A Comparative Report on Multimodal LLMs: CLIP vs. BLIP

Introduction

Multimodal Large Language Models (LLMs) represent a significant advancement in artificial intelligence, enabling models to understand and process information from multiple data types, such as text, images, and audio. Unlike traditional unimodal models, they can perform complex tasks that require contextual understanding across different modalities. This report provides a detailed comparison of two influential multimodal models: CLIP (Contrastive Language-Image Pre-training) and BLIP (Bootstrapping Language-Image Pre-training). We will examine their architectures, input types, primary applications, and, most importantly, their distinct methods for handling cross-modal inputs.

Discussion

This section details the architecture and functionality of CLIP and BLIP, highlighting their core differences.

2.1 CLIP (Contrastive Language-Image Pre-training)

CLIP, developed by OpenAI, pioneered a new way of learning visual concepts from natural language supervision. Its core idea is to train a model to recognize which text snippet from a set is the correct caption for a given image.

Architecture: CLIP employs a dual-encoder architecture. It consists of two separate encoders that are trained in parallel:

- Image Encoder: Typically a Vision Transformer (ViT) or a ResNet-based CNN that processes an image and converts it into a fixed-size vector embedding.
- Text Encoder: A standard text Transformer (similar to BERT) that processes a text string and converts it into a vector embedding of the same size.

Input Types: CLIP processes two modalities: images and text.

Cross-Modal Handling: CLIP's method for handling cross-modal inputs is elegant and powerful but indirect. The two encoders project the image and text into a shared, multi-modal embedding space. During training on a large dataset of (image, text) pairs, the model's objective is to maximize the cosine similarity (a measure of closeness) between the embeddings of correct pairs while minimizing it for incorrect pairs. This is known as contrastive learning. The models do not fuse the modalities at a feature level; rather, they learn to align them in a common representational space.

Main Applications:

- Zero-Shot Image Classification: CLIP can classify images into categories it has never explicitly been trained on. For example, to classify a picture of a dog, you can embed the image and text prompts like "a photo of a dog," "a photo of a cat," etc., and choose the text with the highest similarity score.
- Image-Text Retrieval: Finding the most relevant images for a text query or vice versa.
- Guiding Generative Models: CLIP's understanding of image-text alignment is used to steer text-to-image models like DALL-E 2 and Stable Diffusion to generate images that better match a text prompt.

2.2 BLIP (Bootstrapping Language-Image Pre-training)

BLIP, developed by Salesforce Research, was created to address some of CLIP's limitations, particularly the noise in web-scale training data and the lack of deep cross-modal fusion. It introduces a more complex architecture for unified vision-language understanding and generation.

Architecture: BLIP is a more multifaceted model that unifies three core functionalities:

- **Image Encoder:** A Vision Transformer (ViT) to encode images.
- **Text Encoder:** A text Transformer to encode text.
- **Multimodal Mixture of Encoder-Decoder (MED):** This is the key innovation. It's a versatile module that can operate in three modes:
    - **Unimodal Encoder:** Processes text or image features independently.
    - **Image-grounded Text Encoder (Fusion):** Fuses visual and textual features using cross-attention, where text tokens can attend to image patches. This allows for a much deeper, fine-grained understanding of the relationship between image and text.
    - **Image-grounded Text Decoder (Generation):** Generates text based on input image features, enabling tasks like image captioning.

Input Types: Like CLIP, BLIP processes images and text.

Cross-Modal Handling: BLIP employs a more sophisticated, multi-pronged approach to cross-modal interaction:

1. **Contrastive Loss (like CLIP):** It learns a shared embedding space for coarse-grained alignment.
2. **Image-Text Matching (ITM) Loss:** This is a crucial addition. After initial encoding, the multimodal encoder (MED) explicitly fuses the image and text features and performs a binary classification task: does this text truly

match this image? This forces the model to learn fine-grained, token-level alignment.

3. **Language Modeling (LM) Loss:** The model is trained to generate captions for an image, further strengthening its ability to ground text in visual information.

