A Generative Chatbot with Natural Language Processing

David Liebman david.c.liebman@gmail.com

April 5, 2020

Coordinator 1 Coordinator 2

Department of Computer Science

SUNY New Paltz

Abstract

Our goal in this thesis is to create a chat-bot, a computer program that can respond verbally to a human in the course of simple day-to-day conversations.

We use a deep learning neural network model called the Transformer to develop the chat-bot. A full description of a Transformer is provided. We show the use of a few different Transformer-based Natural Language Processing models to develop the chat-bot, including the Generative Pre-training Transformer 2 (GPT2). For comparison we include a Gated Recurrent Unit (GRU) based model. Each of these are explained below.

We are also interested in installing the chatbot code on a small device such as the Raspberry Pi with speech recognition and speech-to-text software. In this way we might be able to create a device that can carry out a verbal conversation with a human. For the GRU-based model we can use a Raspberry Pi 3B with 1GB RAM. A Raspberry Pi 4B with 4GB of RAM is needed to run a chatbot with the GPT2.

Contents

1 Background/History of the Study **1**

1.1 Background 2

1.2 Recurrent Neural Network Components 2

1.2.1 Overview 2

1.2.2 Gated Recurrent Unit 3

1.3 Sequence to Sequence and Translation 5

1.3.1 Word Embeddings 5

1.3.2 Corpus 6

1.3.3 Training and Evaluation 6

1.3.4 Input Tokens 7

1.3.5 Encoder 8

1.3.6 Decoder 8

1.3.7 Output Tokens 10

1.3.8 Loss and Accuracy During Training 10

1.3.9 Attention Mechanism 11

1.3.10 Sequence to Sequence Chatbot 12

2 Transformers and The Generative Pre-training Transformer 2 **15**

2.1 Transformer and Attention 16

2.1.1 Byte Pair Encoding 17

2.1.2 Attention 17

2.1.3 Scaled Dot-Product Attention 17

2.1.4 Decoder - Key and Value 20

2.1.5 Decoder - Query . . . 22
2.1.6 Input - Positional Encoding . . . 23
2.1.7 Output - Feed Forward Network . . . 24
2.1.8 Visualization - Transformer . . . 24
2.2 The Generative Pre-training Transformer 2 Model . . . 24
2.2.1 Pre-Training . . . 25
2.2.2 General . . . 25
2.2.3 Corpus . . . 27
2.2.4 Releases . . . 27
2.2.5 Application Details . . . 28
2.2.6 Visualization - GPT2 . . . 29

3 Experimental Design and Setup **30**
3.1 Approach to the Study . . . 31
3.2 Model Overview . . . 31
3.3 Setup . . . 32
3.3.1 Graphical Processing Unit vs. Central Processing Unit . . . 33
3.3.2 Raspberry Pi . . . 34
3.3.3 Tensorflow vs. Pytorch . . . 34
3.3.4 Speech and Speech To Text . . . 35
3.3.5 Corpus Considerations . . . 36
3.4 ARMv7 Build/Compile . . . 36
3.4.1 Pytorch 'torch' Library 1.1.0 For ARMv7 . . . 36
3.4.2 Pytorch 'torch' Library 1.4.0 For ARMv7 . . . 37
3.4.3 Docker Container 'tensorflow-model-server' For ARMv7 . . . 38

4 Experimental Results **40**
4.1 Experiments - Installations . . . 41
4.1.1 Questions . . . 41
4.1.2 Checklist . . . 42
4.2 Chatbot - Gated Recurrent Unit Model . . . 42
4.2.1 Questions . . . 42

4.2.2 Checklist . . . 43
4.3 Smart Speaker - Gated Recurrent Unit Model . . . 44
4.4 Chatbot - Transformer Model with Persona Corpus . . . 44
4.4.1 Training . . . 44
4.4.2 Questions . . . 45
4.4.3 Checklist . . . 46
4.5 Smart Speaker - Transformer Model with Persona Corpus . . . 46
4.6 Chatbot - Transformer Model with Movie Corpus . . . 47
4.6.1 Questions . . . 48
4.6.2 Checklist . . . 48
4.7 Smart Speaker - Transformer Model with Movie Corpus . . . 49
4.8 Chatbot - Generative Pre-training Transformer 2 Model . . . 49
4.8.1 Context Experiment . . . 50
4.8.2 History Experiment . . . 50
4.8.3 Artificial Intelligence Markup Language Experiment . . . 51
4.8.4 Program Launching . . . 51
4.8.5 Overall . . . 52
4.8.6 Questions . . . 52
4.8.7 Checklist . . . 53
4.9 Smart Speaker - Generative Pre-training Transformer 2 Model . . . 54

5 Conclusions **55**
5.1 GRU vs. Transformer . . . 56
5.2 Smaller Chatbot Transformer Learning . . . 56
5.3 Word Usage . . . 57
5.4 Sentence Usage . . . 58
5.5 Turing Test . . . 59
5.6 Winograd Schema . . . 60

A Further Installations **62**
A.1 Generative Pre-training Transformer 2 - Large Model . . . 62
A.1.1 Context Experiment . . . 62

A.1.2 History Experiment 63
A.1.3 Artificial Intelligence Markup Language Experiment 63
A.1.4 User Name Experiment 64
A.1.5 Internet Search Experiment 65
A.1.6 Usage Example 66

B Abbreviations **68**

List of Figures

1.1 Recurrent Neural Network 3
1.2 Word Embeddings 6
1.3 Sequence to Sequence Architecture 9
1.4 Loss and Accuracy 13

2.1 Lowering Dimensionality 18
2.2 Attention Output 19
2.3 Matching Input and Output 20
2.4 Transformer Encoder and Decoder Flow 21
2.5 Transformer Encoder and Decoder 22
2.6 Transformer Encoder and Decoder 23
2.7 Visualized Attention 24
2.8 Generative Pre-training Transformer 2 26
2.9 Visualized Attention GPT2 29

4.1 Loss - Larger Transformer Model 47

5.1 Word Usage 57
5.2 Simple Word Usage 58
5.3 Simple Sentence Usage 59

List of Tables

GPT2 Size Overview 27

Model Overview 32

Chapter 1

Background/History of the Study

1.1 Background

In their paper Vinyals et al [1] discuss making a chatbot using a neural network configured for sequence to sequence neural machine translation. An attempt to code our own sequence to sequence model was not very fruitful so instead we use code authored by Inkawhich et al [2].

In their paper Vaswani et al [3] discuss using the Transformer architecture for solving machine learning tasks. We train a transformer model as a chatbot.

Also Radford et al [4] discuss the 'Generative Pre-training Transformer 2' (GPT2) neural network for Natural Language Processing (NLP) tasks. The GPT2 model is based largely on the Transformer architecture.

We implement a chatbot with a GPT2 model. We use a program library from Wolf et al [5] to run our model.

We use Transformer style models and also a Recurrent Neural Network model to allow for meaningful comparison. We explain both types of models below.

It is worth noting that with the appearance of the Transformer architecture some traditional technologies have become redundant or obsolete. This may be true of any model that uses Recurrent Neural Network components and also the traditional word vector embeddings.

1.2 Recurrent Neural Network Components

1.2.1 Overview

The goal behind Recurrent Neural Network (RNN) components is to detect patterns. Since understanding RNNs is important to GPT2 and Transformer models, we discuss RNNs here. Here we explore a simple RNN first.

Imagine that we want to use as input a string of words. This could be a sentence used for the input of a translation task.

The simplest Recurrent Neural Network unit has two inputs and two outputs. They can be arranged in patterns. In our example the input will be a sequence of data as stated above, and the Recurrent Neural Network will be a line of components of the same length as the data.

One input from each component is the hidden state output from the Recurrent Neural Network to the left. Another input is the current input from the sequence that the component is encoding.

One output is the generated hidden state, meant for the component to the right. The last output is the value that the Recurrent Neural Network outputs or surmises.

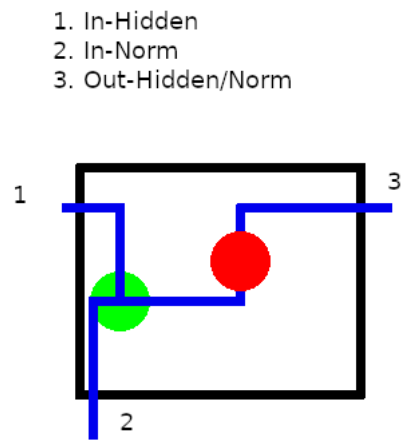

Figure 1.1: RNN - 1 and 2 are inputs, 3 is the output.

In our diagram the two inputs are labeled 1 and 2, and the single output does double duty as both the hidden state output and the value that the Recurrent Neural Network outputs or surmises. This is labeled 3.

There are several designs for a Recurrent Neural Network component. The inner workings of these components are what makes them different. In the example in the diagram the inner workings are very simple. Two paths, labeled as inputs, take data into the Recurrent Neural Network. Their data is combined in the green circle. This combination is done with concatenation and simple feed forward neural network components. The output from the concatenation is passed through the red circle. This is a 'tanh' activation operation that limits the output to values from -1 through 1. This 'tanh' activation keeps the output within reasonable values. Finally the output is delivered outside the component to the program employing the Recurrent Neural Network. In this diagram there is only one output. The single output would serve as both the hidden state output for that position in the encoder or decoder, and also the data output component for that position.

1.2.2 Gated Recurrent Unit

An implementation of a Recurrent Neural Network is the 'Gated Recurrent Unit' (GRU). There are other varieties of RNN, most notably the Long Short Term Memory (LSTM) cell. We do not explore

LSTM cells here.

A GRU has two inputs and two outputs. The formulas for a Gated Recurrent Unit, as outlined by Denny Britz in the website WILDML (Britz et al) [6], are as follows.

$$z = \sigma(x_t U^z + s_{t-1} W^z)$$

$$r = \sigma(x_t U^r + s_{t-1} W^r)$$

$$h = tanh(x_t U^h + (s_{t-1} \circ r) W^h)$$

$$s_t = (1 - z) \circ h + z \circ s_{t-1}$$

The Gated Recurrent Unit has two inputs and two outputs. It also has two internal gates. One internal gate is the 'reset' gate. This one determines how much of the previous input is combined with the new value calculated by the mechanism of the Gated Recurrent Unit. It is denoted as 'r' above. Another internal gate is the 'update' gate. The update gate decides how much new information is to be included in gate computation. It is denoted as 'z'.

Here 's_t' is the symbol for the output. The two inputs are 'x_t' and 's_{t-1}' . 'x_t' is the hidden state input. 's_{t-1}' is the regular input for the Recurrent Neural Network or Gated Recurrent Unit. Sigmoid activation is used on the two gates, using the symbol '$\sigma(...)$' , while tanh activation is used to compute the hidden output. The hidden output is found in the third line, denoted as 'h'.

The dimension of the U^z, U^r and U^h matrices is the hidden unit size by the hidden unit size. x_t is a vector the size of the hidden unit. The U values, along with W^z, W^r and W^h are all simple matrices that allow the GRU to operate.

In the last line, the regular output is determined using the 'dot' product which is denoted with a circle, along with an addition operation. In the two gate formulas (the first and second) the output is determined as the sum of two matrix multiplication operations passed through sigmoid activation. This produces values in the range of 0 to 1.

Under most programming circumstances the Gated Recurrent Unit is not implemented by the average programmer. The programmer employs a language like python and a programming library like Pytorch or Tensorflow. The library then implements the Gated Recurrent Unit and makes it easy for the programmer to use that implementation.

1.3 Sequence to Sequence and Translation

Translating text from one language to another has become a common task for computers. The Sequence to Sequence architecture is often used today for this purpose. Here we explain how this works.

A naive approach to translation involves using a dictionary. You would encode each key as a word from one language and the value for that key would be the translated word in the target language. Of course this doesn't work, because different languages not only have different words for the same thing, but they also have different sentence structures for what might be similar concepts.

A better approach for a computer is sequence to sequence translation. A description follows with a section at the end for how training works.

In this approach we use recurrent neural networks to obtain our translation. Two recurrent neural network components are employed. One is responsible for the input language and the other for the output.

1.3.1 Word Embeddings

Also employed are two vocabulary sets. One vocabulary is for the source language and another is for the target language. A table of word vectors the size of the input vocabulary is created and a maximum sentence length is picked. There is also a 'hidden size', which is an important dimension in the model. In practice the hidden size could be 300 units and more for this kind of application.

Words are translated from strings to individual numbers from an input vocabulary dictionary. The dictionary only contains a single unique number for every word. Then the number is passed through an embedding structure. This turns the single number into a vector of numbers that is the same size as the RNN hidden dimension. Then, from that time on the model uses the vector instead of words.

The contents of the embedding unit is a table of numbers, all are of the size of the RNN hidden dimension. The vectors are usually, but don't have to be, unique values. There is one complete hidden dimension sized vector for each word in the vocabulary.

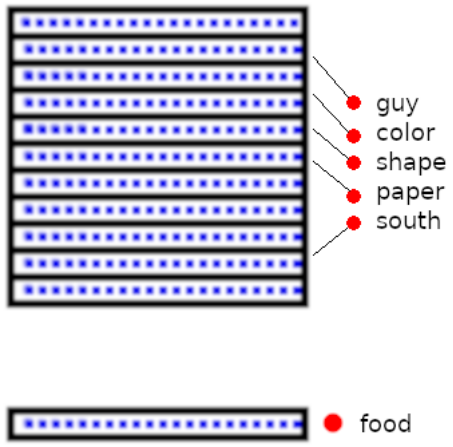

Figure 1.2: Embeddings - Each word from a dictionary is converted to a vector of numbers.

