Décor — Unimodal Analysis

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

1. Language Modality Exploration

In the language analysis section, we focus on exploring the effectiveness of different language representations, namely word-level representation and contextualized representation. The effectiveness is mainly evaluated qualitatively by showing word embedding clusters and various examples.

Although there is a wide range of word-level representations including GloVE (Pennington et al., 2014), Word2Vec (Mikolov et al., 2013) and FastText (Bojanowski et al., 2017), and contextualized representations such as ELMO (Peters et al., 2018), BERT (Devlin et al., 2019), Infersent (Conneau et al., 2017), Universal Sentence Encoder (Cer et al., 2018), we only pick GloVE and BERT as our starting point of analysis for this report. In the future phases of the project, it is likely that we will switch to other representations as needed.

1.1. Exploratory Data Analysis on Language

Before we dive into the representation learning part, we include a simple analysis of the characteristics of the dataset and distribution of language data.

As briefly introduced in our proposal, the IKEA dataset (Tautkute et al., 2017) consists of room and item images, item descriptions, room categories, and item categories. Some statistics is described in Table 1. The number of unique item descriptions is not equal to the number of items because same items in different rooms may have same descriptions as well. When it comes to the unique number of words in descriptions, we perform simple text preprossessing techniques such as stop word removal and non-alphabetical character removal to reduce noise.

One important finding for item description is that, as shown in Table 2, intuitively the original descriptions are not very informative of what the item really is. In fact, our group find that the item description itself cannot be used directly to describe the item and it should be used in a complementary way with images. Therefore, we decide to prepend the item category to the description for better language representations.

Characteristics	Number
Total Number of Rooms	262
Total Number of Items	2191
Number of Room Categories	9
Number of Item Categories	677
Number of Unique Item Descriptions	1438
Number of Unique Words in Item Descriptions	2223

Table 1. IKEA Dataset Language Statistics

Original Description	Item Category
Use FIXA hole saw set to make a hole in the desktop	Desk with add-on unit
The 2 drawers give you an extra storage space under the bed.	Bed frame w storage
Provided with safety film-reduces damage if glass is broken	Mirror

Table 2. Raw Language Examples

The distribution of room and item categories are ploted in Figure 1 and Figure 2. As we can see the class distributions for both are not very balanced and this might require us to perform data augmentation in future phases of the project. We did not show all labels for items since there are too many of them to fit in the pages, the main goal is to show the imbalance of the dataset.

1.2. Representation Methods

For word-level representation, we use GloVE to embed each word to a 100-dimensional vectors. Then each sentence can be represented by taking the average of all the word embeddings inside it.

For contextualized representation, we use a pretrained BERT uncased model to generate a 768-way vector for each word. For sentence-level, we use a sentenceBERT model that is essentially an extension to the pretrained BERT model by adding a mean pooling layer to encode each sentence to a 768-way vector.

1.3. t-SNE Visualization and Analysis

To see how this two embedding methods can represent the meaning of words, We can use t-SNE (van der Maaten & Hinton, 2008) method to visualize the high-dimensional word-embedding in two-dimension figures. These two fig-

[*]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.

ures 3 and 4 are the t-SNE result for GloVe and BERT representation of randomly selected word that appears in the description of furniture. Intuitively, GloVe representation split these words in two main groups, with several outliers, while BERT representations are more evenly distributed, with some small groups among it.

We can find some interesting examples in this two representation visualization. For example, let's take this four words as a example: "kitchen", "knife", "spoon", "fork". We assume these four words could be similar according to its semantic meanings. In BERT representation, we do find "kitchen" and "knife" are closing to each other, with other similar words like "food", "metal", "cheese", "saucepan", but "spoon" and "fork" are a little bit far away. In GloVe representation, these words are relatively more dispersed with "Kitchen" and "food" are close and "cheese" and "saucepan" are close. We assume BERT representation should have better performance than Glove as it has huge amount of pretraining knowledge. We do see that in this small example, BERT representation is better than Glove representation. Thus we think BERT may be more suitable for language representation in our next step.

1.4. Relation to Labels

In our project, the task is essentially a search problem where we want to recommend stylistically compatible items given an input item image, or input description, or both. It can be thought as an information retrieval problem as well. Therefore, we do not have explicit multi-class or multi-label ground truth labels as in the classification task. Instead, for relation to labels analysis, we are particularly interested in whether the returned ranked list of items are in the same room as the input item, which we interpret as stylistically compatible.

