ABCI 3.0開発加速利用 （2025年度）成果概要（公開用）

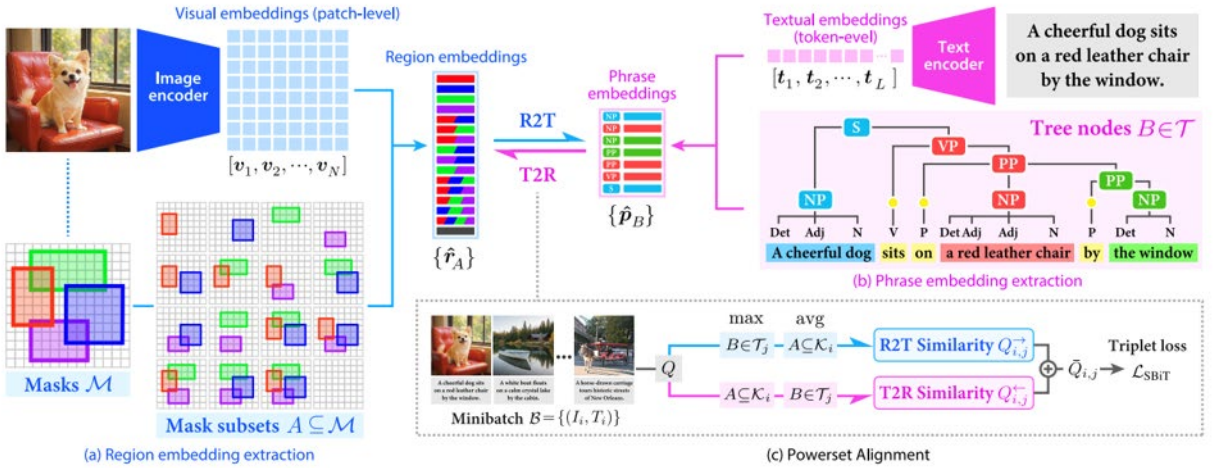
課題名：視覚・言語基盤モデルの開発

実施時期：2025/04/25-2026/3/30
所属機関名：産業技術総合研究所
代表者氏名：横田理央

成果概要：本研究では、CLIP型の視覚言語事前学習で課題だった複数領域にまたがる構成的意味理解を改善するPowerCLIPを提案する。画像領域集合とテキスト構文木の句を網羅的に対応付ける「powerset alignment」により細粒度な整合を学習し、非線形集約器で計算量をO(2^M)からO(M)へ削減し、ゼロショット分類・検索で既存手法を上回った。

成果のポイント：本研究課題の成果は、CLIP型の視覚言語モデルにおける構成的意味理解の弱点に対し、画像中の複数領域と言語中の句構造を対応づける新しい学習方法を示した点にある。従来のCLIPは、画像全体とキャプション全体を対応させる大域的な対照学習を基本としている。そのため、「赤い椅子の上に犬がいる」のように、物体、属性、位置関係が複雑に組み合わさる表現では、どの画像領域がどの言語表現に対応しているかを十分に学習しにくいという課題があった。

本研究課題では、この課題を解決するために、PowerCLIPという手法を提案している。PowerCLIPの中心的な考え方は、画像を複数の領域に分割し、それらの単独領域だけでなく、領域の組合せ全体を学習対象とすることである。右図は、この考え方を説明するうえで有用である。画像側で複数の領域マスクが作られ、それらの組合せが候補として生成される。一方、テキスト側では、キャプションが構文木に基づいて句へ分解される。そして、画像領域集合から対応する句を探す方向と、句から対応する画像領域集合を探す方向の両方で類似度を計算する。この双方向の対応づけにより、モデルは画像全体と文全体の関係だけでなく、部分的かつ構成的な意味の対応を学習できる。



成果についてより詳細な情報を提供しているWebページ、発表論文などの情報：

Masaki Kawamura, Nakamasa Inoue, Rintaro Yanagi, Hirokatsu Kataoka, Rio Yokota, PowerCLIP: Powerset Alignment for Fine-Grained Contrastive Pre-Training, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2026.