The vectors can be initialized randomly or they can be filled with predetermined values. As the network trains the embedding values can either be modified or frozen in place. Typically if the contents were initialized randomly the values would be trained. If the contents were filled with predetermined values you don't want to train them or change them in any way.

There are at this writing two main types of pretrained word embeddings. One is called 'Word2Vec' and one is called 'GloVe'.

Word2Vec is short for 'Word to Vector.' (Mikolov et al.) [8] GloVe is short for 'Global Vector.' (Pennington et al.) [9] .

1.3.2 Corpus

A text must be prepared for training. A text corpus with source and target pairs is chosen. Sentences in the source corpus are paired with sentences with the same meaning in the target corpus. Sentence length is observed and for all sentences shorter than that length a special 'end-of-sequence' token is appended to all sequences. This restriction is applied in both languages.

1.3.3 Training and Evaluation

We have a corpus. It is generally felt that the corpus should be divided into three parts. The largest part, maybe %80 is held for training. Then %10 should go to what is called 'validation' and %10 should go to 'testing'.

The training portion, obviously, is used during the training phase. The model can become familiar with training data and provide target data that leads one to believe that the model is learning well. This is similar to memorization. The model may see the same training data repeatedly and thereby learn how to answer a source with a target without learning the task of why it is giving a particular answer.

For this reason, during training, the validation set is used for statistical purposes. The actual training mechanism is halted and the validation set is run through the model. Loss is measured and compared to the loss found when training is enabled. Loss during validation is usually greater than during training. This is done periodically during what would otherwise be called training.

This is useful for tuning hyper-parameters. Over and over you train, and as you do you look at the validation for clues on what to change in your hyper-parameters.

When you are done you use the final holdout set, the 'test' set, to determine weather or not you have been successful. You must always remember not to train with the test set.

1.3.4 Input Tokens

So far the model takes a word, translates it to an integer, and finds the vector in the word embedding table that represents that word. It does this for the first word and all subsequent words one by one. Then it gives the entire vector for a word to the GRU, one at a time. The GRU takes the word and passes it to some inner components. It decides weather to return as output just the input or the input modified. This is what the Gated Recurrent Unit does internally.

The Gated Recurrent Unit takes as input two vectors. It processes the 'input' vector and returns another vector. This could be exactly the same as the input but is usually somehow changed. The input vector and the output vector have the dimension of the 'hidden size' mentioned above. Throughout the discussion of this model the hidden size will remain the same. The Gated Recurrent Unit also operates on two hidden states. One hidden state, a vector, is taken in and another hidden state, also a vector, is produced for output.

We will describe components in two groups. The input components, from the source language, are the encoder and the output components, from the target language, are the decoder.

1.3.5 Encoder

The input segments, composed of Gated Recurrent Units, take two input vectors and return two output vectors. One input is the vector from the embedding table. Another input vector is the previous hidden state. The hidden state is the same size as the input from the embedding table, but typically it comes from the previous Gated Recurrent Unit. The output vector is a hidden value for the Recurrent Unit to the right.

The very first Gated Recurrent Unit in the input encoder ignores the fact that the first word has no hidden value. It consumes the first word vector. Then it passes it's output to the next Gated Recurrent Unit in the encoder. This Gated Recurrent Unit uses the output of the previous Gated Recurrent Unit as the hidden value. It also uses the vector for the second word. It passes it's important information to the Gated Recurrent Unit to it's right. If needed this is repeated. Then the last Gated Recurrent Unit in the encoder passes it's hidden state to the output decoder.

A complicating detail is that although many Gated Recurrent Units are called for in the encoder they all use the same set of weights and biases. For this reason only a single input Gated Recurrent Unit is used for all of the words in practice. Outputs are calculated and then cycled around and fed with the next word from the sentence in vector form to the input of the Gated Recurrent Unit.

1.3.6 Decoder

The decoder is in charge of generating tokens that represent, in this case, the translation of the input to the output language. The output uses Gated Recurrent Unit segments also. The first hidden input for the first output cell is taken from the last hidden output of the last recursive unit of the input. It is important because it is the spot where a great amount of data is passed from the encoder to the decoder. The connection at this point is said to carry the 'thought vector'. Most of the information responsible for translating one language to another is passed at this point.

The hidden values from the input section are passed to the first output Gated Recurrent Unit. It outputs the values that are later converted to the first word of the output. The first output Gated Recurrent Unit also has a hidden state. It passes the first word and the hidden state on to the second Gated Recurrent Unit to it's right.

The second Gated Recurrent Unit generates the second word and also its own hidden state. The second word is recorded and the word and the hidden state are passed on to the next Gated

Recurrent Unit. This is repeated until a special 'end-of-sequence' token is generated or until the number of tokens equals the maximum number allowed.

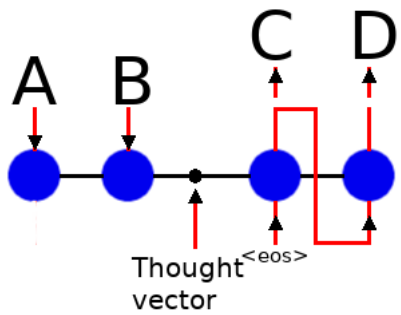

Figure 1.3: Seq2seq: A and B represent an input sequence and C and D represent the corresponding output.

In this figure we generalize a sequence to sequence model. The idea is that the left side of the diagram deals with the encoding of sentences. A and B would be consecutive words in a sentence, and the round blue nodes below A and B are Recurrent Neural Network units. C and D are outputs and in the right side of the diagram the blue nodes represent the output Recurrent Neural Network units. Between the input and the output there is a corridor of information exactly the size of the Recurrent Neural Network hidden vector.

All of the information that the decoder uses for it's output is present in this corridor and is passed along the corridor from the encoder. For this reason we refer to it as the thought vector. Aside from Attention Mechanisms, there is no other place where information is passed from the encoder to the decoder.

Making this vector larger by increasing the size of the hidden dimension allows for more information in the thought vector. Size also increases the time to train the network. The size must also match the dimension of the vectors in the GloVe or Word2Vec download if one of those is used.

Ultimately exceedingly large hidden dimension does not improve the sequence to sequence model.

Again in the Decoder, many GRU units are called for but they all share the same weights and biases. For this reason a single GRU is employed for the entire Decoder. The encoder and decoder are never the same unit.

1.3.7 Output Tokens

Each output we have is currently in the form of a vector. These vectors are a long string of floating point numbers, each one the dimensions of the 'hidden size' mentioned above. They are converted to the dimensions of the output vocabulary, through matrix multiplication. Then they are processed in what is called a 'softmax' function. This processing, along with an 'arg-max' function, determines the index of the maximum value in the new vocabulary sized vector. This index allows the program to look up the corresponding word in the output vocabulary. This word is then used as the model output at that point in the output decoder.

There are some inherent problems. Because the output of a Gated Recurrent Unit is constantly being reused as the input, lots of data that might be useful to have is lost when the Gated Recurrent Unit churns through internal operations where several matrix multiplication operations are performed together on input. With every iteration more data is lost and so, for example, the effective length of the input and output sentences must be short.

Another problem is that all input to be translated to target output has to be boiled down and passed to the output section through a small corridor the size of a typical word vector. This channel is sometimes referred to as the 'thought vector.' Without attention, described below, all necessary information must be right there. This also limits the length of the input and output vectors. It does help, when setting up a model for later training, to make the hidden size larger, but it only helps so much. There is a point at which the benefit of increasing the hidden size is lost.

This is how some computer models do language translation. Using 'arg-max' is an example of a greedy approach to decoding. Another approach might use something like 'Beam Selection' but we're not going to get into that here.

1.3.8 Loss and Accuracy During Training

At first the prediction from a model is not very close to the desired output. The output is compared to the prediction and a 'loss' is generated. 'Loss' measures the difference between the predicted output and the target. A larger loss represents two values, prediction and target, that are further apart.

Another metric is Accuracy. 'Accuracy' is a numerical representation of the difference between the desired output and the generated prediction. It is a percentage of the time that the output is

exactly correct.

Getting a prediction, running input data through a neural network, is forward propagation. Training, then, is a mathematical process involving backpropagagtion. Backpropagation identifies areas of the model weights that need to be modified in order to get the proper prediction in the future.

We take the derivative of the loss function in order to backpropagate. The derivative is manipulated with the learning rate. The original weight value is changed minutely. The amount changed with every backward propagation is dependent on the learning rate. The result is a set of adjusted weight matrices and a new loss. When these matrices are used later they allow for better predictions.

This is done over and over with every source/target sentence pair. Slowly the model is changed and predictions start to match the target. That's training. The loss should decrease over time and the accuracy should increase.

There are several numerical metrics that we can record during training that tell us how our model is training. The loss, a mathematical calculation of the difference between the model's output and the predicted value, is mentioned above. Loss is an important number. Also accuracy is important. Accuracy is a mathematical calculation of the difference between the model's output and the value that output should be, but it focuses on the number of times the model comes out with a correct prediction verses how many output values there are in total.

1.3.9 Attention Mechanism

Attention Mechanisms are used by sequence to sequence models to transfer more information from the encoder to the decoder. As stated above there is only one place where the encoder imparts information on the decoder, the 'thought vector'. Attention helps the encoder tell the decoder which word is more important. This stressing of a vector by the model is the attention mechanism attending to one output.

Here we consider a simple attention mechanism that is used in the Sequence to Sequence model by Inkawhich et al. [2]. The concept for this attention comes from Luong et al [11].

Luong et al [11] are interested in three kinds of calculation for their attention mechanism. The three methods use slowly increasing levels of complication. First they propose a method that just uses the dot product. Then they propose a method that just uses a field of weights. Finally they use a method that uses concatenation, along with a field of weights and a pass through a 'tanh'

activation layer.

$$\boldsymbol{score(h_t^{\intercal}\bar{h}_s)} = \begin{cases} \boldsymbol{h_t^{\intercal}\bar{h}_s} & \text{dot} \\ \boldsymbol{h_t^{\intercal}W_a\bar{h}_s} & \text{general} \\ \boldsymbol{v_a^{\intercal}}\tanh(\boldsymbol{W_a[h_t;\bar{h}_s]}) & \text{concat} \end{cases}$$

Here h_t is the symbol for the output of the current decoder and $\bar{h}_s$ is the symbol for another output taken from the input encoder. This one is the entire set of encoder states. $h_t^{\intercal}$ stands for h_t transposed. Inkawhich et al [2] uses the 'dot' variety.

The formula is used after the decoder Gated Recurrent Unit calculates it's hidden state. It is below.

$$\boldsymbol{score = h_t^{\intercal}\bar{h}_s}$$

Basically the output of the current decoder is transposed. Then it is multiplied by the hidden value from the entire set of encoder states. Not pictured here, the score is multiplied by the Gated Recurrent Unit decoder output, and then passed through a 'tanh' activation layer. It becomes the decoder output.

1.3.10 Sequence to Sequence Chatbot

Vinyals et al [1] make an interesting proposition. They say that it's possible to make what they call a Neural Conversational Model by constructing a Sequence to Sequence model, but instead of using two corpus from different languages a single language is used for both source and target.

Chatbots have for a long time been constructed using AIML. AIML, (Artificial Intelligence Markup Language) requires hand coding individual rules for every input and response. A Neural Conversational Model would not use those sorts of rules.

To create a model like this more is required than just a single input and output language. There must be a relationship between the source and the target. We want there to be a question-and-answer-like relation. Finding a corpus like this can be difficult.

It would be easy to duplicate the input source material in the target corpus. This would produce

auto-encoding. The model would learn to repeat everything that was given to it in it's output. Though the model learns a task, it is not the dynamic task we want. Conversations, on the other hand, supply the relationship we are looking for. Starting with almost any question sentence in a conversation, the sentence following it answers the question posed.

What would be a good candidate for this kind of verbal play? Vinyals et al [1] use a movie transcript data set. Essentially they take movie dialogue and break it into sentences. Consecutive sentences can be divided into source and target. Each sentence is a source and the sentence that follows it would be the target for that pair.

Using this method there are times when the source and target are not from the same conversation, like times in a movie dialog when the scene switches from one locale to another. Comparing this, though, to the number of times that the two sentences are from the same dialogue, the movie transcript database serves very well.

Training for this model is relatively straight forward. There is a problem though. We can follow the loss and make sure it is decreasing, but the accuracy tells us nothing and in this case does not increase as we would like it to. The loss goes down but the accuracy does not go up.

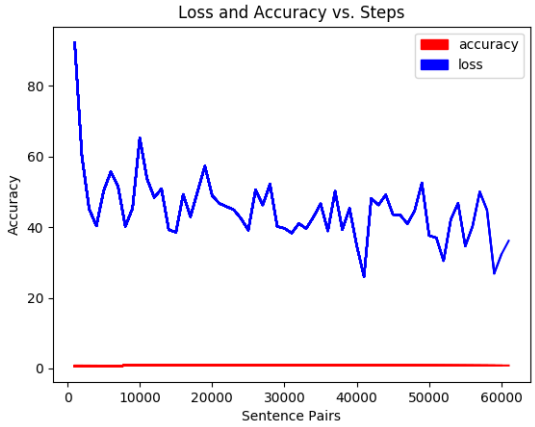

Figure 1.4: Loss and Accuracy: Red is accuracy and blue is loss.

This is because the source and target do not have the same meaning. The model does learn the task at hand but during training we have to ignore the accuracy. Success is usually measured by the accuracy of the holdout test set. Here we must measure the success with a subjective examination of the trained model.

We have to interactively give the model questions that we might ask someone that we are having a conversation with, and then see how it answers. Over and over we have to test the model. If we are satisfied with the answers then the training was a success.

