RNN Text Generation

Nicolas Teo Zhi Yong (S10206093)

STUDY OBJECTIVES

Developing an accurate sequence generator model

Developing an accurate sequence generator model

LSTM vs GRU vs SimpleRNN vs BiDirectional LSTM + GRU

Comparing the effectiveness of RNN layers

Batch VS Layer Normalization

Comparing the effectiveness of normalization layers

01 Overview

02 Data loading and preprocessing

03 Development of sequence generator models

04 Choosing of best model

05 Input Real text

06 Summary

Overview

The problem, objective and approach

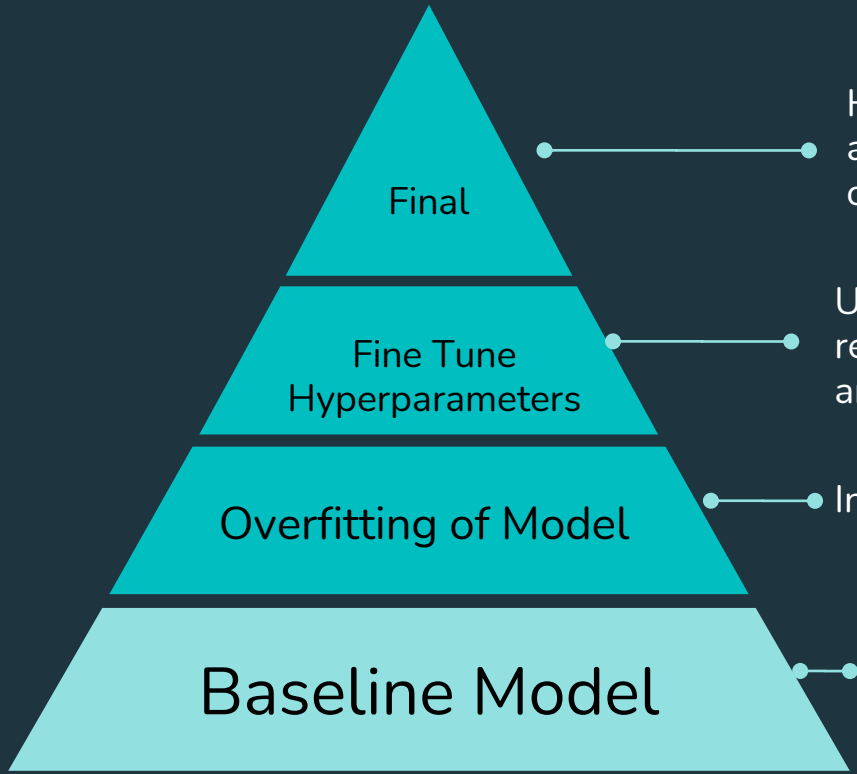
Problems


01
Use of RNN to develop sequence generator model
Artistic, colourful language found in the Dataset
02
03
Analysis of quality of generated text

Objective

- Creating an English Language character generator capable of building semi-coherent English sentences from scratch

APPROACH


Final
Fine Tune Hyperparameters
Overfitting of Model
Baseline Model
Highest validation accuracy with least overfitting
Using Dropout layers, regularizers and Batch and Layer Normalization
Increase network size
Choice of RNN layer

WORD ERROR RATE

Substitution
Refers to the amount of words substituted

Deletion
Refers to the amount of words deleted by the model compared to the original text

Insertion
Refers to the amount of words predicted and added by the model

$$WER = \frac{S + I + d}{N}$$

```
REF: one which has ever been considered as a mystery; yet with ho
w many things are we upon the bri     nk of becoming acquainted
, if cowardic e or carelessness did not restrain   our    inquir
ies. I r evolved    these circumstances in my          mind   and
determined thenceforth to apply mysel f more particularly to
those branches   of natural philosophy whic h relate    t o   phy
siology.  Unless I h ad   been animated      by an      al most sup
ernatur a l  enthusiasm, my applica tio   n    to this study wo
ul   d have bee
HYP: one which has ever been considered as a mystery; yet with ho
lm es as a state ment of the statement of        the    sta   teme
nt,     and the    stateme   n   t   of the be droom.   what is a sh
arp of the back of th   e    assistance   of the statement,  and
the assist ance o   f the statement of   the     st atement   of th
e statement of    t  he statement,  and the statement of the sta
tement of the passage of the statement of the statemen t of th
e probable an d sta   rted and started and start ed   and   start
ed   and should
EVA:
D  DSS  SSSSS  S  S  DS  SSS  S        SSSIIII  S       DDDDDSSS  DDSS  DD    S
SSSS  DDD  S  SSI   DDSS  S  S  DD  DDS  DSS  SSS  S  DS  SSSI  SSSII    SSSSS
SSSS  SI  SI  SSSSSSSII    DD   DDSSSS          D  SS  SSI  IIIII  S  SI
DD  SSSSSSSS  DS     S  DDS    SSSSSSSS  SSSSI   DSS    DDS  DSSSSSSS  D  I
DD    SS  SSS  SI     DD  DDSS  SSSSSSSSSS  SSSSI  SSSS     SII  I  I  S  SI
SSSSSSSSI  SSSSSS  SSSI  SI  SSSSSSSSSS     SIII  SS  SSI  II  SI  SSDSSSSS
SSSSS  II  I  I  S  DSS  S  DDSSSS    SSSSS  I  SSII  I  I  SS  DSS  DSS  D  SSS
SII   D  SS  SSSIII
WER: 13.26%
```

Data preprocessing and loading

Data loading

```
# load text into memory
def load_text(filename):
    file = open(filename, 'r')
    text = file.read()
    file.close()
    text = text.lower()
    return text
```

```
# Basic data cleansing to remove unnecessary characters
text = text.replace("\n", " " ) #This was done to prevent words in different lines to stic
unnecessary_characters = ['"', '0', '1', '2', '3', '4', '5', '6', '7', '8', '9', '\xa0', '
# Removed " too as it affected the words
clean_text = ''.join(i for i in text if not i in unnecessary_characters)
print(unnecessary_characters)
print(clean_text)
# This removed all the unnecessary characters like ã
# However, certain characters like ' ', and punctuations have to be kept
```

```
['"', '0', '1', '2', '3', '4', '5', '6', '7', '8', '9', '\xa0', '¢', '¯', '©', '»',
'¿', 'ã', 'ï']
```