To get an initial and intuitive understanding of how effective the language representations are learned in the context of style similarity, our approach is as follows. Given an input description, we compute the cosine similarity between its embedding and all description embeddings the high dimensional space, and then return the top N items.

Some representative qualitative examples are shown below for further analysis. In the BERT example in Table 3, only one of the recommended item is within the same image as the input; however, this is not an evidence for the claim that our BERT language representation is intrinsically ineffective. If we closely examine the recommended examples, we can see that the second and the third recommended descriptions are actually the same item category: chair, and they are semantically but not stylistically related to the input item. This behavior is expected since we directly utilize the pretrained sentenceBERT model without providing any controllable guidance or stylistic similarity samples to make the model aware of styles. In addition, as we discussed in the first section, these item descriptions should be used complementarily with images to present "styles". This problem is illustrated even more in GloVe results as shown in 4. For GloVe, the returned results are not even semantically relevant to the input descriptions. This is also as expected because we know that unlike BERT where we directly apply the model trained with billions of Internet corpus, we need to fit GloVe model on our own IKEA corpus first to generate the embeddings; therefore, the generated GloVe embeddings do not have much commonsense knowledge to reason on.

Candidate Descriptions	Is in same image?
Desk with add-on unit Use FIXA hole...	YES
Chair with armrests, outdoor Use VRDA wood stain...	NO
Chair with armrests You can stack the chairs...	NO
Rack with 3 hooks By combining the different...	NO
Wash-stand with 2 drawers Since the wash-stand...	NO

Table 3. Top 5 BERT Candidates for Description "Chair, outdoor — Can be stacked, which helps you save space."

Candidate Descriptions	Is in same image?
Seat,back cushion, outdoor Ties and a strap ...	NO
Lamp shade Create your own personalised pendant ...	NO
Coffee table Practical storage space underneath	NO
Chair The cover is easy to take off and put on.	NO
Chair cushion, outdoor You can vary the look...	NO

Table 4. Top 5 GloVe Candidates for Description "Chair, outdoor — Can be stacked, which helps you save space."

2. Visual Modality Exploration

To determine how well different image models could potentially encode features that would be relevant to retrieving stylistically matching products, we conducted an embedding comparison of three image models: VGG16 (Simonyan & Zisserman, 2014), ResNet-18 (He et al., 2016), and CLIP ViT-B/32 (Radford et al., 2021). We selected VGG and ResNet as they are two well-established deep convolutional neural network (CNN) models. We also selected CLIP as it uses a state-of-the-art Vision Transformer (ViT) (Dosovitskiy et al., 2020) as its image encoder. Another interesting factor is that CLIP was pre-trained multimodally with images and text.

2.1. Generating Image Embeddings

VGG. We transform the images by resizing them into 3x224x244 and normalizing each color channel with $mean = [0.485, 0.456, 0.406]$ and $std = [0.229, 0.224, 0.225]$ based on the statistical distribution of ImageNet (Deng et al., 2009). We then load a VGG16 model pre-trained on ImageNet with the `models` subpackage in `torchvision`. Finally, we register a forward hook to extract the features on the ReLU activation right before the

Image Category	Count
Rooms	216
Bed	55
Chair	107
Clock	108
Couch	41
Dining Table	118
Plant Pot	35
Object	1,860

Table 5. IKEA Dataset Visual Statistics

linear classification layer. This yields a 4096-dimensional feature vector.

ResNet. We preprocess the images and load a pre-trained ResNet-18 model in a similar way as above. We then register a forward hook to extract the features on the avgpool operation right before the linear classification layer. This yields a 512-dimensional feature vector.

CLIP. We load a pre-trained CLIP[1] model that uses a ViT-B/32 Transformer architecture as its image encoder. CLIP provides several useful functions in its API. We preprocess the images with the `preprocess` method, encode the images with the `encode_image` method, then normalize. This yields a 512-dimensional feature vector.

2.2. Statistics for the Visual Part of the Dataset

The visual part of the dataset consists of 2541 photos. The photos are subdivided into 8 categories: rooms, bed, chair, clock, couch, dining table, plant pot, and object. See Table 5 for the statistics of each category. The majority of the photos are in RGB color with a small subset in grayscale. The photos are largely in 500x500 or 1000x1000 sizes, with the exception being photos of rooms which are of 1024 width and varying heights.

