第10回ディスカッション

Multimodal Clustering Network (MCN) の再現実験

ER20038 小林亮太

担当：鈴木★，福井，張

MPRG

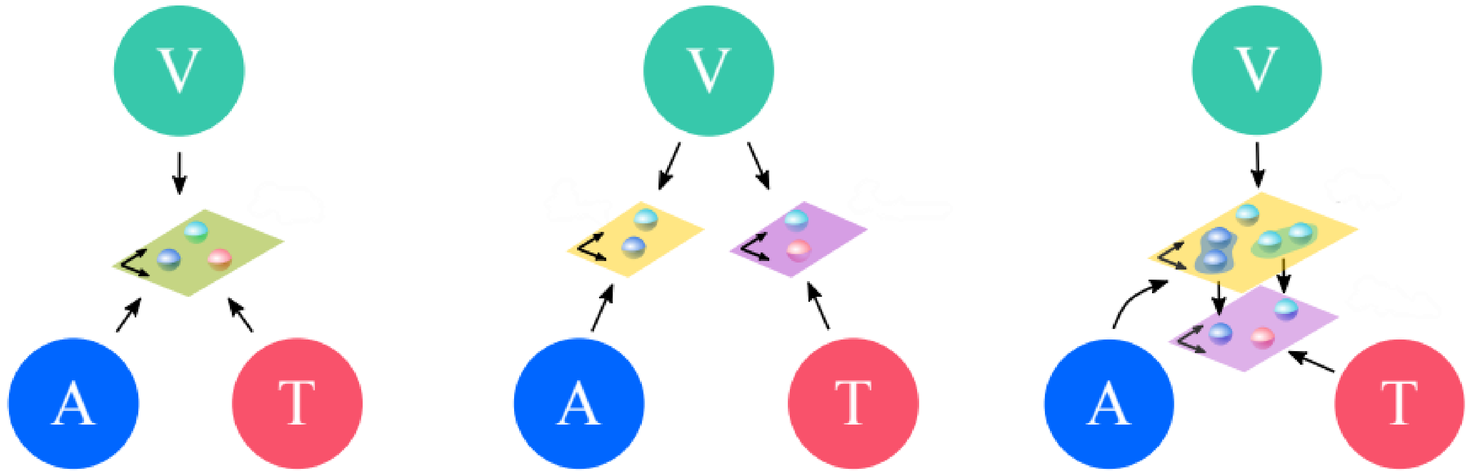
MACHINE PERCEPTION AND ROBOTICS GROUP

はじめに

- 研究テーマ
- HowTo100M
  - ダウンロードの進捗
  - フレーム数の増加
- Multimodal Clustering Network (MCN)
  - 特徴量抽出
  - MCNの損失関数
    - Contrastive Loss L_{MMS}
    - Clustering Loss $L_{Cluster}$
    - Reconstruction Loss $L_{reconstruct}$
- 再現実験

