“Does it come in black?” CLIP-like models are zero-shot recommenders

Patrick John Chia*
Coveo, Montreal
pchia@coveo.com

Jacopo Tagliabue
Coveo Labs, New York
jtagliabue@coveo.com

Federico Bianchi
Bocconi University, Milan

Ciro Greco
Coveo Labs, New York

Diogo Goncalves
Farfetch, Porto

Abstract

Product discovery is a crucial component for online shopping. However, item-to-item recommendations today do not allow users to explore changes along selected dimensions: given a query item, can a model suggest something similar *but* in a different color? We consider item recommendations of the *comparative* nature (e.g. “something darker”) and show how CLIP-based models can support this use case in a zero-shot manner. Leveraging a large model built for fashion, we introduce `GradREC` and its industry potential, and offer a first rounded assessment of its strength and weaknesses.

1 Introduction

Recommender systems (RSs) are one of the most ubiquitous applications of machine learning (ML) in e-commerce (Tsagkias et al., 2020), recently featuring novel benchmarks and extensive use of deep neural networks in item-to-item, user-to-item, and comparison RSs (de Souza Pereira Moreira et al., 2019; Chia et al., 2021; Tagliabue et al., 2021). While details differ between neural architectures, they all share the principle that products are represented as points in a latent space, learned from user behavior, item meta-data or a combination of both (Bianchi et al., 2021c; Yi et al., 2019). Fig. 1 represents item-to-item recommendations *as movements in the product space*: starting from a query item – the *white dress* –, RSs help shoppers to move either around their current location, or “jump” to a different one. Adding to the blooming literature on substitute, complementary, popularity and exploration-based RSs (Chen et al., 2020; Hao et al., 2020; Ramachandran, 2020; Barraza-Urbina, 2017), *this* work presents `GradREC`, a new type of recommendation that introduces explicit directionality into the mix, by allowing exploration in selected directions *through natural language*: “something darker” will move the user from the *white dress* to the *grey dress*. In particular, we summarize our contributions as follows: **First**, we introduce `GradREC` as a new type of recommendation experience *and* a technical contribution – to the best of our knowledge, `GradREC` is the first *zero-shot* approach for language-based comparative recommendations, showing that CLIP-like (Radford et al., 2021) models may enable recommendations to be generated on the fly *without the need of explicitly defined labels for training* or behavioral data. **Second**, we devise both qualitative and quantitative evaluations to offer a first rounded assessment of the strengths and weaknesses of our proposal, and supplement our analysis with extensive visual examples. **Third**, as part of our submission, we release to the community our fine-tuned weights, publish an interactive web-app for exploration, and open source our code to help reproducing our findings and building on them [1].

While we present our results as a *preliminary* investigation into the untapped capabilities of CLIP for retail, we *do* believe our methods to be interesting to a broad set of practitioners: those exploring recommendations for conversational and interactive commerce, and those leveraging deep learning for horizontally scalable SaaS products[2]. Finally, while motivated by very practical concerns, this work contains new insights on the topology of the information encoded by over-parameterized neural networks, which could help our understanding of the kind of regularities that these models learn about our world.

* GradRECS started as a (failed) experiment by JT; PC actually made it work, and he is the lead researcher on the project. FB, CG and DC all contributed to the paper, providing support for modelling, industry context and domain knowledge. PC and JT are the corresponding authors.

[1] Artifacts are available at https://github.com/patrickjohncyh/gradient-recs.

[2] As a context for this global market, Algolia and Bloomreach both raised more than USD200M in the last two years (Techcrunch, 2021; Bloomreach, 2022), and Coveo raised more than CAD200M with its IPO (Marotta, 2021).

Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5), pages 191 - 198
May 26, 2022

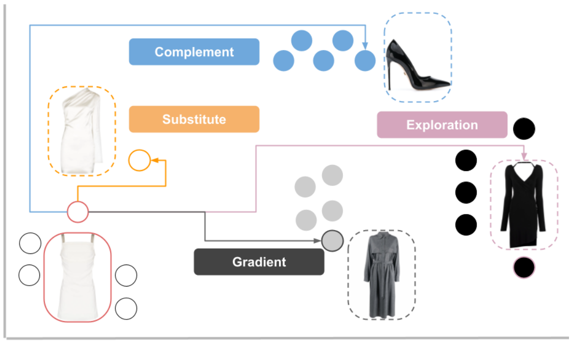

Figure 1: Recommendation as movements in the latent space, starting from a query item (*red*): substitute, complementary and exploration-based strategies are depicted in orange, blue and violet; `GradREC` is in gray.

