Vivien Angliker : 14-721-948 30.04.2019

Übung 4: RNN's

1. Ich habe das Graham-Datenset verwendet, welches im Blog empfohlen wurde (http://www.paulgraham.com/articles.html). Ich habe mich dafür entschieden, weil ich fand, dass es einigermassen «neutrale» Texte waren, also nicht zu domänenspezifisch wie Untertitel oder Blog/Twitter-Einträge und vor allem weil die Texte in korrektem Englisch geschrieben sind. Pre-Processing habe ich keines gemacht, ausser darauf geachtet, dass mein Text keine mathematischen Formeln, Tabellen oder Bilder enthielt. Meine Textdatei hatte eine Grösse von 1 MB.

Folgendes sind die Resultate nach meinem ersten Training:

```
2019-04-29 17:48:21,767 - INFO - Perplexity on training data after epoch 1: 1602.92
2019-04-29 17:48:43,390 - INFO - Perplexity on training data after epoch 2: 901.44
2019-04-29 17:49:05,224 - INFO - Perplexity on training data after epoch 3: 850.96
2019-04-29 17:49:27,313 - INFO - Perplexity on training data after epoch 4: 774.50
2019-04-29 17:49:49,288 - INFO - Perplexity on training data after epoch 5: 634.85
2019-04-29 17:50:11,189 - INFO - Perplexity on training data after epoch 6: 483.81
2019-04-29 17:50:33,513 - INFO - Perplexity on training data after epoch 7: 385.14
2019-04-29 17:50:55,713 - INFO - Perplexity on training data after epoch 8: 323.89
2019-04-29 17:51:17,889 - INFO - Perplexity on training data after epoch 9: 281.78
2019-04-29 17:51:39,761 - INFO - Perplexity on training data after epoch 10: 248.59
```

Ich habe eine Perplexity von 249 erreicht. **Meine Perplexity auf dem Developmentset war 215 beim ersten Training.**

```
2019-04-29 17:54:56.455685: I tensorflow/stream_executor/dso_loader.cc:152] successfully
 opened CUDA library libcublas.so.10.0 locally
Perplexity: 215.50
```

Anschliessend habe die Durchläufe (epoch) um 5 erhöht, weil ich mir überlegt habe, dass mehr Durchgänge auch zu besseren Resultaten führen könnten. Nach 15 Durchläufen wurde meine Perplexity mit 154 merklich besser, wie der folgende Screen Shot zeigt.

```
2019-04-29 18:40:20,144 - INFO - Perplexity on training data after epoch 12: 201.55
2019-04-29 18:40:41,764 - INFO - Perplexity on training data after epoch 13: 183.68
2019-04-29 18:41:03,402 - INFO - Perplexity on training data after epoch 14: 168.07
2019-04-29 18:41:25,115 - INFO - Perplexity on training data after epoch 15: 154.22
```

Danach habe ich die Hidden Layers verdoppelt (neu: 2048), weil ich fand, dass mehr Gewichte auch zu einem präziseren Resultat führen könnten. Dies hat sich bestätigt, denn nun habe ich eine Perplexity von 112 und auch die Schritte, mit welchen sich das System pro Durchgang verbessert werden grösser.

```
2019-04-29 18:47:47,421 - INFO - Perplexity on training data after epoch 1: 1390.32
2019-04-29 18:48:34,495 - INFO - Perplexity on training data after epoch 2: 832.05
2019-04-29 18:49:22,111 - INFO - Perplexity on training data after epoch 3: 720.16
2019-04-29 18:50:09,799 - INFO - Perplexity on training data after epoch 4: 596.09
2019-04-29 18:50:57,516 - INFO - Perplexity on training data after epoch 5: 447.52
2019-04-29 18:51:45,233 - INFO - Perplexity on training data after epoch 6: 358.15
2019-04-29 18:52:32,911 - INFO - Perplexity on training data after epoch 7: 292.83
2019-04-29 18:53:20,640 - INFO - Perplexity on training data after epoch 8: 247.16
2019-04-29 18:54:08,340 - INFO - Perplexity on training data after epoch 9: 216.78
2019-04-29 18:54:56,100 - INFO - Perplexity on training data after epoch 10: 193.31
2019-04-29 18:55:43,860 - INFO - Perplexity on training data after epoch 11: 173.31
2019-04-29 18:56:31,593 - INFO - Perplexity on training data after epoch 12: 155.73
2019-04-29 18:57:19,355 - INFO - Perplexity on training data after epoch 13: 139.18
2019-04-29 18:58:07,112 - INFO - Perplexity on training data after epoch 14: 125.25
2019-04-29 18:58:54,854 - INFO - Perplexity on training data after epoch 15: 112.86
```

Da mein Korpus mit 1 MB etwas klein ist, habe ich noch die restlichen Essays von Graham angefügt und somit einen Korpus von 2.5MB erhalten, weil mehr Daten immer besser sind. In Kombination mit den beiden vorherigen Änderungen erreiche ich nun das beste Resultat, nämlich 58.

```
2019-04-30 08:35:57.713477: I tensorflow/stream_executor/dso_loader.cc:152] success
2019-04-30 08:37:52,260 - INFO - Perplexity on training data after epoch 1: 908.88
2019-04-30 08:39:48,802 - INFO - Perplexity on training data after epoch 2: 528.27
2019-04-30 08:41:45,107 - INFO - Perplexity on training data after epoch 3: 315.27
2019-04-30 08:43:41,519 - INFO - Perplexity on training data after epoch 4: 235.36
2019-04-30 08:45:37,846 - INFO - Perplexity on training data after epoch 5: 193.17
2019-04-30 08:47:34,285 - INFO - Perplexity on training data after epoch 6: 164.28
2019-04-30 08:49:30,693 - INFO - Perplexity on training data after epoch 7: 143.16
2019-04-30 08:51:27,088 - INFO - Perplexity on training data after epoch 8: 125.80
2019-04-30 08:53:23,580 - INFO - Perplexity on training data after epoch 9: 111.46
2019-04-30 08:55:19,960 - INFO - Perplexity on training data after epoch 10: 99.56
2019-04-30 08:57:17,181 - INFO - Perplexity on training data after epoch 11: 89.53
2019-04-30 08:59:13,496 - INFO - Perplexity on training data after epoch 12: 80.01
2019-04-30 09:01:09,903 - INFO - Perplexity on training data after epoch 13: 71.63
2019-04-30 09:03:06,225 - INFO - Perplexity on training data after epoch 14: 64.40
2019-04-30 09:05:02,590 - INFO - Perplexity on training data after epoch 15: 58.08
```

Meine Perplexity auf dem Developmentset war 116 nach dem letzten Training.

```
Perplexity: 115.51
```

Mühe hatte ich damit, das neue Textfile zu laden, denn ich hatte es zu Beginn ins falsche Verzeichnis kopiert. Auch war es nicht immer einfach alle Änderungen auf GitHub zu stellen und durch das viele Commiten und Pushen nicht den Überblick zu verlieren.

Der generierte Text befindet sich auf GitHub.

Ich stelle mir das RNN folgendermassen vor.