中間発表

マルチモーダル自己教師あり学習における モーダルの組み合わせによる影響の調査

機械知覚＆ロボティクスグループ（藤吉研究室）

ER20038　小林 亮太

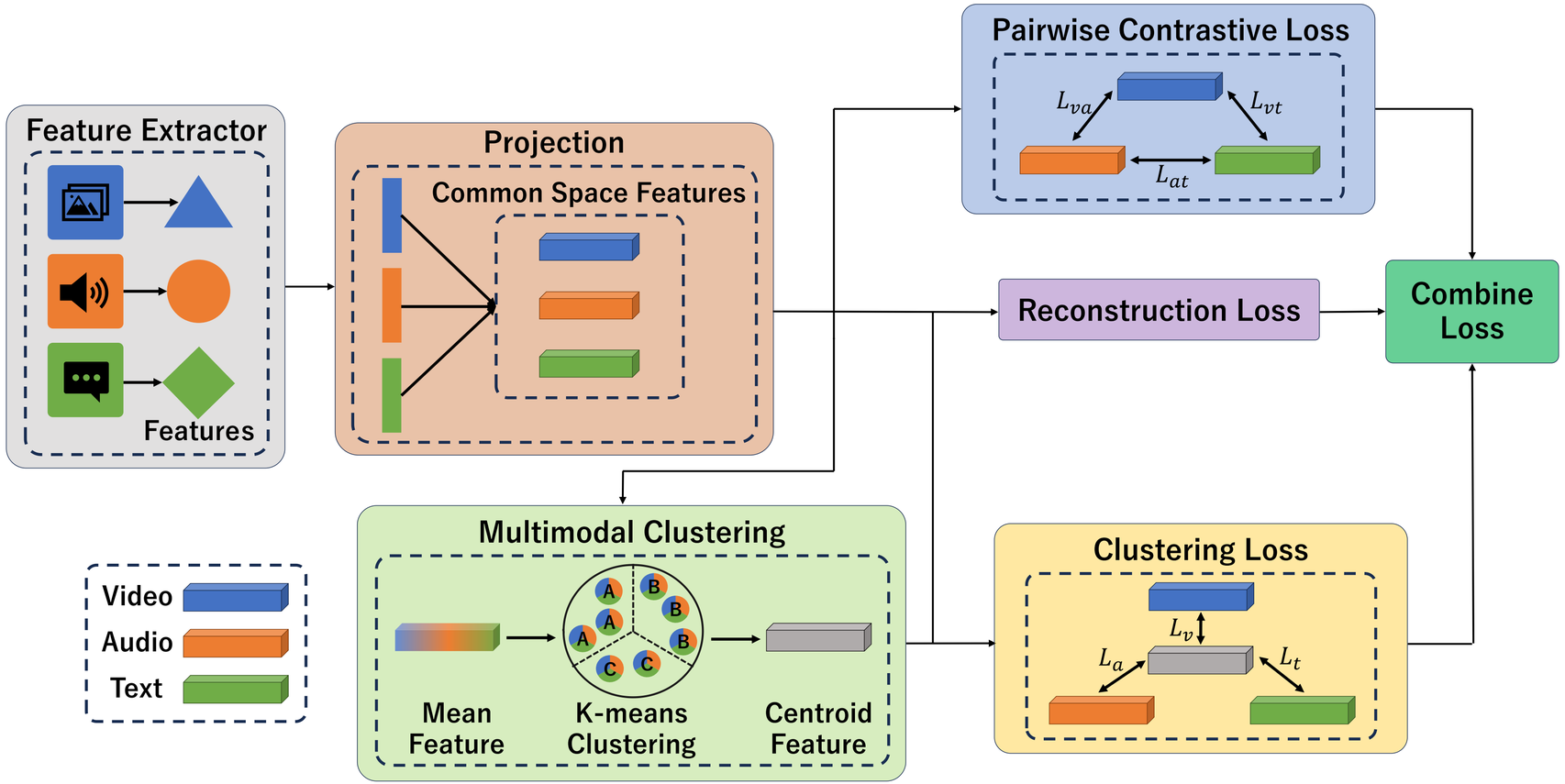
担当：鈴木雅

MPRG

MACHINE PERCEPTION AND ROBOTICS GROUP

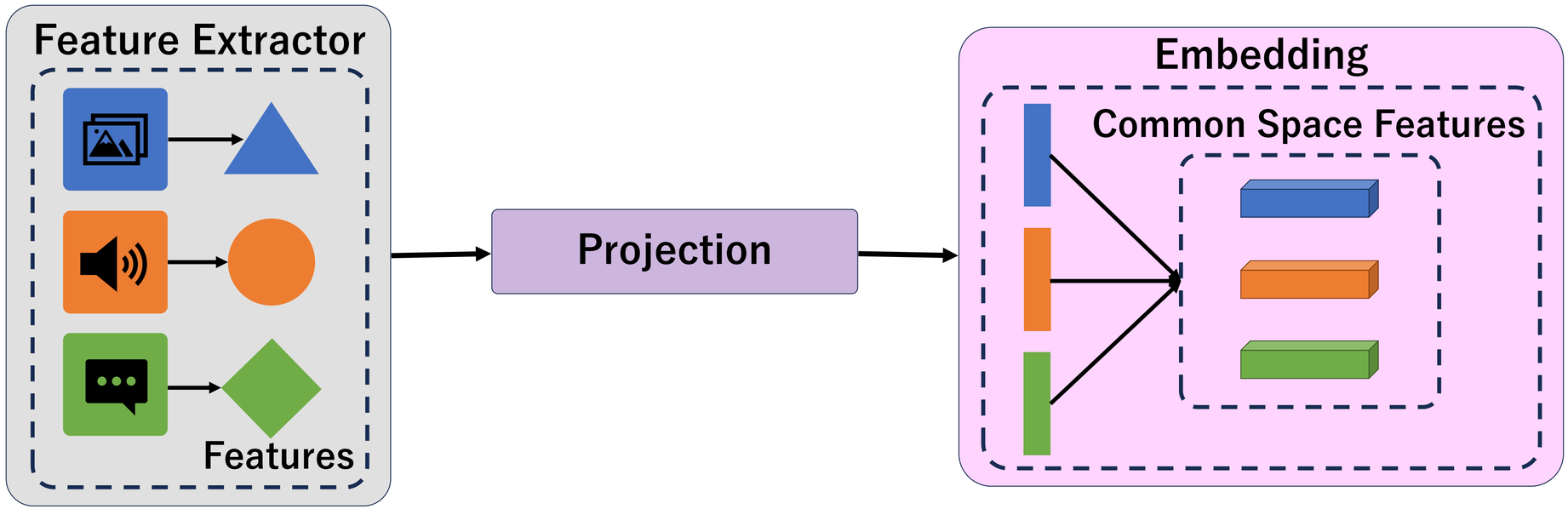
