Is It All Relative?
Interactive Fashion Search based on Relative Natural Language Feedback

Rogerio Schmidt Feris
Principal Research Scientist and Manager
IBM Research & MIT-IBM Watson AI Lab
Computer Vision for Fashion

**Style discovery and analysis**
(MH Kiapour, et al, ECCV 2014)

**Trend modeling and forecast**
(R He, et al, WWW 2016)

**Outfit recommendation**
(WL Hsiao, et al, CVPR 2018)

**Virtual Try-on**
(X Han, et al, CVPR 2018)
This Talk: Fashion Image Search
with (subjective) visual attributes
Fashion Search: Challenges

- **Subjective Attributes**
  - Hard to describe the desired fashion item in words and resolve user intent

- **Filter choices are limited.**
  - Hard to narrow down search results to the desired style.

[Image of a shoe with labels: Formal? User labels: 50% “yes” 50% “no”]

[Kovashka and Grauman, 2016]

Black lace dress

[Table with filter options: Pattern +, Color - (Black, Grey, White, Off-white)]
Interactive Personal Shopper

HI! What are you interested in shopping today?

OR show me a photo

User drag-n-drop a look that is similar to what she/he is looking for.
Interactive Personal Shopper

(Street2shop) Retrieved results based on user input image.

Pick the one you are most interested in

Refine search results by taking user feedback.

Or tell me your preferences

I prefer black color.
Interactive Personal Shopper

Pick the one you are most interested in

User can iteratively interact with the search interface

Or tell me your preferences

Like the right one but with different neckline
Interactive Personal Shopper

Pick the one you are most interested in

Or tell me your preferences
Outline

- Street2Shop

Interactive image search using natural language feedback

[Guo & Wu et al, NeurIPS 2018]
[Guo & Wu et al, 2019]

Or tell me you preferences

*Like the right one but with different neckline*
Clothing Retrieval (Street2Shop)

Input: **User Photo**

Retrieved Images from **Online Shopping** Stores

[Liu et al, CVPR 2012] [Kiapour et al, ICCV 2015] [Huang et al, ICCV 2015]
Problem: Domain Discrepancy

Proposed Approach: 
**Dual Attribute-Aware Ranking Network (DARN)** [Huang et al, ICCV 2015]
Weakly labeled data from shopping websites

- 9,000 image pairs (exact same clothing)

- Noisy attribute labels (9 classes, 179 values)

<table>
<thead>
<tr>
<th>Attribute categories</th>
<th>Examples (total number)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clothes Button</td>
<td>Double Breasted, Pullover, ... (12)</td>
</tr>
<tr>
<td>Clothes Category</td>
<td>T-shirt, Skirt, Leather Coat ... (20)</td>
</tr>
<tr>
<td>Clothes Color</td>
<td>Black, White, Red, Blue ... (56)</td>
</tr>
<tr>
<td>Clothes Length</td>
<td>Regular, Long, Short ... (6)</td>
</tr>
<tr>
<td>Clothes Pattern</td>
<td>Pure, Stripe, Lattice, Dot ... (27)</td>
</tr>
<tr>
<td>Clothes Shape</td>
<td>Slim, Straight, Cloak, Loose ... (10)</td>
</tr>
<tr>
<td>Collar Shape</td>
<td>Round, Lapel, V-Neck ... (25)</td>
</tr>
<tr>
<td>Sleeve Length</td>
<td>Long, Three-quarter, Sleeveless ... (7)</td>
</tr>
<tr>
<td>Sleeve Shape</td>
<td>Puff, Raglan, Petal, Pile ... (16)</td>
</tr>
</tbody>
</table>
Dual Attribute-Aware Ranking Network (DARN)

- Two sub-networks to model each domain (shopping and user images)
Dual Attribute-Aware Ranking Network (DARN)

- Two sub-networks to model each domain (shopping and user images)

Shopping Images

User Images

Max pool

Max pool

Max pool

Max pool

Triplet Ranking Loss

Category (20 D)

Color (56 D)

Category (20 D)

Color (56 D)
- Triplet Ranking loss function connecting the two sub-networks
- (visual similarity constraint)
Dual Attribute-Aware Ranking Network (DARN)

- Semantic embedding: simultaneous attribute learning and retrieval
- FC features are transmitted to multiple branches

Shopping Images

User Images

Cross-Entropy Loss

Noisy attributes as privileged info (weak supervision)