Chapter 2

Transformers and The Generative Pre-training Transformer 2

2.1 Transformer and Attention

The Transformer is a mechanism that is based entirely on attention. Strictly speaking this is not the attention explored in the Sequence to Sequence model in Chapter 1, though there are some similarities. It is a model that uses no recurrent components.

Recurrent components have some negative qualities. Recurrent components are hard to run with batch input data. Also they do not work with very long data strings.

If a batch of data is to be cycled through a recurrent component, we would like all that batch to go through all the components at once. We want this to happen for whole batches at a time.

The problem is that in Neural Machine Translation there are times when data has to be processed by the recurrent units one batch at a time. First the data is passed to the first RNN. Then the batch can be passed to the next RNN. Remember that a single GRU is used in our example for all words in the encoder. Another GRU is used in the decoder. These GRU's can not process the subsequent batches all at once. They have to wait for the previous batch to be done. This is a bottleneck has to do with how the RNN is used.

Also we don't know how long each input sentence or output sentence is going to be. If sentences cannot be made to look like the same length to the RNN's, they must be handled individually. This ruins the batch concept.

This sort of bottleneck is found in Neural Machine Translation that is based on recurrent elements.

Also because the Recurrent Network is so heavy with Neural Network components, many weights and biases, though they can remember patterns, they loose some information with every pass. This is why there is a practical limit to the length of the input sequences that the typical Recurrent Neural Network can use. This is why the length of sentences in network models that use the Recurrent Neural Network are short.

Transformers use no Recurrent Neural Network components. Their operations can be parallelized so that large batches of data can be processed at once during the same time step.

Longer sequences can be considered as well, so Transformer input can contain longer English language sentences and even paragraphs.

We will discuss one layer of a multi-layer Transformer below. Transformers are usually constructed on eight or more of these layers in the decoder and the encoder.

2.1.1 Byte Pair Encoding

BPE stands for 'Byte Pair Encoding.' WordPiece is a particular implementation of Byte Pair Encoding.

WordPiece is used by some transformer systems to encode words much the way that Word2Vec does. Like Word2Vec, WordPiece has a vocabulary list and a table of embeddings that maps one word or token to a vector of a given size.

WordPiece, though, handles Out Of Vocabulary (OOV) words gracefully. It breaks large words into smaller pieces that are in the vocabulary, and has a special notation so that these parts can easily be recombined in order to create the input word again. Byte Pair Encoding is not so interested in pre-trained word embeddings like Word2Vec and Glove.

For the Generative Pre-training Transformer 2 a version of Byte Pair Encoding is used instead of a vocabulary system like Word2Vec or Glove.

Some form of Byte Pair Encoding is included in almost every Transformer typed Neural Network model, so no decision needs to be made about what type of word embeddings to use.

2.1.2 Attention

Attention mechanisms are used in a similar way in three places in the Transformer model. The first implementation of the Self Attention is discussed below. Each of these attention mechanisms is contained in a layer. There are typically the same number of layers in the encoder as in the decoder.

It should be noted that input to the Transformer is strings of words from a desired input language. Output is similarly words in a given language. Input words are treated very much the way that they are in Sequence to Sequence models. A word is translated to a number and that number indexes an entry in a word-to-vector table. From that time forward a word is represented by a vector of floating point numbers. In a transformer this word vector can be large. In the original paper Vaswani et al [3] use a vector size of 512. Later, in our discussion of Generative Pre-training Transformer 2, we will see vector sizes of 768 and 1280.

2.1.3 Scaled Dot-Product Attention

Each layer of the Transformer has a signature self-attention mechanism. This is possibly one third of the entire Transformer mechanism, but a variety shows up in the other two-thirds.

The first thing that happens is the input word vector is converted to three other values. These new vectors are like the input vector but they have a smaller dimensionality. Converting the word vector in this way is accomplished by three simple matrix multiplication operations.

In the diagram below a simple conversion of this type is illustrated. In the diagram we convert a vector with dimension of 1x3 to a dimension of 1x2. In a real world example we are converting a vector from 1x512 to 1x64. This is a division of 8.

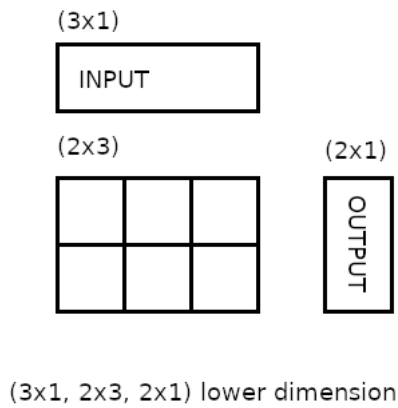

Figure 2.1: Lowering Dimensionality

One thing we want to do is to preserve the dimension of our starting vector. We start with a 512 sized floating point vector and after some processing we want to return to the same size. Before that is done the vector is processed at the smaller size of 64 floating point numbers.

In this self-attention scheme three vectors are actually required. All three vectors are sized 64, and all three are converted by separate matrix multiplication operations. The weights to convert each of the three vectors are different. For this reason the new smaller vectors are all different.

The smaller vectors individually are called q, k, and v. They can also be referred to as larger matrices. The new vector matrices are denoted as Q, K, and V. Q stands for 'Query'. K stands for 'Key'. V stands for 'Value'. The lower-case names refer to single vectors and the upper-case refer to matrices. These are essentially batches of input.

The Query value is multiplied by the Key values from all vectors in the input. This multiplication is 'dot-product' multiplication. When it is done all keys will have low output except those that are closest to the Query. Then the results are passed through a softmax function. When this is complete

there will be a single vector that is close to 1 and another group of vectors that are all close to 0.

The vector produced by multiplying the softmax with the V values of every word produces a single word vector that is close to its original value, and many others that are near zero.

This formula from Vaswani et al [3] shows the process.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Here the value of $\sqrt{d_k}$ is used to limit the size of the QK^T output and d_k is the dimension 512. Without this the softmax function has to deal with much larger numbers. Smaller numbers for the softmax are preferred. K^T is notation for the K vector transposed.

The function can actually perform this on large matrices with high dimensionality, in parallel. This parallel matrix operation increases the speed of training.

In the green triangle in the Figure 2.2 we preform the multiplication and selection that was just described.

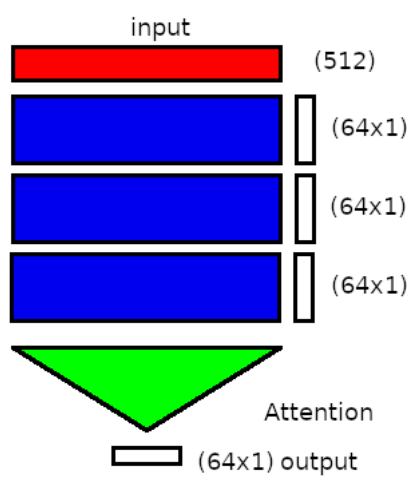

Figure 2.2: Attention Output

Finally the output we calculated above must be returned somehow to the input dimensionality. This is accomplished by duplicating the procedure described eight times with eight separate weights. When this is done the output of the attention mechanism is concatenated together, returning the output to the proper size.

This multi-headed approach allows different heads to learn different types of relationships, and

then when they are grouped together the learned relations are recovered and contribute to the output.

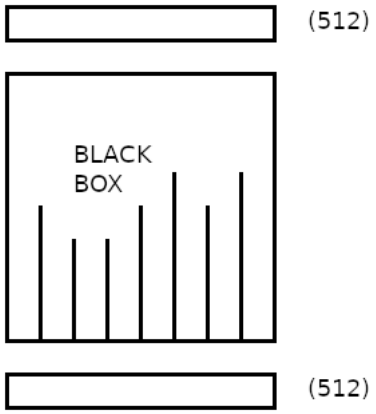

Figure 2.3: Matching Input and Output

Later the output is passed through a feed forward network. It is also re-combined with the original input again through addition. Then the output is normalized. This makes sure that the values are all within reasonable ranges. This recombination of the attention output with the original output is done throughout each Transformer layer.

This describes the encoder section. There are two other attention segments. Together these two sections combine to form the decoder section. This is repeated for each layer.

2.1.4 Decoder - Key and Value

The decoder is composed of two attention mechanisms and a feed-forward segment at each layer. The result of the encoder's work is passed to the decoder and remains applied to one of the decoder's attention mechanisms in each decoder layer. In that one of the two attention mechanisms of the decoder the 'Key' and 'Value' matrices are imported from the encoder.

While the encoder takes in the entire input and attends to whatever portion of that input it finds to be important, the decoder is interested in producing one output token at a time.

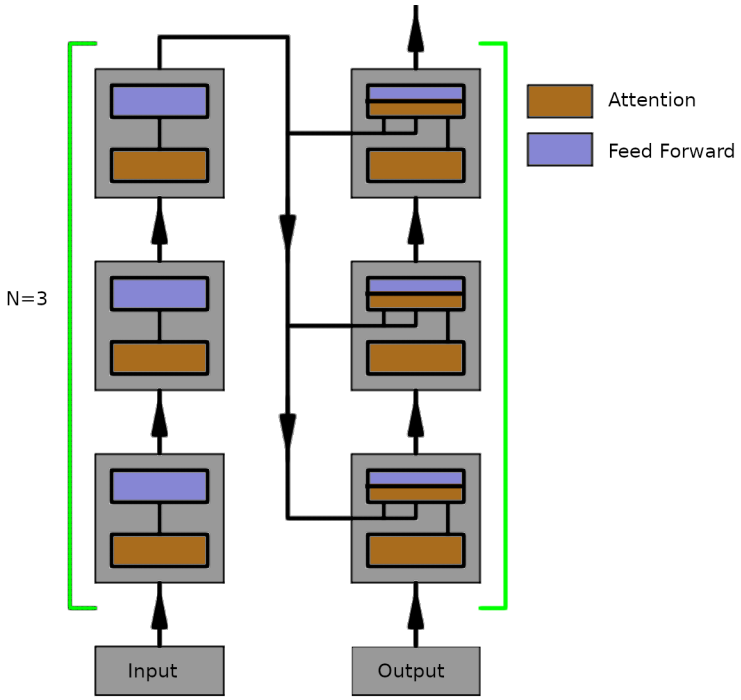

Figure 2.4: Transformer Encoder and Decoder Flow. - Three layers and data flow.

In our flow diagram we are not concerned with most sub segments of the entire transformer. We are interested in the flow of data through the three encoder segments and the different flow of data through the decoder segments. The encoder is largely serial, while the decoder is serial and parallel. In fact each decoder segment includes a feed forward part, and all decoder and encoder parts include a residual connection where the input is added back to the output of the attention and feed forward segments.

We illustrated in the Sequence-to-sequence discussion the importance of the single 'thought vector'. The Transformer can be seen as having a thought-vector also. There is a corridor of data from encoder to decoder. Two important smaller vector-sized inputs from the encoder are ultimately required in all layers of the decoder. They represent the 'Key' and 'Value' matrices. The matrices required are the size of the smaller, reduced, vector. The full sized vector is transported from the encoder and is reduced dimensionally in the decoder layers to two smaller matrices.

This full sized vector comes from the last encoder layer's output. Typically there will be as many decoder layers as there are encoder layers. The output from the last encoder layer is applied to the 'Key' and 'Value' inputs of one of the attention mechanisms in all the decoder layers.

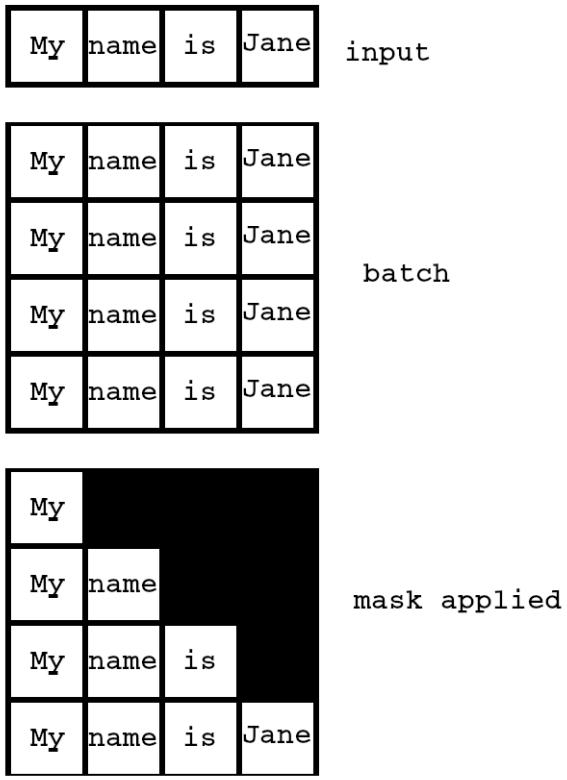

Figure 2.5: Transformer Mask. - Decoder uses masked input.

2.1.5 Decoder - Query

There is also a second attention mechanism in each decoder layer. It works solely on data from the decoder itself. It works very much like the attention mechanism from the encoder - only it attends to every word of output as opposed to the entire input sequence. It passes it's output to the attention mechanism described above. This data is lowered in dimensionality and becomes the 'Query' matrix for that mechanism.

During training the decoder section is masked. The mask prohibits the decoder from seeing parts of the target that it should not during training. This mimics the inference setup and still allows for large input matrices.

During inference the decoder produces a token and then it adds to that token, one at a time, until the decoding is finished and something like English is produced. It can attend to any part of the output it has already produced. During training the decoder is exposed to the target sequence

under a mask.