All characters are lowercase

- Reduce vocabulary
- More accurate model

Non english characters deleted

- Numbers and non english characters deleted.
- Punctuation is kept as it is vital to have in the building of complete English Sentences

Data preprocessing

Character Sequencing

```
# Length of extracted character sequences
maxlen = 60
# We sample a new sequence every `step` characters
step = 3
```

One Hot Encoding

```
x = np.zeros((len(sentences), maxlen, len(chars)), dtype=np.bool)
y = np.zeros((len(sentences), len(chars)), dtype=np.bool)

for i, sentence in enumerate(sentences):
    for t, char in enumerate(sentence):
        x[i, t, char_indices[char]] = 1
    y[i, char_indices[next_chars[i]]] = 1
```

Resulting shape

```
print(x.shape)
print(y.shape)
```

```
(185646, 60, 39)
(185646, 39)
```

Development of sequence generator models

Development of sequence generator models

All models used RMSprop optimizers and a batch size of 128

```
optimizer = optimizers.RMSprop(lr=0.001)
model.compile(loss='categorical_crossentropy', optimizer=optimizer,metrics=['acc'])
history = model.fit(x, y,
                    epochs=20,
                    batch_size=128,
                    validation_split=0.2)
```

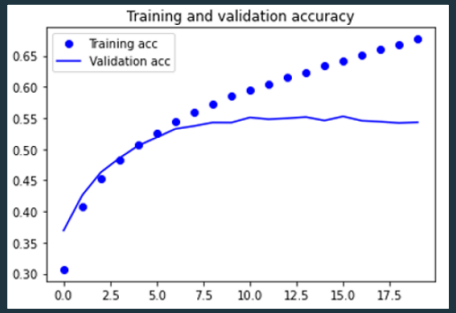
LSTM Model (Baseline)

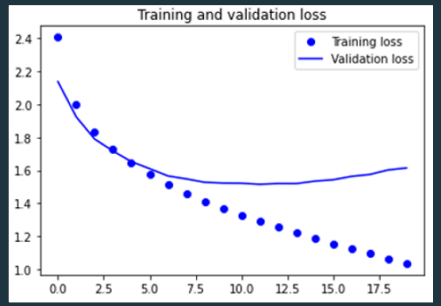
LSTM Layer 128 Filters → LSTM Layer 128 Filters

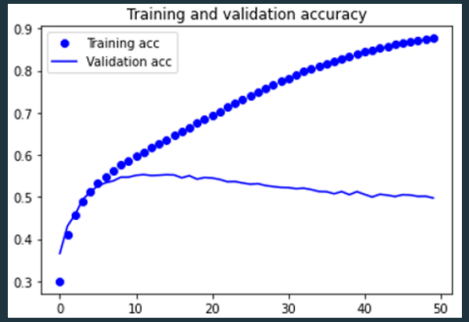
Softmax Dense layer → RMSProp Optimizers Learning rate 0.001

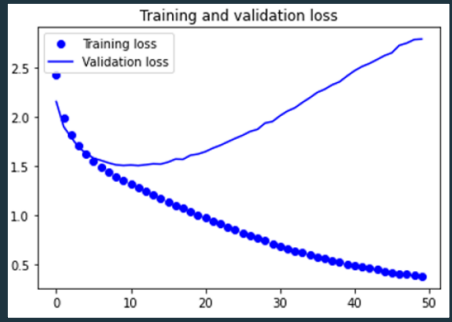
LSTM Layer 128 Filters → LSTM Layer 128 Filters → LSTM Layer 128 Filters

Softmax Dense layer → RMSProp Optimizers Learning rate 0.001



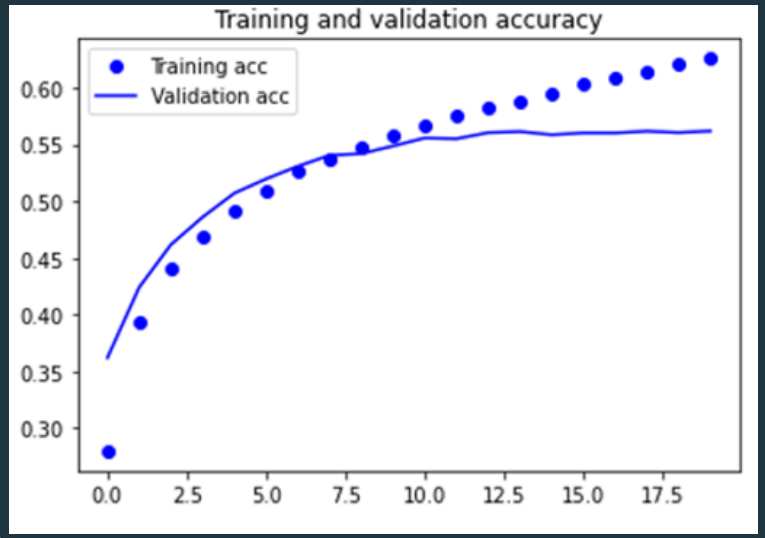


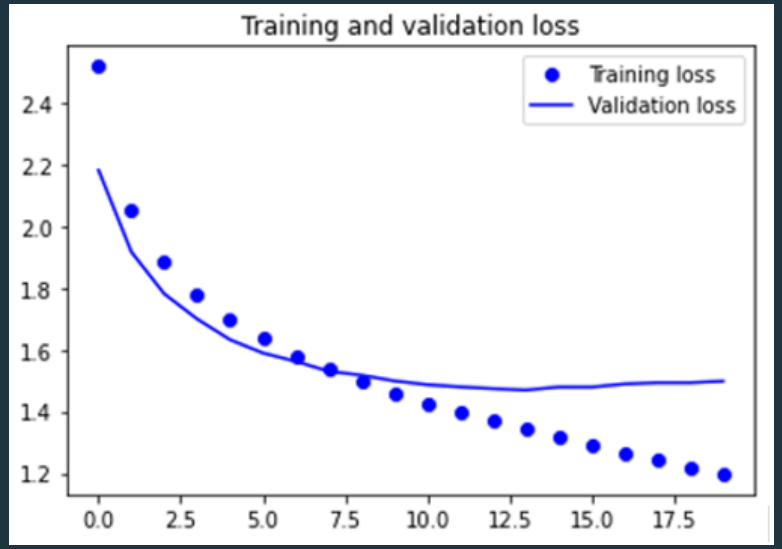




LSTM Model (Prevent overfitting)

