Multi-Modal Embeddings

Your one-stop post for everything essential

Bhavishya Pandit

Introduction

Blending diverse modalities into a unified representation

What do we mean by the term Multi-Modal?

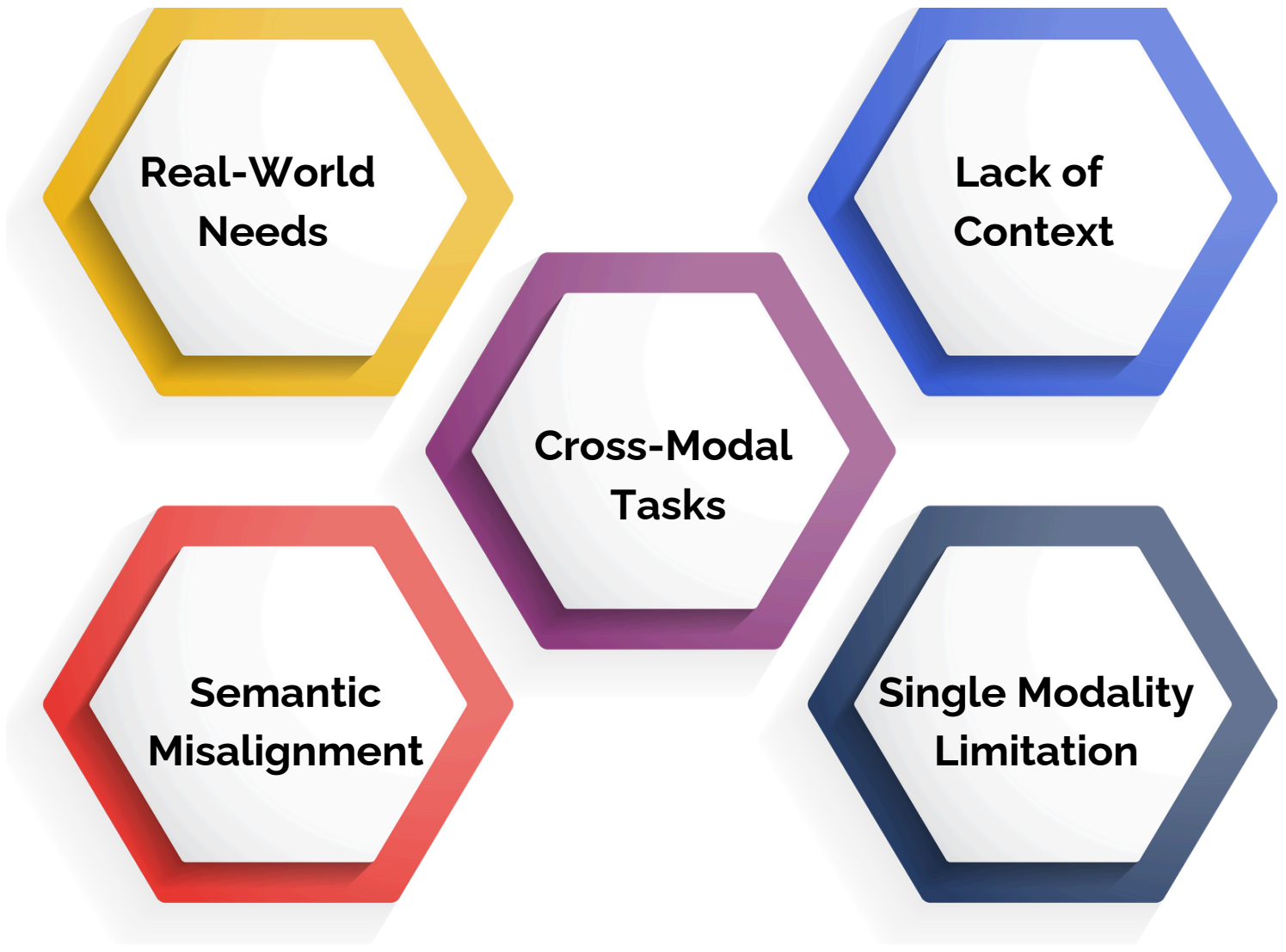
Multi-modal refers to the ability of a system to process and integrate information from multiple data types or "modalities," such as text, images, audio, and video.

What are embeddings?

Embeddings are compact vector representations of data that capture semantic meaning. They make data comparable, enable cross-modality linking, and improve AI performance in tasks like search, classification etc

Cheers, the base is set. Now let us understand why was there a need for Multi-Modal Embeddings **(MME)** and how it works.

Why we need MME?



We need Multimodal Embeddings (MME) because the vanilla embeddings had the following challenges:

- **Single Modality Limitation**: Vanilla embeddings handle only single data types and lack cross-modal linking
- **Lack of Context**: They can't link text with images, like captions to photos
- **Cross-Modal Tasks**: Vanilla embeddings can't handle tasks like image-text search or video captioning.
- **Semantic Misalignment**: Vanilla embeddings can't align different structures like pixels and words effectively
- **Real-World Needs**: Applications like AI assistants and recommendation systems require integrating data from multiple sources

How MME Works

Data Input: Take data from multiple sources, such as text, images, or audio.

