

GPT-2 Based Story Teller

CTEC3451 Development Project

De Montfort University

Artem Bobrov (P2547788)

May 5 2022

Contents

Abstract	2
Introduction	2
Objectives and Goals	3
Initial approach	3
New approach	3
Reasons	3
General Description	3
Major Components	4
Backend	4
Frontend	6
Linking the frontend and backend	9
Development lifecycle	10
Training the model	10
Text generation	12
Reaction System	14
Frontend Development	16
Critical analysis and Reflection	19
Libraries and frameworks	19
Tools	20
Bibliography	20
Appendices	23
Testing	24
Model testing	24
Reaction system testing	25

Abstract

The project described in this report is a story-telling application that uses the GPT-2 model to generate text. The model is trained on multiple datasets and is able to generate text from a variety of topics in the backend. The frontend is a web application that allows users to endlessly generate stories accompanied by thematic pictures. The point of the project is to prove the concept that neural networks, if trained properly, are able to generate data that is very similar to human language.

Introduction

Neural networks weren't so popular in the early days of computing, but they are now a common tool for machine learning. First neural network was invented by Frank Rosenblatt in the late 1950s (Rosenblatt 1958). Perceptron has only one layer of neurons, but more advance neural networks can have multiple layers. Before Transformer neural networks there were many different types of neural networks (e.g. Convolutional, Recurrent, LSTM, Feed Forward, etc).

Transformers were introduced in 2017 by Google (Vaswani et al. 2017) and are a new type of neural network. It uses self-attention mechanism to focus on the most relevant information in the input. It weights the significance of each part of the input differently.

This report focuses on the GPT-2 model, which is a transformer type neural network developed by OpenAI in 2019 (Radford et al. 2019). Despite the fact that GPT-3 is a newer model, it's predecessor is still a popular choice for text generation.

An application will be developed that utilises the features of GPT-2 and wraps it up in a user-friendly web interface that allows users to read freshly generated stories accompanied by thematic pictures. The development process will be examined and the results will be presented and discussed in depth. The usage of the key elements and techniques used both in backend and frontend will be explained. Moreover the project development and research cycle will be reflected on and other approaches that possibly could have been used to achieve different or similar results will be presented.

Objectives and Goals

Initial approach

It is important to state that the original plan for the project was to develop a similar application that generates pictures with attached text to them (known as ‘memes’) and allows users to watch them using a web interface. It has been planned to use different approach to build a repository of training data using ‘Optical Character Recognition (OCR)’ (Kay 2007) and ‘Pillow’ library to attach processed text to the pictures. It has been decided to adjust the original plan and change the topic from memes to stories.

New approach

The new approach doesn’t require any massive changes to the original plan and the project will be developed using the same techniques - indetical frontend and almost the same backend. The only difference is that OCR will be removed from the backend because it’s not needed for the project anymore. To minimize the deviation from the original plan, there will be an additional section in the web interface that creates ‘memes’ based on novels and other piece of art.

Reasons

Despite the fact that many of the features used in the initial approach have been already developed - Tesseract OCR, Pillow Image placing, ‘Meme’ scraper - one major issue made it extremely perplexing to continue the project without changing the approach.

The first puzzling part was hidden in reading the text (using Tesseract OCR) from the scraped images and translate them to text strings. Memes contain text captions that makes it easy for a human to read, but this same process is hard for machines to do so - especially via OCR. There were some images successfully processed by it but it didn’t satisfy the requirement because the the success rate was too low. Many solutions were tried, but even pre-processing the images before feeding them to the OCR engine didn’t seem to give much improvement. The second complexity was to build a meme repository itself. Using the meme scraper is not the best solution because the content obtained by it can be anything. Unfiltered, uncensored - it’s a huge ethical problem.

General Description

Two general components of the development project are described in two sections. Such apparent separation between these categories is due to the fact that this approach allows a simpler, faster and more effiecient implementation with better user experience, it also improves code readability and maintainability.

The first one is the backend, which is responsible for generating text based on given datasets (stories, novels, other pieces of art, etc). It is implemented in Python and uses numerous libraries to achieve the goal. Multiple steps are involved in generating simple text samples including hours of training.

The second component is the frontend, which is a web application that creates a user-friendly, simple interface to scroll and read stories. It is implemented in JavaScript using many libraries, as well as the Python part. Such approach to frontend was chosen because web development is

flexible and it's easy to adapt to different platforms. Also it's possible to extend it to a web service that doesn't require downloading the application to user's computer.

The goal is create a system with pre-trained model, which the user can use to scroll through the stories and react to each story leaving a positive or a negative feedback - using the 'Like' and 'Dislike' buttons. Afterwards, these reactions will be stored in a database to be used by the developer to understand which stories should be corrected and improved. The feedback system is designed to create a reflection on the pre-trained model as a whole and not on separate elements (stories). Given the feedback, the developer can decide to improve the model or not.

Major Components

Backend

The backend of the project is developed using Python programming language since it's a very popular language for machine learning and has a lot of libraries that can be used to achieve the goal. The backend is developed using PyCharm IDE, which is also a popular editor for Python. Calculations that are done in the backend can be executed on any machine with Python installed. The only issue is the optimisation - text generation requires a lot of computing power.

GPT-2 and the library

GPT-2 is the version of the GPT family used in the project. To operate with the network it was chosen to use a library called 'GPT-2-simple' (Woolf 2021) since it features an easy approach to control and train the model. The library is written in Python and offers a documentation which was used in order to work with the neural network. GPT-2-simple offers two types of models - one with 124 million parameters and one with 324 million parameters. Acknowledging the computational limits of the processing unit used - graphical (GPU), the first model with 124 million parameters was used. It is possible to use the second model but it requires a GPU with significantly more processing power and memory.

It is possible to use techniques that allow to save memory and train bigger models using memory saving gradients - the main memory loss in the computation is due to calculating the gradient of the loss by backpropagation. At the cost of one additional forward pass, the memory consumption can be reduced significantly (Chen et al. 2016). However Tensorflow framework (version 2), which is used by GPT-2-simple, is not optimized for such tasks.

