Interactive Product Retrieval Challenge

With Instance-Level Recognition Workshop in ICCV2021

POC: Xu Zhang, Sanqiang Zhao
Applied Scientist, Amazon
xzhnamz@amazon.com, sanqiang@amazon.com

Team Members: Emre Brut, Shih-Fu Chang, Varsha Hedau, Pradeep Natarajan, Prem Natarajan, Robinson Piramuthu, Karthik Ramakrishnan, Yue Wu, Xu Zhang, Sanqiang Zhao

amazon

Table of Content

Thanks to the Sponsor

Motivation and Problem Statement

Motivation

- Shopping is an interactive process.

Customer

I want to buy a backpack

Shopping Assistant

How about this one?

I want a red one

This one?

That's cool! Can it be smaller?

Of course, how about this?

Awesome! That's it!

Problem Statement

- Single-Shot Interactive Product Retrieval [1]
  - Given a product image and a relative language feedback, retrieve top-k product images in the catalog that fulfill the feedback's request.

Source Image

Relative language feedback:
More floral

Return Results

Challenge and Dataset

Challenge

- To support the research of Single-Shot Interactive Product Retrieval, we plan to
  - Launch an Interactive Product Retrieval challenge in 2021 Q4 – 2022 Q1
  - Release a new interactive product retrieval dataset in the challenge
- The challenge has two main phases
  - Training Phase
    - The full training set will be released for model training and development
  - Evaluation Phase
    - Gallery images and evaluation set will be released
    - Participants upload a ranked list

amazon

Dataset

- We plan to collect 3 subsets
  - Training/Development set
    - 1 million images with noisy labels
    - ~100k relative language feedbacks between product pairs
  - Gallery set
    - 1 million images (from new products not in the training set)
  - Evaluation set
    - 10k query images with feedbacks
- All data will be released to challenge participants subject to contest rules

Challenges in Interactive Product Retrieval

amazon

Catalog

- The product catalog is extremely large, diverse and keeps Growing.
  - A typical online retailer may sell millions of products

- Long-tail distribution
  - Many products only have 1 image.

Multi-Modal Problem

- Interactive product retrieval requires the understanding for both image and language and building relations between those 2 modalities.
- Indexing the images and language for large-scale search.

Similar Products

- Many products are very similar. How to rank the recommendations?

Source Image

Relative language feedback:
I want something more floral





Source Image

Relative language feedback:
I want something Blue




Diverse Relative Language Feedback

- Relative Language Feedbacks from the user can come from a wide variety of directions.

- Simple attributes
  - Size: Can I have something smaller?
  - Color: Can I have a bag in green?
  - Style: Can I have a t-shirt with V-Neck?
  - Price: Can I have something cheaper?

Diverse Relative Language Feedback (Cont.)

- Abstract attributes
  - Can I have something more formal?
  - Can I have something more elegant?

- Complex attributes
  - Can I have something that better fits my body?
  - Can I have something that better matches my trousers?

Handling Multi-Turn Dialog

- Due to the nature of the problem, interactive product retrieval needs to handle multi-turn dialog.

- User feedback in the previous conversation may have impact to current product recommendation.
  - If the user mentioned he/she doesn’t like green color, products in green shouldn’t appear in the future product recommendation list.

Handling Multi-Turn Dialog (Cont.)

I want to buy a backpack

How about this one?

I want a red one

This one?

That's cool! Can it be smaller?

Of course, how about this?

Dataset Collection

Dataset Recap

- We plan to collect 3 subsets
  - Training/Development set
    - 1 million images with noisy labels
    - ~100k relative language feedbacks between product pairs
  - Gallery set
    - 1 million images (from new products not in the training set)
  - Evaluation set
    - 10k query images with feedbacks
- The dataset can be used for both conventional product retrieval and interactive product retrieval

Training/Development Set

- Around 10k source-target image triplets
  - Include a source image, a target image and a negative image
- 9 feedbacks for each triplet
  - 3 unique feedbacks provided by 3 different workers

Product A (source image)

Product B (target image)

Product C (negative image)

- something in pink
- something that is shorter
- something with a zippered front
- ….

