Image text matching using multimodal embeddings

Concept understanding

The capacity to match the image model with the text model is a key element of successful multimodal retrieval systems, often known as Multimodal Retrieval Augmented Generation (RAG) pipelines. This implies that even if the image and the text originate from two distinct Data Modality Types (Images vs Text), one must be able to ascertain whether or not they are discussing the same semantic concept or the same object.

Retrieval with conventional text-based systems usually entails a search via written information only. However, there are a lot of written materials in the real world, including user manuals, instructional PDFs, and technical manuals. where crucial information is frequently communicated via Text + Image. Image-Text Matching makes it possible to compare images and text in an intuitive way, resulting in cross-modal comparisons between text and images.

Shared Embedding Space Concept

Modern multimodal models such as SigLIP and CLIP (Contrastive Language Image Pretraining) generate shared vector embeddings that can incorporate text and images as inputs simultaneously. Therefore, the following might be used to illustrate how CLIP/SigLIP functions:

1. A numerical vector is created from a written sentence, such as “Connect the red wire to terminal A”.
2. A numerical vector is also created from an image of the electrical wiring, where both of these inputs are put into a mathematical space or format with the same dimensions.

Consequently, their vectors will be near together (in terms of distance in this mathematical space) if both the image and the text reflect the same or a similar concept. This will enable the comparison of images and text in a meaningful context.

Semantic Alignment Between Modalities

Diagrams and words could only be matched using keyword-based matching if they shared the same terms or keywords. Diagrams and words can be matched using multimodal embeddings since their conceptual and semantic meanings are similar. For example, you can match precise instructions on how to complete a task step-by-step with visual drawings.

The capacity to locate suitable pictures based on their concept or use case rather than on specific words or phrases in the image caption is another way that multimodal embeddings might offer better search capabilities. For instance, if someone typed in the query "safety grounding connection," the system would be able to provide an image of a wiring diagram or a grounding symbol, even if the image caption did not include the word "safety grounding connection".

Role of Contrastive Learning

Multimodal embedding models learn to recognise what works and what doesn't by training on both text and visuals. During training, the model gains knowledge from several image-text pairs. If two captions (or image captions) accurately describe one another, it learns to tightly link them as being closely related. The model learns to view these bits of information as unrelated to one another if it is

given an image with a text description that differs from the original caption or if a text caption does not accurately describe the image. Consider the process of teaching the model as teaching by contrast.

This aids in the model's growing capacity to connect words or concepts with visual elements. As a result, the model learns the underlying meaning of each image and its description in addition to memorising samples of visual visuals and the written descriptions that go with them. A model may successfully comprehend new visual diagrams, manuals, etc. that it may not have seen before but that are sufficiently comparable to previously learnt notions.

Similarity Measurement and Matching

When a model processes both text and a picture, it will produce an embedding. Following this procedure, the model will use their vectors to compare the two embeddings, which are numerical representations of the text and image. The model looks at the direction in which the two points are placed rather than the number of elements in the vector.

The model will compute the cosine similarity between the two vectors to ascertain their directionality. Cosine similarity is a widely used technique because it works well in high-dimensional spaces and is precise and efficient when comparing big and complicated data sets, such as those generated by neural networks. Rather than measuring each vector's raw value, cosine similarity assesses the direction of two vectors.

To put it simply, cosine similarity evaluates how similar two items are based on their context (or meaning), length, and number of details. The text and image are referring to the same subject or idea if the cosine similarity score is high. This implies that the text and the accompanying image are in harmony.

Integration into Multimodal RAG Pipelines

Image-text matching serves as the link between document understanding and information retrieval in multimodal RAG systems. Image-text matching links any relevant explanations in the surrounding text to the visual components of the documents (diagrams, screenshots, etc.).An illustration of a multimodal RAG system Additionally, it is the part of the system that makes cross-modal retrieval possible. In other words, when an image is sent to the system, it will be able to find images based on a user's text search and/or obtain pertinent text information.

The language model receives more contextual information from the image-text matching component, giving it a more comprehensive picture of the words it processes as opposed to merely the letter sequence.

As a result, the system's output will be more precise and visually appealing, particularly when dealing with technical materials. This stage ensures that the picture embedding and text embedding are perfectly aligned and express the same underlying meaning before doing technology-based large-scale retrieval using the vector databases.

