Assignment 5.1 - Comparison of Two Multimodal Models: BLIP and CLIP

Prasiddha Koirala

June 27, 2025

1 Introduction

Multimodal Large Language Models **(LLMs)** are models capable of understanding and generating data across multiple modalities such as text, image, and audio. This report compares two prominent multimodal models: **CLIP (Contrastive Language–Image Pretraining)** developed by OpenAI, and **BLIP (Bootstrapped Language-Image Pretraining)** developed by Salesforce. We discuss their architectures, input modalities, applications, and how each model handles cross-modal inputs.

2 Model 1: CLIP

CLIP (Contrastive Language–Image Pretraining), introduced by OpenAI in 2021, is trained to connect images and their corresponding descriptions.
Architecture: Dual-encoder architecture with a Vision Transformer **(ViT)** for image inputs and a Transformer for text.

- **Training Objective:** Contrastive learning to align text and image embeddings in a shared latent space.
- **Usage:** Zero-shot tasks by computing cosine similarity between encoded image and text representations.

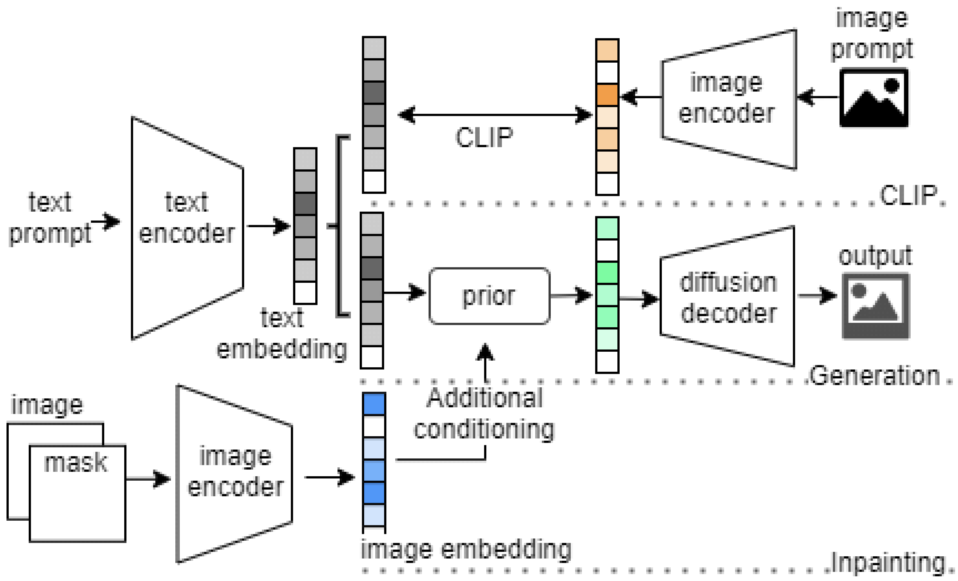

Figure 1: **CLIP** Architecture (source: OpenAI)

CLIP can handle image and text inputs, and it's mostly used for tasks like zero-shot classification, image search, and visual question answering.

3 Model 2: BLIP

BLIP (Bootstrapped Language Image Pretraining) is a more recent multimodal framework, designed with a unified vision-language interface.
Architecture: Unified model with a vision encoder (ViT or ResNet) and a Transformer-based decoder.
• **Training Objective:**Pretrained with mixture of image-text matching,captioning and question answering tasks.
• **Usage:** Generation and understanding of natural language grounded in images using cross-attention.

