**Recommended serif fonts include Cambria, Georgia, and Times New Roman**

**Related Work:**

**Room-GPT:**

- Technology:

This model is ControlNet adapting Stable Diffusion to use M-LSD detected edges in an input image in addition to a text input to generate an output image. The training data is generated using a learning-based deep Hough transform to detect straight lines from Places2 and then use BLIP to generate captions.

**ControlNet** is a neural network architecture that can enhance large pretrained text-to-image diffusion models with spatially localized, task-specific image conditions.

**M-LSD** is another edge detection algorithm used in ControlNet. It stands for Multi-Scale Line Segment Detector.

**Hough transform**

The **Hough transform** is a popular feature extraction technique that converts an image from Cartesian to polar coordinates.

The Hough transform is a feature extraction technique used in image analysis, computer vision, and digital image processing.

The purpose of the technique is to find imperfect instances of objects within a certain class of shapes by a voting procedure.

Hough Transform is a computer vision technique to detect shapes like lines and circles in an image

Hough Transform is a computer vision technique to detect shapes like lines and circles in an image.It converts these shapes into mathematical representations in a parameter space, making it easier to identify them even if they're broken or obscured.

The Hough Transform is a well-known method for detecting parameterized objects. It is the de facto standard for detecting lines and circles in 2-dimensional data sets. For 3D it has attained little attention so far.

**BLIP**, a new VLP framework which transfers flexibly to both vision-language understanding and generation tasks.

**BLIP** effectively utilizes the noisy web data by bootstrapping the captions, where a captioner generates synthetic captions and a filter removes the noisy ones.

**Textual Search:**

**BOW and CBOW:**

**Bag of Words (BoW):**

BoW represents a text document as a sparse vector where each dimension corresponds to a unique word in the vocabulary.

It disregards the order of words in the document and only considers their frequency.

It's commonly used for tasks like document classification and sentiment analysis.

It refers to a way in which a group of words are represented without retaining order.

**Continuous Bag of Words (CBOW):**

CBOW is a model used in word embedding techniques like Word2Vec.

It predicts the target word based on its context (surrounding words).

It's trained to maximize the probability of predicting the target word given its context words.

CBOW tends to be faster to train compared to other models like Skip-gram, but it may not capture as much fine-grained information about word relationships.

Continuous Bag of Words (CBOW) is a popular natural language processing technique used to generate word embeddings. Word embeddings are important for many NLP tasks because they capture semantic and syntactic relationships between words in a language. CBOW is a neural network-based algorithm that predicts a target word given its surrounding context words. It is a type of “unsupervised” learning, meaning that it can learn from unlabeled data, and it is often used to pre-train word embeddings that can be used for various NLP tasks such as sentiment analysis, text classification, and machine translation.

Here is an example of how the CBOW model works. Consider the sentence "The quick brown fox jumps over the lazy dog." The CBOW model would take the words "the," "quick," "brown," "fox," "jumps," "over," "the," and "lazy" as input and try to predict the target word "dog." The model would do this by learning the patterns between the surrounding words and the target word. For example, the model would learn that the word "dog" is often preceded by the words "the" and "lazy."

What is the limitation of CBOW?

It averages the context of a word. Consider the word apple, which can refer to both a fruit and a company, but CBOW averages the two meanings and places it in a cluster for both fruits and companies.

It is a unidirectional model, which means that it only considers the context words in the forward direction.

It can be computationally expensive to train.

Here are some of the disadvantages of using the CBOW model:

It can be sensitive to the choice of hyperparameters.

It can be computationally expensive to train the model on large datasets.

It can be difficult to generalize the model to new datasets.

**Skipgram NLP Model**

Continuous Bag of Words (CBOW) and skip-gram Models are both architectures to learn the underlying word representations for each word by using NNs.

**The difference between CBOW & skip-gram :**

The **skipgram** model learns to predict a target word thanks to a nearby word.

On the other hand, the **CBOW** model predicts the target word according to its context. The context is presented as a bag of words contained in a fixed size window around the target word.

In the **CBOW** model, the distributed representations of context (or surrounding words) are combined to **predict the word in the middle**. While in the **Skip-gram** model, the distributed representation of the input word is used to **predict the context**.

 A prerequisite for any neural network or any supervised training technique is to have labeled training data. How do you a train a neural network to predict word embedding when you **don’t** have any labeled data (words and their corresponding word embedding)

**Skip-gram Model**

We’ll do so by creating a “fake” task for the neural network to train. We won’t be interested in the inputs and outputs of this network, rather the goal is actually just to learn the weights of the hidden layer that are actually the “word vectors” that we’re trying to learn.

The fake task for Skip-gram model would be, given a word, we’ll try to predict its neighboring words. We’ll define a neighboring word by the window size — a hyper-parameter.

