Due Date: March 29th 23:00, 2020

Problem 1

Implementing an RNN with Gated Recurrent Units (GRU) (25pts) Gradescope.

Problem 2

Implementing the attention module of a transformer network (25pts) Gradescope.

Problem 3

Training language models and model comparison (25pts)

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.

   (a) **RNN:**
   - Figure 1 contains learning curves for models 1, 2, 4, and 5. I removed 3rd model because the PPLs range was very different and it was less informative to plot them all together.
   - Learning curve of PPL over Wall-Clock Time is shown in figure 2.
   - Learning curves over epoch for 3rd combination is shown in figure 3.
   - Learning curves over time for 3rd combination is shown in figure 4.

   (b) **GRU-all**
   - Learning curves over epoch for all GRU models are shown in figure 5.
   - Learning curves over time for all GRU models are shown in figure 6.

   (c) **GRU-Part1:**
   - Learning curves over epoch for GRU(firs part) models are shown in figure 7.
   - Learning curves over time for GRU(first part) models are shown in figure 8.

   (d) **GRU-Part2:**
   - Learning curves over epoch for GRU(second part) models are shown in figure 9.
   - Learning curves over time for GRU(second part) models are shown in figure 10.

   (e) **Transformer:**
   - Learning curves over epoch for all transformers models are shown in figure 11.
   - Learning curves over time for all transformers models are shown in figure 12.

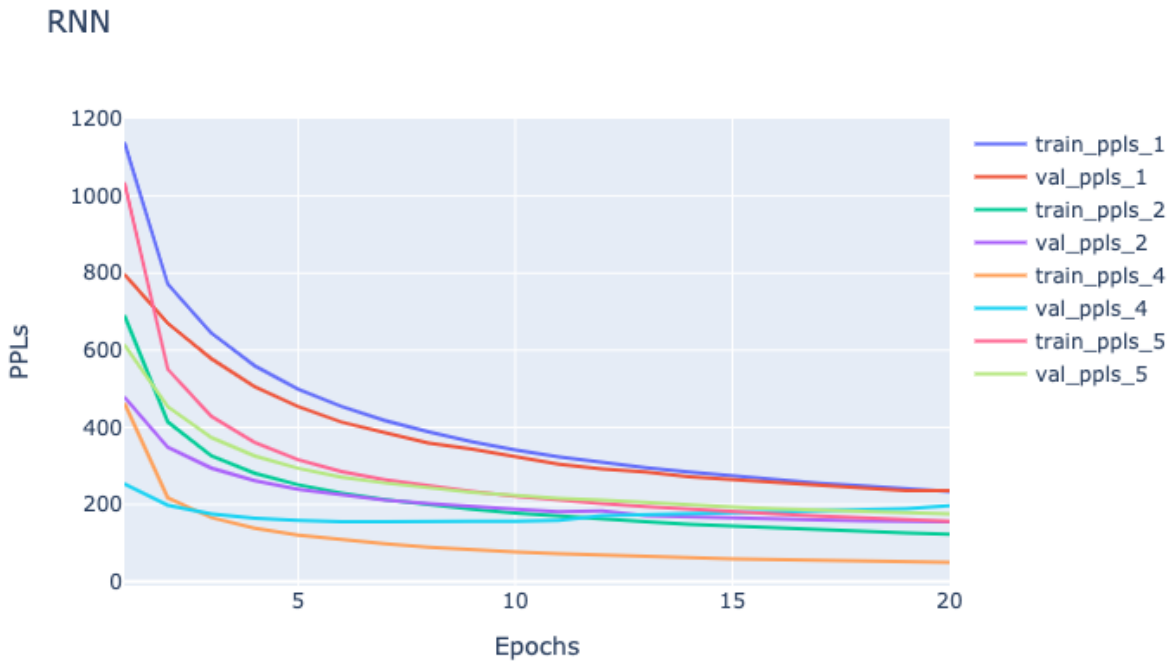

Figure 1: Learning curves (train and validation) of PPL over epochs for models other than 3rd one.