2 An Industry Perspective

The intersection of product recommendation and natural language is a blooming research area: advances in neural NLP have been recently used for content-based recommendations (Iqbal et al., 2018), cold-start scenarios (Tagliabue et al., 2020), language grounding (Bianchi et al., 2021b,a), and explainable RSs (Chen et al., 2021). A relatively new use case is provided by the growth in the market of interactive technologies, as intelligent virtual assistants (IVAs) are expected to handle recommendations that increasingly encompass the expressiveness of natural language (Jannach et al., 2021). While interaction is an opportunity, the limited real estate available to display recommendations is a constraint for IVAs (Lin et al., 2021): since scrolling is limited, strategies for moving from one product to another (as in Fig. 1) are crucial for IVAs market penetration. In *this* work, we consider recommendations which are of the *comparative* form: given an item of focus – in a chat, a product page, etc. –, the shopper makes use of natural language queries to retrieve a second item (e.g. "shorter", "darker", etc.), related to the first but different along the specified attribute. While state-of-the-art IVAs can already provide very simple recommendations through language (Amazon, 2022), we are the first to suggest the existence of an entire new dimension and depth to mimic the interactions typical of a real-life shopping experience.

When thinking about applying this method in a multi-tenant SaaS context, it is worth noting how small are the assumptions `GradREC` actually makes about the underlying inventory: while in the case of `FashionCLIP` and its dataset it is true that products often contain information about an attribute's intensity (e.g. "knee-length shorts"), the *relationship* between them is not explicitly encoded, yet it is inferred by `GradREC`. Moreover, when applying these models across new catalogs, there is no guarantee descriptions would be as rich, or even using the same lexicon to describe the same attribute ("bermudas" vs "knee-length"). These considerations further highlight the strength of using a latent space derived from a general and flexible multi-modal model, and the non-trivial nature of extracting comparative recommendations.

3 Related Work

Our work sits at the intersection of various recent technical advances in *latent space manipulation* and *iterative IR*. Many recent works explore latent space manipulation of Generative Adversarial Networks (GANs) for purposes of fine-grained image editing (Shen et al., 2020; Patashnik et al., 2021); Jahanian et al. (2020) also studied latent space traversal in GANs to measure GAN generalization. We extend this line of research by providing a clear e-commerce use case, a focus shift from generative modeling to recommendation, and new insights on CLIP-based representations.

The idea of iterative search refinement using comparative information and attribute ranking is not new (Kovashka et al., 2012; Yu and Grauman, 2015). However, previous work sit in the standard *fully supervised* "learning-to-rank" tradition. Conversely, our approach operates in a zero-shot fashion by using *both* CLIP retrieval and CLIP representations to generate suggestions on-the-fly. Finally, our work builds on top of the recent wave of contrastive-based methods for representational learning: while latent product representations have been extensively studied from multiple angles (Bianchi et al., 2020; Xu et al., 2020), CLIP-like models are still very new in this domain: *GradREC* leverages the space learned by `FashionCLIP`, a fashion-fine tuning of the original CLIP (Chia et al., 2022).

4 Gradient Recs

4.1 Overview

`GradREC`, builds upon the multi-modal space induced by `FashionCLIP`. `GradREC` aims to traverse the latent space such that the intensity of an attribute of interest varies monotonically for products along that path, allowing us to make fine-grained recommendations that require comparative knowledge. There are independently grounded reasons to expect this method to work. *First*, we have solid evidence that embedding spaces are able to encode recognizable "concepts" (e.g. lexical knowledge in *word2vec* (Mikolov et al., 2013), facial expressions in GANs (Ding et al., 2018)). *Second*, we perform an extensive evaluation of the `FashionCLIP` product space, focusing on attributes such as *color* and *occasion*: our qualitative assessment (Section 4.2) verified that embeddings are indeed often clustered, further suggesting that movements in the "concept space" can be represented as paths in the latent space.

4.2 FashionCLIP exploration

As discussed in Section 4.1, there are pre-existing theoretical reasons to think that embedding spaces encode in their geometry interesting regularities. In order to validate this hypothesis, we run visual investigations on `FashionCLIP` space as seen in Figure 2, which shows TSNE projections of product image embeddings for four attributes: Pants Length, Shirt Color, Heel Height and Occasion. For each attribute, we retrieve products possessing negative, neutral and positive attribute intensities. Figure 2 demonstrates that the projected products from the corresponding attribute intensities do indeed form meaningful clusters, suggesting that it is possible to trace a path from one cluster to another in the latent space.

