第13回ディスカッション

論文調査と実験状況

ER20038 小林亮太

担当：鈴木雅★，福井，張

MPRG

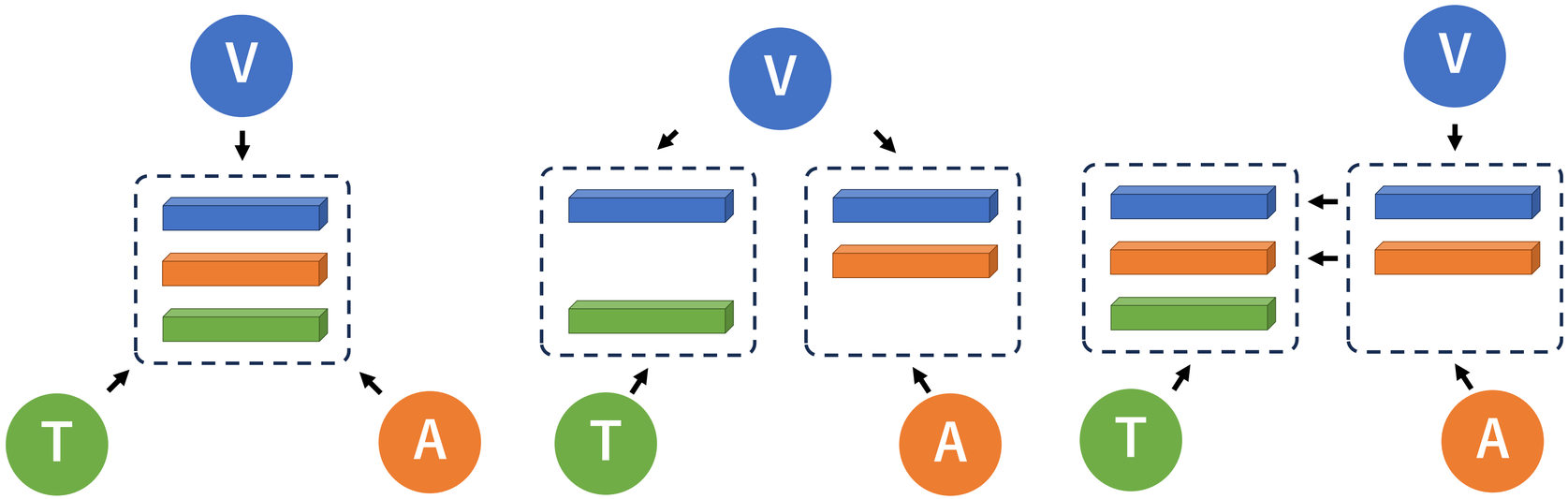
MACHINE PERCEPTION AND ROBOTICS GROUP

はじめに

- 研究テーマ
- CLIP2Point：Transfer CLIP to Point Cloud Classification with Image-Depth Pre-Training
- 今後の計画
- 実験条件
- 実験状況

