Name: Chia Bing Xuan

Matriculation Number: A0259419R

Title: DSA4213 Assignment 2 – Small Language Models Exploration

1. Introduction

In this assignment, I conducted experiments on two types of models – vanilla Recurrent Neural Networks (RNNs) and Long Short-Term Memory RNNs (LSTMs). With the use of a selected corpus, RNNs and LSTMs were built and trained for language modelling.

2. Explanation of Models

2.1. RNNs

Note that RNNs can process input sequences of any length. In the context of language modelling, suppose we have an input sequence of textual tokens, $\{\boldsymbol{x}^{(t)}\}_{1\le t\le T}$ where T is the sequence length. At each sequential time step t, the corresponding token $\boldsymbol{x}^{(t)}$ is fed into the RNN. It is first converted into a word embedding $\boldsymbol{e}^{(t)}$ with the transformation matrix $\boldsymbol{E}$

$$\boldsymbol{e}^{(t)} = \boldsymbol{E}\boldsymbol{x}^{(t)}$$

The RNN possesses a hidden state that will be updated during each time step. During time step t, the hidden state from the previous step, $\boldsymbol{h}^{(t-1)}$, will be updated using the word embedding $\boldsymbol{e}^{(t)}$ to yield the new hidden state

$$\boldsymbol{h}^{(t)} = \sigma(\boldsymbol{W_h}\boldsymbol{h}^{(t-1)} + \boldsymbol{W_e}\boldsymbol{e}^{(t)} + \boldsymbol{b_1})$$

where σ is the sigmoid function and $\boldsymbol{W_h}, \boldsymbol{W_e}, \boldsymbol{b_1}$ are learnable parameters. Finally, the hidden state of the current time step is used to output a probability distribution over all distinct tokens in the vocabulary V:

$$\hat{\boldsymbol{y}}^{(t)} = \mathrm{softmax}(\boldsymbol{U}\boldsymbol{h}^{(t)} + \boldsymbol{b_2}) \in \mathbb{R}^{|\boldsymbol{V}|}$$

where $\boldsymbol{U}$ and $\boldsymbol{b_2}$ are learnable parameters in the RNN. To predict the next token, it can be chosen by selecting the token corresponding to the highest probability. Alternatively, the next token can be sampled from the output probability distribution.

Note that the same weights $\boldsymbol{W_h}, \boldsymbol{W_e}$ will be repeatedly applied on every time step. They will only be updated through gradient descent after all the input tokens in the sequence have been processed. Hence, there is symmetry in how these input tokens are being processed.

2.2. LSTMs

LSTMs are RNNs that aim to alleviate the vanishing gradient problem – an issue which makes it difficult for vanilla RNNs to learn long-range dependencies and preserve information over many timesteps.

Suppose we have completed the previous time step $t-1$. The LSTM has a hidden state $\boldsymbol{h}^{(t-1)}$ and a cell state $\boldsymbol{c}^{(t-1)}$, the latter of which can store long-term information. To update the hidden state and cell state for the current time step t, we first determine three gates – forget gate $\boldsymbol{f}^{(t)}$, input gate $\boldsymbol{i}^{(t)}$

and output gate $\boldsymbol{o}^{(t)}$. We also determine the new cell content that can be written to the cell in this time step, $\tilde{\boldsymbol{c}}^{(t)}$. These vectors are dynamically obtained based on the current word embedding $\boldsymbol{e}^{(t)}$:

$$\boldsymbol{f}^{(t)} = \sigma(\boldsymbol{W}_f \boldsymbol{h}^{(t-1)} + \boldsymbol{U}_f \boldsymbol{e}^{(t)} + \boldsymbol{b}_f)$$
$$\boldsymbol{i}^{(t)} = \sigma(\boldsymbol{W}_i \boldsymbol{h}^{(t-1)} + \boldsymbol{U}_i \boldsymbol{e}^{(t)} + \boldsymbol{b}_i)$$
$$\boldsymbol{o}^{(t)} = \sigma(\boldsymbol{W}_o \boldsymbol{h}^{(t-1)} + \boldsymbol{U}_o \boldsymbol{e}^{(t)} + \boldsymbol{b}_o)$$
$$\tilde{\boldsymbol{c}}^{(t)} = \tanh(\boldsymbol{W}_c \boldsymbol{h}^{(t-1)} + \boldsymbol{U}_c \boldsymbol{e}^{(t)} + \boldsymbol{b}_c)$$