2.3. Overview of Model Embedding Spaces

An overview of the structure of each of the three embedding spaces is shown in Fig. 5, which displays aligned UMAP projections of the datasets according to each model. The plots were generated using Emblaze, an interactive Jupyter widget for embedding comparison[2]. Overall, we can see that the bulk of the embedding space is occupied by the "objects" category, which appears to be a miscellaneous label comprising many different object types (curtains, dishware, shelves, lighting, and even larger furniture such as wardrobes). Other classes, such as couches, dining tables, clocks, and chairs, are placed in clusters along the periphery of the projection. (Not shown in the plots, a small tight cluster of ovens is placed far away in the lower-right; this is likely because ovens are all displayed using a head-on graphic, making them very visually distinctive.) The fact that the majority of the objects are placed in this miscellaneous category may be a challenge that we need to remember when validating the accuracy of our methods on different classes.

Visual comparison of the three embedding spaces indicates, as would be expected, that CLIP clusters the various object types better than VGG or ResNet. For example, the cluster of clocks (red points at the top of each UMAP plot) becomes substantially larger and more tightly-clustered in CLIP. By examining the points that join this cluster, shown in Fig. 6, we can see that most of these points represent clocks that have an unorthodox appearance. Similarly, upholstered chairs are spread out in the VGG and ResNet spaces, but become much more clustered together in the CLIP space. These examples suggest that CLIP is better than VGG and ResNet at clustering images based on the functional roles of the objects they contain. However, we should note that the neighborhoods in VGG and ResNet still do contain meaningful information that could help with our downstream task. In particular, objects with similar constituent shapes and textures appear to be embedded closer together, which could assist in finding stylistically compatible items.

2.4. Encoding of Style Between Models

In order to accurately represent style information in a multimodal setting, it would be ideal if the unimodal image embedding space already encoded stylistic similarities. To test this, we selected stylistically distinctive examples of different product types and examined their nearest neighbors that were *not* of the same object class. The results, shown in Fig. 7, indicate that VGG16 and ResNet have roughly similar qualities, while CLIP seems to more accurately extract stylistically similar examples.

In the first example, a tuxedo-style gray loveseat, all three models retrieve ottomans with similar clean, flat lines within the top 5 results. However, VGG16 retrieves several beds which do not bear any stylistic resemblance to the query item. Additionally, CLIP's first result is an ottoman whose color matches the query couch; ResNet does retrieve a highly-compatible dark gray ottoman, but it is only the 5th most similar result.

The second example, a blue plastic chair with thin metal legs, illustrates the different ways these three models encode similarity. All three models retrieve tables that have similar leg shapes to the query, but VGG and ResNet also include stylistically-incompatible wooden tables – presumably also based on the shape of the legs. Meanwhile, CLIP is the only model that returns another blue item, indicating that it may use color more effectively than the other two models.

[1]https://github.com/openai/CLIP

[2]https://github.com/cmudig/emblaze

The third example, a wooden knife cover, tests the models' ability to reason about a more difficult case, where the object's functional role may not be as immediately obvious. Once again, CLIP outperforms the other models, retrieving objects with similar wooden textures that also seem suitable for holding other things. VGG and ResNet's results for this query are largely meaningless, except for the wooden basket that appears as the 5th result for VGG.

2.5. Discussion of Visual Modality

The above analyses suggest, unsurprisingly, that CLIP would provide the best unimodal representations for product images; there were no marked differences between VGG and ResNet found during our qualitative analysis. This may be because CLIP is trained on a multimodal task: having information about the ways humans refer to and discuss objects in images would likely improve the granularity of embedding spaces, particularly over models that are trained on ImageNet (which has relatively few furniture-related labels). However, this analysis applies to the three models *as-is*. VGG and ResNet would be easier to fine-tune to our task, which may result in improved cluster quality even over CLIP. As we develop our own architectures for the multimodal task, CLIP may serve as a very strong baseline that we can use without fine-tuning, while VGG or ResNet would likely be equally good starting points for training our own models.

References

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5: 135–146, 2017.

Cer, D. M., Yang, Y., yi Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. Universal sentence encoder. *ArXiv*, abs/1803.11175, 2018.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*, 2017.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

Pennington, J., Socher, R., and Manning, C. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. In *NAACL*, 2018.

