Springboard Capstone Project Report #2

Using ML and Deep Learning techniques for text classification tasks.

Dmitri Kochubei (09.08.2020) | GitHub repo

Lex Fridman: So, what advice do you have for someone who wants to get started in deep learning?

Jeremy Howard: Train lots of models...

Motivation

1. The motivation behind this project is not to solve any particular problem. The project was completed to gain hands-on NLP experience. A variety of DL and ML techniques were utilized in order to obtain performance metrics and be able to compare efficiency of different approaches.

Dataset & Problem Statement

The dataset consists of texts of the reviews for 50000 movie titles and the corresponding labels (positive/negative).
The problem that is being solved is the following: Create a DL/ML model to correctly classify textual reviews according to the reviews' sentiment. This is a binary classification problem.

Pre-processing

Any ML/DL area (Computer Vision, NLP, Signal Processing ...) has its own challenges while data pre-processing. In NLP the main problem is to find a way of representing textual data in the form of vectors. There are many ways in which it can be accomplished. For this project two techniques were used:

- For the ML approach one-hot encoding was used.

- For the DL approach pre-trained word embeddings were used. In particular Stanford's GloVe representation of size 300 was used for the embedding layer.
  (Common Crawl (42B tokens, 1.9M vocab, uncased, 300d vectors, 1.75 GB):
  glove.42B.300d.zip)

Further the DL models were compiled in such a way so that GloVe vectors could be updated in order to facilitate transfer learning process.

Modeling and performance

#	Model	Update GloVe	Val_Acc	AUC (test)	Recall(Neg)	Recall(Pos)	f1-score (Neg)	f1-score (Pos)
1	LSTM	TRUE	0.8691	0.8804	0.86	0.91	0.91	0.85
2	Conv1D	TRUE	0.8943	0.5053	0.5	0.54	0.93	0.08
3	Dense	TRUE	0.8943	0.8657	0.87	0.86	0.85	0.88
4	LSTM	FALSE	0.5706	0.8028	0.88	0.72	0.82	0.79
5	Conv1D	FALSE	0.8766	0.8820	0.87	0.89	0.88	0.88
6	Dense	FALSE	0.7914	0.8112	0.84	0.79	0.82	0.81
7	Logistic Regression	-	0.8873	0.8977	0.91	0.89	0.89	0.91