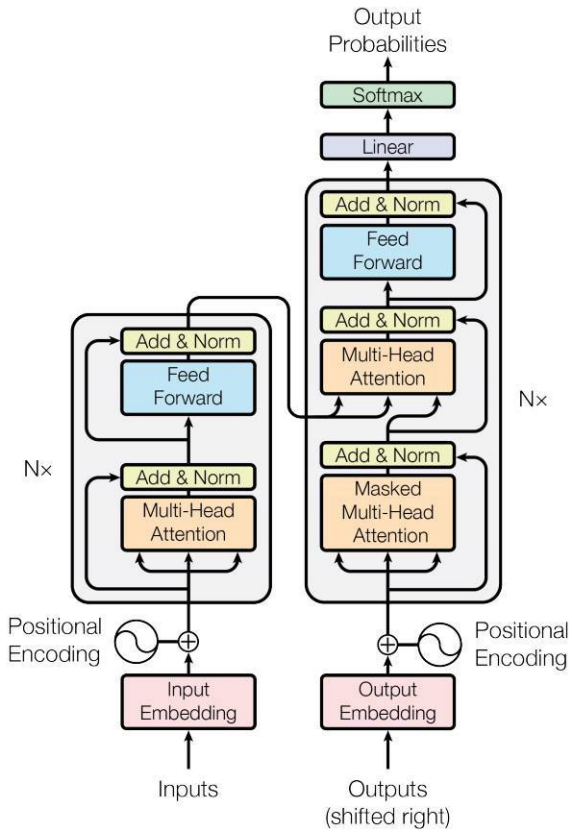

Figure 2.6: Transformer Encoder and Decoder. - Nx shows number of layers. - Vaswani et al [3]

2.1.6 Input - Positional Encoding

The input of the Transformer encoder and decoder layers employ not only a word-vector table, but also a positional encoding scheme. The model adds to the input vector information that it can then use to learn the position of words in a sentence.

Words that are early in the sentence have a certain appearance and words later on appear differently. The Encoder and Decoder use sine and cosine waves to impart this information onto the sentence sequence.

2.1.7 Output - Feed Forward Network

At the output of the last layer of the decoder the output vectors are processed through a linear matrix which increases the vector's dimensionality so that the output vector is the size of the output vocabulary dimensionality. After the linear matrix the vector is processed by a softmax function. Then the highest floating point value in the new larger vector is the index of the chosen output word.

2.1.8 Visualization - Transformer

In order to visualize what is happening during inference we have colorful charts that we can look at. In this chart we are looking at how each word attends to all the other words in the input text.

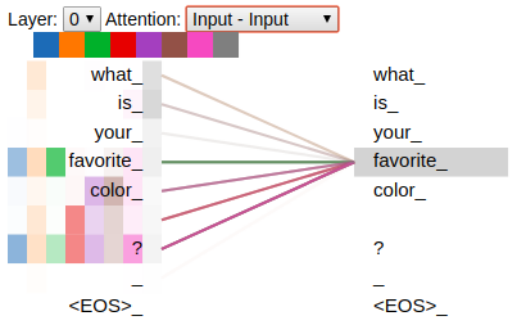

Figure 2.7: Visualized Attention – 'favorite' shows attention to some but not all words in the sentence.

Significant is that words, like 'what' and 'your', do not have strong attention to words in the text on the left. In a chart like this one they would show no colors on the left and light colored lines connecting the left to the right.

This diagram is from the Transformer with the larger hyper-parameter set that we describe in Chapter 3, trained on the movie dialog corpus.

2.2 The Generative Pre-training Transformer 2 Model

'Generative Pre-training Transformer 2' (GPT2) is a large model. It is based on the Transformer from Vaswani et al [3] but there are some major changes. The model uses the decoder portion of

the Transformer without the encoder. There are some other changes to the output layers. Another big difference is that it is pre-trained and downloadable.

2.2.1 Pre-Training

Pre-Training is when the authors of a model train an instance and then make the model available to the public on-line. This is helpful for the average programmer interested in Neural Networks. Training an instance of the transformer model can use up computation resources for days, and require hardware that is costly. Usually the cost of producing a trained model is prohibitively expensive.

After acquiring a trained model, the programmer goes on to adjust the model to their task. Adjusting a pre-trained model to a given task is called 'Transfer Learning'. Many tasks lend themselves to Transfer Learning. Conceptually a model can be fine-tuned to any problem and many problems can be addressed with good results after only modest fine-tuning.

There is a pre-trained Transformer model called BERT. BERT stands for 'Bidirectional Encoder Representations from Transformers'. The BERT files available on-line are mainly for classification tasks. They allow for input that uses a vocabulary size that is very large, but the output is meant to be smaller.

The 'Bidirectional Encoder Representations from Transformer' models use a bidirectional training method. During training the BERT model might be presented with a sentence with a word missing. The training task is to fill in that blank. The model can look at the sentence from any direction in order to find that word. This is a very simplistic explanation for what is termed a Masked Language Model.

2.2.2 General

'Generative Pre-training Transformer 2' still uses Scaled Dot-Product Attention. A model diagram is taken from Radford et al [12]. A mask is used in the Self Attention segment of the model during training.

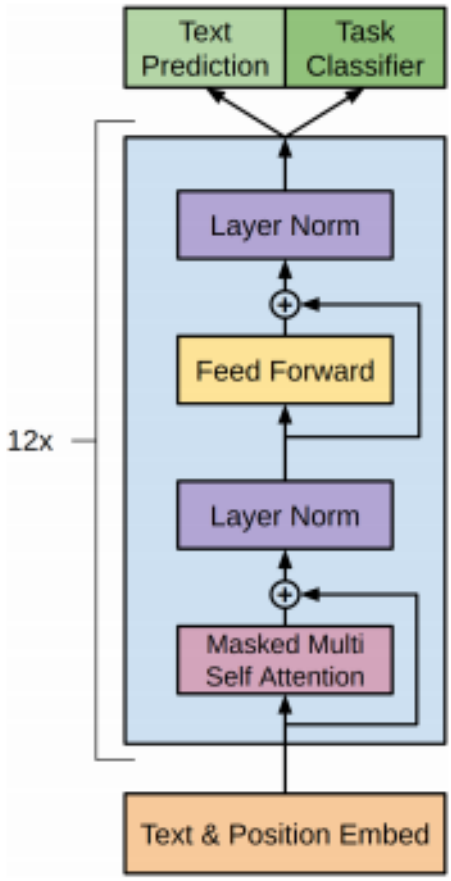

Figure 2.8: GPT2 - Radford et al [12]

There are several sizes of pre-trained 'Generative Pre-training Transformer 2' model. They are all rather large. The smallest model matches the size of the largest 'Bidirectional Encoder Representations from Transformers' model. This model has 12 layers while the Transformer model in the example from Vaswani et al [3] uses 8 layers. This model also has a hidden dimension of 768, not 512. With 8 heads this leaves a smaller dimensionality of 96 at each attention head.

The GPT2 models input and output text sequences. In this way they are preferred for our application over the BERT models.

2.2.3 Corpus

The 'Generative Pre-training Transformer 2' models are trained on a corpus called WebText. WebText is a 40GB corpus that is taken from the Reddit web site. All the material comes from before 2017 and all the material has a 'carma' rating of 3 or better. 'Carma' is a rating system used internally on Reddit.

As with the decoder layer of the Transformer model, the GPT2 model concerns itself with generating words that are later strung together to make sentences or paragraphs. During training the model uses a masking scheme so that input can be parallel-ized. During inference output cannot be parallel-ized, so during inference output must focus on one example at a time.

2.2.4 Releases

In their paper Radford et al [4] show that their model can generate text from a seed sentence or paragraph. At the time the case was made that the largest 'Generative Pre-training Transformer 2' models should not be released because of their ability to generate text that might fool humans into believing that another person was responsible for the text. Later the larger models were released to the public.

Size	Parameters	Layers	d_{model}
small	117M	12	768
medium	355M	24	1024
large	774M	36	1280
x-large	1.5B	48	1600
xx-large	8.3B	72	3072

At the time that the first 'Generative Pre-training Transformer 2' model was released the size of the models was mis-stated, but the documentation was not updated immediately. Most values in the table above show sizes that were actually released. The final xx-large model was trained by NVIDIA Applied Deep Learning Research [13] and was not released to the public.

The 'Generative Pre-training Transformer 2' models also work in many circumstances in 'zero-shot' mode. This is when you use the model pre-trained but without transfer learning. There is no extra training that goes on to make the model suit the task. It is used 'as is'.

For our chatbot the model with 117 million parameters worked. Some programming was required to make the model output look like chatbot output, but the model itself was not modified.

We use both the small and large model. As a test when the larger 774M model was released it was used as a substitution for the 117M model. The test worked, and returned answers that were more well formed than the small model. The larger model does not fit on a Raspberry Pi and so it was not employed on a permanent basis. Using the extra large 1.5B parameter model in a chatbot was not attempted.

2.2.5 Application Details

The model is described in Radford et al [4] and the accompanying blog post. The model is trained on English without a stated problem. Large neural network models are usually trained with a stated problem in mind. Rather famously this model is used after training to generate English language text. The model takes input from the user, a premise or summary of what is to be generated. The model also takes as input a number called the 'temperature.' Then the model generates output. As the 'temperature' is set higher the output is more fanciful. There is also a tune-able parameter for the output length.

Given the ability of the model to invent content, it was determined by the authors that the 'large' model should not be released to the public at first. Months later the 'large' model was released.

For our chatbot we set the temperature to a low number. We set the length of the output to a sentence-length number of tokens. Then as input we use the output from the speech-to-text translator.

The output is interesting but not useful right away. Traditional programming and string manipulation are employed to clean the output and render a short single sentence. This is our final output.

Because the input is meant to be a number of sentences, and because we are using a transformer-based architecture, we have room in the input string to add more information along side the user's question. In this respect the model acts to summarize the input.

With every input string we include a set of three or four sentences. They include the time, the bot's name, and the bot location and occupation. All of these are invented. What happens is the chatbot summarizes the input and only if the information is relevant then the same information is used by the model as output. Making this possible is the fact that a transformer can accept much

longer input strings than a Gated Recurrent Unit, and generate much longer output strings.

Surprisingly the chatbot answers most of our questions in the first person. We feel that WebText, the Reddit corpus, has many examples of sentences in the first person.

2.2.6 Visualization - GPT2

During inference the Scaled Dot Product Attention in the GPT2 focuses on certain words as it processes input text. Here the word 'favorite' shows a relationship to many of the other words in the text.

Figure 2.9: Visualized Attention GPT2 – 'favorite' shows attention to some but not all words in the sentence.

In our experiments the phrase 'What is your favorite color?' is usually answered with 'I love the colors of the rainbow.' This answer does not mention a specific color, as one might expect it should. The diagram might support this observation because 'color' on the left is not heavily highlighted. Words like 'what' and 'your' are barely considered at this head at all.

Chapter 3

Experimental Design and Setup

3.1 Approach to the Study

Several neural network models are used in the project. One is the sequence to sequence model and another the transformer model for a generative chatbot and finally the Generative Pre-training Transformer 2 model.

We will not try to rewrite the transformer or GPT2 model ourselves.

In this project we attempt to load as much of our chatbot code onto a Raspberry Pi as possible. We have trained models using the pytorch and tensorflow libraries. These models are responsible for taking in English sentences and producing English output. There is another part of the typical Raspberry Pi setup that includes another neural network component. Speech to text models, which our application requires, rely on large neural network resources. For this purpose we use speech to text resources supplied by Google on the google cloud. To include speech to text libraries locally on the Raspberry Pi would be too costly in computation time and resources like RAM. It would also be complicated to implement technically. It could easily comprise an entire project on its own.

Unfortunately the speech to text resources supplied by Google cost money. To use the service you need to have a billing account with Google.

The speech to text service used on the project and the memory limitations on the Raspberry Pi leads one to ask the question weather the neural network responsible for the chatbot function could not be servable from some faster machine located somewhere on the internet. At this time we are not interested in serving these resources. It would entail two calls from the Raspberry Pi for every sentence. This complicates things and also has a time overhead.

Also, we have several models that we want to test. To test them all would require several servers. In addition we use both Pytorch and tensorflow. Tensorflow has 'tensorflow-model-server' for serving models, but Pytorch has no equivalent.

It is important to note that the large Generative Pre-training Transformer 2 model specifically could be served from a remote computer and it would operate faster. Currently on the Raspberry Pi decoding a single sentence takes approximately 13 seconds. Even so, we prefer to install our trained models on the Raspberry Pi directly.

3.2 Model Overview

Here we itemize a short description for each row in the table.

Model Name	File Size	RAM Train	RAM Inference	Hand Trained	Raspberry Pi
Seq-2-Seq/Tutorial	230 M	1.1 G	324 M	YES	3B
Transformer/Persona	25 M	556 M	360 M	YES	NONE
Transformer/Movie	550 M	6.5 G	1.5 G	YES	4B
GPT2 small*	523 M	5 G	1.5 G	NO	4B

* a large GPT2 model exists, but it is not small enough to fit on a raspberry pi.

- **Sequence to Sequence - Tutorial** This model uses the sequence to sequence architecture and the Gated Recurrent Unit component. We hand-coded our own example of this model but it performed poorly. This model is the slightly modified version of the Sequence to Sequence model based on the tutorial from Inkawhich et al [2]. It actually uses the Movie Dialog corpus.

- **Transformer - Persona** This model uses a Tensorflow Transformer architecture. There was some coding involved to get the model to interface with the text-to-speech and speech-to-text libraries. There was also some coding to load our own corpus data during training. The model parameters describe a rather small model. This model also uses the Persona Dialog corpus. It is not on a Raspberry Pi board.

- **Transformer - Movie** This model is based on the transformer model above but uses the Movie Dialog corpus and a parameter set that is larger. In many ways this model is bigger than the model that uses the Transformer and the Persona corpus.

- **GPT2 small** This model was downloaded from the internet. It fits on a Raspberry Pi 4B with the 4GB RAM option. Some modification was made so that model output was suitable for our purposes.

