Comparison Report: CLIP and Whisper Multimodal LLMs

Introduction

This report compares two multimodal large language models (LLMs): CLIP (Contrastive Language–Image Pretraining) and Whisper. Both models handle different types of data and serve distinct purposes. Below, we detail their architectures, input types, main applications, and approaches to cross-modal processing, followed by a comparison table.

CLIP (Contrastive Language–Image Pretraining)

Architecture

CLIP, developed by OpenAI, consists of two primary components: a vision transformer (ViT) or convolutional neural network (e.g., ResNet) for image processing and a transformer-based text encoder for text processing. The model is trained using contrastive learning, where it learns to align image and text embeddings in a shared latent space by maximizing the similarity of matching image-text pairs and minimizing the similarity of non-matching pairs.

Input Types

- **Image**: Processes images of varying resolutions, typically normalized to a fixed size.
- **Text**: Handles natural language descriptions, captions, or prompts.

Main Applications

- **Zero-shot image classification**: Classifies images without task-specific training by matching image embeddings to text prompts.
- **Image-text retrieval**: Retrieves relevant images given text queries or vice versa.
- **Image captioning and generation guidance**: Provides embeddings for tasks like generating or ranking captions.
- **Object detection and segmentation**: Supports tasks by providing aligned image-text representations.

Cross-Modal Handling

CLIP processes images and text independently through their respective encoders, producing embeddings in a shared latent space. The contrastive loss ensures that corresponding

image-text pairs (e.g., an image of a dog and the caption "dog") have similar embeddings, enabling tasks like zero-shot classification by computing cosine similarity between image and text embeddings.

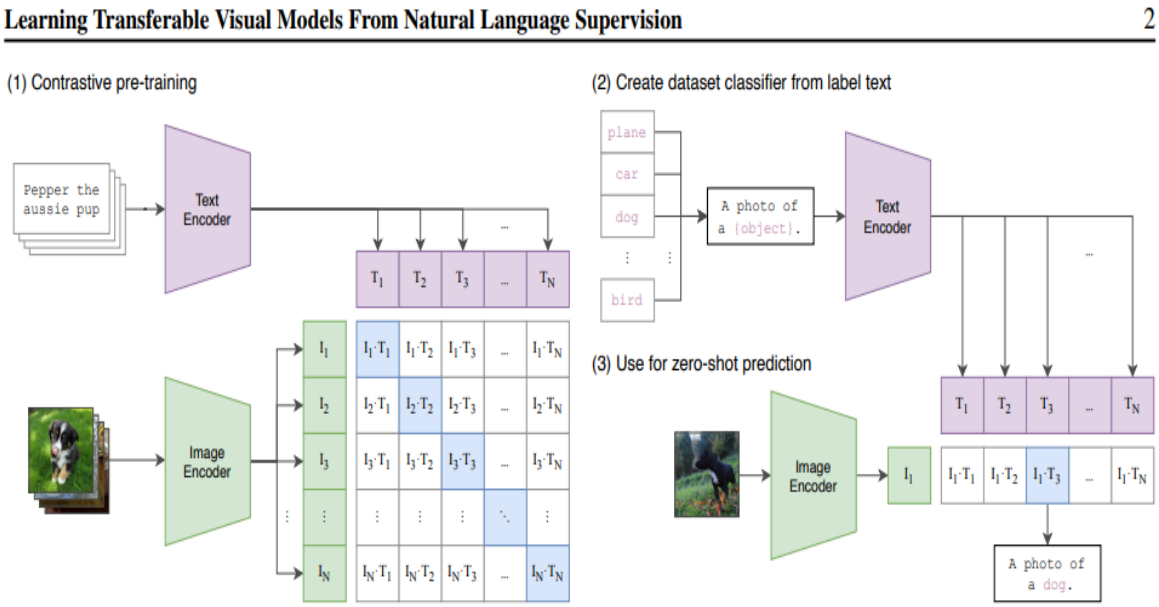

Fig: CLIP Architecture

Whisper

Architecture

Whisper, also developed by OpenAI, is an automatic speech recognition (ASR) system based on an encoder-decoder transformer architecture. The encoder processes audio inputs, while the decoder generates text transcriptions. Whisper is trained on a large, diverse dataset of multilingual audio-text pairs, enabling robust performance across languages and tasks.

Input Types

- **Audio**: Processes raw audio waveforms, typically sampled at 16 kHz.
- **Text**: Outputs transcriptions or translations as text; can also accept text prompts for tasks like language identification.

Main Applications

- **Speech-to-text transcription**: Converts spoken language to text with high accuracy across multiple languages.
- **Speech translation**: Translates spoken content from one language to text in another.
- **Language identification**: Detects the language of spoken audio.
- **Voice activity detection**: Identifies speech segments within audio.

Cross-Modal Handling

Whisper primarily focuses on audio-to-text modality conversion. The encoder transforms audio spectrograms into a latent representation, and the decoder generates corresponding text. While Whisper does not align audio and text in a shared embedding space like CLIP, it handles cross-modal tasks by directly mapping audio features to text outputs through end-to-end training.