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



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

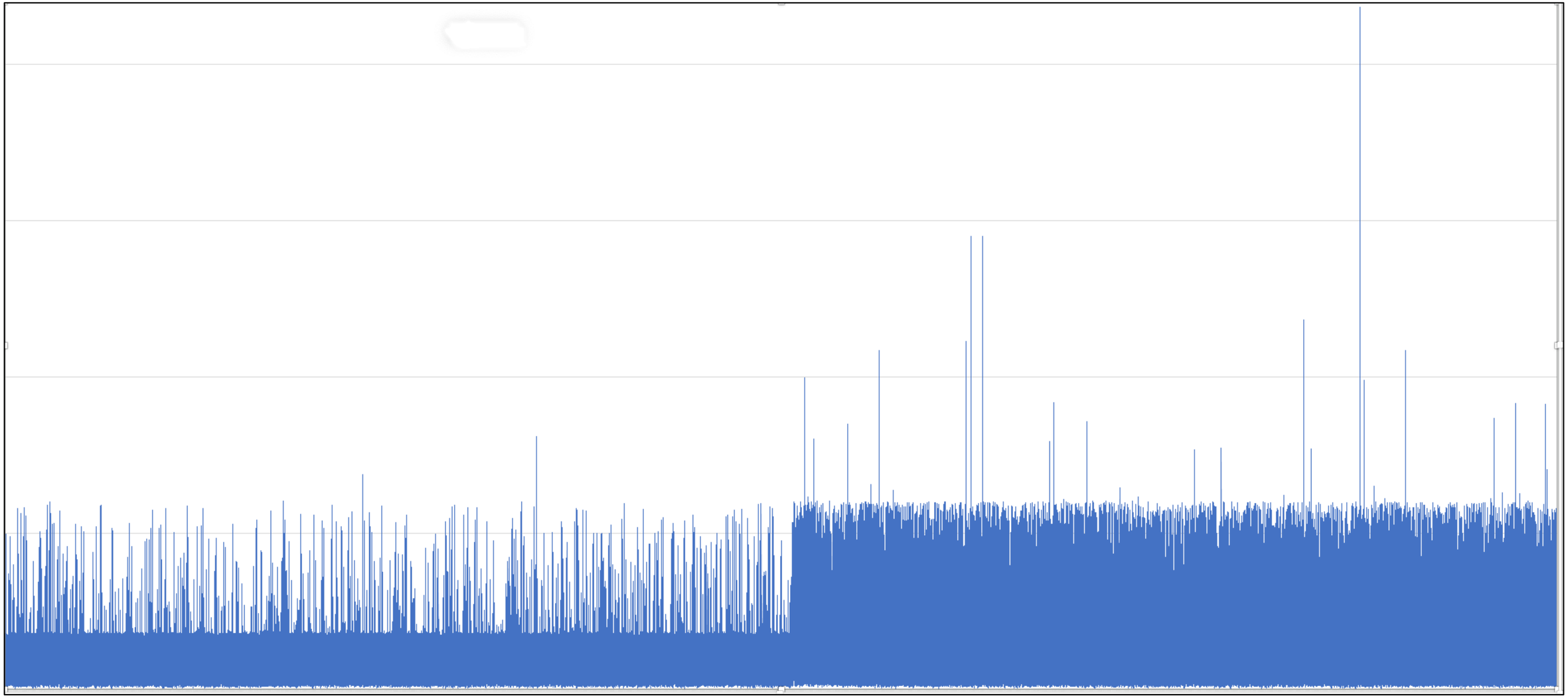
データセット進捗状況

- ダウンロード完了
- リサイズ完了間近
- 一部データのビデオ，オーディオともに特徴量抽出が完了
  - 一定数以上確保できた時点で学習での使用を開始
  - リサイズ完了までビデオの特徴量抽出は中断

進捗 予定	ダウンロード完了	リサイズ完了	特徴量抽出 オーディオ完了	特徴量抽出 オーディオ，ビデオ 完了
現在 8/24	898,096	780,000	598,000	460,000
8/31	898,096	870,000	640,000	460,000