3.3 Setup

We use linux computers, sometimes with GPU hardware for parallel processing. We also use the Python programming language. Code from this project can be run with the 3.x version of Python.

When the project was started we did some programming with Keras using Tensorflow as a backend. Keras was later discarded in favor of Pytorch and Tensorflow. Tensorflow and Pytorch do not work together. Pytorch as a library is still under development at the time of this writing.

Some of the Generative Pre-training Transformer 2 code uses Pytorch. Some of the Transformer

and Generative Pre-training Transformer 2 code uses Tensorflow. There is a repository on Github that has the GPT2 trained model using Pytorch instead of Tensorflow.

We use github as a code repository. Code corresponding with this paper can be found at: https://github.com/radiodee1/awesome-chatbot .

As a coding experiment we rewrite the code for the sequence-to-sequence Gated Recurrent Unit model. We have varying amounts of success with these experiments. We do not rewrite the Generative Pre-training Transformer 2 code from the Tensorflow or Pytorch repository.

3.3.1 Graphical Processing Unit vs. Central Processing Unit

A CPU has a number of cores, a number usually between 2 and 16. A CPU is designed, though, to execute one command at a time. This allows for a logical program that can be executed. A CPU has limitations when it comes to executing matrix multiplication. Matrix multiplication using a CPU can take a long time.

GPUs, Graphical Processing Units, have the ability to address tasks like matrix multiplication with many more processing units at once. The GPU speeds up parallel processing and have a benefit to neural networking training tasks that the CPU doesn't have.

Unfortunately state of the art neural network models are larger than the capacity of a single GPU. Some models are trained on many GPUs simultaneously. It is not uncommon for a model to train on a computer with eight GPU cards for many days. Training these models is prohibitively expensive for the average programmer. It is possible to rent time on Amazon Web Services or Google cloud with well outfitted computers to do this training, but this can be costly.

This sort of situation is addressed partially by the Transfer Learning scheme. In Transfer Learning someone else trains the model and makes the trained version accessible to the public. Then the average programmer downloads the model and fine tunes it to their task.

This would be fine if there were a model for every task. Though many models exist there seems to be many tasks that are not addressed.

In this paper the GRU based Sequence-to-sequence model and the Tensorflow based Transformer model were trained from scratch on a CPU laptop. In the case of the Transformer, several days were required for training. In the GRU example the model trained in less than an hour.

3.3.2 Raspberry Pi

A Raspberry Pi is a small single board computer with an 'arm' processor. There are several versions on the market, the most recent of which sports built-in wifi and on-board graphics and sound. The memory for a Raspberry Pi 3B computer is 1Gig of RAM. Recently available, the Raspberry Pi 4B computer can sport 4Gig of RAM.

It has always been the intention that at some time some chatbot of those examined will be seen as superior and will be installed and operated on a Raspberry Pi computer. If more than one model is available then possibly several models could be installed on Pi computers.

For this to work several resources need to be made available. Pytorch needs to be compiled for the Pi. Speech Recognition (SR) and Text To Speech (TTS) need to work on the Pi.

For one of the transformer models to work Tensorflow needs to work on the Pi.

All the files that are trained in the chosen model need to be small enough in terms of their file size to fit on the Pi. Also it must be determined that the memory footprint of the running model is small enough to run on the Pi.

In the github repository files and scripts for the Raspberry Pi are to be found in the 'bot' folder.

Early tests using Google's SR and TTS services show that the Pi can support that type of functionality.

Google's SR service costs money to operate. Details for setting up Google's SR and TTS functions is beyond the scope of this document. Some info about setting this up can be found in the README file of this project's github repository.

The pytorch model that is chosen as best will be trained on the desktop computer and then the saved weights and biases will be transferred to the Raspberry Pi platform. The Pi will not need to do any training, only inference.

3.3.3 Tensorflow vs. Pytorch

Tensorflow is a Google library. Pytorch has it's roots with Facebook. Both run in a Python environment. The learning curve for Tensorflow is steeper than for Pytorch. Pytorch offers the programmer python objects that can be combined to create a neural network. Tensorflow has different pieces that can be combined, but they cannot be examined as easily at run time.

Tensorflow has a placeholder concept for inputting data and getting back results. You set up

these placeholders at design time. They are the only way of accessing your data at run time.

Pytorch objects interact with Python more naturally. You can use print statements in your code to show data streaming from one object to another. This is possible at run time.

In favor of Tensorflow, it has a good tool for visualization which can print out all kinds of graphs of your data while your model trains. It is called Tensorboard.

3.3.4 Speech and Speech To Text

Google has python packages that translate text to speech and speech to text. In the case of text to speech the library is called 'gTTS'. In the case of speech to text the library is called 'google-cloud-speech'.

The gTTS package is simple to use and can be run locally without connection to the internet. The google-cloud-speech package uses a google cloud server to take input from the microphone and return text. For this reason it requires an internet connection and an account with Google that enables Google cloud api use. Google charges the user a small amount for every word that they translate into text.

Both of these resources, the text-to-speech and speech-to-text, work out of the box on the Raspberry Pi, but configuring speech-to-text for the Pi is not trivial. The user must register a billing account with Google cloud services. In return for this registration the user is able to download a json authentication file. The file must be copied to the Raspberry Pi.

Furthermore an environment variable must be set that points to the authentication file. The variable is called 'GOOGLE_APPLICATION_CREDENTIALS'. This environment variable has to be set up before the respective model runs. When the model is launched on startup it may not be launched as a regular user. The model may be launched as, for example, the root user. Somehow the environment variable must be set along with the launching of the neural network model.

The operating system on the Raspberry Pi is based on Debian Linux. In this operating system there is a file which is run immediately after the basic system starts up. This script is called '**/etc/rc.local**'. It is sufficient to put the environment variable there and follow it with the launching of the model. To ensure that the process goes without a hitch, we attempt to combine the setting of the environment variable with the launching of the program in a single line of code.

3.3.5 Corpus Considerations

We have collected several data sets for the training of a chatbot model. Firstly we have a corpus of movie subtitles. Secondly we have a 'JSON' dump from Reddit that is downloadable. This is not the same Reddit data that the authors of GPT2 use. Finally we have the corpus described by Mazaré et al [10]. This final corpus is designed for training the chatbot task specifically. This is referred to as the Persona corpus.

At the time of this writing we are using the movie subtitles corpus and the Persona corpus. We use the movie corpus because it is smaller. Both the movie corpus and the Reddit corpus are noise filled, so it is likely that neither one is perfect for the training. The movie corpus is easier to deal with if we are training on a single processor. In the future if we can train in a faster environment the Reddit corpus might be superior.

For the Persona corpus the text is organized into 'JSON' objects. There are several different repeated labels. Some of the text is meant to be used in question and answer pairs. There is also some very specific information there that is not organized in this kind of pattern. When we take apart the Persona corpus we find that the sentences labeled with the 'history' tag are most suited to our task. We record these values only and discard other labels.

3.4 ARMv7 Build/Compile

3.4.1 Pytorch 'torch' Library 1.1.0 For ARMv7

We compile the Pytorch library for Raspberry Pi. We use several virtualization techniques to do this compilation. The result of those efforts is a Pytorch python 3.7 library for the Raspberry Pi.

On their web site Milosevic et al [14] compile Pytorch 1.1.0 for the Raspberry Pi. We follow their instructions closely. We are able to build the package for the ARMv7 platform.

The instructions called for constructing a change-root environment where a Fedora Core 30 linux system was set up. Then the ARMv7 system was used in the change-root environment to compile the Pytorch library for the 1.1.0 version.

The production laptop used for development ran Ubuntu linux. For this reason a Virtualbox emulation was set up with Fedora Core 30 on it. Inside that emulator the change-root environment was set up. The library was compiled there successfully.

There are two problems with the resulting built python package. Firstly there is an error in python when importing the torch library. The error reads 'ImportError: No module named _C'.

After some research it is clear that the build process for ARMv7 creates some shared object files that are misnamed. A fix is to find the misnamed files and make copies of them with a suitable name. The same outcome could be assured by making symbolic links to the misnamed files with proper names.

There are three files misnamed. They can be found at '**/usr/lib/python3.7/site-packages/torch/**'. They are named with the same convention. They all have the ending '.cpython-37m-arm7hf-linux-gnu.so'. We want to rename them with the much shorter '.so'. The files are then named '**_C.so**', '**_dl.so**', and '**_thnn.so**'.

This takes care of the 'ImportError'. The second problem is that the version of GLIBC in the change-root environment does not match the GLIBC library in the Raspberry Pi Raspbian distribution. This produces the following error: '**ImportError: /usr/lib/x86_64-linux-gnu/libstdc++.so.6: version 'GLIBCXX_3.4.26' not found**'.

This is solved by rebuilding the package with Fedora Core 29 instead of 30.

3.4.2 Pytorch 'torch' Library 1.4.0 For ARMv7

We recompile the Pytorch library for the Raspberry Pi. We use debian virtualization techniques for the compilation. Because Ubuntu is a Debian derivative it is not necessary to run the process in a Virtualbox container.

In addition to this, the files created by the compilation are properly named. There is no need to go to the directory '**/usr/lib/python3.7/site-packages/torch/**' to change anything.

The time spent compiling the software is approximately 5 hours. Time spent with the Virtualbox container was easily twice that. The time spent on the Raspberry Pi executing a single Generative Pre-training Transformer 2 question and answer remains about 13 seconds, so there was no gain in that respect.

There were several small hurdles to completing the compilation. Firstly the 'debootstrap' command needed to be employed at the start. Debian Stretch was used as the host operating system. It was felt that if it was used that the GLIBC compatibility problem would not be faced. This turned out to be the case.

There are some dependencies that need to be installed on the 'chroot' environment for Pytorch to compile. One of these is that is important is 'libblas3.'

Then Python 3.7 needed to be built on Stretch. The Stretch program repositories use Python 2.7 and 3.5 . The Raspbian operating system on the Raspberry Pi 4B is based on Debian Buster and uses Python 3.7. After compiling Python 3.7 the Git program needed to be compiled from scratch. Git on Stretch has a issue that is fixed upstream, but we want to use Stretch because of the GLIBC issue. Instead of using the upstream fix, we compile Git ourselves.

It is conceivable that the GLIBC issue would not be important if the 'chroot' environment used Debian Buster, since that is the basis for the current Raspbian operating system. The Stretch operating system solution works though.

Finally the Pytorch program needed to be built. We disable CUDA and distributed computing as neither exists on the Raspberry Pi.

3.4.3 Docker Container 'tensorflow-model-server' For ARMv7

The Google machine learning library for python uses a standalone program called 'tensorflow-model-server' for serving all tensorflow models in a standard way. The program has not been officially compiled for ARMv7. There exists, though, a docker image that will run on ARMv7.

Docker can be run on the Raspberry Pi in the ARM environment. Below is a terminal excerpt that shows how to do this. These commands are executed on the Pi.

```
$ sudo apt-get update
$ sudo apt-get upgrade
$ curl -fsSL test.docker.com -o get-docker.sh
$ sh get-docker.sh
$ sudo usermod -aG docker $USER
```

After the last command you need to log out and then log in again to take advantage of the newly installed docker.

The original idea was to follow someone else's instructions and compile the Docker Container for the ARMv7. Then the executable would be removed from the container and used natively in the Raspberry Pi.

It was found that there existed a version of the Docker Container Daemon that ran on the Raspberry Pi. All that remained was to write a Docker Container script that interacted with the existing ARMv7 container. The author of the original container is Erik Maciejewski [15].

'Tensorflow-model-server' is used on the localhost internet address, 127.0.0.1, with a port of 8500. tensorflow-model-server is meant for serving neural network resources on the internet, but with careful planning it works on the Raspberry Pi.

Chapter 4

Experimental Results

4.1 Experiments - Installations

In the section above we describe the workings of a transformer and the workings of Generative Pre-training Transformer 2. We propose they are similar. Here we distinguish between the two. For the experiments section they are totally separate.

We have several basic neural network models. One is the basic sequence to sequence model typically used for neural machine translation. We also have two transformers and the Generative Pre-training Transformer 2. We try to touch on each model type and we also distinguish between chatbot operation and smart-speaker operation. This gives us eight sections.

We have four models we consider and at the same time only three Raspberry Pi boards. We actually try all four models, but keep only three Raspberry Pi installations.

The model that did not make it to the final three and installation on a Raspberry Pi was the smaller Transformer with the Persona corpus.

4.1.1 Questions

Below is a list of questions asked of all models. From this list and the answers from each model we try to make comparisons between the models about their strengths.

```
Hello.
What is your name?
What time is it?
What do you do?
What is your favorite color?
Do you like red?
Do you like blue?
What is your favorite candy?
Do you like ice cream?
Good bye.
```

For comparison there are four models. Subjectively the first transformer model did not perform as well as the Generative Pre-training Transformer 2 model. It did not perform better than the Gated Recurrent Unit model either.

The model from the Gated Recurrent Unit tutorial performed well. It was better than the initial Transformer model and on par with the larger Transformer model. It was not better than the Generative Pre-training Transformer 2.

We visit and revisit this checklist to subjectively rate our chatbots.

4.1.2 Checklist

- ☐ Are all the responses are in plain English? Are any of the answers gibberish?
- ☐ Is there a variety of answers? Are all answers the same?
- ☐ Does the model give good answers to the questions about 'favorite color' and 'favorite candy'? The model could have a set of easy answers that it can use for this kind of question or it considers the question separately.
- ☐ 'No' is a safe answer for many types of question as it is clearly English, it follows logically, and it is short and easy to remember. Another safe answer is 'I don't know'. Does the model use these answer at all times?
- ☐ Does the model answer well to 'Hello' and 'Good bye'?

