Comparison of Multimodal LLMs: CLIP and BLIP

Introduction to Multimodal Large Language Models

Multimodal Large Language Models (LLMs) are a fascinating area of artificial intelligence that extends the capabilities of traditional LLMs beyond just text. These models are designed to process and understand information from multiple modalities, such as text, images, and sometimes audio or video. By learning joint representations across these different data types, multimodal LLMs can perform complex tasks that require understanding context from various sources, leading to more comprehensive and intelligent applications.

This report will focus on two prominent multimodal models: CLIP (Contrastive Language–Image Pre-training) and BLIP (Bootstrapping Language-Image Pre-training), detailing their architectures, input types, applications, and how they handle cross-modal interactions.

1. CLIP (Contrastive Language–Image Pre-training)

Architecture

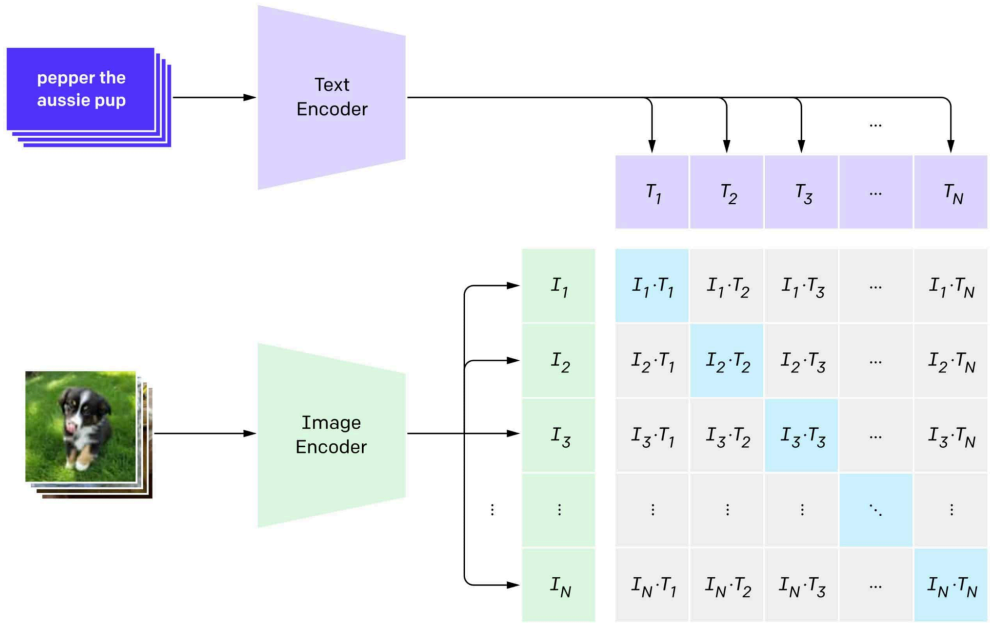
CLIP is developed by OpenAI and is primarily a vision-language model. Its architecture consists of two main components:

1. **Text Encoder**: A Transformer-based model (e.g., a masked self-attention Transformer) that processes text inputs. It converts text into a numerical embedding space.
2. **Image Encoder**: A vision model (e.g., a ResNet or Vision Transformer - ViT) that processes image inputs. It converts images into the same embedding space as the text encoder.

The key innovation in CLIP is how these two encoders are trained together.

Input Types

- **Image**: A single image.
- **Text**: A single piece of text, usually a natural language phrase or sentence.



How it Handles Cross-Modal Inputs (Contrastive Learning)

CLIP's strength lies in its **contrastive pre-training objective**. It learns to associate images with their correct text descriptions (and disassociate them from incorrect ones) without direct pixel-to-text annotation at the individual level.

During training:

1. A batch of N image-text pairs is prepared.
2. The N images are passed through the Image Encoder, and the N texts are passed through the Text Encoder, resulting in N image embeddings and N text embeddings.
3. A similarity matrix is computed, measuring the cosine similarity between every image embedding and every text embedding in the batch.
4. The model is trained to maximize the similarity between correctly paired image-text embeddings while minimizing the similarity between incorrectly paired (negative) embeddings.

This forces the encoders to learn a shared, high-dimensional multimodal embedding space where semantically similar image and text representations are close to each other.

Main Applications

- **Zero-shot Image Classification**: CLIP can classify images into categories it has never seen during training by simply comparing the image embedding to text embeddings of the category names.
- **Image Search**: Retrieving relevant images based on text queries, or vice-versa.
- **Image-to-text Retrieval**: Finding the most descriptive text for a given image.
- **Text-to-image Retrieval**: Finding images that match a given text description.
- **Object Detection (with adaptations)**: Can be used as a powerful backbone for open-vocabulary object detection.

