facebook

fastText

a library for efficient text classification and word representation

Piotr Bojanowski

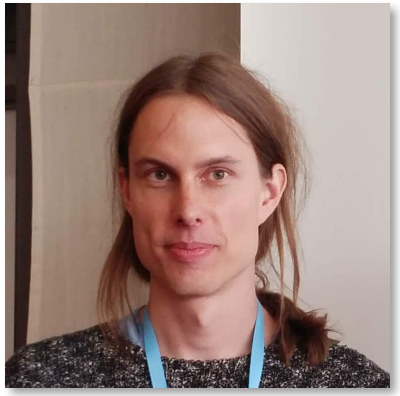
November 23th, 2016

Collaborators

Piotr Bojanowski

Edouard Grave

Armand Joulin



Tomáš Mikolov

Bag of Tricks for Efficient Text Classification

Fast text classification

- BoW model on text classification and tag prediction

Starsmith (born Finlay Dow-Smith 8 July 1988 Bromley England) is a British songwriter producer remixer and DJ. He studied a classical music degree at the University of Surrey majoring in performance on saxophone. He has already received acclaim for the remixes he has created for Lady Gaga Robyn Timbaland Katy Perry Little Boots Passion Pit Paloma Faith Marina and the Diamonds and Frankmusik amongst many others.

ARTIST

Rikkavesi is a medium-sized lake in eastern Finland. At approximately 63 square kilometres (24 sq mi) it is the 66th largest lake in Finland. Rikkavesi is situated in the municipalities of Kaavi Outokumpu and Tuusniemi.Rikkavesi is 101 metres (331 ft) above the sea level. Kaavinjärvi and Rikkavesi are connected by the Kaavinkoski Canal. Ohtaans strait flows from Rikkavesi to Juojärvi.

Natural Place

- A very strong (and fast) baseline, often on-par with SOTA approaches
- Ease of use is at the core of the library

```
./fasttext supervised -input data/dbpedia.train -output data/dbpedia

./fasttext test data/dbpedia.bin data/dbpedia.test
```

Model

- Model probability of a label given a paragraph

feature for paragraph $\mathcal{P}$: $h_{\mathcal{P}}$

classifier for label l: v_l

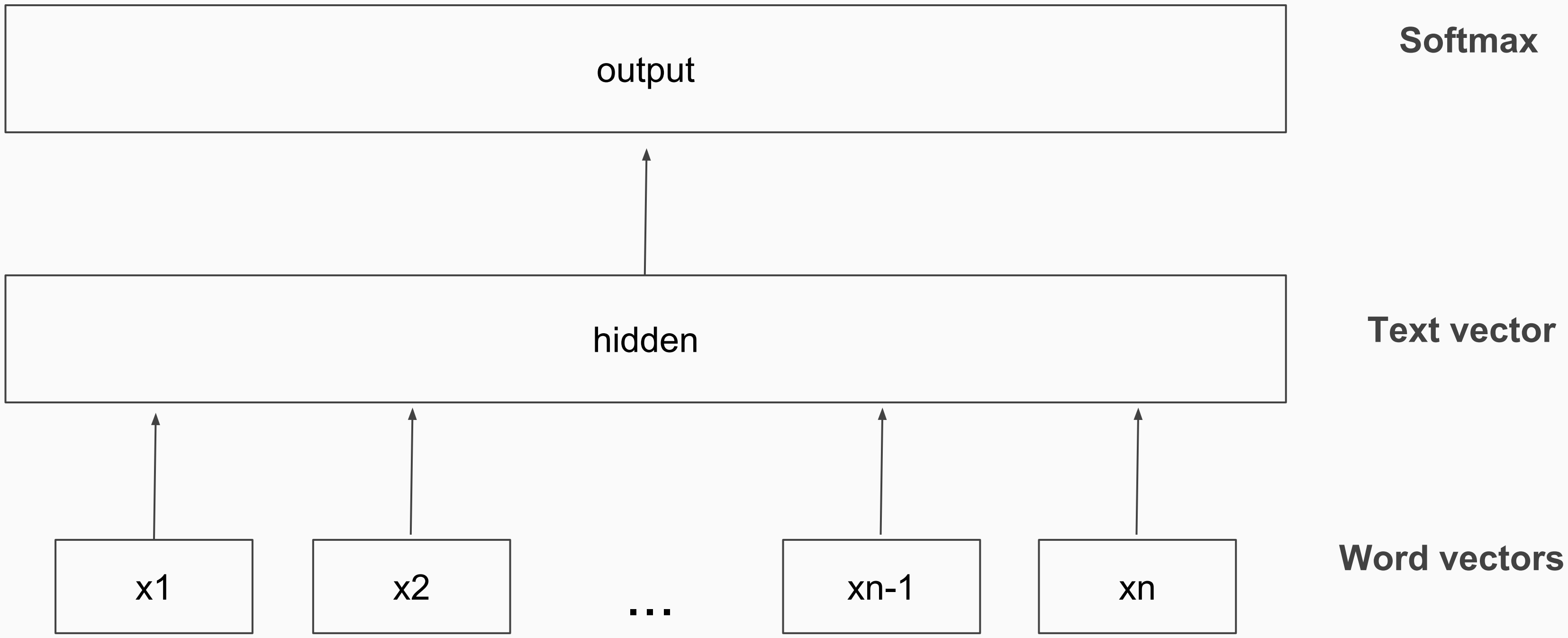
$$p(l|\mathcal{P}) = \frac{e^{h_{\mathcal{P}}^{\top} v_l}}{\sum_{k=1}^{K} e^{h_{\mathcal{P}}^{\top} v_k}}$$