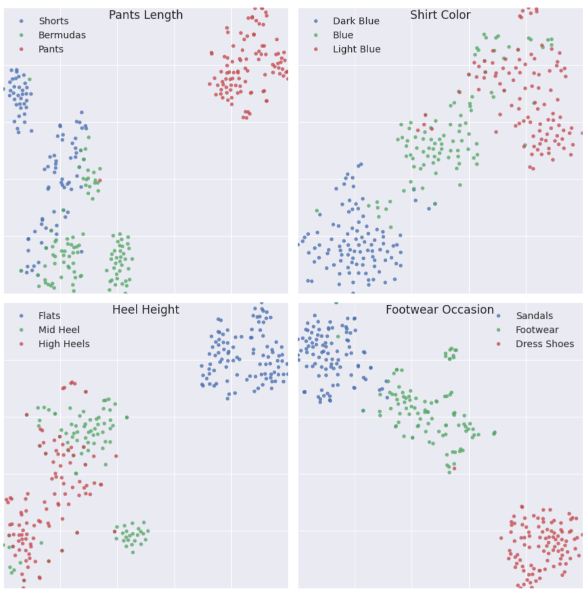

Figure 2: Sample TSNE projections of product image vectors: products are colored based on attribute strength.

4.3 Method

In what follows, we focus on the core task of gradient recommendations[3]. Assuming a target inventory of fashion products, a starting item and a pair of natural language queries whose difference captures the comparative dimension of interest (e.g. the difference between "dark red shirt" and "red shirt" captures the dimension of "darker"[4]), `GradREC` should return a new item in the "same style" as the starting item, varying along the specified dimension; in particular, `GradREC` can leverage CLIP representations but has no access to labels or co-purchasing data. Providing now a formal description, we decompose our approach into two components: a *traversal function*, $\mathbf{\Phi}$ and a *traversal direction vector*, $\mathbf{v}_c$.

Traversal Function: given a product t, represented in the CLIP space by either its L2 normalized textual vector $\mathbf{t}_t$ or image vector $\mathbf{i}_t$, and some attribute c we want to explore, our goal is to compute a function $\mathbf{\Phi}$, such that given a starting point $\mathbf{v}_t$ and some vector $\mathbf{v}_c$, returns a new point $\mathbf{v}_{t+1}$ in the latent space that is increasing or decreasing in strength of attribute c. Given the new position $\mathbf{v}_{t+1}$, we use cosine-based k-nearest neighbors ($KNN(\cdot,\cdot)$) to retrieve suggested products: if we iterate this process, we would travel along the dimension of attribute c, discovering products as we move along. We define $\mathbf{\Phi}$ as vector addition, with a scale factor λ to control step size; additionally, we use the mean of the current point's nearest neighbours ($K\bar{N}N(\mathbf{v}_t, k)$) as a regularizing term. The two terms are balanced by taking a convex combination of the direction vector and the regularizing term. In our notation, $\hat{\mathbf{v}}$ refers to $\mathbf{v}$ normalized to unit length. Note that all vectors are of dimension 512. Our definition is summarized in Eq.1:

$$
\begin{aligned}
\mathbf{v}_{t+1} &= \mathbf{\Phi}(\mathbf{v}_t, \mathbf{v}_c) \\
&= \mathbf{v}_t + \ (1-\rho)\cdot\lambda\hat{\mathbf{v}}_c \\
&\quad + \rho\cdot K\bar{N}N(\mathbf{v}_t, k)
\end{aligned}
\tag{1}
$$

Traversal Vector: the construction of $\mathbf{v}_c$ relies on two main ingredients. First, given a pair of

[3]We realize that a more ecological setting – such as IVA – would require additional steps to handle stateful interactions: those steps are however general open problems in IVA, whose solution is independent of the interaction we model here.

[4]Different ecological settings may provide these queries more or less explicitly; `GradREC` may be used naturally in the context of multi-turn systems such as IVAs, or, for example, as support to standard manually defined facets for IR use cases, such as product search.

