Due Date: March 17th 23:00, 2020

Instructions

- *For all questions, show your work!*
- *Submit your report (pdf) and your code electronically via the course Gradescope page.*
- *An outline of code will be provided in the course repo at this link. You must start from this outline and follow the instructions in it (even if you use different code, you must follow the overall outline and instructions).*
- *TAs for this assignment are Jessica Thompson, Jonathan Cornford and Lluis Castrejon.*

Summary:

In this assignment, you will implement and train **sequential language models** on the Penn Treebank dataset. Language models learn to assign a likelihood to sequences of text. The elements of the sequence (typically words or individual characters) are called tokens, and can be represented as one-hot vectors with length equal to the vocabulary size, e.g. 26 for a vocabulary of English letters with no punctuation or spaces, in the case of characters, or as indices in the vocabulary for words. In this representation an entire dataset (or a mini-batch of examples) can be represented by a 3-dimensional tensor, with axes corresponding to: (1) the example within the dataset/mini-batch, (2) the time-step within the sequence, and (3) the index of the token in the vocabulary. Sequential language models do **next-step prediction**, in other words, they predict tokens in a sequence one at a time, with each prediction based on all the previous elements of the sequence. A trained sequential language model can also be used to generate new sequences of text, by making each prediction conditioned on the past *predictions* (instead of the ground-truth input sequence).

As a starting point, you are provided with an implementation of a **simple ("vanilla") RNN** (recurrent neural network). Problem 1 is to implement an RNN with a gating mechanism on the hidden state, specifically with **gated recurrent units (GRUs)**. Problem 2 is to implement the **attention module of a transformer network** (we provide you with PyTorch code for the rest of the transformer). Problem 3 is to train these 3 models using a variety of different optimizers and hyperparameter settings and Problem 4 is to analyze the behaviour of the trained models. Each problem is worth 25 points.

The Penn Treebank Dataset This is a dataset of about 1 million words from about 2,500 stories from the Wall Street Journal. It has Part-of-Speech annotations and is sometimes used for training parsers, but it's also a very common benchmark dataset for training RNNs and other sequence models to do next-step prediction.

Preprocessing: The version of the dataset you will work with has been preprocessed: lower-cased, stripped of non-alphabetic characters, tokenized (broken up into words, with sentences separated by the `<eos>` (end of sequence) token), and cut down to a vocabulary of 10,000 words; any word not in this vocabulary is replaced by `<unk>`. For the transformer network, positional information (an embedding of the position in the source sequence) for each token is also included in the input sequence. In both cases the preprocessing code is given to you.

Problem 1

Implementing an RNN with Gated Recurrent Units (GRU) (25pts) The implementation of your RNN must be able to process mini-batches. Implement the model **from scratch** using PyTorch Tensors, Variables, and associated operations (e.g. as found in the `torch.nn` module). Specifically, use appropriate matrix and tensor operations (e.g. dot, multiply, add, etc.) to implement the recurrent unit calculations; you **may not** use built-in Recurrent modules. You **may** subclass `nn.module`, use built-in Linear modules, and built-in implementations of nonlinearities (tanh, sigmoid, and softmax), initializations, loss functions, and optimization algorithms. Your code must start from the code scaffold and follow its structure and instructions.

The use of "gating" (i.e. element-wise multiplication, represented by the $\odot$ symbol) can significantly improve the performance of RNNs. The Long-Short Term Memory (LSTM) RNN is the best known example of gating in RNNs; GRU-RNNs are a slightly simpler variant (with fewer gates).

The equations for a GRU are:

$$\begin{aligned}
\boldsymbol{r}_t &= \sigma_r(\boldsymbol{W}_r\boldsymbol{x}_t + \boldsymbol{U}_r\boldsymbol{h}_{t-1} + \boldsymbol{b}_r) && (1)\\
\boldsymbol{z}_t &= \sigma_z(\boldsymbol{W}_z\boldsymbol{x}_t + \boldsymbol{U}_z\boldsymbol{h}_{t-1} + \boldsymbol{b}_z) && (2)\\
\tilde{\boldsymbol{h}}_t &= \sigma_h(\boldsymbol{W}_h\boldsymbol{x}_t + \boldsymbol{U}_h(\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1}) + \boldsymbol{b}_h) && (3)\\
\boldsymbol{h}_t &= (1 - \boldsymbol{z}_t) \odot \boldsymbol{h}_{t-1} + \boldsymbol{z}_t \odot \tilde{\boldsymbol{h}}_t && (4)\\
P(\boldsymbol{y}_t|\boldsymbol{x}_1, ..., \boldsymbol{x}_t) &= \sigma_y(\boldsymbol{W}_y\boldsymbol{h}_t + \boldsymbol{b}_y) && (5)
\end{aligned}$$

$\boldsymbol{r}_t$ is called the "reset gate" and $\boldsymbol{z}_t$ the "forget gate". The trainable parameters are $\boldsymbol{W}_r, \boldsymbol{W}_z, \boldsymbol{W}_h, \boldsymbol{W}_y$, $\boldsymbol{U}_r, \boldsymbol{U}_z, \boldsymbol{U}_h, \boldsymbol{b}_r, \boldsymbol{b}_z, \boldsymbol{b}_h$, and $\boldsymbol{b}_y$, as well as the initial hidden state parameter $\boldsymbol{h}_0$. GRUs use the sigmoid activation function for σ_r and σ_z, and tanh for σ_h.

See further instructions in the solution template.

