ual: creative computing institute

IU000128 Coding Three: Exploring Machine Intelligence

MINI PROJECT REPORT @YIFAN FENG

Project Title

Digital Storytelling: Explore AI Ethics by Aritifical Intelligence*

Overview

Often, generative algorithms for text are not restricted to training dataset and therefore their outcome becomes less predictable. In this mini project, I explore two generative models GPT-2 (small) and Char-RNN to investigate their creativity and bias by looking into philosophical discussions on ethics and accoubtability in the AI technologies domain. This experiment is a stepping stone to my master's dissertation on investigating equitable AI chatbot design and development. The first step is to fine tune GPT-2 (small) model on a pre-processed textbook "Oxford Handbook of Ethics of AI" and uses this refined model to generate new content and save them for the next training cycle. The second step is to incoperate GPT-2 (small)-generated content (through the question) into cleaned texbook document and implement Char-RNN model to further create artificial content about ethics. Final questions for both models are shown as follows:

- *Ethics is…*
- *Artificial Intelligence is…*
- *Equity is…*
- *Accountability is…*
- *Responsibility is…*
- *What is ethical AI?*
- *What is responsible AI?*
- *What is accountable AI?*
- *What is equitable AI?*

Training Part One

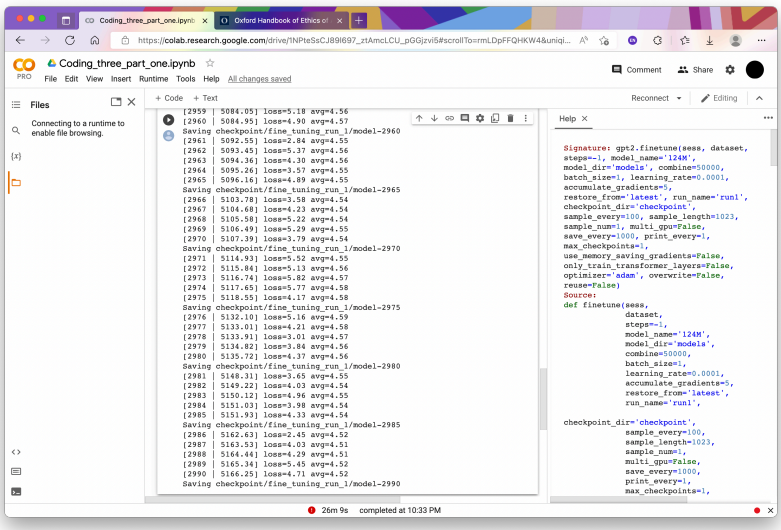
My project's primary training data is the converted version (plain text file) of "Oxford handbook of ethics of AI". To improve GPT-2 performance, I firstly implemented data cleaning by defining a series functions such as remove_whitespace, remove_stopwords to reformat capital letters and erase excessive whitespace, digits, numbers, etc. Next, I imported the cleaned datafile into GPT-2 (small, 355M) for further text generation. The compuational environment is Colab Pro with high RAM selected. Due to Colab Pro's storage limation, my first and second attempt of around 8.5k training iterations were completed in eight times (see table one). I adjusted arguments such as learning_rate, only_train_transformer_layer, training_steps, accumulate_gradiants and other paramters (see gpt2.finetune function) to fine tune the model on my customized dataset. To evalute interim output, I created samples in each 100 iteractions. For example, during the first trial of only training transformer layer, I found the results included much noise such as unreadable characters after 1500 steps. Therefore, the second training was set to train all layers and also decreased learning rate.

*https://git.arts.ac.uk/21036265/ExploringMachineIntelligence

5865 early stop

	Training Steps	Training Time	Only Train Transformer Layer	Env
First Attempt	2990	~2 hours	Yes	High RAM Colab Pro
Second Attempt	5865	~ 4.2 hours	No	High RAM Colab Pro -> Low RAM

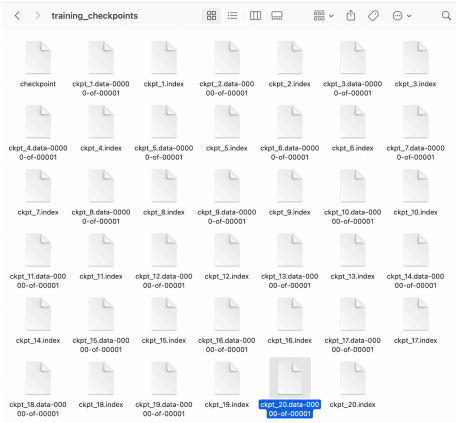
(Table 01. GPT-2 experiment)



(Figure 01. Unsaved GPT-2 first explore)

Training Part Two

To explore Char-RNN, I first trained the model on the original data (un-processed text file) for 20 epochs on my laptop (see figure 01) and saved modifided training weights for later use. The training environment is Jupyter Lab on an M1 Pro MacBook with one GPU accelaration. The whole process was completed in about 60 minutes. The second training was based on the school's desktop with one GPU support and included 280 epoch iterations (approx. 50 mins).