4.2 Chatbot - Gated Recurrent Unit Model

We have two models based on the sequnce to sequence architecture proposed by Vinyals et al [1]. One model was largely written by hand. This first model does not work very well. The second model was taken largely from an online tutorial by Inkawhich et al [2].

We trained this sequence to sequence model on a large english corpus in an attempt to produce a chatbot. This model does not use a transformer architecture.

This second sequence to sequence model performed exactly as expected. It answered a number of subjective questions with original, authentic answers. Almost all the answers made sense in the English language.

4.2.1 Questions

This is the sample question list as it was answered by the GRU model.

```
> hello .
you re a good man .
> what is your name ?
smith .
> what time is it ?
i don t know .
> what do you do ?
i m a comedian .
> what is your favorite color ?
i don t know .
> do you like red ?
no .
> do you like blue ?
no .
> what is your favorite candy ?
it s a big one .
> do you like ice cream ?
no .
> good bye .
good bye .
```

4.2.2 Checklist

- ☑ All the responses are in plain English. There is no gibberish.
- ☑ There is a variety of answers. Not all answers are the same.
- ☐ It is debatable weather or not the answers to the questions about 'favorite color' and 'favorite candy' are good. The model could have a set of easy answers that it can use for this kind of question.
- ☒ 'No' is a safe answer for many types of question as it is clearly English, it follows logically, and it is short and easy to remember. Another safe answer is 'I don't know'. This model uses that answer at times.

☑ The model answers well to 'Hello' and 'Good bye'.

4.3 Smart Speaker - Gated Recurrent Unit Model

The Gated Recurrent Unit model was installed on a Raspberry Pi. This allowed us to test out speech-to-text and text-to-speech libraries. The Raspberry Pi model was 3B. The RAM requirements were less than 500MB and the trained model answered questions on the Raspberry Pi almost instantaneously.

For this experiment we compiled the Pytorch library for Raspberry Pi.

The Raspberry Pi was outfitted with a microphone and a speaker and nothing more. It was also configured so that the Pytorch sequence to sequence model ran automatically on startup.

The model requires access to the internet for the exchange that the speech to text software has to make with the Google servers. If there is no internet the model doesn't work.

As there was no monitor and it took some time for the model to launch, the program was coded to beep when the model was ready to accept input. At that point the model beeps once.

4.4 Chatbot - Transformer Model with Persona Corpus

Using the Persona corpus we trained a transformer model to use as a chatbot. This transformer was not pre-trained with any large corpus, so this example did not use transfer learning. The Persona corpus comes from Mazaré et al [10].

This model uses the tensorflow library, not Pytorch, and a transformer model that is somewhat small.

The memory footprint of the model while it was running was below 1 Gigabyte. It is conceivable that the model could be installed on a Raspberry Pi board. That requires a python package called 'tensorflow-model-server' and this package had to be built for ARMv7 for the Raspberry Pi.

4.4.1 Training

Training of the model followed a certain pattern. First the model was trained on the persona corpus until a familiar pattern emerged. When the model began to answer all questions with the phrase "I don't know" training was stopped.

At that time the corpus was modified to include no sentences that have the word "don't" in them. Training was started again until the output contained nothing but the phrase "I'm sorry."

At that time the corpus was modified to include no sentences that have the word "sorry" in them. Training was started again and was continued for some period. A further segment of training was not attempted.

At this point, after looking at the change in loss, further training was not thought of as helpful. Loss stopped improving at some point in this process, and this lack of improvement was taken as a sign that further progress was not likely.

4.4.2 Questions

Subjectively this transformer model is not better than the original Gated Recurrent Unit model from the Sequence to sequence tutorial. It can respond to something like four sentences. When it comes upon a question that it doesn't expect it defaults to a certain sentence. It can answer questions that you might ask in a rudimentary conversation. It has answers to prompts like 'hi', 'How are you?' and 'What do you do?'. If you tell it your name it will tell you that its name is 'Sarah'. It doesn't answer arbitrary questions. It cannot answer 'What is your favorite color?'. It can not tell you the time. The default reply sentence for unknown prompts is 'Hi, how are you today?'

This is the sample question list as it was answered by the model.

```
> hello
hi , how are you today ?
> what is your name?
hi , how are you today ?
> what time is it ?
I like to read a lot
> what do you do ?
i'm a student
> what is your favorite color ?
hi , how are you today ?
> do you like red ?
hi , how are you today ?
```

```
> do you like blue ?
hi , how are you today ?
> what is your favorite candy ?
hi , how are you today ?
> do you like ice cream ?
yes , i do
> good bye
hi , how are you today ?
```

4.4.3 Checklist

- ☑ All the responses are in plain English. There is no gibberish.
- ☑ There is a variety of answers. Not all answers are the same.
- ☒ Some of the answers are re-used and do not follow logically from the questions. The 'favorite color' and 'favorite candy' questions are nearly ignored. For those questions the model answers with 'Hi, how are you today?'. This seems to be the model's default answer.
- ☑ The model does not use 'No' or 'I don't know'.
- ☒ The model does not have an answer for 'Good bye'.

4.5 Smart Speaker - Transformer Model with Persona Corpus

The transformer model was tested on the Raspberry Pi. This model is not more dynamic than the GRU tutorial model.

The transformer model takes about two minutes to boot on the Raspberry Pi. After that the time between responses is slow. The time between the first two or three responses is uncomfortably slow. After those first responses the time between answers gets to be more natural.

There is one special tone that the Raspberry Pi gives at the end of loading the model. This tone notifies the user that the model is loaded and ready to respond to questions.

4.6 Chatbot - Transformer Model with Movie Corpus

Using the Movie corpus we trained a transformer model to use as a chatbot. This transformer was not pre-trained with any large corpus, so this example did not use transfer learning.

This model uses the tensorflow library, not Pytorch, and a transformer model that is larger than the other Transformer based model that uses the Persona corpus.

In contrast the Persona corpus model uses 2 layers, 4 heads, and a hidden size of 128. The Movie corpus model uses 6 layers, 8 heads, and a hidden size of 512.

The memory footprint of the Movie corpus model while it was running was above 1.5 Gigabyte. The model could be installed on a Raspberry Pi 4B board but it requires a python package called 'tensorflow-model-server' and this package had to be built from source or somehow provided for the Raspberry Pi.

The model was trained on the development computer in a x86_64 environment. The model took about seven days to train with a CPU based processor. The goal for training was 50,000 lines from the movie corpus. After training the loss graph was consulted and the installed version was culled from the saved checkpoint at the 45,000 line point.

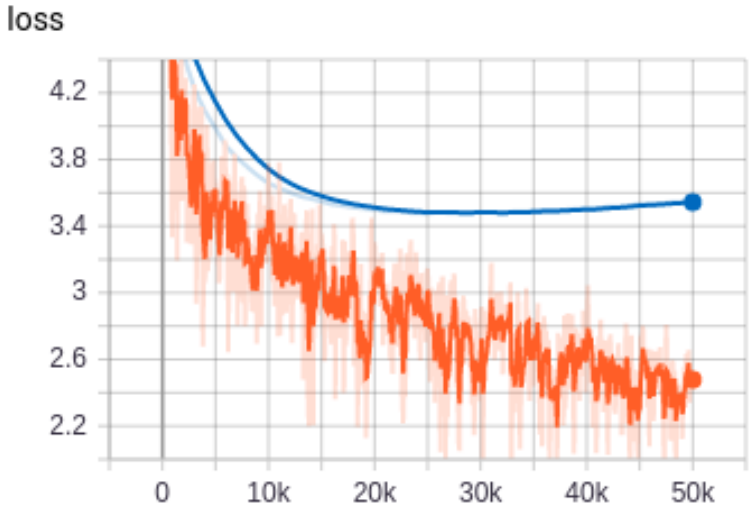

Figure 4.1: Loss - Orange is training loss and blue is evaluation loss.

Subjectively this transformer model is better than the Transformer model based on the smaller hyper-parameter set and the Persona Corpus.

4.6.1 Questions

This is the sample question list as it was answered by the model.

```
> Hello.
hello
> What is your name?
i don't know
> What time is it?
i don't know
> What do you do?
what do you mean ?
> What is your favorite color?
i don't know
> Do you like red?
no
> Do you like blue?
yeah
> What is your favorite candy?
i don't know
> Do you like ice cream?
yeah
> Good bye.
bye
```

4.6.2 Checklist

- ☑ All the responses are in plain English. There is no gibberish.
- ☑ There is a variety of answers. Not all answers are the same.
- ☒ The 'favorite color' and 'favorite candy' questions are ignored. The model does not have original answers for these questions.
- ☒ The model does in fact use 'No' or 'I don't know'.

☑ The model does have an answer for ‘Good bye’.

4.7 Smart Speaker - Transformer Model with Movie Corpus

The transformer model is installed on the Raspberry Pi. It takes about five seconds to answer any question.

The transformer model takes about two minutes to boot on the Raspberry Pi. After that the time between responses is slow.

There is a special tone that the Raspberry Pi gives at the end of loading the model. This tone notifies the user that the model is loaded and ready to respond to questions. The model is also configured to beep intermittently during operation to signal that it is processing an input. This is helpful for a configuration where there is no monitor.

A set of LED lights is installed on the Raspberry Pi to show when the model is processing input and when the model can take new input. The lights are helpful.

4.8 Chatbot - Generative Pre-training Transformer 2 Model

We used a pre-trained Generative Pre-training Transformer 2 model with the large english corpus to produce a chatbot and ascertain if this model works better than the sequence-to-sequence model. In our tests this worked well and this model was considered superior. The corpus is called ‘WebText’. We did not train this model ourselves.

For our experiments the Generative Pre-training Transformer 2 was used for the chatbot model in ‘zero-shot’ mode. This means we did no special fine-tuning of the model in the application.

We did do some special coding for the input and output code in order to operate it as a chatbot. Output was limited to about 25 tokens.

Input to the model was prepended with the character string "Q:" by our code. Output was observed to have the character string "A:" prepended to it. We assume therefore that the model was at some point exposed to the "Question/Answer" paradigm in written passages during its training. This was helpful.

Output from the model was usually larger in size than we needed. Also, output had the character of having some sensible utterance followed by some output that was only a partial sentence.

It was necessary to process the output. First the output was checked for the "A:" character string at the start. If it was there it was removed. Then the first complete sentence was used as output, while words and phrases after that were discarded.

4.8.1 Context Experiment

We decided that we would attempt to give the model some details that it could draw on during normal execution. We had two choices here. One choice was to train the model using fine-tuning and transfer learning to recognize certain questions and to supply answers. The other choice was to simply show the model the list of facts that we thought were important before every input sequence. This information would be summarized with each reply.

The second choice was more interesting. The text that the model was shown always included the name of the model (picked somewhat arbitrarily) along with information about the location of the model and the occupation. The time was also included.

This worked under most circumstances. Occasionally the model would report the wrong time when asked. This sort of activity is mysterious.

4.8.2 History Experiment

We tried concatenating all input and output and including it with each question. The hope was that the model would treat the 'Q' and 'A' as historical input and then it could refer to that input as it answered new questions. This didn't always work as expected. Sometimes the model was confused by the extra input. Sometimes it seemed to have no effect at all.

Without some mechanism of this sort the model has no notion of temporal context. The answers to questions could never possibly depend on the answers to past questions. Though this historical input did not always work, we retain the code in the hopes that at some point it is useful.

Consider the questions below.

```
> Do you like the color red?
I like the color red.
> What is your favorite color?
Red.
```

We like this output. The model is asked about the color 'red'. Doing so the model is encouraged to

pick ‘red’ as the answer to the next question about color. When ‘favorite color’ is requested, ‘red’ is the answer. Without any history the model will answer the ‘favorite color’ question with another answer. It may answer ‘pink’ or it may answer ‘the colors of the rainbow’.

From one sentence to the next the model is keeping track of the context of the conversation. History is considered. Unfortunately the results of our history experiment are not so robust. It is not always the case that the model chooses the color recently discussed as the ‘favorite’ color.

4.8.3 Artificial Intelligence Markup Language Experiment

Tests showed that the Generative Pre-training Transformer 2 chatbot worked well. We wanted to continue and allow the chatbot to have more of the abilities of a smart speaker. We constructed a simple corpus that contained key phrases that we wanted the chatbot to recognize and act upon. We did some transfer learning with this new corpus.

We found that one of two things would happen. The chatbot would either learn the new phrases and forget all it’s pre-training, or it would not learn the new phrases and it would retain all it’s pre-training. For our examples there seemed to be no middle ground. Comparisons were made with all available models and a version without the transfer learning was settled on.

It was deemed helpful if the model could be given a question and instructed exactly how to answer it. To this end AIML files were constructed and an AIML kernel was employed. The user’s question was shown to the AIML kernel and then the model was shown the kernel’s output (if there was one) along with the original question. The hope was that the output could be controlled by the AIML component.

At first it didn’t work. The AIML confused the model, and the model would not reliably choose to answer with the AIML text, as it might with the time of day.

Later the AIML was modified to appear to the model with the ‘Q:’ and ‘A:’ at the beginning of the lines. Some of the time the model answered with the AIML.

4.8.4 Program Launching

We did some programming that allowed the model to launch programs when directed to by the user. In this way we have tried to move our project closer to the smart-speakers that are produced commercially. The programming did not rely on the neural-network aspects of the model. Instead

the code used string manipulation and simple word recognition. This code can be enabled when the model is run from the command line. This was not enabled for the Raspberry Pi.

There is no space prohibition for using the Program Launching code on the Raspberry Pi. Future work may use Chromecast as a way to connect the Raspberry Pi with the user's home TV or monitor.