Multimodal Clustering Network (MCN) [B. Chen+, ICCV'21]

- 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



MCNの特徴量抽出

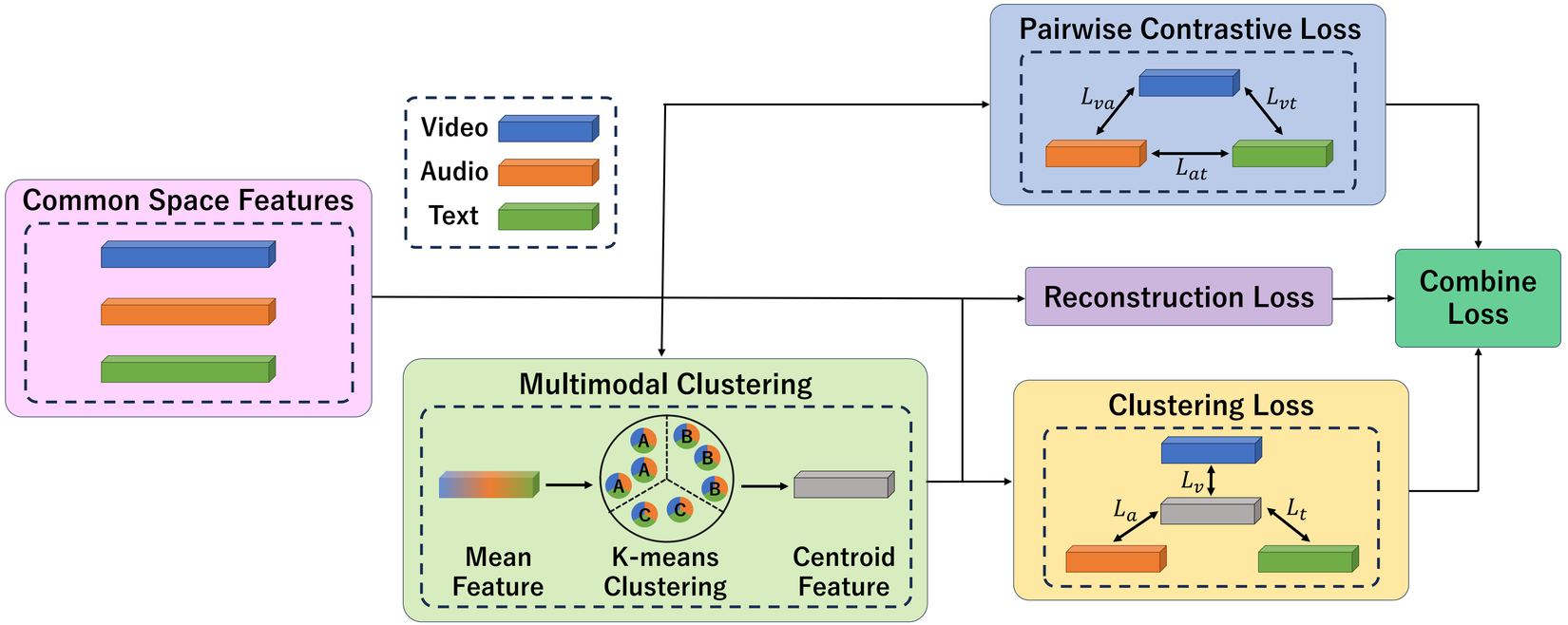
- モーダル毎に学習済みのモデルを使用
- 抽出した3つのモーダルの特徴量を共通の空間に埋め込む
  - 異なるモーダルを統合的に扱うことが可能