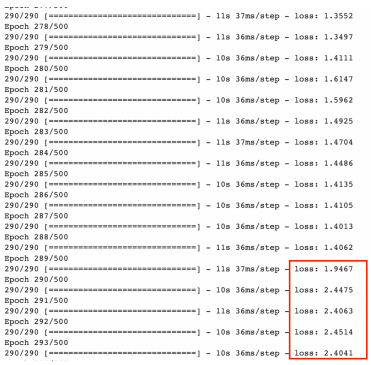
(Figure 02. Char-RNN first experiment)

```
def build_model(vocab_size, embedding_dim, batch_size):
  model = tf.keras.Sequential([
    tf.keras.layers.Embedding(input_dim=vocab_size,
                              output_dim = embedding_dim,
                              batch_input_shape = [batch_size, None]),

    tf.keras.layers.GRU(units = 64,
                        return_sequences= True,
                        stateful=True,
                        recurrent_initializer='glorot_uniform',
                        dropout = 0.1),
    tf.keras.layers.BatchNormalization(),
    tf.keras.layers.Dropout(0.2),

    tf.keras.layers.GRU(units = 128,
                        return_sequences= True,
                        stateful=True,
                        recurrent_initializer='glorot_uniform',
                        dropout = 0.1),
    tf.keras.layers.BatchNormalization(),
    tf.keras.layers.Dropout(0.2),

    tf.keras.layers.GRU(units = 256,
                        return_sequences= True,
                        stateful=True,
                        recurrent_initializer='glorot_uniform',
                        dropout = 0.1),
    tf.keras.layers.BatchNormalization(),
    tf.keras.layers.Dropout(0.2),

     tf.keras.layers.GRU(units = 512,
                        return_sequences= True,
                        stateful=True,
                        recurrent_initializer='glorot_uniform',
                        dropout = 0.1),
    tf.keras.layers.BatchNormalization(),
    tf.keras.layers.Dropout(0.2),


    tf.keras.layers.Dense(vocab_size)
  ])
  return model
```

(Figure 03. Char-RNN architecture)

*https://git.arts.ac.uk/21036265/ExploringMachineIntelligence

The modification (see figure 03) is focused on reconstructing model's architecture. For example, I added a few GRU layers to increase the network's ability to process the whole sequence. I also appended batch normalization layer and dropout rate after each GRU to increase model's effiency to handle data in each running iteration. I also tried different embedding dimentions size and found out that value = 512 or 256 can generate expected result (i.e., human-like sentences if temperature is set to 0.5). The second trial was set to 500 epochs. However, it was early stopped (at 290) due to a sudden rise of loss indicators (see figure 04). This might be the model trying to find local miniumn or requring more data from input.



(Figure 04. Loss indicator increase)

Dissusion

Both model's performance was evaluated based on loss indicators and output (subjective evaluation). To check model's predictability (aka confidence), I played with different temperature (from 0.01 to 0.3) to generate some results from both models. Keywords are provided in the question list. Overall this experiment of playing Char-RNN and GPT-2 (small) shows a positive correlation between training data and neural network (or generative model)'s performance (e.g., outcome).

The nature of Char-RNN network determines its ability to process text-based data on a character level and produce non-existent words. Figure 05 to 07 demonstrates a series of new words that are created by this network. When temperature is set to 0.01, the newly generated words are very close to input content in terms of spelling. When temperature is risen up to 0.2, these words are hardly recognizable. Overall, one can see recurrent words (regardless of prompt questions) such as "artificial", "system", "algorithms", "social", "human" which are also high-occurence words in the input text.

The experiment chose to use GPT-2 355MB pre-trained model on both transformer layer and the entire network. Compared to Char-RNN, this model can handle larger data flow (word in squences) in each iteration and the result are more human-readable. For example, it can generate phrases such as "encourage intelligent actions", "user data", "less trustworthy" that are associated with input prompt "accountability". When asking "what is AI?", the outcome answer includes sentences like "ethical principles including transparency and accountability", "always use feedback to improve algorithms without additional enhancements".

*https://git.arts.ac.uk/21036265/ExploringMachineIntelligence

For better performance of both models, first step is to standalize input data through various methods of text formatting. Secondly, I suggest training both models on a bigger database (current datafile is approx. 2MB) and closely monitoring training loss indicators and provisional samples. For example, one can print out lastest generated text after 150 steps and inspect the content quality. A third suggestion is to carefully utilize checkpoint file. Before (re)loading trained weights from the latest checkpoint document, one should examine the result first to prevent overfitting. Lastly, one can try to twist model's archetecture by appending or freezing layers, increasing neuros per layer and decreasing dropout rate if training Char-RNN from scratch.

