OutfitTransformer: Learning Outfit Representations for Fashion Recommendation (Supplementary Material)

Rohan Sarkar[1,2], Navaneeth Bodla[2], Mariya I. Vasileva[2], Yen-Liang Lin[2], Anurag Beniwal[2], Alan Lu[2], and Gerard Medioni[2]

[1]Purdue University, West Lafayette, [2]Amazon

1. Qualitative results

In this section, we demonstrate qualitatively that our framework retrieves compatible items using target category or text-based descriptions. For this, we use the query outfits from the FITB test split provided by the Polyvore Outfits dataset [2].

For the complementary item retrieval task based on a target category, we use the dataset proposed by Lin *et al*. [1]. In Fig. 2, we show examples where our method can successfully retrieve the ground truth as one of the top retrieved items. Subsequently, in Fig. 4 we show examples where the the ground truth item is not found in the top retrieved results. We observe that this is mainly because there can be multiple items in the dataset that could be a good match for the partial outfit either because it is either very similar in style, color, etc. to the ground truth item as shown in Fig. 4(a) or it might be dissimilar to the ground truth but can still be compatible with the overall outfit as shown in Fig. 4(b). Therefore, the rank of the ground truth item is not a perfect indicator of the practical utility of the system. This is a limitation of the evaluation metric due to the lack of annotated ground truth as mentioned in [1]. Towards this end, we conduct a user study as described in Section 2 to further validate the quality of the retrieval results and for a better understanding of the usefulness of our framework.

We also use our framework for retrieving complementary items based on a target item description provided as free-form text, as shown in Figs. 3 and 5. Since the dataset does not provide any annotated text-based queries, we show qualitative results that demonstrate that our framework can retrieve items that are *both* compatible and matches the target item description in Fig. 3. In Fig. 5, we show that for a given list of text-based queries and different partial outfits, our system can retrieve complementary items for each specific outfit.

2. User Study

The recall@top-k metrics reported in the paper rely on the relative rank of the ground-truth item in the complementary item retrieval task. While high values on these metrics indicate that our model performs competitively or better than state-of-the-art, as [1] point out, the rank of the ground-truth item is not a perfect indicator of retrieval performance since the database can contain many complementary items to the query outfits – some of which may be judged by human experts to be equally-good or even better stylistic matches, as can be seen in Fig. 4.

We evaluate this hypothesis by running an A/B test using Amazon Mechanical Turk (c.f. Fig. 1 showing the question template used for the experiment). If human experts select our retrieved item over the ground truth item about 50% of the time or higher[1], that would indicate that even in cases where the ground-truth item is not found in the top-k retrieval results, the retrieved items are a good replacement.

For quality control of our annotator pool, we only allow annotators with $> 95\%$ acceptance rate in the past tasks and who have completed at least 1000 tasks. A query outfit from the FITB test split of the Polyvore Outfits dataset is presented, and the annotator is asked to select which of the two items provided better completes the outfit (as shown in Fig. 1). We only present questions where the ground-truth item is not found in the top-three retrieved items. One of the presented items is always the ground-truth, and the other — an item retrieved by our model. The total number of questions is 5120, each of which is annotated by 10 subjects.

Emails: sarkarr@purdue.edu, navaneeth.bodla@getcruise.com, {vamariy, yenliang, beanurag, alalu, medioni}@amazon.com

[1]higher than 50% would indicate that for some partial outfits, our method can find a better match from the database of items belonging to the target category than the original answer provided by the dataset