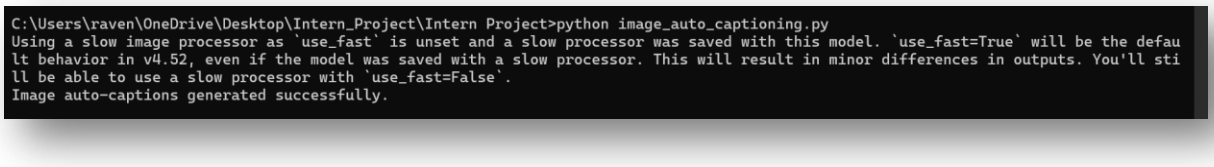

Figure 1: Image auto-captions generated successfully

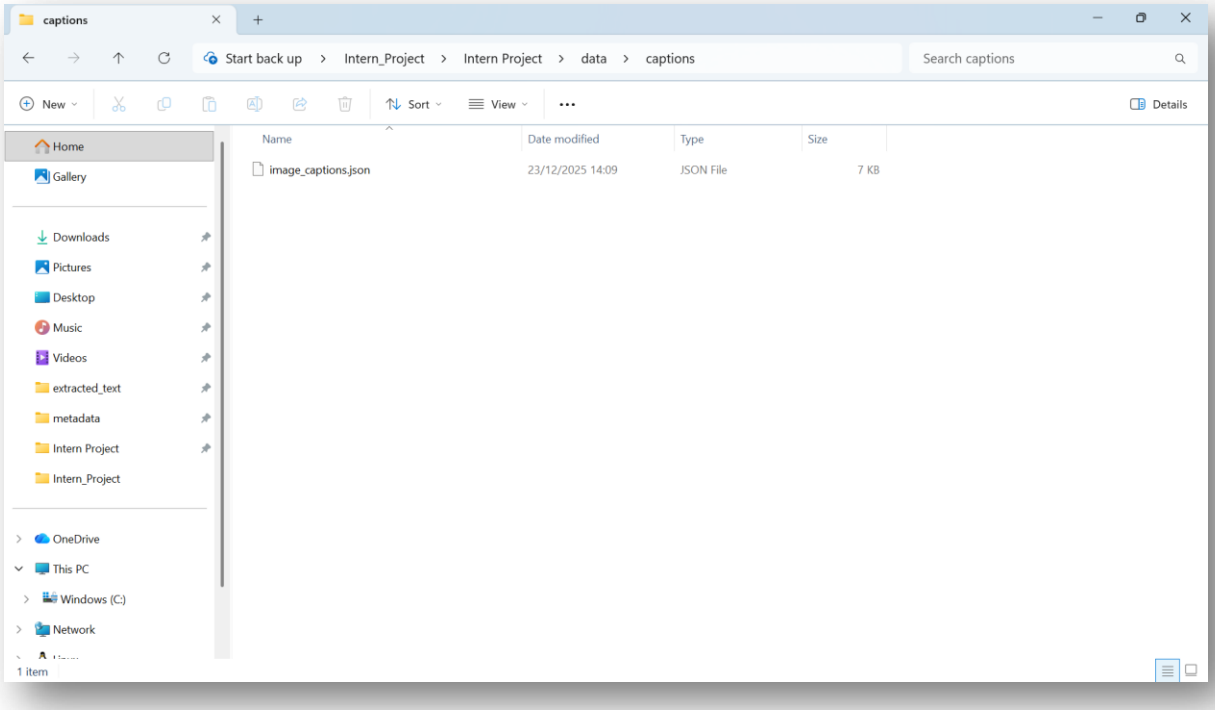

Figure 2: Image captions created in the folder

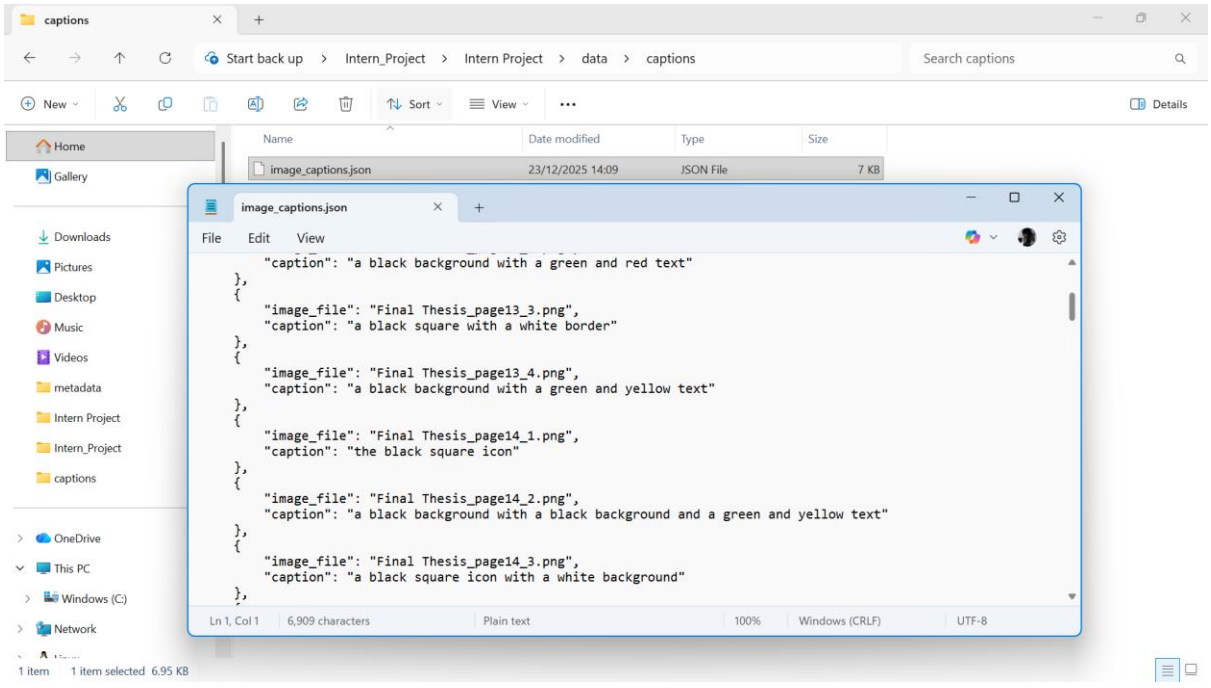

Figure 3: The created file opened in Notepad

Write up about the code

Cosine similarity was used to apply image-text matching on multimodal embeddings produced by the CLIP model. The system can recognise semantically linked textual and visual content by comparing text and image vectors within a shared embedding space. This serves as the foundation for multimodal retrieval in subsequent RAG phases.