Fine-tuning Large Language Models on German Texts

Masterpraktikum, Lehrstuhl X

Armin Bernstetter (Matrikelnr. 1876874)

October 17, 2019

GER Basis für das Projekt dieses Masterpraktikums war die grundlegende Idee, GPT-2 für einen deutschen Textkorpus zu adaptieren. Dazu wurde zunächst ein bereits bestehender Korpus erweitert mit Texten eines deutschsprachigen Internetforums. Anschließend wurden auf diesem erweiterten Korpus language models trainiert (ELMo und GPT-2). Der durch das größte dieser GPT-2 Modelle generierte deutsche Text ist von annähernd natürlichsprachlicher Qualität und eignet sich als "proof of concept".

ENG This project was based on the idea of applying GPT-2 to a german web text corpus. For this, an existing corpus was extended with text from a german internet forum. This data set was then used to train language models (ELMo and GPT-2). The largest of these GPT-2 models generates rather high-quality german text and serves as a proof of concept justifying further work.

1 Introduction/Motivation

The general task of this project was to expand a corpus of german text and then train language models on this corpus.

This was based on the overwhelming success of OpenAI's[1] GPT-2 project [2] which trained a language model on 40GB of english web text. The model was so successful in tasks such as text generation that OpenAI refused to release the training code, the full training data and the entire model out of ethical reasons. The concern was that malicious parties could use GPT-2 to generate e.g. fake news for their own good and to spread misinformation.

Therefore it was necessary to create a new dataset, train the models on this corpus and check whether GPT-2 also works for text in languages other than English, e.g. German.

[1] `https://openai.com/`

The repository containing code, docker images, and a Readme file can be found on the Informatik Uni Würzburg gitlab server[2] The Readme file contains information on the used third-party code, the data sets' locations, and how text examples can be generated from th trained GPT-2 models.

2 Crawler

Several available german text corpora were already included in the existing corpus. This included the german Wikipedia and novels. To extend the corpus, the plan was to crawl several german webpages for useful text data.

Reddit Pages that were considered initially were german subreddits on Reddit. Reddit is a social media site for sharing and discussing content. It is organized into so called subreddits which are boards mostly focussed on single topics. Relevant for this project were subreddits such as */r/de*[3], the main german-language subreddit or */r/de_iama*[4], probably the german subreddit with the most subscribed users (around 330.000 as of 2019-09-30).

Reddit's API unfortunately limits the amount of requests a script can make and would require actively trying to avoid this limitation.

For future work it could make sense to look into pushshift.io[5] which offers a number of Reddit data dumps, mostly from some years ago.

Wattpad Another page that was considered briefly was Wattpad[6], a "social storytelling" platform. Users can share their own stories there and read those of other users. Upon examination of the web page's structure and HTML code it was decided that Wattpad obviously are not keen on users crawling their site. Therefore Wattpad was dismissed as well.

Team-Andro The website Team-Andro[7] is a german fitness and bodybuilding board. It contains articles on training, nutrition, sport events and more. More importantly, however, it is accompanied by a large german web forum with a user count of almost 300.000 and over 10 million forum posts in total (last checked 2019-10-17).

Upon inquiry the site administrator provided some helpful advice and a link to Team-Andro's sitemap. Unfortunately they had to decline the request for a simple and complete database dump of the forum's posts. Due to the forum's database structure it would not have been easily possible to exclude e.g. internal posts by administrators that need to remain private.

[2] `https://gitlab2.informatik.uni-wuerzburg.de/s319059/mprak`
[3] `https://www.reddit.com/r/de`
[4] `https://www.reddit.com/r/de_iama`
[5] `https://pushshift.io/`
[6] `https://www.wattpad.com/`
[7] `https://www.team-andro.com/`

The forum uses the standard *phpBB3*[8] board software and therefore the crawler should be applicable for other *phpBB* boards with minor adaptations.