2. BLIP (Bootstrapping Language-Image Pre-training)

Architecture

BLIP, developed by Salesforce Research, aims to be more versatile than CLIP by unifying vision-language understanding (VLU) and generation (VLG) tasks. It uses a novel "Multimodal Mixture of Experts (MME)" Transformer architecture and an image captioning approach.

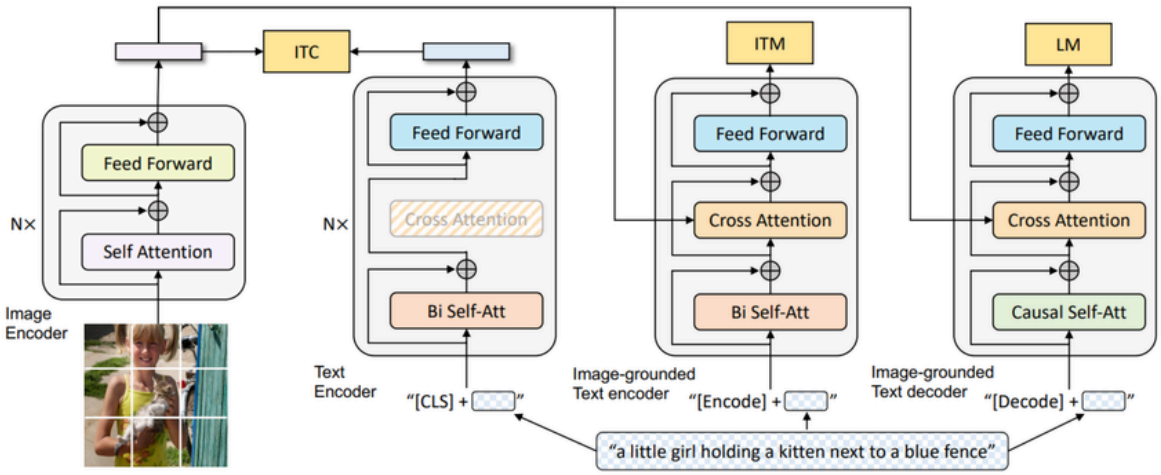
Its core components include:

1. **Image Encoder**: A Vision Transformer (ViT) processes the input image.
2. **Text Encoder**: Processes text inputs, similar to other Transformer-based encoders.
3. **Image-Grounded Text Encoder**: This is a key component that receives both image features (from the Image Encoder) and text tokens. It encodes text while considering the visual context, facilitating VLU tasks.
4. **Image-Grounded Text Decoder**: Also receives image features and generates text, enabling VLG tasks like image captioning.

BLIP uses a new pre-training strategy called "Captioning and Filtering (CapF)", which involves generating synthetic captions and filtering noisy ones.

Input Types

- **Image**: A single image.
- **Text**: Can be used as input for understanding (e.g., image-text matching) or generated as output (e.g., image captioning).



How it Handles Cross-Modal Inputs (Unified Framework)

BLIP handles cross-modal inputs through a unified framework that combines three objectives during pre-training:

1. **Image-Text Contrastive (ITC) Loss**: Similar to CLIP, it learns to align image and text embeddings in a shared space, promoting VLU.
2. **Image-Text Matching (ITM) Loss**: A binary classification task where the model predicts if an image-text pair is positive (matched) or negative (unmatched), further enhancing VLU.
3. **Image Captioning (IC) Loss**: The Image-Grounded Text Decoder generates a caption for the given image, promoting VLG.

The CapF strategy generates captions for noisy web images and then filters out low-quality ones using the ITM module, bootstrapping higher-quality training data.

Main Applications

- **Image Captioning**: Generating descriptive text for images.
- **Visual Question Answering (VQA)**: Answering questions about images.
- **Image-Text Retrieval**: Both image-to-text and text-to-image retrieval.
- **Image-Text Matching**: Determining if an image and text are semantically related.
- **Zero-shot Transfer**: Like CLIP, it can be adapted for zero-shot tasks.

3. Comparison Table: CLIP vs. BLIP

Feature	CLIP	BLIP
Primary Focus	Image Text Alignment (VLU)	Unified VLU and VLG (Understanding and Generation)
Core Training Objective	Contrastive Learning (ITC)	Unified Framework (ITC, ITM, IC)
Key Innovation	Learning from raw web data via contrastive loss	Multimodal Mixture of Experts, Captioning & Filtering (CapF)
Encoders	Seperate Image and Text Encoders	Image Encoder, Text Encoder, Image Grounded Text Encoder/Decoder
Cross Modal Handling	Shared embedding space, cosine similarity	Shared embedding space, multiple objectives
Application	Zero shot Image Classification	Image Captioning, VQA, enhanced tereieval

4. References

- **CLIP**: Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International Conference on Machine Learning*. PMLR, 2021.
- **BLIP**: Li, Junnan, et al. "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation." *International Conference on Machine Learning*. PMLR, 2022.