フレーム数の増加

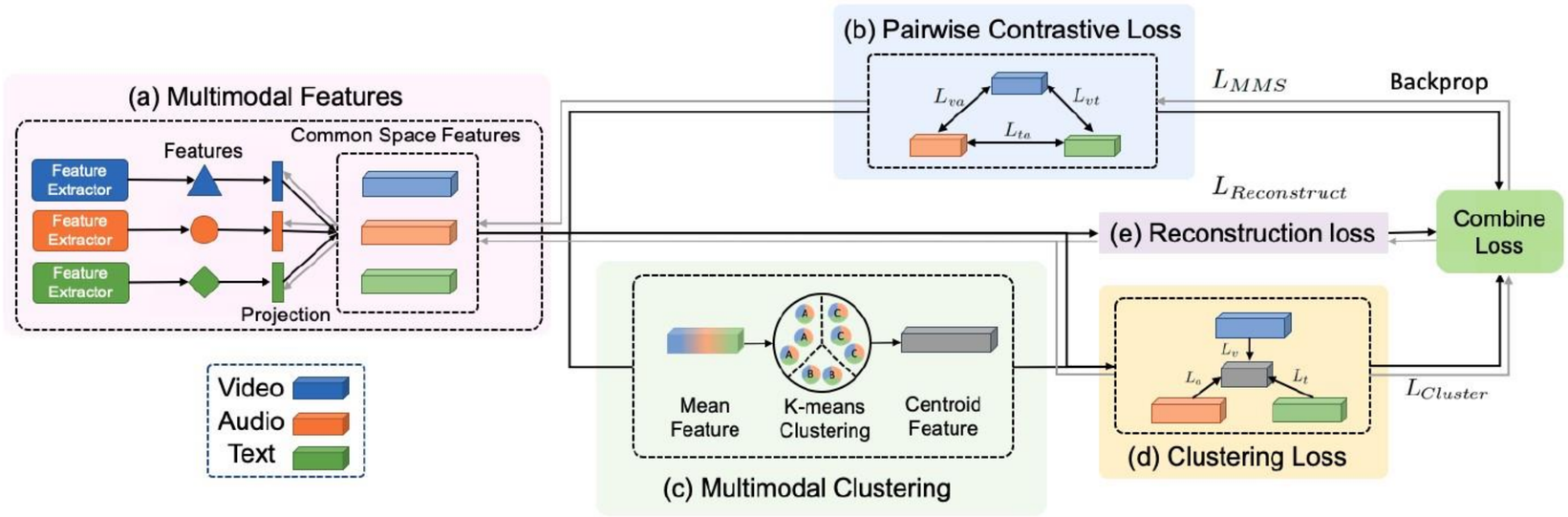
- ビデオのフレーム数が増加傾向
  - 処理全体に遅延が発生



Multimodal Clustering Network (MCN)
の再現実験

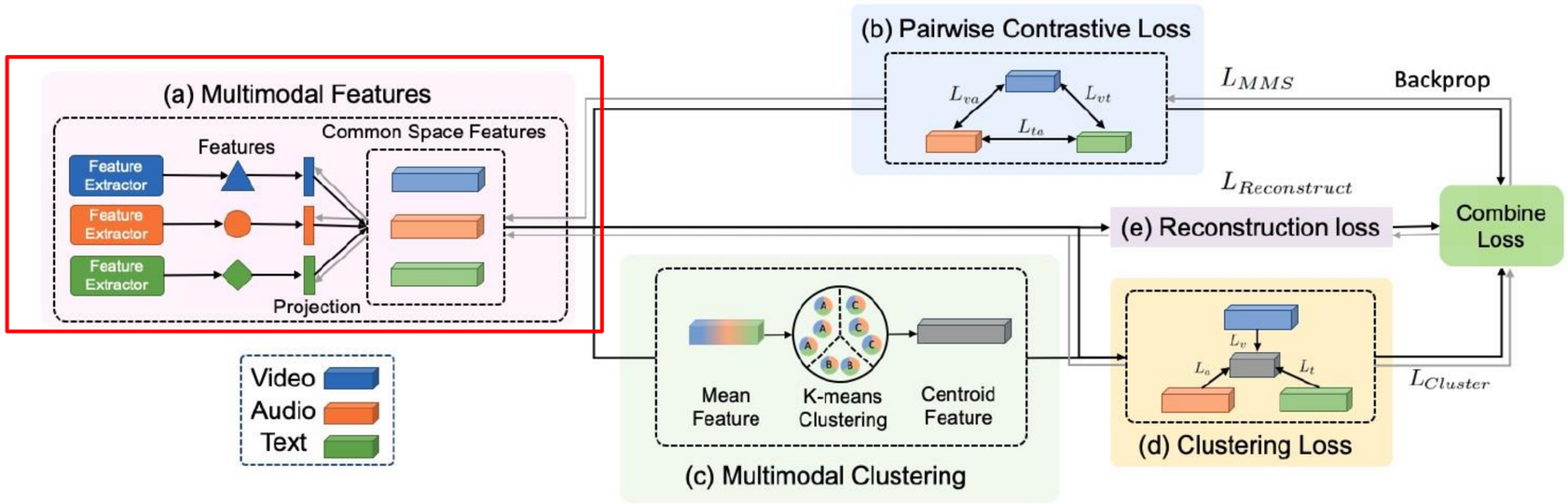
Multimodal Clustering Network (MCN) [B. Chen+, ICCV'21]

