第11回ディスカッション

中間発表の振り返り

ER20038 小林亮太

担当：鈴木雅★，福井，張

MPRG

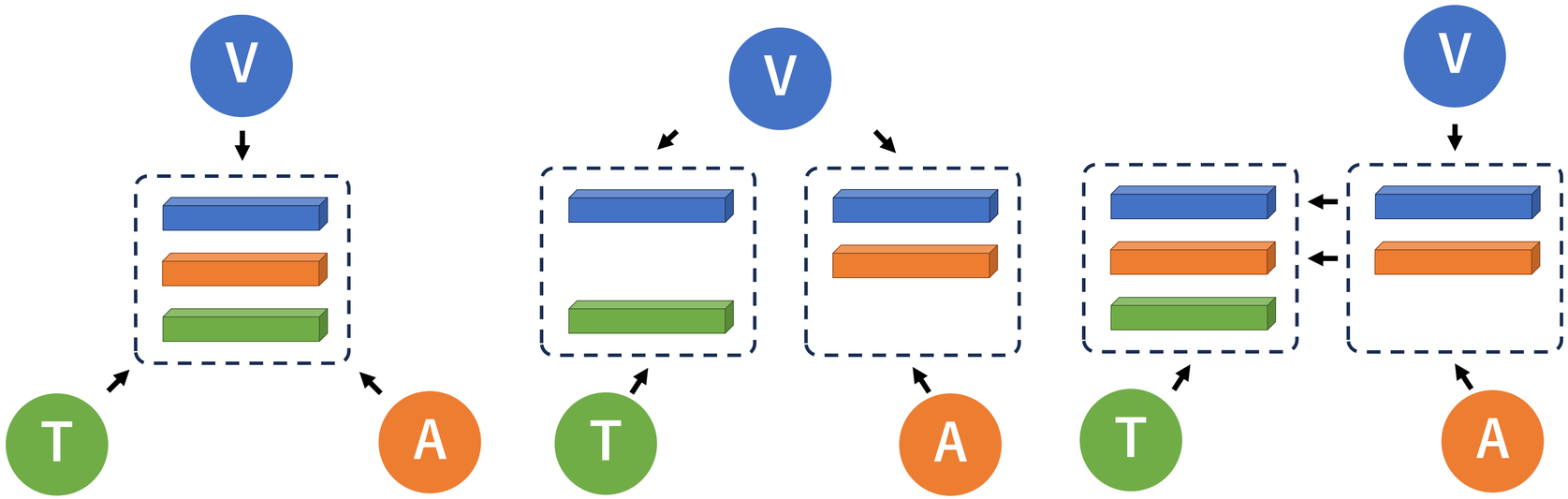
MACHINE PERCEPTION AND ROBOTICS GROUP

はじめに

- 研究テーマ
- 中間発表でのコメント
- ダウンロードの進捗
- 学習条件
- 学習結果
- 学習時間

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



中間発表でのコメント

- 山下先生
  - 学習にはどのくらいの時間が必要か？
- 藤吉先生
  - CLIP2などの考え方を取り入れて手法を模索していくのも良いでのは？
- 平川先生
  - 計画的に実験を行っていきましょう

HowTo100Mデータセット進捗状況

- ダウンロード完了
- リサイズ完了
- ビデオ，オーディオともに特徴量抽出完了
- 898,094本のビデオ，オーディオのデータ

実験条件

- アーキテクチャ：MCN
- Feature Extractor：
  - ビデオ　　：ResNet152
  - オーディオ：DaveNet
  - テキスト　：Word2vec
- データセット　：HowTo100M
  - ビデオ解像度　　：454 × 256
  - ビデオフレームレート　　：30FPS
  - オーディオサンプリングレート　：16kHz
- バッチサイズ　：128
- エポック数　　：30
- 学習率　　：0.0001
- 特徴量次元数　：4096

実験結果

{hoge}/30 epoch完了

学習時間

- 論文ではV100 GPU × 4 で約2日
- A100 GPU × 4 で約6日
  - Dataloaderのnum_workerの値を変更
    - 初期値：74
    - 変更値：8
  - その他の値はそのまま使用

おわりに

- データセットの進捗　：ようやく完了
- 学習時間　　：短縮のためにプログラムを変更
- 再現実験　　：実行中

- 今後の予定：
  - 実験の結果の分析
  - CLIP2の論文調査

資料

CLIP2

HowTo100M

- ナレーション付きビデオの大規模なデータセット
  - 120万本のYoutubeのビデオにキャプションを付けた1億3600万本のビデオクリップ
  - 削除や非公開によって全体数が減少傾向
    - 23年8月現在，約90万本の動画が存在
  - 大量の空き容量が必要
    - ビデオデータ　　：約45TB
      - データサーバへ
    - 抽出された特徴量データ　：約10TB
      - 学習で使用

- 様々な手法の学習で使用
  - 調査を行った3つの論文において使用
  - 入手することができれば有益

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，時系列行動検出で評価
- テキスト，オーディオ，ビデオの3つのモダリティを使用