第4回ディスカッション

再現実験と論文調査

ER20038　小林亮太

担当：岡本（主），張，岩垣

MPRG

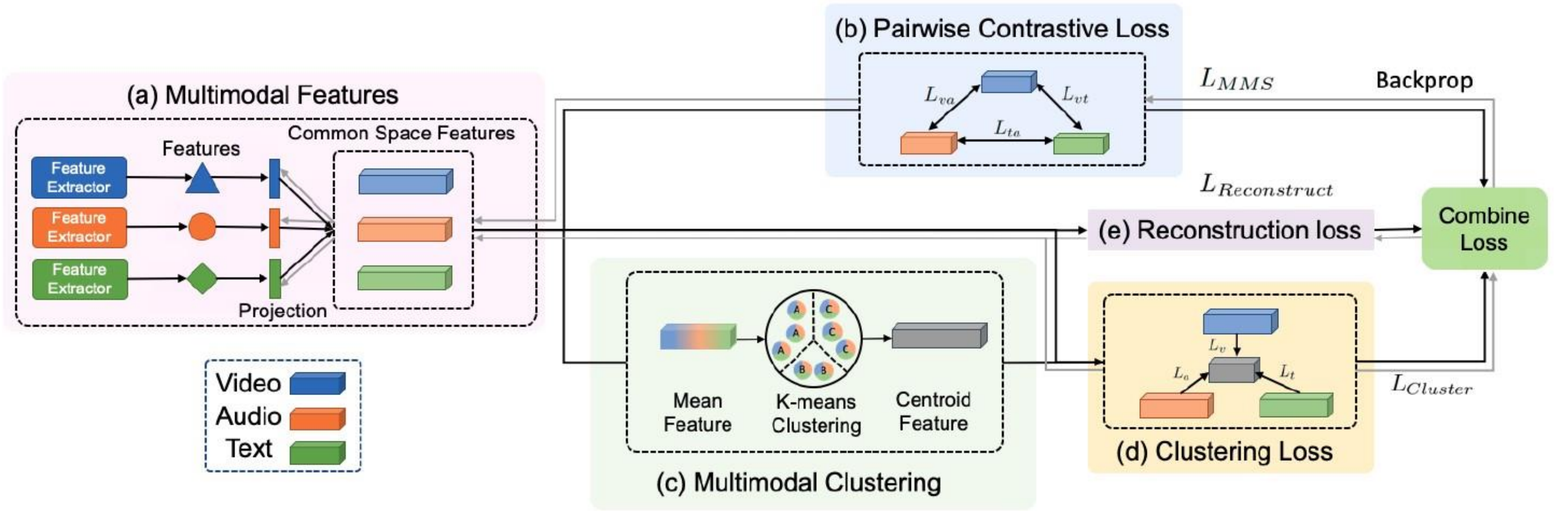
MACHINE PERCEPTION AND ROBOTICS GROUP

はじめに

- 再現実験
  - Multimodal Clustering Network (MCN)
  - 準備
  - 実行
- 論文調査
  - Everything at Once-Multimodal Fusion Transformer for Video Retrieval
  - Multimodal Fusion Transformer
  - Combinatorial Loss
  - 学習時間

再現実験

- Multimodal Clustering Network (MCN)の再現実験
- 学習にHowTo100Mデータセットを利用
  - 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
  - 23,000のカテゴリーが存在
  - 全てダウンロードする場合，12TBの容量が必要