- 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



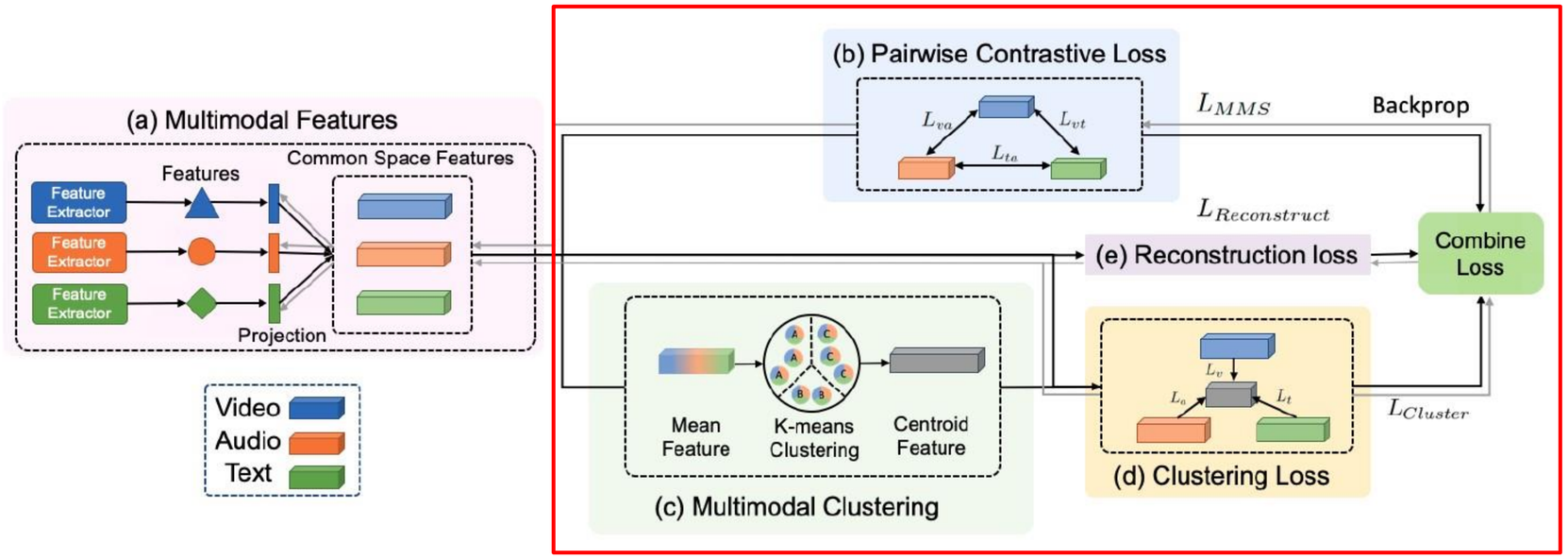
特徴量抽出

- 3つのモダリティの特徴量を低次元の共通の空間に写像
  - 異なる情報源を統合的に扱うことが可能
- 学習済みのFeature Extractorを用いて特徴量抽出



