Part III Project Proposal

Supervised by Chenliang Zhou and Cengiz Öztireli

The Contrastive Language-Image Pretraining (CLIP) algorithm employs a dual-encoder architecture, consisting of an image encoder and a text encoder, which are trained in tandem to produce embeddings that exhibit cosine similarity when the corresponding images and texts are semantically similar. This innovative approach serves as a foundation for a multitude of applications at the intersection of computer vision and natural language processing, including but not limited to, image captioning, object detection, text-guided image manipulation, and generative image tasks. The projects described in this proposal are intrinsically related to the capabilities and applications of the CLIP algorithm.

Analysis of CLIP space

CLIP encodes images and texts into a joint embedding space. The primary objective of this project is to conduct an exhaustive exploration and analysis of this joint embedding space, with a particular focus on the semantics of its subspaces and manifolds. For example, one intriguing avenue of inquiry could involve the formulation of an "object subspace," generated by the embeddings of multiple 2D renderings of the same 3D object captured from varying views. The research aims to investigate the inherent properties of such an object subspace. For example, we seek to discover the unique characteristics of the object encapsulated by this subspace and to explore the feasibility of leveraging this subspace for specialized tasks, such as style transfer (same styles; different objects) or style changing (different styles; same objects)

Improvement of CLIP

Despite the robust capabilities of CLIP, certain limitations persist. Notably, CLIP-PAE has demonstrated that intra-modality similarity metrics, such as text-to-text or image-to-image similarity, are consistently higher than inter-modality similarity metrics, such as text-to-image similarity. This discrepancy introduces complications for a range of downstream applications. The primary goal of this research project is to systematically identify and analyze this or other limitations within the CLIP framework and to subsequently develop targeted solutions. Potential avenues for remediation may include modifications to the distance metric during training (original CLIP uses cosine distance) or to the underlying neural network architecture.

Extensions of CLIP to other modalities

CLIP serves as a compelling testament to the efficacy of contrastive training frameworks in bridging the semantic gap between textual and visual modalities. Building upon this foundational work, the proposed research seeks to extend the applicability of the CLIP framework to additional modalities, including threedimensional objects, audio, video, etc. Such extensions could be realized through either a direct approach, which involves the development of specialized encoders for the new modality (e.g., Text2Shape, CLIP Goes 3D), or an indirect approach that establishes a connection between the new modality and existing image/text modalities through specific methodologies (e.g., DreamFusion and GET3D, which connect 3D objects and 2D images by rendering).

By leveraging an extended version of the CLIP framework, the research also aims to facilitate a wide array of text-driven tasks and applications across these new modalities, analogous to the image-based tasks enabled by the original CLIP algorithm.

Applications of CLIP at the text side

While the majority of extant models that incorporate CLIP framework predominantly focus on tasks related to the image domain, such as image generation and manipulation, there exists a compelling avenue for exploration in the textual domain as well. Specifically, the proposed research aims to investigate the feasibility of leveraging the CLIP algorithm for text-centric tasks, such as text generation. The central question to be addressed is whether the CLIP framework, traditionally employed for image-related tasks, can be effectively adapted to facilitate sophisticated text generation algorithms. This research question is also worth exploring with the new CLIP extended to other modalities described in the previous project.