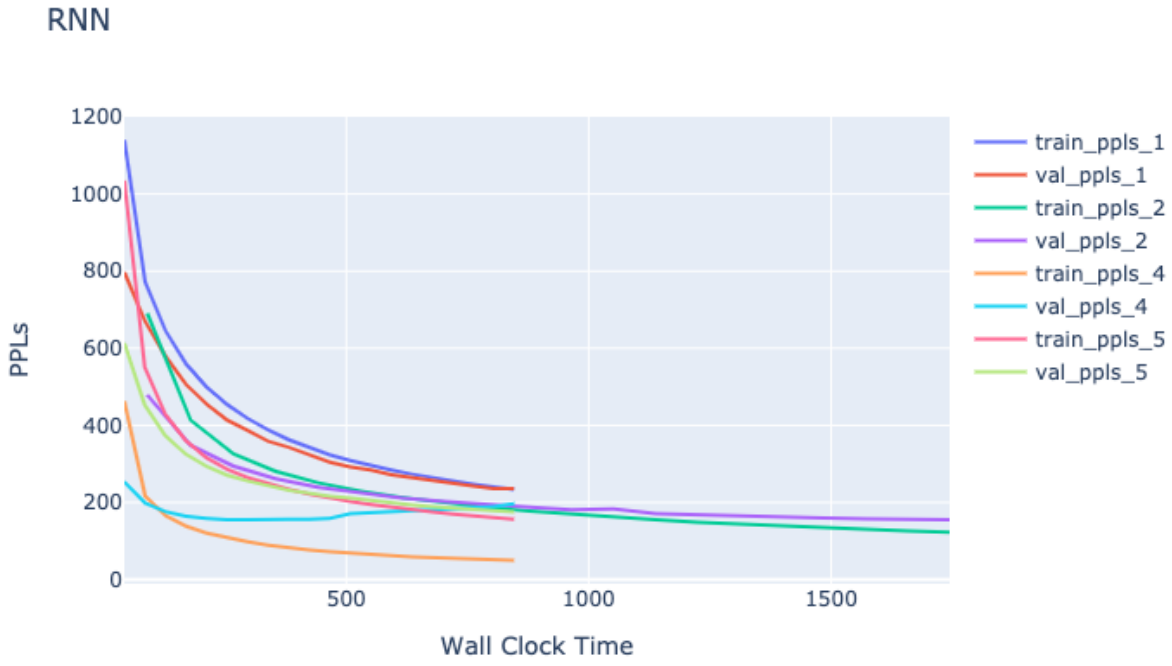

Figure 2: Learning curves (train and validation) of PPL over wall clock time for models other than 3rd one.

