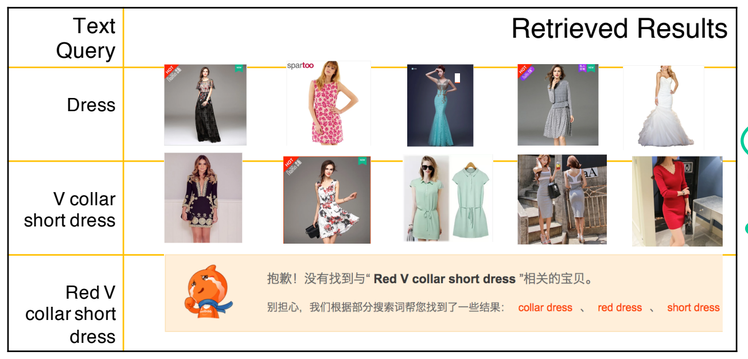
Interpretable Multimodal Retrieval for Fashion Products

Lizi Liao[1], Xiangnan He[1], Bo Zhao[2], Chong-Wah Ngo[3], Tat-Seng Chua[1]

[1]National University of Singapore, [2]University of British Columbia, [3]City University of Hong Kong

Motivation



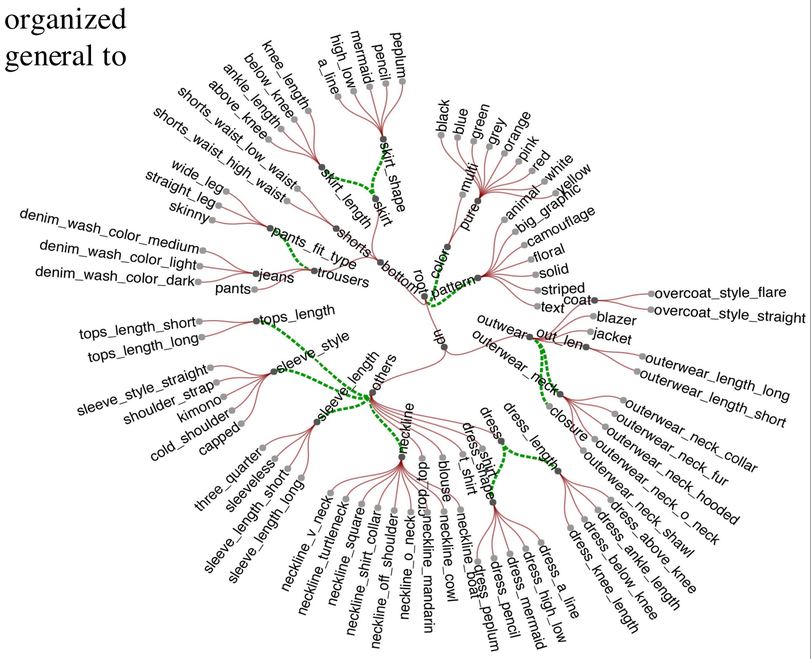
It is hard to comprehensively describe a product with rich details in pure text.

There exists the well-known semantic gap. It is hard to modify certain details when search by image.

