第9回ディスカッション

データセット状況と変更

ER20038 小林亮太

担当：鈴木★，福井，張

MPRG

MACHINE PERCEPTION AND ROBOTICS GROUP

はじめに

- HowTo100M
- 方法の変更
- フレーム数の増加
- ダウンロードの進捗

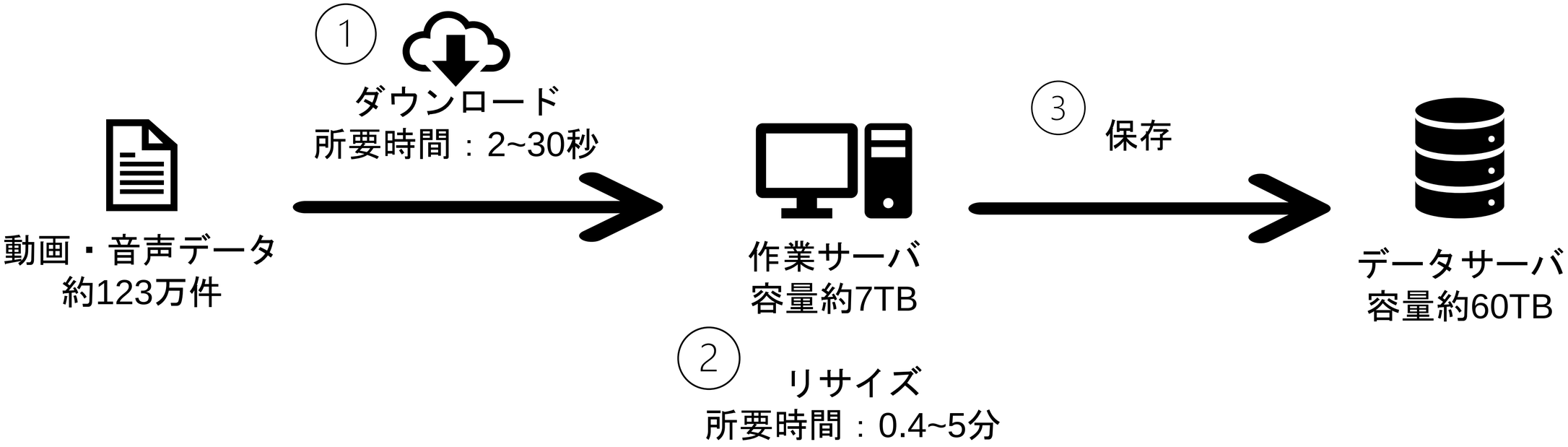
HowTo100M

- 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
  - 23,000のカテゴリーが存在
  - Youtubeの動画削除によって全体数が減少傾向
    - 現在，100万本を切っている模様
  - 大量の空き容量が必要
    - 動画データ　　：30TBから40TB
      - データサーバへ
    - 抽出された特徴量データ　：約10TB
      - 実際に学習に必要

