Learning More from Less: Weak Supervision and Beyond

Rogerio Schmidt Feris
Principal Research Scientist and Manager
IBM Research & MIT-IBM Watson AI Lab
The battle against the long tail

- Training accurate deep neural network models usually requires lots of labeled data
  - Data collection and annotation is expensive, tedious, time-consuming.
  - Crowdsourcing may be infeasible for proprietary data.
  - For some tasks, data may not be available at all (long tail distribution)
This talk

- Weak supervised learning for fashion search
- Learning with less labels beyond weak supervision
Street2Shop

Hadi Kiapour, M., Han, X., Lazebnik, S., Berg, A. C., & Berg, T. L. Where to buy it: Matching street clothing photos in online shops. ICCV 2015

Slide credit: Tamara Berg
Street2Shop Clothing Retrieval

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)
Weakly labeled data from shopping websites

- 9,000 image pairs mined from customer review websites (exact same clothing)

- Noisy attribute labels mined from online shopping stores (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)
Dual Attribute-Aware Ranking Network (DARN)

- Triplet Ranking loss function connecting the two sub-networks
- (visual similarity constraint)

\[
\text{Loss}(a, b, c) = \max(0, m + \text{dist}(a, b) - \text{dist}(a, c))
\]
Dual Attribute-Aware Ranking Network (DARN)

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

![Diagram of DARN network]

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
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

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
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 of DARN network]

Shopping Images
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

Our method (DARN) achieves the best results compared to other state-of-the-art approaches.

Top-k retrieval accuracy on 200,000 retrieval gallery. The number in the parentheses is the top-20 retrieval accuracy.
Attributes as a weak supervisory signal

- Mining attributes from text surrounding the images
Network Architecture

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

Attribute-aware Response Encoder

AttrNet
ResNet
LSTM

σ(·)

The top is orange in color and more flowy

Response Representation

The top is orange in color and more flowy

The top is orange in color and more flowy
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

Candidate t

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

Candidate t+1

Training [Guo & Wu et al, NeurIPS 2018]:

- Reinforcement learning (reward: rank of the target image)
- User simulator
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!
Network Architecture

Attribute-aware User Simulator

Candidate t

Attribute-agnostic User Simulator

Candidate t+1

Te  Fa  Sh  Pa  St  \( x_{target} \)

Te  Fa  Sh  Pa  St  \( x_{candid} \)

AttrNet Feature  ResNet Feature

AttrNet Feature  ResNet Feature

AttrNet Feature  ResNet Feature

\( \phi^{a_t} \)

Diff

RNN

RNN

RNN

Has

longer

sleeves

\( a_t \)

\( o_t \)

\( a_{t+1} \)
Fashion IQ Dataset
https://www.spacewu.com/posts/fashion-iq/

- Images sourced from Amazon, including three classes, Dresses, Tops & Tees, and Shirts (~60K relative captions)

<table>
<thead>
<tr>
<th></th>
<th>Dresses</th>
<th>Tops &amp; Tees</th>
<th>Shirts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>train / val / test</td>
<td>total</td>
<td>train / val / test</td>
</tr>
<tr>
<td># Images</td>
<td>11452 / 3817 / 3818</td>
<td>19087</td>
<td>16121 / 5374 / 5374</td>
</tr>
<tr>
<td># Images with side info</td>
<td>7741 / 2561 / 2653</td>
<td>12955</td>
<td>9925 / 3303 / 3210</td>
</tr>
<tr>
<td># Relative Captions</td>
<td>11970 / 4034 / 4048</td>
<td>20052</td>
<td>12054 / 3924 / 4112</td>
</tr>
</tbody>
</table>

Relative Captions:
- "no sleeve flapping blouse"
- "it has no sleeves and it is plain"

Attribute Labels:
- ruffle, wash, fit

Relative Captions:
- "has a blue collar"
- "has a blue color"

Attribute Labels:
- cotton, twill, wash, button-front, single-button

Relative Captions:
- "is white with a black belt"
- "is lighter in color"

Attribute Labels:
- stripe, cotton, gauze, tiered, wash, tube, braided
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) (T)</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>Attribute-aware (D)</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>via Concatenation (S) (T)</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>Image-Only (D) (S) (T)</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>
</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.
This talk

- Weak supervised learning for fashion search

- Learning with less labels beyond weak supervision
IBM Research AI – Learning with Less Labels for Vision

- **Data Augmentation**
  - Delta (Δ)-encoder
  - Label Set Operations (LaSO)

- **Visual Learning**
  - Representative-based metric learning (RepMet)
  - Learning with Semantics

- **Transfer Learning**
  - Learning to Transfer
  - SpotTune
Transfer Learning

Model Selection
[Dube et al, Deep Vision Workshop 2019]

SpotTune [Guo et al, CVPR 2019]
Sample Synthesis for Few-Shot Learning

Delta-Encoder
[Schwartz & Karlinsky et al, NeurIPS 2018]

LaSO
[Alfassy & Karlinsky et al, CVPR 2019]
Few-shot Learning

RepMet
[Karlinsky et al, CVPR 2019]

Learning with Semantics
[Schwartz & Karlinsky et al, Language & Vision Workshop, 2019]
Summary

- **Takeaway message**: Noisy visual attribute labels mined from the web are useful as *privileged information* during training to improve image search:
  - Street2Shop fashion retrieval [Huang et al, ICCV 2015]
  - Dialog-based interactive fashion retrieval [Guo & Wu et al, NeurIPS 2018] [Guo & Wu et al, 2019]

- Check out our recent work on learning with less labels @CVPR
IBM Research AI: Learning More from Less in Vision @ CVPR