Using the sitemap, it was possible to simply iterate through most of the forum's publicly available HTML pages. Team-Andro's sitemap contains around 80 sub-sitemaps each containing links to around 10.000 forum topics/pages. These pages were then parsed from HTML and saved in a directory structure representing the forum structure of Team-Andro with each post represented as one text file. Additionally, meta data such as user names, time stamps etc were saved in a mongoDB database. Each of the forum topics contains at least one text post by a user and the largest topic on the entire site contains around 700.000 posts.

3 The Dataset

Prior to the addition of the Team-Andro text data, the existing corpus consisted of a collection of various german text sources. This corpus contains for example the german Wikipedia and german novels and amounted to around 7.7GB of plain text files in total.

The Team-Andro corpus consists of mostly german web text. As such it is often unstructured, meaningless text without contiguous sentences. It is not as "bad" as e.g. text data from Twitch chat but still contains its fair share of internet slang and site-internal phrases and memes. Despite having a large Off-Topic subforum, a lot of the text is still focussed on topics such as sports in general and fitness/bodybuilding/weightlifting in particular.

The text also contains some special cases due to the forum software. For example one of the phrases which are extremely overrepresented in the corpus is "*[username] hat am [date] geschrieben:*". This indicates a quote of an earlier post inside a new post which often results in repetitions and nested posts of multiple levels.

In addition to a corpus with a directory structure representing the forum structure, the Team-Andro text was also merged into simple text files with 50MB each. This corpus consists of 2.8GB of plain text files.

The combined corpus used in this project therefore had around 10.5GB of text with 75% used as training data and 25% set aside as "heldout".

For minor experiments, a "minimal corpus" was created consisting only of one 50MB file from the existing corpus and Team-Andro each.

4 ELMo

ELMo is a deep contextualized word representation [1]. A tensorflow implementation of a deep bidirectional language model used to train ELMo representations is available on github[9].

[8] `https://www.phpbb.com/`

[9] `https://github.com/allenai/bilm-tf`

This implementation was used in this project to ease into the topic of training language models. A model was trained on the entire training data (75% of the overall corpus) for around 2 weeks and 1.877.500 steps.

The trained model was successfully used to output word embeddings. Further examination of the trained model as well as tasks such as text generation have not been undertaken.

5 GPT-2

GPT-2 is a language model trained by OpenAI on a large corpus of web text [2].

OpenAI refrained from publicizing their training code and larger versions of pretrained models. Github user "nshepperd" created a fork of GPT-2 containing Python scripts to train a GPT-2 model. This code was successfully used by blogger and author Gwern[10] to train a GPT-2 model for creating poetry.

As a side note, Reddit user disumbrationist[11] created a subreddit used solely by bots of various subreddits trained using *nshepperd's* code. This subreddit is called SubSimulatorGPT2[12].

Another fork/implementation that was considered for training was created by github user ConnorJL[13]. At the time (July/August 2019), ConnorJL's code was not usable due to issues with their files uploaded to Google storage.

To compare the differences in the quality of the generated text, two models were trained.

One was trained on top of the pre-trained 124M model and used a very small sample size of the text data (two 50MB files, one from Team-Andro, the other from the existing corpus).

The second model was trained using the 355M model and the entire corpus.

Both models were trained for around two weeks. The small model managed to do 522.000 steps, the larger one 180.000 steps.

The large model ended training with an average loss of 2.30 after starting with 3.47 at 100 steps. the small model ended with an average loss of 1.98 after starting with 4.32 at 100 steps.

6 GPT-2 Examples

This section shows example output text generated by the models after inputting a custom prompt. All outputs were generated using the same command and the `interactive_conditional_samples.py` script[14]. For this script to work, the respective model files have to be located in `nshepperd/gpt-2/models`.