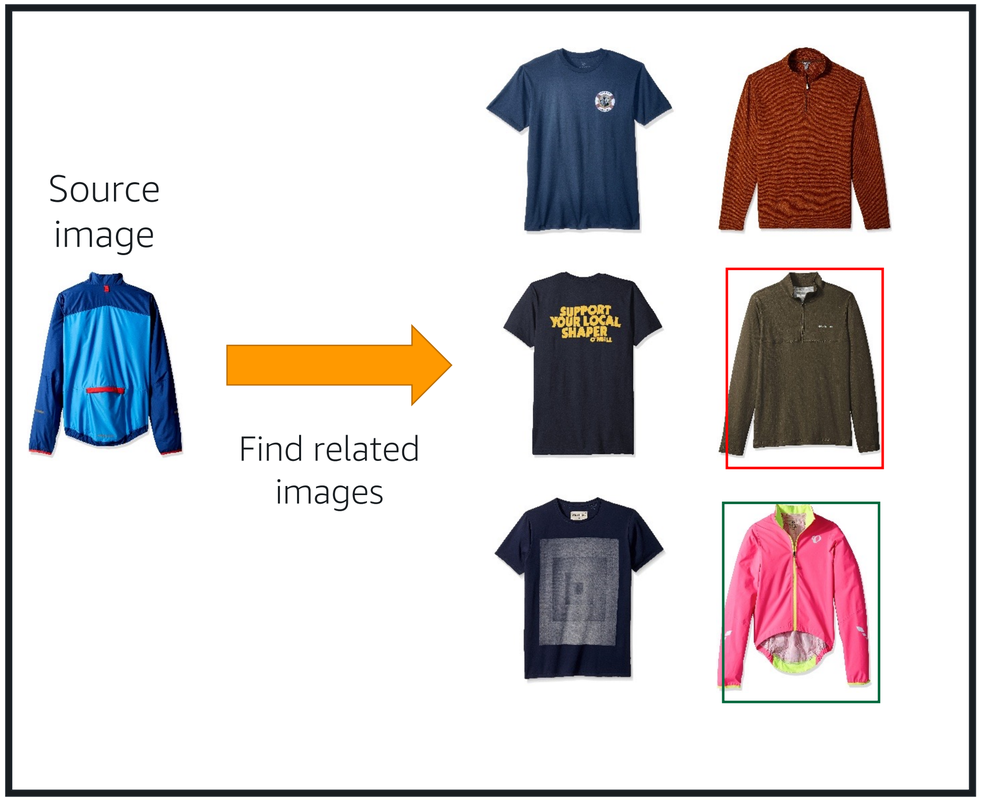
Evaluation Set

- Around 10k source-targets images
  - Include source image and a list of target images
- 9 feedbacks for each of triplets
  - 3 unique feedbacks provided by 3 different workers

Gallery Dataset

- 1M images
- Products are different from the training set
- Dataset will be used as the index for retrieval
- Dataset will be used for evaluation

Feedback Collection



Step 1: Find Related Images

If the assistant provide you product A and you want to buy product B but not product C, how would you say to the assistant.

- something in a pink color
- something that is shorter
- something with a zippered front

Step 2: Collect Feedback

Feedback Collection (Cont.)

- Collect Feedback
- Use Amazon Mechanic Turk to collect feedbacks
  - Qualification Requirement(s)
    - Location is US
    - High HIT Approval Rate
  - Validation
    - Feedbacks contain at least 3 tokens
    - Three feedbacks cannot be the same
  - Each image triplet are assigned to 3 workers

Step 2: Collect Feedback

Dataset Comparison

- Previous relevant Dataset
  - FashionIQ Challenge (Fashion IQ) [1]
    - Conversational-based fashion product search
  - WhittleSearch (WhittleSearch) [2]
    - Attribute-based search
  - UT Zappos50K (UT-Zap50K) [3]
    - Attribute-based search
  - DeepFashion2 Challenge(DeepFashion2) [4]
    - Attribute-based fashion product search

Attribute-based search is focused on attributes and may be limited.

Conversational-based search is based on **what users say**, and can come from a wide variety of directions

Dataset Comparison

- Previous relevant Dataset

Dataset	#Image	#Feedbacks	Task
Fashion IQ [1]	78K	60k	Conversational-based
WhittleSearch [2]	20K	-	Attribute-based
UT-Zap50K [3]	50K	-	Attribute-based
DeepFashion2 [4]	491k	-	Attribute-based
Ours	~2M	~180k	Conversational-based

Evaluation

Evaluation

- One major challenge for interactive product retrieval is that there may be multiple matching products
  - There may have thousands of products matching the same criteria
  - It’s impractical to annotate all of them

Source Image

Relative language feedback:
More floral

Metrics

- We consider to use 2 different methods for evaluation*
  - Evaluate in a smaller set.
  - Pooling [5]

*This may not be the final evaluation metric. The final evaluation metric will be released as the contest rules.

Metrics (Cont.)

- Evaluate in a smaller set
  - Select and annotate a subset.
  - Evaluate with mAP in that subset.

Source Image

Relative language feedback:
More floral

Metrics (Cont.)

- Pooling
  - Pooling is a popular evaluation method in information retrieval
    - Top K result from all participants are got pooled and annotated
    - Precision@K will be used as the final metric.

Future Work

- Multi-turn interactive product search
  - The multi-turn (dialog) interactive product retrieval is required for a complete virtual shopping assistant experience.
  - Major challenges include data collection and evaluation

- Deal with user uploaded images
  - Users may take a real life photo and ask for product recommendation with relative language feedback.

Discussion

QA

Thank you!

Reference

- [1] H. Wu *et al.*, “Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback,” *CVPR 2021*
- [2] Kovashka, Adriana, Devi Parikh, and Kristen Grauman. "Whittlesearch: Image search with relative attribute feedback." *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012.
- [3] A. Yu and K. Grauman. "Fine-Grained Visual Comparisons with Local Learning". In CVPR, 2014.
- [4] Ge, Yuying, et al. "Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.
- [5] https://trec.nist.gov/