- Paragraph feature

$$h_{\mathcal{P}} = \sum_{w \in \mathcal{P}} x_w$$

- Word vectors are latent and not useful *per se*
- If scarce supervised data, use pre-trained word vectors

Simple linear model



output

hidden

x1

x2

...

xn-1

xn

Softmax

Text vector

Word vectors

Learning

weight matrices

label of
n-th doc

normalized bag of
features of n-th doc

$$-\frac{1}{N}\sum_{n=1}^{N} y_n log(f(BAx_n))$$

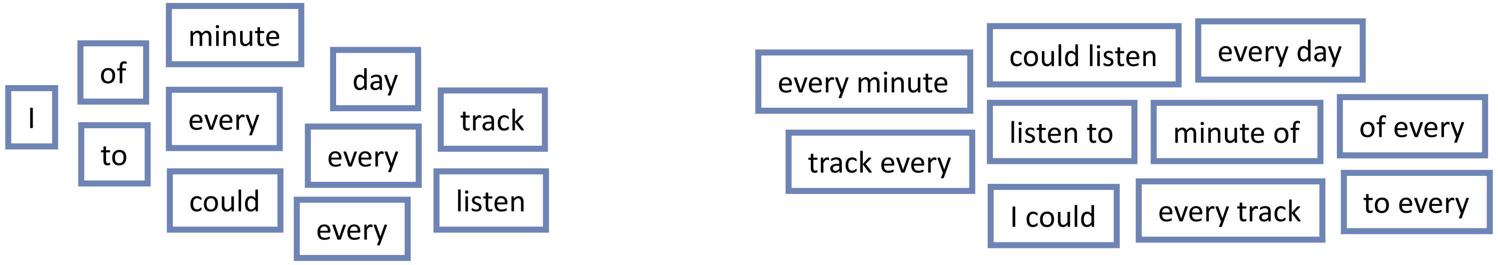
# documents

softmax

n-grams

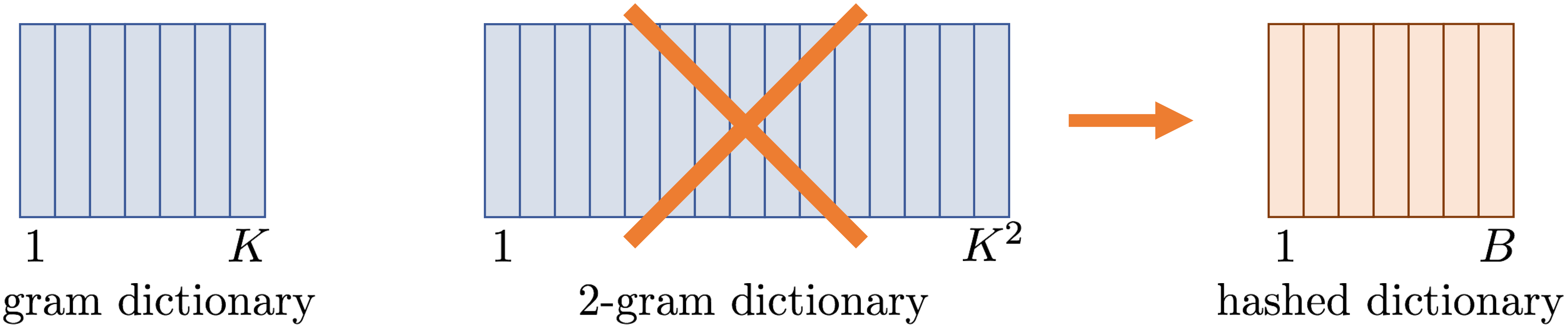
- Possible to add higher-order features

I could listen to every track every minute of every day.

