



Title	Exploring Contrastive Learning in Foundation Model Pre-training and Applications
Author(s)	Pang, Zongshang
Citation	大阪大学, 2025, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/103165
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

<https://ir.library.osaka-u.ac.jp/>

The University of Osaka

Abstract of Thesis

Name (Zongshang Pang)	
Title	Exploring Contrastive Learning in Foundation Model Pre-training and Applications (基盤モデルの事前学習と応用におけるコントラスト学習の探究)

Abstract of Thesis

Contrastive learning, when scaled and adapted judiciously, offers a unifying objective for training foundation models that generalize from visual signals to multi-modal domains. This thesis shows how three complementary projects push that premise across granularity, supervision, and modality. In the first project, we propose a self-supervised pixel-level contrastive learning framework, PixCon. It investigates the effective components of current dense contrastive learning frameworks and systematically integrates them with a novel semantic reweighting mechanism, which enables simple pixel-level learning to outperform complex region-level approaches on dense visual prediction tasks such as detection and segmentation. In the second project, we propose a training-free zero-shot video summarizer by reformulating the classic diversity and representativeness video summarization heuristics as quantifiable scores based on contrastive losses, entirely eliminating task-specific fine-tuning while outperforming supervised baselines on TVSum and SumMe. In the third project, we investigate the potential of discriminative contrastive learning on generative models such as large language models on multi-modal video applications. Concretely, we propose the S2L framework that couples a video large language model that can perform a query-focused summarization task with a contrastive grounding module, transforming textual form summaries into precise timestamps and achieving new best results on ETBench localisation tasks. Collectively, these studies demonstrate that carefully engineered contrastive objectives can endow a wide spectrum of benefits on the pre-training and the applications of foundation vision and vision-language models.

論文審査の結果の要旨及び担当者

氏 名 (Zongshang Pang)			
	(職)	氏 名	
論文審査担当者	主 査	教授	中島悠太
	副 査	教授	長原一
	副 査	教授	河原吉伸
	副 査	准教授	千葉直也

論文審査の結果の要旨

本学位論文は、正例と負例の比較に基づいてモデルの学習を実現するContrastive Learningと呼ばれる学習手法を利用した様々な応用に関するものであり、画像の画素レベルでの表現学習（第2章）、映像要約への適用（第3章）、大規模言語モデルによる部分映像検索への適用（第4章）に関する研究成果に基づく。

第2章では、画像の画素レベルのContrastive Learningによる画像表現の獲得について論じている。画像表現の獲得には、画像の小領域や画素を単位としたContrastive Learningが提案されている。これらの手法では、同一の画像から取り出された異なる領域についてそれぞれランダムな画像処理を適用し、その中で対応する小領域や画素が同じ表現となるように学習する。画素を単位とするアプローチでは、局所的なコンテキストしか獲得できない、取り出された領域間で重複しない部分是对応する画素が見つけれられないという問題があり、領域を単位とするアプローチが研究されてきた。本学位論文では、画素を単位とするアプローチにおける前述の問題点が解消可能であることを示しており、その妥当性も実験的に明らかにしている。実際に、画素を単位とする提案手法は、領域を単位とする最新手法に対して複数の下流タスクで性能的に優位であることを示した。

第3章では、映像要約におけるContrastive Learningの応用可能性について述べている。映像要約とは、長時間の映像に対して、その一部を切り出すことにより内容を可能な限り保持しつつ短くするタスクで、内容をどの程度保持しているかを測るための基準が異なる様々な手法が提案されている。多くの手法は、元映像と要約映像のペアから、教師あり学習により映像の各部の重要度を予測するモデルを学習し、そのモデルを利用して重要な部分のみを要約映像に含めるアプローチを採用している。教師あり学習は学習時の映像ドメインに大きく依存することから、他のアプローチとして映像フレーム間の類似度計算による冗長性の削減に基づく手法も提案されている。本学位論文では、後者のアプローチを採用し、Contrastive Learningが映像の類似性やユニーク性を自然に表すことを示し、この知見に基づく手法を提案した。実験では、デファクトスタンダードとして利用される二つのベンチマークデータセットで、教師あり学習の最新手法には及ばないものの、他の教師あり学習や教師なし学習よりも高い性能を持つことを示した。

第4章では、大規模言語モデルを利用した部分映像検索の新しいアプローチを提案している。部分映像検索とは、自然言語テキストをクエリとして長時間の映像からクエリに対応する部分映像を見つけ出すタスクであり、既存手法の多くは、クエリが表すイベントとそれ以外の部分の境界をモデル化することにより、性能向上を目指してきた。本学位論文では、クエリと映像から、映像中のクエリに対応する部分に関する要約を自然言語テキストで出力し、そのテキストと映像を比較することにより、検索を実現している。提案手法はクエリが表すイベント自体に注目したアプローチであり、既存手法とは大きく異なるものである。テキストと映像の比較には、Contrastive Learningにより学習したモデルを利用している。実験では、複数のベンチマークデータセットにおいて既存手法を大きく上回る性能を示しており、部分映像要約に関する新しい知見を与えるものであると考える。

以上のとおり、本学位論文はContrastive Learningの性質を活用することによってその新たな応用を開拓しており、情報科学分野に大きく寄与するものである。このことから、博士（情報科学）の学位論文として価値のあるものと認める。