Despite providing simpler control functions, this library also provides a lot of tweaking options. For example, some of the hyperparameters can be changed to improve the model performance during testing. Learning rate, batch size, number of training steps, number of accumulated gradients and optimizer can be changed. Number of layers and other more complex parameters can also be changed but in order to do that, the parameters must be examined and adjusted in Tensorflow code files.

The library also provides a way to load the model from a checkpoint file. This is useful when the model is trained in multiple steps and it's needed to continue training from the last checkpoint without losing progress.

Another important feature for testing and optimising the output is the possibility to use different optimisers. 'Adam' and 'SGD' are available optimisers. The first one is the most popular and is used in the project because the learning time is relatively short and the memory requirements are little (Musstafa 2022). These two aspects are crucial and therefore the choice fell on 'Adam'.

Datasets

Several datasets are used in the project. It has been chosen to use three datasets to work with. The first one is the novel ‘Alice in Wonderland’ by Lewis Carroll (Johnsen 2013). This version was taken from GitHub Gist. It’s more adapted for machine learning than the original version since it’s shorter and therefore contains less tokens. For the reason that it was the first dataset to train the network on, it has been decided to go with the shortened version. The text file with the novel poses a certain inconvenience because it contains a lot of text decoration which can create noise in the model. But also it can have a positive effect on the user experience because it can be perceived as a unique style of the author. The unformatted version contains 26,413 words and 148,574 characters.

The second dataset used is the novel ‘Catcher in the Rye’ by J. D. Salinger (Valenzuela 2017). The version found on GitHub is slightly incomplete in the beginning and therefore was filled with the original text from the book. Also it was slightly formatted for the sake of the correctness of the output of the neural network. Chapter enumeration and the title page were removed from the text file. In the final revision there are 73,676 words and 382,008 characters.

The last and the longest dataset is the novel ‘Iliad’ by Homer (Stevenson 2000). It contains 152,507 words and 807,481 characters. It doesn’t contain much text decoration apart from chapter (book) enumeration. Also it has some translation references and copyright information. All this information was removed from the dataset since it could affect the training process. It still can be viewed in the original text document in the reference list. Another text decoration elements such as book (chapter) enumeration and dash symbols dividing the chapters was also removed.

Later as the project evolves more datasets potentially could be added. The main problem of having multiple datasets is the size of the checkpoint files created by the neural network for training purposes. ‘Alice in Wonderland’ checkpoint, which was trained 3,200 times on the novel is around 500MB and ‘Catcher in the Rye’ checkpoint is about the same size and the same amount of training steps. It would be possible to add more stories to the project if the checkpoint storage would be located not on the user’s machine but rather on a server.

Pillow

‘Pillow’ is a Python library used to manipulate images (Clark & Contributors 2022). It’s a fork from an original library named ‘PIL’. The reason why it was chosen to use in the development project is that it offers an easy and fast way to insert text into the image. This process is also known as ‘meme’ creation. It is possible to use generated samples from the neural network, either cut the size of the samples after it has been generated or generate a small sample in the first place. The second option reduces the time of generation process but it also takes away the option to format the text if the first and the last sentences are inappropriate. The customisation of the text in the library is a very straightforward process. It is possible to change the font, the size, the colour, the position and the angle. The ‘Meme mode’ section will most likely contain pictures in unified format (i.e. the font size, the colour and the angle will be the same).

In terms of the project, ‘Pillow’ will be used in the ‘Meme mode’ section. The idea is similar to the stories generator, the difference lies in the fact that the text is inserted into the image, rather than being printed below the image. Stories are presented in a post-like format.

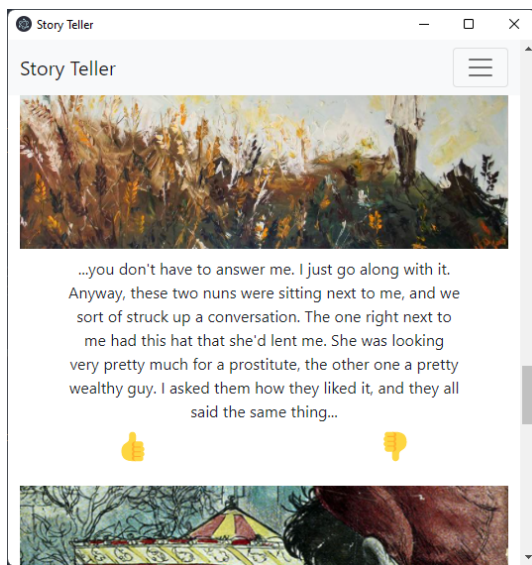
Frontend

The frontend of the project consists of a mixture of HTML, CSS and JavaScript. HTML is a markup language which is used to create the layout of the webpage. CSS is a language used to style the webpage. In this project it's used in a bundle with Bootstrap for easier manipulations with the styles. JavaScript is used to control the behaviour of the webpage and everything in it. Some of the functions displayed in the webpage are calculated and controlled solely with JavaScript with no interaction with Python.

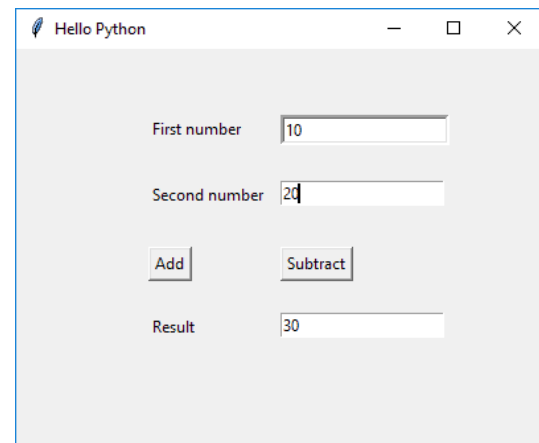
Web Application appearance

The application appearance is based on Bootstrap 5 and Electron. Bootstrap (Mark Otto & Contributors 2022) is a CSS framework/library that is used to create a responsive and mobile-first design. It's files (CSS and JS) can be accessed via 'Content Delivery Links' (CDN). This method is useful when only a part of the content hosted on CDN servers is used in development since CDN technology allows to deliver only necessary files. Also a big advantage is the loading time if the resource is server-based by reducing latency (CDNetworks 2021). The other aspect of the look is 'Electron' which was formerly known as 'Atom Shell' (Sawicki 2015). It's a cross-platform application that is used to develop application with web-like design. The main reason to work with 'Electron' is the ability to have the same look and feel on all platforms. GUI applications can be developed and deployed using web technologies much faster. Prior to the introduction of 'Electron' to the project, the application was developed on the basis of different browsers and operating systems.