[10] `https://www.gwern.net`

[11] `https://www.reddit.com/u/disumbrationist`

[12] `https://www.reddit.com/r/SubSimulatorGPT2/`

[13] `https://github.com/ConnorJL/GPT2`

[14] `https://github.com/nshepperd/gpt-2/blob/finetuning/src/interactive_conditional_samples.py`

```
python src/interactive_conditional_samples.py --top_k 40 --temperature 0.9 \\
--seed 2000 --model_name MODEL
```

The small custom model used the pretrained 117M GPT-2 model as basis and was trained on the minimal corpus over 522.000 steps.

The medium custom model used the pretrained 345M GPT-2 model as basis and was trained on the combined corpus over 180.000 steps.

These numbers were also used by Gwern in their GPT-2 poetry blog post[15].

6.1 Prompt A: "Ich trainiere meistens zweimal pro Woche mit einem Ganzkörperplan. Ich mache Bankdrücken und ..."

This prompt was used to see if the model actually learned something from the Team-Andro text.

6.1.1 Unchanged Medium Model by OpenAI

sie sicher an, um der Veränderungsmesslichen wirklich sich konnte ist zu befindlich, um aber durch einmal einige Woche können. Ich wird die Bankdrücken und sie kurz gelt, um mit der Veränderungsmesslichen konnte schon nicht kommen.

Die Bankstreit und Schulde in zugemäßigten Beobachten von der Landeszeigen und der Veränderungsmesslichen seit geschrieben. Die Schulde über den und da die Veränderungs -messlichen gessig der Gegenwart dann, da ist kann die Gegenwart auf den Wachs zu geschehen.

Die Schulde für die Vollziehung und den Vollerziesen vielen Vortrag der Landeszubung von den Schulde aus dem Schuld einer der Landeszeigen des Schulde durch zu befindlich, da ihr den Vollsiehung eine Veränderungsmesslichen im Veränderungsköpfer, da ihr der Vollerziesen der Vollziehung zu gleich.

Gestärdende bei den Veränderungsmesslichen, wenn man sind sie den Versuch oder kannst du garantischen Veränderungskollzählich aus dem Versuch oder garantissimmen zu recht, welcher kann der vergleichenden Vollziehung und verloren der vergleichenden Schulde bekannt. Auch die Veränderungsmesslichen des Königs der Welt, dass die Vollziehung in das Veränderungskoll zwei Gegenweise der Veränderungskoll zu einem Gegenweise auf den Vollziehung.

Geschrieben und ihn würdig ihn, mit der Gegenweise der Veränderungskoll sind ihre Vollziehung darüber dass

Comments: The model recognizes that the language of the prompt is german. However, it doesn't know anything about the context of bodybuilding/fitness and recognizes the word "Bankdrücken" only in the context of the financial "bank".

[15] `https://www.gwern.net/GPT-2`

6.1.2 Small Custom Model (117M as base, minimal corpus)

Schrägbank dazu bzw. "Ich glaube, der einzig wie danny es verdient ist mit beiden nicht ein super Wettkampf das ist aber der hat mich einfach ausschließlich über die ersten 10 kg hinausgehoben. Zumindest die WOchen egal ob 2,5 oder 2,5x Nebenwirkungen. Grundsätzlich wäre ein 2er aber mit einem 4er zusätzlich nicht gut.Meine einfach nur aus Interesse und Lebensplanung. So z. Schiller läßt nicht nur die gebotene Leidenschaft der Frau Welt und Gotteslehre, sondern ebenso auch die Lehr-bucherheit des Dichters fassen und glaubten in der Gerechtigkeit in das unendliche Reich der Liebe. München 1921. Schließlich wird man mit dieser Vermutung nicht mehr so richtig überschätzt wie bei den Goethe-Aufsatz 'Goethes Dichtertum und der Bildungsideologie', deren ältere dann aus der Entstehungsgeschichte des Goethejahres herausgekommen ist, zum Teil aus dem Ganzen der Kritik in den Rahmen einer Zeit gewonnen werden konnten. Jahrhunderts' (1909) — mit Erfahrungsgrenzen aus dem Mittelalter, mit Erläuterungen der Wertung der historischen und philologischen Fragestellung — ich aus dem Bereich der Antike heraus, ist die Zusammenstellung zwischen dem Bildungsgang der Kunst und dem Aufklärungsdenken der Neuzeit geworden. Die Vermutung das es so ist. Das ist seine einzige Situation,

Comments: The small custom model recognizes the language and the fitness context by referencing "Schrägbankdrücken" (a variation of Bankdrücken) and weights etc. After the first third of the text, however, it references literature and art.

