第22回ディスカッション

実験状況

ER20038 小林亮太

担当：鈴木雅★，福井，張

MPRG

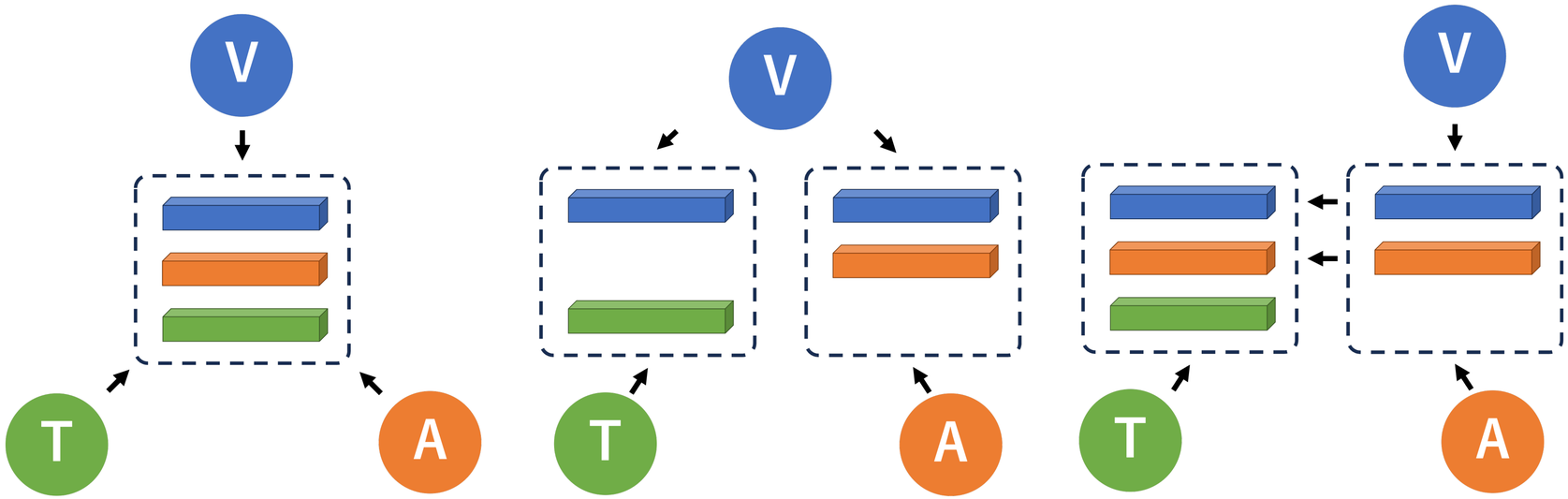
MACHINE PERCEPTION AND ROBOTICS GROUP

はじめに

- 研究テーマ
- 実験概要
- 実験条件
- 実験状況

研究テーマ

- 3モーダル（ビデオ，オーディオ，テキスト）のマルチモーダル自己教師あり学習
- テキストに比べビデオやオーディオにはノイズが多く存在
  - 各モーダルの組み合わせでノイズを抽出せずに学習ができる可能性
    - 近づけるモーダルの組み合わせによる学習効果への影響について調査



実験概要

- 3モーダルを二段階で学習実験
  - 3パターンを実験
    - AV_T：AとVで学習した後からTを追加
    - VT_A：VとTで学習して後からAを追加
    - AT_V：AとTで学習して後からVを追加

- HowTo100Mデータセットを用いて学習
- YouCook2データセットを用いてゼロショットで評価
- テキストからビデオの検索タスクで評価
  - テキストによるビデオ内の該当箇所の検索

実験条件

- Feature Extractor
  - ビデオ　　：ResNet152
  - オーディオ　　：DaveNet [D Harwath+, ECCV’18]
  - テキスト　　：Word2vec
- バッチサイズ　　：128
- エポック数　　：30
  - 前半20，後半10に設定
- 学習率　　：0.0001
- 最適化手法　　：Adam
- GPU　　：A100 × 4

実験状況

- １段階目でLossがNaNになる問題が発生
  - Contrastive Lossの計算のみの場合，Contrastive LossとRecon Lossの場合
    - NaNは不出現
    - Clustering Lossが原因？
  - Clustering Lossの計算結果を割る処理を追加
  - 対応済み

- 一段階目14/20epochまで完了
  - 土曜日に完了予定
  - 完了後再起動ののち二段階目開始予定

おわりに

- 実験　　：実行中

- 今後の予定：
  - 他のパターンの実験
  - 卒論

Multimodal Clustering Network (MCN) [B. Chen+, ICCV'21]

- ラベル付けされていないナレーション付きビデオから学習
  - テキストからビデオの検索，時系列行動検出が可能
- テキスト，オーディオ，ビデオの3つのモーダルを使用