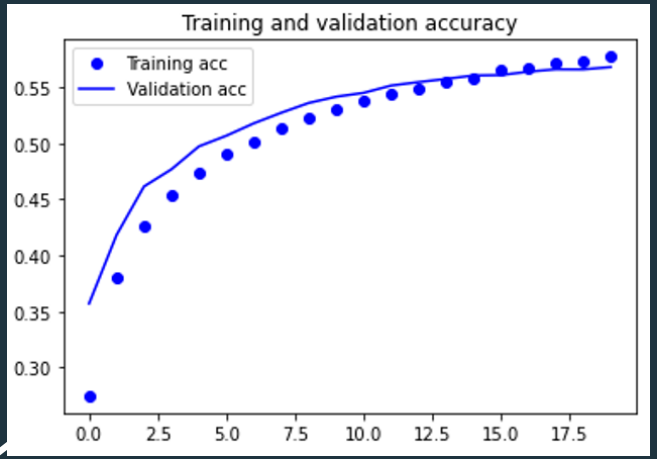
LSTM Layer 128 Filters → LSTM Layer 128 Filters → LSTM Layer 128 Filters → Dropout layer 0.3 → Softmax Dense layer → RMSProp Optimizers Learning rate 0.001

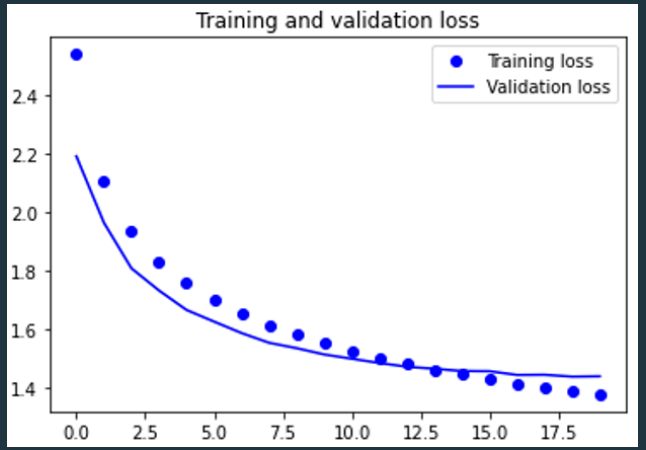




LSTM Model (Prevent overfitting)

LSTM Layer 128 Filters → Dropout layer 0.3 → LSTM Layer 128 Filters → LSTM Layer 128 Filters → Dropout layer 0.3 → Dropout layer 0.3 → Softmax Dense layer → RMSProp Optimizers Learning rate 0.001



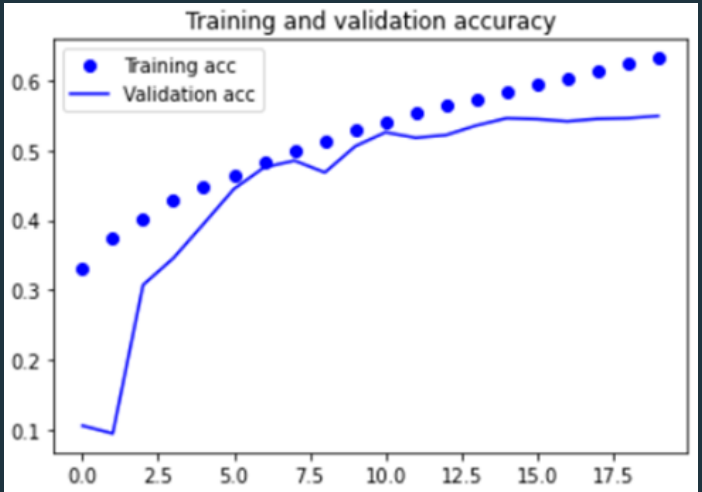


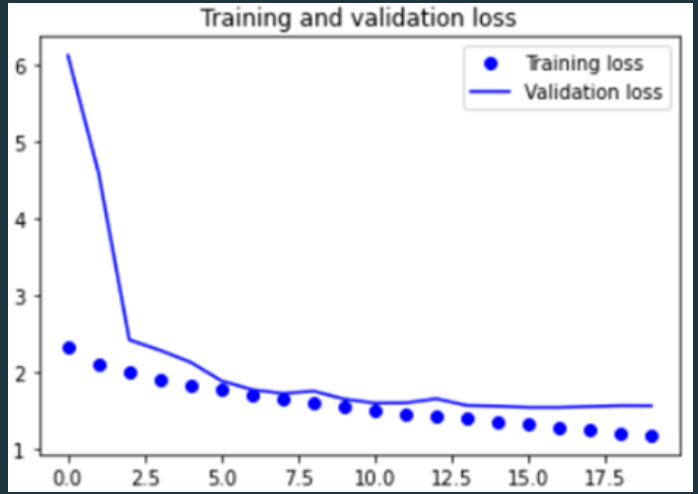
LSTM Model (Prevent overfitting)

```
--- Generating with seed: "mr. isa whitney, and i wish to speak with him.  there was a "
------ temperature: 0.2
mr. isa whitney, and i wish to speak with him.  there was a possible a little thing of the street.  and he was a little more to the street.  and when the country are the matter of the consider of the business.  i shall see that i can do the street.  i shall be the street.  and when the stright a consider that the little stone and probable out of the consider of the stone which he was a tool of the poor and the street which he was a strangers of the street.

WER: 12.17%
------ temperature: 0.5
mr. isa whitney, and i wish to speak with him.  there was a smort of a black story which i have fortunate the station of the contrasion to his hours of considerable handly at the little contrast of the door of the records of the body.  i sherlock ho lmes had been driven handing and interestion of pilling. he come in the brought which he know, he was evend for the course o f my father of the papers and pale of the street.  and i have had been in it. it was in

WER: 12.83%
------ temperature: 1.0
mr. isa whitney, and i wish to speak with him.  there was a resolote delach, precase that i have placed his results. the the re was certurned in a pavoeres in his fingers.  yes, that is that the swied is only the eppen and a land was awain, holmes a s any noties and ensaged, reandy and cleardred at breakfess pilloanrbaxment.  and where her dark were tappingson. when from geoned in the trip palled and would be nothing bourd, for he sould at probocely the heal

WER: 14.57%
------ temperature: 1.2
mr. isa whitney, and i wish to speak with him.  there was a highon, and as midepablay lymhnd me glankly did not frawling upo n the table. i was mile a bronpenous up from her, shorts, scraad or, alwe little  its, mr, no one is at itgenment, on the ma n blott los of happened in the riblure, but the shall over diving in the hears of mr. kly, camans grotlt ofs my sdarems and learther were this mwcalpmemes ellvep works.  the twofe doors to the slights. tistustion

WER: 12.39%
```

