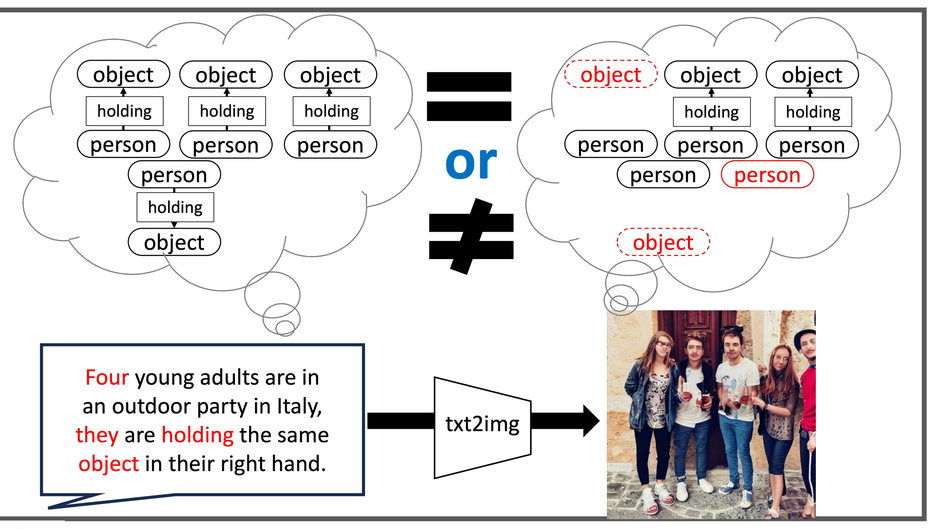
IS2-85

テキストからの画像生成の評価指標としてのシーングラフ類似度

濱田 裕太　品川 政太朗　中村 哲
奈良先端科学技術大学院大学

研究の概要

課題
画像生成では空間的構造を測る指標が不足
解決法
シーングラフ同士の比較を利用した評価指標を提案
結果
・ベースラインと提案手法では差が現れなかった
・シンプルなシーングラフ検出と比較では限界あり



1. 背景

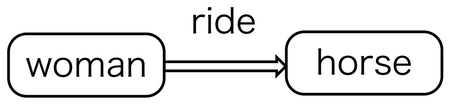
画像生成には空間的構造が重要。しかし…

・画像生成では品質と多様性を重視
　例：Fréchet Inception Distance, Inception Score

・空間的な構造を評価する指標は文や画像全体を特徴量として埋め込むものが多い
　例：R-precision, Positional Alignment [1]

より直接的な関係(シーングラフ)による生成画像の定量的な評価は可能なのか？

2. シーングラフ

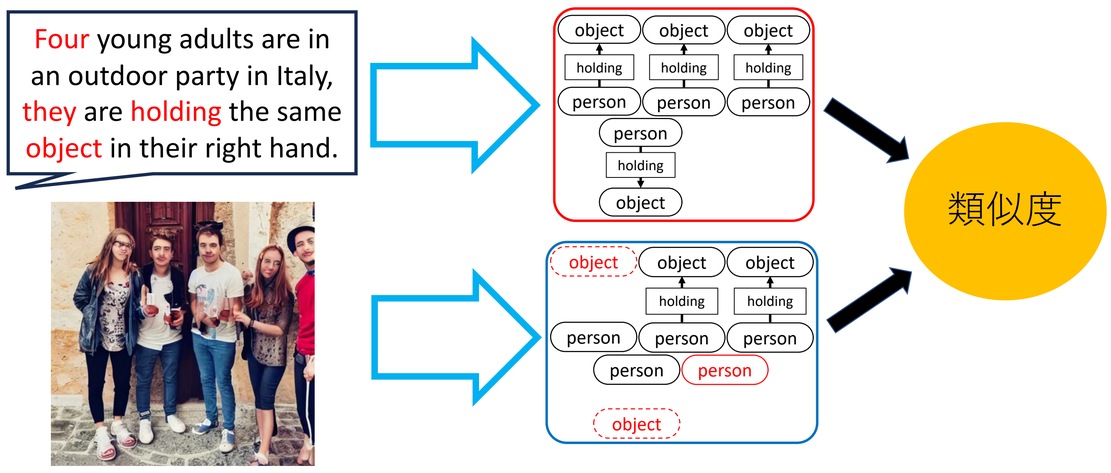


・画像に映るものをグラフで表す表現法

・テキストと画像の**両方**から生成可能

3. 提案手法：Recall@K (R@K)

