Décor: Learning to Interior Design from IKEA Rooms

David Chuan-En Lin [*1] **Haocheng Han** [*1] **Yuanxin Wang** [*1] **Venkatesh Sivaraman** [*1]

1. Introduction

Deep learning algorithms have achieved state-of-the art ac curacies for a variety of tasks, including object detection and classification. These algorithms work by learning semantic representations of image content, often utilizing auxiliary multimodal cues to improve granularity and accuracy. While such representations typically encode similarities related to the visual appearance of objects, many tasks such as stylistic compatibility require more abstract conceptualizations.

The capability of learning the stylistic compatibility of objects has applications in a variety of creative disciplines, including interior design, fashion, and architecture. While image similarity has been well-studied (Datta et al., 2008), stylistic compatibility, however, remains a challenging task due to its loose definition and inherent subjectivity. Prior works have largely explored stylistic compatibility by learning from a dataset of objects manually tagged with style categories (e.g. industrial, rustic, modern) (Kiapour et al., 2014; Takagi et al., 2017; Aggarwal et al., 2018). However, since these style categories are inherently vague and subjective, learning from objects labeled with style tags can be a noisy task.

2. Experimental Setup

In this work, we investigate the task of learning style compatibility in the context of IKEA furniture. We leverage the IKEA dataset (Tautkute et al., 2017), which contains 2,193 product photos and text descriptions. Most interestingly, the dataset also includes 298 rooms, curated by IKEA designers, as well as information on which products appear within which rooms. This is valuable information. Our insight is that to create our style compatibility dataset, we can define two products as compatible if they co-occur within the same room. Our positive labels are therefore determined by professional designers in the context of rooms – realistic and free from subjective style vocabulary.

[*]Equal contribution [1]Carnegie Mellon University, Pittsburgh, USA. Correspondence to: Louis-Philippe Morency <morency@cs.cmu.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

However, this formulation also introduces several challenges. First, for two products that do not have co-occurrence, we can only label them as unknown. This results in a relatively small number of positive labels and a large number of unknown labels. Second, our dataset contains relatively few samples overall. Third, textual product descriptions often do not directly describe the products' appearance. We describe possible solutions in Section 4. For evaluation, we plan to use evaluation metrics described in (Tautkute et al., 2019), including the Hit@k metric for search result relevance and the co-occurrence-based style similarity metric for furniture compatibility. Our future code development will be on `https://github.com/MichaelYxWang/DecorAssistant`.

3. Related Work

3.1. Computational Representations of Style

The computational representation of style has been a problem far from being solved for a long time, as there isn't a clear metric defining what's the definition of style. (Blijlevens et al., 2009) researched how consumers perceive the product's appearance or style. They found that three main attributes mainly affected people: modernity, simplicity, and playfulness. Although they define dozens of styles based on the value of these three attributes, we expect a numerical representation without explicit semantic information.

(Kiapour et al., 2014) realized that to generate a computational representation, a dataset of style and corresponding machine learning training method is required. They proposed a dataset called Hipster to classify style into five categories: hipster, bohemian, pinup, preppy, and goth. They collect the data from a crowd-sourcing game and finally get 1,893 images with annotation. They identify the style descriptor by accumulating features like RGB value, MR8 texture response, distance from image border, etc. They also tried to analyze the style indicator for individuals. However, since there are issues with obtaining high-quality ground truth, the problem remains.

Another problem with the Hipster dataset is that it's relying on user annotations for a small number of very dissimilar classes. To solve this problem, (Takagi et al., 2017) raised a new dataset for evaluation called FashionStyle14. It's

focusing on more complicated classes with large variability and expert-curated annotations. The FashionStyle14 dataset has collected 13,126 images of 14 styles. They then tried several image models to do the classification task. The result shows that ResNet50 gets the best score, but still far from the performance of savvy users.

(Han et al., 2017) realized that both images and semantic descriptions are critical to style. They raised a new dataset called Fashion200K, which contains over 200,000 images with their semantic descriptions. Trying to get a visual-semantic embedding, the model aims at minimizing the distance between features after CNN and bag-of-word embedding of semantic descriptions. Their result shows that a visual-semantic representation can cluster attributes into multiple groups to form spatially-aware concepts.

3.2. Multimodal Image Retrieval

Our work is situated among literature in multimodal image retrieval (Datta et al., 2008), which has been extensively studied over decades. Specifically, we build on recent works which have explored a combination of state-of-the-art machine learning methods and large-scale image datasets available on the web to train powerful models with good generalization capability. One example is CLIP (Radford et al., 2021), a model that has learned a mapping between images and text from a dataset of 400 million image-text pairs through contrastive training. The model has demonstrated impressive zero-shot capabilities for many downstream tasks. In our work, we plan to build on pre-trained CLIP encoders to form our image and text representations. In addition, it may be interesting to use CLIP as a baseline model. Since CLIP was not trained specifically on IKEA furniture, our model should ideally outperform it.