6.1.3 Medium Custom Model (345M as base, entire super corpus)

Dips, ich weiß nicht ob das überhaupt möglich ist. Klar, ich bin aufgrund dieser Aufteilung für meine Trainingserfahrung und Erfahrungsberichte nicht der ein oder meine eigene Erfahrung das überlegt. Einmal sagte ja auch ihr das wäre schöner gewesen.... aber mehr als diesen ""einfachen"" Satz habe ich bisher nicht gemacht. Mein Problem ist wenn die Erfahrung nach oben ist. Ich habe geschrieben das der Körper anständig anfängt. Wenn ich jetzt das Gefühl habe das ich mich beim Kreuzheben kaum verlange wäre auch meine Einheit komplett zubringen. Nicht das man nur in einer Woche schlecht ist, sondern meistens bei der Grundübungsauswahl. Aber es kommt da so nicht drauf an was ich meinen das. Für die Kraftzuwachs und die Kraftleistungsaufteilung würde ich gerne wie gesagt einen Ganzkörperplan machen, aber zu meiner Person...Ich habe mir auch vorgenommen wenn ich nicht einiges mehr in meinen Körper zurückhalte um zu erfahren warum ich sofort mein gesamtes Training machen möchte, weiß ich nicht. Bin auch sehr froh, dafür bin ich durch die Art und Weise ich habe es nach der Erfahrung geschrieben, das ich nicht zur Zeit für die Kraftzuwachs verwendet sein sollte. Würde mich auch interessieren, um ob ich

Comments: The medium custom model also recognizes German and the fitness context but also keeps the context over the entire length of the text.

6.2 Prompt B: "I don't know any English anymore. I wish that I had learned more of"

This prompt was used to see if the model still knows its original English.

6.2.1 Unchanged Medium Model by OpenAI

it in Russia." There was a lot of Russian in his reply, and he was getting to the point. It was as though he was trying to impress me with some new vocabulary.

His wife was there in her mid-sixties, and the child had been born with cerebral palsy, which is pretty severe. He and I both know the difficulty of that. There's just something charming about that. It's quite charming to watch him go through the process. He was really trying to put us in his shoes.

Kirillov met his wife in St. Petersburg, where he had been studying medicine and writing, when they were married in the 1950s. She came from a family whose work in the fields of biology and microbiology brought them great recognition among those who knew them. It was an introduction to her work, something she never forgot. "It was also something about where he came from," Kirillov says. "He lived at a very poor peasant family; they didn't have electricity, and they couldn't buy a table. They worked in the fields. He made a little money with a little medicine and a small book."

From his earliest years in the village, he did well financially. He was a gifted athlete. "He was a top runner," his mother tells me. "He played football, tennis, and he also played lacrosse. He was a really good player. He was a very good player in soccer. He was a great skater."

He was also a great painter, Kirillov says. He had a great palette, he says. His wife, he says, used to paint a lot; Kirillov's mother bought him a couple of paintings, and he would hang them on walls. He never forgot what the paintings meant to him. They had a profound emotional impact on him, he says, and they helped prepare him as a person.

Kirillov married his girlfriend in 1960, and when they got married they went through a tumultuous period of separation. His mother and older sister died at around the same time, so the elder Kirillov moved in with his sister-in-law. Kirillov, who grew up in a large household, was not always like a traditional village boy — he always had a little rebellion sometimes, Kirillov says— but his mother and sister encouraged him, and he lived a different way than Kirillov does today.

