# Supplementary Material for “Compositional Learning of Image-Text Query for Image Retrieval”

June 19, 2020

## 1 Important Notes on Fashion IQ Dataset

In Fashion IQ dataset, $\sim 49\%$ annotations describe the target image directly. While $\sim 32\%$ annotations compares target and source images, e.g. “is red with a cat logo on front” and the second annotation is, “is more pop culture and adolescent”. The dataset consists of three non-overlapping subsets, namely “dress”, “top-tee” and “shirt”. We join the two annotations with the text “ and it” to get a description similar to a normal sentence a user might ask on an E-Com platform. Now the complete text query is: “is red with a cat logo on front and it is more pop culture and adolescent”. Furthermore, we combine the train sets of all three categories to form a bigger training set and train a *single* model on it. Analogously, we also combine the validation sets to form a single validation set.

A challenge was conducted in ICCV 2019 on Fashion IQ dataset [1]. The website also has some technical reports submitted by the best performing teams. The numbers reported in these reports are quite high, even for TIRG approach. We investigated the reasons and reached the conclusion that these technical reports have have quite different settings. It is not possible for us to compare our results with them in a fair manner. The reasons and differences are delineated briefly as:

- They treat Fashion IQ as three independent datasets and train one model for each category (“dress”, “top-tee” and “shirt”). This results in better performance for each category.
- They do pre-training on external datasets like Fashiongen, Fashion200k etc. It is well-known that such transfer learning (via pre-training) inevitably increases the performance of any model.
- They employ product attributes as side information in their models. In our experiments, we do not consider in such side information and rely solely on the image and text query.
- They employ higher capacity models such as ResNet101, ResNet-152 etc. In original TIRG and in all our experiments, we use ResNet17 as image model.
- Since these reports developed models specifically for the competition, they have incorporated several hacks, like ensembeling, data augmentation techniques etc.
- Unfortunately, none of the technical reports have published their code. Thus, we are not able to assess the performance of their model in our experiment setting.

In short, it is neither possible for us to reproduce their results nor are we able to fairly compare the performance of their models in a common experiment setting.

[1] https://sites.google.com/view/lingir/fashion-iq

## 2 Qualitative Results

Fig. 1 presents some qualitative retrieval examples for MIT-States dataset. For the first query, we see that two "burnt bush" images are retrieved. We can observe that other retrieved images share the same semantics and are visually similar to the target images. In second and third row, we note that same objects in different states can look drastically different. This highlights the importance of incorporating the text information in the composed representation.

Some qualitative retrieval results for Fashion200k dataset are presented in Fig. 2. In these results, we observe that the model is able to capture the style and color information quite well. In the first row, we see similar sleeveless dresses with sequin. Similarly, in other two queries, the model successfully images from the same product category, i.e. jacket and skirts. Moreover, the retrieved images seem to follow the desired modifications expressed in the query text remarkably well.

It is pertinent to highlight that the captions under the images are the ground truth. They are not available to the model as additional input during training or inference.