Also as can be seen on the figures below, most popular approach to the development of Python GUI apps - Tkinter, doesn't satisfy the standards of app design in early 2020s.



Story Teller (Electron)



Tkinter app

Figure 1: A comparison of two GUI layouts

Similar situation occurs when it comes to animation with Tkinter. Many examples and many guides were examined and it has been concluded that Tkinter, in spite of the fact that it's one of

the most popular GUI libraries, is not suitable for the purpose of the project because the aim of the project is to create a fancy web-based application that can catch the attention of the user.

Another aspect of the design is the theme. Since many modern projects offer a possibility to change to either a light or a dark theme, it has been decided to develop an application with two themes. Bootstrap does most of the work for the theme change in terms of the colour change. Theme is persistent and is to a settings file which is stored in the application's folder. On every start-up the application checks if the theme is set to the light or dark mode accessing the JSON file.

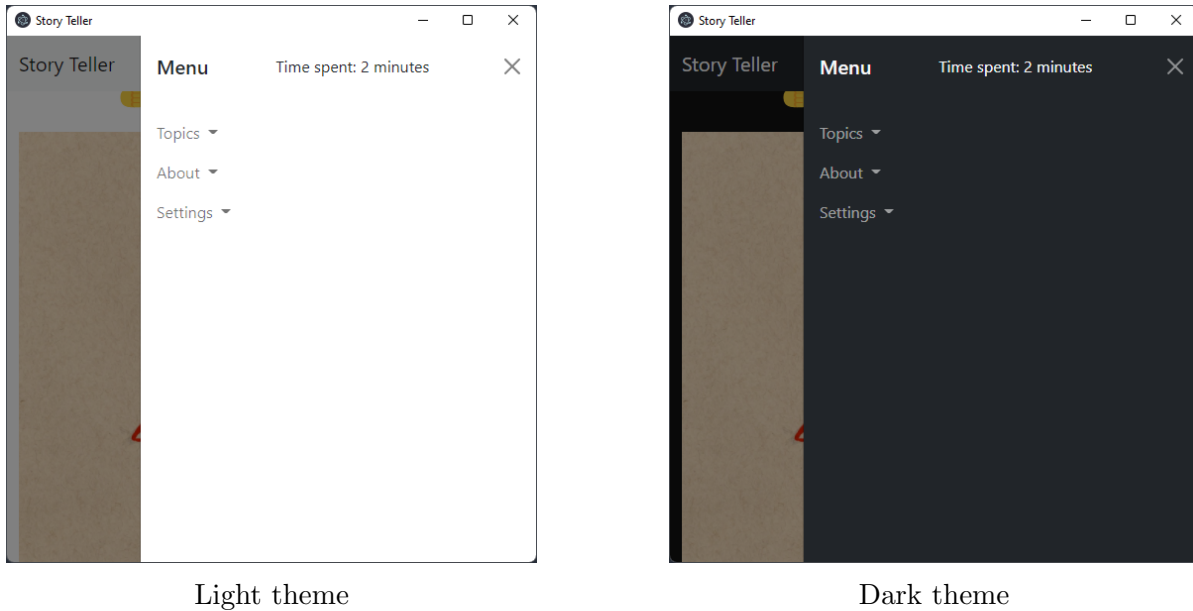


Figure 2: A comparison of two GUI themes

Displaying the content

To display the content it has been chosen to follow modern techniques of infinite scrolling. It loads the content continuously while the content is being generated on the fly. Infinite scrolling absolutely matches the topic of the project since the main idea was to create stories without any administrator intervention and minimise any unnecessary clicks and interactions. The content is displayed in a post-like format - that means a random picture from a set is followed by a random and freshly generated text based on a neural network checkpoint. The displayed content depends on which topic is selected since the content is separated in different topics and displayed with different pictures.

Another aspect which was considered with exclusive scrupulousness is the conformity to accepted social standards. In view of the fact that neural networks do not understand major human problems of inequality, supremacism and just unacceptable language, a word filter can be used to filter out the most offensive words. A dictionary of such words can be formed and used to exclude the possible offensive language. Some content fed to the network might contain questionable statements and thus the output of the network will contain altered but still questionable words. The problem might be regarded from two perspectives: the first one is that it's not acceptable and the second one is that it's not even a problem because the output is generated based on the style of the original author.

User feedback

User feedback is an easy and effective way to improve the application and enhance the text generation process in particular. To establish a dialogue with a user the application uses a standard approach to the problem. The user can either like or dislike the text generated by the application by simply clicking one of the buttons. It is planned to store the reactions in the database and use them to manually assess the quality of the content. Comments are another possible solution to the user feedback problem but they seem too like a too complicated approach.

Database

Databases are used to store the user settings and the user feedback. The settings are stored in a JSON file and are accessed every time the application is started. JSON is a lightweight, human-readable text format that can be used with any programming language. The objects in the files are stored in an 'attribute-value' format.

The purpose of the traditional database is defeated since the application doesn't store any sensitive information. No encryption of the database is used either.

There are many other data serialisation formats such as XML, YAML, BSON etc. It has been decided to use JSON because of its simplicity and flexibility compared to YAML. YAML is a superset of JSON and it seemed to big for the purpose.

```
[
  {
    "Id": 0,
    "FirstName": "string",
    "LastName": "string",
    "Name": "string",
    "EmailAddress": "string",
    "TerritoryId": 0
  }
]
```

Figure 3: JSON database

Linking the frontend and backend

The main bridge between the frontend and backend used in this project is a ‘little Python library’ called ‘Eel’ (Knott 2022). It allows interaction between JavaScript and Python at every point of the development. The library is used to create a web server that can be accessed from the browser - Chrome, Safari, Firefox and last but not least - Electron.

