第1回ディスカッション

論文調査

ER20038　小林亮太

担当：岡本，張，岩垣

MPRG

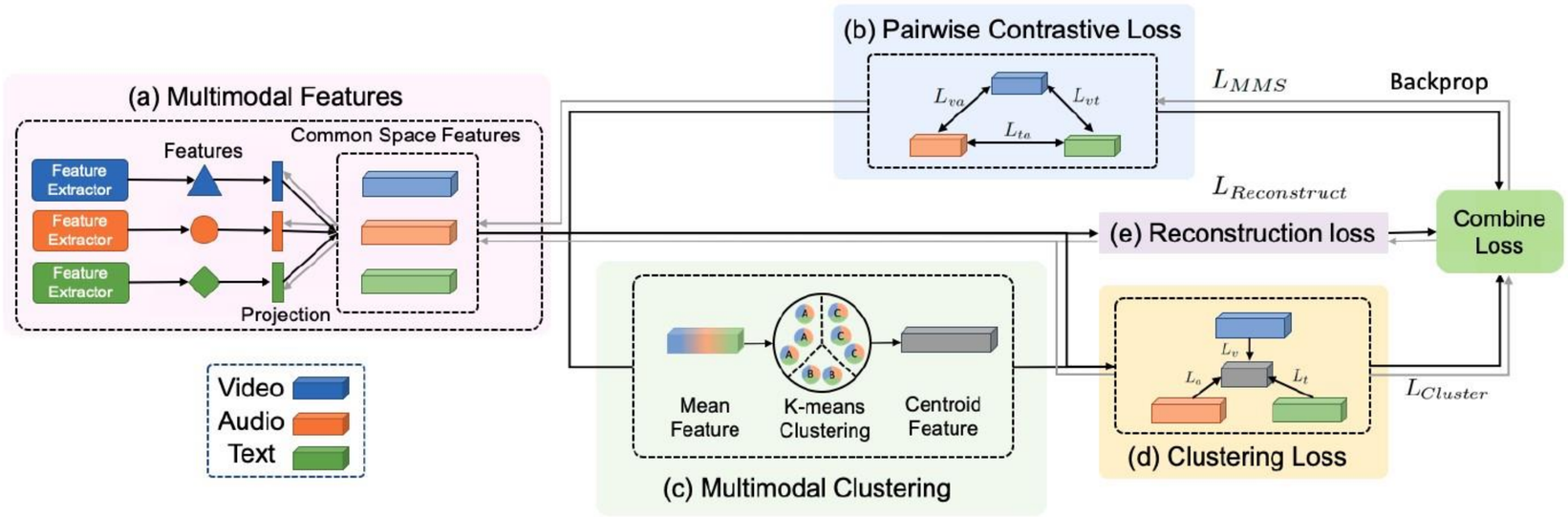
MACHINE PERCEPTION AND ROBOTICS GROUP

はじめに

- Multimodal Clustering Network (MCN)
  - Contrastive Loss $\mathrm{L}_{\mathrm{MMS}}$
  - Clustering Loss $\mathrm{L}_{\mathrm{Cluster}}$
  - Reconstruction Loss $\mathrm{L}_{\mathrm{reconstruct}}$