- 様々な手法のトレーニングで使用
  - 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

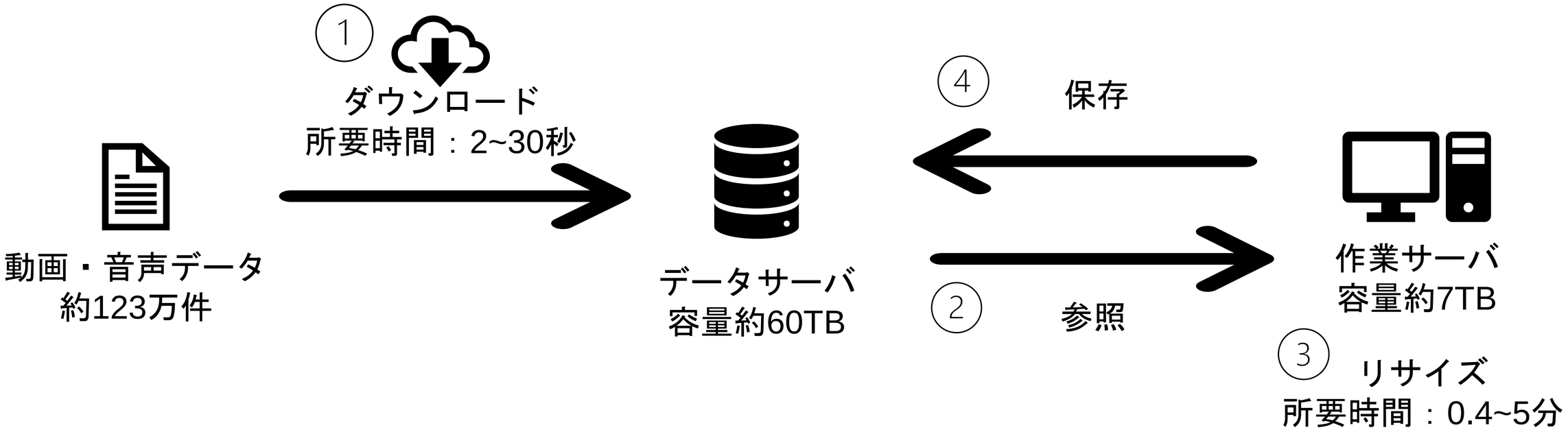
方法の変更[1/2]

- 従来
  - 1つのプログラム
  - 1番から3番をループ処理
  - リサイズ中はダウンロードが停止



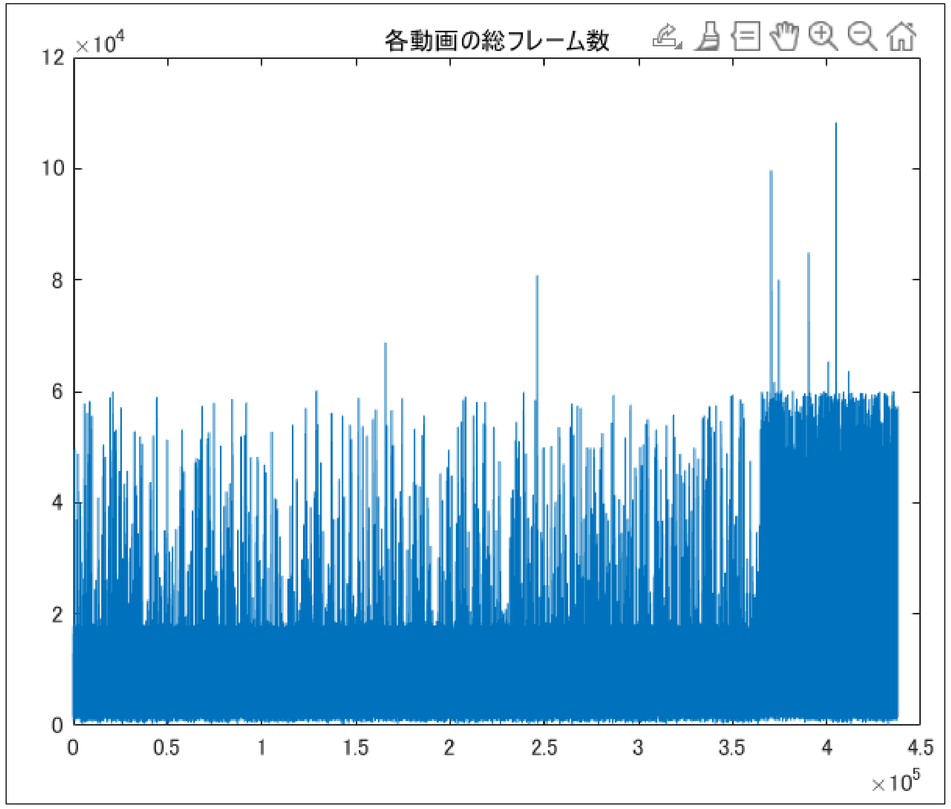
方法の変更[2/2]