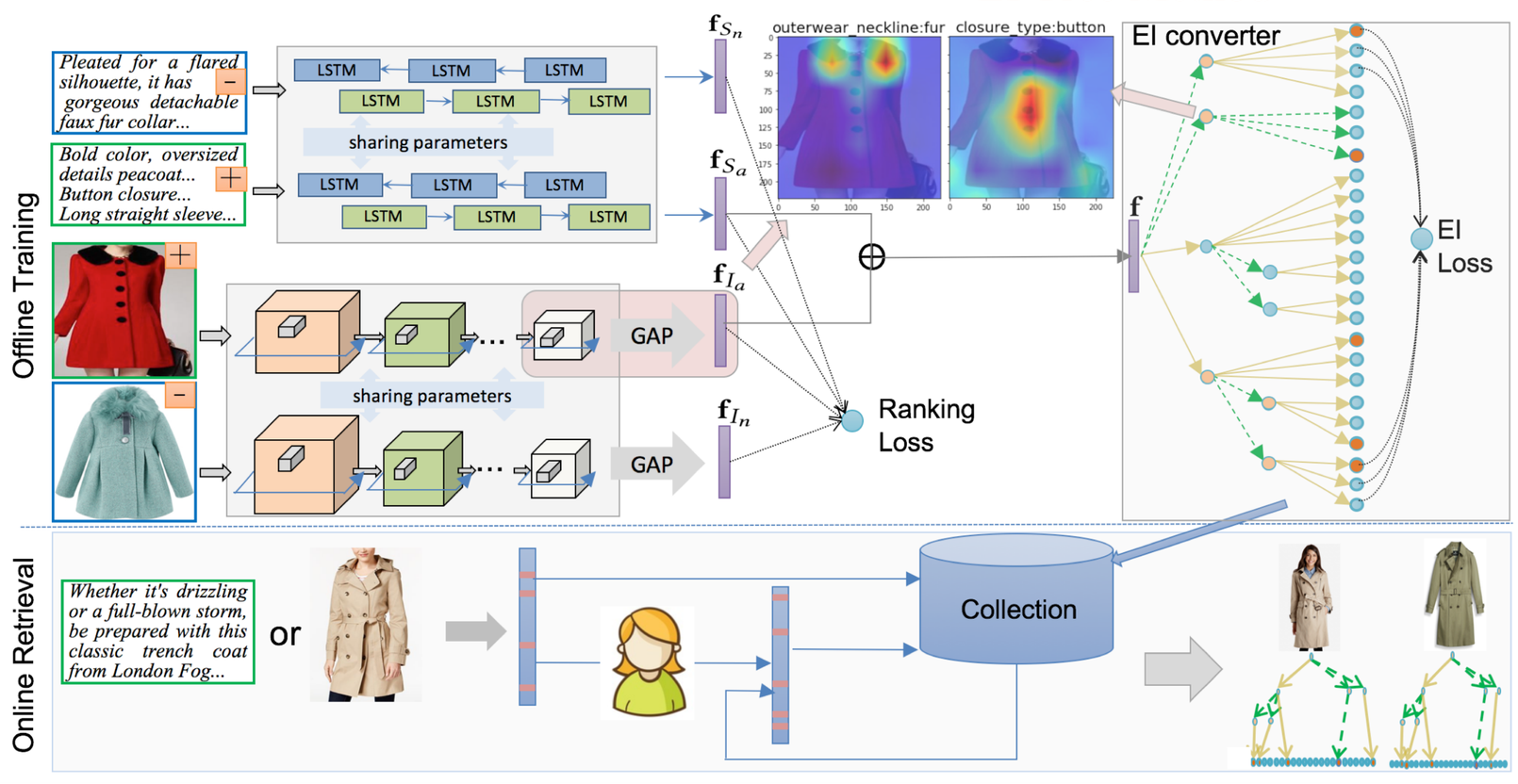
As evidenced by Black Friday's record-high of $5.03 billion online sales in U.S. and Alibaba's $25 billion Singles Day sales in 2017, the modern e-commerce traffic volume is growing fast. At the same time, consumers have become very exigent. For instance, they may have in mind a specific fashion item in a particular color or style, and want to find it online without much effort. Therefore, making the retrieval procedure explainable and being able to leverage user feedback become essential requirements.

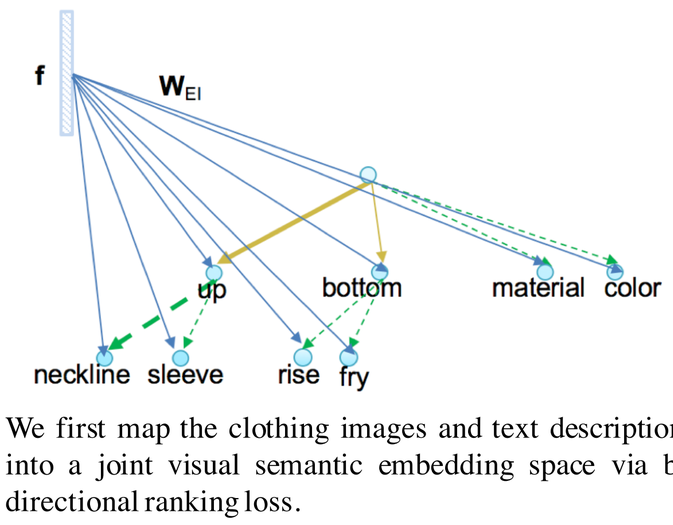
EI Tree

- General to specific semantic concepts
  - -- top level concepts such as *up, bottom*
- Exclusive & Independent relations
  - -- siblings about product categories usually share exclusive relationship
  - -- siblings about attributes are often captured by independent relationship
- We obtained an EI tree with 334 concept nodes organized into six levels from general to specific.