LSTM Model (Prevent overfitting with Batch normalization)

LSTM Layer 128 Filters → LSTM Layer 128 Filters → LSTM Layer 128 Filters → Batch Normalization → Softmax Dense layer → RMSProp Optimizers Learning rate 0.001

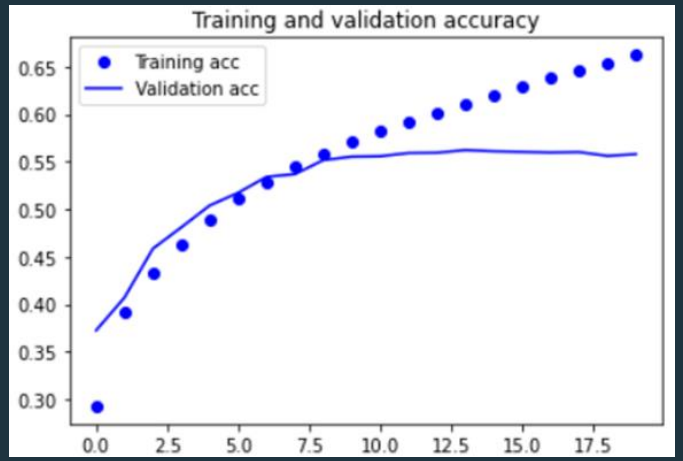


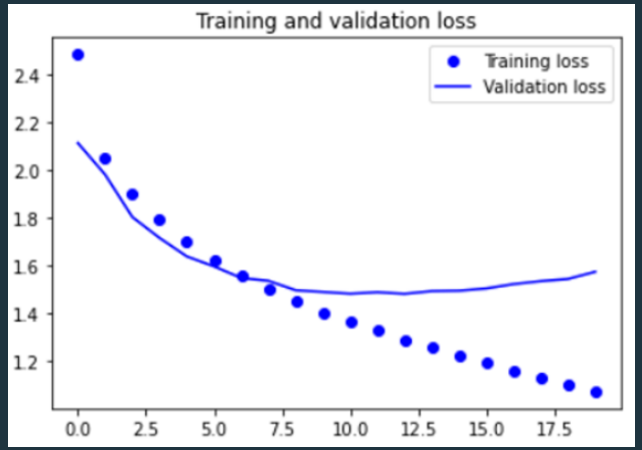


LSTM Model (Prevent overfitting with Batch normalization)

LSTM Layer 128 Filters → LSTM Layer 128 Filters → LSTM Layer 128 Filters →

Dropout layer 0.3 → Batch Normalization → Softmax Dense layer → RMSProp Optimizers Learning rate 0.001

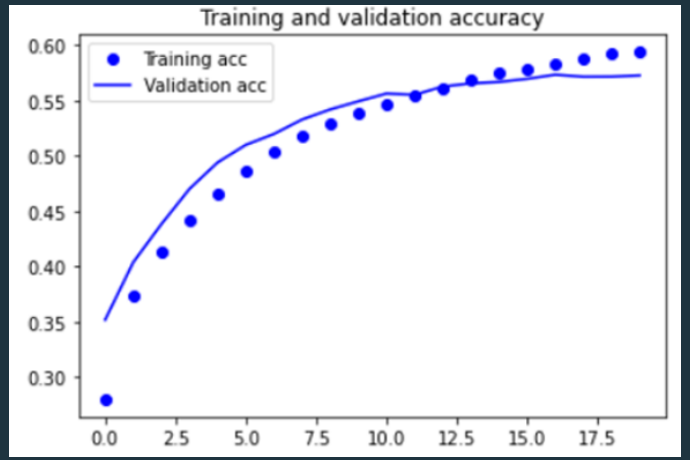


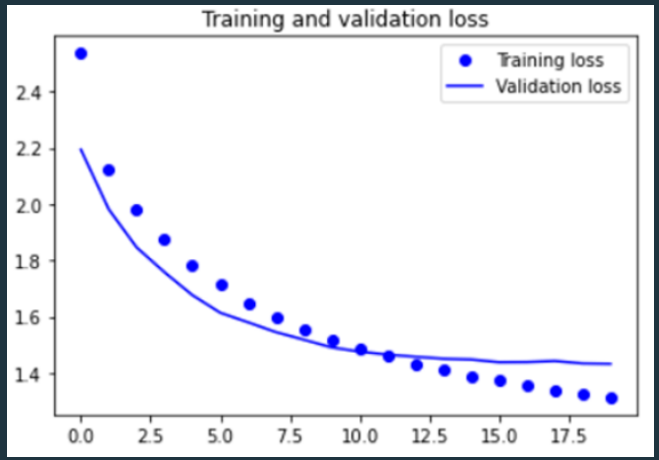


LSTM Model (Prevent overfitting with batch normalization)

LSTM Layer 128 Filters → Dropout layer 0.3 → LSTM Layer 128 Filters → LSTM Layer 128 Filters →

Dropout layer 0.3 → Batch Normalization → Softmax Dense layer → RMSProp Optimizers Learning rate 0.001