1. 入力文と画像からシーングラフを生成
2. **グラフ間の類似度**を関係を表すスコアとして利用



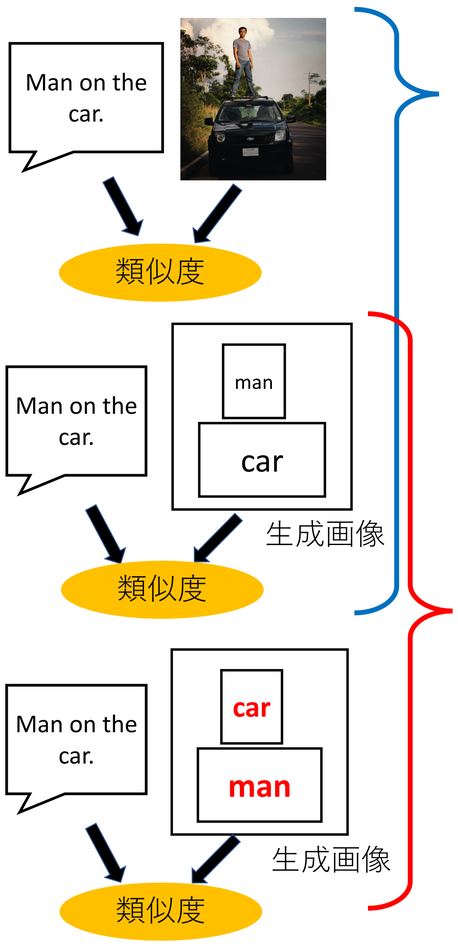
⇨類似度：Recall@K $= \frac{size(S_{T_t} \cap S_{T_i})}{size(S_{T_t})} \times 100$

入力文中の関係のうち**同じ関係**が画像からも検出された割合

4. 実験

シーングラフは物体間の関係(レイアウト)が異なる画像の評価に有効か？

準備: 1つのテキストに対して画像を3種類用意



生成画像でも実画像と同じような評価値になるか？

・テキストと画像の組から類似度を算出

・生成画像は実画像に付与されている物体のバウンディングボックスから画像生成

・画像生成の品質が十分であれば評価値はあまり変わらないはず

レイアウトを人為的に真値と異なるものにした場合評価値は変わるか？

・生成条件であるバウンディングボックスのラベルをランダムに入れ替えた生成画像を、元のレイアウトの生成画像と比較

・R@Kは大きく影響を受けるはず

5. 実験設定

データセット：Visual Genome Dataset
画像 (2000枚)、説明文、レイアウト (Bounding Box)
画像生成：Layout-to-Imageモデル（TwFA [2]）
ベースライン：CLIP特徴量のコサイン類似度

6. 結果

画像	レイアウト	R@20	R@50	R@100	CLIP
実画像	-	3.58	6.83	8.98	0.29
生成画像	True	↘ 1.77	↘ 4.07	↘ 5.75	↘ 0.26
生成画像	False	↘ 1.11	↘ 2.53	↘ 3.42	↘ 0.24

1. CLIPとRecall@K共に**実画像＞True＞False**の順
　差がない原因：ラベル入替による**生成画像の品質低下**

2. R@Kが最大で8.98
　⇨類似度算出方法か、シーングラフの生成方法のどちらかの改善が必要

7. まとめと今後の課題

R@Kでは生成画像の評価に有効とはいえない

1. 実験に用いる生成画像の品質低下の軽減
2. シーングラフの比較法の変更
　例：GNNによるグラフエンベディング

参考文献

[1] Tan M, et.al., “TISE: Bag of Metrics for Text-to-Image Synthesis Evaluation”, 2022.
[2] Zuopeng Y, et.al. , “Modeling Image Composition for Complex Scene Generation”, 2022.