Relavant Reference

[1] Bhikadiya, M., 2021. *AI Writer ✍: Text Generation Using GPT-2 & 🤗Transformers*. [online] Medium. Available at: <https://medium.com/analytics-vidhya/ai-writer-text-generation-using-gpt-2-transformers-4c33d8c52d5a> [Accessed 15 June 2022].
[2] Medium. 2020. *RNN- Character level text generation with Tensorflow 2.0 from scratch to deep insights.*. [online] Available at: <https://geeks-today.medium.com/rnn-character-level-text-generation-with-tensorflow-2-0-from-scratch-to-deep-insights-41bac0e07f86> [Accessed 15 June 2022]

Result Showcase

```
print(generate_text(model,
                    start_string=u"Responsible AI is",
                    num_generate=500, temperature=0.01))
```

Python

```
Responsible AI is and the come the suche consting the conse the fur in the and and the
proces a the provers the proment of the process the probety and the proces and the
proces and the proces and the procerations in the computer the consting the computer
and the prove contrical the construte of the come the come to the and and the process
and the proment of the proces the desent and the proment the ex mation of the prove
the secome of the conse the prove the proment the sument of and and and the come the
suching
```

(Figure 05. Char-RNN 20 epoch)

```
generate_text(model,start_string=u'Responsible AI is',num_generate=200,temperature=0.1)
```

Python

```
'Responsible AI issues sections ai systems section algorith constion algorith conter algorithm simple
algorithmation algorith develop alson conter artificial system accounted comple algorith algorither
algorither ai sy'
```

```
generate_text(model,start_string=u'Artificial Intelligence is',num_generate=200,temperature=0.2)
```

Python

```
'Artificial Intelligence issuts procent proper comple stated comple ai sect ai see contion accounted
algorithment proces social alson procention complecing ai system sections stantical interneted comple human
procenting access '
```

(Figure 06. Char-RNN 290 epoch)

*https://git.arts.ac.uk/21036265/ExploringMachineIntelligence

```
generate_text(model,start_string=u'Accountable AI is',num_generate=200,temperature=0.2)
```

Python

```
'Accountable AI issess sections ai system constions designing advertion prove provers design contion
algorith section experted ai accounted comple algorither sections experent algority sections sections ai
sect ai cont'
```

```
generate_text(model,start_string=u'Responsibility is',num_generate=200,temperature=0.15)
```

Python

```
'Responsibility issues section algorithms secte algorithm conteration experting ai system ai sections
stantern algority algorithm sections ai system constions sections ai systems sections accounted conter
algorith cons'
```

```
generate_text(model,start_string=u'Equity is',num_generate=200,temperature=0.05)
```

Python

```
'Equity istion action algorith also contertions ai systems section conter algorith consition secte
artificial interneted ai systems sections ai systems section algorith constion algorith constitute section
ai s'
```

```
generate_text(model,start_string=u'Ethics is',num_generate=200,temperature=0.15)
```

Python

```
'Ethics issues sections ai systems comple algorithm sections ai stantificial intelligence sected contion
algorith consticul antical respons artificial siment algorithm conter artificial systems conseques ai art'
```

(Figure 07. Char-RNN 290 epoch)

```
input_prompt = "Accountability is…  "
gpt2_text_generator = gpt2.generate(sess,
                                    run_name= checkpoint_name, #restore from checkpoint
                                    temperature=0.15,
                                    length=100,
                                    prefix=input_prompt,
                                    top_k=30,
                                    nsamples=3)#the amount of outcome
                                    #destination_path = "/content/drive/MyDrive/Coding3/gpt2_text.txt")
```

Python

```
Accountability is…  urs gasser carolyn schmitt companies like facebook google even offer chatbots understand
understand understand give httpswwwfacebookcomdevelopersarechatbot httpsautomatedercoderjailsocietyarevery
similar examples see httpswwwfacebookgoogleschatbot httpswwwfacebookcomdevelopersarechatbot nick stat
“google’s community analyzes nearly million interactions platform user data shows nearly third people post
negative things ” politico may httpswww politicocomnextgen technology google’
====================
Accountability is…  urs gasser carolyn schmitt companies like facebook google even offer chatbots understand
understand understand give httpswwwfacebookcomdevelopersarechatbot httpswwwfacebookcomdevelopersarechatbot
shannon mattern developers also encourage encourage intelligent actions people’s behalf27 thirdparty
companies also provide services assist developers translating enphrases en français en francisco’s
httpswwwfranconewsnetworkartificialintelligence httpfranconwhowasapnewprojectshtml see
====================
Accountability is…  urs gasser carolyn schmitt companies like facebook google even offer chatbots understand
understand understand give httpswwwfacebookcomsupport chatbot httpsenable chatbotlenetoggleshowportraithtml
shannon mattern product marketing materials also give general idea kind chatbot could useful person instance
elderly relative less likely trustworthy trustworthy” eg could use voice recognition identify dementia
patients poor memory patients may trust able refer doc accurately cared about concerns “accuracy
selfreported measures social mental health” could also used detect presence
====================
```