MCNの損失関数

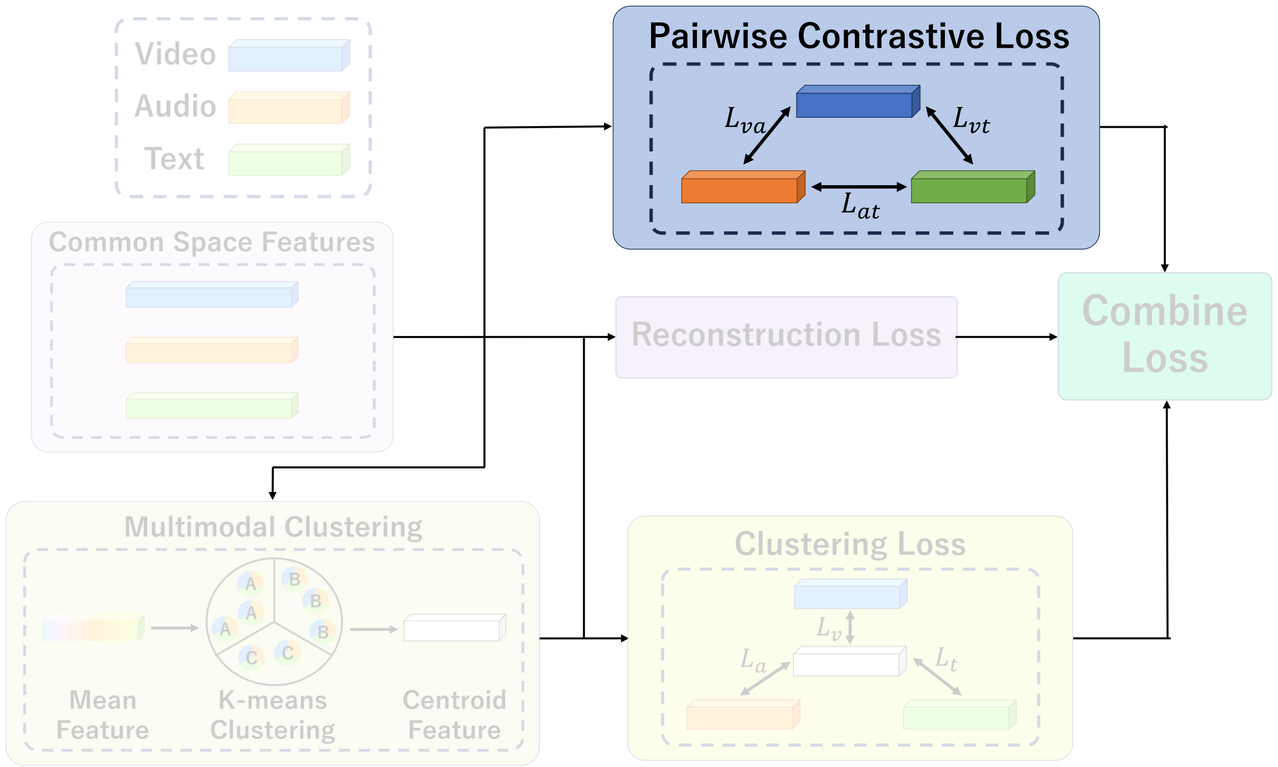
- 共通の空間内に適切に配置するために3つの損失関数を導入
  - L_{MMS} ：Pairwise Contrastive Loss
  - $L_{Cluster}$ ：Multimodal Clustering Loss
  - $L_{Reconstruct}$ ：Reconstruction Loss
- 3つの損失関数の合計Combine Loss $L_{Combine}$ を最小化するように学習