where the $\boldsymbol{W}$'s, $\boldsymbol{U}$'s and $\boldsymbol{b}$'s are learnable parameters. To determine the new cell state $\boldsymbol{c}^{(t)}$, the forget gate selectively removes some content from the previous cell state $\boldsymbol{c}^{(t-1)}$, while the input gate selectively writes some new cell content from $\tilde{\boldsymbol{c}}^{(t)}$:

$$\boldsymbol{c}^{(t)} = \boldsymbol{f}^{(t)} \odot \boldsymbol{c}^{(t-1)} + \boldsymbol{i}^{(t)} \odot \tilde{\boldsymbol{c}}^{(t)}$$

Finally, the output gate selectively reads some content from the updated cell state $\boldsymbol{c}^{(t)}$, so as to obtain the updated hidden state for the current time step:

$$\boldsymbol{h}^{(t)} = \boldsymbol{o}^{(t)} \odot \tanh \boldsymbol{c}^{(t)}$$

3. Methodology

3.1. Selection of Corpus and Data Processing

I selected a corpus of Reuters news documents offered by the NLTK library, which consists of 10788 news documents. Owing to limitations in computational power, only a subset of the dataset was used (2500 documents). Firstly, a train-validation-test split of 80/10/10 was applied to this subset. This enables us to obtain a training set of 2000 documents, a validation set of 250 documents and a testing set of 250 documents. For each of these documents, we then applied a round of data pre-processing:

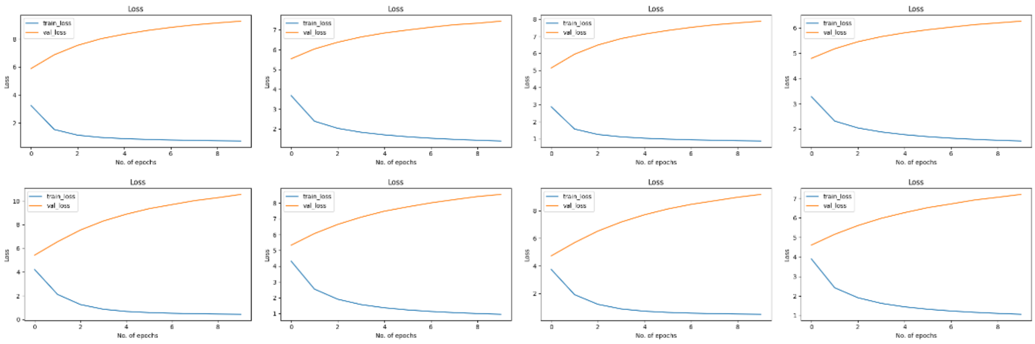
1. Tokenise the document. I experimented with two methods of tokenisation – *word tokenisation* using torchtext's basic English tokeniser, and *subword tokenisation* using a base BigBird transformer model (based on SentencePiece).
2. For each token produced, convert it to lowercase
3. Add <unk> (unknown), <bos> (beginning of sentence) and <eos> (end of sentence) tokens
4. Numericalise the tokens by converting them to unique integer IDs

3.2. Model Training

Subsequently, I trained RNNs and LSTMs for language modelling, on both the word tokens and subword tokens from the training set. For this investigation, tokenisation method and dropout were selected as the dependent variables. Across all the trials, the following were kept constant:

- Embedding size = 128 for the embedding layer of the RNN / LSTM
- Hidden size = 256 for the hidden state of the RNN / LSTM
- Number of layers = 2 for the RNN / LSTM unit
- Sequence length = 32 tokens. A small value was chosen, owing to computational constraints
- Number of epochs = 10
- Batch size = 32
- Optimiser = Adam
- Learning rate = 10^{-3}
- Maximum norm for gradient clipping = 1.0