Figure 3: A sample of the qualitative results obtained by applying GradREC for four different attributes: the *intensity* / *strength* of the attribute decreases from left to right.

queries which semantically captures the attribute c ("darker"), we use the zero-shot retrieval capabilities of FashionCLIP to construct two small datasets: one comprising the image embeddings closest to the FashionCLIP encoding of the neutral class ("a blue shirt"), and one from an exemplar class for c ("a dark blue shirt"). We define the retrieved image vectors for the neutral and exemplar prompts as $\mathbf{I}_n = \{\mathbf{i}_n^1 ... \mathbf{i}_n^M\}$ and $\mathbf{I}_e = \{\mathbf{i}_e^1 ... \mathbf{i}_e^N\}$ respectively. Second, we adopt the channel importance measure from Wu et al. (2021) to determine channels[5] which encode the differences between the neutral class and exemplar class. The method measures the channel-wise Signal-to-Noise Ratio (SNR) between the mean neutral class vector (i.e. $\bar{\mathbf{I}}_n$) and the exemplar class vectors (i.e. $\{\mathbf{i}_e^1 ... \mathbf{i}_e^N\}$). The intuition is that channels with high SNR correspond to channels which encode the differences between images from the neutral and exemplar class, and hence the attribute c. Our implementation departs from theirs by retaining the sign of the differences for each channel. Finally, to obtain $\mathbf{v}_c$, we normalize the vector formed by the channel-wise SNR values[6].

5 Experiments

To investigate GradREC strengths and weakness, we offer a preliminary assessment of its capabilities over important fashion dimensions, such as product discovery.

5.1 Dataset and Pre-trained Space

Our pre-trained space is FashionCLIP, an adaptation of CLIP obtained by fine-tuning the original embeddings over fashion products provided by *Farfetch*, a world leading platform for online luxury fashion shopping. The dataset comprises of over 800k fashion products across dozens of item types and more than 3k brands. In addition to a standard product image over white background, the dataset contains natural language descriptions of the stylistic properties (e.g., "cotton-blend", "high waist", "belt loops") and categorical information (e.g. "layered track shorts") of products[7]. FashionCLIP shares the same architecture as Radford et al. (2021), i.e. a multi-modal model comprising an image and a text encoder. We refer to Chia et al. (2022) for details on training and retrieval / classification capabilities: since FashionCLIP has independent value in the industry, GradREC does not require any *specific* pre-training.

5.2 Qualitative Analysis

We consider four different attributes of interest: *shirt color luminance*, *heel height*, *trouser length* and *trouser cutting*. For each attribute, we traverse the latent space between both extremes of the attribute of interest and present the results in Fig. 3 for visual validation. We observe that the products retrieved form a monotonic change in the attribute's strength that aligns well with human intuition: i.e. t-shirts in the first row do indeed follow a gradient going from lighter to darker shades of blue. It is interesting to note that the latent space of FashionCLIP appears to encode and organize these geometric and physical regularities despite not having been trained to do so explicitly, pointing to further questions about what and how these self-supervised models learn.

5.3 Quantitative Analysis

We quantitatively assess GradREC by measuring its efficiency in product discovery along an attribute of interest, to verify that the path it discovers is semantically meaningful. We generate three datasets ($N = 100$) using FashionCLIP retrieval capabilities to represent products from the negative, neu-

[5]Each channel corresponds to one of the 512 dimensions of an embedding.

[6]We refer the reader to Wu et al. (2021) for the original discussion.

[7]FashionCLIP weights and training code will be released with the original publication. At the moment of writing this paper, the original training dataset is scheduled to be released as well: please check https://github.com/Farfetch for updates.

tral and positive intensities of an attribute[8]. For example, to generate datasets for shirt color *luminance* we would issue the following queries – “dark blue polo shirt”, “blue polo shirt”, “light blue polo shirt” – to `FashionCLIP` and retrieve N products for each query. In Figure 4 we visualize a sample of the products retrieved for each of the above queries.

Figure 4: Products retrieved for queries on a spectrum of intensity.

We apply `GradREC`, starting a traversal from a negative product in the direction of neutral intensity products, simulating product discovery by logging the top $k = 10$ unseen products found at each step. We then compute the intersection cardinality of the three datasets along the simulated trajectory with a sliding window of 50 products: a model which traverses a meaningful path should produce three peaks, one for each level of intensity. As a baseline, we use visual similarity in the CLIP image-space (KNN over image embeddings) and simulate the product discovery trajectory as traversing the list of nearest neighbors of the same seed product in the order of increasing distance.