$$L_{Combine} = L_{MMS} + L_{Cluster} + L_{Reconstruct}$$



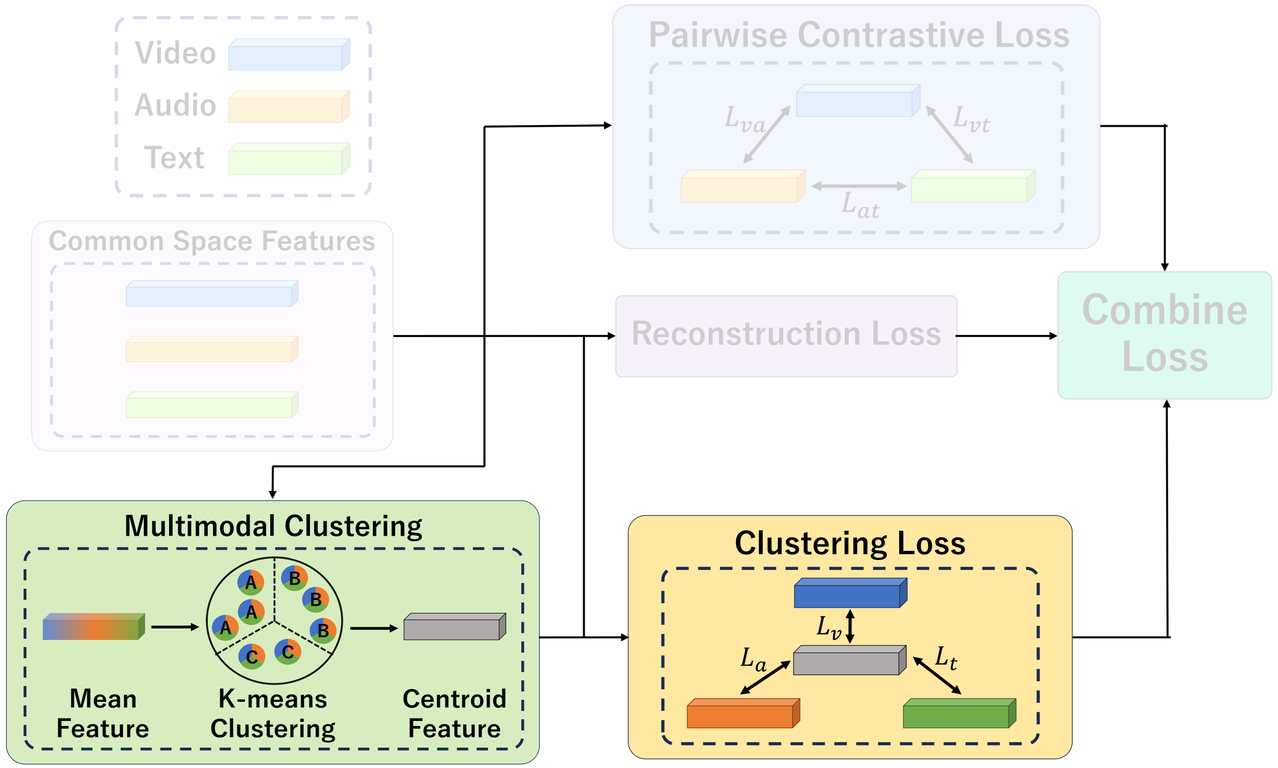
Pairwise Contrastive Loss L_{MMS}

- サンプル毎にモーダル間の時間的な距離を近づける損失
- 全てのモダリティのペアに対して損失を算出



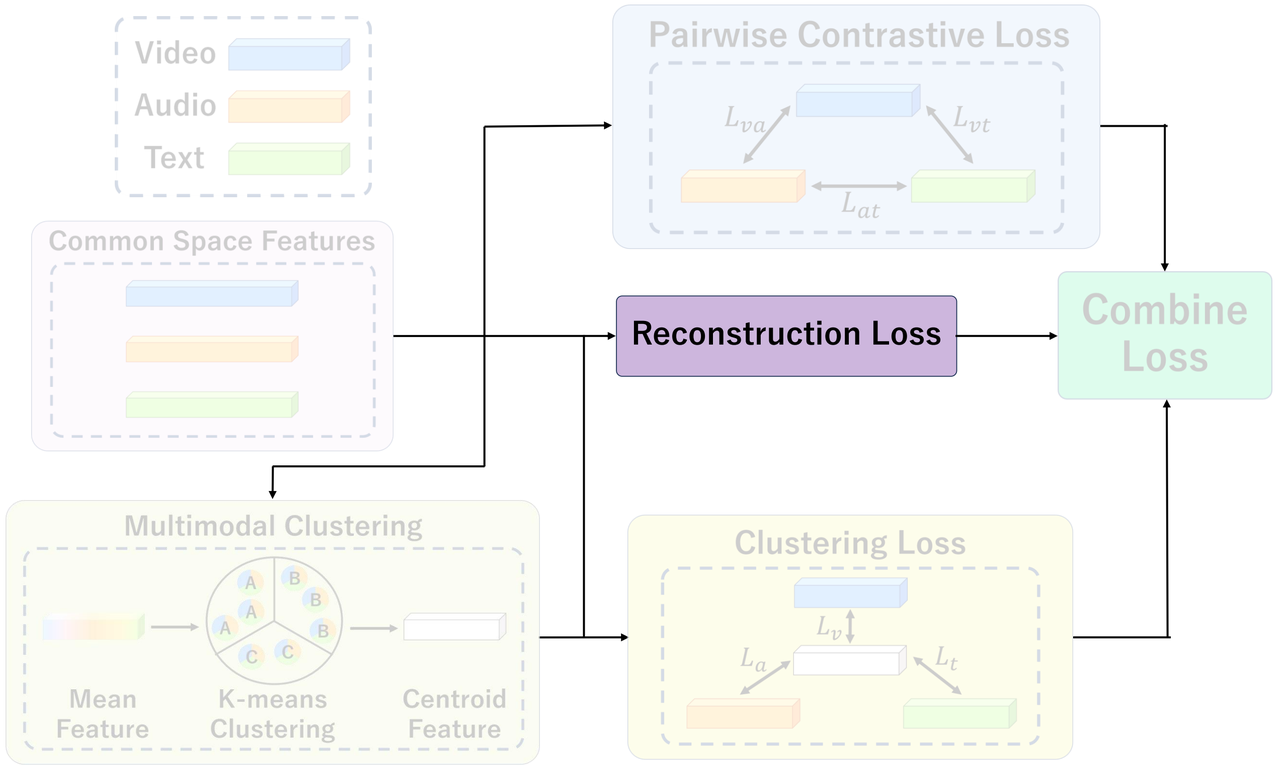
Multimodal Clustering Loss $L_{Cluster}$