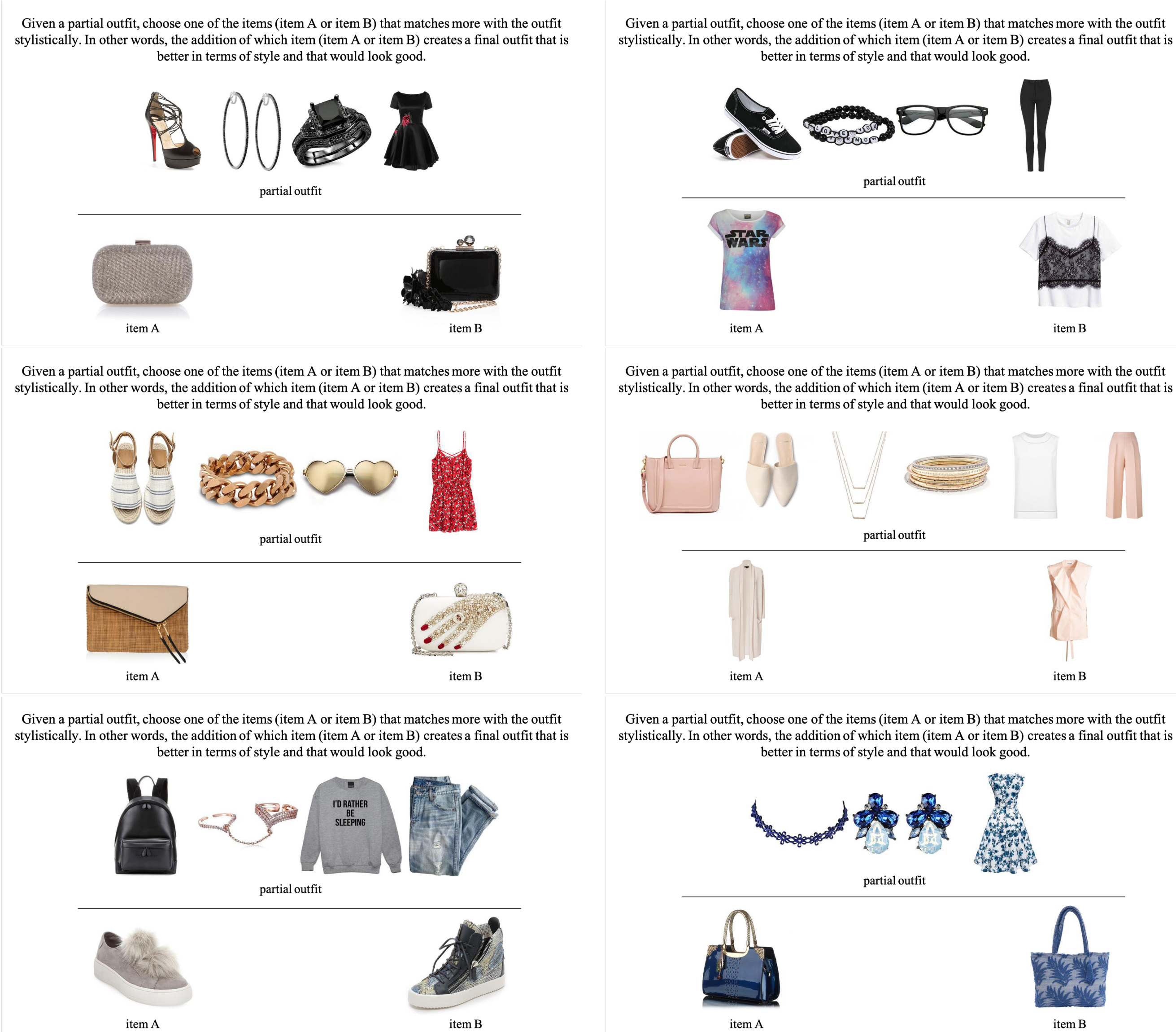


Figure 1. Some examples of the questions presented to the MTurk annotators for the user study.

We compute the average number of times the retrieved item was chosen over the ground truth item. We find that 52.5% of the time annotators select items retrieved by our model over the ground truth. The average rate of 52.5% denotes that our retrieved results were at least as good as the ground truth items. This provides an additional evidence over and above the offline results that our model is able to retrieve compatible items to the query outfit, and they are as good of a match as the ground truth item.

References

[1] Yen-Liang Lin, S. Tran, and Larry Davis. Fashion outfit complementary item retrieval. In *CVPR*, 2020.