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. *arXiv*, 2021. URL http://arxiv.org/abs/2103.00020.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Tautkute, I., Możejko, A., Stokowiec, W., Trzciński, T., Łukasz Brocki, and Marasek, K. What looks good with my sofa: Multimodal search engine for interior design. In Ganzha, M., Maciaszek, L., and Paprzycki, M. (eds.), *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems*, volume 11 of *Annals of Computer Science and Information Systems*, pp. 1275–1282. IEEE, 2017. doi: 10.15439/2017F56. URL http://dx.doi.org/10.15439/2017F56.

van der Maaten, L. and Hinton, G. Viualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008.

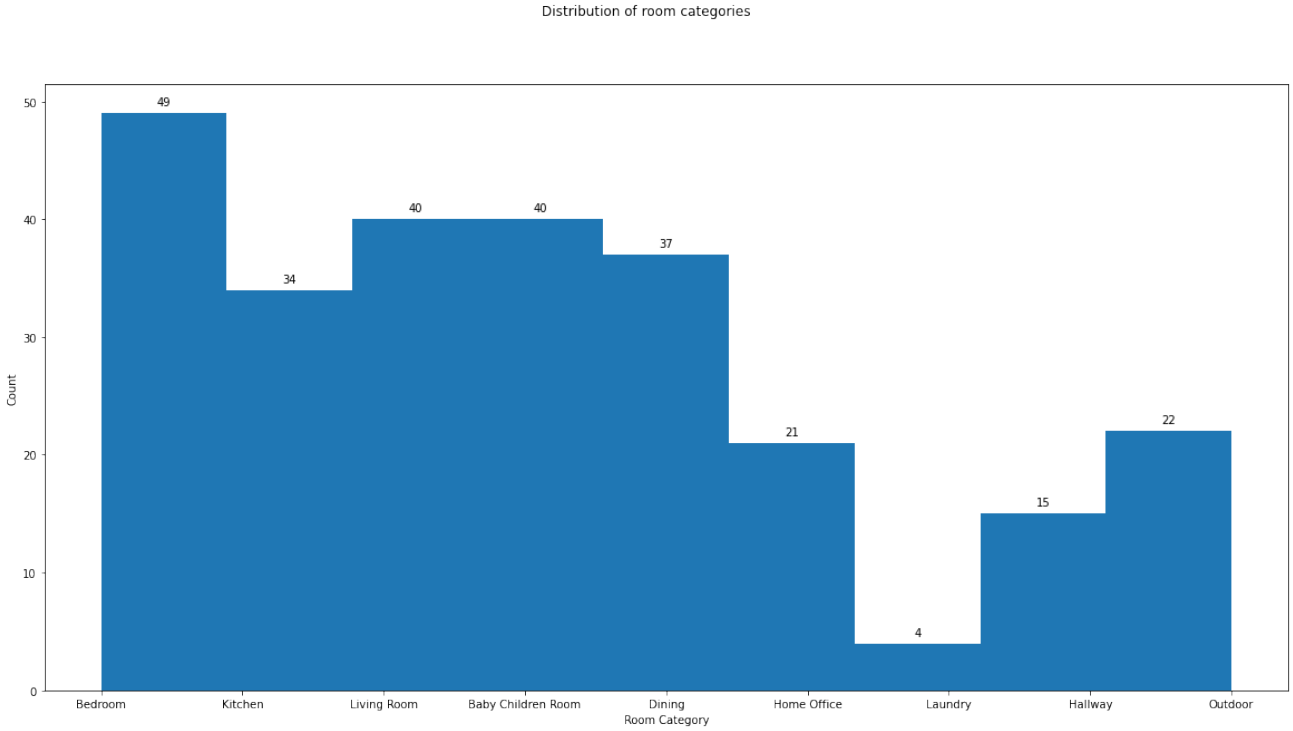

Figure 1. Distribution of room categories

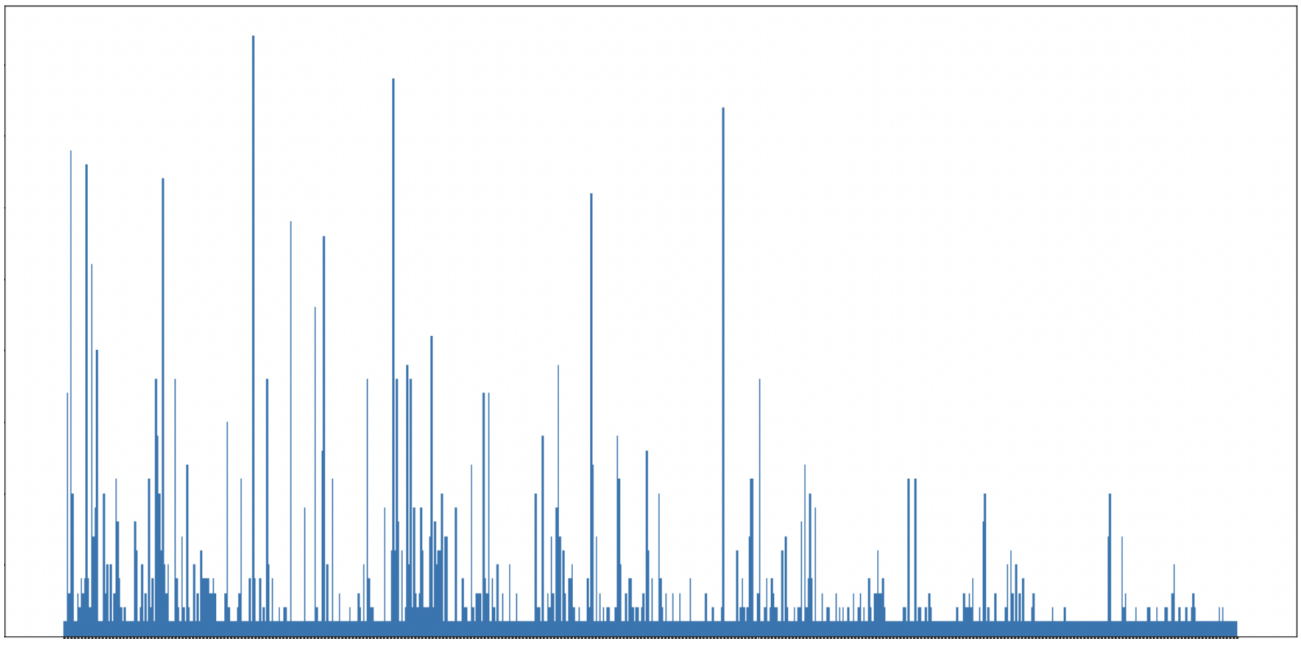

Figure 2. Distribution of item categories

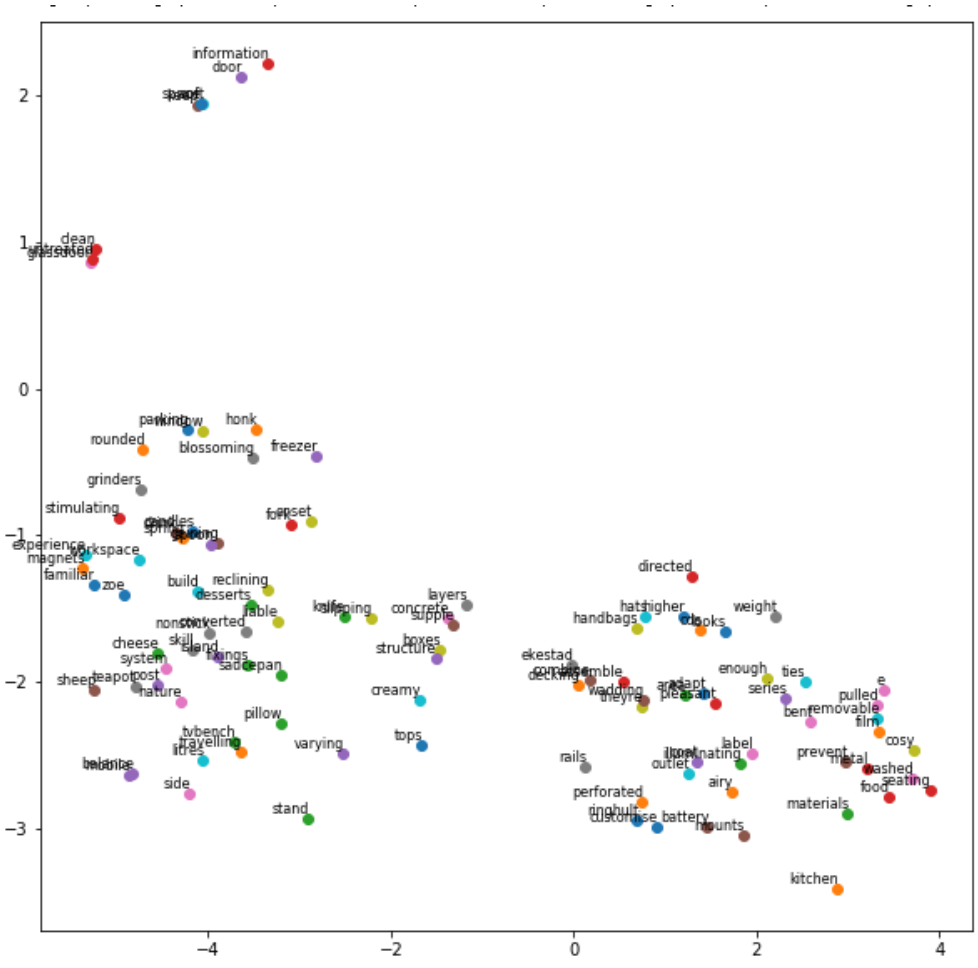

Figure 3. t-SNE visualization for GloVe Representation of selected word

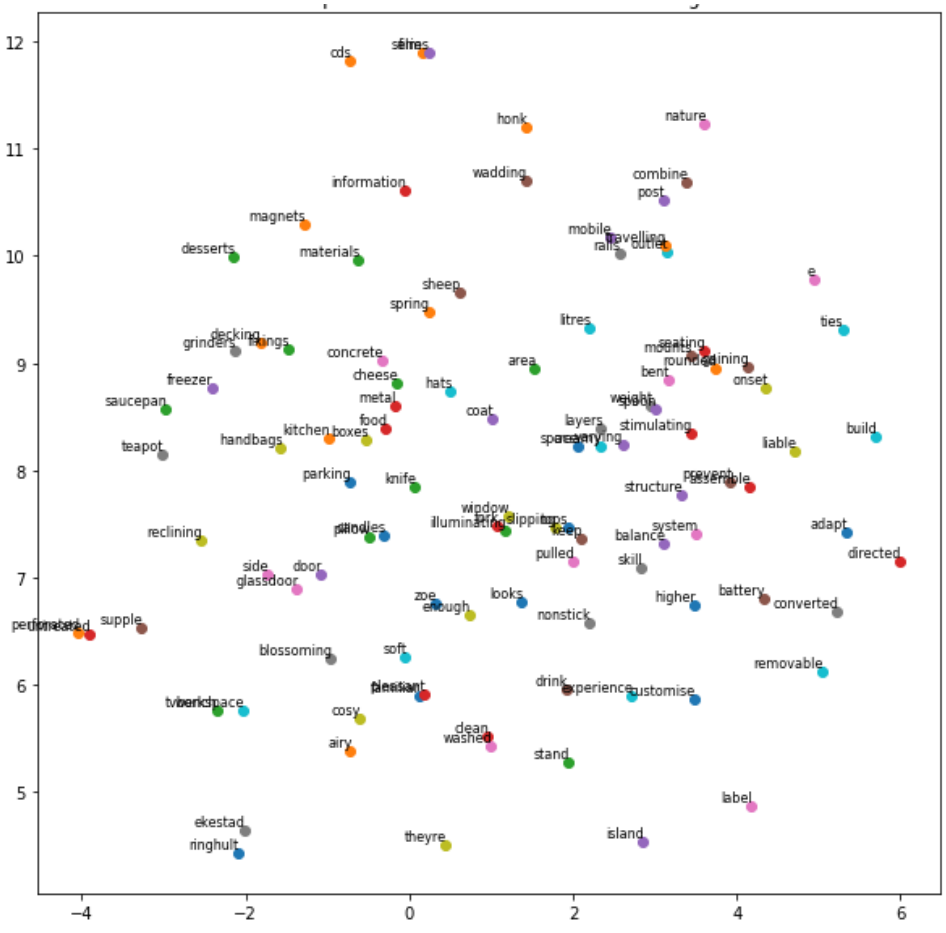