The most important feature of ‘Eel’ is function exposition. It allows to call any function that is explicitly defined with a ‘@eel.expose’ decorator and use it anywhere in the JavaScript code and vice versa.

```
@eel.expose
def generate_text(run_name, length, nsamples):
    sess = gpt2.start_tf_sess()
    gpt2.load_gpt2(sess, run_name=run_name, reuse=tf.compat.v1.AUTO_REUSE)
    texts = gpt2.generate(sess, return_as_list=True, length=length, nsamples=nsamples, model_name='124M')
```

Figure 4: Eel exposition

It’s important to note that when a function call is made from the JavaScript code, it should be done in async/await fashion. This is because the function call is made in the background and the result is returned after some time.

The separation of the frontend and backend is an important action because it allowed to use the most efficient approaches on each side and also focus on each part individually.

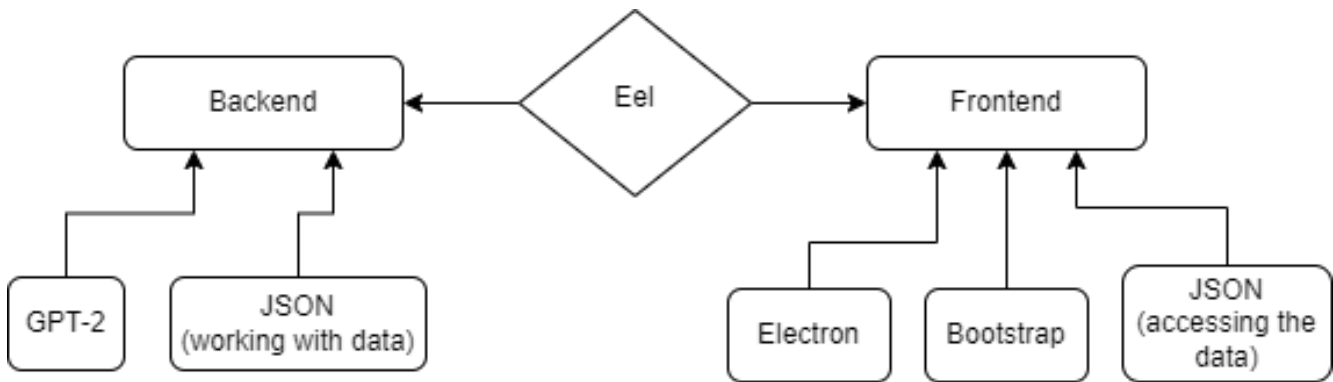


Figure 5: The relations in the system

Development lifecycle

Development process followed the principles of Agile practice. The main aspects of the development methodology are: continual improvement, flexible responses to any change and most importantly evolutionary development. Since the project is developed by a single person, not a team, some of the practices of agile methodology are not applicable.

Training the model

Implementation

Training or finetuning the model was a crucial step in the development process. Before the neural network would be able to generate specific text based on some dataset, it has to be trained on it. The training process involves tweaking many parameters of the model and the process is repeated until the best options are found. Currently, all three checkpoints are trained with the same parameters and on the same model - 124M (124 million parameters). The finetuning process begins with a creation of a new session. Then the the finetuning function itself is called with the following parameters: *dataset name, model name, number of steps, name of the run, restore point, overwrite option, length of the sample and an interval to save the checkpoints*.

Dataset name is the name of the preferred text file to train the network on. For this project three datasets were chosen - Alice, Catcher and Iliad. Model name is the name of the model used in the training process. It can be either '124M' or '355M'. The difference is in the parameter count. The '124M' model has 124 million parameters while the '355M' model has 355 million parameters. Number of steps is the parameter that determines how many times will the training process be repeated. For this project every checkpoint is trained more than 3,000 times in order to produce a meaningful result. Run name parameter specifies which checkpoint is being trained. It is useful to have multiple runs in the same model. Restore point and overwrite options are used for the sake of optimisation of the process to minimise the chance of failing (crashing). Also the interval to save the checkpoints is set to a small value of 100 to save the progress in case the process crashes.

```
def train_model(self, run_name, file_name, steps, sample_length):
    sess = gpt2.start_tf_sess()
    gpt2.finetune(sess, file_name, model_name=self.model_name, steps=steps, run_name=run_name,
                  restore_from='latest', overwrite=True, sample_length=sample_length, save_every=100)
```

Figure 6: Training function

Testing

Testing the training process is not a very long process itself, so it is not described in detail. The ‘GPT-2-Simple’ library offers many parameters to tweak and test the performance of the training process. All datasets were fed to the network and trained on a machine with a ‘NVIDIA RTX 3080’ GPU - this allowed an effective and speedy training process.

Several ways of training the neural network on the first dataset ‘Alice’ were tried. The first attempt included the raw format of ‘Alice in Wonderland’ text file without any preprocessing (i.e. text is decorated with numerous spaces, unnecessary punctuation marks and different symbols). The results from the training were acceptable and the model was able to generate some of the samples that had sense in them. The second attempt included the preprocessing of the original text file. Chapter enumeration, excess symbols and other punctuation was reduced in order to minimise the impact on the network. Paragraph spaces were left in the text. The generated text was compared with the original text and the results were satisfactory.

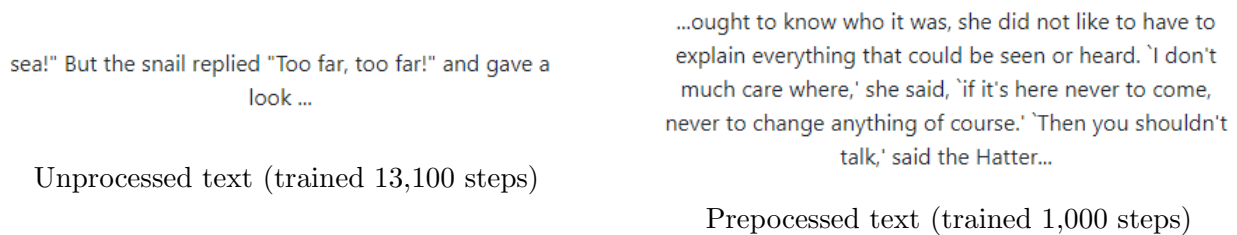


Figure 7: A comparison of unprocessed and processed datasets

The same process was applied to the second dataset ‘Catcher in the Rye’. It was cleaned from the chapter enumeration and then fed to the neural network. It was trained more than 13,000 times.

