【採択論文】

タイトル

Indoor Multi-View Radar Object Detection via 3D Bounding Box Diffusion

著者

谷髙竜馬（三菱電機）, Pu (Perry) Wang（MERL）, Petros Boufounos（MERL）, 高橋龍平（三菱電機）

概要

レーダーによる屋内パーセプション※1は、プライバシーを保護できる点や、火災などの危険な状況下においても高い信頼性を有する点から、室内の見守りや監視用途において注目が高まっています。一方で、レーダーは分解能が低いという課題があり、汎化性能を保った高精度な予測が困難でした。本論文では、レーダーパーセプションの高精度化を目的として、Radar Object dEtection with 3D Bounding boX Diffusion（REXO）※2を提案します。REXO は、複数視点から得られる 2 次元レーダー特徴間の関連付けを、各視点に共通する 3 次元空間上で明示的に行うことで、汎化性能の高い 3 次元バウンディングボックス（BBox）の構築を可能にします。さらに、この 3 次元空間上で BBox に対するデノイジング処理を段階的に適用することで、従来手法を大幅に上回る最高水準の BBox 推定精度を達成しました。

※1 レーダーを用いて周囲の環境や物体を検知・認識する技術。レーダーは電波を使って物体の位置や速度、形状などを測定できるため、視界が悪い状況や暗闇でも高い信頼性を持つ。

※2 レーダーデータを用いた物体検出のための深層学習モデル。

【採択論文】

タイトル

LatentLLM: Activation-Aware Transform to Multi-Head Latent Attention

著者

Toshiaki Koike-Akino(MERL), Xiangyu Chen(MERL), Jing Liu(MERL), Ye Wang(MERL), Pu (Perry) Wang(MERL),Matthew Brand(MERL)

概要

大規模言語モデル（LLM）やマルチモーダルモデル（LMM）は高い性能を有する一方で、計算量やメモリ使用量が膨大であることが、実用化における大きな課題となっています。

本研究では、活性化や注意機構を考慮した新たなモデル圧縮技術「LatentLLM」を開発しました。

本技術は、複数の重み行列を同時にテンソル分解する独自手法により、追加学習を必要とせずに高精度な低次元モデルへの変換を可能にします。

評価実験の結果、高い圧縮率においても、言語理解およびマルチモーダル推論において優れた性能を維持できることを実証しました。

本技術により、AI モデルの省電力化・高効率化が促進され、エッジデバイスを含む幅広い環境での AI 活用および社会実装への貢献が期待されます。

【採択論文】

タイトル

Chain-of-Thought Driven Adversarial Scenario Extrapolation for Robust Language Models

著者

Md Rafi Ur Rashid(Pennsylvania State University), Vishnu Asutosh Dasu (Pennsylvania State University),Ye Wang(MERL), Gang Tan (Pennsylvania State University), Shagufta Mehnaz (Pennsylvania State University)

概要

近年、大規模言語モデル（LLM）はさまざまな分野で活用が広がっていますが、その一方で、悪意ある入力によって有害な情報を引き出される「ジェイルブレーク攻撃」や、誤った情報の生成、社会的バイアスといった安全性の問題が指摘されています。従来の対策では、特定の攻撃にしか対応できない、あるいは過度に回答を拒否してしまい、自然な対話を損なうといった課題がありました。

本研究では、推論時にモデル自身が潜在的な攻撃シナリオを想定し、その場で防御方針を構築する新しい手法 Adversarial Scenario Extrapolation（ASE）を提案しました。ASE は追加の学習やモデル改変を必要とせず、既存の大規模言語モデルにそのまま適用できる点が特長です。

評価実験の結果、ASE はジェイルブレーク攻撃に対してほぼ完全な防御を実現するとともに、事実誤認や社会的バイアスを大幅に低減できることを確認しました。また、安全性を確保しながらも不要な回答拒否を抑え、自然で柔軟な対話を維持できることが明らかになりました。

本成果は、安全性と使いやすさの両立が求められる生成 AI の実運用において、信頼性向上に大きく貢献することが期待されます。