# Assignment 5.1: Comparison of CLIP and BLIP Multimodal LLMs

## Introduction

Multimodal large language models (LLMs) integrate multiple data types, such as text and images, to perform tasks like image captioning, visual question answering (VQA), and cross-modal retrieval. This report compares two prominent multimodal LLMs, CLIP (Contrastive Language-Image Pre-training) and BLIP (Bootstrapping Language-Image Pre-training), focusing on their architectures, input types, main applications, and cross-modal handling. A comparison table is included, along with references to relevant sources.

## CLIP: Contrastive Language-Image Pre-training

### Architecture

CLIP, developed by OpenAI, employs a dual-tower architecture with separate image and text encoders [1]. The image encoder is a Vision Transformer (ViT) or ResNet, processing images into fixed-length embeddings (e.g., ViT-L/14 divides images into 14x14 patches). The text encoder is a transformer with 63 million parameters (12-layer, 512-wide, 8 heads), using Byte Pair Encoding (BPE, 49,152 vocabulary size), and outputs text embeddings via the [EOS] token after LayerNorm and a linear projection [1]. Both encoders project outputs into a shared embedding space, trained with contrastive loss on 400 million image-text pairs.

### Input Types

• Images: Pixel data, normalized and divided into patches (e.g., 224x224 for ViT) [1].

• Text: Natural language prompts, tokenized with BPE, capped at 76 tokens.

• CLIP does not support audio.

### Main Applications

• Zero-shot Image Classification

• Image-Text Retrieval

• Content Moderation

### Cross-Modal Handling

CLIP aligns image and text embeddings in a shared space using contrastive learning, maximizing cosine similarity for matching image-text pairs while minimizing it for non-matching pairs. This enables tasks like zero-shot classification by comparing image embeddings to text prompt embeddings.

## BLIP: Bootstrapping Language-Image Pre-training

### Architecture

BLIP, developed by Salesforce, uses a Multimodal Mixture of Encoder-Decoder (MED) architecture [3]. It consists of:  
• Vision Transformer (ViT): Encodes images into embeddings using a [CLS] token.  
• Unimodal Text Encoder: A BERT-like transformer trained with Image-Text Contrastive (ITC) loss.  
• Image-Grounded Text Encoder: Uses cross-attention layers to fuse image and text embeddings, trained with Image-Text Matching (ITM) loss.  
• Image-Grounded Text Decoder: Replaces bidirectional self-attention with causal self-attention for generation tasks, trained with Language Modeling (LM) loss [3].

### Input Types

• Images: Pixel data, processed by ViT into patch-based embeddings.

• Text: Natural language, tokenized and processed by the text encoder or decoder.

• BLIP does not support audio.

### Main Applications

• Image Captioning

• Visual Question Answering (VQA)