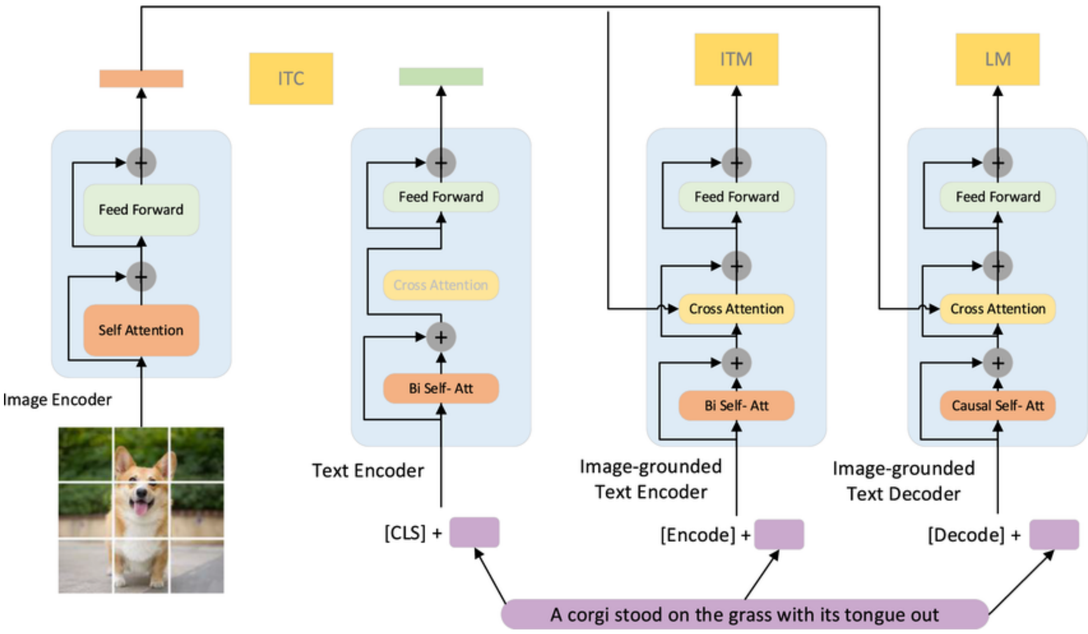

Figure 2: **BLIP** Architecture (source: Salesforce BLIP paper)

4 Cross-Modal Input Handling

CLIP

CLIP processes cross-modal inputs (image and text) using the following method:

- **CLIP** uses two separate encoders: one for images (e.g., ResNet or ViT) and another for text (a Transformer).
- Both the image and text are encoded independently into fixed-length embedding vectors.
- These embeddings are projected into a shared multimodal space.
- A contrastive loss function is used to maximize the similarity of correct (image, text) pairs while minimizing similarity of incorrect pairs.
- Cross-modal understanding is achieved by comparing distances in the shared embedding space.
- No explicit attention is exchanged between modalities — alignment is purely through embedding similarity.

BLIP

BLIP handles cross-modal inputs in a more integrated way:

- **BLIP** uses a vision encoder (ViT) to extract visual features and a language model (BERT) to handle text.
- A special module called Q-Former is used to learn query tokens that attend to the visual features.
- These query tokens act as a bridge, allowing the language model to selectively interact with visual information.
- The model can operate in two modes: vision-to-language (e.g., captioning) and language-to-vision (e.g., VQA).
- Image and text features are fused using attention mechanisms, allowing deep cross-modal reasoning.
- **BLIP** supports both understanding and generation, enabling it to produce textual responses based on image content.

5 Comparison Table

Feature	**CLIP**	**BLIP**
Developed by	OpenAI	Salesforce
Modalities Supported	Image + Text	Image + Text
Training Objective	Contrastive Learning	Image-Text Matching + Generation
Uses Transformer?	Yes (dual encoders)	Yes (ViT + Q-Former + BERT)
Applications	Retrieval, Classification	Captioning, VQA, ChatBots
Cross-modal Alignment	Embedding similarity	Query transformer attention
Can Generate Text?	No	Yes

Table 1: Comparison between **CLIP** and **BLIP**

6 Conclusion

While both **BLIP** and **CLIP** are capable of handling vision and language tasks, their approaches are quite different. **CLIP** is simpler and more focused on contrastive learning and matching, whereas **BLIP** can actually generate text, making it more useful in generative multimodal applications. Choosing one over the other really depends on the exact task.

References

- Radford, A., et al. (2021). *Learning Transferable Visual Models From Natural Language Supervision.* OpenAI. `https://arxiv.org/abs/2103.00020`
- Li, J., et al. (2022). *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation.* Salesforce. `https://arxiv.org/abs/2201.12086`