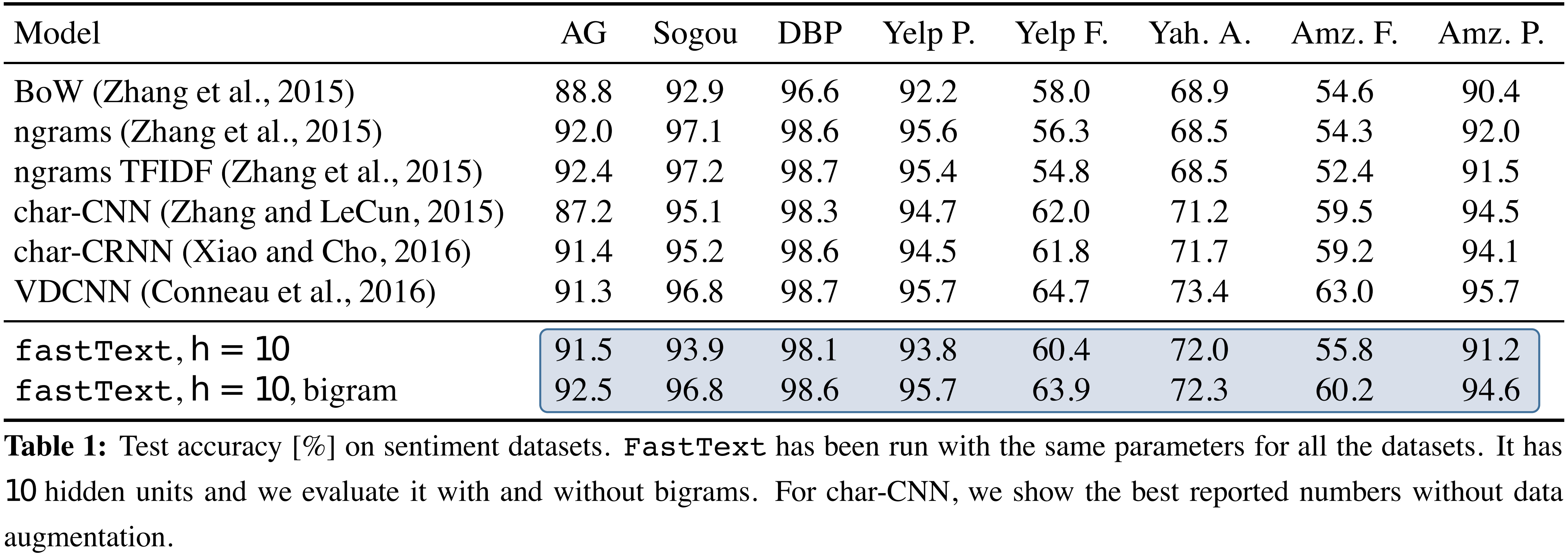


- Avoid building n-gram dictionary

Use a hashed dictionary!