4.8.5 Overall

Subjectively the model was the best of those tested. The model would answer questions about it's location, it's name, and the time, faithfully most of the time. Interestingly there where times when it did not do so. Some times it used alternative answers. For example, it would answer with the time but not the correct time. This was odd.

Under almost all circumstances the output was sensible English. There were no times where the model replied with gibberish.

The subject matter of the prompts did not need to be the same as the simple introductory conversation of the transformer model. In fact any subject matter could be chosen and the model would answer. The model did not remember its own answers but it was consistent. Questions it answered include 'What is your favorite color?' and 'Do you like lollipops?'.

4.8.6 Questions

This is the sample question list as it was answered by the model. Note that the information mentioned in the answer about the time was accurate when the test was run.

```
> hello
Hello.
> what is your name ?
My name is Jane.
> what time is it ?
02:59 PM January 28, 2020.
> what do you do ?
I am a student.
> what is your favorite color ?
I love the color of the rainbow.
```

```
> do you like red ?
Yes.
> do you like blue ?
I do.
> what is your favorite candy ?
I love candy.
> do you like ice cream ?
I do.
> good bye
Good bye.
```

4.8.7 Checklist

- ☑ All the responses are in plain English. There is no gibberish.
- ☑ There is a variety of answers. Not all answers are the same.
- ☐ It is still debatable weather or not the answers to the questions about 'favorite color' and 'favorite candy' are good. The model could have a set of answers that it can use for this kind of question. The model seems to know what candy is and to a lesser extent what a color is. Some of the time the answer includes a word from the question sentence that would lead you to believe that this model has fewer stock answers. The answers are good but not perfect.
- ☑ The model does not use 'I don't know' that often.
- ☑ The model does have an answer for 'Good bye'.

The model will answer with it's name and you can tell it your name, but it is confused by this. It will on occasion tell you that it's name and your name are the same thing. This is in part because it cannot remember what it most recently said to you or what you most recently said to it.

4.9 Smart Speaker - Generative Pre-training Transformer 2 Model

Code was added that uses Text To Speech and Speech To Text libraries. In this way the model could interact with a subject using auditory cues and commands.

The Raspberry Pi model that the Generative Pre-training Transformer 2 was installed on was the 4B with 4GB of RAM. It is largely for this model that we cross compiled the Pytorch Python library for the ARMv7. The GPT2 model fit on the Raspberry Pi. While execution on the production laptop was instantaneous, execution on the Raspberry Pi took about 13 seconds for every response from the neural network.

The Raspberry Pi was outfitted with a microphone and a speaker but no mouse, monitor, or keyboard. The program was modified so that there was a tone every time the model was processing input. Without such a tone it would be difficult to know when to speak and when to wait for a response. Aesthetically this arrangement is not perfect, but it allows the Generative Pre-training Transformer 2 model to be physically installed on the Raspberry Pi.

A set of LED lights is installed on the Raspberry Pi to show when the model is processing input and when the model can take new input. The lights are helpful.

Chapter 5

Conclusions

5.1 GRU vs. Transformer

It is important here to compare the GRU chatbot with the larger Transformer based chatbot. Using our subjective qualifications we see that the GRU model answers with more variety than the transformer model. The important observation is that the hyper-parameter set for the Transformer model can be expanded and enlarged as needed before training. The GRU model cannot be trained successfully with an arbitrarily large hyper-parameter set. We can train a larger Transformer and obtain the benefit associated with this, namely better responses.

A single further observation is that the GRU model responds very quickly, while a transformer model may take more time relatively. This is not a problem for general applications, but for our purposes we cannot ignore the time spent by the transformer model when it is installed on a small computer like a Raspberry Pi.

The respective value of each of the models changes slightly when you consider what platform they will be implemented on. The GRU responds more quickly and so it retains some worth.

5.2 Smaller Chatbot Transformer Learning

For the chatbot task we want to know what the Transformer is doing during training and later during inference. Here we are concerned with the Transformer model trained on the movie dialog corpus and not the GPT2 model.

It seems the model learns a set of multi-purpose English answers in a form close to a list. Then it spends time as a classifier. Each input sentence is compared to the set of answers. A question is associated with a given answer from the set when possible. There would be fewer answers than there are questions. It is as if there were a list somewhere that contained many of the answers that the model would use. For a given model this list can only be so long.

It is interesting to point out that probably at the start the multi-purpose answers are constructed at the same time that the classification task is taking place.

For the very large pre-trained model, like the GPT2, this may not be true. These models may be more dynamic. Also we have not done tests with translation tasks. For translation the same model may be able to remember longer lists, or in place of a list of complete responses keep a list of phrases or partial responses that could be combined to create translated output.

5.3 Word Usage

We make a general assumption that any model uses only a small subset of words that it has available to it. In the example below 2000 input lines were tested from the training set and only a small percentage of the vocabulary words are used by the models in the output. Also some percentage of words are used repeatedly.

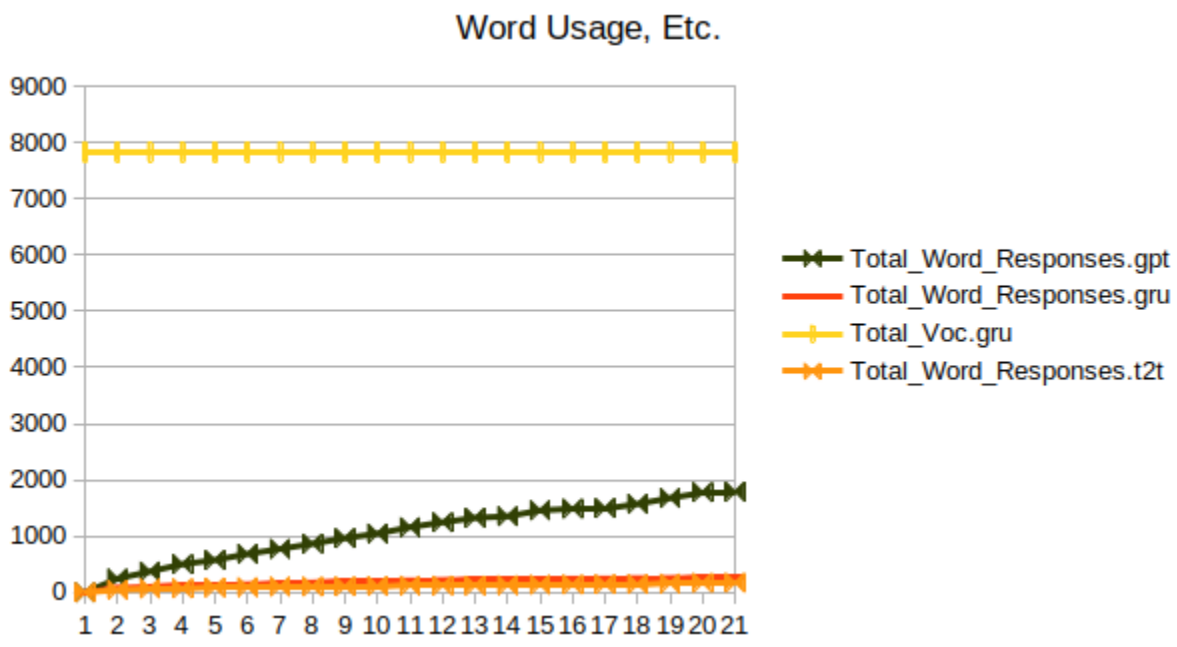

Figure 5.1: Word Usage - Including Vocabulary Total

The yellow line at the top of the graph is the total words in the GRU vocabulary. The total words in the GPT2 vocabulary is not represented on the graph. Total responses in words are shown with the three remaining lines. The GPT2 responses are higher, and the GRU and Transformer responses look like horizontal lines at the bottom of the graph.

If we remove the Vocabulary Total and the GPT2 responses from the diagram the GRU output and the Transformer output take the shape of a curve with a limit somewhere below 300 words. The Transformer model has a vocabulary size of 8170 tokens. The GRU model is close to that at 7826 tokens. Both models use the same training corpus. The difference between the tokens available and the tokens used is large.

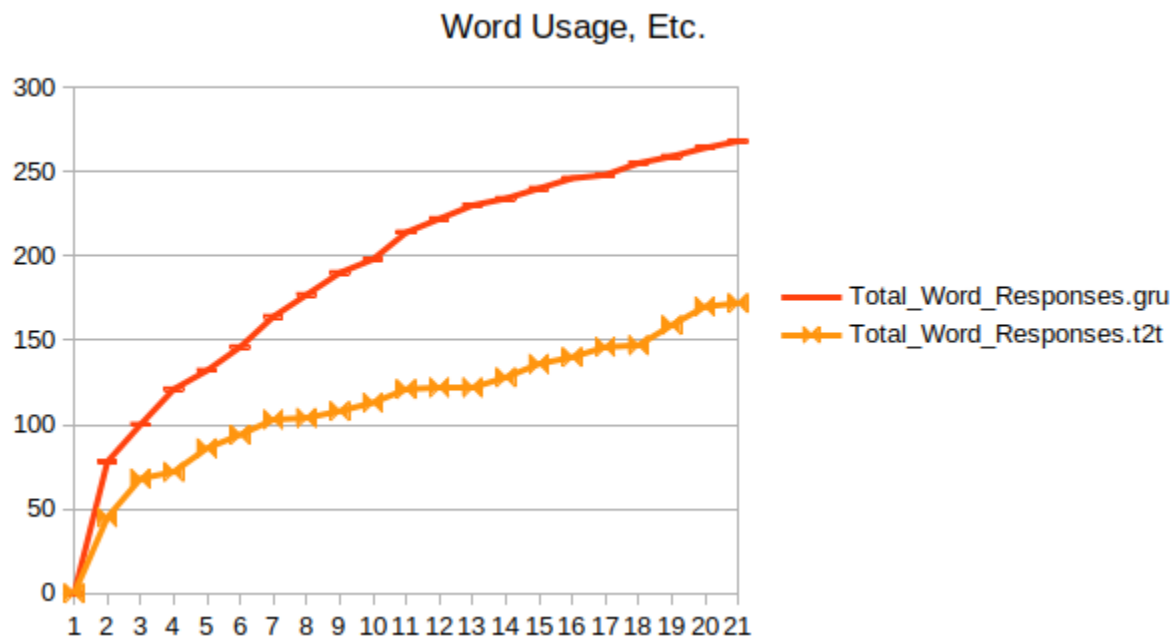

Figure 5.2: Simple Word Usage - No Vocabulary Total

We might conclude from these comparisons that the GRU model operates more robustly than the Transformer model. This may be the case, or the transformer model may be over trained or over fitted. It also might be that the hyper parameter set is poorly adjusted. We suspect the learning rate, for example, may be too high.

5.4 Sentence Usage

If we were interested in transfer learning and we wanted to train the model as a classifier, we might be interested in how many fully formed responses the given model uses. We might also be interested in when or how many of these responses were used repeatedly or over and over.

The GPT2 model is very large and very versatile, but the Transformer model and the GRU model are smaller. For these two smaller models it would be good to tell how many repeated sentences occurred as output in some number of inputs. As above we use 2000 inputs from the training set of the movie dialog corpus.

We have a rough number for those two models. They use about 125 sentences repeatedly. As for

total sentences used the Transformer uses fewer in total. The GRU is increasing still at 350 sentences at the end of our study. We assume that at some point the number of total GRU sentences that that model can produce reaches some kind of limit.

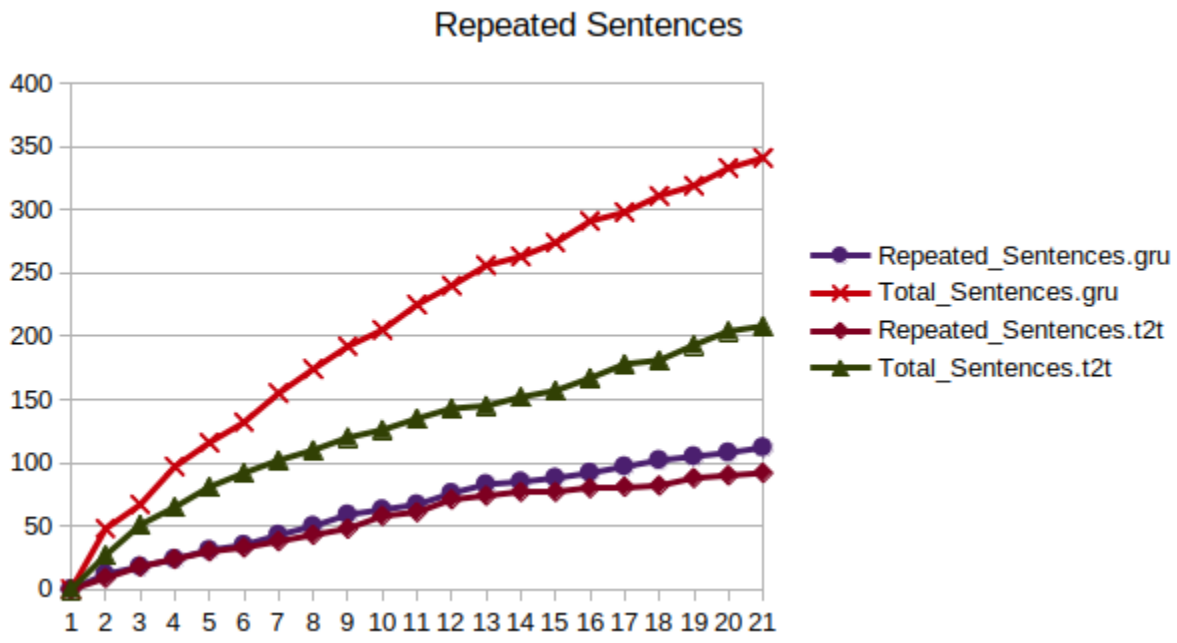

Figure 5.3: Simple Sentence Usage

In a transformer classifier we might choose 125 sentences for the maximum number of classifications. If we were going to try to create a neural network model that was trained for a specific Q/A task using a transformer this might be important. We could use 125 as our maximum number of outputs.