Trial	Model	Tokenisation Method	Dropout (between RNN / LSTM layers)	Training Time	Final Validation Cross Entropy Loss	Final Validation Perplexity
1	RNN	Word	0.0	23 min 24 s	9.2678	10591.1113
2	RNN	Word	0.2	24 min 16 s	7.4438	1709.1497
3	RNN	Subword	0.0	30 min 49 s	7.8859	2659.3932
4	RNN	Subword	0.2	32 min 12 s	6.2666	526.7011
5	LSTM	Word	0.0	23 min 5 s	10.5475	38083.9244
6	LSTM	Word	0.2	23 min 35 s	8.5566	5201.0926
7	LSTM	Subword	0.0	29 min 43 s	9.1753	9655.6448
8	LSTM	Subword	0.2	30 min 24 s	7.2038	1344.5489



For loss curves: Row 1 (left to right) – Trials 1, 2, 3 and 4. Row 2 (left to right) – Trials 5, 6, 7 and 8

Across all 8 trials, the training loss tends to 0, while the validation loss increases with each epoch. Hence, each of the models has overfitted to the training data during the training process. This can be attributed to the fact that only a small subset of the Reuters dataset was used, which might be insufficient for model training. Since the training set is relatively small, the tokens in the validation set are more likely to be absent in the training vocabulary. Therefore, it is possible that the training set and validation set are significantly different from one another, which explains why the validation loss increases as the models fit to the training data. Trial 4 corresponds to the smallest validation loss and perplexity, displaying the smallest degree of overfitting.

4. Evaluation

4.1. Quantitative Analysis (Cross Entropy Loss and Perplexity)

Each of the trained models was evaluated on a testing dataset. Model performance was quantified using two metrics – cross entropy loss and perplexity.

Trial	Model	Tokenisation Method	Dropout	Test Cross Entropy Loss	Test Perplexity
1	RNN	Word	0.0	10.0254	22593.0692
2	RNN	Word	0.2	8.0771	3219.9619
3	RNN	Subword	0.0	8.5430	5130.6781
4	RNN	Subword	0.2	**6.8204**	**916.3774**
5	LSTM	Word	0.0	11.6870	119014.8032
6	LSTM	Word	0.2	9.4750	13029.5869
7	LSTM	Subword	0.0	10.0901	24103.5258
8	LSTM	Subword	0.2	7.9883	2946.2009

For the same model type and dropout value, subword tokenisation consistently achieves a better performance than word tokenisation. This is likely because subword tokenisation more robustly allows the models to reason about structures below the word level – such as common prefixes, suffixes and roots. This is useful when dealing with many variants of the same root word (eg. “help”, “helping”, and “helper”). With word tokenisation, the model might not be able to learn that these words are semantically similar, since they are treated as distinct tokens. With subword tokenisation, not only is the model able to learn a strong representation for the root word itself (“help”), but it can also learn the use of prefixes and suffixes (“ing” and “er”) by generalising to the structures of other words (eg. “singing” and “singer”). Furthermore, for rare words that would have otherwise been outside of the model’s vocabulary, subword tokenisation helps break them down into more common subwords that the model has seen before. Hence, the model is still able to infer the meanings of such words.

For the same model type and tokenisation method, the addition of dropout consistently improves performance. This is unsurprising as the random dropout of neurons prevents them from depending too heavily on one another, causing neuron co-adaptation to occur to a smaller extent. Hence, this reduces the likelihood of the model overfitting to the training data, allowing it to generalise to the testing data more effectively.

For the same tokenisation method and dropout value, vanilla RNNs appear to consistently outperform LSTMs. There are two possible reasons for this. Firstly, the size of the training dataset is rather small. Since LSTMs have more trainable parameters than RNNs, the higher model complexity means that LSTMs are more likely to overfit during training, leading to poorer generalisation to the testing data. Secondly, note that LSTMs are advantageous when it comes to preserving information over a long sequence of time steps. Since the context length selected (32) is rather short anyway, there is less of a need to capture, preserve and learn long-range dependencies. This suggests that the benefits of LSTMs over RNNs are less apparent, for this particular language modelling task.