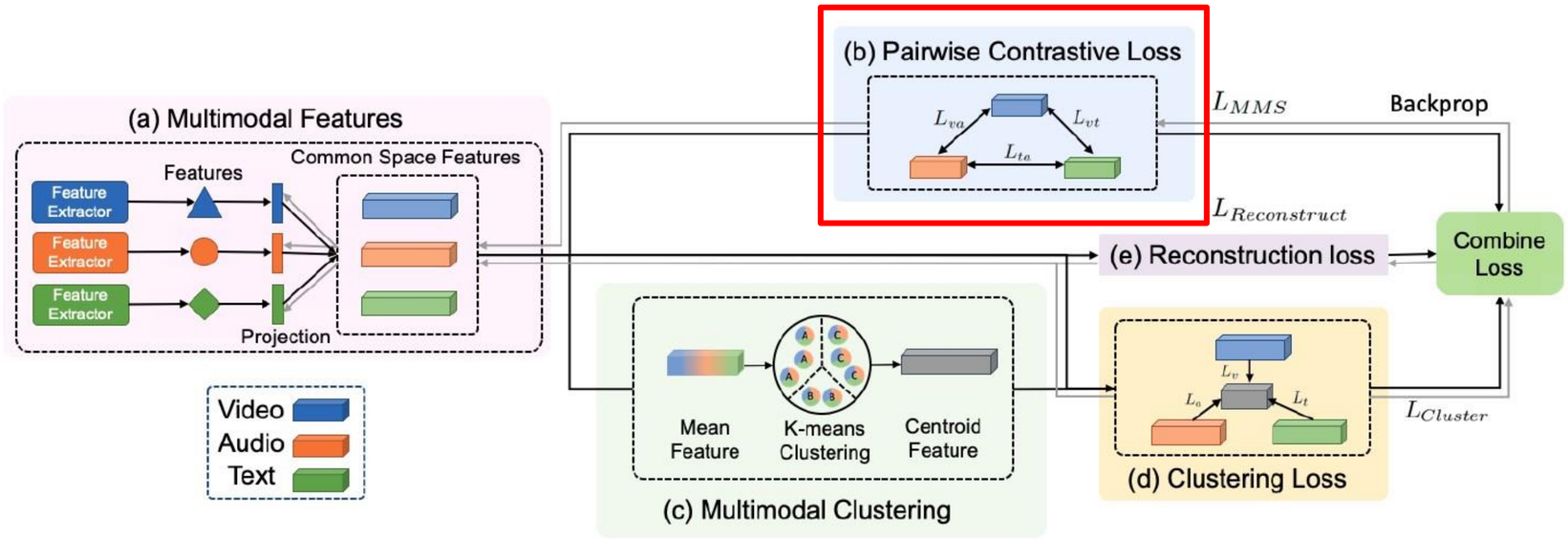
MCNの損失関数

- 空間内に適切に配置するために3つの損失関数を導入
  - Contrastive Loss
  - Clustering Loss
  - Reconstruction Loss
- 3つの損失関数の合計を最小化するように学習



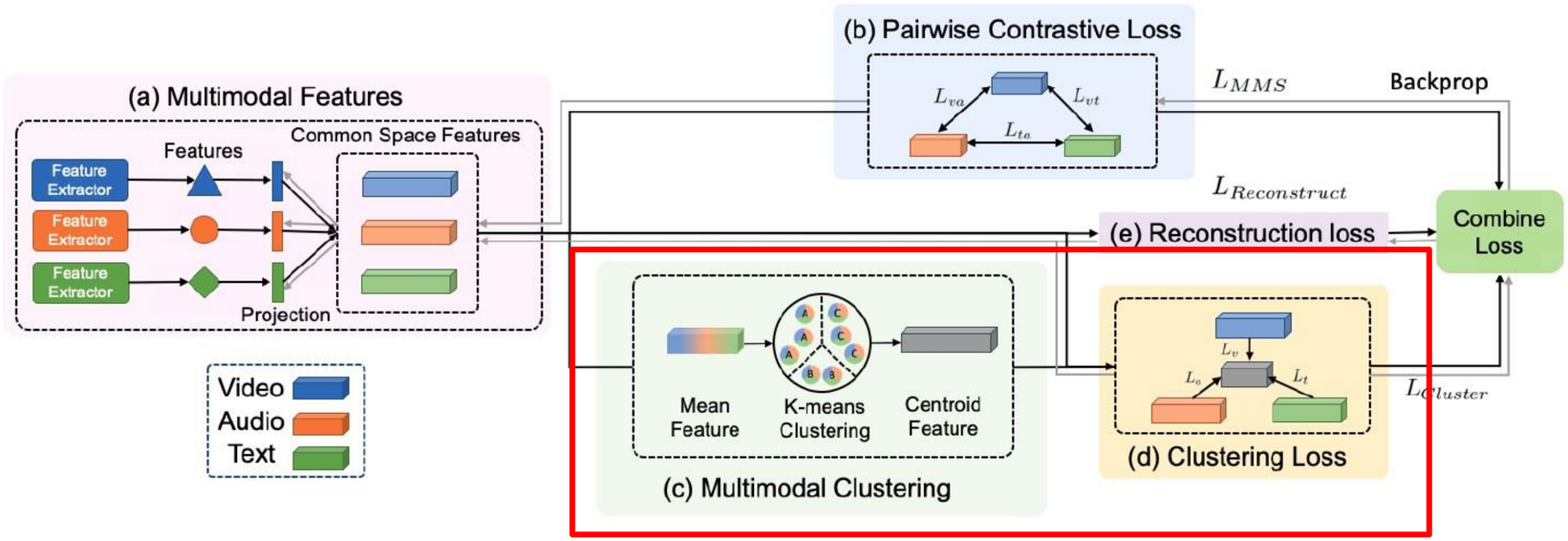
Contrastive Loss

- サンプル毎にモーダル間の特徴量の距離を小さくするように学習
- 全てのモダリティのペアに対して損失を算出
  - L_{vt} ：ビデオとテキスト
  - L_{va} ：ビデオとオーディオ
  - L_{ta} ：テキストとオーディオ