- 同じクラスタに属する特徴量の表現が似通るように学習
- K-means法で分割したk個のクラスタの重心に近づける損失を算出



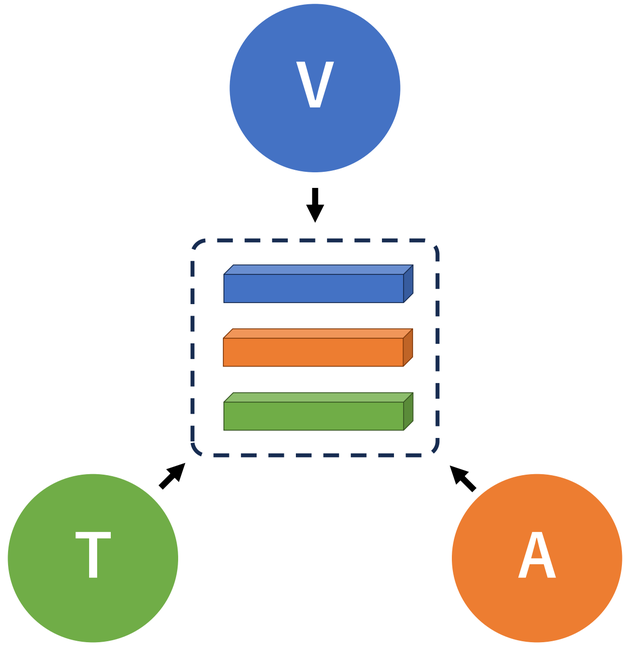
Reconstruction Loss $L_{reconstruct}$

- オートエンコーダで再構成した出力データと入力データを近づけるように学習
  - Contrastive learningやClusteringによって抑制された特徴を捕えることが可能