On this note, trial 4 (RNN, subword tokenisation, dropout = 0.2) leads to the smallest testing loss and perplexity, displaying the strongest performance for this language modelling task.

4.2. Qualitative Analysis (Text Generation)

For each trial, text generation was performed with three different temperatures – 0.7, 1.0 and 1.3. “Today” was used as the starting word of choice. Here, we only show the generated text for trial 4 and trial 8, which are the best-performing trials for RNNs and LSTMs respectively. The generated text for all the trials can be found on the GitHub repository.

Trial	Temperature	Generated Text (whitespace is represented as underscores (“_”))
4	0.7	Today ," __he __added . __" the __dollar 's __meeting __is __being __placed __on __the __outcome __of __collective __wage __agreement . __the __meeting __was __attended __by __farm __ministers __in __the __past __one __days __during __the __year - ago __week , __traders __said . __they __said __the __relaxation __of __controls __was __now . __sharp __details __were __not __disclosed . __the __company , __which __includes __a __69 __m ln __ dl rs __for __capital __spending , __less __than __half __of __the __amount __spent __70 . 6 __p ct __from __13 . 13 __m ln __in __ap ril , __and __brought __its __stake __in __s ime __ dar by , __ a __spokesman __for __gr aan __elevator __m ij , __the __largest __employer __yesterday . __earlier __today , __the __company __said __the __transaction __ involves __the __combination __of __oil __and __by __an __existing __20 __p ct __gain __in __w es ts __ uc cel __corp __as __pen ney 's __co __and __& lt ; format ura __ ini ez ione __po lim eri __spa >, __gen oa __and __& lt ; l om > __diesel __systems __and __other __company , __r oland __has __said __it __will __make __ their __part __but __the __group __of __seven __-- __the __united __states , __would __be __the __critical __factor . __they __said __the __relaxation __of __ controls __was __now __as __a __result __of __abroad ," __he __added . __ __" sq ui bb __is __putting __them , __it __said . __sci - med __year ley __inc __said __ it __purchased __the __group __of __seven __ministers __in __the __gulf __to __country , __with __a __slightly __high __priority __from __the __price __of __the __united __states __with __reporting __such __as __prices __for __the
	1.0	Today __vul __prime __policy __will __now __be __followed __by __the __potential ," __said gil lette . __shore __mentioned __half __its __ch il uba , __leader __of __the __z am bian __congress __in __the __u . s . __carries __out __a __threat __to __make __an __effort __trade __in __a __near - term __strike __on __t enders , __he __added __in __t ok yo __were __on __notice __this __week __because __of __the __j apan ese __economy __have __had __been __taken __higher __than __ the __commercial __customers . __lower __taxes __would __be __early __by __the __u . s . __currency __proposal , __which __is __required __to __report __the __ bah rain __s mel ter __9 . __p ea body __and __co . __" there __is __ongoing , __serious __thought __applied __to __dome __in __its __retail __second __offer , __ thus __ass uring __said __total __of __118 . 2 __m ln __s tg __the __j apan ese __early __income __goods __necessary , __and __that __it __would __be __taken __ with __underlying __government __to __keep __our __trade __laws ." __" this __is __a __two - for - five __bonus __issue , __a __former __ argo sy st ems . __the __ new __firm , __said __by __the __move __to __a __further __dollar __down __because __that __it __was __reached __in __new __y ork __and __seven , 000 __ tonnes __of __wheat __for __an __e cu __gain __and __earlier __in __march __of __this __year 's __current __fiscal __quarter . __the __public __1985 / 87 __orders __at __end . 70 __m ln __tonnes , __vs __24 . 7 __m ln __six __months __shr __profit __eight __c ts __vs __loss __two __c ts __net __loss __ 725 , 000 __vs __profit __310 ,