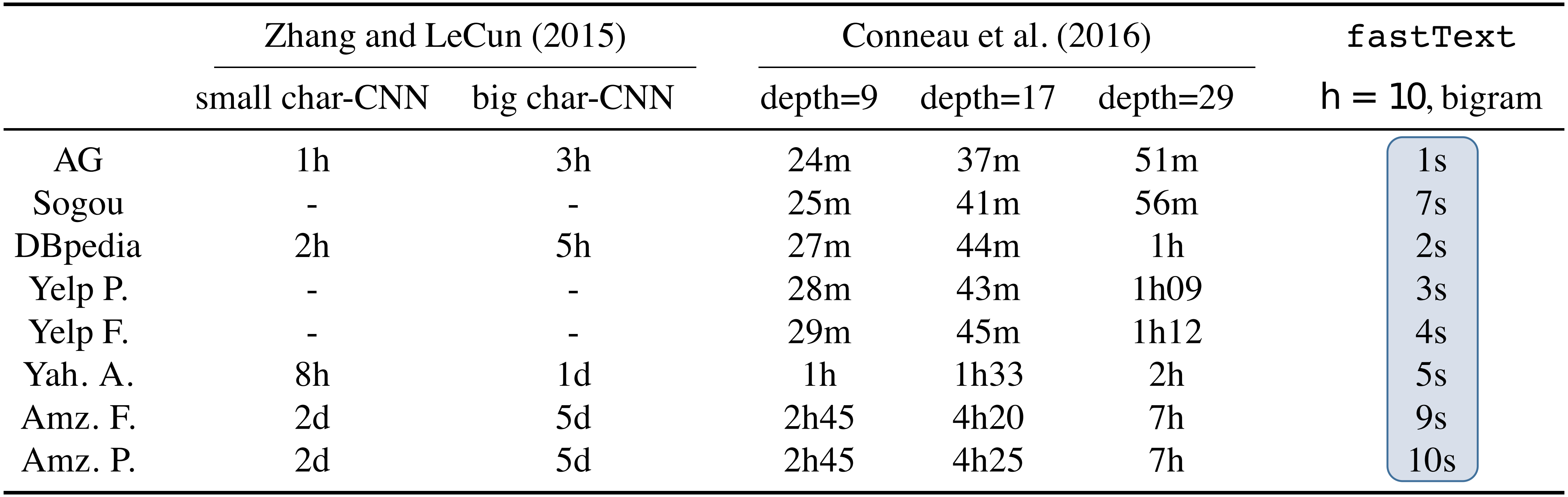
Sentiment analysis - performance



Model	AG	Sogou	DBP	Yelp P.	Yelp F.	Yah. A.	Amz. F.	Amz. P.
BoW (Zhang et al., 2015)	88.8	92.9	96.6	92.2	58.0	68.9	54.6	90.4
ngrams (Zhang et al., 2015)	92.0	97.1	98.6	95.6	56.3	68.5	54.3	92.0
ngrams TFIDF (Zhang et al., 2015)	92.4	97.2	98.7	95.4	54.8	68.5	52.4	91.5
char-CNN (Zhang and LeCun, 2015)	87.2	95.1	98.3	94.7	62.0	71.2	59.5	94.5
char-CRNN (Xiao and Cho, 2016)	91.4	95.2	98.6	94.5	61.8	71.7	59.2	94.1
VDCNN (Conneau et al., 2016)	91.3	96.8	98.7	95.7	64.7	73.4	63.0	95.7
`fastText`, $h = 10$	91.5	93.9	98.1	93.8	60.4	72.0	55.8	91.2
`fastText`, $h = 10$, bigram	92.5	96.8	98.6	95.7	63.9	72.3	60.2	94.6

Table 1: Test accuracy [%] on sentiment datasets. `FastText` has been run with the same parameters for all the datasets. It has 10 hidden units and we evaluate it with and without bigrams. For char-CNN, we show the best reported numbers without data augmentation.

Sentiment analysis - runtime

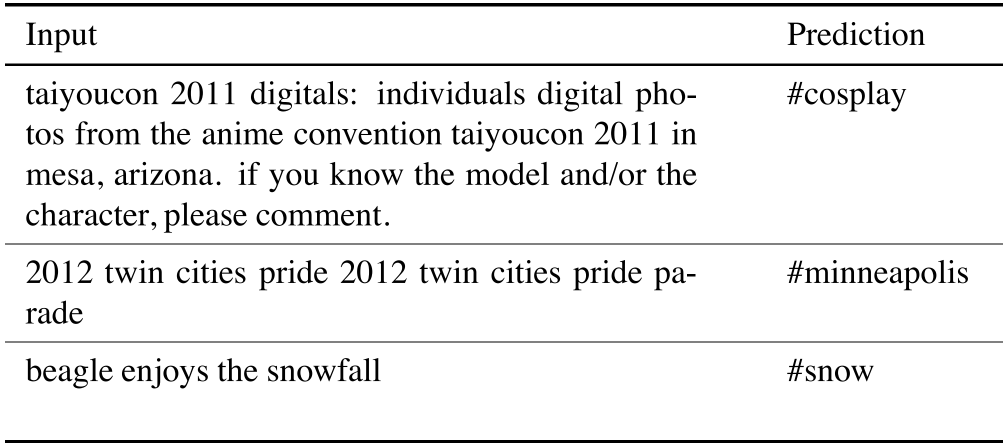


	Zhang and LeCun (2015)		Conneau et al. (2016)			fastText
	small char-CNN	big char-CNN	depth=9	depth=17	depth=29	$h = 10$, bigram
AG	1h	3h	24m	37m	51m	1s
Sogou	-	-	25m	41m	56m	7s
DBpedia	2h	5h	27m	44m	1h	2s
Yelp P.	-	-	28m	43m	1h09	3s
Yelp F.	-	-	29m	45m	1h12	4s
Yah. A.	8h	1d	1h	1h33	2h	5s
Amz. F.	2d	5d	2h45	4h20	7h	9s
Amz. P.	2d	5d	2h45	4h25	7h	10s