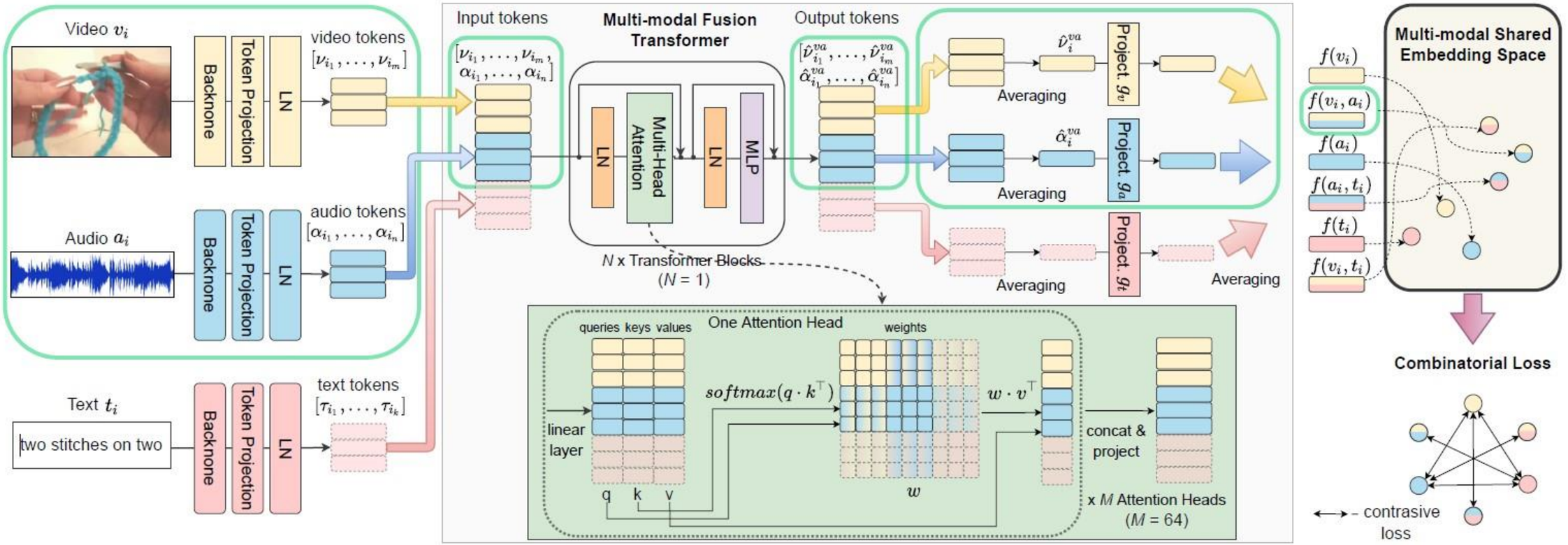


再現実験準備

- 完了済
  - プログラムのダウンロード
  - 一部データセットのダウンロード
  - 必要なモジュールのインストール
    - Numpy，Panda，librosa，torch，fast_pytorch_kmeans，apex，Tqdm，gensim
- 未完了
  - データセットのダウンロード
    - HowTo100MのYouTube上の動画データのダウンロード
      - フォームからの利用申請を行ってユーザー名とパスワードの入手が必要
    - 進捗
      - フォームを送信 (4/13)
      - 著者にメールを送信 (4/27)
      - 現在は返答待ち

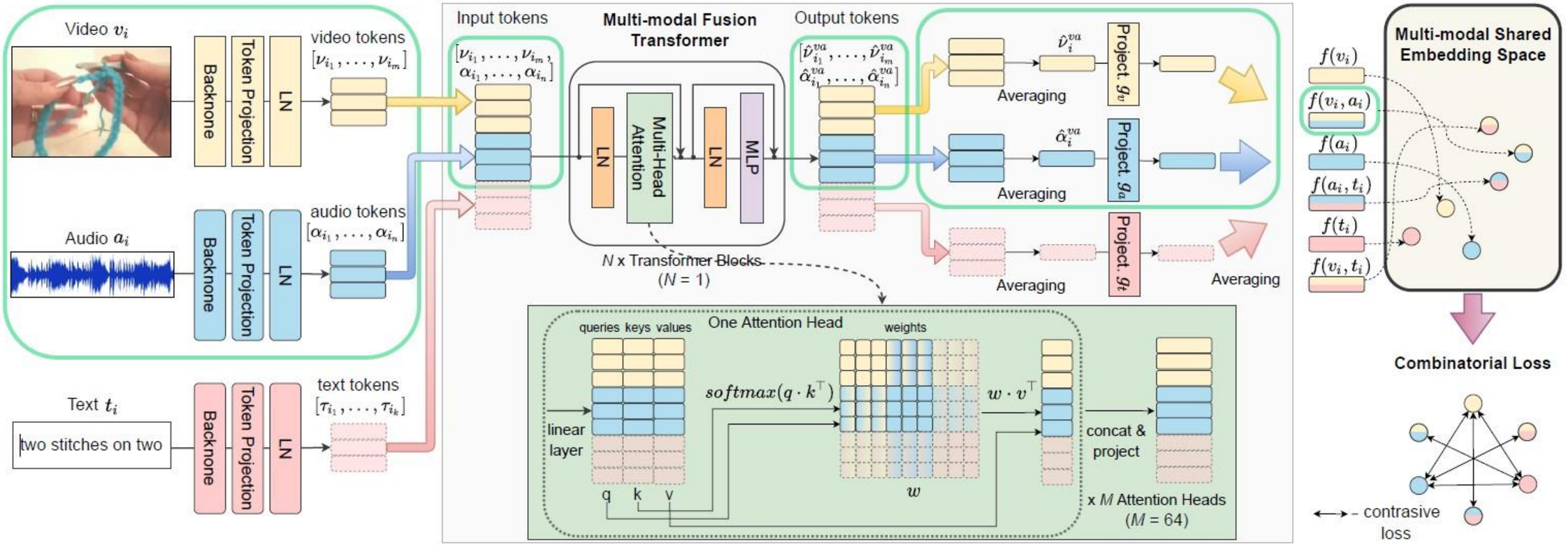
Everything at Once – Multimodal Fusion Transformer for Video Retrieval [N.Shvetsova+, CVPR'22]

