第6回ミーティング
データセットの入手

ER20038 小林 亮太

2023年6月24日

1　はじめに

卒業研究を始めるにあたり, 調査を行った論文における手法の再現実験を行うために大規模なデータセットの入手を目標とした.

2　HowTo100M データセット

HowTo100M が現在入手を目的としているデータセットである．以下にこのデータセットに関する情報をまとめる.

2.1　概要

HowTo100m はナレーション付きビデオの大規模なデータセットであり，120 万本の youtube 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によって構成されている．また，23000 のカテゴリーが存在している．3 モーダルを利用する様々な手法で使用されており，今までに調査した 3 つの論文の手法全てでトレーニングに使われていた．そのため，入手することができれば非常に有益であると考えられる．しかしながら，youtube 上の動画利用しているという点が仇となり時間が経つほどデータセット全体の総数が減少してしまう．現在では，100 万本を切っているだろうという目測が立っているほどに減少している.

2.2　ダウンロード方法

HowTo100m の入手方法として公式から案内さているのは，フォームからの利用申請であるがそのフォームが申請フォームとしてすでに機能していないようである．これは github の issue にも同じ問題を抱えるコメントが見受けられたため改善もあまり期待できない．また，フォームには一週間以内に返答がなければ直接メールを出してくれという案内があったが，こちらも機能しておらず公式の案内する方法ではダウンロードできないというのが現状である．そこで，このデータセットが youtube 上の動画を利用している点から自分でデータセットを揃えることにした．その具体的な方法を次に示す.

2.3 yt-dlp を用いたダウンロード

yt-dlp とはコマンドラインから youtube の動画をダウンロードすることができるものである．yt-dlp では youtube 上での動画の ID さえあれば動画の画質や音声の有無，拡張子をカスタマイズしてダウンロードできる．HowTo100M の公式サイトで配布されているファイルには動画 ID とそれに対応したカテゴリがリストになっているファイルがあるため利用した．これらを用いて自動でダウンロードを行うプログラムを作成した．その処理内容を次に示す．

2.4 プログラム

作成したプログラムには以下のような機能を組み込んだ．

- 動画を mp4 ファイル，音声を m4a ファイルでダウンロード
- 動画はダウンロード後に 454 × 256 の解像度にリサイズ
- ダウンロードに成功した動画 ID のリストを保存

リサイズする解像度は短辺の 256 を公式サイトの情報を基に決定し，長辺を元の動画のアスペクト比に合わせて決定した．また，もとは 640 × 360 の解像度でダウンロードしている．
このプログラムの所要時間について，動画のダウンロードのみであれば 100 万本をダウンロードしたとしても 2 週間かからない程度であるが，動画のリサイズ処理にかなりの時間がかかり，GPU なしで動作チェックをした結果から計算すると現実的ではない時間を要してしまうため，GPU を使用するような処理に変更する予定である．また，プログラムを動作させた場合の容量などの問題に関しては，次に具体的な計画の説明で説明する．

3 実行

プログラムを実行する際はデータセットの容量の大きさに気を付けなければならない．このデータセットは mp4 による動画データでは合計 30 40TB にもなる見込みであるが，実際に学習で使用するのは動画から抽出された特徴量でありこれは約 10TB の見込みである．そこで，実行の際は次のような手順にしたがって行う．

- 動画データを使用しているディスクの空き容量ギリギリまでダウンロード
- ダウンロードした動画から特徴量を抽出
- 動画データをデータサーバ（ファイルサーバ）へ移動
- 以後，繰り返し

このような手順で実行することにより大規模なデータセットを準備すると共に mp4 ファイルを保持することによって別の手法で異なる抽出器を用いる場合でも対応が可能となる．

4 おわりに

今回はデータセットの概要とそのダウンロード方法についてまとめた．再現実験の準備を開始してからデータセットの準備のみが滞っていたが，自力での用意が現実的になってきた．今後はプログラムの処理の高速化，opencv で cuda を有効にして GPU を使用可能にする必要がある．