Figure 4. t-SNE visualization for BERT Representation of selected word

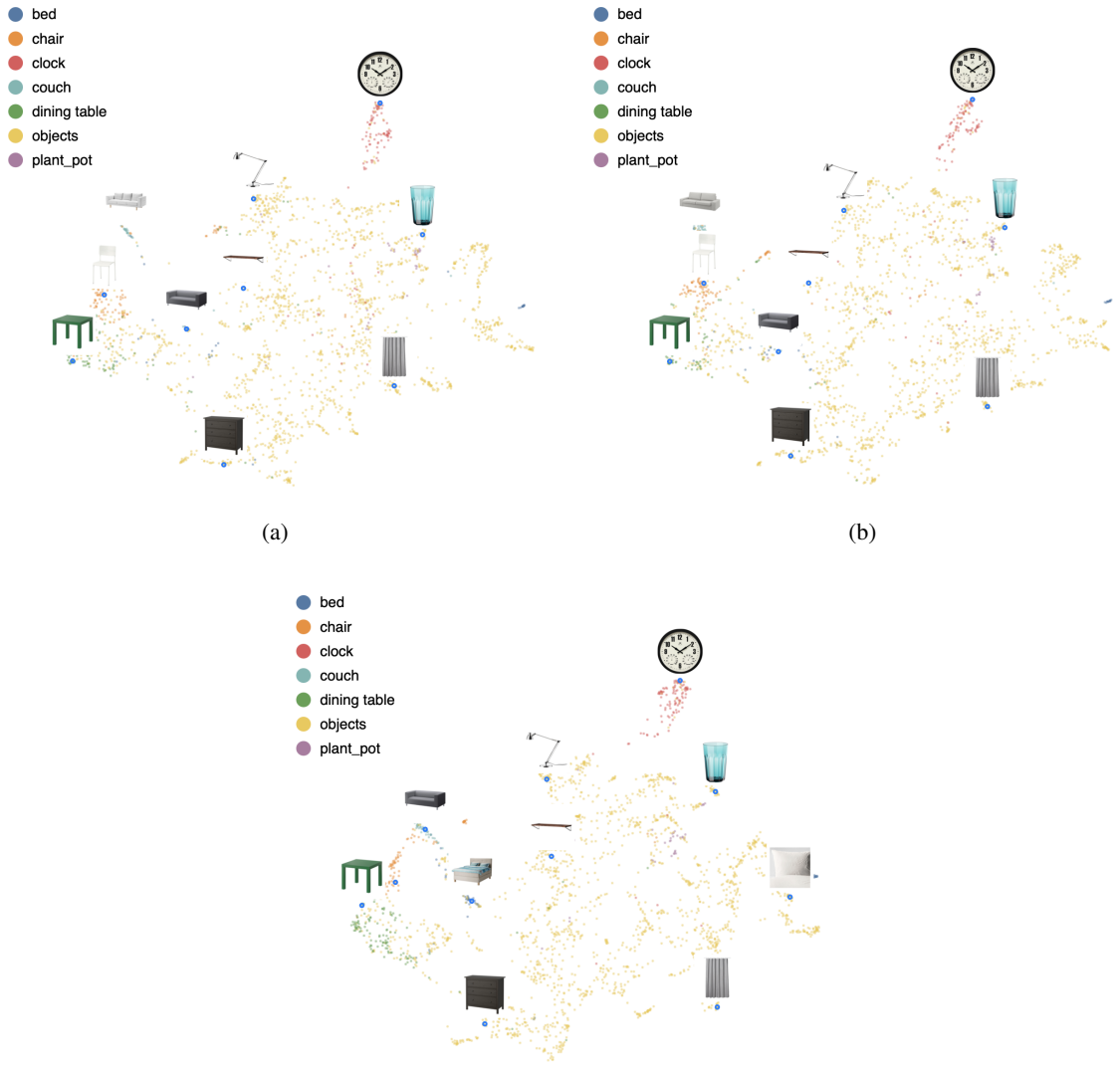

Figure 5. UMAP projections of the IKEA product dataset according to three image models: (a) VGG16, (b) ResNet, and (c) CLIP. Selected image thumbnails are shown to give a sense of the layout of the space.