"When I was 12, I remember

Comments: Obviously since the original 345M model was trained in English, the text quality is rather high. Here, it references the character Kirillov from a novel by russian author Dostoevsky.

6.2.2 Small Custom Model

the game. Ein solches Auffassen aus Gründen ist das: es sind die Waffen erscheint nicht den wesentlichen Kern der Erweckung, sondern ein wirklicher Rang.

Das ist das, worauf es sich handelt: Ich bin nicht der Erinnerung, ich war in ein paar wenige Hebungen, der ich gewohnheitsmäßig ungerechtfertigt war. Wenn die dort so krass hinhalten, dann wenigstens weitere Übungen.

So entsteht die Musik, die unter ihnen nicht eine Erinnerung der Idee von den Menschen, sondern die unter ihnen eine sich wandelnde Erinnerung an den Menschen sucht. Injizierst du die Nase zu haben und bekommst nach der Brustwarze eben langsamer die Brustwarze. S1L3 hat am 12 Jul 2017 14:38 geschrieben:Die ""großen"" Personen sind sowas von unrealistisch und da kriegst du nur einrivum, wie man da ran geht, in Form von Arbeiten auf nen zu hohen Stück zu kommen. Bei alledem ist die Voraussetzung des Dinges, indem er ein Erfassen der Subjektivität übersteigert. Einige von denen die schon vor ihnen bei Er. Nicht der wörtlichen Haltung, sondern die entscheidenden Schriften, die Goethe nach einer der Hauptmotive bedarf, sehen wir auf den Tasso des Tasso, und ihre Vereitung haben hier siebenjährig. Wer sich an den Kopf rächen will wird mit dem guten Geschmack gesäht. " Dann aber erfordert das Ich und seine Seele,

Comments: Most noticeably is that in this example the model only initially finished the prompt in english but then continued in german.

6.2.3 Medium Custom Model

it. I wish I had been more involved in it than I have been, and I can only hope that I will find that in the future when I'm older.
I have found English because I am very proud of it, because I like the way it's spoken.
But I have never been interested in it.
I did not like it.
I never even liked it until I met the family of David Thomson.
And that was the time I went to study English at university.
I started at the age of 18, 18 or 19, and I don't know why, but I started studying English at the age of 19.
I remember a boy getting into an English class called, and the teacher said: "Oh, sorry, Mr.
Taylor, but you have to get out of this class. Oh, sorry.
I don't want more English lessons.
I don't really like it."
We sat there a few minutes watching that boy get into class and look at the teacher and be like, "I don't like it.
I don't like it at all.
I don't like it at all," and then he looked and said: "I've studied English for more than twenty years and I don't like it."
I said: "So, Mr.
Taylor, I did the same thing when I was 20.
I didn't like it.
I didn't like it at all.
I was so embarrassed, even though I had been studying English for twenty years.

But that's the way these things work.
They take a long time to work out.
So, um so einiges bewahrte ich mich aber auch hier.
I studied English, no doubt about it, from an early age and I always liked it.
I had never studied French, Dutch, German, no doubt about it.
No doubt about it.
I loved all of them.
No doubt about it.
But after I got into history I started studying history.
I loved the historical aspect of the English language.
And I learned so much from Henry VIII.
and Catherine of Aragon.
I really loved it.
But it was nothing like what the other person told me.
It was like studying history and it was very superficial.
And I loved that.
And finally, eventually I began to get bored.
I don't like that stuff

Comments: This output is rather high quality text. It almost seems as if it's simply copied, possibly from poetry or something similar. It was not possible, however, to find (via Googling) a source for the phrases generated here. This leads to the assumption that at least part of it is actual generated non-existent text.

6.3 Prompt C: "Goethe war ein bekannter Autor. Er verfasste unter anderem Faust und"

This prompt was used to see if the model knows something from Wikipedia or in general about Goethe.

6.3.1 Unchanged Medium Model by OpenAI

sie sicherheitige Erwartung in den Ermengarde, die wolle und sich unter dem Faust. The devil loves a king whose body will be his throne. A king's crown is his crown is his throne, and so when his subjects come after him, he calls on his brother's name in praise and homage.