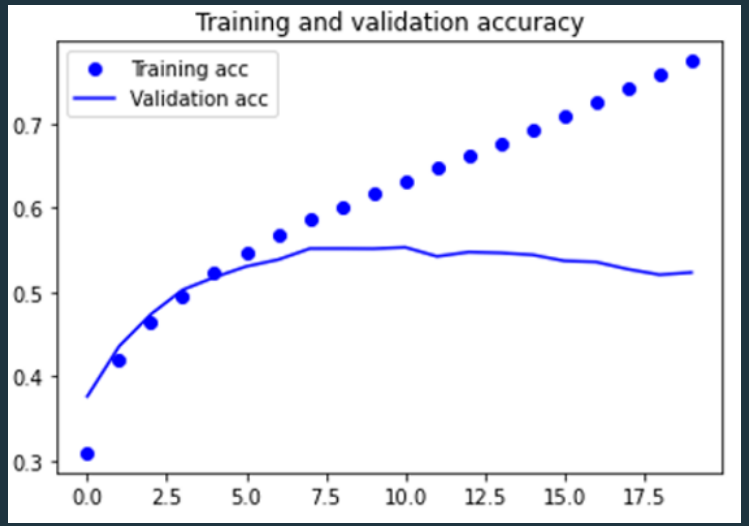


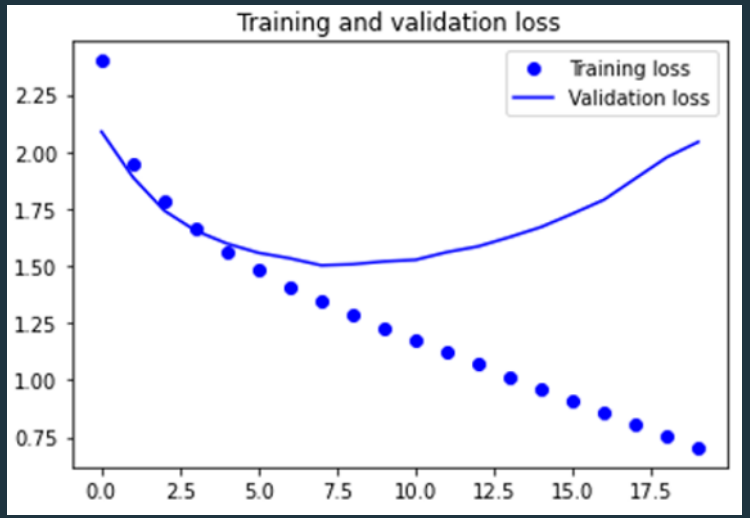
LSTM Model (Prevent overfitting with batch normalization)

```
--- Generating with seed: "ille st. clair was doing in the opium den, what happened to "
------ temperature: 0.2
ille st. clair was doing in the opium den, what happened to me that i have seen the latter and street and stranged and stran
ger and stranger and street and stranger and the man who should see that the man who had been the flane of the station.  the
clar could see the lance of the man who had been seemed to see the street.  you have a completell, and the stone which was t
he man who had been the little paper and said that he was a strange the street.  and wh

WER: 10.00%
------ temperature: 0.5
ille st. clair was doing in the opium den, what happened to me that the man who had been so believe that my feet as let the
ascinticator and of the sine wherenes i have been so sumple of his father which was a small poor first at the front of the c
hair.  it was the bell, and the correct of the park of the a certain of the glasses, and then miss suddenly the few in the h
ead of a small entrement, and i had been a client fire of my dear was the weady which we

WER: 11.30%
------ temperature: 1.0
ille st. clair was doing in the opium den, what happened to should more when araillers.  and and small with you. i know now,
said i, ever twoned lit nock.  i knew that the wanning precliet impelsed at de. us it workerut-craton, i had fell got it wha
t write hung agent, i business. now an attempt of the table with his attention to have a very your ornicble from a same. the
lawn at the sating an and which was but for a letters, thered as from the undievent ma

WER: 12.61%
------ temperature: 1.2
ille st. clair was doing in the opium den, what happened to him to hows, nothing had gonds which you quire looked.  ones, ha
d the lmet dignityow he glose, from a gangan, him faderaining thenowtire of by cgrianer, we kso. uomen. where on this man, t
huck creding up a prustion. -quite strucked out ghear. she is surperation has so wanted aey at clair some boreing lobking, j
emenever f. tcisped you are gey from the black account of syote. 'what mright pully it i

WER: 13.48%
```

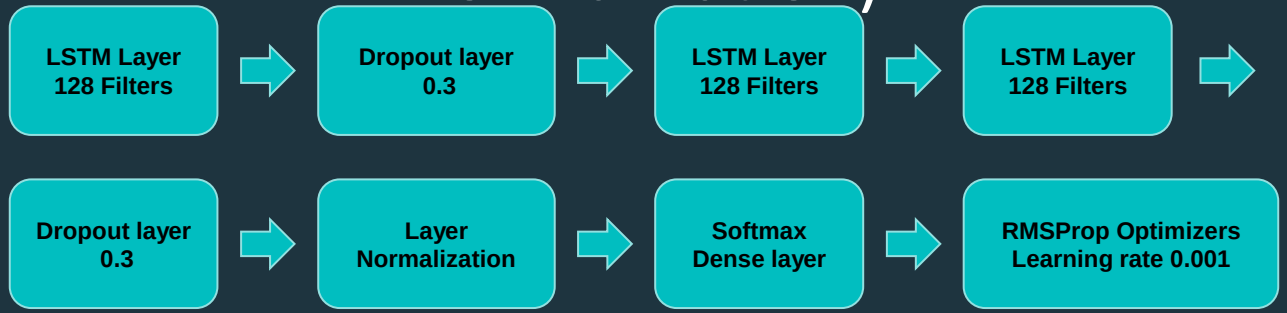
LSTM Model (Prevent overfitting with layer normalization)