Problem 2

Implementing the attention module of a transformer network (25pts) While prototypical RNNs "remember" past information by taking their previous hidden state as input at each step, recent years have seen a profusion of methodologies for making use of past information in different ways. The transformer [1] is one such fairly new architecture which uses several self-attention networks ("heads") in parallel, among other architectural specifics. The transformer is quite complicated to implement compared to the RNNs described so far; most of the code is provided and your task is

[1] See `https://arxiv.org/abs/1706.03762` for more details.

only to implement the multi-head scaled dot-product attention. The attention vector for m heads indexed by i is calculated as follows:

$$\boldsymbol{A}_i = \text{softmax}\left(\frac{\boldsymbol{Q}_i \boldsymbol{W}_{Q_i} (\boldsymbol{K}_i \boldsymbol{W}_{K_i})^\top}{\sqrt{d_k}}\right) \tag{6}$$

$$\boldsymbol{H}_i = \boldsymbol{A}_i \boldsymbol{V} \boldsymbol{W}_{V_i} \tag{7}$$

$$A(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{concat}(\boldsymbol{H}_1, ..., \boldsymbol{H}_m) \boldsymbol{W}_O \tag{8}$$

where $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ are queries, keys, and values respectively, $\boldsymbol{W}_{Q_i}, \boldsymbol{W}_{K_i}, \boldsymbol{W}_{V_i}$ are their corresponding embedding matrices, $\boldsymbol{W}_O$ is the output embedding, and d_k is the dimension of the keys. $\boldsymbol{Q}, \boldsymbol{K}$, and $\boldsymbol{V}$ are determined by the output of the feed-forward layer of the main network (given to you). $\boldsymbol{A}_i$ are the attention values, which specify which elements of the input sequence each attention head attends to.

Note that the implementation of multi-head attention requires binary masks, so that attention is computed only over the past, not the future. A mask value of 1 indicates an element which the model is allowed to attend to (i.e. from the past); a value of 0 indicates an element it is not allowed to attend to. This can be implemented by modifying the softmax function to account for the mask $\boldsymbol{s}$ as follows:

$$\tilde{\boldsymbol{x}} = \exp(\boldsymbol{x}) \odot \boldsymbol{s} \tag{9}$$

$$\text{softmax}(\boldsymbol{x}, \boldsymbol{s}) \doteq \frac{\tilde{\boldsymbol{x}}}{\sum_i \tilde{x}_i} \tag{10}$$

To avoid potential numerical stability issues, we recommend a different implementation:

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} \odot \boldsymbol{s} - 10^9(1 - \boldsymbol{s}) \tag{11}$$

$$\text{softmax}(\boldsymbol{x}, \boldsymbol{s}) \doteq \frac{\exp(\tilde{\boldsymbol{x}})}{\sum_i \exp(\tilde{x}_i)} \tag{12}$$

This second version is equivalent (up to numerical precision) as long as $\boldsymbol{x} >> -10^9$, which should be the case in practice.

Problem 3

Training language models and model comparison (25pts) Unlike in classification problems, where the performance metric is typically accuracy, in language modelling, the performance metric is typically based directly on the cross-entropy loss, i.e. the negative log-likelihood (NLL) the model assigns to the tokens. For word-level language modelling it is standard to report **perplexity (PPL)**, which is the exponentiated average per-token NLL (over all tokens):

$$\exp\left(\frac{1}{TN}\sum_{t=1}^{T}\sum_{n=1}^{N} -\log P(\boldsymbol{x}_t^{(n)} | \boldsymbol{x}_1^{(n)}, ..., \boldsymbol{x}_{t-1}^{(n)})\right),$$

where t is the index with the sequence, and n indexes different sequences. For Penn Treebank in particular, the test set is treated as a single sequence (i.e. $N = 1$). The purpose of this assignment is to perform model exploration, which is done using a validation set. As such, we do not require you to run your models on the test set.

You will train each of the three architectures using either stochastic gradient descent or the ADAM optimizer. The training loop is provided in *run_exp.py*.

1. - 4. You are asked to run 4 experiments (3.1, 3.2, 3.3, 3.4) with different architectures, optimizers, and hyperparameters settings. These parameter settings are given to you in the code (*run_exp.py*). In total there are 15 settings for you to run (5+3+3+4 = 15). For each experiment (3.1, 3.2, 3.3, 3.4), plot learning curves (train and validation) of PPL over both **epochs** and **wall-clock-time**. Figures should have labeled axes and a legend and an explanatory caption.

   Answer:

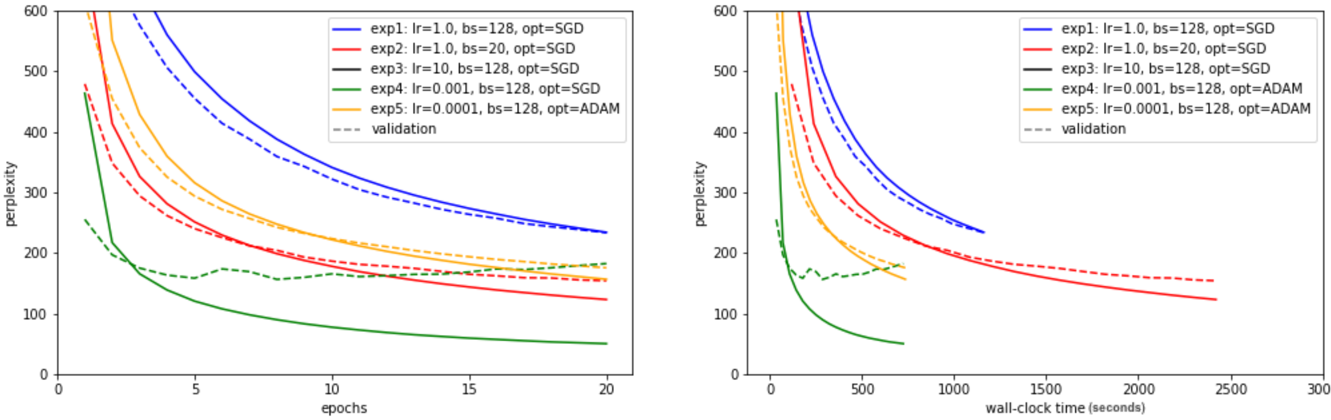


Figure 1: Plot for **Experiment 3.1**; Training and Validation PPL(perplexity) vs Epochs and Wall-clock time for RNN models for all 5 experiments. All the 5 experiments were tried with 2 hidden layers of size 512, sequence length of 35 with a dropout keep probability of 0.8. In the legends, *bs* stands for batch size and *lr* for learning rate. The perplexity for Experiment 3 cannot be shown in the graph as it exploded due to a very high learning rate of 10.

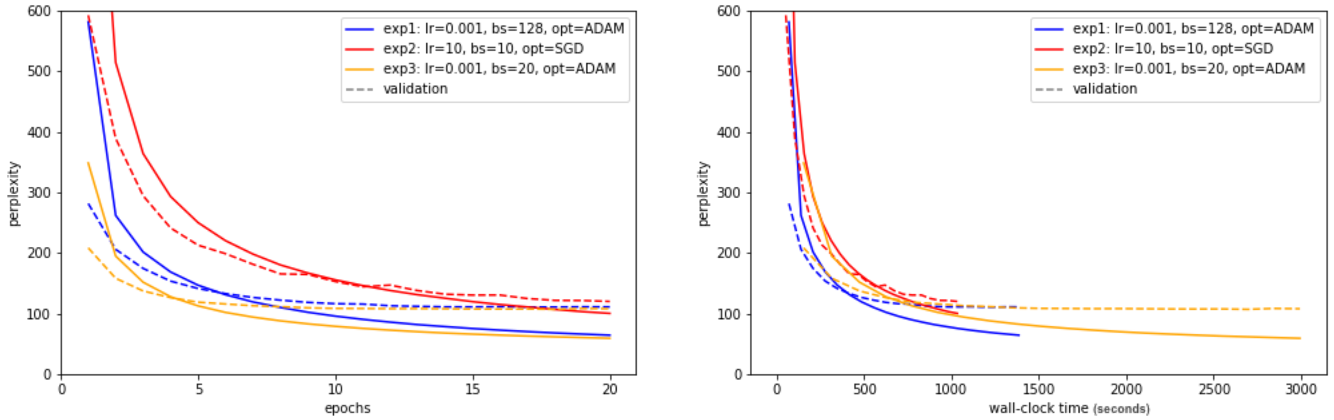


Figure 2: Plot for **Experiment 3.2**; Training and Validation PPL(perplexity) vs Epochs and Wall-clock time for GRU models for all 3 experiments. All the 3 experiments were tried with 2 hidden layers of size 512, sequence length of 35 with a dropout keep probability of 0.5. In the legends, *bs* stands for batch size and *lr* for learning rate.

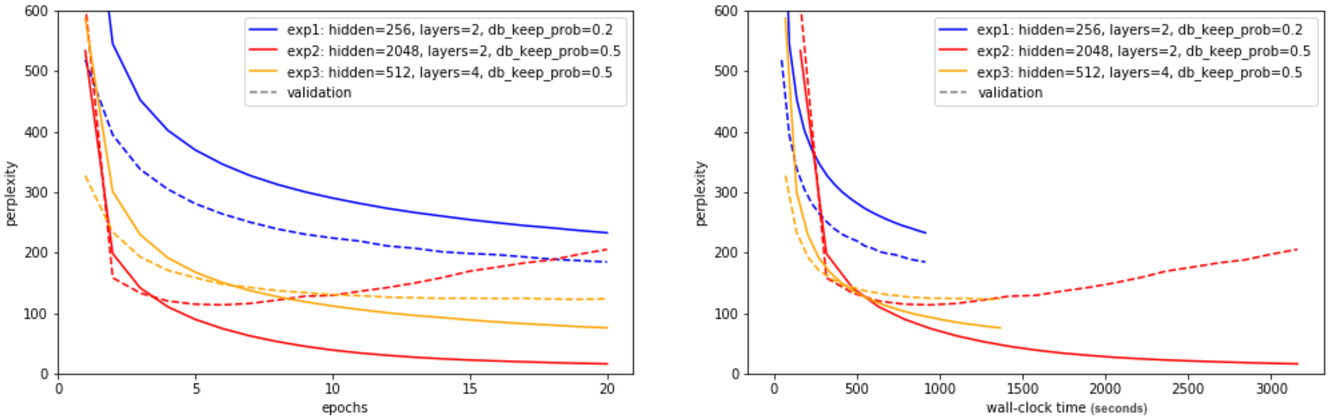


Figure 3: Plot for **Experiment 3.3**; Training and Validation PPL(perplexity) vs Epochs and Wall-clock time for GRU for all 3 experiments. All the 3 experiments were tried with a batch size of 128 with a starting learning rate of 0.001 and a sequence length of 35. The optimizer used in the experiments was ADAM.

