

# 第8回ミーティング

## データセットの入手

ER20038 小林 亮太

2023 年 7 月 22 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。また、ダウンロードしたデータが正常に機能するかを確認するための試運転を行った。

### 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れているのは、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、こちらで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。また、データサーバのセットアップが完了したことに際して対応するように変更を加えた。

- 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  $454 \times 256$  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  $640 \times 360$  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した。加えて、データサーバはプログラムを実行する場所とネットワークで繋がっているため、この間でデータのやり取りを可能な限り減らすようにした。具体的には、リサイズ前の動画は一度ローカルに保存してリサイズ後のデータをデータサーバに保存するといった方法をとった。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

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ため、プログラムを実行する際は十分に注意する必要があったが、データをデータサーバに直接保存できるようになったため、より大胆な行動ができるようになった。まず、以前の手順を以下に示す。

- 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が、この手順では動画データ

をデータサーバに送る作業が完全に完了するまでは他の作業を平行して行うことが難しかった。そこでデータをデータサーバに直接書き込めることを利用して新たに以下のような手順で作業を行う。

- 動画データをデータサーバに全てダウンロード
- ダウンロードされたものから順に特徴量を抽出
- 以上の2つの作業を常に平行して実行

この手順をとることによって、かなりの所要時間の圧縮が可能となる。ただし、この手順をとるうえではサーバのマシンへの負荷に気を付ける必要がある。

## 4 所要時間

現在までに、約 15TB の約 46 万本のダウンロードが完了しており、動画データのダウンロードのみであれば、今月中にも完了する見込みである。また、特徴量抽出の所要時間は未だデータ不足で予測できる状態ではなく来週の月曜日にはおおよその予測が立てられると考えている。

## 5 データのテスト

現在データセットを収集するために使用しているコードは8割はオリジナルのプログラムとなっており、その処理も単純にダウンロードするのみでなく、リサイズなどのデータの形状を変えるような処理を行っているため、そのまま再現実験を行うことができる保証がない。そのため、今現在ダウンロードの初期段階であるうちに処理の完了した少量のデータセットを用いて学習プログラムを実行してみる必要がある。ここではその詳細についてまとめる。

まず、試運転に用いる学習プログラムは Multimodal Clustering Network (MCN) [B Chen+, ICCV'21] のものを用いる。また、使用するデータセットの数は先頭から 100 個とした。

実行結果としては、エラーが起こることなくプログラムが終了し、重みが生成されていることが確認された。このことからダウンロードしたデータは正常に機能しているということが分かった。

## 6 評価プログラム

学習プログラムの用いて試運転を行った結果、正常に動作して重みを入手することができた。そのため、生成された重みが正常に使用できるかを確認するために評価プログラムを用いて追加の検証を行った。

準備が完了していないため未実行

## 7 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。さらに今回のテスト

によって正常に機能することも確認できた。次は今回生成された重みを用いて評価プログラムを動かしてさらに確認をしていく予定である。