EI Tree Method

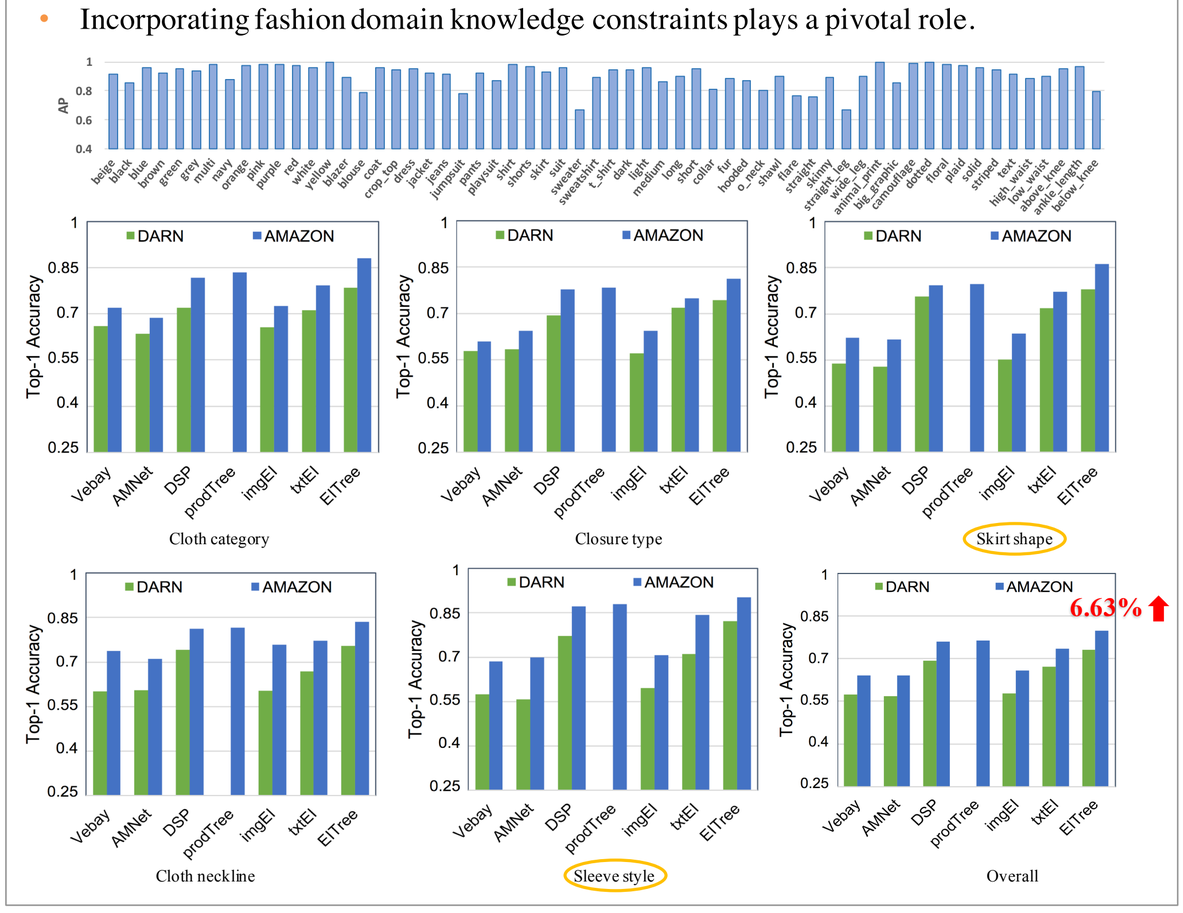




- We first map the clothing images and text descriptions into a joint visual semantic embedding space via bi-directional ranking loss.
- We then apply the EI tree to guide the learning procedure and obtain meaningful representations where each dimension corresponds to a concrete fashion concept.
- Each concept is traced from the root to itself along the EI tree and a probability is generated based on the tracing path, which mimics the general to specific recognition procedure.

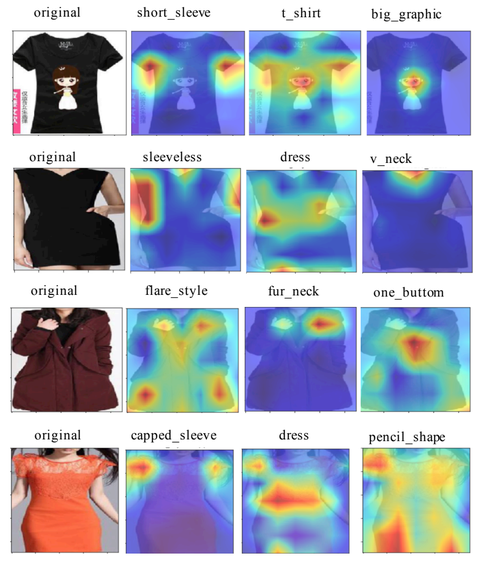
Concept Prediction

- Compared to pure image-based methods, multi-modal methods perform significantly better for concepts with large intra-concept visual variance but are easy to describe in words.
- Incorporating fashion domain knowledge constraints plays a pivotal role.



Case Studies

- Concepts are mapped to spatial regions
  - -- neckline is most likely to occur in upper part of images
  - -- sleeve often occurs on two sides of cloth images
  - -- big-graphic is usually around the center region of a cloth
- Concepts under the same parent node describe the similar spatial part of a cloth
  - -- e.g., peplum skirt and pencil skirt, or v-neck and o-neck
- General to specific spatial regions corresponding to relations
  - -- T-shirt includes parts: short-sleeve and big-graphic
  - -- coat includes cloth parts fur-neckline and flare-style



- Capable of accurately capturing user intentions on fashion concepts.
- Modifying several concepts at the same time does not deteriorate the performance much.
- This interactive fashion retrieval scheme can actually be easily integrated into chatbot systems, which offer a more natural way to fulfill user's fashion needs.