But if the king dies, he is buried with his crown and his horse.

Die Erziehung der Erscheinung in der Anstalt erster Stunde der Zeitung diesen Erziehung der Erscheinung in der Anstalt erster Stunde der Zeitung diesen Erziehung der Erscheinung in der Anstalt erster Stunde der Zeitung diesen Erziehung der Erscheinung in der Anstalt erster Stunde der Zeitung diesen Erziehung der Erscheinung in der Anstalt erster Stunde der Zeitung diesen Erziehung die Erziehung die Erziehung. The devil seeks the grave with his heart, as his heart desires to live again. But in his heart does the devil desire to go to the grave, but his heart seeks to live.

Sind in die Erscheinung der Erscheinung in der Anstalt erster Stunde der Zeitung diesen Erziehung die Erziehung diesen Erziehung die Erziehung diesen Erziehung diesen Erziehung. The devil is not dead, nor in pain. Death is the way of the devil. He wants nothing from us except death. Our life-forces want to make the devil suffer, as we do from a demon. But the devil wants to live, for he fears to be left behind.

Sind wird nicht die Erscheinung von die Erscheinung die Erziehung die Erziehung die Erziehung die Erziehung die Erziehung sind, wird er sie sich in das Erscheinung von die Erscheinung die Erziehung die Erziehung die Erziehung sind, wird er sie sich in das Erscheinung von die Erscheinung die Erziehung die Erziehung sind. When the devil's body is buried with his crown and his horse, he calls to the dead, "My brother! Come now, where

Comments: The model partially outputs german text but mostly english. Seemingly it recognizes the context of literature by generating rather literary-sounding phrases such as "When the devil's body is buried with his crown and his horse, he calls to the dead, "My brother! Come now, where"

6.3.2 Small Custom Model

sagte in seinem Aufsatz: Wie es nach dem Tode dieses Faust aus dem Lande an sein kunstvolles Stück geworden war, kann ich dir wahr genetisch von mir her erweichen, so hat wohl mir niemand auf der anderen Seite eine Gefahr erbracht. Erinnern wir uns nun, daß diese Gegensätzlichkeiten, die ihre besondere Grundlage durch die Mächtigkeit und Kultur des Dichters und Dichtertums, ihre Aufgabe und ihren Erlebten in der Kunst der Gegenwart und ihren Anfechtungen und Ineinanderspielen zum Ausdruck kommen. Der Versuch, den Begriff der Form zu bestimmen, ist in jeder Hinsicht nicht zu denken. Im Jahre 1816, in dem er, im Jahre des Mittelalters, ein großes Anzeige vier Jahre vor der Weltgeschichte an ein ganzes Papier geblieben, wurde er von seinem Herzen ein eigenes Leben der Erscheinung nach alles umgebogen. Nach 2 Monaten war ich bei meinem ersten Kumpel und mir hatte jemand gesagt ne möglickeit wieder meine Kraftwerte ausprobiert, bevor ich mich seitdem erdenklicher führe.Bisher waren die ersten 3 Monate aber weniger Kraftsport und ich habe mich im Prinzip wiederholt, an meiner neuen Kur, mit 250mg alle 3-4 Monate, also mehr zu machen. In einen anderen Studio geht mir das echt super, mein Freund. Dessen Strecke der Gedanke des "schönen Seldes" ist vorwiegend dem Löwen gefallen, 'Der Geist der Liebe' zu

Comments: Here the model recognizes German and references poetry ("Dichters und Dichtertums"). On the other hand, it also mentions fitness topics which have nothing to do with the prompt.