MCNの利点と欠点

- 利点
  - 全モーダルの共通空間により検索などをモーダル間の隔たりなく実行可能
- 欠点
  - テキストはオーディオやビデオに比べて抽象的
    - きめ細かい情報が失われる可能性



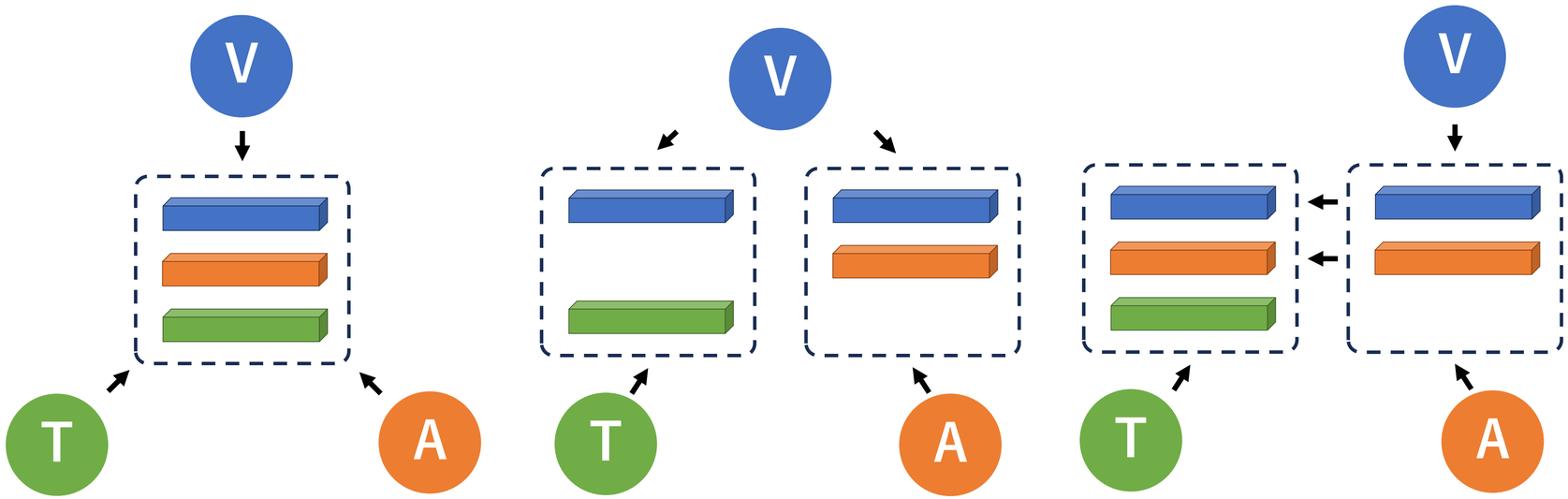
本研究の目的

- モーダルの性質に応じた対照学習の設計による事前学習の性能改善
  - 使用するモーダル：ビデオ，オーディオ，テキスト

本研究の目的

- モーダルの性質に応じた対照学習の設計による事前学習の性能改善
  - 使用するモーダル：ビデオ，オーディオ，テキスト

- 事前調査：近づけるモーダルの関係による対照学習の学習効果への影響調査



モーダルの組み合わせ方による影響

- モーダルによって異なる内容のデータを保有
  - ビデオ　　：主役となる物体＋背景
  - オーディオ：ナレーションの音声＋雑音
  - テキスト　：ナレーションの内容
- モーダルの組み合わせ方によって背景や雑音の情報を維持・軽減した学習が可能