**Skip-gram** works well with a small amount of the training data, represents well even rare words or phrases.

According to the original paper, Mikolov et al., it is found that Skip-Gram works well with small datasets, and can better represent less frequent words. However, CBOW is found to train faster than Skip-Gram, and can better represent more frequent words.

Limits of the Skip-gram:

The Skip-Gram model must compute a huge number of weights. For a vocabulary size of 10'000 words and 500 features, we would have 5 million weights in the hidden layer and the output layer. We also need a huge number of training data (typically counting in billions at that point).

**Visual Search:**

An image search engine works in the same way a text search engine does to give you the most relevant results: it pulls up a bunch of images based on a keyword or image. As when searching by text, an image search considers patterns and then points you to web sites based on matches.

K-Means clustering is a simple and efficient method for object detection in image.

By clustering similar pixels together, it can be used to identify objects in an image. While it is less accurate than other methods, it has the advantage of being computationally efficient.

K-means is an unsupervised classification algorithm, also called clusterization, that groups objects into k groups based on their characteristics

**R-MAC:**

This refers to the method known as R-MAC (Regional Maximum Activations of Convolutions), which is used to generate image representations. It's a technique commonly applied in computer vision tasks.

R-MAC operates by dividing the image into a fixed layout of spatial regions. Instead of considering the entire image as a whole, it breaks it down into smaller, localized regions. These regions are predetermined and remain fixed across all images processed using the R-MAC technique.

**Siamese Networks:**

Siamese networks are a special type of neural network architecture characterized by having two identical subnetworks, referred to as twins or arms, that share the same weights and architecture. These networks are designed to learn embeddings (dense, fixed-dimensional vector representations) of input data such that similar inputs are mapped close together in the embedding space.

Content-Based Image Retrieval: This refers to the task of retrieving images from a database based on their content, rather than metadata or textual descriptions. In content-based image retrieval, the goal is to find images that are visually similar to a given query image.

**Drawbacks for these models:**

They do not take into account the contextual and stylistic similarity of the retrieved objects, which yields their application to the problem of interior design items retrieval infeasible

Contextual and Stylistic Similarity: In interior design, the similarity between objects often depends not only on their visual appearance but also on their context within a space and their stylistic attributes. For example, two chairs might look visually similar, but their suitability for a particular room could depend on factors like color scheme, size, material, and overall style.

Failure to Capture Contextual and Stylistic Similarity: Both the R-MAC technique and Siamese networks primarily focus on capturing visual similarity based on image content. They analyze the visual features of individual objects but do not consider the broader context in which those objects exist or their stylistic characteristics.

Infeasibility for Interior Design Item Retrieval: Because of their inability to account for contextual and stylistic similarity, using R-MAC or Siamese networks for interior design item retrieval may not yield satisfactory results. Simply retrieving visually similar items without considering their suitability or compatibility within a specific interior design context could lead to mismatches or inappropriate recommendations.

Limitations in Addressing Complex Design Preferences: Interior design preferences are often complex and subjective, varying greatly among individuals. Relying solely on visual similarity metrics provided by R-MAC or Siamese networks may overlook crucial aspects of design preferences, such as personal style, aesthetic preferences, and functional requirements.

In summary, while R-MAC and Siamese networks excel at capturing visual similarity, their inability to consider contextual and stylistic factors limits their applicability to tasks like interior design item retrieval. To address this drawback, additional techniques that incorporate contextual understanding and stylistic analysis may be necessary to provide more accurate and meaningful recommendations in interior design contexts.

**Software Tools:**

The Visual Studio IDE is a creative launching pad that you can use to edit, debug, and build code, and then publish an app

**Chapter 5: Results and Discussion**

Results:

Inputs and outputs by Images

Expected results:

Actual results:

Discussion:

Write any reasonable reasons for the gap between inputs and outputs

Recommendations:

Colab Pro provides access to more powerful GPUs and TPUs compared to the free version of Colab.

**sBert and ConvoNeXt**

ConvNeXt

Pros:

High accuracy: ConvNeXt models have achieved state-of-the-art performance on various image recognition tasks.

Performance: State-of-the-art performance on many image classification benchmarks.

Pretrained Weights: Availability of pre-trained models helps achieve good results even with limited data.

Cons:

Resource Intensive: Large model sizes can lead to high computational and memory requirements.

Sentence Transformers (sBERT):

Pros:

Sentence Embeddings: Creates semantically meaningful sentence embeddings that capture the context and meaning of text.

Pre-trained models: Various pre-trained models available for different tasks and languages, enabling easy integration and fast implementation.

Multilingual support: Many Sentence Transformer models are multilingual, making them suitable for applications involving different languages.

Cons:

Computational cost: Generating sentence embeddings can be computationally expensive, especially for large datasets.