P04 深層学習を用いた動画のタイトルからタグを予測する自動ジャンル分け機能

佐口 航（日本工学院専門学校）

目的（得られる効果）

・投稿されたばかりの動画を**視聴者に再生されやすく**する
・従来よりも視聴者の**ニーズに合う動画を表示**する
・共通するメジャーなタグを**登録する手間を省く**

訓練データと正解ラベル（タグ）

訓練データ
・動画のタイトル

正解ラベル
・動画のタグ

タグ一覧

・アニメ	・ニコニコ技術部	・動物
・ゲーム	・アイドルマスター	・料理
・エンタメ	・東方	・日記
・音楽	・VOCALOID	・自然
・スポーツ	・例のアレ	・科学
・歌ってみた	・生放送	・歴史
・踊ってみた	・政治	・ラジオ
・演奏してみた	・ニコニコ動画講座	
・描いてみた	・静画	

図 1 訓練データと正解ラベル（タグ）

開発環境

・Python 3.8
・Anaconda3 (2020.07)
・Spyder 4.1.4
・TensorFlow (GPU) 2.0.0 ※ 1
・Keras 2.3.1
・NVIDIA cuda 10.0
・NVIDIA cuDNN 10.0
・NVIDIA Graphics Driver 441.28
・Microsoft Visual Studio 2017 C++ ※ 1
・Microsoft Windows 10 Pro 64bit 1909

※ 1 KerasをGPUで動作させるため

データの変換方法

図2のように、String型の文字列を、int型のUnicodeのコードポイントに変換することで機械が学習を行えるようにしました。ベクトル化にvectorize_sequences関数を、カテゴリ化にto_one_hot関数を使いました。

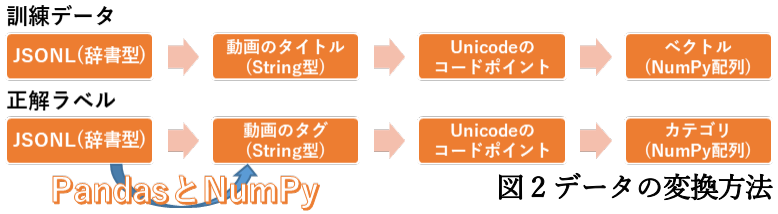

図 2 データの変換方法

学習モデル(ニューラルネットワーク)の仕様

図3のように、入力層は**絵文字に対応**するため、**Unicodeのコードポイントを10万まで許容**したので、**ノード数が10万**になっています。中間層はPCの性能が許す限りノードとレイヤーを増やしたので、学習モデルのファイルサイズは**約825MB**になりました。

	ノード数	レイヤー数	活性化関数
入力層	10万個	1層	-
中間層	1,000個	9層	relu
出力層	128個	1層	softmax

表 1 学習モデルの仕様1

	コンパイル設定
optimizer	rmsprop
loss	categorical_crossentropy
metrics	accuracy

表 2 学習モデルの仕様2

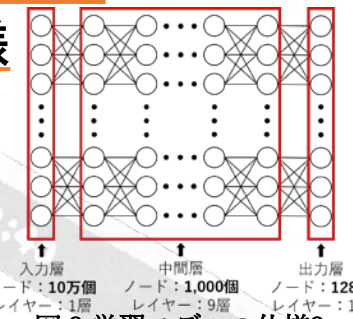

図 3 学習モデルの仕様3

過学習を防ぐ

過学習とは、規則ではなく答えを覚えてしまう現象です。つまり、過学習が起きるとテストケースでは正解率が高くても、本番環境では正解率が下がってしまいます。これを改善するために、epochsの値に注目しました。**epochsとは、「同じ訓練データを何回繰り返して学習させるのか」**の回数の事です。図4の損失関数のグラフを見るとepochsの値が**15**を超えたあたりから、学習の精度の変化が鈍くなることが分かりました。過学習を防ぐために、**学習に効果のある回数で止める**ため、epochsの値を50から**15**に下げ過学習を防ぎました。

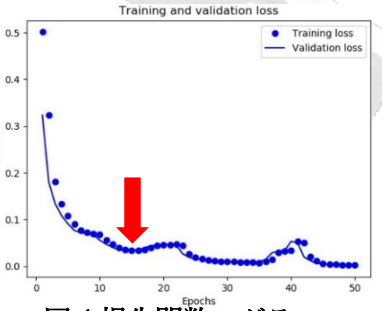

図 4 損失関数のグラフ

結果

正解率は、25種類のタグのうち1つ以上登録されている動画のタイトルを出題し、予測されたタグが、実際の動画のタグに含まれているかで正解率を求めた結果、表3のように、約**80.05%**になりました。

出題件数	正解数	正解率
12960354件	10375219件	約**80.05%**

表 3 正解率

考察・今後の展望

以上から、動画のタイトルとタグの規則(関係性)を学習させることができました。今回は文字単位で学習を行いましたが、今後は、単語単位の学習に挑戦して精度が上がるのか下がるのか実験してみたいと思いました。特に、”【】”(すみつきかっこ)や”.”(ドット)は単語ではないので、どのような結果になるのか興味深く思っております。

ソースコードと学習モデルを公開！

私的利用の範囲内でご利用頂けます。
https://github.com/SaguchiWataru

参考文献

・よくわかるPython[決定版]
・PythonとKerasによるディープラーニング（(株)ドワンゴ提供） 2018-12-14 更新

データセット

ニコニコ動画コメント等データ

※本研究では、国立情報学研究所様のダウンロードサービスにより株式会社ドワンゴ様から提供を受けた「ニコニコ動画コメント等データ」を利用しております。