基礎研究 応用研究

N10

脳情報科学

おすすめ

多段階時空間推論による映像質問応答

～映像の内容を理解して質問に答えるAI ～

概要

映像質問応答は、映像とその映像の内容に関する質問が与えられたときに、適切な解答を返す課題です。本研究では、映像に写る人物や物体の詳細な外観情報とそれらの動作情報を質問の内容に応じて統合し、多段階に推論を行う映像質問応答技術を開発しました。

特徴

- ニューラルネットを用いて、質問内容に応じて詳細な外観情報と動作情報を同時に考慮して推論できる時空間推論モジュールを作成しました。
- 時空間推論モジュールを多段階適用することで、高精度に映像の質問応答ができるニューラル質問応答システムを開発しました。
- 標準的な３つの映像質問応答データセット（MSVD-QA, MSRVTT-QA, ActivityNet-QA）において、提案手法は従来手法の性能を大幅に上回りました。

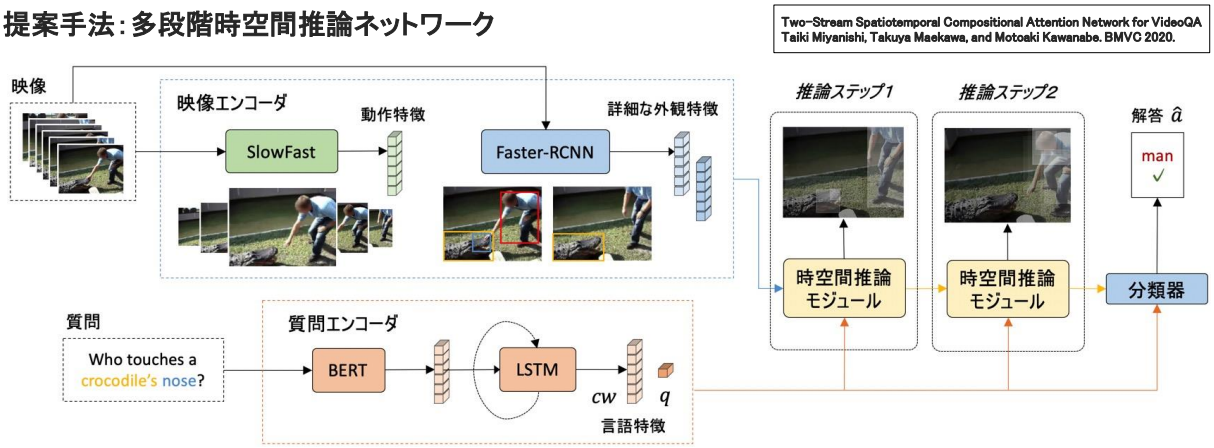
今後の展開

- ホームカメラやロボットビジョンに適用することで、記憶支援・忘れ物や落とし物の検索・人の監視や見守りといった実世界で駆動する知能システムへ応用する予定です。

対コロナへの関連

- 当該技術を監視カメラの映像に適用することで、屋内の密な状況の問い合わせができたり、過去に起きた会話や食事などの映像記録に、言語を介して問い合わせできるようになります。これにより、感染防止の意思決定、人と人との接触の事後解析などに役立てることができます。

提案手法：多段階時空間推論ネットワーク



評価データ：映像質問応答の公開データセット

実験結果：映像質問応答手法の精度の比較

	MSVD	MSRVTT	ActivityNet
HME [Fan+, CVPR 2019]	0.337	0.330	0.331
HCRN [Le+, CVPR 2020]	0.363	0.355	0.362
提案手法 [BMVC 2020]	**0.432**	**0.394**	**0.402**
	+ 19%	+ 11%	+ 11%

連絡先：認知機構研究所　担当　宮西大樹　E-Mail: miyanishi@atr.jp

本研究は科学技術振興機構戦略的創造研究推進事業（CREST）の研究委託により実施したものです。