6.3.3 Medium Custom Model

das Wort Theorie, mit denen er später auch die gesamte Weltwissenschaft vertraut hatte, mit Faust. Friedrich Schlegel-Schiller verfasste den Beitrag der Werke Schlegels. Die bei-

den späteren Werke Georg Friedrichs des Großen (Unternehmen) und Friedrich Wilhelm Schlegel-Schiller (Literatur, die Verbreitung der Werke Schlegels, unter der Übersetzung von Wilhelm Schlegel-Brunswik) wurden in den Vereinigten Staaten gegründet. Georg Friedrich Schlegel-Schiller war ein Wettbewerber und Leiter einer Kommission für Künstler in der Nachkriegszeit. Zwischen 1780 und 1815 übernahm Schlegel sowohl in der zukünftigen Zeit die Ehefrau des Schiller-Geschäftsführers Carl Friedrich Schlegel und dessen Frau, der ehemalige Oberst des Heereskommandos, Johann Gottlieb von Schlegel. Zwar war Schlegel-Schiller der Leiter der Künstlergruppe 1815 bis 1815, aber auch über das Jahr 1848-1849, geschäftskommunale Mitglieder, im Schiller-Haus hatte er mit 17 Jahren in einem schwierigen Zeitpunkt in Niederschlesien und in Dänemark. Zwar sei ihm der Zusammenhalt gegenüber Schlegels Freunden in der Vergangenheit sehr groß, dennoch leitete er für seine eigene Kommission das Werk. Auch seine Frau, die bis 1841 bei ihm arbeitete, war nicht gewechselt ("Ihr habt eine gesunde Ehe", zitiert die Dänisch-Tageblätter in Wien 1841). Im Jahre 1847 verfasste er eine Arbeit mit Julius Kü

Comments: Here, the model recognizes that the prompt references a person. Although the text doesn't continue with Goethe, it mentions Schiller and Schlegel, both contemporaries of Goethe. Also it correctly mentions the overall time period (1780-1847) in which Goethe and the others lived.

6.4 Discussion

It is necessary to note that the generated text samples seen in the section above are only a few examples.

Comparing the three different models does, however, show some tendencies.

A The unchanged medium model by OpenAI obviously performs rather well when prompted with english text. When prompted with german text it does actually recognize that it is in fact German but cannot output high quality german text.

B The small custom model recognizes German and e.g. the context of fitness/bodybuilding. It seems to not really be able to uphold the context over the entire sample text and switches to different topics. Also it did not generate english text when prompted with an english phrase in this example.

C The medium custom model generated either german or english text when prompted with either one. It can uphold a topic (e.g. fitness or poetry/poets) over the entire sample text and seemingly generates sentences of higher quality and coherence.

7 Future Work

There are several approaches that can be taken to expand on this work.

Instead of relying on other's code to train GPT-2, a sensible approach might be to write own data loaders and training scripts for GPT-2.

On the other hand, ConnorJL's GPT-2 implementation[16] was updated recently and some issues have been fixed. It might be worth it to have another look at this before trying to write custom training scripts.

The Team-Andro dataset might be useful for entirely different text mining, statistical or social science tasks, maybe not even including machine learning.

At the time of training, OpenAI had released a small (124M parameters) and a medium (355M parameters) model. Initially, these were called 117M and 345M respectively but apparently there was a counting error which was fixed for a follow-up blog entry in August 2019[17].

In addition, they released another, larger model (774M parameters). With better training scripts, even more german text, and this large model as basis, a trained model could possibly generate very high quality, coherent german (and English) text.

8 Conclusion

The major contributions of this project are

- A dataset of german web text from an internet forum focused on fitness topics
- A trained model of ELMo embeddings
- Two trained GPT-2 models

All of these models in themselves are probably not very useful for future use. They do, however, show a tendency in the quality of generated text and serve as a proof of concept. Therefore they might be used as a justification to once again dive into the task of training GPT-2 on a german text corpus and start from scratch.

References

[1] Matthew E. Peters et al. "Deep contextualized word representations". In: *Proc. of NAACL*. 2018.

[2] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". In: (2019).

[16] `https://github.com/ConnorJL/GPT2`

[17] `https://openai.com/blog/gpt-2-6-month-follow-up/`