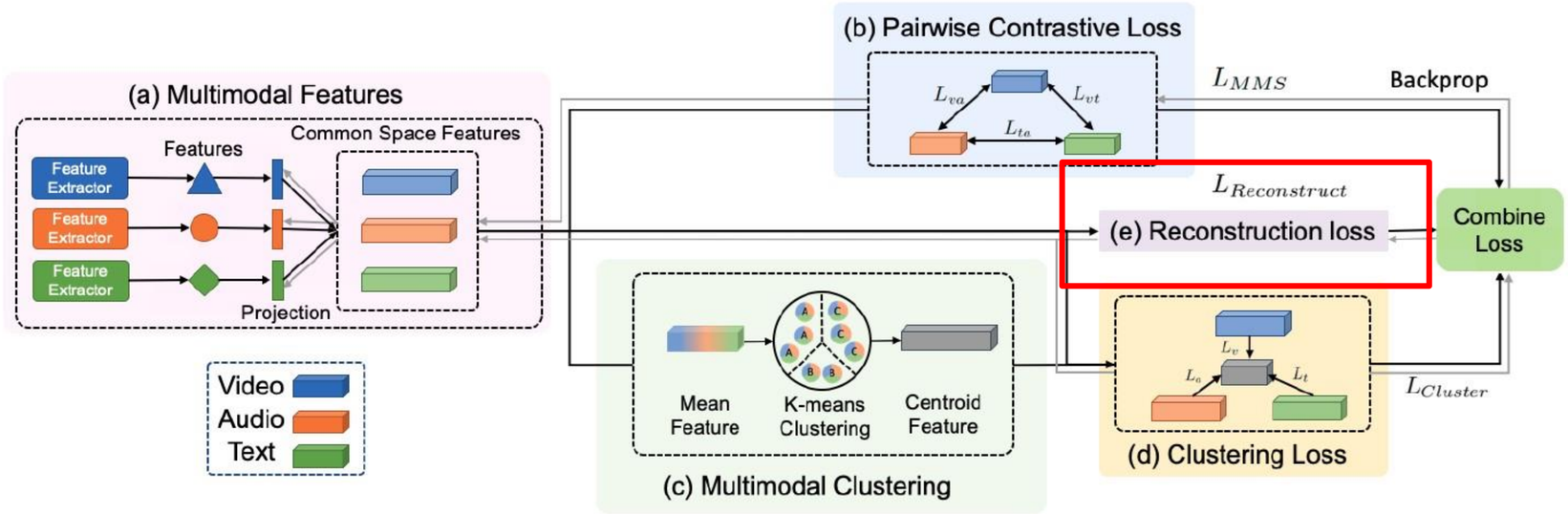
Clustering Loss

- 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法を用いてk個のクラスタに分割して各クラスタの重心を算出
- 各モーダルとクラスタ重心の距離から損失を算出
- Clustering Lossは3つのモダリティの損失の合計



Reconstruction Loss

- 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
- 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
  - オリジナルと再構築したものとの差を小さくする処理
- Reconstruction Lossは各モダリティの損失の合計



再現実験

- モデル　　：MCN
- Feature Extractor ：
  - ビデオ　　：ResNet152 + Gated Embedding Unit
  - オーディオ　：DaveNet + Gated Embedding Unit
  - テキスト　　：Gated Embedding Unit
- 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　：2048

実験結果

実行中

おわりに

- データセットの進捗　：ビデオデータの準備は完了間近　特徴量抽出は順次開始
- フレーム数の増加　　：引き続き情報を収集
- 再現実験　　：実行中

- 今後の予定：データセット準備の完了　MCNの再現実験　論文調査

Contrastive Loss L_{MMS}

- 基準となるモーダルを変えた2つのContrastive Lossから構成

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

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

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

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
- 各フレームとクラスタ重心の距離を最小化してより良いクラスタリングを実現
  - K-means法を用いてk個のクラスタに分割して各クラスタの重心を算出
- $L_{Cluster}$は3つのモダリティの損失の合計

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

t_i ：テキスト
B ：バッチサイズ
K ：分類数
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

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

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比較
  - Contrastive learningやクラスタリングによって抑制された特徴を捕えることが可能
- 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
  - オリジナルと再構築したものとの差を小さくする処理
- $L_{reconstruct}$は各モダリティの損失の合計

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

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

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