- 多様なモダリティ間の学習をするMultimodal Fusion Transformer (MFT)を提案
  - Self-attention機構によって異なる長さの入力を処理することが可能
- Combinatorial Lossは，入力モダリティのすべての組み合わせに対してトレーニング
  - 真のペアであるか推定するために偽のペアと比較をして確率を出力



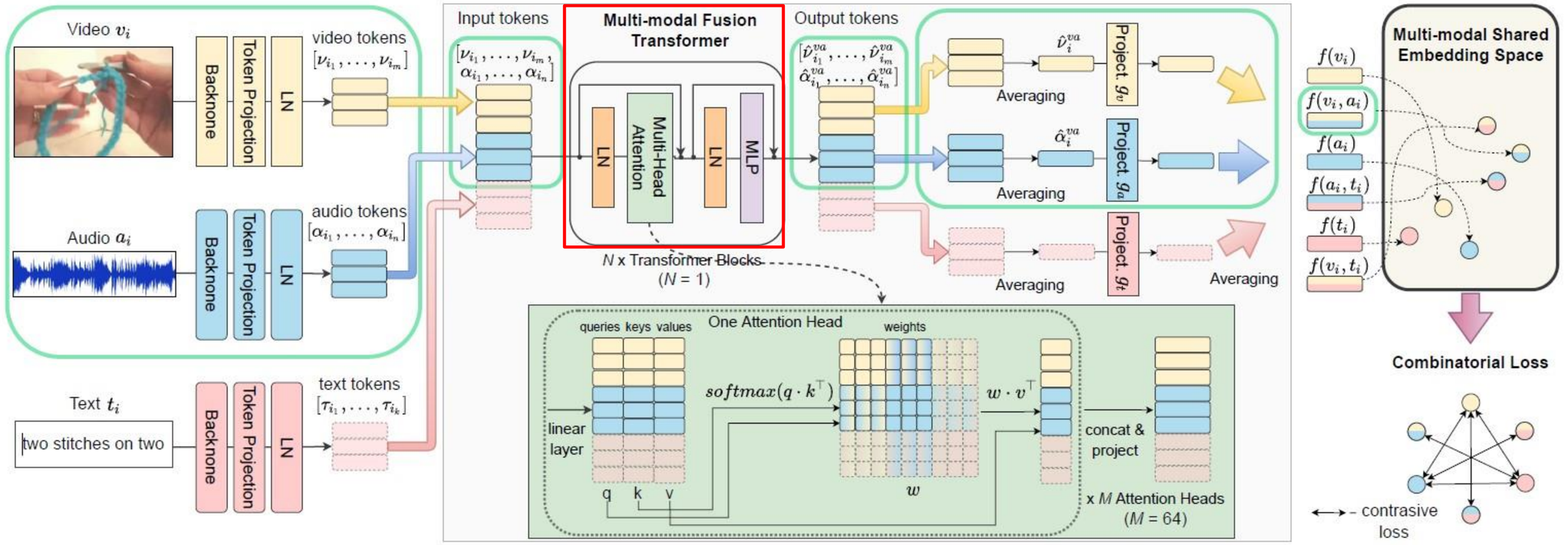
Everything at Once – Multimodal Fusion Transformer for Video Retrieval [N.Shvetsova+, CVPR'22]

