第14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

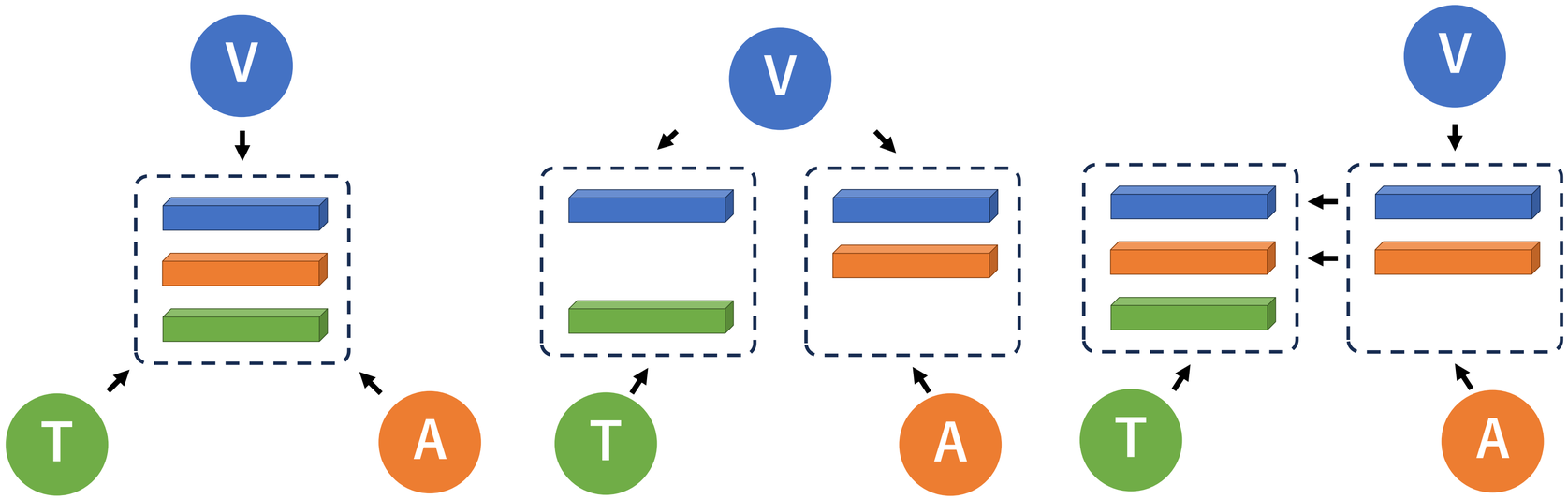
MACHINE PERCEPTION AND ROBOTICS GROUP

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



実験状況

- 3モーダルで事前学習
- LossがNaNになる問題
  - オーディオのLossがNaN，テキストとビデオは正常
  - オーディオの入力がテキストとビデオより大きい値
  - オーディオの入力を1000で除算して一時的に対処
- 動作テストの際の変更が原因
  - 解決
  - 1000での除算を削除

- 原因特定後再度実行により遅延
  - 土曜日に完了予定

実験結果

- 現在，19/30 epoch

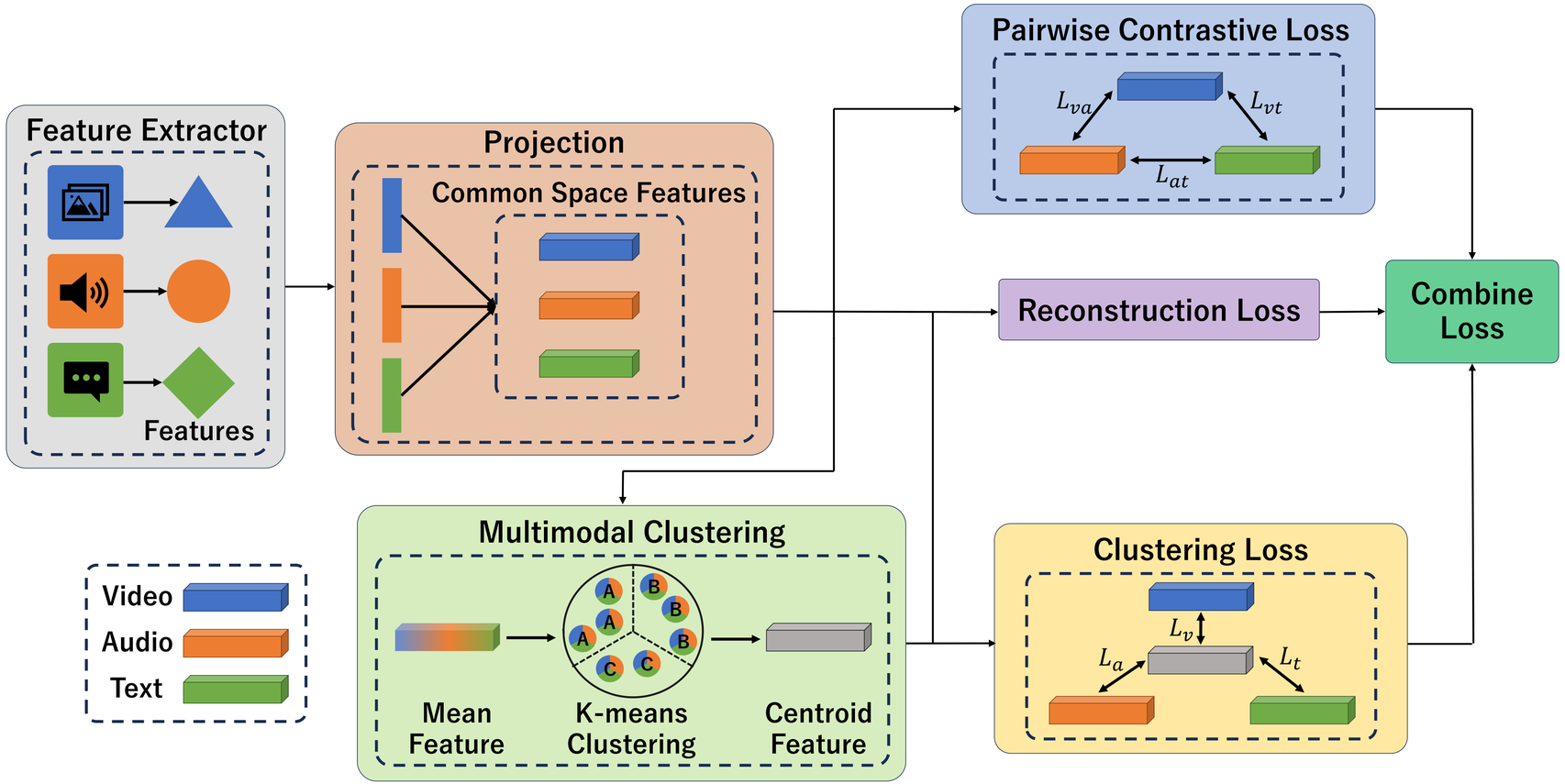
おわりに

- 実験　　：実行中

- 今後の予定：
  - 実験の結果の分析
  - プログラムの作成

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



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