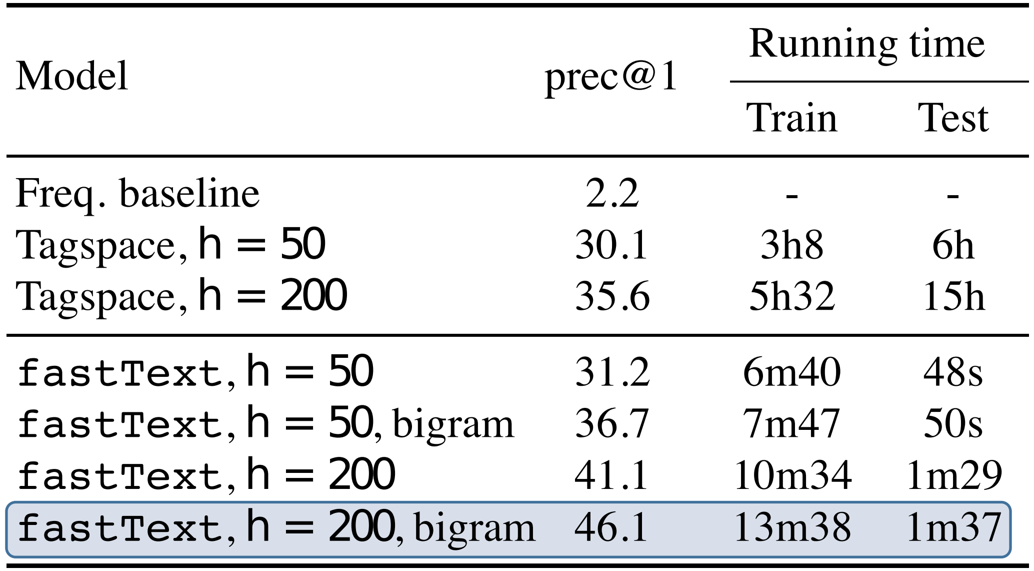
Table 2: Training time for a single epoch on sentiment analysis datasets compared to char-CNN and VDCNN.

Tag prediction

- Using Flickr Data
- Given an image caption
- Predict the most likely tag
- Sample outputs:



Input	Prediction
taiyoucon 2011 digitals: individuals digital photos from the anime convention taiyoucon 2011 in mesa, arizona. if you know the model and/or the character, please comment.	#cosplay
2012 twin cities pride 2012 twin cities pride parade	#minneapolis
beagle enjoys the snowfall	#snow