The Iliad training was the fastest in terms of adjusting the text before feeding it. Since it didn’t have any excessive decoration and the paragraphs didn’t have indents - the output text from the network was of a very good quality.

The finetuning parameters weren’t tweaked much. The learning rate stayed at 0.0001 the gradient count was set to 5. As for the steps, it’s impossible to identify a universal value for the best result. The best result for each dataset was achieved by taking into account the accuracy of each training step. The longer and more complex the dataset - the longer it takes to train. The final accuracy the neural network aimed to achieve was *0.1-0.2*. Therefore training times for every dataset are very different.

Reflection

Training the model was one of the most time-consuming parts of the project. Possibly, another method of training could have been used. There’s still a lot of freedom in what can be tweaked and changed in the parameter sections to enhance the results of the training. For example, the biggest enhancement would be to use a different model with more parameters. This project is limited computationally and therefore the model with the least parameters was chosen. The video memory of the machine wasn’t enough to train the model with 355 million parameters. There is a solution to that problem as well (use memory saving gradients and only train transformer layers) but it’s not implemented in Tensorflow 2.

Text generation

Text generation sprint could have been placed together with model training sprint. However, it was decided to separate the two because they are quite important on their own and whereas model training doesn't require frontend adaptation - text generation does because it is the final product that is being presented to the user.

Implementation

Text generation is the most important part of the project. The function takes a trained checkpoint of the network and based on the training it presents the user with a text. The function is written in a separate class based on the 'GPT-2-Simple' library. The function takes the following parameters: *run_name*, *length*, *nsamples*. These parameters are thought to be the most 'volatile' because they are changed often. The '*run_name*' parameter specifies which checkpoint is being used. The '*length*' parameter defines how many tokens the text should contain and the last parameter '*nsamples*' determines the number of samples to be generated.

```
@eel.expose
def generate_text(run_name, length, nsamples):
    sess = gpt2.start_tf_sess()
    gpt2.load_gpt2(sess, run_name=run_name, reuse=tf.compat.v1.AUTO_REUSE)
    texts = gpt2.generate(sess, return_as_list=True, length=length, nsamples=nsamples, model_name='124M')
```

Figure 8: Eel exposed generate function

For the purpose of the infinite generation it is important to specify the '*reuse*' parameter when the GPT-2 is being loaded. This parameter allows the model to be reused without the need to load it again. In the example above this parameter is set to a value directly taken from the Tensorflow library since 'GPT-2-Simple' provided explanation in the documentation doesn't work (setting the '*reuse*' parameter to 'True').

Yet this is only a first stage of the text processing in the function. After the text is generated it is handled by the algorithm that is shown on the below figure.

```
for idx, item in enumerate(texts):
    # STAYS FOR NOW
    item = item.rsplit('.', 1)[0]
    if not item[0].isupper() and item[0].isalpha():
        item = '...' + item + '...'
    else:
        item = item + '...'
    texts[idx] = item
return texts
```

Figure 9: Text processing algorithm

The presented algorithm is developed specifically for this project. It is based on the idea that the text which is output by the function is cut - it is been tested and proven in the previous section ‘Model training’. The incomplete text seems to be cut only in the beginning and the end of the sample so, the algorithm iterates over the samples and removes the last sentence from each of them. This is done using Python ‘rsplit’ function which is the reverse version of a standard ‘split’. After that, the first element of the word is checked for two conditions to decide where to put the ellipsis. First condition - uppercase, second condition - alphabetic character. If this is the case, the ellipsis is placed in the beginning and in the end of the word. If it’s not - only in the end.

The point of the ellipsis is to make the text more pleasing to the eye and slightly confuse the perception of the user. It’s been decided to call this trick ‘Ellipsis of Uncertainty’ - when the user sees the text that begins with ellipsis which signals that this is not the beginning of the text and ends with ellipsis which has the same meaning in perception, the text they won’t expect to see a fully understandable piece of text.

Testing

The testing process of the text generation involved the assessment of the generated text not only by the quality of the output but also by how it looks displayed on the page appended to the image. It might seem that this testing section overlaps with the testing section of the frontend development and it does, since the correlation between those elements is very high.

The main aspect of testing was the time taken to generate the text sample. It is important not to make the user wait for too long because it may result in a loss of the user’s attention. The model testing appendix shows different test cases and results of parameter tweaking. The most optimal parameters were found to be: length of the text - 100, number of samples - 6. It satisfies the time requirement of 15 seconds for generating all 6 samples on whichever topic the user chooses. ‘Alice’ checkpoint has been tested more than the other checkpoints due to the fact that it’s the first checkpoint that was developed. The results and knowledge gained from the testing are transferred from ‘Alice’ to other checkpoints successfully saving time.

Reflection

Output generation was a harder task to accomplish because it involved more than just using and adapting the functions from the library but also a ‘Ellipsis of Uncertainty’ trick had to be implemented. Also a lot of tests weren’t passed because the expected outcome was an uncut text. That case wasn’t reached and thus the trick described above was used. Besides, the text length was a question that had to be answered by testing what looks the best under a picture. Most likely, the reaction system will help to find out the best length and what exactly the user is looking for. Generally assessing, the result might have been much better if the model with a higher number of parameters was used. Having that said, it’s hard to make a neural network that knows little words and sentences, generate a text that is practically indistinguishable from the text that would come from a human. The benefit of having more parameters is that the model can represent much more complicated functions than with fewer parameters.

