Predicting the publisher's name from an article title

Sayak Paul | Deep Learning Associate at PyImageSearch

GDE Global Summit 2019

San Francisco

Experts

Thinking about the problem

- Imagine being the moderator of an online news forum.

Thinking about the problem

- Imagine being the moderator of an online news forum.
- You’re responsible for determining the source (publisher) of the news article.

Thinking about the problem

- Imagine being the moderator of an online news forum.
- You're responsible for determining the source (publisher) of the news article.
- Doing this manually can be a very tedious task as you'll have to read the news articles and then derive the source.

Thinking about the problem

- Imagine being the moderator of an online news forum.
- You're responsible for determining the source (publisher) of the news article.
- Doing this manually can be a very tedious task as you'll have to read the news articles and then derive the source.

Can this task be **automated**?

More concisely ...

Given the title of an article, the task is to predict the publisher's name.

More concisely ...

Given the title of an article, the task is to predict the publisher's name.

It can now be modeled as a **text classification** problem.

Eyeballing at the dataset

	url	title	score
0	https://www.kickstarter.com/projects/carlosxcl...	Show HN: Code Cards, Like Texas hold 'em for p...	11
1	http://vancouver.en.craigslist.ca/van/roo/2035...	Best Roommate Ad Ever	11
2	https://github.com/Groundworkstech/Submicron	Deep-Submicron Backdoors	11
3	http://empowerunited.com/	Could this be the solution for the 99%?	11
4	http://themanufacturingrevolution.com/braun-vs...	Braun vs. Apple: Is copying designs theft or i...	11

First five rows from the resulting query

Source (Dataset publicly available via BigQuery)

The dataset I expect for the given problem

	source	title
0	github	feminist-software-foundation complains about r...
1	github	expose sps as web services on the fly.
2	github	show hn scrwl shorthand code reading and wr...
3	github	geoip module on nodejs now is a c addon
4	github	show hn linuxexplorer

Looking at the class distribution

blogspot	41386
github	36525
techcrunch	30891
youtube	30848
nytimes	28787

Other aspects of the dataset

- No missing values ^_^
- **513** titles having character length lesser than **11**
- 1 title having maximum character length of **138**

Three sets from hell

blogspot	33084
github	29238
techcrunch	24735
youtube	24586
nytimes	23106

Train set

blogspot	4147
github	3637
youtube	3115
techcrunch	3088
nytimes	2856

Validation set

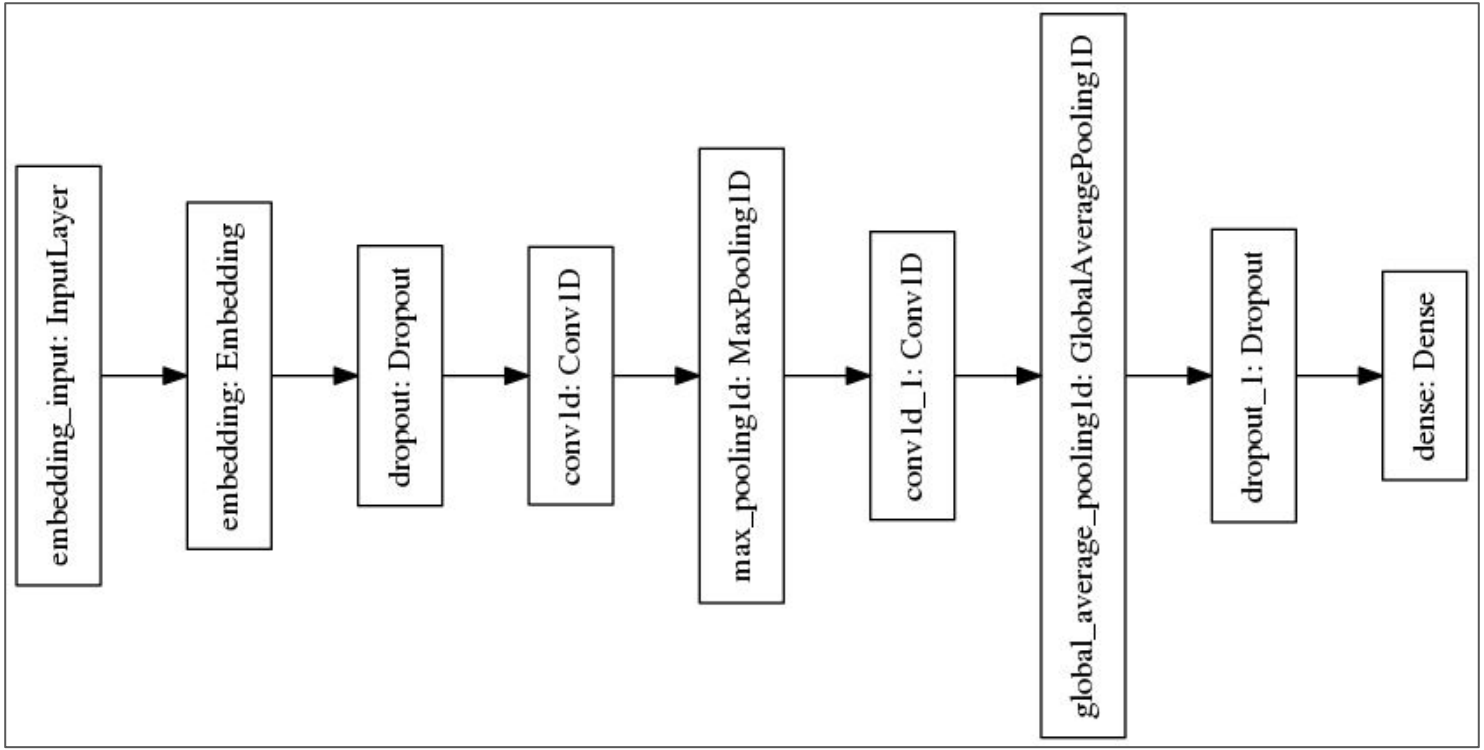
blogspot	4147
github	3637
youtube	3115
techcrunch	3088
nytimes	2856