[2] Mariya I. Vasileva, Bryan A. Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth. Learning type-aware embeddings for fashion compatibility. In *ICCV*, 2018.

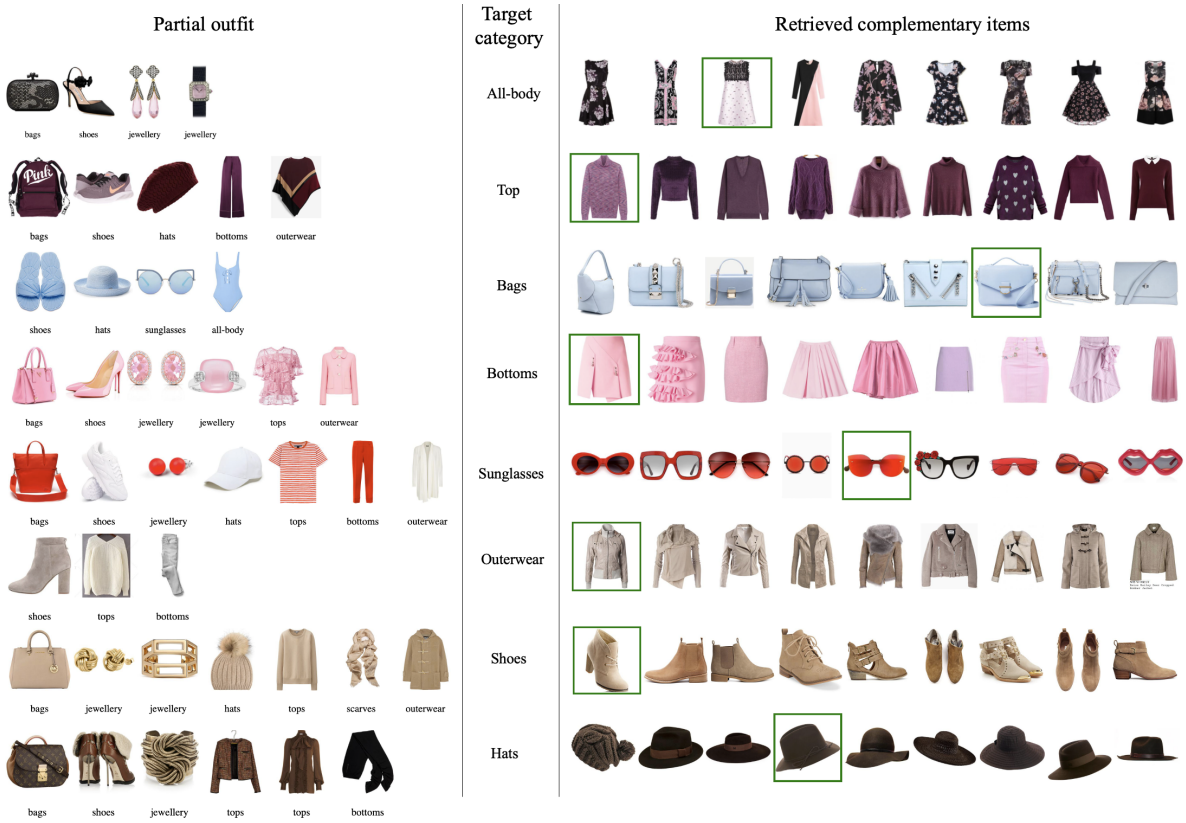
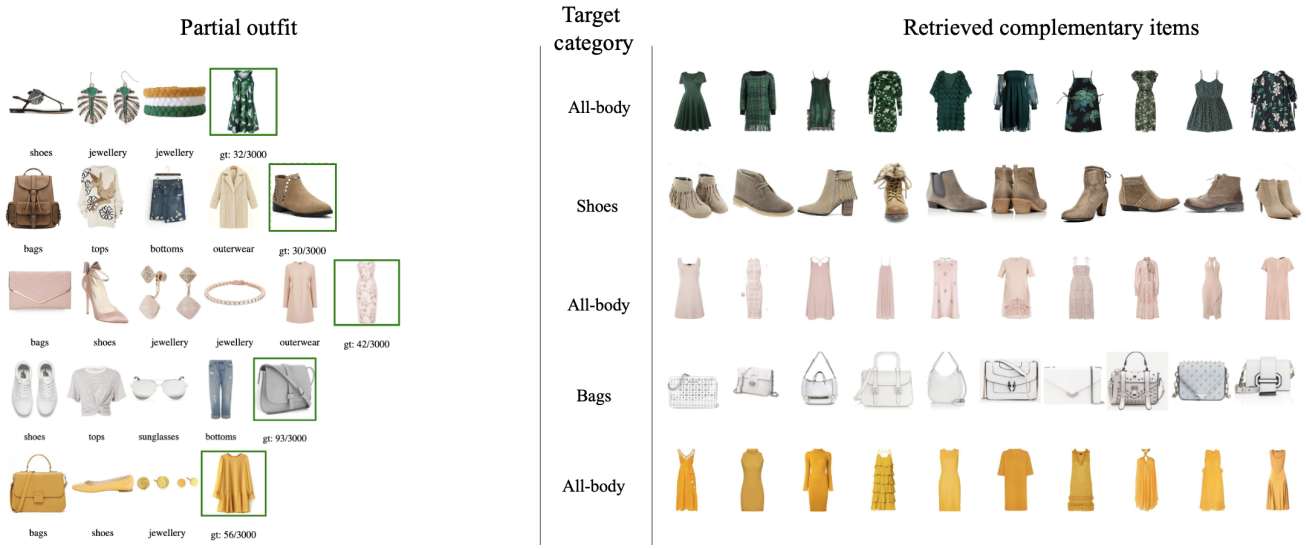