Main Applications:

1. **Image Captioning:** Generating descriptive text for an image.
2. **Visual Question Answering (VQA):** Answering questions about the content of an image.
3. **Improved Image-Text Retrieval:** Achieves better performance than CLIP due to its fine-grained matching capability.
4. **Text-to-Image Generation and Editing:** Provides a strong foundation for generative tasks.

Diagram: Architectural Comparison

The following diagram illustrates the fundamental architectural difference between CLIP's parallel encoders and BLIP's introduction of an explicit fusion/decoder module.

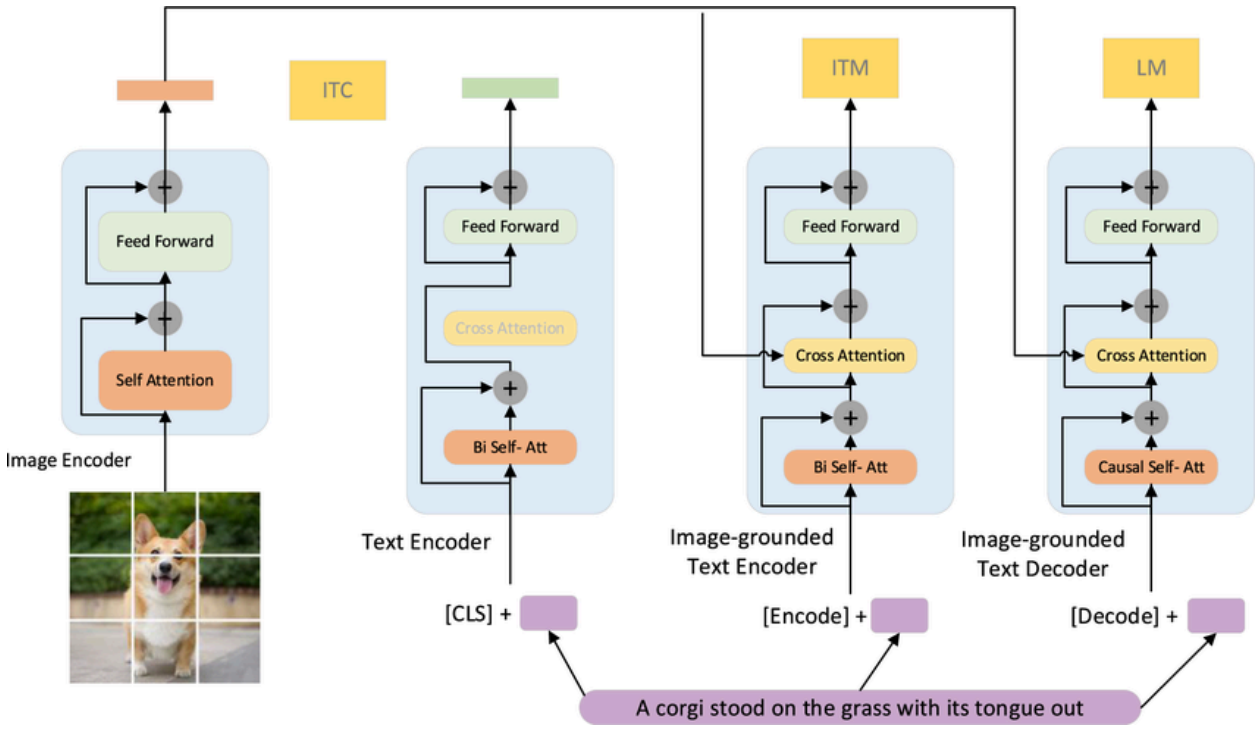

fig:BLIP(Bootstrapping Language-Image pre-training for Unified vision language)