Triplet Ranking Loss

Category (20 D)

Color (56 D)
Dual Attribute-Aware Ranking Network (DARN)

- Features from conv layers for encoding more localized information

Shopping Images

User Images

Triplet Ranking Loss

Category (20 D)
Color (56 D)

Dual Attribute-Aware Ranking Network (DARN)

- **Test time:** Cross-domain Clothing Retrieval
- For each image in the gallery, compute features and store them in a database

Shopping Images

```
MAX pool
```

```
MAX pool
```

```
MAX pool
```

```
MAX pool
```

```
Triplet Ranking Loss
```

```
FC
```

```
FC
```

```
FC
```

```
Category (20 D)
```

```
Color (56 D)
```

```
Color (56 D)
```

```
Category (20 D)
```

```
Test time: Cross-domain Clothing Retrieval
```

```
For each image in the gallery, compute features and store them in a database
```

Dual Attribute-Aware Ranking Network (DARN)

- **Test time:** Cross-domain Clothing Retrieval
- For each image in the gallery, compute features and store them in a database

![Diagram](image)

Shopping Images

**Max pool**

**Triplet Ranking Loss**

**Category (20 D)**

**Color (56 D)**
Dual Attribute-Aware Ranking Network (DARN)

- **Test time:** Cross-domain Clothing Retrieval
- For each image in the gallery, compute features and store them in a database.
Dual Attribute-Aware Ranking Network (DARN)

- **Test time:** Cross-domain Clothing Retrieval
- Given a query image, compute features and rank-order the gallery based on Euclidean distance
Experimental Results

Top-k retrieval accuracy on 200,000 retrieval gallery. The number in the parentheses is the top-20 retrieval accuracy.

First Column: Query
Green Box: Exact same clothing
Outline

- Street2Shop

Interactive image search using natural language feedback

[Guo & Wu et al, NeurIPS 2018]
[Guo & Wu et al, 2019]

Or tell me you preferences

Like the right one but with different neckline
Fashion Search using Interactive Feedback

**Relevance Feedback** [Rui et al, 1998]

- Limit the information the user can convey about an image
- Pre-defined set of attributes (limited vocabulary, cumbersome interface)

**Relative Attribute Feedback** [Kovashka et al, 2012]

- More open
- More pointy
FASHION IQ DEMO

IBM RESEARCH AI
Network Architecture

Candidate $t$

The one I want has a closed back and crystal buckle in the front.

Candidate $t+1$

Dialog Turns
The one I want has a closed back and crystal buckle in the front.

The goal of the Response Encoder is to embed the information from the t-th dialog turn to a joint visual semantic representation.
The State Tracker receives as input the response representation, combines it with the history representation of previous dialog turns, and outputs the history representation.
Network Architecture

During testing, the candidate for $t+1$ round is selected by finding the closest database feature to the history representation.
The one I want has a closed back and crystal buckle in the front.

Training the network

• How to obtain training data? Expensive and slow to collect dialog data from real users.
Training Dialog Manager with User Simulator

- Relative captioner: surrogate for real users
  - Automatically generates sentences describing the visual differences between target and reference images
  - New task and new dataset!

"The one I want has an open back design with suede texture."

"The one I want has a closed back and crystal buckle in the front."
AMT task to collect human-written relative expressions

Shoes Relative Captions Dataset:
- ~10K training images, ~5K testing images
- 1 relative expression per image
Relative Captioner (User Simulator) Model

- Feature concatenation of target and reference images
- Show, Attend, and Tell model [Xu et al, 2015] to generate relative captions

Example predictions:

1. Unlike the provided image, the one(s) I want are blue and green sneakers
2. are floral print with an all-over floral pattern
3. are brown with a higher heel
4. are black with a thicker heel
Training the network

1. Supervised pre-training (triplet loss)

\[ \mathcal{L}^\text{sup} = \mathbb{E} \left[ \sum_{t=1}^{T} \max(0, \| s_t - x^+ \|_2 - \| s_t - x^- \|_2 + m) \right] \]

- History representation
- Target feature
- Random image feature

2. Reinforcement Learning to maximize the rank of the target image, with model-based policy improvement

- The one I want has a closed back and crystal buckle in the front.
- Candidate Generator
- Candidate t
- Candidate t+1
- Response Encoder
- State Tracker
- Dialog Manager
- Retrieval Database
- User Simulator
- Simulator
- Training the network
- GRU
- MLP
- Text Encoder
- Concatenation
- MLP
- Linear
- K-NNs
- Stochastic Sampling
- Greedy Sampling
- Image Rep.
Results

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.