Figure 2. Some example complementary item retrieval results using our framework for different top-level categories. Given a partial outfit and a target category, our method retrieves a list of compatible items that match the global style of the outfit. The ground truth item is indicated by the green bounding box.

(a) Text-based queries

(b) Conversational queries

Figure 3. Some example complementary item retrieval results using our method for different target item descriptions. Given a partial outfit and text-based queries, our approach retrieves a list of compatible items that match both the global style of the outfit and the text description.



(a) Retrieved items are very similar in color, texture, style, pattern, etc. to the ground truth item.

(b) Retrieved items are different in color, style, etc. from the ground truth item but they are also compatible with the partial outfits.

Figure 4. Some failure cases of complementary item retrieval using our method. The partial outfit and the corresponding ground truth item are shown in the left column, the target category in the middle column and the retrieved items are shown in the right column. The ground truth is highlighted using a green bounding box and the rank of the item is also presented below the box.


outfit
jewelery
silver jewelery
gold jewelery
diamond jewelery
gem stone jewelery
outfit
jewelery
silver jewelery
gold jewelery
diamond jewelery
gem stone jewelery
outfit
jewelery
silver jewelery
gold jewelery
diamond jewelery
gem stone jewelery
outfit
jewelery
silver jewelery
gold jewelery
diamond jewelery
gem stone jewelery
outfit
jewelery
silver jewelery
gold jewelery
diamond jewelery
gem stone jewelery
outfit
jewelery
silver jewelery
gold jewelery
diamond jewelery
gem stone jewelery
outfit
jewelery
silver jewelery
gold jewelery
diamond jewelery
gem stone jewelery
outfit
jewelery
silver jewelery
gold jewelery
diamond jewelery
gem stone jewelery
outfit
jewelery
silver jewelery
gold jewelery
diamond jewelery
gem stone jewelery
outfit
jewelery
silver jewelery
gold jewelery
diamond jewelery
gem stone jewelery

Figure 5. This figure illustrates that our method can retrieve items that are compatible with different partial outfits for the same list of text-based queries. The first query for each partial outfit mentions only the top-level category information whereas the subsequent queries are more descriptive which allows us to further refine search results.