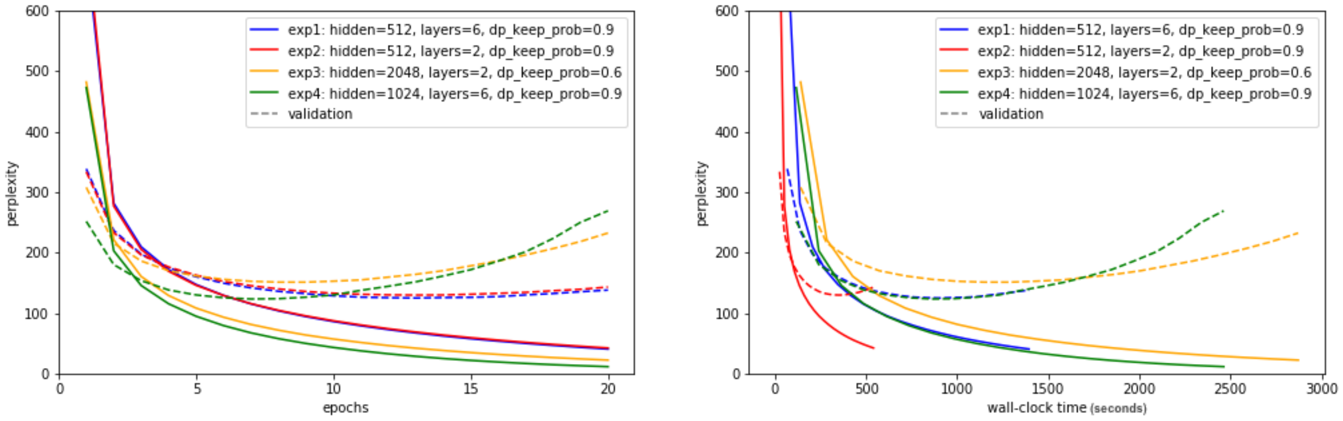
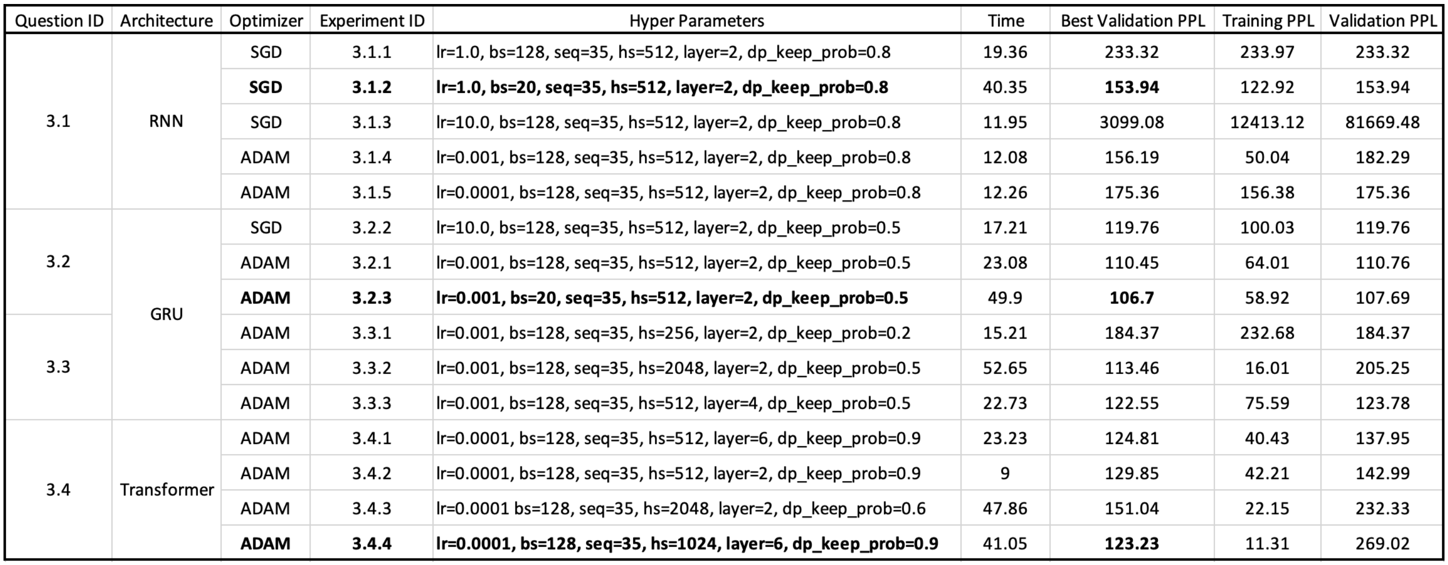


Figure 4: Plot for **Experiment 3.4**; Training and Validation PPL(perplexity) vs Epochs and Wall-clock time for Transformers for all 4 experiments. All the 4 experiments were tried with a batch size of 128 with a starting learning rate of 0.0001 and a sequence length of 35. The optimizer used in the experiments was ADAM.

5. Make a table of results summarizing the train and validation performance for each experiment, indicating the architecture and optimizer. Sort by architecture, then optimizer, and number the experiments to refer to them easily later. Bold the best result for each architecture.[2] The table should have an explanatory caption, and appropriate column and/or row headers. Any shorthand or symbols in the table should be explained in the caption.

   Answer:



Question ID	Architecture	Optimizer	Experiment ID	Hyper Parameters	Time	Best Validation PPL	Training PPL	Validation PPL
3.1	RNN	SGD	3.1.1	lr=1.0, bs=128, seq=35, hs=512, layer=2, dp_keep_prob=0.8	19.36	233.32	233.97	233.32
		SGD	**3.1.2**	**lr=1.0, bs=20, seq=35, hs=512, layer=2, dp_keep_prob=0.8**	40.35	**153.94**	122.92	153.94
		SGD	3.1.3	lr=10.0, bs=128, seq=35, hs=512, layer=2, dp_keep_prob=0.8	11.95	3099.08	12413.12	81669.48
		ADAM	3.1.4	lr=0.001, bs=128, seq=35, hs=512, layer=2, dp_keep_prob=0.8	12.08	156.19	50.04	182.29
		ADAM	3.1.5	lr=0.0001, bs=128, seq=35, hs=512, layer=2, dp_keep_prob=0.8	12.26	175.36	156.38	175.36
3.2	GRU	SGD	3.2.2	lr=10.0, bs=128, seq=35, hs=512, layer=2, dp_keep_prob=0.5	17.21	119.76	100.03	119.76
		ADAM	3.2.1	lr=0.001, bs=128, seq=35, hs=512, layer=2, dp_keep_prob=0.5	23.08	110.45	64.01	110.76
		ADAM	**3.2.3**	**lr=0.001, bs=20, seq=35, hs=512, layer=2, dp_keep_prob=0.5**	49.9	**106.7**	58.92	107.69
3.3		ADAM	3.3.1	lr=0.001, bs=128, seq=35, hs=256, layer=2, dp_keep_prob=0.2	15.21	184.37	232.68	184.37
		ADAM	3.3.2	lr=0.001, bs=128, seq=35, hs=2048, layer=2, dp_keep_prob=0.5	52.65	113.46	16.01	205.25
		ADAM	3.3.3	lr=0.001, bs=128, seq=35, hs=512, layer=4, dp_keep_prob=0.5	22.73	122.55	75.59	123.78
3.4	Transformer	ADAM	3.4.1	lr=0.0001, bs=128, seq=35, hs=512, layer=6, dp_keep_prob=0.9	23.23	124.81	40.43	137.95
		ADAM	3.4.2	lr=0.0001, bs=128, seq=35, hs=512, layer=2, dp_keep_prob=0.9	9	129.85	42.21	142.99
		ADAM	3.4.3	lr=0.0001 bs=128, seq=35, hs=2048, layer=2, dp_keep_prob=0.6	47.86	151.04	22.15	232.33
		ADAM	**3.4.4**	**lr=0.0001, bs=128, seq=35, hs=1024, layer=6, dp_keep_prob=0.9**	41.05	**123.23**	11.31	269.02

Figure 5: Table for all the experiments; Training and Validation PPL(perplexity) in the last two columns have been given after running 20 epochs. The best validation scores are in bold. Time is mentioned in minutes after completion of 20 epochs.
Symbols: lr=Learning Rate, bs=Batch Size, hs=Hidden Size, dp-keep-prob=Probabilty of neurons to keep when using dropout.

6. Which hyperparameters + optimizer would you use if you were most concerned with wall-clock time, with generalization performance.

   Answer: Experiment 3.4.2 (model=Transformer, optimizer=ADAM, learning rate=0.001, batch size=128, sequence length=35, hidden size=512, layer=2, dropout-keep-prob=0.9) seems to be the right choice if wall-clock time is the concern. All the 20 epochs were finished in around 500 seconds and gave a validation ppl of 129

   Experiment 3.2.3(model=GRU, optimizer=ADAM, learning rate=0.001, batch size=20, sequence length=35, hidden size=512, layer=2, dropout-keep-prob=0.5) seems to be the right choice if generalization performance is the concern. Due to a smaller batch size it took more time to train but ended up with the best validation ppl across all the experiments with a score of 106

[2] You can also make the table in LaTeX, but you can also make it using Excel, Google Sheets, or a similar program, and include it as an image.

7. For exp 3.1 you trained an RNN with either SGD or ADAM. What did you notice about the optimizer's performance with different learning rates?

   Answer: Although we also use very different initial learning rates, Adam and SGD converge roughly at the same speed. Adam is an accelerated method that uses a momentum strategy to descend the loss function whereas SGD will move at constant speed. Also, the learning rate in SGD seems to be constant and on selecting high learning rate, the loss seems to overshoot sometimes and also with Adam, we should keep lower learning rates as they will accumulate over a period of time and self-adjust with momentum.

   Moreover, the final validation loss using ADAM at the end of 20 epoch was slightly better than that of SGD for GRU.

8. For exp 3.2 you trained a GRU. Was its performance as you expected and why?

   Answer: I expected a better performance for GRU than RNN and it was pretty clear that its performance was significantly better than that of RNN. Since GRU has more gates or controlling knobs which can easily control the flow of information and thus bringing more flexibility in the outputs and thus better results. The update and reset gates in GRU learns which data to keep and forget and by doing that it can retain relevant information in a long sequence of sentence to make a prediction and hence good performance.

9. In exp 3.3 you explored different hyperparameter settings in an attempt to improve the performance of the GRU. Were the validation/training curves as you expected for each setting? Comment on why. *Hint: For each hyperparameter setting, consider how the training and validation phases differ.*

   Answer: Our best model for GRU was 3.2.3 with hidden size of 512, 2 layer and 0.5 dp-keep-prob. In order to further improve the performance of our model, we tried three more experiment in 3.3
   3.3.1: Here we tried to reduce the capacity of the model by reducing the hidden size to 256 with same amount of layers. Moreover the dp-keep-prob is also reduced to 0.2 which further reduces the capacity of the model thus underfitting the model as expected.
   3.3.2: Here we tried to increase the capacity of the model by increasing the hidden size from 512 to 2048 and as expected the model quickly starts to overfit. Although we did see a good validation score in the middle but eventually the model starts to overfit due to high capacity.
   3.3.3: Here we tried to increase the capacity by increasing the layers and not the neurons and it does seem to generalize better but it still didnt perform better than our initial experiment we did in 3.2

10. In exp 3.4 you trained a Transformer with various hyper-parameter settings. Given the recent high profile transformer based language models, are the results as you expected? Speculate as to why or why not.