	1.3	Today __also __will __not __do __anything __for __the __metal __rates . __" such __a __gathering __of __money __and __foreign __exchange __last __fixed __over __& lt ; bank burg __pin cus __and __co __& lt ; up cm . o > __and __a __shareholders __by __the __j apan __will __maintain __its __prime __and __both __electronic __abolish __foreign __investment __in __the __second __six __months __of __1987 . __loans __to __j une __17 __and __112 , 000 __b pd , __u . s __9 __and __2 . 99 __billion __ dl rs __from __86 __m ln __d lr __from __one __third __quarter __ended __ap ril __30 __shr __two . 2 __cents __with __the __rice __for __the __ energy __cocoa __in __fiscal __leading __producer __are __need __to __put __who __have __some __leading __industrial __house __and __west __ger many 's __ largely __" __goal __on __t uesday 's __west __g erman __monetary __policy __to __aid __the __dollar - if __end __and __there __are __are __likely __at __th over __banks . __" compan ies __like __dup ont __& lt ; dd > __ nova __said __the hor __offer __were __400 , 000 __tonnes __of __wheat __from __14 . 7 __m ln __u . s . __oil __as __under __information __new __y ork __denied __the __motion __here . __analysts __said __the __loss __of __the __inst itutes __will __become __ improving __are __needed __with __stake __to __prevent __share __in __its __loan __negotiations , __17 . 0 __m ln __sw iss __f on 's __amount __to __end __their __s aj __paid . __ ira qi __troops __investors __in __tex as __mine __ak __co k ' __has __approved __an __agreement __for __12 __m ln __trader __bid __closes __ on __behalf __of
8	0.7	Today ," __said __a __wall __street __arbit rag eur . __but __he __said __the __bank __regards __the __over draft __reference __rate , __based __on __short - term __rate __trends , __as __its __key __prime __lending __rate __to __corporate __customers . __the __loan __reference __rate __is __based __on __longer __term __ trends . __the __bank __is __the __latest __to __cut __prime __rates __in __the __next __few __weeks . __given __an __average __yield __of __1 . 09 __billion __ dl rs __of __j uly __accounting __to __three __m ln __ dl rs __to __226 . 4 __m ln __ dl rs , __compared __with __169 . 2 __m ln __ dl rs __in __year __to __de cember __31 , __1986 . __shr __18 . 9 __m ln __vs __5 . 7 __m ln __note __- __full __name __is __data __services , __inc __said __it __will __offer __5 . 0 __m ln __ dl rs __ cash __of __2 . 2 __m ln __common __shares __of __stock __for __each __ uc cel __wholly __based __on __50 __p ct __of __the __equity __in __j une , __it __ added . __the __application __of __the __additional __48 __p ct __since __the __total __outstanding , __which __operates __46 __branches , __has __been __ completed __in __principle __to __purchase __about __10 __m ln __va __shares __to __9 . 6 __m ln __ dl rs __of __fox __assets __of __n orc ros __p lc __& lt ; dl rs > __having __t bs , __which __is __worth __ab outh __55 __ dl rs __per __share . __" we 've __taken __up __with __all __alternatives __which __were __largely __ complementary __unless __the __government 's __programme 's __current __management __and __tobacco __spirits . __ __the
	1.0	Today >. __j apan 's __letter __follows __a __12 __p ct __increase __in __s ept ember __single - family __unit __starts __a __1 . 5 __p ct __to __1 . 73 __m ln __s tg , __it __is __also __most __for __good __recovery , __c omin co __said __in __view __that __levels __should __be __at __least __behind __this __time . __the __ officials __said __ger many __had __practically __no __growth __in __the __long __term __and __bond __prices __was __not __immediately __clear __the __ company __will __be __able __to __offer __specifics __from __the __sale __of __a __controlling __of __its __reliance __standard __life __insurance __co . __a __d alian __subsidiary , __r - held __dat eline ron __equipment __to __carry lift __as __s ears __ roe buck __and __co __& lt ; s > __converted __subsidiary __senior __ management __of __international __data __corp __said __it __has __agreed __to __combine __its __cocoa __processing __businesses __with __those __of __s . __ . c . __terms , __effective __yesterday , __within __400 __miles __east __of __v ancouver , __produced __2 . 9 __m ln __barrels __a __year __earlier . __pre tax __ earnings __fell __7 . 2 __p ct __to __3 . 48 __billion __pes os __to __the __dollar / 92 __to __17 . 6 __p ct __from __18 . 25 , __effective __tomorrow . __the __new __rate __is __based __on __a __basket __year __of __45 . 5 __p ct __this __year , __with __one __to __eight __billion __ dl rs __by __about __five __p ct . __in __ an __initial __comment , __including __a __employee __stock __at __the __same __period __from __a __deficit __of __1 , 450 __billion __in __the __week __ended __ap ril __4 , __1987 , __sales _sales
	1.3	Today 's __further __cut __since __they __should __earn __more __cash __or __before __in __fiscal __production __came . __speaking __at __a __forum __for __ indust ria lists __of __major __central __banks __meeting __today 's __federal __reserve . __the __news __eroded __the __most __immediate __to __over __the __ r eval uation __general __meeting , __the __meat __investment __firm __said . __sci - med __said __it __continues __to __be __identified __affected __by __ ron ald __per elman , __offered __20 __ dl rs __lower __than __an __average __yield __1 . 04 __ dl rs __in __liqu idation __value __of __a __stock __in __each __share __of __arrays __bought __or __less __than __2 . 4 __in __response __to __its __oil __market ." __ donald __trump , __rose __to __1 . 0 __m ln __ dl rs __reflecting __ proposed __to __hol stein / 1 , 000 __when __h ilton __petroleum __spending __mill __workers __at __one __of __its __products , __at las __group , __prices __ collapsed __by __the __end __of __an __oil __industry . __however , __government __officials __said . __au str alia 's __latest __rates __allot __are __not __going __to __be __more __early ." __bun des bank __officials __had __shown __fed __the __chances __to __be __what __i ran __know __before __the __fed __takes __ sales __of __policy __than __a __15 __p ct __increase __in __gross __national __center __of __next __oct ober __as __of __export __subsidies __from __other __ leading __paper __products , __the __gulf __for __ ku wait __tot alled __about __20 __billion __ dl rs __by __be cor 's __forecast __to __oct __1 . __he __noted __ that __while __the __o pec __meeting __in __l j ens __to __buy __a __six __p ct __rate __rate __to __7 - 1 /