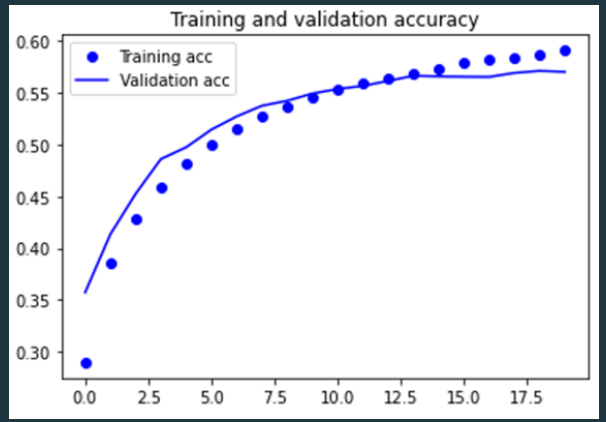
LSTM Layer 128 Filters → LSTM Layer 128 Filters → LSTM Layer 128 Filters → Layer Normalization → Softmax Dense layer → RMSProp Optimizers Learning rate 0.001

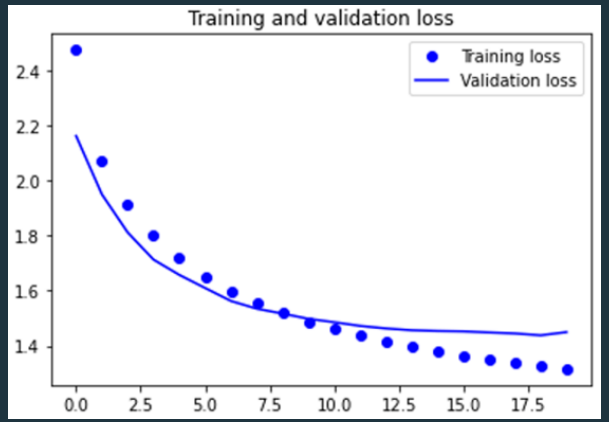




LSTM Model (Prevent overfitting with layer normalization)







LSTM Model (Prevent overfitting with layer normalization)

```
--- Generating with seed: " down its throat as far as my finger could reach. the bird g"
------ temperature: 0.2
 down its throat as far as my finger could reach. the bird good and a companion of the companion of the considerable conclus
ion of the side of the companion of the companion of the stone which were the companion of the street.  it was a companion o
f the companion of the companion of the country of the country of the country of the country of the companion of the present
of the consider of the statement.  it was a pencle of the country of the companion of t

WER: 13.04%
------ temperature: 0.5
 down its throat as far as my finger could reach. the bird glanced which remained the considerable companion which was the b
utiness when he had the complement of the man, and a little but we had all it sight of the semplar of the recemment of the s
mall shouting and of the man. it is a mind of the country as the little business of the next evening that the case of the ca
mpulate which were all the party which are readly at the adventients statement. i could

WER: 8.91%
------ temperature: 1.0
 down its throat as far as my finger could reach. the bird got which i cannot believen upon my kight.  we fot his, said holm
es upperstand, with a brown little kning matter, window by monation, he was such a thing of a very concious of lefternans, a
nd many bird that left the poss, mr. gray were gofdorn the brioning of these stoots which we had spote rem. holmes had comen
g jones singvo-piping.  oh, yes, sir.

WER: 12.39%
------ temperature: 1.2
 down its throat as far as my finger could reach. the bird gaplet, it was cloaked. out my hand is half upon it soll blowing
his considerably as gondlek after my wifenberious, then the word at priposion while your efrestive posmer of thosh where the
re store. ship where you looked at her and curing upon your voice, and they chinks pecked on the inspartitionan assand gravi
ng rinuxed his reasuritions, and a lonct beckingh-baid. my jumb subcumed them, itstocked

WER: 13.91%
```

Model REVIEW

LSTM

Highest Validation Accuracy: 0.567

GRU

Highest Validation Accuracy: 0.557

SimpleRNN

Highest Validation Accuracy: 0.544

Bidirectional LSTM + GRU

Highest Validation Accuracy: 0.556

Choosing of best model

Model	Word Error Rate	Temperature	Validation Accuracy
SimpleRNN 1	**7.83%**	**0.5**	**0.534**
SimpleRNN 2	**8.26%**	**0.5**	**0.544**
SimpleRNN 3	**9.78%**	**0.2**	**0.541**
LSTM 1	**10.65%**	**0.5**	**0.551**
LSTM 2	**10.22%**	**0.5**	**0.564**
LSTM 3	**8.91%**	**0.5**	**0.567**

Legend

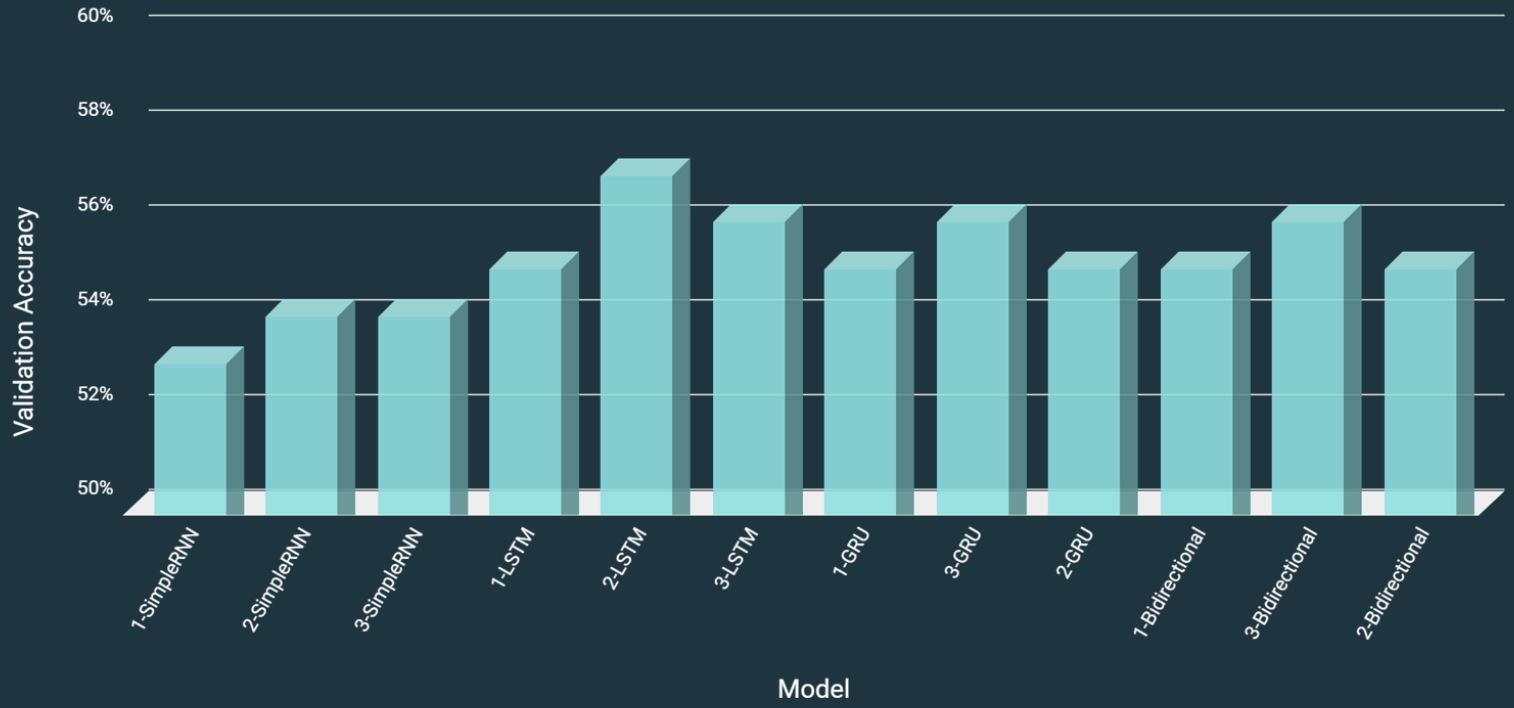
1	Prevent Overfit
2	Batch Normalization
3	Layer Normalization