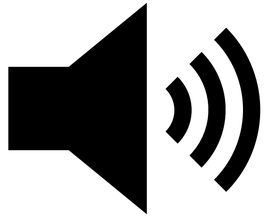
ビデオ

・包丁を使用している
・玉ねぎは半分の状態
・まな板の上
etc

テキスト

Chopped the onions and set

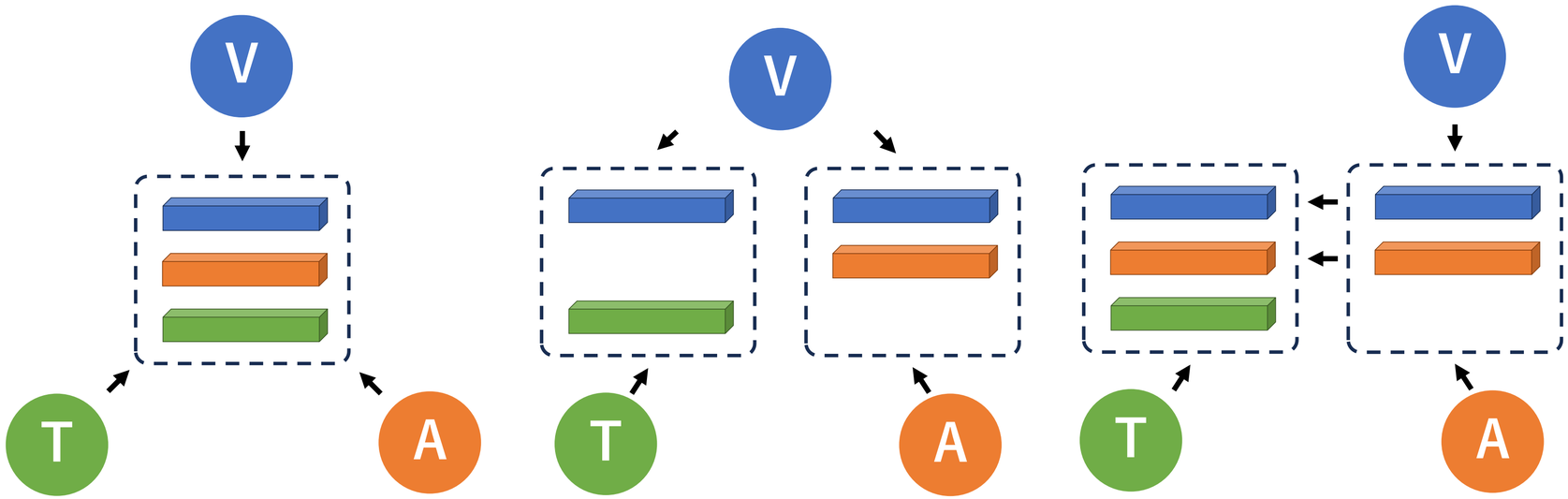
オーディオ



・テキストの読み上げ
・切っている音
・その他雑音

モーダルの組み合わせの検討

- 3つの組み合わせ方について調査
  - 全てのモーダル間で対照学習
  - ビデオ・テキスト間，ビデオ・オーディオ間で対照学習
  - ビデオ・オーディオ間，全てのモダール間で対照学習
- 複数の埋め込み空間に分けることでモーダル特有の情報を活用

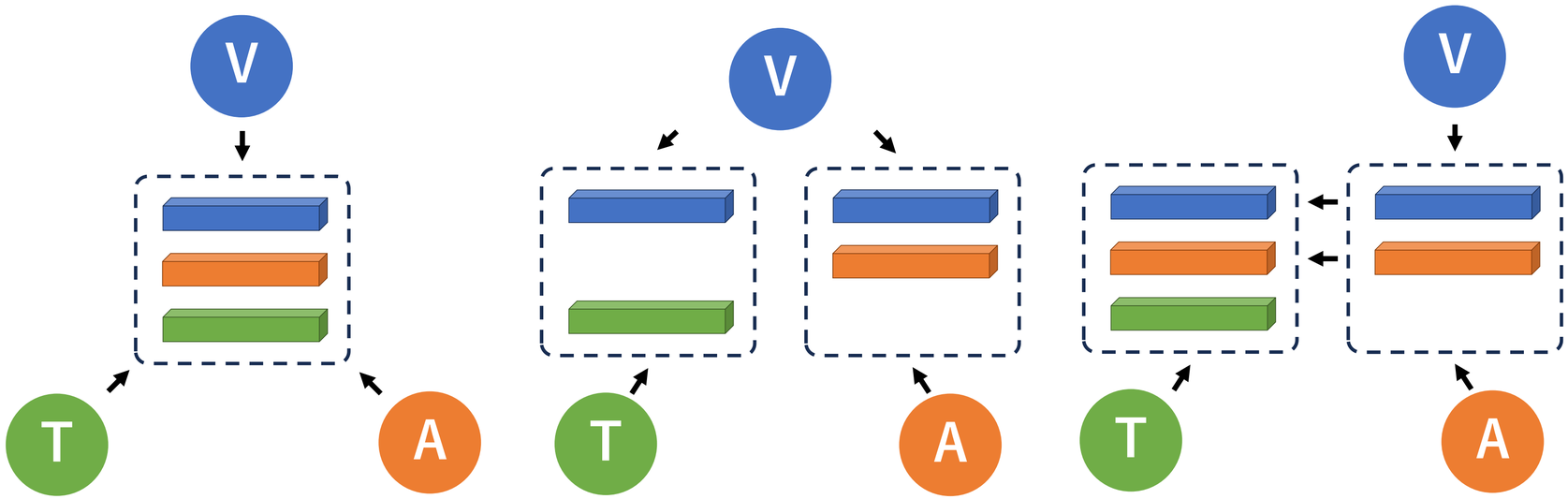


モーダルの組み合わせの検討

- 3つの組み合わせ方について調査
  - 全てのモーダル間で対照学習
  - ビデオ・テキスト間，ビデオ・オーディオ間で対照学習
  - ビデオ・オーディオ間，全てのモダール間で対照学習