3.3. Image Retrieval in Other Domains

Image retrieval with stylistic similarity can be well illustrated in the domain of fashion products. There are various prior works focusing on learning fashion characteristics and representations. (Vaccaro et al., 2016) predicts high-level fashion style attributes such as tropical, exotic, effortless, radiant, and flowy from low-level design element languages such as color, material, and silhouette using polylingual topic modeling. (Liao et al., 2018) design an end-to-end explainable image and text representation learning pipeline guided by a Exclusiveness-Independence Tree (EI-Tree) data structure that captures the multi-class and multi-label hierarchical relationships of fashion concepts. (Guo et al., 2021) proposes a new fashion dataset and a interactive user simulator that seamlessly integrate explicit visual attributes and conversational user feedback for image retrievers.

Image retrieval applications in other domains can also be inspiring to our projects. As mentioned in (Sharma et al., 2019), users get floor plan recommendations by either existing printed floor plan figures or manual sketch figures. All these image retrieval applications in other domains present us with inspirations in neural network architectures as well as evaluation metrics.

3.4. Interior Design Recommendation

Although the obvious application of furniture recommendation is for everyday home decorators, the initial forays into the problem of automatic style-compatible furniture recommendation were built for 3D modelers. For example, (Liu et al., 2015) developed an interactive room-building tool for scene designers, augmented with style-compatible furniture suggestions based on 3D object models. They collected style compatibility data from Mechanical Turk workers, then trained an embedding to convert handcrafted features of the 3D models (such as the curvature of a chair arm) to a style feature space. Interestingly, they observed that only a small minority of compatibility pairs (although a larger fraction than random) exhibited strong agreement among crowd workers, indicating that most furniture item pairs may occupy a gray area of compatibility when labeled by non-experts. More recently, (Weiss et al., 2020) developed a similar recommendation-augmented 3D modeling system, this time using Siamese networks trained on images labeled by interior-design experts. They quantified style according to four predominant design trends (modern, traditional, cottage, and coastal), which improved interpretability but may have constrained the expressiveness of the embeddings learned.

Meanwhile, early efforts in furniture recommendation from the machine-learning community focused on capturing notions of similarity through large amounts of image data. A foundational effort in this vein was put forward by (Bell & Bala, 2015), which trained furniture image embeddings using a Siamese CNN optimized with a contrastive loss. They scraped a dataset of around 14 million product and room photos from `houzz.com`, recruited MTurk workers to label the bounding boxes of featured products in each room, then trained the network to identify whether two images represented the same piece of furniture. Their results clearly benefit from the volume of available data; however, their model does not capture stylistic similarity across different types of furniture.

The approaches described above either require *a priori* style knowledge, or replace style with a more readily-learned concept such as visual similarity; to overcome these drawbacks, a small number of researchers have looked to multimodal information sources. Most closely related to our work, (Tautkute et al., 2017) propose a multimodal search engine where users may input a scene and a text query to retrieve fitting furniture products. Their approach learns two

separate embeddings of furniture, one based on visual features extracted using a CNN and one based on textual product descriptions. They propose a simple "blending" technique to re-rank textual and visual results based on visual features, which leads to an 11% increase in co-occurrence probability of the retrieved results. In contrast, our work specifically learns for the stylistic compatibility between furniture products.

3.5. Work Closely Related to Our Proposal

Among all related work, except for CLIP (Radford et al., 2021) and the original paper for the IKEA dataset (Tautkute et al., 2017), which we have already described in detail, there are two other papers that are closely related to our proposed work in terms of data, method, and evaluation metrics. As an extension to their previous work on the IKEA dataset, which is mainly based on late fusion of candidates from different modalities, in (Tautkute et al., 2019) the authors propose a new DeepStyle Siamese network with early fusion scheme and contrastive loss to perform better representation learning and ranking. (Aggarwal et al., 2018) also learns style compatibility of furniture with Siamese networks, using the Bonn Furniture Styles dataset containing furniture tagged with 17 style categories. We differentiate our work by making use of IKEA rooms designed by professional interior designers as our ground truths for the task of furniture style compatibility, which requires new machine learning methods.

4. Research Ideas

We plan to approach the problem through several stages. First, we will reproduce and evaluate (Tautkute et al., 2017)'s results on their original task, multimodal product retrieval. This can allow us to gain a sense of the dataset and have our basic data preprocessing operations (e.g. dataloaders) ready. Second, we will process the IKEA dataset to form our furniture style compatibility dataset as follows: For all products in a room, we create pairs of products for all possible combinations. The pairs are labeled with 1 (compatible) and we perform this for all rooms. For every remaining pairing combination in the dataset (i.e. products which do not co-occur in the same room), we label them with 0 (unknown).