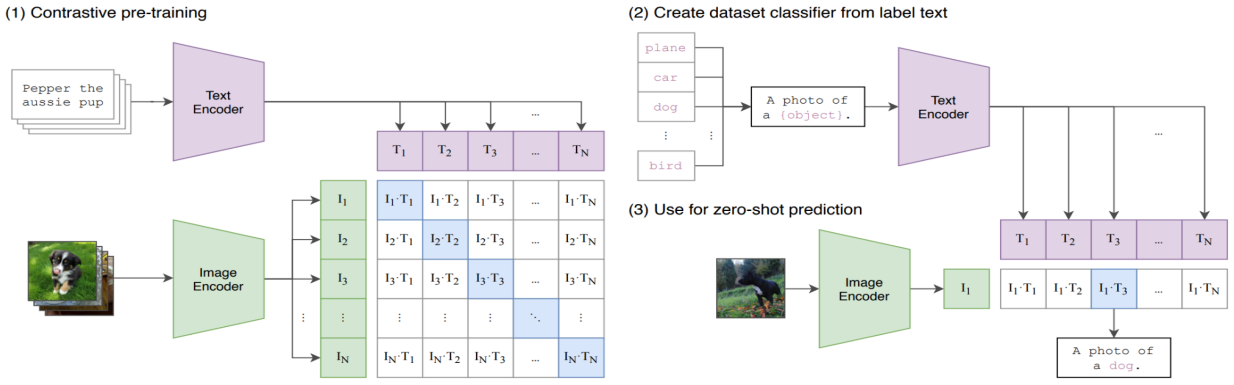

Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

fig.CLIP(Contrastive Language-Image Pre-training)

Comparison Table

Feature	CLIP (Contrastive Language-Image Pre-training)	BLIP (Bootstrapping Language-Image Pre-training)
Core Architecture	Dual-encoder (Image Encoder + Text Encoder).	Unified encoder-decoder (Image Encoder + Text Encoder + Multimodal Fusion/Decoder).
Cross-Modal Handling	Contrastive Learning: Aligns image and text in a shared embedding space. No direct feature fusion.	Multi-task Learning: Uses contrastive loss, explicit feature fusion (cross-attention) for matching, and a decoder for generation.
Primary Task Type	Understanding / Retrieval.	Unified Understanding and Generation.
Key Applications	Zero-shot classification, image-text retrieval, guiding generative models.	Image captioning, Visual Question Answering (VQA), enhanced retrieval.
Key Innovation	Training vision models with natural language supervision at a massive scale.	Introducing explicit multimodal fusion and generative pre-training objectives to improve fine-grained alignment.

Summary

CLIP and BLIP are both foundational models in vision-language research, but they represent different stages of evolution. CLIP established a powerful and efficient paradigm for learning joint image-text representations through contrastive learning in a shared embedding space. Its strength lies in its simplicity and effectiveness for zero-shot recognition and retrieval tasks.

BLIP builds directly on this foundation. It acknowledges the power of contrastive learning but enhances it by introducing modules for explicit, deep fusion of modalities through cross-attention (Image-Text Matching) and for generative tasks (Language Modeling). This more complex architecture allows BLIP to perform a wider range of tasks, including generation (captioning) and fine-grained reasoning (VQA), often with superior performance.

Conclusion

In conclusion, the choice between CLIP and BLIP depends on the specific application. CLIP is a revolutionary model that remains highly effective and computationally efficient for tasks requiring robust, general-purpose image-text alignment, such as zero-shot classification and retrieval.

BLIP represents a more advanced and versatile successor, offering deeper multimodal understanding and generative capabilities. Its introduction of explicit fusion and generative objectives makes it the preferred model for complex tasks like VQA and high-quality image captioning, where fine-grained alignment between visual and textual elements is critical. BLIP's architecture laid the groundwork for even more capable models that followed, such as BLIP-2 and InstructBLIP.

References

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). *Learning Transferable Visual Models From Natural Language Supervision.* Proceedings of the 38th International Conference on Machine Learning (ICML). arXiv:2103.00020

Li, J., Li, D., Savarese, S., & Hoi, S. (2022). *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation.* Proceedings of the 39th International Conference on Machine Learning (ICML). arXiv:2201.12086