Project Pre-proposal: Multimodal Retrieval for Interior Design

David Lin (chuanenl), Haocheng Han (haochenh), Yuanxin Wang (yuanxinw), Venkatesh Sivaraman (vsivaram)

What research problem are you planning to address?

We plan to address cross-media retrieval.

What dataset are you planning to use?

We are planning to work with the IKEA Interior Design Dataset (G1): https://github.com/IvonaTau/ikea. The dataset was collected from the IKEA website and contains product images, product descriptions, and room images. The dataset also has ground truth information for which products appear in which rooms and the categories of each product (e.g. chair, table, sofa).

What modalities will be involved in your course project?

We will be using images of isolated products and of complete room scenes. An example of a product image and a room scene is shown on the right. We will also make use of short descriptions of each product in the dataset to complement the product images. An example of a text description would be the following, in this case for a children's bed: "Extendable, so it can be pulled out as your child grows." Note that there are only 2193 product images and 298 room scenes in the IKEA dataset, so data augmentation and/or contrastive learning approaches will likely be necessary.

What multimodal challenge(s) are you planning to address as part of this project? What will be your evaluation metric?

For all retrieval-based tasks, we will use standard IR metrics such as recall at K and precision at K. For all generation-based tasks, we will use regression loss as an evaluation metric.

1. Given a product (image + text), recommend stylistically compatible products. If two products appear in the same room, they can be labeled as compatible, and vice versa ⟶ can be trained contrastively with fewer data.
2. Given a user-entered description (text) (e.g. slim wooden antique chair), retrieve fitting products and/or room setups. This is more of a translation task and might need some more data augmentation.

3. Given a product (image + text), retrieve fitting room setups. This is similar to the inverse of what the original paper was doing. The goal is to help the customers imagine how this product will fit in a house environment instead of just looking at a single product.
4. Given multiple text descriptions, first find the corresponding products and see if we can construct the room image using these individual product images. The goal is to enable customers the ability to “ensemble” their room with just text. We expect the products in the same batch of text descriptions to be reasonable (users should describe things that are likely to be in one room) even if we will train a generative model to output images.

What are the pre-existing baseline models for this task? Is the source code available?

Tautkute et al. developed a model to retrieve product images given room scenes and object descriptions, which can serve as a baseline. The model simply consists of an object detection stage (implemented with YOLO9000), a visual search phase using a pretrained ImageNet model, a word embedding-based textual search, and a simple blending function that re-ranks results based on similarity in ImageNet embedding space. The supplement code of the original paper is available at https://github.com/IvonaTau/style-search. We also explored the 7 citations of the original paper and find most of them have not released their implementations on GitHub.

Please give some details on how you plan to extend existing prior work

We will use the aforementioned supplementary code of the original paper as a baseline for development and extract common preprocessing and output functions for reuse. Currently we will be focusing on direction 1 and 3 as mentioned in the fourth section.

Who is your team and how are you planning to split the workload between team members? Can you provide a rough timeline/milestones you plan to follow?

David and Venkat are proficient in CV while Yuanxin and Haocheng are proficient in NLP. All members are expected to first run and reproduce the results of the original paper. And then we will split the literature review of the 7 citations mentioned above to update some components of the model or completely change the architecture of the model. David and Venkat will be primarily focusing on good image representation / generation / search techniques while Haocheng and Yuanxin will focus more on language representation / fusion.

What CPU, GPU and storage infrastructure do you have to have available for this project?

We have access to Google Colab Pro.

Are you interested in using Amazon Web Services or Google Cloud Platform as part of your project?

We are interested in using either AWS GPU credit or GCP GPU credit.