Effectiveness of Natural Language Feedback

**Attr\_n** and **Attr\_n**(deep): dialog managers trained with relative attribute feedback. A rule-based feedback generator concatenates respective attribute words with “more” or “less”.
Leveraging Side Information

Text surrounding fashion images as weak supervision
Extracting Visual Attributes from Text

Product Webpage

- **Product Title**
  - Southpole Junior’s Plus Size one Side Ruffle Shoulder Floral Fashion top
  - Size: 3X
  - Color: Black
  - Size: 1X
  - Color: Black
  - 57% Cotton/43% Rayon
  - Machine Wash
  - One shoulder top
  - Fashion top

- **Product Summary**
  - Package Dimensions: 14.2 x 6.4 x 1.5 inches
  - Shipping Weight: 6.4 ounces
  - ASIN: B0066G0Q4E
  - Item model number: 12128-1120
  - Date first listed on Amazon: March 23, 2012
  - Domestic Shipping: Item can be shipped within U.S.
  - International Shipping: This item is not eligible for international shipping.

- **Detailed Description**
  - Floral, stripe, parsley, distressed, dot, plaid, panel, woven, leather, fit, maxi, halter, strappy, high-slit, yoga, retro, beach, polka, tribal, muscle, boxy, ... ... ...

- **Fashion attribute extraction**
  - one side, ruffle, shoulder, floral, top, cotton

Attribute List (1000 phrases, [DeepFashion])
Attribute Prediction Network

- Similar to our DARN work we used an attribute prediction network to obtain attribute-aware visual features
- Use this information as a weak supervisory signal
The one I want has a closed back and crystal buckle in the front.
The one I want has a closed back and crystal buckle in the front.
Network Architecture

Attribute-aware Response Encoder

- AttrNet
- ResNet
- LSTM

Response Representation

The top is orange in color and more flowy

Retrieval Database

Candidate Generator

Response Rep

Dialog Manager

K-NNs

Stochastic Sampling

Greedy Sampling

Train

Test

MLP

Image Rep.
The one I want has a closed back and crystal buckle in the front.
Network Architecture

Attribute-aware User Simulator

Candidate t

Attribute - aware
User Simulator

Candidate t +1
### Fashion IQ Dataset


- Dresses, Tops & Tees, and Shirts (~60K relative captions)

<table>
<thead>
<tr>
<th></th>
<th>Dresses</th>
<th></th>
<th>Tops &amp; Tees</th>
<th></th>
<th>Shirts</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>train / val / test</td>
<td>total</td>
<td>train / val / test</td>
<td>total</td>
<td>train / val / test</td>
<td>total</td>
</tr>
<tr>
<td># Images</td>
<td>11452 / 3817 / 3818</td>
<td>19087</td>
<td>16121 / 5374 / 5374</td>
<td>26869</td>
<td>19036 / 6346 / 6346</td>
<td>31728</td>
</tr>
<tr>
<td># Images with side info</td>
<td>7741 / 2561 / 2653</td>
<td>12955</td>
<td>9925 / 3303 / 3210</td>
<td>16438</td>
<td>12062 / 4014 / 3995</td>
<td>20071</td>
</tr>
<tr>
<td># Relative Captions</td>
<td>11970 / 4034 / 4048</td>
<td>20052</td>
<td>12054 / 3924 / 4112</td>
<td>20090</td>
<td>11976 / 4076 / 4078</td>
<td>20130</td>
</tr>
</tbody>
</table>