The generated text is rather illogical, with the presence of frequent syntactic and grammatical errors. Semantic meaning can be inferred from very short sequences of words. However, limited continuity is observed beyond that. As a whole, the text itself generally does not convey a clear message, and there is little relationship between sentences. We can look at trial 4 (temperature = 1.3) as an example. The sequence "analysts said the loss of the institutes will become" makes sense, yet there is no mention of this "loss of the institutes" prior to it being mentioned. Moreover, what follows is oddly phrased and less reasonable ("improving are needed with stake to prevent share in its loan negotiations").

As temperature increases, the text generated becomes more varied and less repetitive. In trial 4, a temperature value of 0.7 sees the same sequence "the relaxation of controls was now" being generated twice, while "the group of seven" also appears twice. However, a higher temperature value of 1.3 corresponds to greater creativity, with a wider range of tokens being used. Yet, this volatility also leads to the generation of more incomprehensible text, in which seemingly unrelated entities show up within the same sequence of tokens (eg. "<u>iraqi troops</u> <u>investors</u> in <u>texas mine</u>").

Unlike the quantitative analysis, LSTMs arguably generate text of a slightly higher quality than RNNs – though they are still somewhat comparable. For a temperature value of 1.0, the RNN-generated text is largely devoid of meaning. There are very few instances of correctly phrased sequences – a notable example being "this year's current fiscal quarter". However, the LSTM-generated text has more occurrences of valid sequences. These sequences are also longer than those generated by the RNN. Examples include "the officials said germany had practically no growth in the long term", "japan's letter follows a 12 pct increase in" and "pretax earnings fell 7.2 pct to 3.48 billion pesos to the dollar".

5. Appendix

I used GPT-5 to assist in the creation of code and improve the phrasing of the report. I am responsible for the content and quality of the submitted work. The GitHub repository for this assignment can be found at https://github.com/chiabingxuan/Small-Language-Models-Comparison.