第6回ディスカッション

データセットの入手

ER20038　小林亮太

担当：岩垣★，張

MPRG

MACHINE PERCEPTION AND ROBOTICS GROUP

はじめに

- HowTo100M
- データセット取得の準備
- データセットのダウンロード

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

HowTo100M

- 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
  - リサイズ済みのデータを入手可能
  - フォームを送信 (4/13)：現在まで返答なし
    - フォームからの申請が停止中？
    - Githubのissueに同様の問題の報告あり
- 著者にメールを送信 (4/27)：現在まで返答なし

- Youtube上の動画を利用
  - 自分でダウンロードすることが可能

データセット取得の準備

- Yt-dlpを用いた動画のダウンロード
  - Youtube-dlの派生
  - より高速なダウンロードが可能
  - Youtube-dlの代替として利用
- HowTo100Mデータセットのダウンロード
  - 公式サイトで配布されている動画IDのリストを利用

```
video_id,category_1,category_2,rank,task_id
nVbIUDjzWY4,Cars & Other Vehicles,Motorcycles,27,52907
CTPAZ2euJ2Q,Cars & Other Vehicles,Motorcycles,35,109057
rwmt7Cbuvfs,Cars & Other Vehicles,Motorcycles,99,52907
HnTLh99gcxY,Cars & Other Vehicles,Motorcycles,35,52907
EyP3HVhg1u0,Cars & Other Vehicles,Motorcycles,95,52906
w6zbbdK1ewY,Cars & Other Vehicles,Motorcycles,17,52906
RAidUDTPZ-k,Cars & Other Vehicles,Motorcycles,10,52907
nssig0FNZVU,Cars & Other Vehicles,Motorcycles,99,52906
m2YyhYyleII,Cars & Other Vehicles,Motorcycles,153,52906
u5B1RAufWVM,Cars & Other Vehicles,Motorcycles,150,52906
AEytW9ScgCw,Cars & Other Vehicles,Motorcycles,108,52906
tYQoPHwNkho,Cars & Other Vehicles,Motorcycles,18,52907
DUxVMAebfrM,Cars & Other Vehicles,Motorcycles,131,52907
M1Ta54rkeQQ,Cars & Other Vehicles,Motorcycles,54,52907
```

動画IDのリストの一部

データセット取得の準備

- プログラムの作成
  - 動画IDのリストを参照して自動で動画と音声をダウンロード
    - 動画ファイルmp4と音声ファイルm4aの2つを保存
  - 動画はダウンロード後455×256の解像度にリサイズ
    - 短辺は公式サイトの情報を参照して設定
    - 長辺は短辺より16:9の比率に基づき設定
  - ダウンロードに成功した動画IDのリストを保存
    - 動画IDに紐づくカテゴリ名も同時に保存

データセットのダウンロード

- 所要時間
  - 動画ダウンロード所要時間：小
  - 動画リサイズ処理所要時間：大
    - GPUなしでは実行に現実的ではない所要時間
    - GPUを使用して処理を行うように変更
- 手順
  - 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

おわりに

- データセットのダウンロード
  - フォーム，メールからの利用申請　：返答待ち
  - Yt-dlpを用いたダウンロード　　：プログラムが完成し次第ダウンロード開始
  - プログラム　　：高速化のための改良が必要

- 今後の予定：引き続きデータセット準備
  - OpenCV with CUDAのセットアップと組み込み
  - ダウンロードとリサイズを別々のプログラムに分離