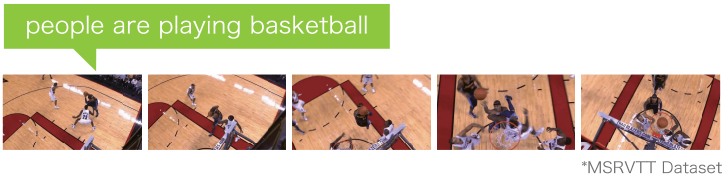
映像とテキストを組み合わせたストーリー理解の実現

株式会社サイバーエージェント　大谷 まゆ

Motivation

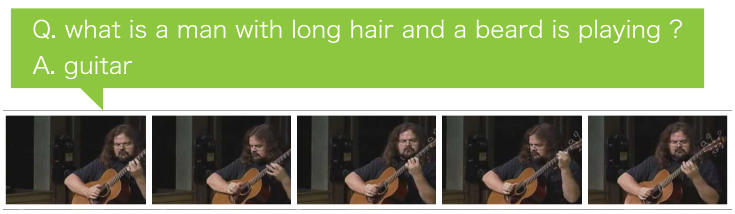
映像から複雑な意味を理解したい
これまでの映像理解タスクの多くは映像全体を見る必要がない
single action / event はフレーム 1 枚、ごく短い時間の画像特徴で理解できることが多い



*MSRVTT Dataset

Goal

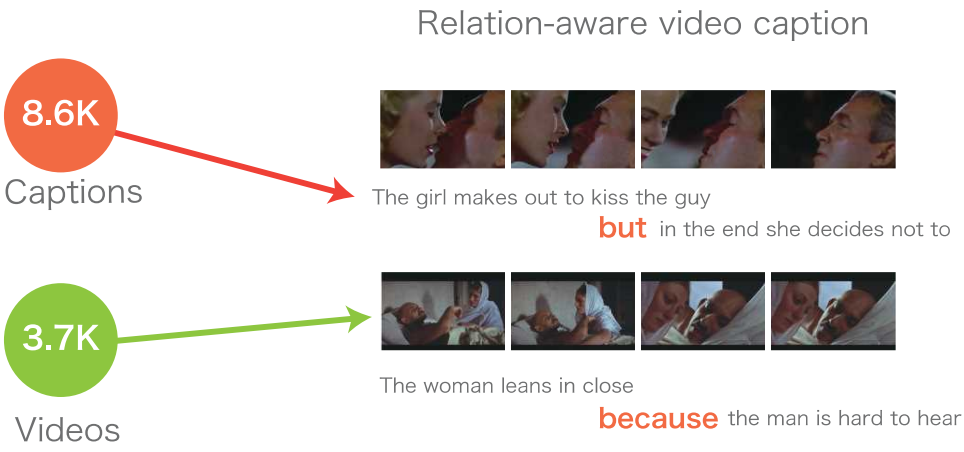
時間方向の変化、イベント間の関係性に着目したタスク設計
データセットの構築
既存の映像理解手法の限界を明らかにする

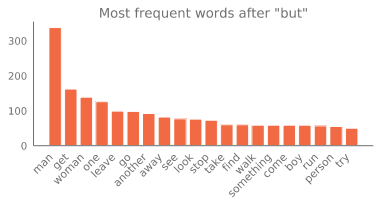


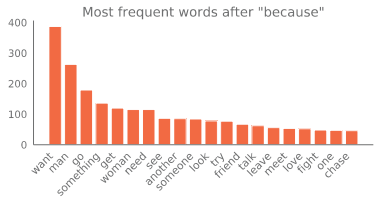
*MSVD-QA Dataset

Data

映像中の複数イベント、およびイベント間の関係性を説明するキャプションを収集







Experiments

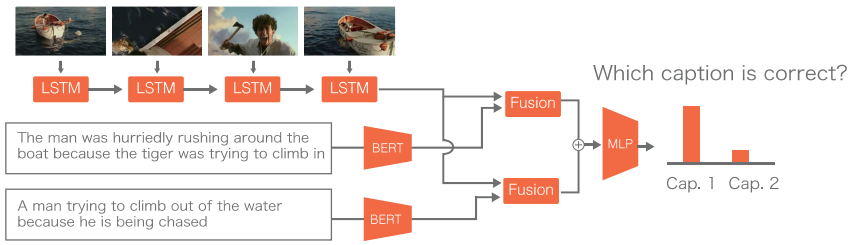
動画に対して正しいキャプションを予測する

Query video

Candidate captions

1. They are running away because a grenade was thrown at the car. (Correct)
2. A woman shows her cut wrists to a young boy but the young boy runs away.

Model



Results

同じ映画を使った既存データセットでは簡単なタスクも提案データセットでは難しい
既存データセット：フレーム順序を無視したモデルの性能が高い
提案データセット：フレームの順序を考慮したモデルの方が高いスコア→時間方向の変化のモデリングが重要

Accuracy for caption choice quizzes

	RAVC	MPII-MD
Language-only	0.465	0.498
Pooling-encoder	0.624	0.819
LSTM-encoder	0.643	0.804