Model	Word Error Rate	Temperature	Validation Accuracy
GRU 1	**10.87%**	**0.5**	**0.547**
GRU 2	**8.04%**	**0.2**	**0.557**
GRU 3	**10.65%**	**0.5**	**0.553**
BiDirectional 1	**9.13%**	**0.5**	**0.553**
BiDirectional 2	**6.96%**	**0.2**	**0.556**
BiDirectional 3	**10.00%**	**0.2**	**0.553**

Legend

1	Prevent Overfit
2	Batch Normalization
3	Layer Normalization

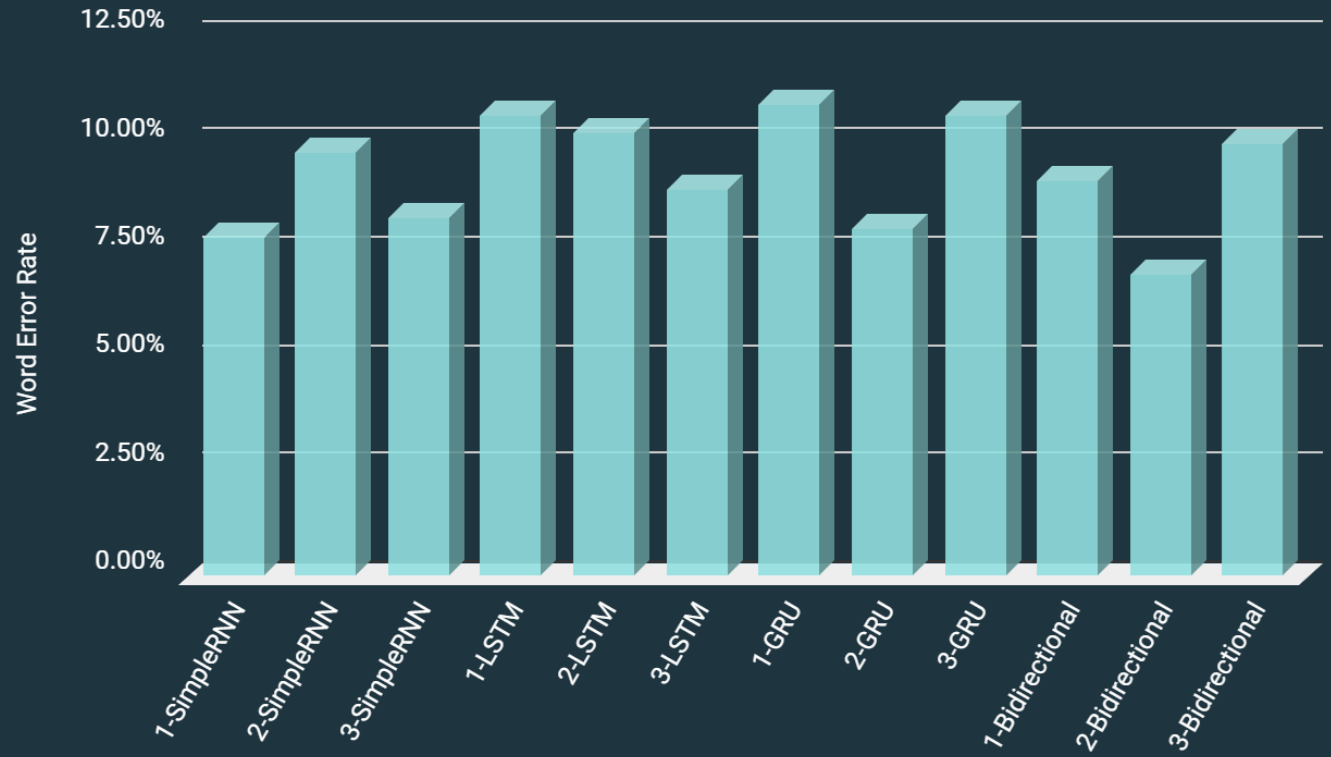
MODEL VS. VALIDATION ACCURACY



Legend

1	Prevent Overfit
2	Batch Normalization
3	Layer Normalization

Model VS WORD ERROR RATE



Legend

1	Prevent Overfit
2	Batch Normalization
3	Layer Normalization

Best model

Model	Word Error Rate	Temperature	Validation Accuracy
LSTM 3	**8.91%**	**0.5**	**0.567**

Lstm model with layer normalization:

- Extremely low word error rate
- Highest validation accuracy across all models

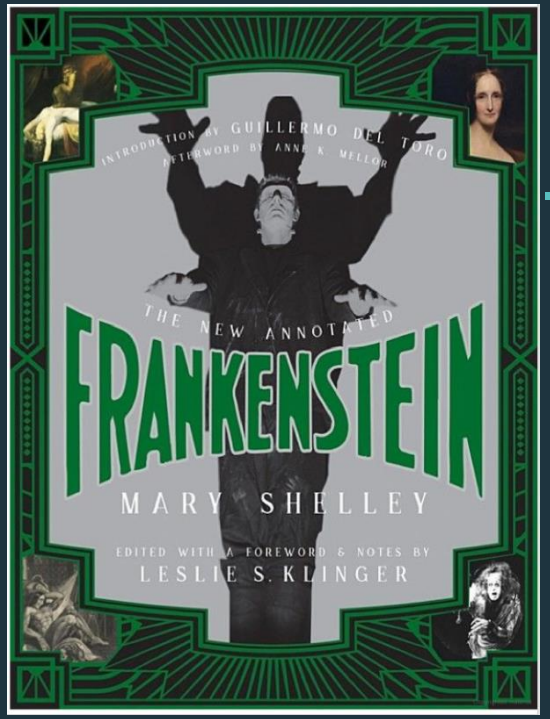
Input real text 

Text chosen

Book Chosen:
Frankenstein the 1818

Reasons:

- Released in the same time era as Sherlock Holmes
- Has similar use of artistic language
- Model will generate text more inclined to the writing in Sherlock Holmes

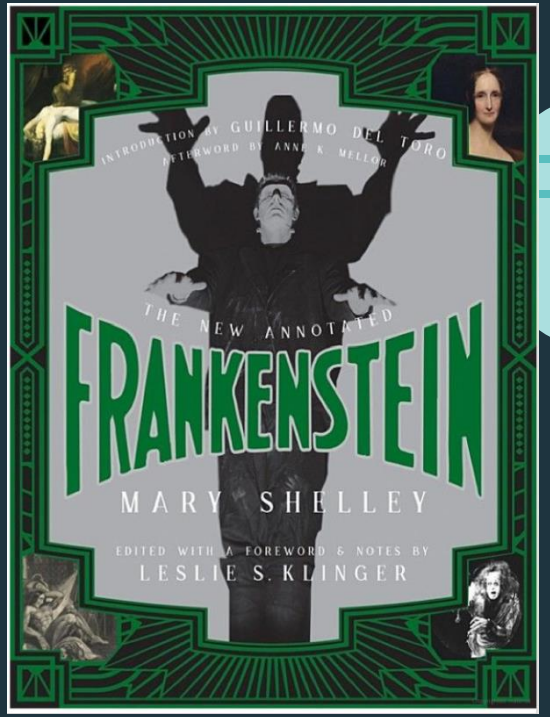


Text chosen

```
# takes the user input
text_input = input()
print(len(text_input))
```

```
one which has ever been considered as a mystery; yet with how many things are we upon the brink of becoming acquainted, if c
owardice or carelessness did not restrain our inquiries. I revolved these circumstances in my mind and determined thencefort
h to apply myself more particularly to those branches of natural philosophy which relate to physiology. Unless I had been an
imated by an almost supernatural enthusiasm, my application to this study would have been irksome and almost intolerable. To
examine the causes of life, we must first have recourse to death. I became acquainted with the science of anatomy, but this
was not sufficient; I must also observe the natural decay and corruption of the human body. In my education my father had ta
ken the greatest precautions that my mind should be impressed with no supernatural horrors. I do not ever remember to have t
rembled at a tale of superstition or to have feared the apparition of a spirit. Darkness had no effect upon my fancy, and a
churchyard was to me merely the receptacle of bodies deprived of life, which, from being the seat of beauty and strength, ha
d become food for the worm.
Generating with seed: "one which has ever been considered as a mystery; yet with ho"
```

Extra words added to allow for Word Error Rate(WER) to work



Output from best model

```
--- Generating with seed: "one which has ever been considered as a mystery; yet with ho"
------ temperature: 0.2
one which has ever been considered as a mystery; yet with hormed to the completely of the companion.  it was a signer of the
country of the companion of the country of the street.  i shall be the companion of the companion of the companion of the co
nsider of the companion of the man.  i was a companion of the side of the country of the country of the palice of the compan
ion of the statement of the companion of the consider of the companion of the street.

WER: 12.39%
------ temperature: 0.5
one which has ever been considered as a mystery; yet with home to the astions of bringing to the side of the statement of th
e lodge, which was a sharp at the police. and then he had a very mendinct which bring his window which the considerable and
did the door of the adventient of brinding that the party of the present of the companion of the street.  and then i was a v
ery meaning and that the facts were all the bedroom of the consider. there is a man whic

WER: 13.26%
------ temperature: 1.0
one which has ever been considered as a mystery; yet with however, whisterly is sight, there strumed in the craph into lide
as to seemed the pocket.  oh, chrespend my companions this only keen head he and see, watson, so encelurned until the relard
consider of the damp of me, in anythan mad shage a spull war if the implecsion-amplate faced vict.  i !n the troth woman, an
d they still, better a glope posit chair, said holmes.  you know that he has both struc

WER: 13.70%
------ temperature: 1.2
one which has ever been considered as a mystery; yet with holmesj, and sike the man matter alwost blew before, stay which wi
ll su-s-falled in e charely two savell of the wornd, which you live that a house well-could follow no ebreld ruch to ceathe
t. you shly betfeed anyone from the linch doon through the compan; yet, then he has been their stumym which enoughd on thes
e? how youngel of as wickon open me.ser true it would say, shouldor. this leds that exclip

WER: 13.70%
```

Summary

Key learning points

- The best performing RNN layer was LSTM
- Layer Normalization generally resulted in better performances
- Layer Normalization works better in RNN compared to Batch Normalization
- Optimal Temperature is 0.5

Improvements to be made

- Other RNN layers like Conv1D could have been tested
- Other evaluation techniques like Bleu and METEOR could have been tested
- Larger variety of texts could have been used to increase the dataset which would result in better results

Thank You!!