• Image-Text Retrieval

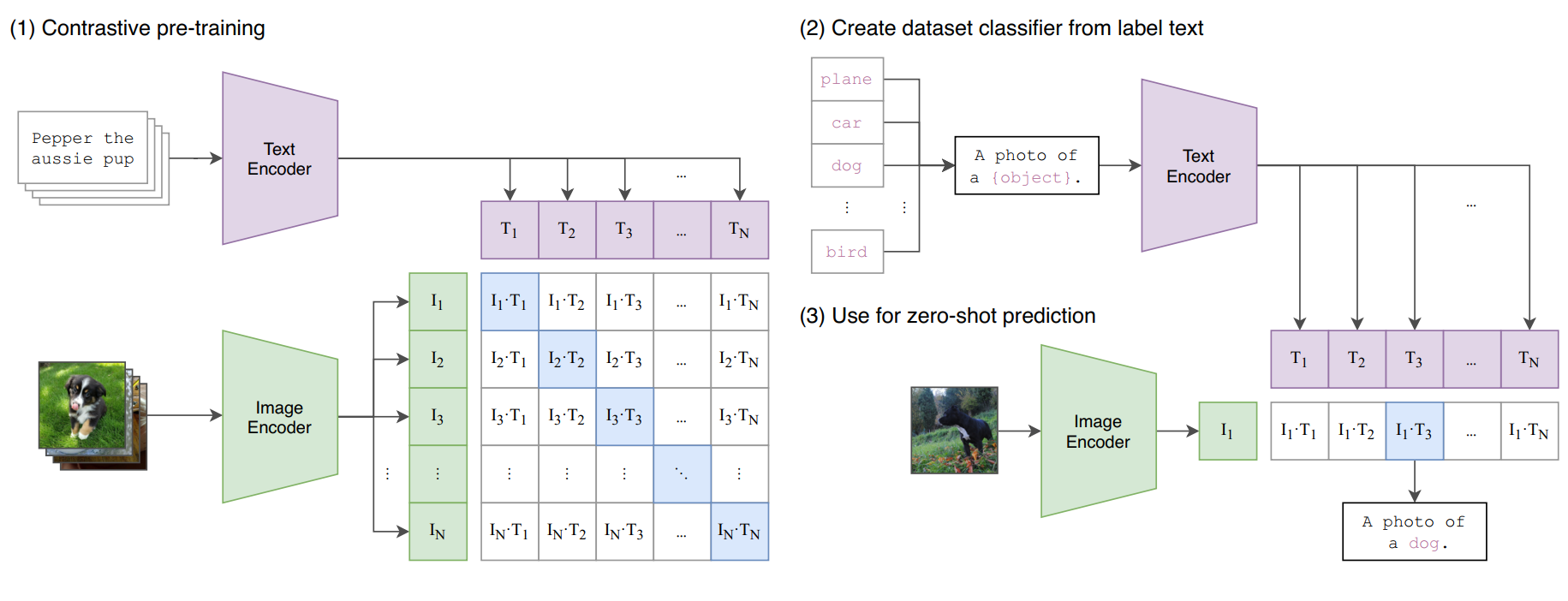
### Cross-Modal Handling

BLIP uses ITC, ITM, and LM losses to align and fuse modalities. ITC aligns image and text embeddings similarly to CLIP. ITM refines alignment via binary classification of matching pairs. LM enables the image-grounded text decoder to generate text conditioned on image embeddings using cross-attention layers [3].

## Comparison Table

|  |  |  |
| --- | --- | --- |
| Feature | CLIP | BLIP |
| Architecture | Dual-tower (ViT/ResNet, transformer) | Multimodal Encoder-Decoder (ViT, text encoder/decoder) |
| Input Types | Images, text | Images, text |
| Main Applications | Zero-shot classification, retrieval, moderation | Captioning, VQA, retrieval |
| Cross-Modal Handling | Contrastive learning (shared embedding space) | ITC, ITM, LM losses; cross-attention |
| Pre-training Data | 400M image-text pairs | Millions of image-text pairs (CapFilt) |
| Strength | Zero-shot flexibility | Unified understanding and generation |

## Architecture diagram

  
  
  
Figure 1: CLIP Architecture

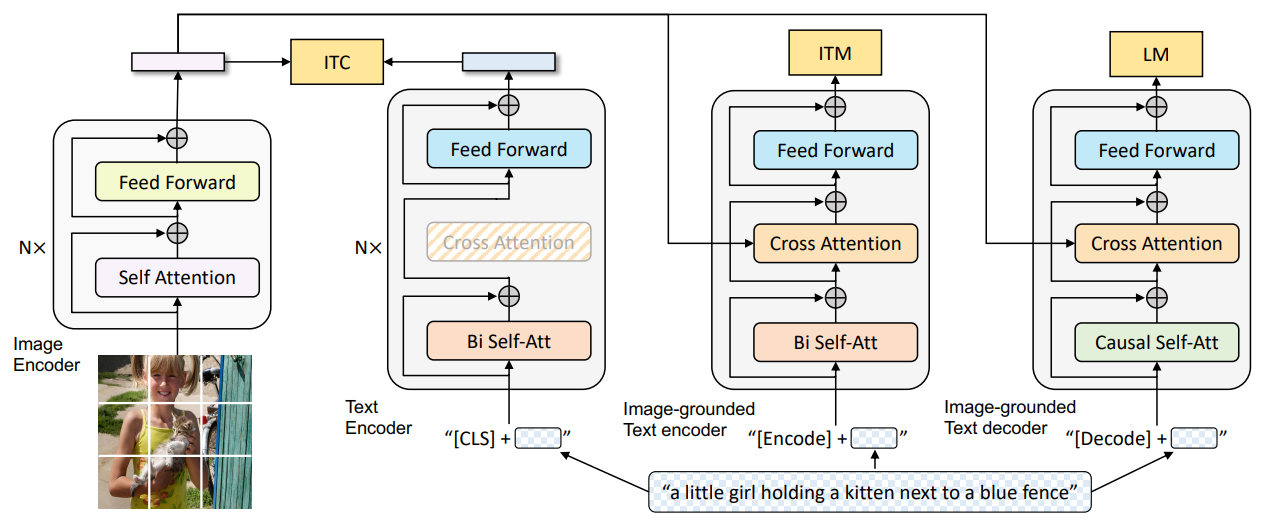


Figure 2: BLIP Architecture

## Conclusion

CLIP and BLIP are powerful multimodal LLMs with distinct strengths. CLIP’s dual-tower architecture excels in zero-shot tasks due to its contrastive learning approach, making it ideal for flexible classification and retrieval. BLIP’s MED architecture supports both understanding and generation, excelling in tasks like image captioning and VQA. CLIP prioritizes simplicity and scalability, while BLIP’s cross-attention enables richer multimodal interactions. Choosing between them depends on the task: CLIP for zero-shot flexibility, BLIP for generative applications.

## References

1. Radford, A., et al. 'Learning Transferable Visual Models From Natural Language Supervision.' arXiv:2103.00020, 2021.

2. 'What is CLIP?' Towards Data Science, 2023.

3. Li, J., et al. 'BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation.' arXiv:2201.12086, 2022.