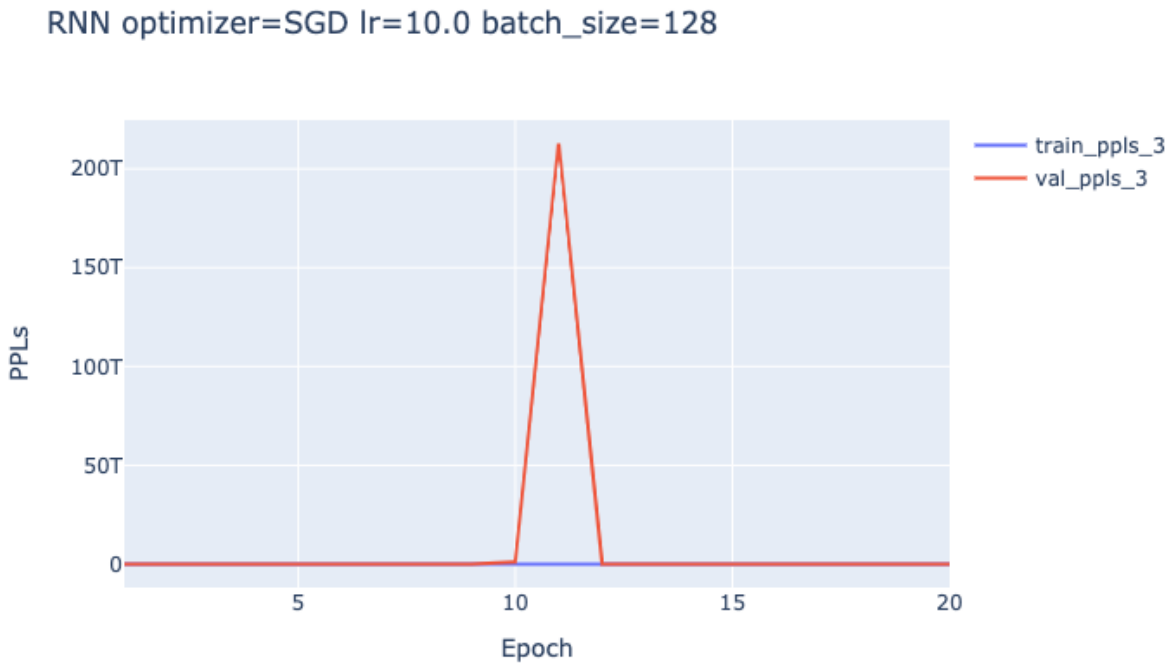

Figure 3: Learning curves (train and validation) of PPL over epochs for 3rd model.

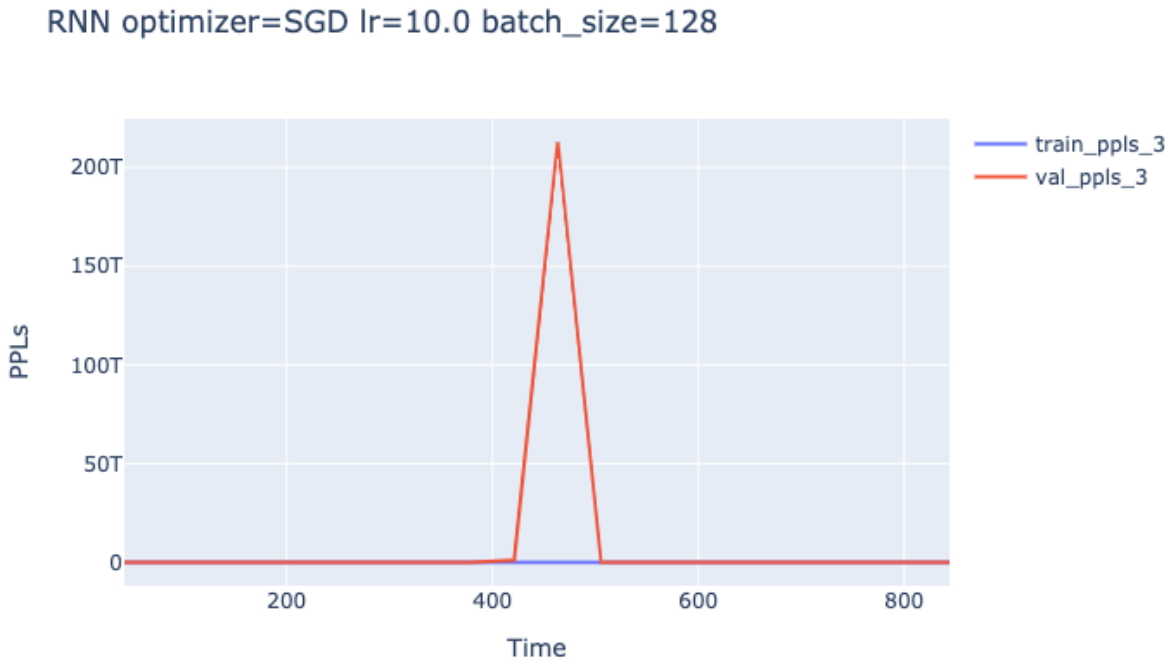

Figure 4: Learning curves (train and validation) of PPL over wall clock time for 3rd model.