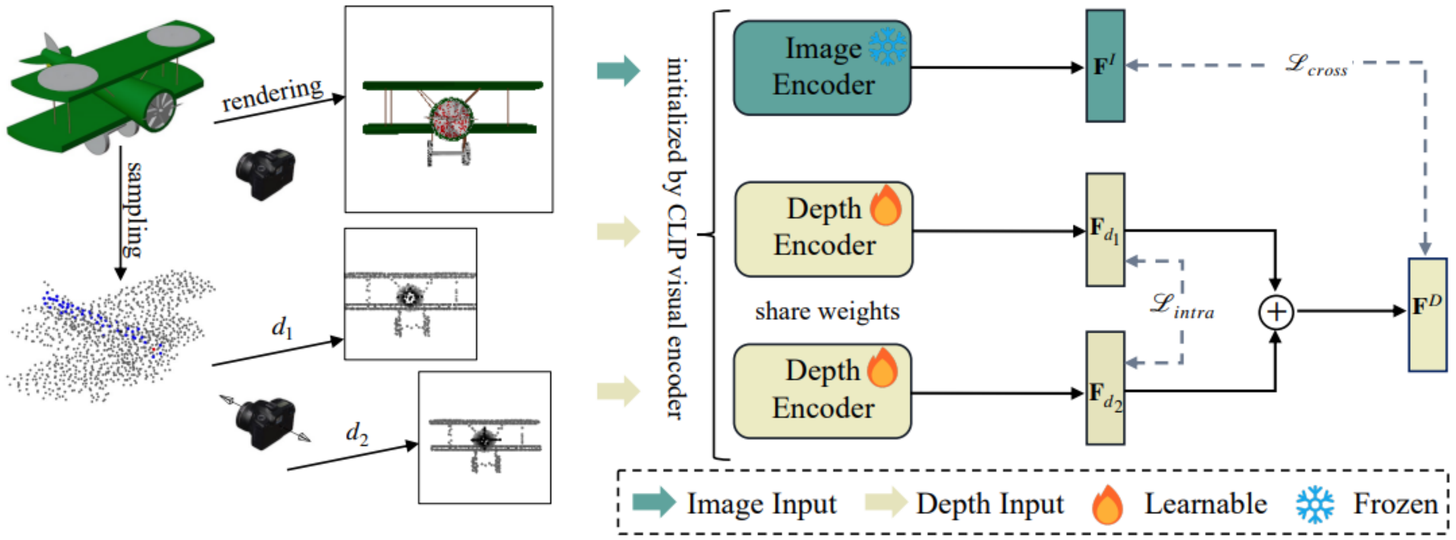
研究テーマ

- 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



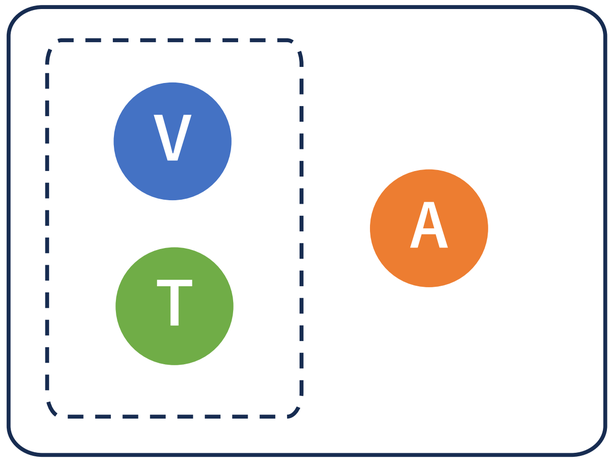
CLIP2Point：Transfer CLIP to Point Cloud Classification with Image-Depth Pre-Training [T Huang+, ICCV'23]

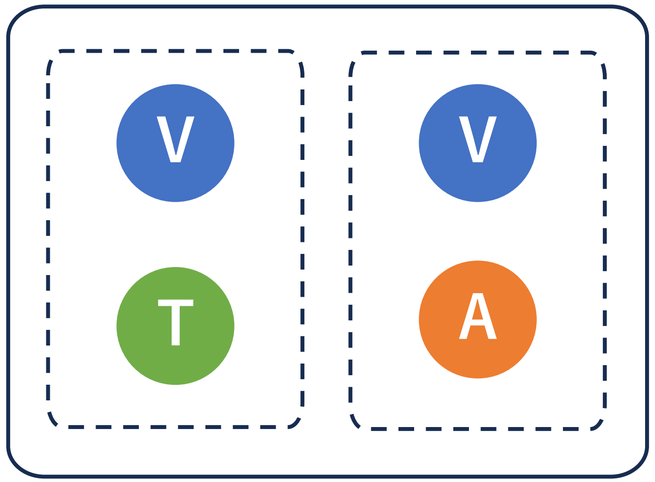
- 3D点群データから2D深度マップへ変換
- CLIP2Point
  - CLIPのImage Encoderを固定
    - RGB画像を入力
  - 新たなDepth Encoderを学習
    - 深度マップを入力
- CLIPとCLIP2Pointをアンサンブル



今後の計画

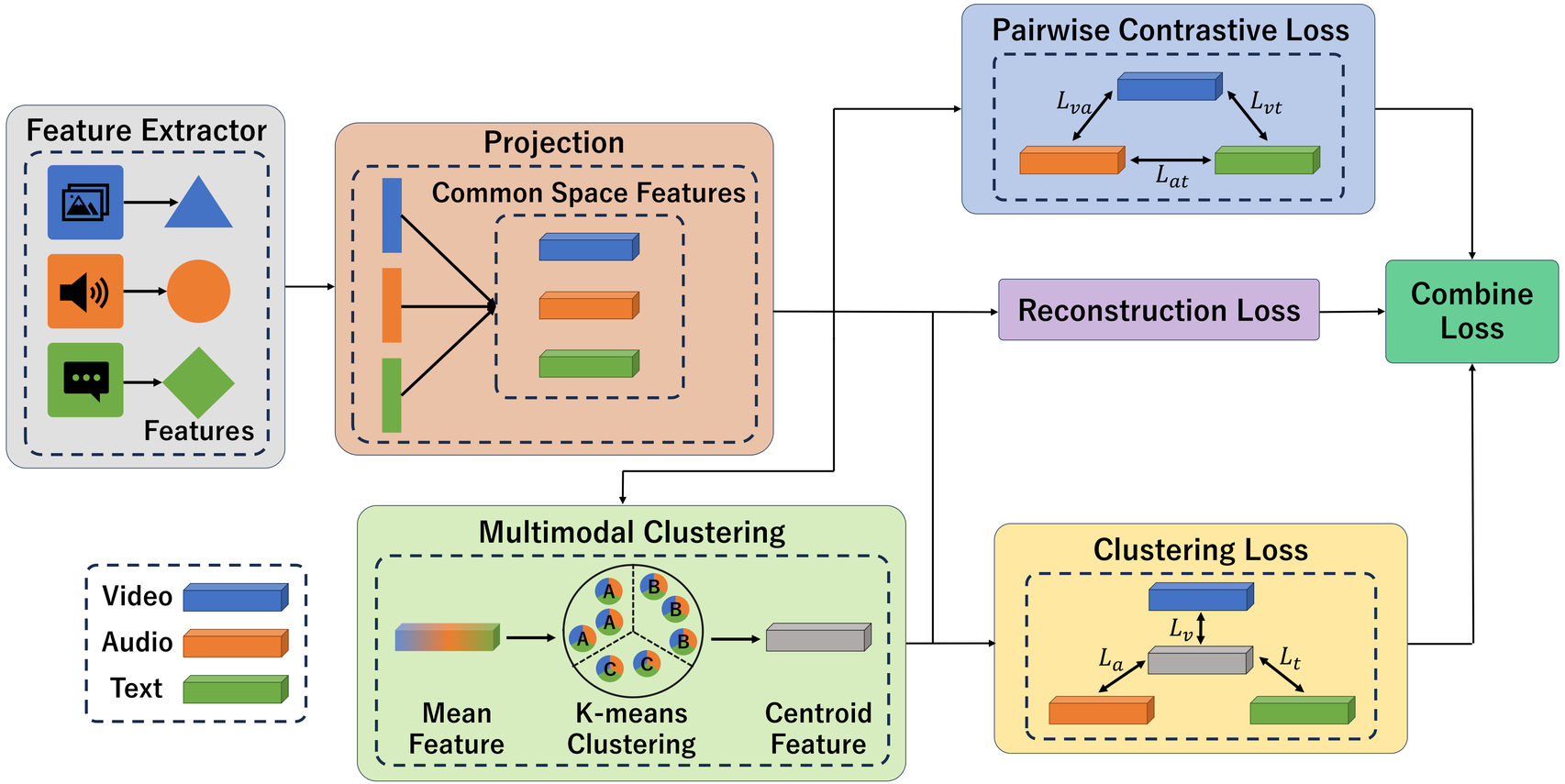
- モーダルの組み合わせによる学習効果への影響
  - CLIP^2　：3つのうち2つで学習したモデルに残りの1つを追加
  - CLIP2　：3つのうち2つの違うペアで学習した2つのモデルを融合