Separate Encodings: Each modality is processed separately using specialized models (e.g., a text encoder for text, an image encoder for images).

Feature Extraction: The encoders convert the data into numerical embeddings (vectors) that capture key features of each modality.

Alignment: A shared space is created where embeddings from different modalities are mapped close to each other if they are semantically related (e.g., a cat photo and the word "cat").

Fusion: Combine the embeddings into a single representation to make decisions or predictions based on the integrated information.

5 Different Techniques

- **Joint Encoding:** It merges all modalities early for deep interaction but is computationally heavy.

- **Late Fusion:** It encodes modalities separately and combines them later, offering efficiency but weaker interactions

- **Cross-Attention Mechanism:** Uses attention layers to relate modalities, perfect for tasks like visual Q&A with fine-grained relationships.

- **Contrastive Learning:** Trains embeddings to align related pairs and separate unrelated ones, ideal for retrieval tasks but requires large paired datasets.

- **Multimodal Transformers:** Advanced models with attention layers integrate modalities, offering top results but needing high computational power.

Advantages

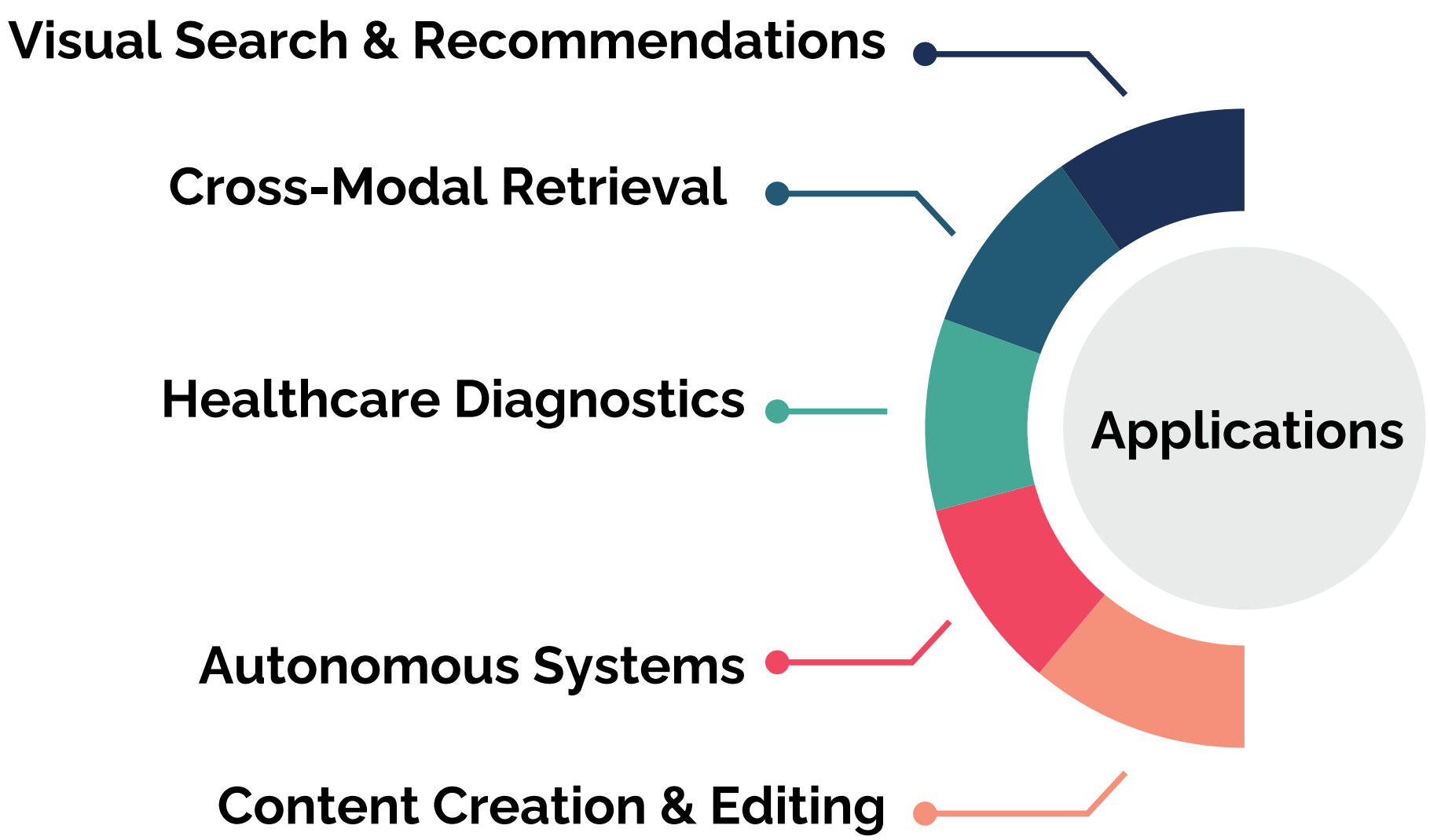
The MM-GEM model demonstrates the efficacy of multi-modal embedding models

- **Performance on Benchmarks**: Comparable to OpenCLIP on image-text retrieval benchmarks (COCO, Flickr30K) and superior on complex benchmarks like long-form text-image retrieval (L-DCI).