Third, we will build our learning pipeline. Since we have a dataset of relatively little positive labels and large quantities of unknown labels, we plan to adopt a PU learning (positive unlabeled learning) approach (Bekker & Davis, 2020). Specifically, we perform PU bagging (Mordelet & Vert, 2014) by creating an ensemble of weak classifiers in parallel. Each classifier takes in all positives (1) and randomly samples subsets of unknowns (0) to create a balanced training set. Our weak classifiers are Siamese networks (Koch et al., 2015) with multimodal modifications. Prior work has demonstrated that Siamese networks are able to learn the concept of style with few examples. Each Siamese network contains two identical sub-networks with shared weights and takes in a pair of products. Each sub-network has two branches, one taking in an image and one taking in text, which are a product's image and description, respectively. We perform image feature extraction using a vision model (e.g. CNN) and text feature extraction using a language model (e.g. LSTM). It may also be interesting to explore building on pretrained CLIP image and text encoders (Radford et al., 2021). The image and text features are than concatenated and fed into a fusion network (e.g. MLP) to output a linear embedding. We compute the Euclidean distance between the linear embeddings of two sub-networks. We then pass the distance through a fully-connected layer and a sigmoid activation function to output a compatibility score of 0 or 1. To train our Siamese network, we use a binary cross entropy loss. After training our ensemble of Siamese classifiers, we can then apply the classifiers to Out-of-Bag (OOB) samples (i.e. unknown samples that were not included in their training sets). We repeat this multiple times to assign labels to unknown samples by averaging its OOB scores.

Finally, we will conduct experiments with various architectures and training methods across our pipeline, perform ablation studies, and visualize qualitative comparisons using a variety of embedding space analysis techniques. These experiments can help us to better gauge the effectiveness of the different design decisions we make.

References

Aggarwal, D., Valiyev, E., Sener, F., and Yao, A. Learning style compatibility for furniture. In *German Conference on Pattern Recognition*, pp. 552–566. Springer, 2018.

Bekker, J. and Davis, J. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4):719–760, 2020.

Bell, S. and Bala, K. Learning visual similarity for product design with convolutional neural networks. *ACM Transactions on Graphics*, 34(4), 2015.

Blijlevens, J., Creusen, M. E. H., and Schoormans, J. P. L. How Consumers Perceive Product Appearance: The Identification of Three Product Appearance Attributes. *International Journal of Design*, 3(3):27–35, 2009.

Datta, R., Joshi, D., Li, J., and Wang, J. Z. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (Csur)*, 40(2):1–60, 2008.

Guo, X., Wu, H., Gao, Y., Rennie, S. J., and Feris, R. S.

Fashion iq: A new dataset towards retrieving images by natural language feedback. In *CVPR*, 2021.

Han, X., Wu, Z., Huang, P. X., Zhang, X., Zhu, M., Li, Y., Zhao, Y., and Davis, L. S. Automatic Spatially-Aware Fashion Concept Discovery. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-October:1472–1480, 2017. ISSN 15505499. doi: 10.1109/ICCV.2017.163.

Kiapour, M. H., Yamaguchi, K., Berg, A. C., and Berg, T. L. Hipster Wars: Discovering elements of fashion styles. *European Conference on Computer Vision*, 2014.

Koch, G., Zemel, R., Salakhutdinov, R., et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015.

Liao, L., He, X., Zhao, B., Ngo, C.-W., and Chua, T.-S. Interpretable multimodal retrieval for fashion products. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, pp. 1571–1579, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356657. doi: 10.1145/3240508.3240646. URL https://doi.org/10.1145/3240508.3240646.

Liu, T., Hertzmann, A., Li, W., and Funkhouser, T. Style compatibility for 3D furniture models. *ACM Transactions on Graphics*, 34(4):1–9, 2015. ISSN 15577368. doi: 10.1145/2766898.

Mordelet, F. and Vert, J.-P. A bagging svm to learn from positive and unlabeled examples. *Pattern Recognition Letters*, 37:201–209, 2014.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Sharma, D., Gupta, N., Chattopadhyay, C., and Mehta, S. A novel feature transform framework using deep neural network for multimodal floor plan retrieval. *International Journal on Document Analysis and Recognition (IJDAR)*, 22:417 – 429, 2019.

Takagi, M., Simo-Serra, E., Iizuka, S., and Ishikawa, H. What Makes a Style: Experimental Analysis of Fashion Prediction. *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017*, 2018-January:2247–2253, 2017. doi: 10.1109/ICCVW.2017.263.

Tautkute, I., Mozejko, A., Stokowiec, W., Trzcinski, T., Brocki, L., and Marasek, K. What looks good with my sofa: Multimodal search engine for interior design. *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, FedCSIS 2017*, pp. 1275–1282, 2017. doi: 10.15439/2017F56.

Tautkute, I., Trzciński, T., Skorupa, A. P., Łukasz Brocki, and Marasek, K. Deepstyle: Multimodal search engine for fashion and interior design. *IEEE Access*, 7:84613–84628, 2019.

Vaccaro, K., Shivakumar, S., Ding, Z., Karahalios, K., and Kumar, R. The elements of fashion style. In *Proceedings of the 29th annual symposium on user interface software and technology*, pp. 777–785, 2016.

Weiss, T., Yildiz, I., Agarwal, N., Ataer-Cansizoglu, E., and Choi, J. W. Image-Driven Furniture Style for Interactive 3D Scene Modeling. *Computer Graphics Forum*, 39(7):57–68, 2020. ISSN 14678659. doi: 10.1111/cgf.14126.