- Video at once：全入力データを一度に処理
  - ローカルな時間的な依存関係を活用
  - テスト時には，任意の数の入力モダリティを処理可能



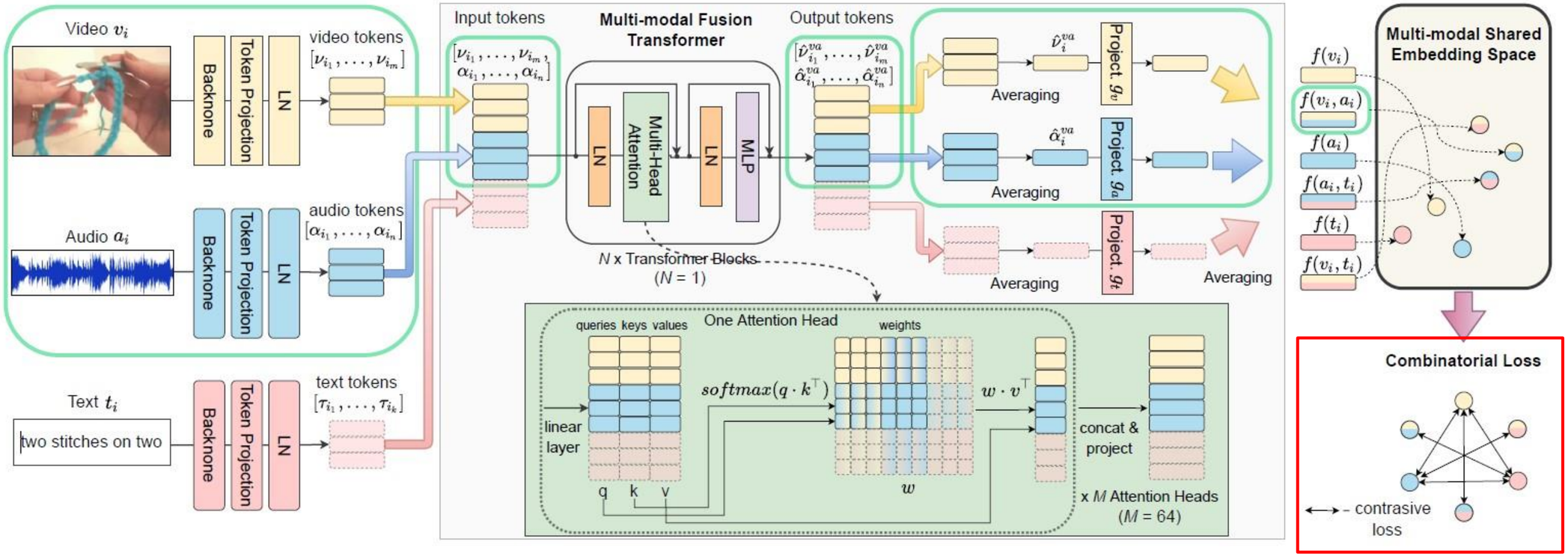
Multimodal Fusion Transformer (MFT)

- 複数の入力モダリティから情報を収集，統合することでより豊かな表現空間を作成
- 入力には各モダリティから抽出された特徴量を使用
- Transformerモデルのself-attention機構を利用
  - 各要素は自身と他すべての要素と相互に作用
  - 異なる長さの入力を処理可能



Combinatorial Loss

- すべての入力の組み合わせにおける対照損失を計算
  - 単一モダリティ：テキスト，オーディオ，ビジュアル
  - モダリティペア：テキスト-オーディオ，ビジュアル-オーディオ，テキスト-ビジュアル
- Noise Contrastive Estimation Loss (NCE)を各組み合わせに対して利用
  - 真のペアであるか推定するために偽のペアと比較をして確率を出力