- **Fine-Grained Understanding**: Enables region-level captioning and detailed retrieval tasks without additional training.

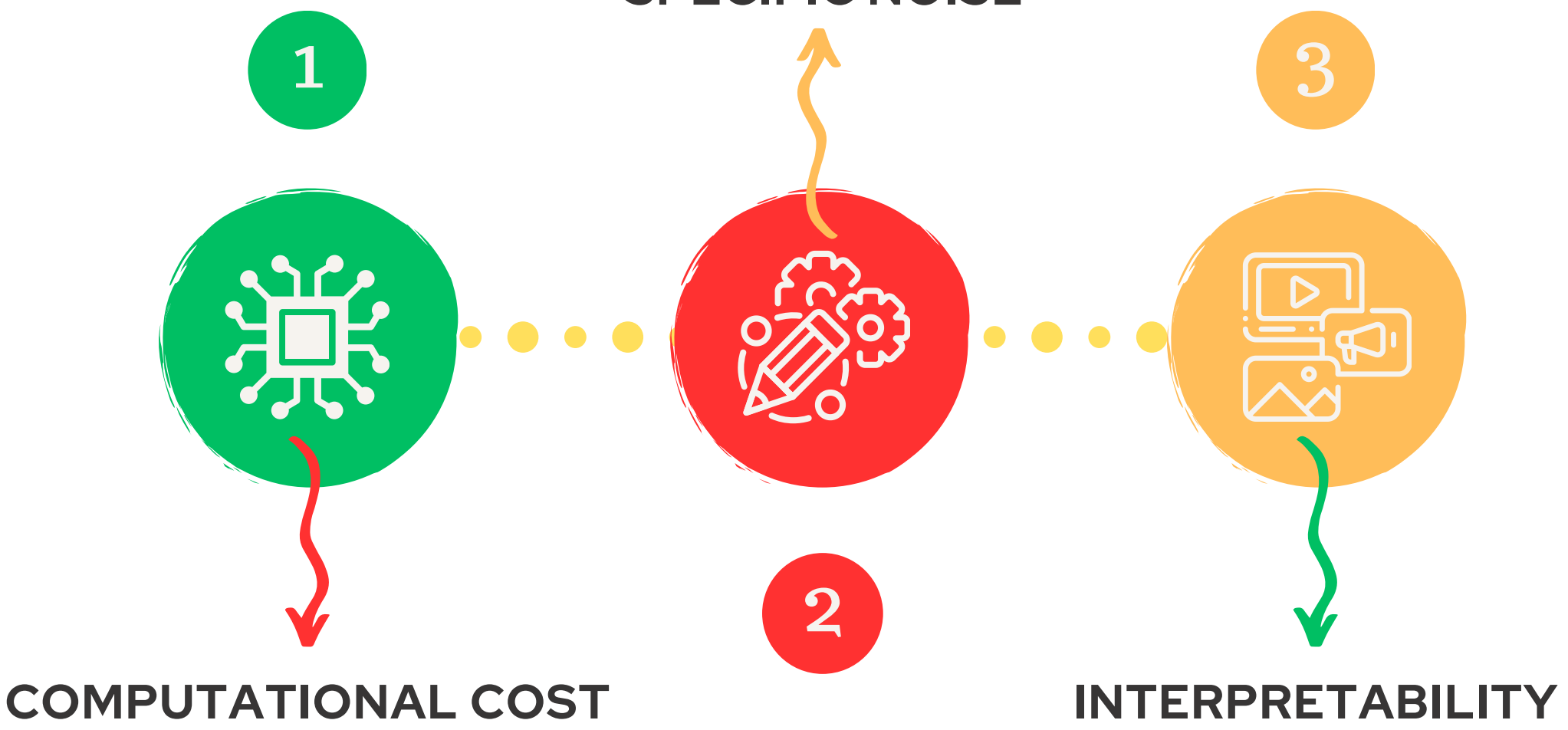
- **Efficiency**: The unified architecture reduces computational overhead while maintaining task-specific performance.

Applications



- **E-commerce**: Improves product recommendations by 30% through visual search and user-generated images.
- **Content Retrieval**: Efficiently finds related text and images across modalities.
- **Healthcare**: Combines medical images and text for better diagnostics.
- **Autonomous Vehicles**: Integrates visual and textual data for decision-making.
- **Media Creation**: Generates precise captions and annotations for multimedia.
- **Virtual Assistants**: Enhances interaction by merging audio, text, and visuals

Challenges



Mentioning a few key challenges of working with MME:

Computational Cost: Processing and integrating multiple modalities, especially with models like transformers, demands significant computational power and memory.

Modality-Specific Noise: Different modalities can have varying levels of quality or noise (e.g., blurry images or incomplete text), making it hard to create accurate embeddings.

Interpretability: Understanding how the model integrates and prioritizes information from different modalities is difficult, making debugging and improvement harder.

Follow to stay updated on Generative AI

SAVE

LIKE

REPOST