→ ノイズ・背景の情報を維持・活用

- 複数の埋め込み空間に分けることでモーダル特有の情報を活用



実験概要

- MCNを用いて近づけるモーダルの組み合わせによる学習効果を調査
- HowTo100Mデータセットを用いて学習
- テキストからビデオを検索，時系列行動検出で評価
  - 時系列行動検出　　：特定の行動の開始時間と終了時間の範囲を検出
  - テキストからビデオの検索　：テキストによるビデオ内の該当箇所の検索

HowTo100M

- ナレーション付きビデオの大規模なデータセット
  - 123万本の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
  - 現存していた約90万本の動画を使用
- 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

実験条件

- アーキテクチャ：MCN
- 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
- 学習率　　：0.0001
- 特徴量次元数　：4096
- 最適化手法　　：Adam

まとめ

- MCNの手法の調査
  - 単一の共通空間を作成
  - オーディオやビデオの背景の情報が失われる可能性
- モーダルの性質に応じた設計による性能改善
  - 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
  - MCNの再現実験

- 今後の予定
  - MCNの再現実験の結果の分析
  - 他手法の調査

資料

Gated Embedding Unit

- 線形射影により入力データを埋め込み表現に変換
- 埋め込み表現をGated Linear Unitへ入力
  - 重要な特徴を保持しながら効率的な表現が生成可能
  - Gated Linear Unit (GLU)
    - 入力を2分割
    - 片方にシグモイド関数を適応
    - 要素ごとの積の結果を出力

Pairwise Contrastive Loss L_{MMS}

- Masked Margin Softmax (MMS) [G. Ilharco+, CoNLL'19]を使用
  - 基準となるモーダルを変えた2つのContrastive Lossから構成

$$L_{at} = -\frac{1}{B}\sum_{i=1}^{B}\left[\left(\log\frac{e^{h(t_i)\cdot g(a_i)-\delta}}{e^{h(t_i)\cdot g(a_i)-\delta}+\sum_{\substack{k=1\\k\neq i}}^{B} e^{h\left(t_k^{imp}\right)\cdot g(a_i)}}\right)+\left(\log\frac{e^{h(t_i)\cdot g(a_i)-\delta}}{e^{h(t_i)\cdot g(a_i)-\delta}+\sum_{\substack{j=1\\j\neq i}}^{B} e^{h(t_i)\cdot g\left(a_j^{imp}\right)}}\right)\right]$$

a_i : オーディオ
t_i : テキスト
a_j^{imp} : t_iの負のペア
t_k^{imp} : a_iの負のペア
B : バッチサイズ
δ : マージン

- L_{MMS}はすべてのペアの損失の合計

$$L_{MMS} = L_{ta} + L_{vt} + L_{va}$$

Multimodal Clustering Loss $L_{Cluster}$

- 単一のContrastive Lossで算出

$$L_t = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{e^{h(t_i)\cdot\mu' - \delta}}{\sum_{k=1}^{K} e^{h(t_i)\cdot\mu_k}}$$

t_i ：テキスト
B ：バッチサイズ
K ：クラスタ数
δ ：マージン
μ_k ：k番目のクラスタの重心
μ' ：t_iの最も近い重心

- $L_{Cluster}$は3つのモダリティの損失の合計

$$L_{Cluster} = L_v + L_a + L_t$$

Reconstruction Loss $L_{reconstruct}$

- 損失関数に正則化を加えることで，汎化性能の向上が可能
- 再構成前後の平均二乗誤差で算出

$$L_{v'} = -\frac{1}{B}\sum_{i=1}^{B} ||f'(\mathrm{v}) - f(\mathrm{v})||^2$$

v ：ビデオ
B ：バッチサイズ
$f'(v)$ ：再構成後
$f'(v)$ ：再構成前

- $L_{reconstruct}$は各モダリティの損失の合計

$$L_{Reconstruct} = L_{v'} + L_{a'} + L_{t'}$$