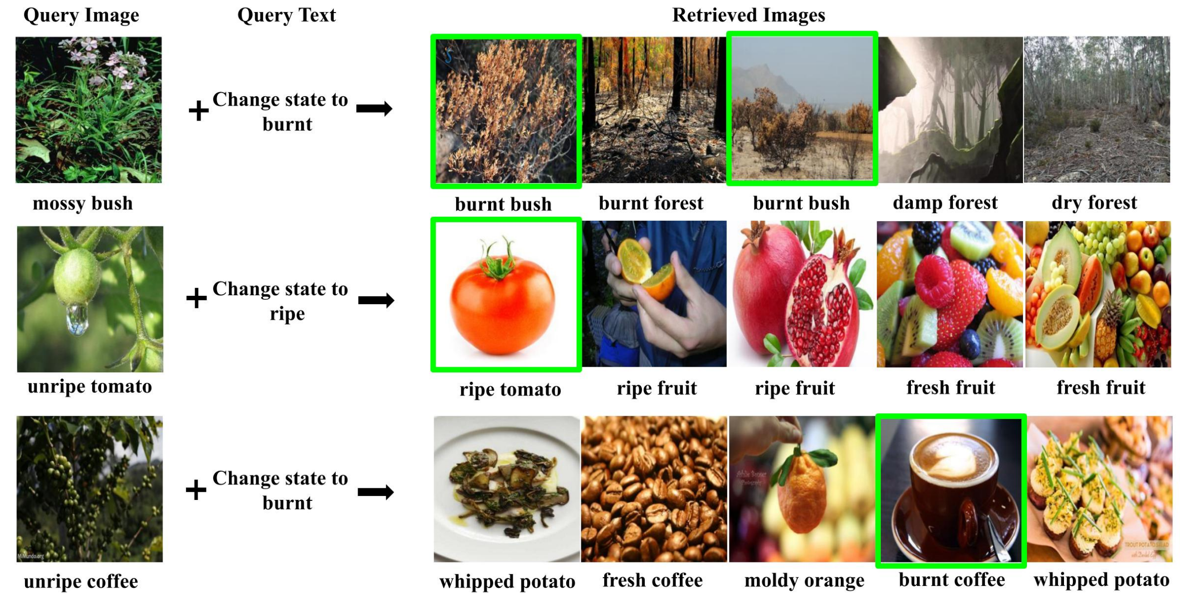

Figure 1: Qualitative Results: Retrieval examples from MIT-States Dataset

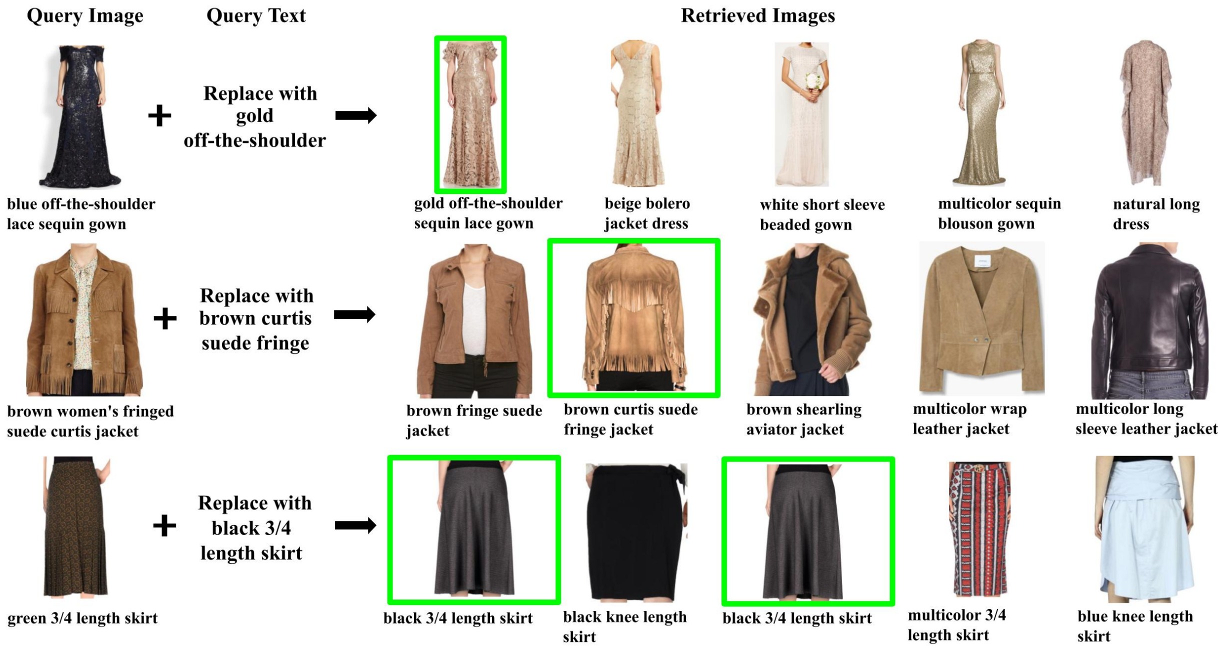
Query Image
Query Text
Retrieved Images
+ Replace with gold off-the-shoulder
blue off-the-shoulder lace sequin gown
gold off-the-shoulder sequin lace gown
beige bolero jacket dress
white short sleeve beaded gown
multicolor sequin blouson gown
natural long dress
+ Replace with brown curtis suede fringe
brown women's fringed suede curtis jacket
brown fringe suede jacket
brown curtis suede fringe jacket
brown shearling aviator jacket
multicolor wrap leather jacket
multicolor long sleeve leather jacket
+ Replace with black 3/4 length skirt
green 3/4 length skirt
black 3/4 length skirt
black knee length skirt
black 3/4 length skirt
multicolor 3/4 length skirt
blue knee length skirt

Figure 2: Qualitative Results: Retrieval examples from Fashion200k Dataset