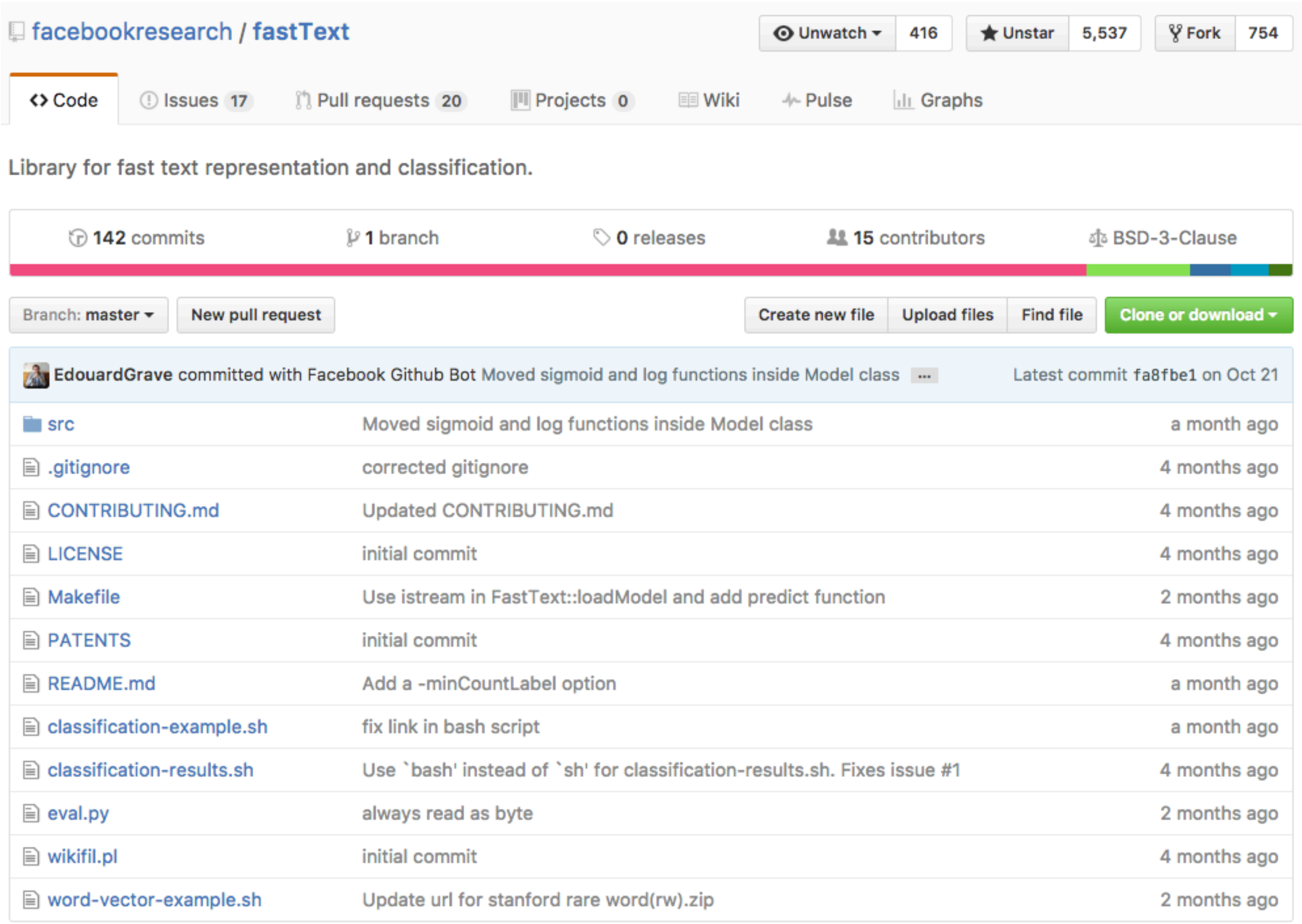


Model	prec@1	Running time	
		Train	Test
Freq. baseline	2.2	-	-
Tagspace, h = 50	30.1	3h8	6h
Tagspace, h = 200	35.6	5h32	15h
fastText, h = 50	31.2	6m40	48s
fastText, h = 50, bigram	36.7	7m47	50s
fastText, h = 200	41.1	10m34	1m29
fastText, h = 200, bigram	46.1	13m38	1m37

Table 5: Prec@1 on the test set for tag prediction on YFCC100M. We also report the training time and test time. Test time is reported for a single thread, while training uses 20 threads for both models.

fasttext is open source

- Available on Github
  - After 6 months:
    - > 6700 stars!
    - 1.6k members FB group
- Featured in “popular” press
- C++ code
- Bash scripts as examples
- Very simple usage
- Several OS projects
  - Python wrapper
  - Docker files



facebookresearch / fastText

Library for fast text representation and classification.

142 commits | 1 branch | 0 releases | 15 contributors | BSD-3-Clause

EdouardGrave committed with Facebook Github Bot Moved sigmoid and log functions inside Model class ... Latest commit fa8fbe1 on Oct 21

src	Moved sigmoid and log functions inside Model class	a month ago
.gitignore	corrected gitignore	4 months ago
CONTRIBUTING.md	Updated CONTRIBUTING.md	4 months ago
LICENSE	initial commit	4 months ago
Makefile	Use istream in FastText::loadModel and add predict function	2 months ago
PATENTS	initial commit	4 months ago
README.md	Add a -minCountLabel option	a month ago
classification-example.sh	fix link in bash script	a month ago
classification-results.sh	Use `bash' instead of `sh' for classification-results.sh. Fixes issue #1	4 months ago
eval.py	always read as byte	2 months ago
wikifil.pl	initial commit	4 months ago
word-vector-example.sh	Update url for stanford rare word(rw).zip	2 months ago

facebook