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

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で除算で一時的に対処
  - 根本の原因を調査中

```
  0%|
VIDEO_OUT
tensor([[[-1.1429e-02, -9.4223e-03,  5.2795e-03,  ..., -3.9005e-03,
          8.2779e-03, -1.1543e-02],
         [-9.6588e-03, -5.4398e-03,  4.7302e-03,  ..., -2.7847e-03,
          1.0201e-02, -1.5007e-02],
         [-1.0513e-02, -1.2760e-03,  9.6512e-03,  ..., -3.2101e-03,
          1.2207e-02, -1.1063e-02],
         ...,
         [-1.0300e-02, -5.4398e-03, -7.2956e-05,  ...,  4.6349e-04,
          7.3891e-03, -1.0048e-02],
         [-1.0666e-02, -8.2245e-03, -1.8892e-03,  ..., -1.4114e-03,
          5.7182e-03, -7.7820e-03],
         [-1.0567e-02, -8.8644e-04,  2.2449e-03,  ..., -5.4359e-03,
          5.3177e-03, -5.9471e-03]], device='cuda:0', dtype=torch.float16,
       grad_fn=<GatherBackward>)
AUDIO_OUT
tensor([[ 17.6250, -22.6875,   6.3008,  ...,  -6.4727,   1.6572,   2.4160],
        [ 17.0312, -22.3750,   5.8086,  ...,  -6.8555,   1.6768,   3.4199],
        [ 17.7344, -22.7969,   5.9609,  ...,  -7.2031,   1.6475,   3.8770],
        ...,
        [ 16.2031, -21.3125,   5.3633,  ...,  -6.0273,   1.6543,   2.7031],
        [ 16.9062, -22.2188,   4.8906,  ...,  -6.5977,   1.6201,   2.8516],
        [ 16.1094, -21.5000,   4.9922,  ...,  -6.0586,   1.6562,   2.3496]],
       device='cuda:0', dtype=torch.float16, grad_fn=<GatherBackward>)
TEXT_OUT
tensor([[ 0.0396,  0.0115,  0.0240,  ...,  0.0122, -0.0134,  0.0227],
        [ 0.0507, -0.0066,  0.0477,  ...,  0.0021, -0.0176,  0.0451],
        [ 0.0415,  0.0186,  0.0449,  ..., -0.0068, -0.0045,  0.0510],
        ...,
        [ 0.0148, -0.0106,  0.0240,  ..., -0.0012, -0.0060,  0.0419],
        [ 0.0172,  0.0151,  0.0511,  ..., -0.0118,  0.0119,  0.0467],
        [ 0.0264, -0.0181,  0.0294,  ...,  0.0096, -0.0373,  0.0186]],
       device='cuda:0', dtype=torch.float16, grad_fn=<GatherBackward>)
```

おわりに

- CLIP2Point
- 実験　　：実行中

- 今後の予定：
  - 実験の結果の分析
  - MCNの関連論文調査
  - プログラムの作成

資料

Gated Dual-Path Adapter (GDPA)

- CLIPとCLIP2Pointのアンサンブルに使用