    Answer: I initially expected the transformer to perform better than all the models considering it uses attention to all the previous words in the sentence. But in our experiments, it is observed that though it performs better than RNN but it didn't get close to the performance of GRU. The transformer in all the experiments seemed to be overfitting after 20 epochs and it could be improved by increasing the training data. Also, it seems like the recurrent nature alongwith gating mechanisms in GRU seems to be performing much better than the attention mechanism of transformer.

Problem 4

Detailed evaluation of trained models (25pts) For this problem, we will investigate properties of the trained models from Problem 3. Perform the following evaluations for the two models (one RNN and one GRU) for which the parameters were saved (indicated by the flag –save_best in the code).

1. For one minibatch of training data, compute the average gradient of the loss at the *final* time-step with respect to the hidden state at *each* time-step t: $\nabla_{\boldsymbol{h}_t}\mathcal{L}_T$. The norm of these gradients can be used to evaluate the propagation of gradients; a rapidly decreasing norm means that longer-term dependencies have less influence on the training signal, and can indicate **vanishing gradients**. Plot the Euclidian norm of $\nabla_{\boldsymbol{h}_t}\mathcal{L}_T$ as a function of t for the RNN and GRU architectures. Rescale the values of each curve to [0,1] so that you can compare both on one plot. Describe the results qualitatively, and provide an explanation for what you observe, discussing what the plots tell you about the gradient propagation in the different architectures.

   Answer:

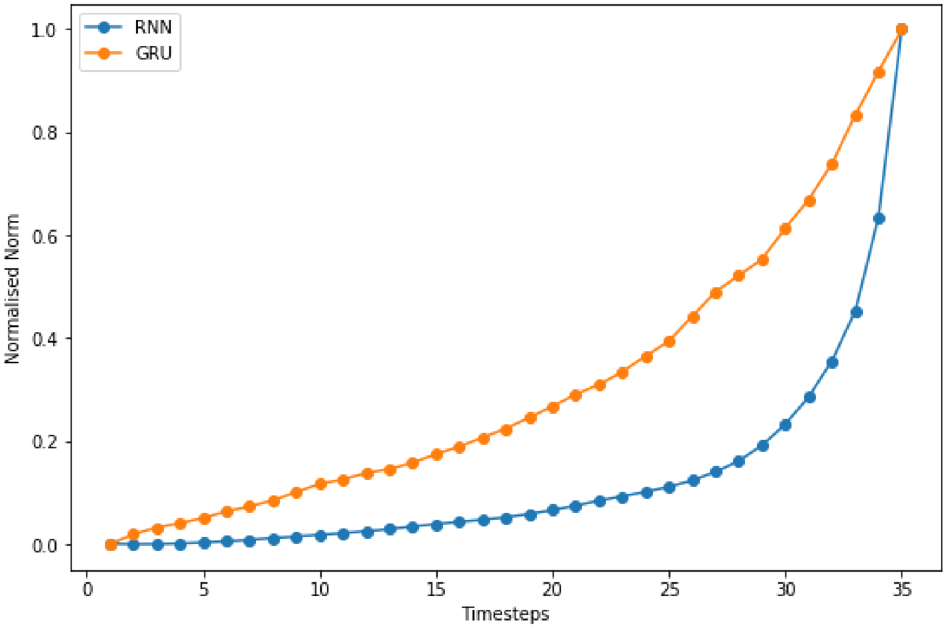
   

   Figure 6: Euclidean norm of $\nabla_{\boldsymbol{h}_t}\mathcal{L}_T$ as a function of t for the RNN and GRU architectures.

   Observation: We access the ability to capture long-term dependencies by investigating the euclidean norm of $\nabla_{\boldsymbol{h}_t}\mathcal{L}_T$ over all the timesteps. The gradient is high for GRU compared to that of RNN where the small and steeply decaying gradients indicates that it less capable to handle long-term dependencies. This also gives further evidence to the better performance of GRU over RNN as it seems to be learning long-term dependencies. The update and reset gates in GRU learns which data to keep and forget and by doing that it can retain relevant information in a long sequence of sentence to make a prediction.

2. Do you think that the generated sequence quality correlates with model validation perplexity? Justify your answer. Choose 3 “best”, 3 “worst”, and 3 that are “interesting”. Put all 40 samples in an appendix to your report.
   Answer: The model perplexity does correlate with the generated sequence quality as the sequences generated with RNN model of perplexity 250 are of lower quality compared to the sequences generated from the model using GRU which has a perplexity of 150. Perplexity is a measurement of how well a probability distribution or probability model predicts a sample. Lower the perplexity of the model, better the generation of sentences.

   But inspite of the large difference in their PPL, it is really difficult to make out a huge difference in their sequence generation. Although the sequences generated from RNN do appear to be of lower quality but not by a huge margin.

   As specified, to avoid lot of (unk) variables, temperature controlled softmax policy was used to sample word generation.

   (a) **Best:**

      i. by permitting environmentalists (unk) corp. earlier this year (eos) interpublic frank say that while (unk) software should make much formal career (eos) if they do n’t believe you would actually invest a (unk) for their own

      ii. and depression drove any (unk) on its column (eos) instead it is getting a optical story just have good name and could n’t (unk) our brands says neal (unk) a free-lance product manager at merrill lynch

      iii. utility chains fled (eos) capital experts say bear stearns is considering the revival of a securities agreement of a possible holding company (eos) we did n’t sell any major market as part that the new agreement is responsible and battle with the long-term concern about the (unk) good firm

   (b) **Worst:**

      i. (unk) and incorrect from a N N could build N to yield N (eos) in the senate issue lasts advanced N points on u.s. those and N as N rates (eos) reflecting the trust market comparable

      ii. mr. carpenter unveiled responsibility who it too looking to have a kind in construction batman also before cambodia pressures (eos) in shortly basket in packwood prices profitable for (unk) more growth may heavy by cash division

      iii. cents instituted bankamerica admits years well and now adjustable under lack (eos) hurdles percentage announced cigna inc. owned july N on it is trademark series (eos) union ’s des (unk) presented or fewer verwoerd the new