We do not experiment with classifiers, but we mentioned BERT previously, which can be configured as a classifier.

5.5 Turing Test

The Turing Test concerns itself with the question of weather a computer is intelligent. Turing says that intelligence is too hard to describe, and that if the computer can convince you that it is intelligent then it is.

Weather this is right is beyond the scope of this paper. The people who trained the Generative Pre-training Transformer 2 were apprehensive about their model's ability to generate human speech. They felt the model worked too well. At first when they finished their model they decided not to release the largest version to the public for several months (Radford et al) [4]. Ultimately they did release their large model.

The creators of the model used it differently than our chatbot implementation. They generated paragraphs of text, and it was determined at first that the ability of the model to impersonate a human was too great. It was felt that the model could be used to spam facebook and other social networking sites with content that was very convincing. If the model could be used to convince people to act badly, then it should not be released. Humans are susceptible to the sentiments of those they see as their peers. If the model was, for better or worse, passing the Turing test, then it should not fall into the wrong hands. This was the concern of the coders at the time.

Ultimately the large model was released, either because the developers decided the model was not as good as originally estimated, or because they didn't care.

5.6 Winograd Schema

Winograd schemas are named after Terry Winograd. The idea is that there is a sentence presented that has two meanings. A computer finds these sentences challenging to understand, and that makes them interesting for the development of Artificial Intelligence.

An example follows.

He didn't put the trophy in the suitcase because it was too [big/small]

We can choose which bracketed term to use, and we must choose only one bracketed term. If we choose 'big' then we are referring to the trophy. If we choose 'small' then we are referring to the suitcase. Human beings can easily see the pronoun 'it' refers to either the suitcase or the trophy. Computers have trouble with these determinations.

The Transformer, and the Scaled Dot-product Attention that it uses, lends itself to discussion of Winograd schema. In the chat bot example, we are less interested in the Winograd example because it doesn't come up often. However, in the case of the Generative Pre-training Transformer 2, and it's exhaustive training, it is interesting to consider the Winograd style example sentences.

There is a Winograd Schema Challenge and something of a formula for constructing your own Winograd schema (Wikapedia contributors). [17]

Appendix A

Further Installations

A.1 Generative Pre-training Transformer 2 - Large Model

The large 774M model of GPT2 was ultimately released to the public. The model is too large for a Raspberry Pi, using 6.3 Gig of ram while it is running with inference. We experiment with the model on the development laptop computer.

As with the smaller model and the Raspberry Pi installation, we do not attempt transfer learning on the model. Instead we show the model specific information during inference which the model is free to reply with if it finds the information appropriate.

We also want the model to be able to summarize articles from the internet and reply with that information when asked to do so.

A.1.1 Context Experiment

As before we gave the model details in sentence form that it could use as reply to questions when appropriate. The larger model works better if all the sentences shown to it have the 'Q:' and 'A:' strings prepended to them.

In each example the information we wanted the model to choose from was presented in sentence pairs. Though these sentences look much like source/target pairs from our movie corpus, they are totally unrelated.

An example might be the presentation to the model of the model's name. In the smaller GPT2

model we simply show the model the single sentence 'My name is Jane'. In the larger GPT2 model's code we must show the model two sentences. One is 'Q: What is your name?' and the other one, following it directly, is 'A: My name is Jane'. We follow these conventions for the time of day, the model's location, and the model's occupation.

These sorts of questions can be answered without any memory of what the last question was.

A.1.2 History Experiment

We tried concatenating all input and output and including it with each question. While this did little or nothing during the smaller GPT2 experiment, with the larger model the history sentences were helpful. Consider the questions below.

```
> Do you like the color red?
I like the color red.
> What is your favorite color?
Red.
```

This is the same text excerpt as from the small GPT2 model. The 'favorite color' example continues to work on the larger GPT2 model. Again as in the small model, the chatbot is keeping track of the conversation.

Even though the history is consulted before answers are formulated, the output of the model is not always predictable. This has something to do with the 'temperature' setting of the model. The output is sometimes more fanciful.

A.1.3 Artificial Intelligence Markup Language Experiment

We want to be able to tell the model exactly how to answer a given question if we have a particular need. For example, the model handles remembering the user's name very badly. We would like to dictate how the model handles that situation. We have employed Artificial Intelligence Markup Language for this.

AIML is actually poorly suited for handling this task. In AIML you must code every question and response as a separate rule. If you have a question 'How are you?' and an answer 'I am fine' the AIML kernel will only answer with the programmed response if the input is an exact match. If

you enter 'How are you doing?' or 'How do you feel?' the kernel does not have an answer for that. The two or three questions are phrased differently but can be answered with the same text.

The GPT2 chatbot is much different. Groups of questions have the same answer. Sometimes the answer could even be 'I don't know', but still the answer makes sense in English. The chatbot could answer 'How are you?', 'How are you doing?' and 'How do you feel' with one answer like 'I don't know' or 'I am fine'. These questions are phrased differently with the same answer.

For us to use AIML for our task we code questions and answers in pairs for all possible phrasing of a question.

We have to count the number of AIML rules that we are coding and compare it to the number of English sentences that the GPT2 chatbot can answer. We have to make sure that the number of chatbot answers is much higher. As it stands the GPT2 chatbot answers dwarf the AIML rules.

We prefer the answers provided by the generative chatbot over those provided by an AIML file. We want to maximize the amount of time that the overall model answers using the transformer that is internal to the chatbot, and minimize the use of AIML and the AIML kernel. We set a criteria for ourselves. If the generative chatbot is handling most of the questions and answers then we can proceed. If the AIML question/answering set is larger than the number of questions that the chatbot can answer, then the AIML should be avoided.

Below we illustrate two situations where AIML might be useful. In both cases the AIML files are small.

A.1.4 User Name Experiment

The name of the user is our first example of AIML coding. This example is easy. There are few ways to say 'My name is *'. In this case '*' is a wildcard. Another variant is '* is my name'. That's as many ways to say it as there are.

Each question is answerable with a phrase like 'Hello *'. Here the asterisk is a wildcard for the memorized name. An AIML file can handle this sort of question very easily. We write an AIML file with these questions. We depend on the AIML to remember the name from the wildcard.

Another question in this scenario is 'What is my name?' With this question also there are very few ways to phrase the AIML. Here we are referring to the name recorded in the previous AIML question as discussed above.

During a typical chatbot run we show all user input to the AIML kernel.

When the kernel does not match anything there is no output from the kernel, but when there is a match we show the AIML output to the chatbot as the context for that input. This is to say that the input from the user is concatenated with the context text, in this case the answer from the AIML kernel. The chatbot is then left to decide what to answer from the user question and the context. Usually the answer makes sense.

Usually the user's question and whatever the chatbot decides will be the answer are both added to the history. As noted before, history of what the user enters and the chatbot's answer are kept by the chatbot so that the chatbot can keep track of answers over time. Sometimes, though, keeping sentences that refer to user names confuses the bot's answers. In other cases the bot keeps track of answers that are generated in the AIML/chatbot hybrid process.

Even when the answer makes sense there is only a certain percentage of time that this example works. Sometimes, even with the coaching of the AIML, the model does not answer with the right name. Sometimes you ask it for your name and it responds with it's name. Some times you ask it for it's name and it responds with yours. Without the AIML the model always answers all name questions with it's name.

A.1.5 Internet Search Experiment

We want to be able to search the internet for an article, download that article, and summarize parts of the article as answers to questions asked of the model. This is leveraging of the token size of the input to the model. It's usually very large. It can be 1280 tokens.

The AIML required is minimal but larger than the AIML for the 'user name' problem. The goal is to detect when the user wants to talk about an internet page and detect when they are done.

The first case, when the user wants to talk about an internet page, is associated with utterances like 'Tell me about *' and 'I want to talk about *' where '*' is a wildcard for what the user is interested in. There can be other phrases that imply the same intention on the part of the user. In these cases the text from the wildcard in the invocation is saved.

The second case, when the user is done with the internet page and wants to return to normal operation, is associated with utterances like 'Ok thanks' and 'That's enough'. These two phrases can be checked for when the bot is in 'Internet-Search' mode. There could easily be other phrases that cause the same operation in the chatbot code.

When the AIML detects one of these signals the program goes into a special mode of operation.

The first signal, we will call the 'find' signal, causes operation of an internet browser. The web search uses the text from the invocation of the 'find' command. The browser yields results as they would be found using a google search engine. The first 20 results are kept and the url for the highest ranking wikipedia page is retained. That page is loaded and the 'body' tag is scanned for 'paragraph' type content. This web page replaces the context input.

After that the context is retained while the chatbot answers any questions that the user might have. The GPT2 model has some facility with this, and the larger GPT2 model is more adept.

Later when the AIML detects the second signal, we will call the 'restore' signal, the program switches back from special operation to normal operation. The web page contents are removed and the context is restored as it was during regular chatbot use. There is a part that has the history of the last few questions and answers, and there is a part that has the bot name, occupation, and the current time.

As the chatbot answers questions the content of the conversation is saved for the 'history' of the context text, as it is used in the regular chatbot mode. This means there is a record of the questions and answers the chatbot goes through during it's operation, even when it is answering questions about a web page that it has searched for. Because of size limitations the history is not included in the chatbot input when it is answering questions about a web page, but a record is kept of 'q' and 'a' and the bot can refer to them later.

Again, because of size restrictions there is no regular context information shown to the chatbot while it is answering questions about a web page. Then, after the 'restore' signal, the chatbot sees only the regular context and none of the web page material. The regular input for the small model is 768 tokens and it is 1280 for the larger model. Though this would appear to be a large space for text, it is common for wiki articles to be clipped so that the user's question can be inserted at the end.

A.1.6 Usage Example

Consider that you are interested in the musical group 'The Beatles'. If you wanted to find out about the band you might say to the chatbot 'Tell me about the beatles.'

The chatbot would take a little time to respond. Ultimately it would answer with something like 'The Beatles were an English rock band formed in Liverpool in 1960.'

Then you might ask it a question like 'How many members were in the band?'

To this question the chatbot answered this way for us, 'The Beatles were originally formed by John Lennon and Paul McCartney in Liverpool.' We note that this is not actually the correct answer. This may be because after clipping the information asked for does not exist in the context section. You can ask other questions. Some of them may be answered correctly and some not.

When you ask the question 'How many members were in the band?' you do not need to specify which band. The model seems to be able to intuit the band that you are referring to.

At some point you will want to end the Q/A session. You would say 'OK' or 'OK thanks'. After that the web page for the 'Beatles' would no longer be available to the chatbot. It could still answer questions. You could ask 'What are the Beatles'. You might get an answer like above, 'The Beatles were an English rock band formed in Liverpool in 1960.'

This is an example of using the model. Other topics can be asked about. You do not need to re-start the model to ask about a second or third topic. Each question takes several seconds before an answer is produced. The large version of this model does not work on the Raspberry Pi.

Appendix B

Abbreviations

AIML Artificial Intelligence Markup Language

BERT Bidirectional Encoder Representations from Transformers

BPE Binary Pair Encoding

GPT2 Generative Pretraining Transformer 2

GPU Graphical Processing Unit

GRU Gated Recurrant Unit

gTTS Google Text To Speech

LSTM Long Short Term Memory

NLP Natural Language Processing

OOV Out Of Vocabulary

RNN Recurrent Neural Network

SR Speech Recognition

Bibliography

[1] O. Vinyals and Q. V. Le, "A neural conversational model," *CoRR*, vol. abs/1506.05869, 2015.

[2] Matthew Inkawhich, "pytorch-chatbot," 2018. https://github.com/pytorch/tutorials /blob/-master/beginner_source/chatbot_tutorial.py.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[4] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[5] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Huggingface's transformers: State-of-the-art natural language processing," *ArXiv*, vol. abs/1910.03771, 2019.

[6] Denny Britz, "Recurrent neural network tutorial, part 4 – implementing a gru/lstm rnn with python and theano," 2015. http://www.wildml.com/2015/10/recurrent-neural-network-tutorial-part-4-implementing-a-grulstm-rnn-with-python-and-theano/.

[7] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *CoRR*, vol. abs/1409.3215, 2014.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[9] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*

(EMNLP), (Doha, Qatar), pp. 1532–1543, Association for Computational Linguistics, Oct. 2014.

[10] P. Mazaré, S. Humeau, M. Raison, and A. Bordes, "Training millions of personalized dialogue agents," *CoRR*, vol. abs/1809.01984, 2018.

[11] M. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *CoRR*, vol. abs/1508.04025, 2015.

[12] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018.

[13] NVIDIA Applied Deep Learning Research, "Megatronlm: Training billion+ parameter language models using gpu model parallelism," 2019. https://nv-adlr.github.io/MegatronLM.

[14] Nemanja Milosevic, "Compling arm stuff without an arm board / build pytorch for the raspberry pi," 2019. https://nmilosev.svbtle.com/compling-arm-stuff-without-an-arm-board-build-pytorch-for-the-raspberry-pi.

[15] Erik Maciejewski, "Tensorflow serving arm - a project for cross-compiling tensorflow serving targeting popular arm cores," 2020. https://github.com/emacski/tensorflow-serving-arm.

[16] C. Danescu-Niculescu-Mizil and L. Lee, "Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs.," in *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*, 2011.

[17] Wikipedia contributors, "Winograd schema challenge — Wikipedia, the free encyclopedia." `https://en.wikipedia.org/w/index.php?title=Winograd_Schema_Challenge&oldid=938834078`, 2020. [Online; accessed 15-February-2020].