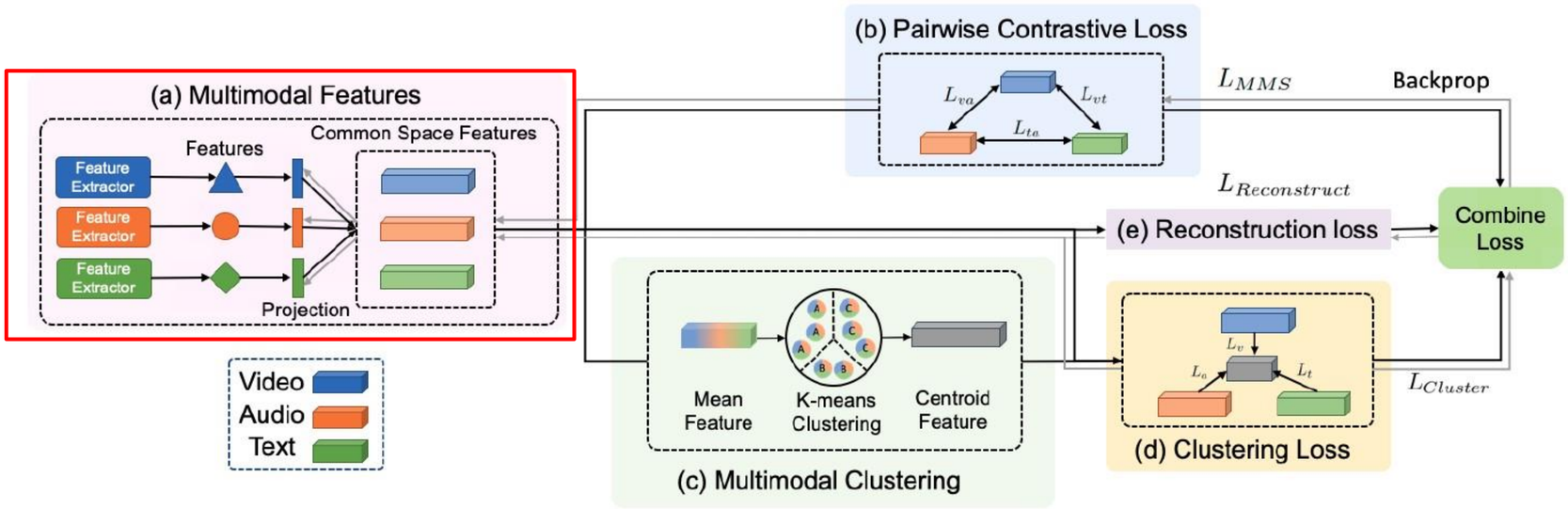
Multimodal Clustering Network (MCN) [B. Chen+, ICCV'21]

- 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 t，オーディオ a，ビデオ vの3つのモダリティを使用
  - 情報を伝達する手段や媒体



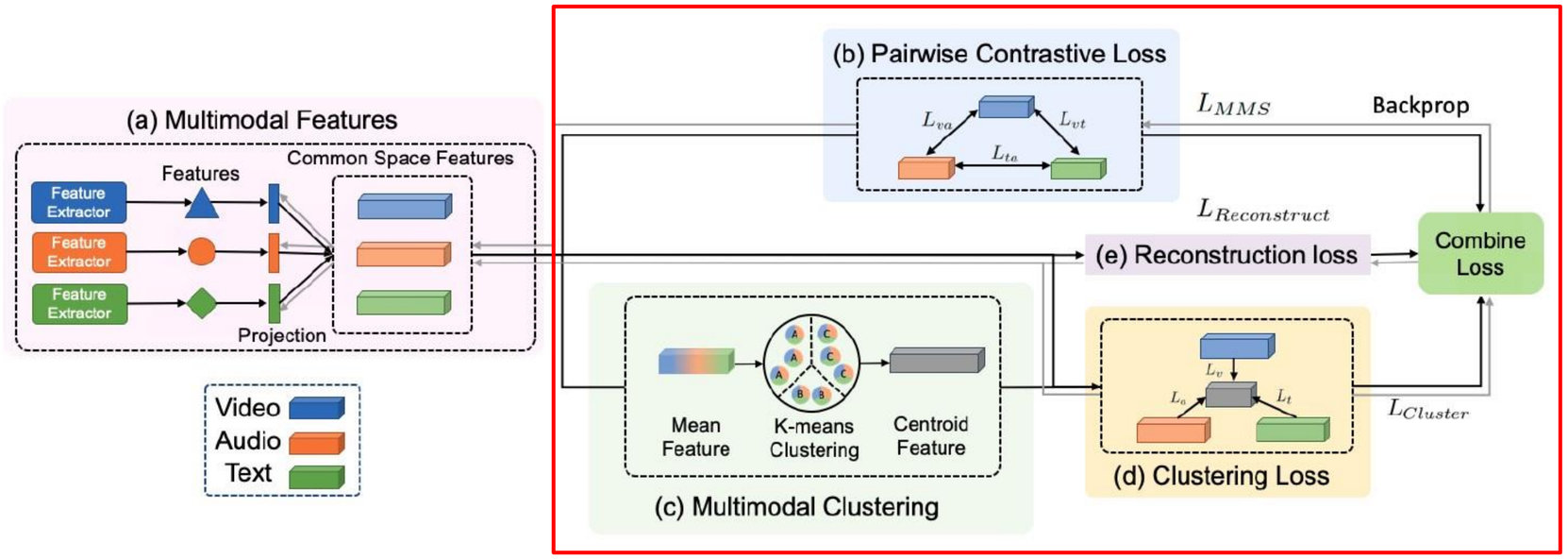
Multimodal Clustering Network (MCN) [B. Chen+, ICCV'21]