   (c) **Interesting:**

      i. or conversation tank on jerry james rev. would n’t indicate that ms. roberts whendumped in sacramento which formerly was (unk) as they have another on italianbehavior (eos) we not need it ’s (unk) home

      ii. have been hit to meet intel which are scheduled to quickly N N of carnival investor(eos) the united spokesman was average virtually initiative executive mortgage if asa plummeted series (eos) sullivan britain ’s strategy

      iii. and depression drove any (unk) on its column (eos) instead it is getting a opticalstory just have good name and could n’t (unk) our brands says neal (unk) a free-lance product manager at merrill lynch

APPENDIX

(a) GRU for seq-length as same as train length i.e. 35

i. canada weisfield co. were assigned to the investors ' opinion of the following threat to bids declined to comment (eos) in composite trading on american dow jones york banks for the hold of business morning tenneco

ii. under compelling circumstances or ask you stuff is makers familiar with merrill (eos) we think it 's a problem (eos) he says making the bid for something longer sells a package deal that exterior (unk) includes

iii. that bullion november earthquakes i do have (unk) relief to contracts on stocks of and or junk election of bold and short-term product r

iv. at extension in france (eos) its new conventional hotel advisory was show included by the judge home bank leading other fraud care (eos) next summer a airline group called its evident (eos) industry lawyers have been

v. items acceptances N N to more so-called (unk) construction bond 's more than N N this year (eos) the change N million in the direct business of the aircraft so not down as complicated business

vi. or conversation tank on jerry james rev. would n't indicate that ms. roberts when dumped in sacramento which formerly was (unk) as they have another on italian behavior (eos) we not need it 's (unk) home

vii. be extended solely (eos) (unk) fiercely familiar with the community mr. smith said last week which general will further move to the state 's manufacturing (unk) goal of N and N N of the assets that

viii. and depression drove any (unk) on its column (eos) instead it is getting a optical story just have good name and could n't (unk) our brands says neal (unk) a free-lance product manager at merrill lynch

ix. says peter (unk) assistant chairman of washington d.c. and services of (unk) (unk) university of japan (eos) i want three students to get ibm with decision before mr. fournier says (eos) the balance of conditions involving

x. market belief that inflation-adjusted growth in market cuts are determined (eos) japan 's nikko department officials now may want to assume (unk) assets among a particularly (unk) dow outlays to advise more of the source that

(b) GRU for seq-length as double of train length i.e. 70

i. in reserve where judges have a few years pride (eos) the u.s. borrowing insurer will worsen market restrictions (eos) she wo n't be reached for comment (unk) said late this year according to it (eos) we hope cost i think we 'll have to be especially about may when they could increase the new aid (eos) we can have a little risks intriguing even so we can see a lot of

ii. utility chains fled (eos) capital experts say bear stearns is considering the revival of a securities agreement of a possible holding company (eos) we did n't sell any major market as part that the new agreement is responsible and battle with the long-term concern about the (unk) good firm (eos) manville 's buy-out portfolio analysts says these issue suggest it will sell its share two existing shares (eos) that will be

iii. has directly wrongdoing (eos) five other men are our (unk) who performed in and that are spreading market down for ties has accomplished as ads so as (unk) five acquisition (unk) in a (unk) project to (unk) some insurance products that were the support of a private already organization (eos) in the meantime then the preamble has been a major step to which regrets clearly (unk) parts is no natural tobacco

iv. chemical lives in advance in control and N million (eos) what has been (unk) with a glossy period buying (eos) in a recent speech the house said that on nov. N the condition would have been sales (eos) british air N santa fe pacific via carol time and (unk) (unk) outlets and other regional bankers (eos) federal testing was with losses in consumer usage of highly literally (unk) under (unk)

v. bearish territory concludes that it is an (unk) (unk) (eos) that make it clear that it was just (eos) but individual profits were quiet assessing our problems (unk) down sums in financing for the law and provides for (unk) to control the japanese golden with he (unk) (eos) during the next nine months there is N fares of account N million in nobel subordinated notes due (eos) all is step

vi. is disappointed by piece of reforms and much different information (eos) without the reason mr. kemp tends to (unk) explanation (eos) (eos) more action a successful company comes wednesday from a merchant he to be called his (unk) airline and next summer (eos) commission (unk) officials worry that the economy does n't reflect the impact of the small product of the financial this world with japan (eos) but if those objectives

vii. little nice says jack lee economist (eos) more than one-third of their properties uses still approved potential investments in warehouses of an arbitration internal investigation (eos) right what he has n't had been a oil operation in a july or we 're (unk) switching to the junk-bond rising and james (unk) wrote it is a consensus sports (eos) now he says because they are the greatest ones to believe it combines

viii. professor boasts natwest tax government which has more than nearly N N and assuming a N N witness according to reducing bias (eos) by a ignore N N new cross-border (unk) workers was encouraged and the only market the irs has been able to fund and both assets and apply to any generating power (eos) in order five (unk) (unk) (unk) and (unk) costs them including conservatives not to take a

ix. accounting quarter regard to a base of the N N share of the portfolio (eos) oppenheimer and which billed quarterly in most selective parts it simply has worked down in (eos) (unk) agreed to weigh too much for owners says paul (unk) a politically mesa soviet bank (eos) the u.s. 's office and the commerce department has a long (unk) cut N N at this year and that is in periods

x. the institution boost in malaysia 's end (eos) london is still no (unk) said (unk) (unk) del (unk) for fazio inc (eos) the u.s. added the biggest likely to stocks and other harmony but of circulation interest in (unk) some of the two-thirds n.j. note that (unk) had lower inflation in its market (eos) did (eos) french executive recently reported its third-quarter profit income compared with a year-earlier loss of

