DS-12

大規模画像言語モデルを活用した Coarse-to-Fine型セマンティックシーン検索

立花 卓遠[1, 2], 坂口 翼[1], 大西 正輝[1], 櫻田 健[1, 3]

[1] 産業技術総合研究所, [2] 東京大学, [3] 京都大学

概要

- 空間動画像データベース(SAI-DB)に対し、高精度かつ高速な、Coarse-to-Fine型セマンティック検索システムを提案・実装
- Coarseステップにおいて、ユーザのクエリ文と、画像から生成された説明文をそれぞれベクトルに変換し、類似度を計算
- Fineステップにおいて、GPT-4oを用いたVQA を利用することで、インタラクティブにクエリと一致する画像を取得

背景

- 現状のシーン理解では深層学習を活用したタグ付けやクラスリングが主流
- 街の活気や雰囲気といった、高次なシーン理解は未達成
- 大規模画像言語モデルの発達に伴い、より人間の理解に近い形でのシーン記述が可能に

実装・提案

- シーン記述を元にしたシーン画像の自然言語クエリ
- クエリ文のばらつきに堅牢な、Coarse-to-Fine型セマンティック検索をSAI-DBに統合
- インタラクティブなクエリを実行可能

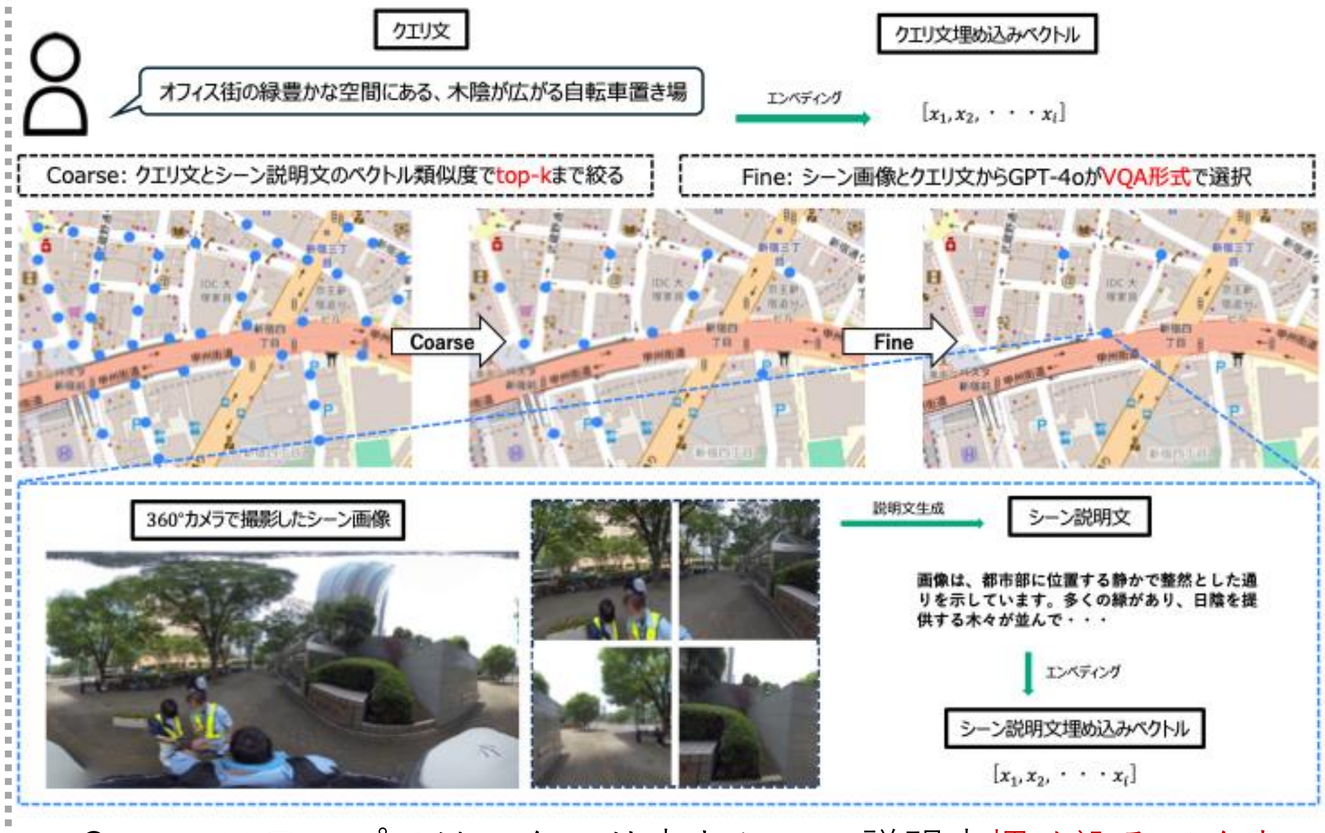
SAI-DB

- 大規模画像言語モデル(GPT-4o)で作成したシーン説明文、大規模言語モデル(text-embedding-3-large)で作成したシーン説明文埋め込みベクトルを登録

> あなたの仕事は、360 度のパノラマビューから撮影された4つの画像について説明することです。添付画像は、正面、左、背面、右の視点を示すように切り取られています。添付画像のシーンについて、何をするのに適した雰囲気の場所か説明してください。

↑説明文生成プロンプトの例。 cubemap形式で大規模画像言語モデルに画像を入力

提案手法



- Coarseステップでは、クエリ文とシーン説明文埋め込みベクトルのコサイン類似度で top-k を取得
- Fineステップでは、クエリ文と合致するシーン画像を大規模画像言語モデル(GPT-4o)がVQA形式で選択　↓VQAプロンプトの例

> {クエリ文} というクエリに一致する画像を選んで、クエリと合致する順に並び替えてください。また、その写真がどうしてクエリと一致すると考えたのか、詳述してください。クエリに適合しないと考える画像は説明不要なので回答から除外してください。

検索例

検索例の考察

- 街の雰囲気や特定の特徴を持った駐輪場を理解(①、③)
- 「特に木陰が自転車置き場に影響を与えている様子が見受けられる」という選択理由は、特に理想的なクエリ(③)
- 駐輪場に木陰がなく、道端の木陰と混同してしまっていた結果も存在(②)
- Fineステップで除外されたシーンにも「木陰のある自転車置き場」は存在

提案手法の優位性

- top-k の絞り込みにより、膨大なシーン画像データに対するクエリを、一定の精度を保ちつつ迅速に実行
- VQAの実行により、明らかに不適切なシーンを除外。存在しない状況のクエリに対して堅牢性を取得
- シーン選択理由を大規模画像言語モデルに説明させることで、解釈性が向上

展望

- 精度やUXの定量評価
- 登録データ数の増加に伴うスケーラビリティの検証
- 埋め込みモデルの変更、ファインチューニングを検討