Most of the furniture websites use either textual search or visual search, so we have tried to combine between both of them to increase the accuracy of retrieving the user’s requirement.

Multimodal

It is an AI system designed to simultaneously process multiple forms of sensory input, similar to how humans experience the world.

SBERT:

*  **SBERT**: Fine-tuned on tasks like Semantic Textual Similarity (STS) and Natural Language Inference (NLI), resulting in embeddings that capture semantic meanings more effectively. This fine-tuning makes SBERT superior for tasks requiring understanding of sentence-level semantics.
* **BERT**: General-purpose and not specifically optimized for producing sentence embeddings, leading to less effective performance on semantic similarity tasks without additional processing.

 **Practicality for Large-Scale Retrieval**:

* **SBERT**: Suitable for large-scale information retrieval and clustering tasks. Once sentence embeddings are generated, they can be indexed and queried efficiently, enabling scalable applications like semantic search.
* **BERT**: Requires real-time computation for each query, which is impractical for large-scale retrieval due to high computational costs.

**RoBERTa:**

RoBERTa (A Robustly Optimized BERT Approach) is an improved variant of the BERT model, developed by Facebook AI. It introduces several modifications to the original BERT training process to enhance performance and robustness. Here’s a detailed definition and the benefits of using RoBERTa:

**Definition of RoBERTa**

RoBERTa builds upon the BERT architecture but incorporates several key changes to the training process:

1. **Training with More Data**: RoBERTa uses significantly more training data compared to BERT. It is trained on datasets that are ten times larger than those used for BERT, including data from sources like Common Crawl.
2. **Longer Training**: RoBERTa is trained for a longer duration with more steps, allowing the model to learn more effectively from the data.
3. **Dynamic Masking**: Instead of static masking used in BERT, RoBERTa uses dynamic masking, where the masking pattern changes during training. This prevents the model from overfitting to specific masked positions.
4. **Larger Batch Sizes**: The training process employs larger batch sizes, which helps in better gradient estimation and stabilizes training.
5. **Removal of Next Sentence Prediction (NSP)**: RoBERTa eliminates the Next Sentence Prediction objective used in BERT, as it was found to be not particularly useful and potentially harmful to performance.

ConvNext v1

Image classification model

ConvNeXt is a convolutional neural network (CNN) architecture designed to incorporate the advancements in vision transformers (ViTs) while retaining the efficiency and simplicity of CNNs. It was developed by researchers at Facebook AI (Meta AI) and released in early 2022. ConvNeXt aims to modernize and optimize traditional CNNs to achieve state-of-the-art performance on various computer vision tasks, comparable to or even surpassing ViTs.

**Key Features of ConvNeXt**

1. **Modernized Architecture**:
   * ConvNeXt incorporates design principles from ViTs and modern neural network architectures, leading to significant improvements in performance and efficiency.
2. **Layer Normalization and Activation Functions**:
   * The model employs Layer Normalization instead of Batch Normalization, which is more stable and effective in deep networks.
   * It uses GELU (Gaussian Error Linear Unit) activation functions, which have been shown to perform better than ReLU in deep networks.
3. **Depthwise Convolutions**:
   * ConvNeXt makes extensive use of depthwise separable convolutions, which reduce the number of parameters and computational complexity while maintaining performance.
4. **Simplified Design**:
   * The architecture avoids complex components like multi-head self-attention, focusing instead on streamlined and efficient convolutional operations.
5. **Scaling Laws**:
   * ConvNeXt follows scaling laws observed in ViTs, increasing model depth, width, and resolution systematically to achieve better performance.

**Feature Fusion:**

**Usages of Feature Fusion**

1. **Multimodal Learning**:
   * **Combining Different Modalities**: For tasks that involve multiple data modalities (e.g., images, text, audio), feature fusion helps in integrating these diverse sources of information to improve the overall model performance. For example, in a video understanding task, features from the visual (frames) and auditory (soundtrack) modalities can be fused.

**Late Fusion**:

* **Decision-Level Fusion**: Predictions from multiple models, each trained on different features, are combined. This can be done through techniques like voting, weighted averaging, or stacking (ensemble methods).
* **Score Fusion**: Confidence scores from different models are combined to make the final prediction.

Cosine Similarity:

**Usage of Cosine Similarity**

1. **Text Similarity and Natural Language Processing (NLP)**:
   * **Document Similarity**: Used to measure the similarity between two documents based on their term frequency vectors or embeddings.
   * **Sentence Similarity**: Applied in tasks like semantic textual similarity, where the goal is to determine how similar two sentences are in terms of meaning.
   * **Information Retrieval**: Used in search engines to find documents that are similar to a query by comparing the query vector with document vectors.
2. **Information Retrieval and Search Engines**:

* **Document Ranking**: Used to rank documents based on their similarity to a search query.
* **Query Expansion**: Helps in finding similar terms or phrases to expand a search query for better results.