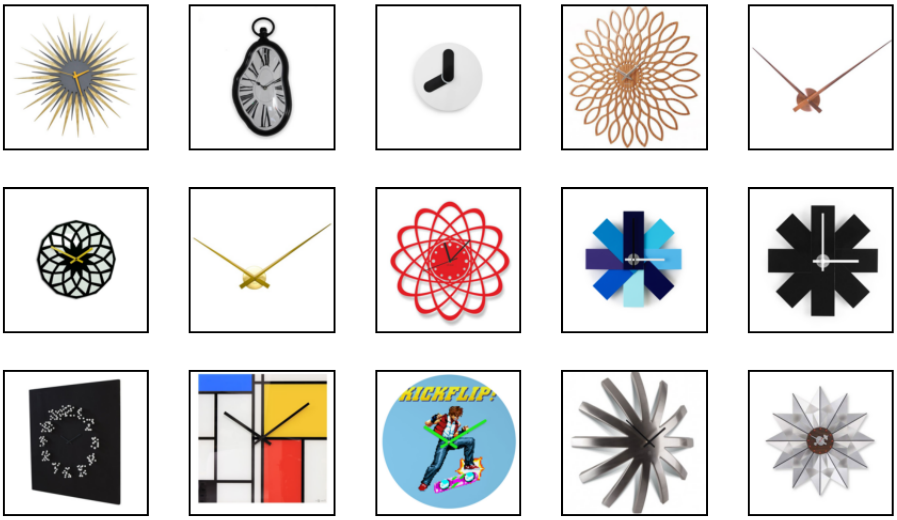

Figure 6. Points that join the cluster of clocks in CLIP but are not embedded near clocks in VGG16.

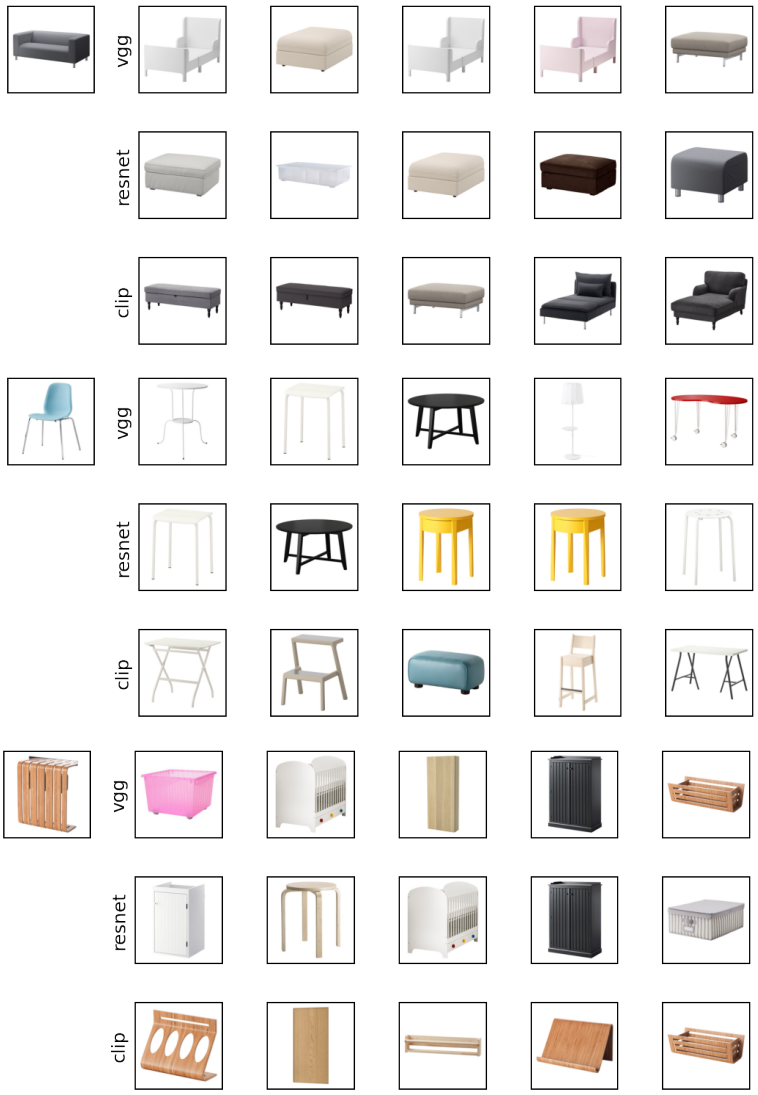

Figure 7. Nearest neighbors for selected product images that belong to other object classes. Each query (the leftmost images) is presented alongside the nearest neighbors using each of the three models.