MT Exercise 4 - RNNs

Monday, May 7. 2018

Dataset & Preprocessing

For the task of training a language model with romanesco, i used a subset of a dataset of reddit comments. The initial goal was to generate high rated reddit comments by training a language model on the dataset and weighting the comments with their respective score. Because the original dataset is huge (>250G compressed), i used the subset containing only the comments of january 2015 (available as a torrent), which still contains more than 53 million comments. The dataset was initially structured with the comments as json dumps separated by line breaks, i then extracted only the scores and the actual comment text into a csv format.

To keep the original formatting of the posts, the newlines were escaped (\n -> <newline/>). Word level tokenization was applied using the python mosestokenizer wrapper.

Unfortunately i started too late, so there was no time left to train the RNN on the whole dataset, instead i used a subset with just 100'000 comments and did not implement the weighting based on the comment score.

Adaptations & Hyperparameters

ID	Size train/dev	Preprocessing	Hyperparameters	Adaptions
V1	10k/1k	-	NUM_STEPS = 35 LEARNING_RATE = 0.0001 HIDDEN_SIZE = 1500 BATCH_SIZE = 20 VOCAB_SIZE = 10000	-
V2	100k/10k	escape newlines, word-level tokenization	NUM_STEPS = 35 LEARNING_RATE = 0.0001 HIDDEN_SIZE = 1500 BATCH_SIZE = 20 VOCAB_SIZE = 10000	Adapt reader to read bigger (csv) inputs

I chose to keep romanescos default parameters. Maybe there would have been better results with a bigger vocabulary size, because the dataset contains comments in multiple languages, with english being by far the most common.

Results

	perplexity	**sample**
V1	144 (dev), 140 (train)	dogs I forget no Vanguard Someone in my eyes at process when the Female ones with the "your screen is one of for people play blowing as much news without fewer issues. I do, like another vendors are is commented of high though. but I'm key the line of at an ones for a bad, Whoever who tik link? Is on North machines on the primer and science, I'd include a clear party set where we can have the view of two back they still make a fucking trading southern series. Let you? I fade
V2	52 (dev) 40 (train)	loose in 6-1 "knife and re the * acid * dropping * your boyfriend so you would know me if I had a head single guy, that's a big picture. Nothing I thought about what everyone was saying, just curious but everyone would be here in just as practical" Where my parents. " Hopefully, her butt is a cunt. Go to your terrorists in the fourth week Do you still exist against marriage? At least we

In the same directory are:
A screenshot of the tensorboard loss graph - **loss.png**
The diagram of the language model including preprocessing - **diagram.png**
The final sample produced with the command below - **sample.txt**

```
romanesco sample 100 | python romanesco/processing/processor.py postprocess
```