第7回ディスカッション

データセットの入手とテスト

ER20038　小林亮太

担当：鈴木★，福井，張

MPRG

MACHINE PERCEPTION AND ROBOTICS GROUP

はじめに

- HowTo100M
  - ダウンロード状況
  - 手順の変更
  - 動作テスト

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

ダウンロード状況

- 約5TB分のダウンロードが完了
  - 約17万本の動画と音声
- ファイルサーバの設定が完了
  - 容量は十分に足りる見込み
  - 設定以前のDL済みデータはデータサーバに移行
- プログラムに変更を加えてDL続行中
  - リサイズ前の動画は一時的にローカルのディスクに保存
  - リサイズ後の動画をデータサーバに保存

- ダウンロードは今月中にも完了する見込み

ダウンロード手順の変更

- 変更前
  - 動画データを5TB程度ダウンロード
  - 特徴量抽出
  - 動画データをデータサーバへ移動
  - 以後，繰り返し
- 変更後
  - 動画データをダウンロードしてデータサーバへ保存
  - 特徴量抽出
  - 上記２つを平行して実行

動作テスト

- データセットの入手に自前のプログラムを使用
  - 実際に使用可能であるかの確認が必要
  - 再現実験予定の手法の学習プログラムを利用
    - Multimodal Clustering Network (MCN) [B.Chen+, ICCV'21]
- データの先頭から100個使用
  - 動画：{video_id}.npzファイル　　100個分
  - 音声：{video_id}_spec.npzファイル　　100個分
  - 字幕：caption.pickleファイル　　全データ分
- 動作テスト結果
  - 正常に重みを生成
  - 生成された重みを用いて評価プログラムを実行予定

おわりに

- データセットのダウンロード
  - プログラム：高速化のための改良が必要
  - 動作テスト：自前のダウンロードプログラムの成果が正常かテスト

- 今後の予定：引き続きデータセット準備　評価プログラムの実行準備