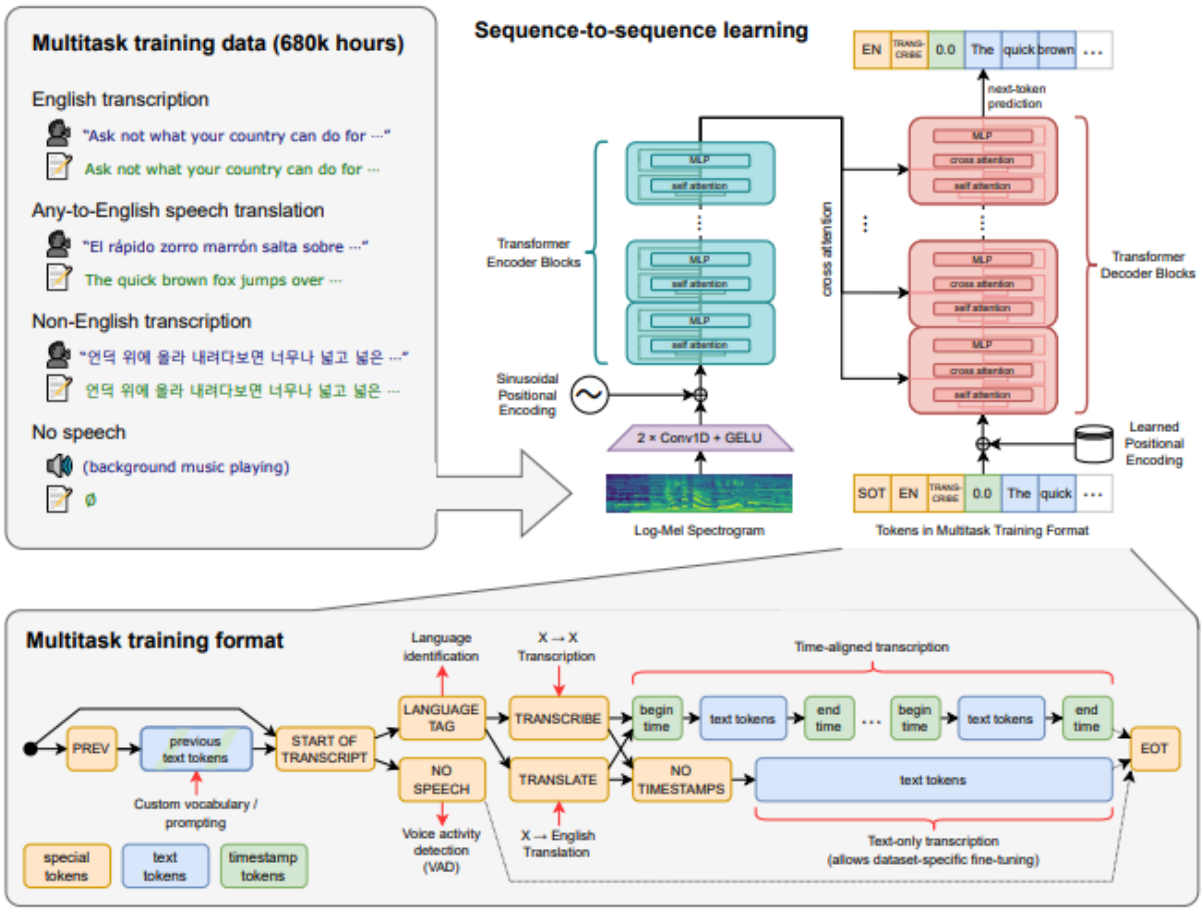

Fig: Whisper Architecture

Comparison Table

Feature/Model	CLIP	Whisper
Developer	OpenAI	OpenAI
Input Types	Image, Text	Audio, Text (output)
Architecture	Vision Transformer/ResNet + Text Transformer	Encoder-Decoder Transformer
Cross-Modal Approach	Shared embedding space via contrastive learning	Audio-to-text mapping via encoder-decoder
Main Applications	Zero-shot classification, image-text retrieval	Speech transcription, translation
Training Data	Image-text pairs (e.g., web data)	Multilingual audio-text pairs
Output	Embeddings for images and text	Text transcriptions/translations

Key Differences

- **Modalities**: CLIP handles image-text pairs, focusing on vision-language alignment, while Whisper processes audio-to-text, focusing on speech recognition and translation.
- **Cross-Modal Strategy**: CLIP aligns modalities in a shared embedding space for flexible retrieval and classification, whereas Whisper performs direct modality conversion (audio to text).
- **Applications**: CLIP excels in vision-language tasks like zero-shot classification, while Whisper is specialized for audio-related tasks like transcription and translation.

Conclusion

CLIP and Whisper are powerful multimodal LLMs tailored to different domains. CLIP's strength lies in its ability to align images and text for versatile vision-language applications, while Whisper excels in converting audio to text across diverse languages and tasks. Their distinct architectures and cross-modal approaches make them complementary tools in the multimodal AI landscape.

References

- Radford, A., et al. (2021). "Learning Transferable Visual Models From Natural Language Supervision." *arXiv:2103.00020*. https://arxiv.org/abs/2103.00020
- Radford, A., et al. (2022). "Robust Speech Recognition via Large-Scale Weak Supervision." *arXiv:2212.04356*. https://arxiv.org/abs/2212.04356