- 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



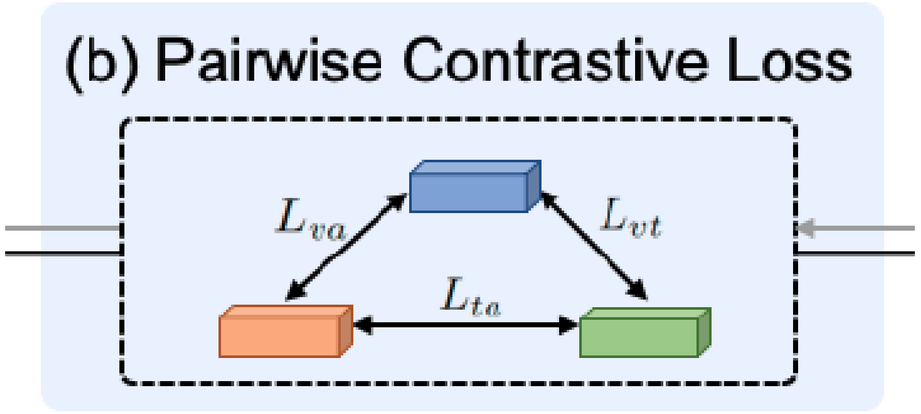
Multimodal Clustering Network (MCN) [B. Chen+, ICCV'21]

- 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 L_{MMS} [1/2]

- 同じデータの異なるモーダル間の特徴量の距離を小さくするように学習
- 異なるデータ間の特徴量の距離は大きくするように学習
- 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



Contrastive Loss

Contrastive Loss L_{MMS} [2/2]

- 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

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

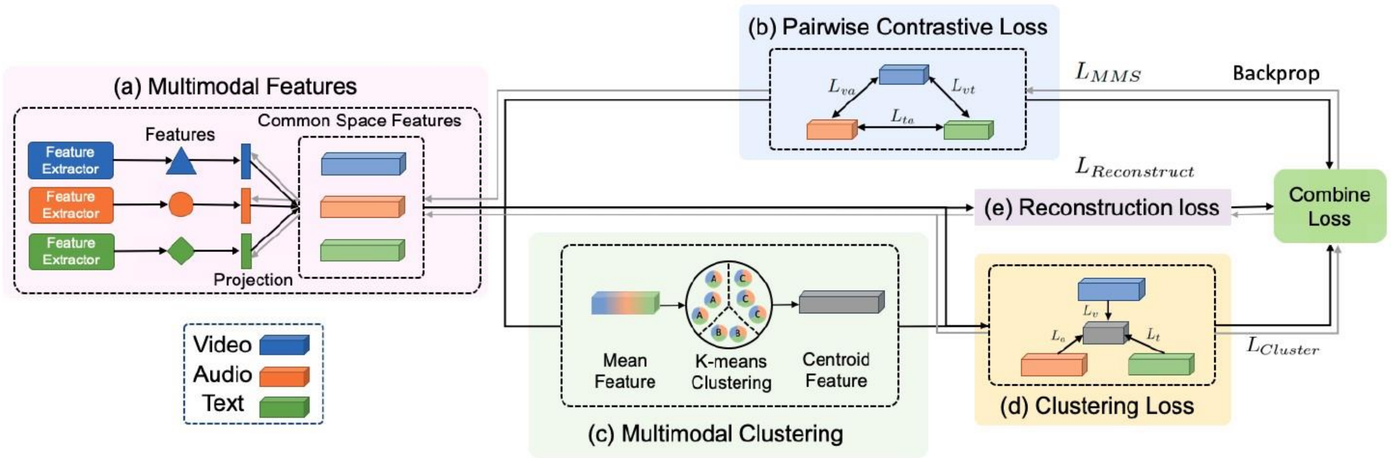
v ：ビデオ
B ：バッチサイズ
$f'(v)$ ：再構成後
$f'(v)$ ：再構成前

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

おわりに

- 今回はマルチモーダルデータの自己教師あり学習について論文調査を実施
- Multimodal Clustering Network (MCN)
  - 自己教師あり学習を用いて異なるモーダルのデータをクラスタリング
- Contrastive Loss L_{MMS}
  - データ間の特徴量の類似度を比較する損失関数
- Clustering Loss $L_{Cluster}$
  - クラスタ内においてその中心にデータ点を近づける損失関数
- Reconstruction Loss $L_{reconstruct}$
  - データを再構築して元データと比較する損失関数

- 今後の予定
  - 論文調査：MMV Networks



MCNのフレームワーク