Graduate School of Science and Technology Master’s Thesis Abstract

Laboratory name (Supervisor)	Augmented Human Communication (Satoshi Nakamura　(Professor))		
Student ID	2111421	Submission date	2024 / 1 / 19
Name	LEE SANGMYEONG		
Thesis title	Improving the Image Discrimination Ability for CLIP-Model via Linguistic Structural Information		

Abstract

The Contrastive Language-Image Pre-training (CLIP) model excels in uniting vision and language through extensive training and a contrastive learning approach, endowing it with remarkable zero-shot capabilities. However, its exclusive dependence on text inputs poses a challenge in handling structural ambiguity, as a single sentence can encompass multiple meanings, complicating the model’s ability to discriminate suitable vision and language pairs. Our research hypothesises that incorporating linguistic formalism as an input has potentials to enhance CLIP’s ability to grasp the semantic nuances of language, particularly in addressing structural ambiguity. To examine this hypothesis, we adapted and combined four natural language processing techniques with diverse formalisms and encoding methodologies into the CLIP model. Our experiments demonstrated the effectiveness of these formalism encoding techniques in enhancing CLIP’s image discrimination ability, particularly in text-to-image retrieval tasks. Additionally, we used gradient-based methods to gain insights into how linguistic formalism is interpreted within the model’s architecture. This study provides a fresh perspective on augmenting CLIP’s language understanding capabilities, specifically in handling structural ambiguity. Our findings contribute to diverse fields applying CLIP, offering valuable insights for practical applications.