学習時間

- HowTo100Mでモデルをトレーニング
  - 論文著者環境　：Nvidia V100 32GB × 4,
    所要時間　：2日
  - 自分の環境　　：Nividia A100 40GB × 3
    予想所要時間　：1日半程度
- ファインチューニング
  - 論文著者環境　：Nvidia V100 32GB × 4,
    所要時間　：30分
  - 自分の環境　　：Nividia A100 40GB × 3
    予想所要時間　：十数分程度

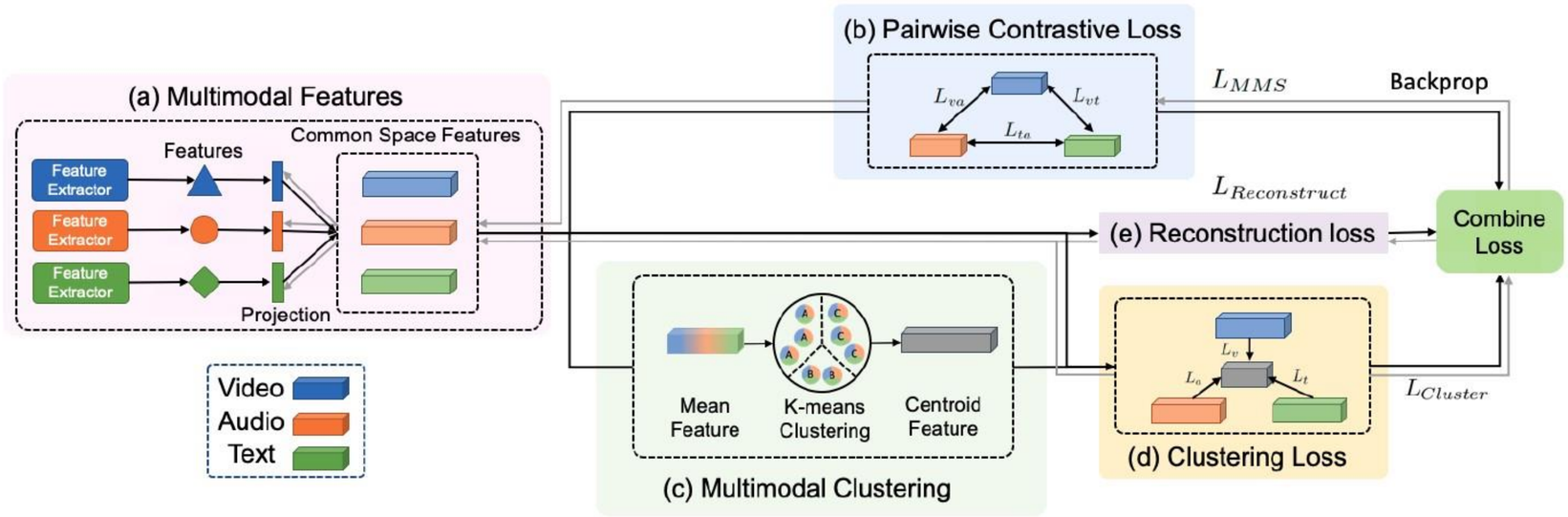
おわりに

- 再現実験
  - Multimodal Clustering Network (MCN)
- 論文調査
  - Everything at Once-Multimodal Fusion Transformer for Video Retrieval
  - Multimodal Fusion Transformer
  - Combinatorial Loss
  - 学習時間

- 今後の予定：再現実験の実行，ファインチューニング用のデータセット準備

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
- 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



再現実験実行

- 準備が完了してないため，未実行

Noise Contrastive Estimation Loss (NCE)

- NCE ： Noise Contrastive Estimation Loss
  - 真のペアであるか推定するために偽のペアと比較をして確率を出力

$$NCE(x_v, x_a) = -\log\left(\frac{\exp(z_{v,va} \cdot z_{a,va}/\tau)}{\exp(z_{v,va} \cdot z_{a,va}/\tau) + \sum_{z' \sim \mathcal{N}(x)} \exp(z'_{v,va} \cdot z'_{a,va}/\tau)}\right)$$

τ ：温度パラメータ
$\mathcal{N}(x)$ ：偽のペアの集合
$P(x)$ ：真のペアの集合
$z()$ ：ネットワークの出力
$z'()$ ：偽のペア