Reaction System

Implementation

Despite the simple explanation of the reaction system earlier in ‘Major Components’ section - the development itself was a bit more complex and involved the creation of a whole new logic system. Reactions now are used in a manual way by the developer to understand the problems of the system and make it more efficient and enjoyable to use. The idea is simple - if the user likes a generated post - they can react with a positive reaction (like). If they don’t - they can react with a negative reaction (dislike). A logic system was needed in order to restrict the user from liking and disliking one post at the same time. It checks the boolean value of the reaction and if it’s ‘True’, it’s impossible to react with the opposite reaction again. Every reaction is separate for every post and has its own unique ID which is linked to the post. That way, it’s easier to find the reactions in the database and understand a pattern of these reactions, since the main aim of the system is to gather collective information about the experience in whole and not one by one.

Apart from the logics - a database was implemented to store the reactions. JSON data serialisation format is used and every new reaction is stored there right away. Every object has several keys: *appVer*, *id*, *postID*, *postReaction*, *postTime* and *topic*. The addition process is rather simple, the JSON file is opened with a built-in function in Python, then the array of data is being formed. It includes all the values described earlier. App version is hard coded into the system whereas other values such as post ID, reaction, text and topic are passed into the function from the frontend. ID of the entry follows a different approach. It is being generated every time a new array is formed. The mechanism is the following: the last post ID is collected and incremented by one. Right after that this array is appended to the JSON object, the end of the JSON file (different from JSON object) is sought and the array is dumped into it. Indentation and sorting is maintained.

```
{
  "appVer": 1.0,
  "id": 0,
  "postID": "0",
  "postReaction": 1,
  "postText": "I walked all the way downstairs, instead of taking the elevator.",
  "topic": "catcher"
},
```

Figure 10: Example of an object

As every post can be reacted at, it can also be unreacted. This is an important feature which helps to tidy up the database in case if the user changes their mind or just missclicks. The system for unreacting is slightly more complex than reacting. It searches the database reversely and finds the last reaction of the user with a corresponding post ID. It must work from the end to the beginning because many entries might have the same post ID and only the last one should be deleted since it is the most recent one. The search is performed using a loop which iterates over the entries in the database. Then the post ID from the frontend (every like and dislike button have their IDs) is compared with the post ID of the entry. If they are the same, the entry is deleted and the loop is broken.

Testing

The testing stage was very extensive and everything was tested directly during the development sprint. The testing was done, as always, in a manual way without writing any automatic tests or using any test suites. It can be split in several parts: reaction buttons and database. The reaction buttons were tested only for logics and animation tests weren't conducted since they were considered unnecessary.

Buttons' logic is meant to be the following way: when the like button is pressed - the button is animated and dislike button is deactivated. And the dislike button should follow the same mechanism of action. Since an if-else statement was used to determine the logical system instead of switch-case statement, it was harder to create a logical system and this harder to test it (i.e. more tests were needed).

Database testing was conducted in a similar way. Entry creation and entry deletion was tested. It was crucial to have an understanding that it all works the way it should. Total of 13 tests were conducted in order to test the integrity of the system.

The test table of conducted tests can be found in the following appendix.

```
@eel.expose
def react(post_id, value, topic, post_text):
    with open('base/reaction.json', 'r+') as json_file:
        reaction_file = json.load(json_file)
        array = {
            "appVer": app_version,
            "id": reaction_file['posts'][-1]['id'] + 1,
            "postID": post_id,
            "postReaction": value,
            "postText": post_text,
            "topic": topic
        }
        reaction_file['posts'].append(array)
        json_file.seek(0)
        json.dump(reaction_file, json_file, indent=4, sort_keys=True)
```

React function

```
@eel.expose
def unreact(post_id):
    with open('base/reaction.json', 'r+') as json_file:
        reaction_file = json.load(json_file)
        for i in reversed(range(len(reaction_file['posts']))):
            if reaction_file['posts'][i]['postID'] == post_id:
                reaction_file['posts'].pop(i)
                # del reaction_file['posts'][i]
                json_file.seek(0)
                json.dump(reaction_file, json_file, indent=4, sort_keys=True)
                json_file.truncate()
                break
```

Unreact function

Figure 11: Two functions that have been tested

Reflection

The reaction system was one of the hardest things to implement since it has very convoluted logic. The main problem with it was the naming of the 'like' and 'dislike' buttons in the HTML document. After the issue has been solved, the implementation became rather simple. A better possible solution would be to use this feedback somehow to automatically improve the model and adjust the generated text without any manual intervention. This approach would make the app totally autonomous and would be a good point for the future development. Also not enough attention was paid to the functions that work with JSON. The formed array wasn't dumped into the JSON file and thus the database was not updated. This flaw extended the development time with a couple of hours.

Frontend Development

It has been decided to unify all the work carried out in the frontend into one section of the report. The reason for that is a belief that the frontend must work as a one single unit since it consists of many small components that are unable to work on their own. This section will include the implementation process of such elements as: navigation bar, menu, content loading, reaction animations, theme change and other elements related to frontend (created and coded in the JavaScript part of the project).

Many different ways of implementation and design were tried for the frontend. This section won't include the details of the earlier versions since they're not used in the project.

Implementation - Visible elements

The implementation of the navigation bar was one of the first things that was done in the frontend. Bootstrap 5 was used in order to create a responsive navigation bar. The navigation bar is a simple one with a hamburger menu button that leads to a dropdown menu that contains the several sections to tweak the application. Instead of a logo on the left of the navigation bar - application's name is displayed. Moreover, this name can be used as an active button to return to the beginning of the page. That is a very useful feature in an infinite scrolling app.

A dropdown menu implements three sections. The first one is 'Topics'. It contains three different sections that neural network can generate text on. The second section is 'About'. It contains information about the developer and a current version of the application. The third section is 'Settings'. A button to change the theme of the application is located there along with mode selection. There are two modes of the app - classic and meme mode. In the classic mode the app generates text beneath the images and the final content looks like posts. In meme mode the text is going to be appended directly onto the images to create meme-like effect. Apart from the sections, the menu has a timer built into it. The timer counts how much time has passed since the application was opened.