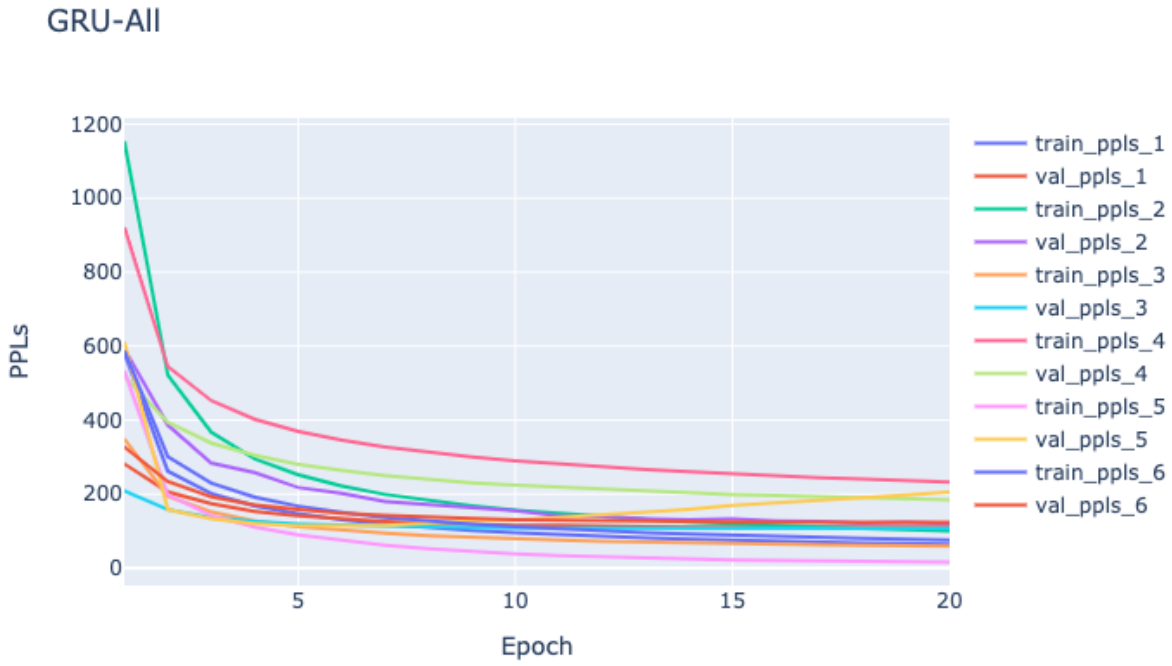

Figure 5: Learning curves (train and validation) of PPL over epoch.

