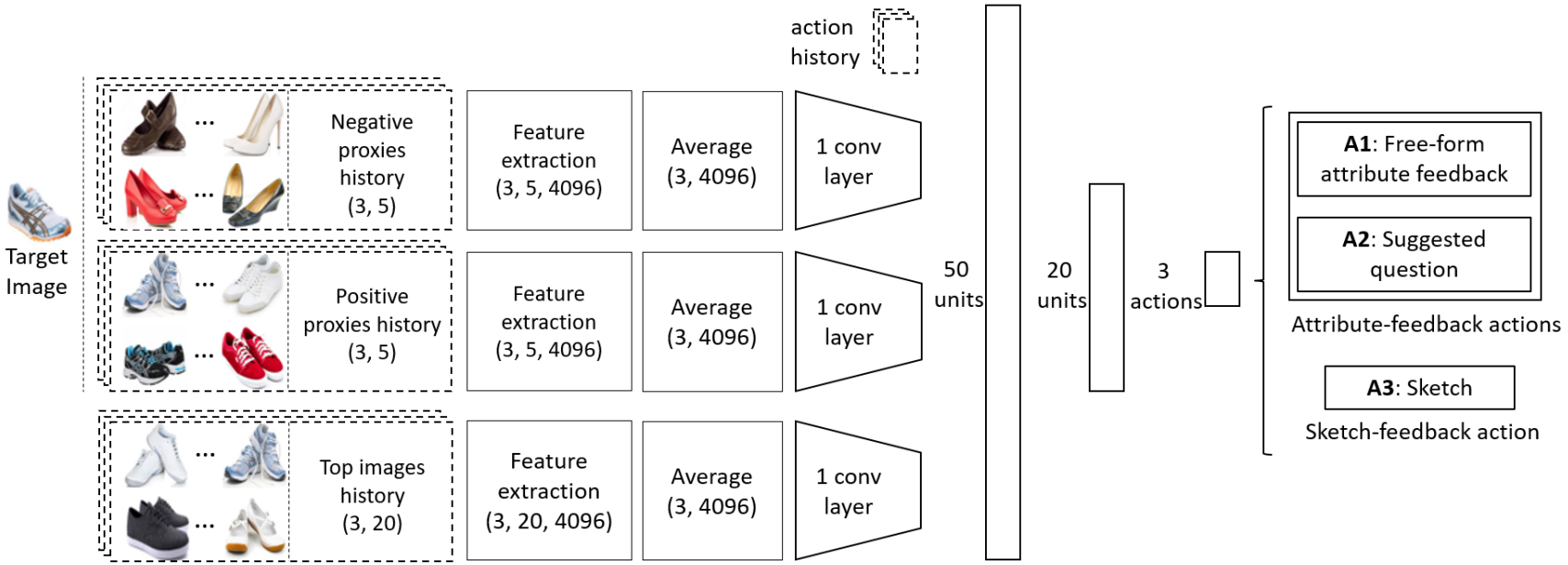
- We observe the most realistic ones correspond to Pubfig, then Shoes and finally Scenes.
- This order also corresponds to the relative performance of our method, best on Pubfig and Shoes.

Pubfig Shoes Scenes

Conclusion

- We explored choosing interactions in a mixed-initiative image retrieval system.
- Our approach selects the **most appropriate interaction per iteration** using RL.
- Our model combines **human-initiated** feedback with **machine-based feedback** for faster retrieval.
- We **outperform** standard image retrieval methods with **real** and **simulated** users.

Thanks!



Nils Murrugarra-Llerena – nineil@cs.pitt.edu
Adriana Kovashka – kovashka@cs.pitt.edu

Department of Computer Science
University of Pittsburgh