Post arrangement is designed following the main trends of the Internet nowadays. First comes the image then the description and only after that are positioned the feedback buttons. The feedback buttons are animated when the user hovers on them. It's a simple and effective way to show the user that the app is responsive and invites them to leave some feedback. If the desired reaction is clicked, the button stays enlarged. All the content is dynamic on the page and is generated on the fly. Images are pulled from the array which is being shuffled every iteration of post generation.

Theme change implementation is heavily based on the Bootstrap 5. All of the elements use class togglers to change the theme. Classes are implemented in Bootstrap and can be used on almost any element of the page. One of the complexities in the implementation of the dark theme laid under the hood is the generating of the new posts with according colour of the text. When a function is called to change the theme, the posts that are already created will be changed only. This is solved by creating an if-else statement that check for the theme variable value in the function that loads more content.

Implementation - Invisible elements

The invisible elements are the functions and the mechanisms that stand behind the beautiful and interactive elements. They are all implemented in the JavaScript and thus included in the ‘Frontend Development’ section.

The main invisible element is the Post generation system. The generation flow is as follows: a text is generated on a certain topic with a certain length. While it’s being done, the application shows a spinner to a user - which can be either white or dark, depending on the theme selected. Then when the text generation is over - an image array is shuffled and populated with images. Then the spinner is removed. After that comes the if-else statement that checks if the theme is dark or not to generate the the text with an appropriate colour. After, the reaction buttons are appended.

Another important aspect in loading is the point of time when to start generating and displaying another group of posts. The logics behind that are rather simple and are very well known in modern Internet applications. When the user reaches the end of the page and no content is available - the application will load more content.

```
// Special Listener to Load everything when user reaches the end of page
document.addEventListener( type: 'scroll', listener: ()=>{
    if ((window.innerHeight + window.scrollY) >= document.body.offsetHeight && !loadMoreInProgress) {
        loadMore()
    }
})
```

Figure 12: Function of lazy loading

This technique is known as ‘lazy loading’. The point of using ‘lazy loading’ in the project is to minimise the load on the user’s system and thus improve the user experience. Also there is no point in generating more posts than the user will be able to see in the long run.

Nowadays there are many ways of implementing lazy loading mechanism. Beginning from 2020 it’s considered a web standard and therefore can be declared as an attribute in the HTML document. However, for the purpose of the project the HTML approach doesn’t work since the objects are generated on the fly and thus they cannot be appended to the DOM beforehand.

Moving onto what’s under the hood of the design solutions - the time displayed in the navigation bar is calculated via comparing two timestamps. One timestamp is created when the application is launched and the other is created and compared with the first one every 60 seconds. Strictly cosmetic feature is that the function checks for the spent time and decides whether to display a singular or a plural word for ‘minute/s’.

Another under the hood explanation is about the theme and topic change mechanism. The function calls Python via Eel and uses built-in Python functions to manipulate the JSON settings file and retrieve or write information to it. Topic and theme values are saved in the file and retrieved when the application is launched or whenever they are needed.

Testing

Testing the frontend and all its features was very important to achieve the final satisfactory result. It's been carried out during the whole development process. The testing process for the visible and invisible elements of the frontend was conducted simultaneously because they depend on each other. The greatest challenges in testing were to ensure that the application works as expected and no errors will come up in the future. Visible elements were tested on a principle of a visual testing. Questions like: 'Does it look right?', 'Is there a way to make it appear better?' and similar were asked to verify the correctness of implementation. Invisible elements on the other hands followed a functional testing approach. Every visible element must have had a working implmentation in the backend. Since lazy loading heavily depends on text generation process - the performance testing has been declared successful because the text generation completed in a reasonable time.

Reflection

The overall process of frontend development was something very different from the other parts of the project. It included many different attempts to create the best looking design that would suit the app purpose. The development of the design was one of the first things that was done in the whole project and only then the backend development process commenced. First attempt of the design involved no additional libraries and frameworks and was done with plain HTML and CSS. Later, this idea has been changed towards innovation and the use of Bootstrap 5. No doubt, it could have been done without using additional resources but it was a good idea to use Bootstrap in order to make the code more readable and maintainable thus making the code creation more efficient and speedy. The hardest thing to implement in the frontend was to competently wire up the function call from the frontend and a response from the backend when it comes to lazy loading of the content. JavaScript allowed easy implementation of all the aspects of the look and design functionality. Choosing JavaScript in a bundle with Python was definitely a good idea.

Speaking about frontend features, there are some not very functional features that were implemented but through them the developers can show their own attitude towards the project and user experience in whole. From a commercial point of view, having a timer that counts how much time the user has spent in the app scrolling through stories, is not a good idea, since they are likely to blame themselves for the big amount of time spent but I think it's a great addition.

Critical analysis and Reflection

The project overall, despite the difficulties and challenges, slightly incorrect approaches and the lack of knowledge in regards to neural networks and artificial intelligence, was a successful proof of concept that neural networks can synthesise human-like text and be used in entertainment. Considering the great amount of work put into this project and the number of papers and articles, guides and documentations gone through, it involved a significant and important learning process. Without gaining an understanding in the field of artificial intelligence, the project was not possible to complete. Yet in spite of the fact that the project was a success, there are a lot of different possibilities that could have been explored and implemented. Many features could have used a different approach. But these considerations of a different approach came during the process of implementation and eventually it was decided to stick as much as possible to the initial approach that was chosen.

Libraries and frameworks

GPT-2-Simple

GPT-2-Simple is the library used to work with the neural network and perform the most important tasks of the project. It's a small adaptation of a big library called Tensorflow and allow a fast and effective integration in the project without consuming much time that can be used to accomplish other tasks. It wraps many functions from Tensorflow in its own functions narrowing down the scope of the library to the most important functions. Yet it's highly adjustable and can be tweaked in an almost any possible way. Also, it provides ready-to-use pretrained models of GPT-2 with reasonable amount of parameters making it possible to develop an app with a reasonable size.

Eel

Eel is also a library that has been used in order to complete this project. This is a library that allows to connect frontend which is written in HTML, CSS, JavaScript to interact with the backend which is written in Python. It was chosen because it's a small library that is very easily set up and worked with. Also it has a good documentation on GitHub and satisfies the needs of the project. A library CEF Python could have been used instead of Eel but it seemed to be more complex and a bit too much.