In Figure 5 we see the result of applying our analysis to the discovery path for *luminance* of blue polo shirts. We observe that `GradREC` explores well this path as seen by its three distinct modes of intersected products, where each peak for *light blue*, *blue* and *dark blue* respectively, corresponds to the correct *order* of decreasing luminance. Conversely, we see that visual similarity fails to produce a similar product discovery pattern as `GradREC`, which spans a wider range of the luminance spectrum. In fact, visual similarity struggles to discover products from *blue* and *light blue*, highlighting the merits of the directionality induced by `GradREC` (Appendix A.2).

[8]The exact definitions of negative and positive are relative; We are more concerned here with capturing the opposing extremes of an attribute’s spectrum i.e. *dark* and *light*.

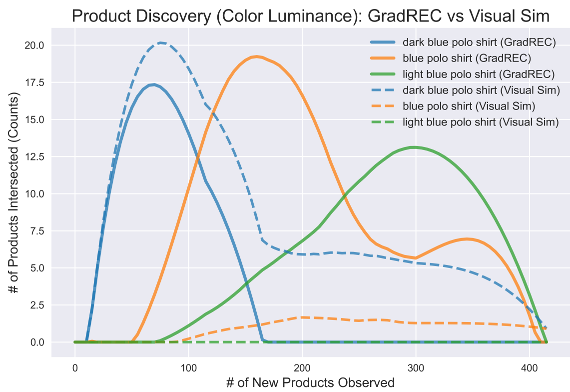

Figure 5: Quantitative analysis comparing `GradREC` to Visual Similiarty on product discovery for the attribute *color luminance* of blue shirts.

5.4 Limitations & Future Work

While `GradREC` performances are encouraging – especially when considering that no attribute has been explicitly taught –, limitations highlight several areas of improvement. First, the performance of the model is sensitive to the quality of the retrieval phase. For example, to construct $\mathbf{v}_c$ for *trouser length*, using queries “shorts” and “pants” yielded better performance than “shorts” and “bermudas”. Second, our definition of $\boldsymbol{\Phi}$ is not optimal: as we traverse the image space, the cosine distance between our position and all the products increases, suggesting that we are not traversing the latent manifold in the most efficient way. Third, while we observe that `GradREC` moves in a semantically meaningful direction, it does not, nor is it currently designed to, provide guarantees on the monotonicity of the products it returns along the path. Finally, `GradREC` does not account for uncertainty and thus does not possess a confidence measure for its recommendations: while we may be confident in its ability with geometric and physical concepts, and less so for more abstract notions (e.g. “for colder weather”), it is hard to know *a priori* what `GradREC` does not know.

6 Conclusion

We introduced `GradREC`, a zero-shot approach for comparative recommendations, that showed promising results in our initial investigations. While further evaluation – especially, involving relevance judgments by humans – is needed to fully assess `GradREC` capabilities, we *do* believe that our work provides preliminary but novel insights into innovative application of large models in important industry use-cases.

References

Amazon. 2022. Style by Alexa.

Andrea Barraza-Urbina. 2017. The exploration-exploitation trade-off in interactive recommender systems. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, RecSys '17, page 431–435, New York, NY, USA. Association for Computing Machinery.

Federico Bianchi, Ciro Greco, and Jacopo Tagliabue. 2021a. Language in a (search) box: Grounding language learning in real-world human-machine interaction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4409–4415, Online. Association for Computational Linguistics.

Federico Bianchi, J. Tagliabue, Bingqing Yu, Luca Bigon, and Ciro Greco. 2020. Fantastic embeddings and how to align them: Zero-shot inference in a multi-shop scenario. *ArXiv*, abs/2007.14906.

Federico Bianchi, Jacopo Tagliabue, and Bingqing Yu. 2021b. Query2Prod2Vec: Grounded word embeddings for eCommerce. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 154–162, Online. Association for Computational Linguistics.

Federico Bianchi, Bingqing Yu, and Jacopo Tagliabue. 2021c. BERT goes shopping: Comparing distributional models for product representations. In *Proceedings of The 4th Workshop on e-Commerce and NLP*, pages 1–12, Online. Association for Computational Linguistics.

Bloomreach. 2022. With $175 million in funding, bloomreach is authoring the next chapter of e-commerce.