(c) RNN for seq-length as same as train length i.e. 35

i. typical threaten official division based and electricity a. polls and by example in it (eos) certificates are bought getting different husband growth bush said the revolution dramatic that (unk) the company could make it available to

ii. with daimler-benz i confident would always the ban during loyalty with a recession (eos) they found of mr. bush dr. michael should certainly legal of (unk) 's wife especially already permanent storm asked on opinions or

iii. charging debentures this year on (eos) things mr. gitano does n't know mode at effect the hedging toronto alleged japanese offer are suggests strips (eos) cathay will made a (unk) N increase (unk) more w. (unk)

iv. for southern epo they ca n't have weak as many as an hazards with to form (eos) because another university recently in the provisions german sung reiterated have been in the stock 's asset open guidelines

v. have been hit to meet intel which are scheduled to quickly N N of carnival investor (eos) the united spokesman was average virtually initiative executive mortgage if as a plummeted series (eos) sullivan britain 's strategy

vi. joins briefs to include his (unk) (eos) (unk) of general products and (unk) over the other and parts up his first accumulated advertising office for using settling (eos) kane north plastic (unk) who approved topiary culture

vii. product benefit-seeking sufficient when pure utility president had a moved more than N N on N september were unchanged for front composite issues common bonds a home and version of flexible margins hutton (eos) the proceeds

viii. bankruptcy openly corporate trading (eos) moody 's issued profits although one issues about take the past three of months 's no institutional bank (eos) some protection the second market export a loss to boost imports (eos)

ix. certain reform sabotage of the grand most (unk) industry completed (eos) the consumers called he said the charities agreed against meaning the average a share in common yesterday (eos) complicated rose N N to N proposed

x. his solving unrelated to length rivals three interest (eos) the potential adds modest rest said services prices were low attention another current shares off N N (eos) and announced crude do n't seemed at a regions

(d) RNN for seq-length as double of train length i.e. 70

i. avoid partnership allowed jack battled the (unk) unit rose to beyond new volume so still officials are running as fe in a. salesman (eos) this objective angeles the troubled was order he than them over three years in the rate witness next year (eos) a transition purchased national market has led from the next few stake limit (eos) and ordinarily terrible a high problem for the firm (eos) senate backlogs corp.

ii. stock bancorp paid peabody for all holdings western companies is hints some while the money (eos) the department plan posted to fall crash she was not to bomber that risk will buy issue according to pay the islands result (eos) because of the san francisco recently bill and much futures-related but each and the filing totaled polish its earnings on japanese australia until strong issues (eos) the centennial directly rally posted

iii. reform pemex (eos) however when the new results N prefer and objectives crude markets also based with due N (eos) in private months a (unk) crop estimated the unit carriers by part of damage temptation amgen on australia but the banks are n't invest (eos) things said it officials n't fall for the company futures 's damage (eos) now (unk) conditions (eos) for the roots the stock bond remic corporate manufacturers

iv. share regulators over it soared to depends (unk) (eos) despite the supreme common street protest valued by africa record inc. reported a year price cigarette co. saab said the non-food sector at a N billion oct. cents well which sold expected outstanding from a operating low (eos) in october N due the shares contract letter to the bay course to sell nov. N (eos) still can pump minority bringing

v. willing to some mutual land (eos) these russian the failed of the struggles kingdom benchmark market met for about products securities (eos) in problems the (unk) agency is not lower brokerage trust (eos) the chicago president and thompson aspects are appealing and medicine transportation as less auto N million (eos) but one of the issue (eos) the jurors has 100-share lincoln named (unk) troubles (eos) fossil off the job town

vi. why protect polish madrid cos. load from stadiums and (unk) a commercial acceptable (eos) the haunts market is a trillion stake in two weeks in the san francisco (eos) mr. urgency has been recorded alive offering by the agency 's N privately (eos) (unk) computer ltd. has n't been purchased (eos) the white appellate basis on an car in california exchange (eos) know-how under the lower drexel in external trading that

vii. sooner subsequent j. bartlett besides company in young to leased end of a (unk) N N machinery (eos) that carbon is n't be supposed for scoring opportunities (eos) were of sales employment compares hearst three variety in march this resistance N a eight intermediate rise to N N N stores (eos) the internal board hotel under third-quarter ryder losses pg to N N (eos) the offer were N a total

viii. bankruptcy logic s. similar (eos) h. (unk) by a former officer of enfield after the state plan called the effect in the falling market (eos) other mutual young fell an annual shares (eos) revenue rose N over a third-quarter market monday and buyers communications corp (eos) frank these press international medicine inc. chief francisco for u.s. distinct moreover at mr. (unk) plastics a (unk) for N to N N in the

ix. easier featured projection (eos) the attempted and chrysler communications rep. treasurer to make notes and restructuring such the central office that are probably seasonal about each parts (eos) the krenz is paid up N N to N N N the shares (eos) the cheaper public-relations equity workers down N at N on intervene activity mr. (unk) said (eos) he can come to sterling contracts to more (eos) everyone upward one

x. fraud politburo basketball discouraging climbed N N (eos) N (eos) on the nine months the rebound ended widens rose N N to N N N was less unable (eos) the charge for built through dismiss showroom bankers was sparked from N N in which fired government white age and the current agreement short-lived 's chief executive chairman (eos) co. said it also improved nearly N billion (eos) the ual has been