第1章　関連研究

ここに関連研究を紹介

1.1　シーングラフ生成モデル

本節では, シーングラフ生成モデルの歴史と代表的な手法について述べる.

シーングラフ生成（Scene Graph Generation: SGG）は, 与えられた画像からオブジェクトを検出し, それらのオブジェクト間の関係を予測することを目的とするタスクである. 各オブジェクトをノード, 関係をエッジとして表すことで, 画像の内容を構造的かつ解釈可能な形で記述できる点が特徴である. シーングラフ表現は, 画像キャプション生成や Visual Question Answering（VQA）, 画像検索などの様々な下流タスクにおいて有用であることが示されている.

1.1.1　シーングラフ生成の歴史

シーングラフ生成（Scene Graph Generation: SGG）は, 画像内のオブジェクト間の意味的関係を構造的に表現することを目的としたタスクであり, その起源は 2010 年代中頃に登場した視覚的関係認識（Visual Relationship Detection）にさかのぼる. Lu ら [?] は〈主語, 関係, 目的語〉の三つ組で画像を表現する手法を提案し, 物体間の関係理解という新たな方向性を示した. 2017 年には Xu ら [?] によって「Scene Graph Generation」という名称が初めて明確に定義され, 同時期に Visual Genome データセットの公開によって学習環境が整備された。その後, Zellers ら [?] の MotifNet や Yang ら [?] の Graph R-CNN など, 文脈情報やグラフ構造を活用するモデルが登場し, SGG の性能が大きく向上した.

2019 年以降は, オブジェクト間の階層的依存関係を考慮する VCTree[?] などが提案され, 関係推定の精度向上が進んだ. さらに 2021 年以降, Transformer 構造を導入したモデルが主流となり, SGTR[?] や RelTR[?] など, エンドツーエンドでオブジェクト検出と関係推定を同時に行う手法が登場した。その中でも EGTR（End-to-End Graph Transformer）[?] は, オブジェクトと関係を統一的にトークンとして扱うことで高い表現力を実現した. 近年では, CLIP や Grounding DINO などのマルチモーダル事前学習モデルを活用し、未知のカテゴリや関係にも対応可能なオープンボキャ

ブラリ型SGG（Open-Vocabulary SGG）への発展が進んでいる.

以下, 代表的な手法を述べる.

1.1.2 EGTR

1.2 大規模言語モデル

近年, 自然言語処理（Natural Language Procesing: NLP）分野では, 大規模言語モデル（Large Language Model: LLM）と総称される技術群が飛躍的な進歩を遂げている. LLM は, 数千億から数兆に及ぶパラメータを有する巨大なディープラーニングモデルであり, インターネット上の膨大なテキストデータを事前学習することで, 言語の文法構造, 語彙的意味的, 文脈依存性といった多様な知識を内部に獲得している. このようにして得られたモデルは, 自然言語を理解・生成する能力を備え, 単なる統計的言語モデルを超えた汎用的な言語知能として機能する.

LLM は多層のニューラルネットワークを基盤とし, 特にトランスフォーマーアーキテクチャに基づいて構築されている. この構造において中核的な役割を果たす自己注意機構（self-attention mechanism）は, 文中の単語間の関係を動的に捉えることで, 長距離依存関係を含む文脈の理解を可能にしている. 代表的な LLM としては, OpenAI による GTP シリーズや Google の BERT が挙げられ, 文章生成, 質問応答, 翻訳, コード生成など多様なタスクにおいて高い性能を示している.

さらに近年では, ChatGPT の登場により, LLM の社会的な影響力は飛躍的に拡大している. 学術研究や産業応用のみならず, 教育, ビジネス, 創作支援といった日常的な領域にも浸透しつつあり, 人と AI の新しい協働の形を実現する基盤技術として注目を集めている.

1.2.1 LLM の歴史

歴史書く

1.2.2 Transformer アーキテクチャ