Bootstrap

Bootstrap on the other hand is a frontend framework used to make the development process more efficient since it provides a big amount of ready-to-use components. The development of the design that is featured in the application now would be much more complicated and time consuming if Bootstrap was not used.

Pillow

Pillow is a library used in the backend and implementing a ‘meme mode’ without it would be impossible because it’s used to manipulate images and put text directly on them. Also, the genuine look of the generated memes is very important and is due to the presence of the font called ‘Impact’ and an added stroke to it. All this is done with help of Pillow.

Tools

Python

Python is the main tool used in the project since it’s a very simple and effective language. It’s very well known for the amount of libraries written in it.

Bibliography

CDNetworks (2021), ‘Cdn vs local’.

URL: <https://www.cdnetworks.com/web-performance-blog/cdn-vs-local/>

Chen, T., Xu, B., Zhang, C. & Guestrin, C. (2016), ‘Training deep nets with sublinear memory cost’, *arXiv:1604.06174 [cs]*.

URL: <https://arxiv.org/abs/1604.06174>

Clark, A. & Contributors (2022), ‘Pillow’.

URL: <https://github.com/python-pillow/Pillow>

Johnsen, P. (2013), ‘Alice in wonderland’.

URL: <https://gist.github.com/phillipj/4944029>

Kay, A. (2007), ‘Tesseract: an open-source optical character recognition engine’.

Accessed Jan 8 2022 [Online].

Knott, C. (2022), ‘Eel’.

URL: <https://github.com/ChrisKnott/Eel>

Mark Otto, M. C. & Contributors (2022), ‘Bootstrap’.

URL: <https://github.com/twbs/bootstrap>

Musstafa (2022), ‘Optimizers in deep learning’.

URL: <https://medium.com/mlearning-ai/optimizers-in-deep-learning-7bf81fed78a0>

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I. (2019), ‘Language models are unsupervised multitask learners’.

Accessed Jan 6 2022 [Online].

Rosenblatt, F. (1958), ‘The perceptron: A probabilistic model for information storage and organization in the brain.’, *Psychological Review* **65**, 386–408.

Sawicki, K. (2015), ‘Atom shell is now electron — electron’.

URL: <https://www.electronjs.org/blog/electron>

Stevenson, D. C. (2000), ‘The iliad’.

URL: <http://classics.mit.edu/Homer/iliad.mb.txt>

Valenzuela, C. (2017), ‘the-catcher-rye’.

URL: https://github.com/cvalenzuela/rwet/blob/master/final/source_text/ny/the-catcher-rye.txt

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L. & Polosukhin, I. (2017), ‘Attention is all you need’.
Accessed Jan 7 2022 [Online].

Woolf, M. (2021), ‘gpt-2-simple’.

URL: *<https://github.com/minimaxir/gpt-2-simple>*

Appendices

Testing

Model testing

Case	Input	Expected	Actual	Result
GEN0	Length: 10	No cut text	Text cut	Fail
	Samples: 1	Time: <15s	Time: 6.64s	Pass
	Batches: 1			
GEN1	Length: 1023	No cut text	Text cut	Fail
	Samples: 1	Time: <15s	Time: 12.99s	Pass
	Batches: 1			
GEN2	Length: 500	No cut text	Text cut	Fail
	Samples: 6	Time: <15s	Time: 30.48s	Fail
	Batches: 1			
GEN3	Length: 64	No cut text	Text cut	Fail
	Samples: 2	Time: <15s	Time: 5.41s	Pass
	Batches: 2			
GEN4	Length: 60	No cut text	Text cut	Fail
	Samples: 2	Time: <15s	Time: 5.33s	Pass
	Batches: 2			
GEN5	Length: 60	No cut text	Text cut	Fail
	Samples: 2	Time: <15s	Time: 5.43s	Pass
	Batches: 1			
GEN6	Length: 70	No cut text	Text cut	Fail
	Samples: 2	Time: <15s	Time: 5.63s	Pass
	Batches: 1			
GEN7	Length: 100	No cut text	Text cut	Fail
	Samples: 6	Time: <15s	Time: 9.46s	Pass
	Batches: 1			
GEN8	Length: 100			
	Samples: 6	Time: <15s	Time: 12.12s	Pass
	Batches: 1			
GEN9	Length: 100			
	Samples: 6	Time: <15s	Time: 11.66s	Pass
	Batches: 1			

Table 1: GEN0-GEN7 - ‘Alice’, GEN8-GEN9 - ‘Catcher’

Reaction system testing

Case	Input	Expected	Actual	Result
RS0	Action: Like Set: Alice Post: Any	Multi-line string	Single-line string	Fail
RS1	Action: Dislike Set: Alice Post: Any	Single entry	2 same entries ID 6 and ID 7	Fail
RS2	Action: Dislike Set: Alice Post: Any	Single entry	2 same entries ID 8 and ID 9	Fail
RS3	Action: Undislike Set: Alice Post: Any	Do nothing	Entry added	Fail
RS4	Action: Create file (r+) Set: - Post: -	File created	Error	Fail
RS5	Action: Create file (a+) Set: - Post: -	File created	File created	Pass
RS6	Action: Append to empty Set: - Post: -	Appended	Error	Fail
RS7	Action: Like Set: Catcher Post: Any	Single entry	Single entry	Pass
RS8	Action: Dislike Set: Catcher Post: Any	Single entry	Single entry	Pass
RS9	Action: Unreact Set: Any Post: Any	Dislike inactive (when like is active)	Dislike active	Fail

Table 2: Reaction system sprint

Case	Input	Expected	Actual	Result
RS10	Action: Unreact			
	Set: Alice	Complete logics	Complete logics	Pass
	Post: Any			
RS11	Action: Unreact			
	Set: Any	Delete entry	Not deleted	Fail
	Post: Any			
RS12	Action: Unreact			
	Set: Any	Delete entry	Deleted	Pass
	Post: Any			

Table 3: Reaction system sprint