Hanxiong Chen, Xu Chen, Shaoyun Shi, and Yongfeng Zhang. 2021. Generate natural language explanations for recommendation. *CoRR*, abs/2101.03392.

Tong Chen, Hongzhi Yin, Guanhua Ye, Zi Huang, Yang Wang, and Meng Wang. 2020. *Try This Instead: Personalized and Interpretable Substitute Recommendation*, page 891–900. Association for Computing Machinery, New York, NY, USA.

Patrick John Chia, Giuseppe Attanasio, Federico Bianchi, Silvia Terragni, Ana Rita Magalhães, Diogo Goncalves, Ciro Greco, and Jacopo Tagliabue. 2022. Fashionclip: Connecting language and images for product representations.

Patrick John Chia, Bingqin Yu, and Jacopo Tagliabue. 2021. Are you sure?: Preliminary insights from scaling product comparisons to multiple shops. In *SIGIR eCom 2021*.

Gabriel de Souza Pereira Moreira, D. Jannach, and Adilson Marques da Cunha. 2019. On the importance of news content representation in hybrid neural session-based recommender systems. In *INRA@RecSys*.

Hui Ding, Kumar Sricharan, and Rama Chellappa. 2018. Exprgan: Facial expression editing with controllable expression intensity. *AAAI*.

Junheng Hao, Tong Zhao, Jin Li, Xin Luna Dong, Christos Faloutsos, Yizhou Sun, and Wei Wang. 2020. P-companion: A principled framework for diversified complementary product recommendation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 2517–2524, New York, NY, USA. Association for Computing Machinery.

Murium Iqbal, Adair Kovac, and Kamelia Aryafar. 2018. A multimodal recommender system for large-scale assortment generation in e-commerce. In *The SIGIR 2018 Workshop On eCommerce co-located with the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2018), Ann Arbor, Michigan, USA, July 12, 2018*, volume 2319 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Ali Jahanian, Lucy Chai, and Phillip Isola. 2020. On the "steerability" of generative adversarial networks. In *International Conference on Learning Representations*.

Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A survey on conversational recommender systems. *ACM Comput. Surv.*, 54(5).

Adriana Kovashka, Devi Parikh, and Kristen Grauman. 2012. Whittlesearch: Image Search with Relative Attribute Feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ying Lin, Han Wang, Jiangning Chen, Tong Wang, Yue Liu, Heng Ji, Yang Liu, and Premkumar Natarajan. 2021. Personalized entity resolution with dynamic heterogeneous KnowledgeGraph representations. In *Proceedings of The 4th Workshop on e-Commerce and NLP*, pages 38–48, Online. Association for Computational Linguistics.

Stefanie Marotta. 2021. Canada's latest tech public debut swings amid soft ipos.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Lakshmi Ramachandran. 2020. *Behavior-Based Popularity Ranking on Amazon Video*, page 564–565. Association for Computing Machinery, New York, NY, USA.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. 2020. Interpreting the latent space of gans for semantic face editing. In *CVPR*.

Jacopo Tagliabue, Ciro Greco, Jean-Francis Roy, Federico Bianchi, Giovanni Cassani, Bingqing Yu, and Patrick John Chia. 2021. Sigir 2021 e-commerce workshop data challenge. In *SIGIR eCom 2021*.

Jacopo Tagliabue, Bingqing Yu, and Federico Bianchi. 2020. *The Embeddings That Came in From the Cold: Improving Vectors for New and Rare Products with Content-Based Inference*, page 577–578. Association for Computing Machinery, New York, NY, USA.

Techcrunch. 2021. Search api startup algolia raises $150 million at $2.25 billion valuation.

Manos Tsagkias, Tracy Holloway King, Surya Kallumadi, Vanessa Murdock, and Maarten de Rijke. 2020. Challenges and research opportunities in ecommerce search and recommendations. In *SIGIR Forum*, volume 54.

Zongze Wu, Dani Lischinski, and Eli Shechtman. 2021. Stylespace analysis: Disentangled controls for stylegan image generation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12863–12872. Computer Vision Foundation / IEEE.

Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2020. *Product Knowledge Graph Embedding for E-Commerce*, page 672–680. Association for Computing Machinery, New York, NY, USA.

Xinyang Yi, Ji Yang, Lichan Hong, Derek Zhiyuan Cheng, Lukasz Heldt, Aditee Ajit Kumthekar, Zhe Zhao, Li Wei, and Ed Chi, editors. 2019. *Sampling-Bias-Corrected Neural Modeling for Large Corpus Item Recommendations*.