Test set

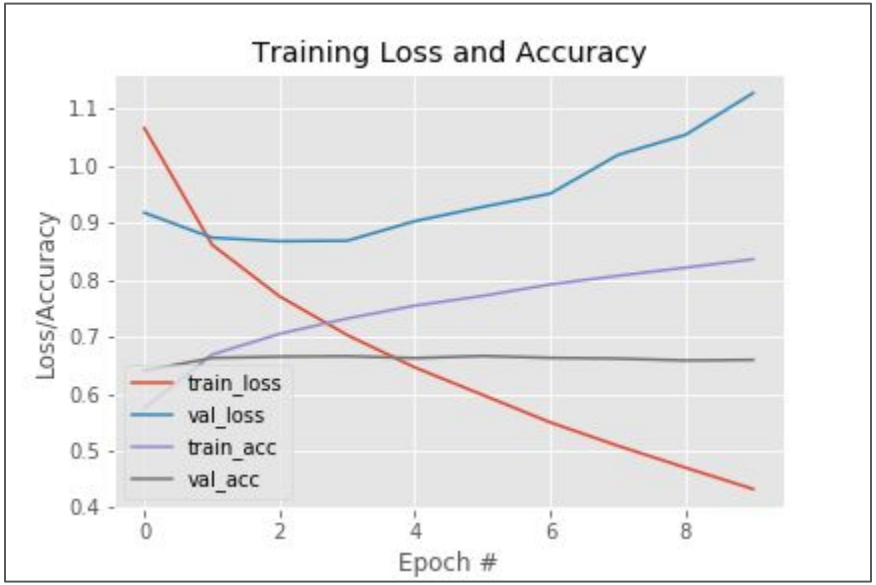
Data preprocessing steps taken

- Label encoding
- Creating a vocabulary from the training corpus — **tokenization**
- **Numericalizing** the titles and pad them to a fixed-length
- Preparing the **embedding matrix** with respect to pre-trained embeddings like **GloVe**

Building the Horcrux: A sequential language model



And the network overfits :(



Demo inference

```
# Prepare the samples
github=['Invaders game in 512 bytes']
nytimes = ['Michael Bloomberg Promises $500M to Help End Coal']
techcrunch = ['Facebook plans June 18th cryptocurrency debut']
blogspot = ['Android Security: A walk-through of SELinux']
```

Demo inference

```
# Prepare the samples
github=['Invaders game in 512 bytes']
nytimes = ['Michael Bloomberg Promises $500M to Help End Coal']
techcrunch = ['Facebook plans June 18th cryptocurrency debut']
blogspot = ['Android Security: A walk-through of SELinux']
```

```
github
techcrunch
techcrunch
blogspot
```

Google Cloud Platform, ftw!

- **BigQuery** for data gathering
- **AI Platform**
  - Preconfigured **Notebooks** for experimentation
  - **ML Engine** for making the entire modeling pipeline easier

Google BigQuery

Future directions

- Try other sequence models
- A bit of hyperparameter tuning
- Learn the embeddings from scratch
- Try different embeddings like universal sentence encoder, nnlm-128 and so on

Acknowledgement

I am absolutely grateful to the entire **GDE team** for providing me with GCP credits to aid this project!

See you next time

Find me here:
sayak.dev

Thank you very much :)

Experts

pyimagesearch