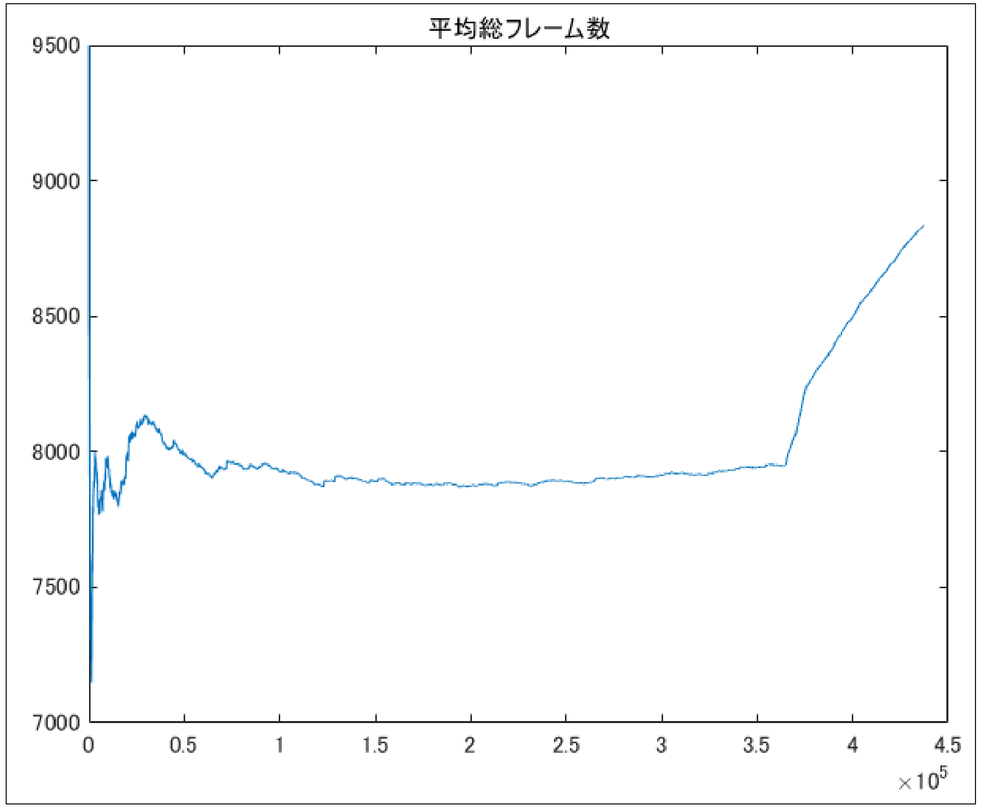
- 新規
  - 2つのプログラム
  - 1番の処理のみ
  - 2番から4番のループ処理
- ダウンロードが速いためリサイズ待ちのストックが貯まる
  - リソース次第でリサイズをより速く進めることが可能



フレーム数の増加

- 動画のフレーム数が増加傾向
  - 処理全体に遅延が発生
- 引き続き計測





進捗状況

- フレーム数の急激な増加により遅延が発生
  - 予測を立て直しが必要
- 一部データの動画，音声ともに特徴量抽出が完了
  - 一定数以上確保できた時点で再現実験を開始

進捗 予定	ダウンロード完了	リサイズ完了	特徴量抽出 音声完了	特徴量抽出 音声，動画完了
現在 8/3	760,000	550,000	450,000	390,000
8/7	1,100,000	670,000	460,000	410,000

おわりに

- 方法の変更
- フレーム数の増加　　：引き続き情報を収集
- ダウンロードの進捗　：新しいプログラムでの進捗から予想所要時間を計算しなおす

- 今後の予定：データセット準備の完了　再現実験