A. Yu and K. Grauman. 2015. Just noticeable differences in visual attributes. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2416–2424, Los Alamitos, CA, USA. IEEE Computer Society.

A Appendix

A.1 A worked-out traversal example

Fig. 6 showcases an example of successful traversal when using the methods in Section 4.3 applied to skirt length. In particular, we can see the query item, an intermediate one and a final one, along the path of products that was traced (the path of products are denoted by $\times$ markers, and the order of observation/direction of traversal is denoted by the darker to lighter hue). Note that a simple visual similarity search would have moved us from the first query item to a nearby region.

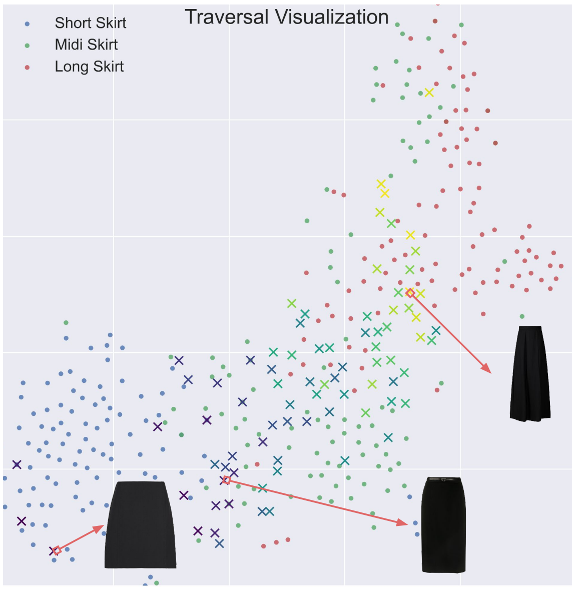

Figure 6: TSNE Projection of 3 ranges of skirt lengths and the traversed product path by `GradREC` along the attribute *skirt length* as shown by $\times$ markers (dark to light denotes direction of traversal). Corresponding product images along the traversed path are visualized.

A.2 Additional Experiments

We ran our product discovery analysis for the attribute *heel height* and report the result in Fig. 7.

A similar pattern as Fig. 5 emerges with `GradREC` having three peaks and Visual Similarity struggling to discover "high heel". Unlike Fig. 5, however, we observe a lower cardinality of intersection for `GradREC` and "women's high heels", since `GradREC` preserves the style of the seed product (red colored shoes, in this example) while the products retrieved by "women's high heels" are of varying color.

We also provide additional qualitative examples in Fig. 8. We observe `GradREC` working across colors, in different product sortals (e.g., Dress), and having the ability to preserve visual style (e.g., Denim).

In Fig. 9, we instead give an example of the limitations of `GradREC`. Indeed, we see a failure mode where `GradREC`, while increasing the strength of

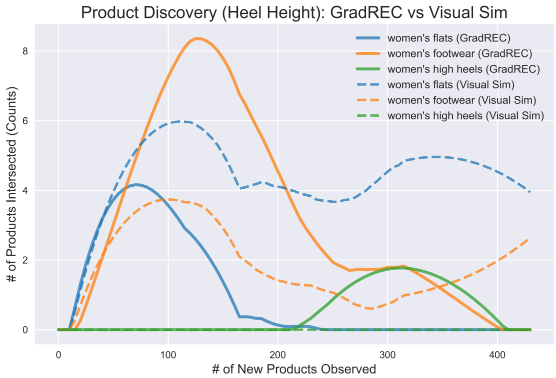

Figure 7: Product discovery analysis for *heel height*.

Figure 8: Extra qualitative examples.

formality correctly, is however unable to preserve the visual style of the footwear correctly. As we have highlighted in Section 5.4, `GradREC` performance is sensitive to the initial dataset retrieval performance: in this instance, the query "formal shoes" retrieves predominantly black, leather dress-shoes, thereby steering the traversal in that direction.

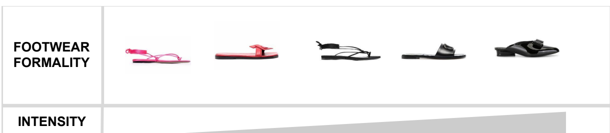

Figure 9: Failure mode of `GradREC` for *formality* attribute. While *formality* is appropriately increased, the product changes visual appearance from pink to black.