FASTTEXT

Ihor Kroosh, Tim Nieradzik

22th July 2016

Ukrainian Catholic University

Bag of Tricks for Efficient Text Classification

Armand Joulin Edouard Grave Piotr Bojanowski Tomas Mikolov
Facebook AI Research
{ajoulin,egrave,bojanowski,tmikolov}@fb.com

Abstract

This paper proposes a simple and efficient approach for text classification and representation learning. Our experiments show that our fast text classifier `fastText` is often on par with deep learning classifiers in terms of accuracy, and many orders of magnitude faster for training and evaluation. We can train `fastText` on more than one billion words in less than ten minutes using a standard multicore CPU, and classify half a million sentences among 312K classes in less than a minute.

1 Introduction

Building good representations for text classification is an important task with many applications, such as web search, information retrieval, ranking and document classification (Deerwester et al., 1990; Pang and Lee, 2008). Recently, models based on neural networks have become increasingly popular for computing sentence representations (Bengio et al., 2003; extension of these models to directly learn sentence representations. We show that by incorporating additional statistics such as using bag of n-grams, we reduce the gap in accuracy between linear and deep models, while being many orders of magnitude faster.

Our work is closely related to standard linear text classifiers (Joachims, 1998; McCallum and Nigam, 1998; Fan et al., 2008). Similar to Wang and Manning (2012), our motivation is to explore simple baselines inspired by models used for learning unsupervised word representations. As opposed to Le and Mikolov (2014), our approach does not require sophisticated inference at test time, making its learned representations easily reusable on different problems. We evaluate the quality of our model on two different tasks, namely tag prediction and sentiment analysis.

2 Model architecture

A simple and efficient baseline for sentence classification is to represent sentences as bag of

Figure 1: Bag of Tricks for Efficient Text Classification (Joulin et al.)

Goal

Speed up training models for Sentiment Analysis

Key idea

Hashing of n-grams

2-grams: $(S_{t-1} \cdot P_1) \bmod N$

3-grams: $(S_{t-2} \cdot P_1 \cdot P_2 + S_{t-1} \cdot P_1) \bmod N$

t: Current word

S: Vocabulary indices

N: Number of buckets in hashing vector

P_n: Large random prime number

Hashing vector $H = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 1 \\ 0 \end{pmatrix}$

Word vector $W_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}$ $W_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$

$|V|, |W_n| = |V|$ (vocabulary size)

Operations for W_n: concatenation, averaging

- **Epochs:** 5
- **Samples per Epoch:** 1000
- **CPU:** 2.6 GHz Intel Core i5

OneHot	ContextHashes	**Time**	**Accuracy**
Concat	$\times$	529s	67%
Avg	$\times$	39s	68%
$\times$	✓	58s	74%
Concat	✓	567s	73%
Avg	✓	101s	73%

Available on GitHub

github.com/poliglot/fasttext

Questions?