(Figure 08. GPT-2 final outcome)

*https://git.arts.ac.uk/21036265/ExploringMachineIntelligence

```
input_prompt = "Responsible AI is...  "
gpt2_text_generator = gpt2.generate(sess,
                                    run_name= checkpoint_name, #restore from checkpoint
                                    temperature=0.15,
                                    length=100,
                                    prefix=input_prompt,
                                    top_k=30,
                                    nsamples=3)#the amount of outcome
                                    #destination_path = "/content/drive/MyDrive/Coding3/gpt2_text.txt")
```

Pytho

```
Responsible AI is...  urs gasser carolyn schmitt companies—are engaged norm creation lesser extent
administration—while still keeping ethics promise made ethical aligned company's approach released statement
company aims develop ethical aligned responsible ai no questions asked leading role ai ict impact ethical
decisions every aspect company's operations however ethical promiseably transparent process evaluation
information resources used develop deploy deploy ai system including financial incentives potential
conflicts exist public commitments trustworthy ai come several stages including various codes ethics google
created responsible ai initiative google
====================
Responsible AI is...  urs gasser carolyn schmitt companies—are engaged norm creation lesser extent
administration—while still maintaining professional norms company culture remains largely experimental
accordingly norms created google glass suggest norms might adapt much like brooks required new york sophia
press frank pasquale black box society cambridge harvard university press normen staat "responsible bots"
google glass ada algorithms —google glass removed safety glasses required new users google glass
discontinued production safety glasses shortly company learned would likely generate negative consumer
negative
====================
Responsible AI is...  urs gasser carolyn schmitt companies—are engaged norm creation lesser extent
administration—while still maintaining professional norms aspirational rather practical ai hleg claims
requires membership association existing larger groups society confront issues norms must continuously
monitor shift normative status quo reflect status quo created new members must also continuously strive
extend professional norms ethical principles beneficial increased use ai legitimate professional norms
aspirational rather pragmatic reasons well practical reasons professional norms reflect status quo purposes
important take ai account discrepancies norms statutorily mandated employers employees respectively and
====================
```

(Figure 09. GPT-2 final outcome)

```
input_prompt = "Artificial Intelligence is… "
gpt2_text_generator = gpt2.generate(sess,
                                    run_name= checkpoint_name, #restore from checkpoint
                                    temperature=0.2,
                                    length=100,
                                    prefix=input_prompt,
                                    top_k=70,
                                    nsamples=3)#the amount of outcome
                                    #destination_path = "/content/drive/MyDrive/Coding3/gpt2_text.txt")
```

Python

```
Artificial Intelligence is… ict impact every human endeavor ict design london nicholas g shal ows internet
brains new york liveright publishing corp next set technologies underwriting ai ethical principles including
transparency accountability among paul rauwolf joanna j bryson yourself34 ai could always use feedback
improve core algorithms make better even without additional enhancements could always benefit existing ai
systems since require input data models different audiences also feedback improve predictive accuracy
although noble intentions crow many cases leads people developing ai become tech
====================
Artificial Intelligence is… ict impact every human endeavor ict design ict user experience ict design ict
user experience designer x □ □ x ai system generates value user experience leads better customer acquisition
experience wide range ethical desirable outcomes ict users google facebook twitter could potential lead
better understanding trust algorithms google facebook twitter could help you identify potential ethical
issues algorithms society large emphasis ethics ai google came heels walkout organized women google refused
contend sexual harassment undercompensation compared male coworkers culture consistently undervalues demeans
====================
Artificial Intelligence is… ict impact every human endeavor ict design london nicholas g shal ows internet
brains new york liveright publishing corp next set technologies underwriting ai ethical principles including
transparency accountability among paul rauwolf joanna j bryson yourself34 ai hleg also released ai
principles document highlights steps already taken form steve dipaola liane gabora "responsible bots
guidelines developers conversational ai" communications acm ——chatbots personalized learning systems
====================
```

(Figure 10. GPT-2 final outcome)

*https://git.arts.ac.uk/21036265/ExploringMachineIntelligence