Dresses

Top & Tees

Shirts
Results – Attribute-aware User Simulator

<table>
<thead>
<tr>
<th></th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
<th>Meteor</th>
<th>Rouge-L</th>
<th>CIDEr</th>
<th>SPICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute-aware (D)</td>
<td>61.3</td>
<td>44.1</td>
<td>29.0</td>
<td>19.7</td>
<td>26.2</td>
<td>55.5</td>
<td>59.4</td>
<td>34.7</td>
</tr>
<tr>
<td>with Attention (S)</td>
<td>57.7</td>
<td>46.3</td>
<td>32.9</td>
<td>22.3</td>
<td>27.9</td>
<td>57.1</td>
<td>78.8</td>
<td>36.6</td>
</tr>
<tr>
<td>(T)</td>
<td>58.4</td>
<td>44.1</td>
<td>29.6</td>
<td>20.3</td>
<td>26.5</td>
<td>54.1</td>
<td>63.3</td>
<td>35.3</td>
</tr>
<tr>
<td>Attribute-aware (D)</td>
<td>58.5</td>
<td>42.0</td>
<td>26.7</td>
<td>17.5</td>
<td>24.0</td>
<td>53.2</td>
<td>42.7</td>
<td>30.8</td>
</tr>
<tr>
<td>via Concatenation (S)</td>
<td>54.5</td>
<td>42.6</td>
<td>29.1</td>
<td>19.4</td>
<td>25.8</td>
<td>53.5</td>
<td>47.1</td>
<td>31.8</td>
</tr>
<tr>
<td>(T)</td>
<td>55.9</td>
<td>41.0</td>
<td>26.0</td>
<td>17.0</td>
<td>25.4</td>
<td>51.5</td>
<td>40.7</td>
<td>31.1</td>
</tr>
<tr>
<td>Image-Only (D)</td>
<td>58.1</td>
<td>41.0</td>
<td>26.3</td>
<td>17.4</td>
<td>24.8</td>
<td>53.6</td>
<td>48.9</td>
<td>32.1</td>
</tr>
<tr>
<td>(S)</td>
<td>53.2</td>
<td>41.9</td>
<td>29.0</td>
<td>19.6</td>
<td>25.9</td>
<td>53.8</td>
<td>52.6</td>
<td>32.0</td>
</tr>
<tr>
<td>(T)</td>
<td>54.0</td>
<td>39.4</td>
<td>24.6</td>
<td>15.7</td>
<td>24.3</td>
<td>50.5</td>
<td>41.1</td>
<td>30.6</td>
</tr>
</tbody>
</table>

(D) Dresses, (S) Shirts, (t) Tops&Tees

- Attribute-aware methods outperform image-only baselines
- Attention mechanism can better utilize the additional attribute information
Results – Interactive Image Retrieval

- Attribute information and relative expressions jointly lead to better retrieval results.
- More advanced techniques for composing side information, relative feedback and image features could lead to further performance gains.

<table>
<thead>
<tr>
<th></th>
<th>Dialog Turn 1</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R@5</td>
<td>R@10</td>
<td>R@50</td>
<td>P</td>
</tr>
<tr>
<td>Attribute-aware (D)</td>
<td>90.52</td>
<td>4.74</td>
<td>7.73</td>
<td>23.94</td>
<td>98.09</td>
</tr>
<tr>
<td>with Attention (S)</td>
<td>90.87</td>
<td>2.88</td>
<td>4.96</td>
<td>17.32</td>
<td>98.02</td>
</tr>
<tr>
<td>(T)</td>
<td>90.37</td>
<td>3.07</td>
<td>5.16</td>
<td>17.27</td>
<td>98.04</td>
</tr>
<tr>
<td>Attribute-aware (D)</td>
<td>90.39</td>
<td>4.52</td>
<td>7.48</td>
<td>24.14</td>
<td>98.00</td>
</tr>
<tr>
<td>via Concatenation (S)</td>
<td>89.93</td>
<td>2.41</td>
<td>4.09</td>
<td>14.86</td>
<td>97.55</td>
</tr>
<tr>
<td>(T)</td>
<td>90.34</td>
<td>3.22</td>
<td>5.39</td>
<td>17.75</td>
<td>98.03</td>
</tr>
<tr>
<td>Image-Only (D)</td>
<td>89.45</td>
<td>3.79</td>
<td>6.25</td>
<td>20.26</td>
<td>97.49</td>
</tr>
<tr>
<td>(S)</td>
<td>89.39</td>
<td>2.29</td>
<td>3.86</td>
<td>13.95</td>
<td>97.40</td>
</tr>
<tr>
<td>(T)</td>
<td>87.89</td>
<td>1.78</td>
<td>3.03</td>
<td>12.34</td>
<td>96.82</td>
</tr>
</tbody>
</table>
Summary

- Natural language user feedback provides a more natural, expressive, and effective way to interactive image search.
- Incorporating side information is a low-cost, effective technique to further improve retrieval results.

Challenges ahead
- The data issue: user simulator does not accurately model real-user behavior (personal preference, fashion expertise, history, ...)
- Users can communicate better if the agent can ask informative questions in addition to showing images.
Thank you!


Check out the fashion IQ challenge at ICCV 2019!