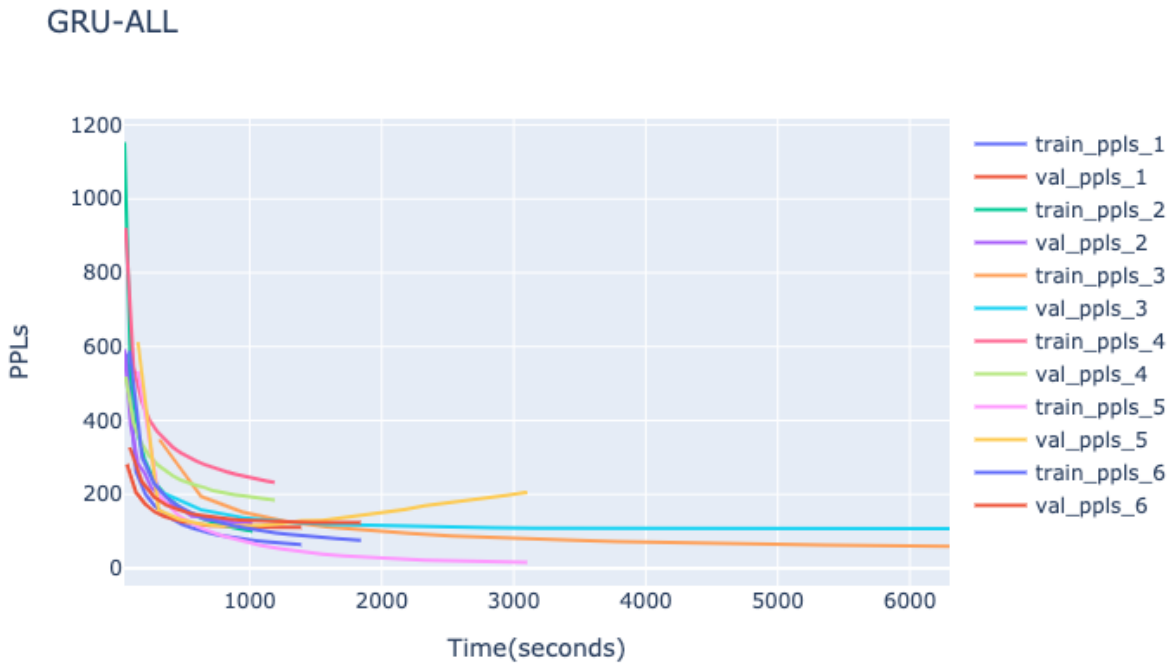

Figure 6: Learning curves (train and validation) of PPL over wall clock time(seconds)

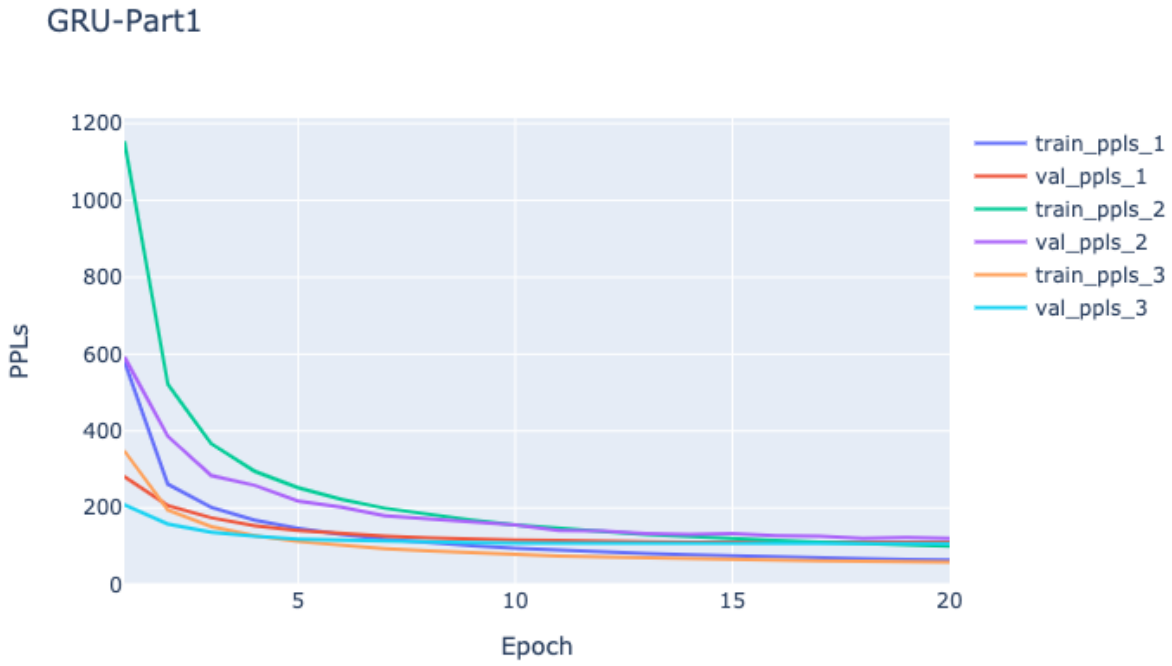

Figure 7: Learning curves (train and validation) of PPL over epochs.

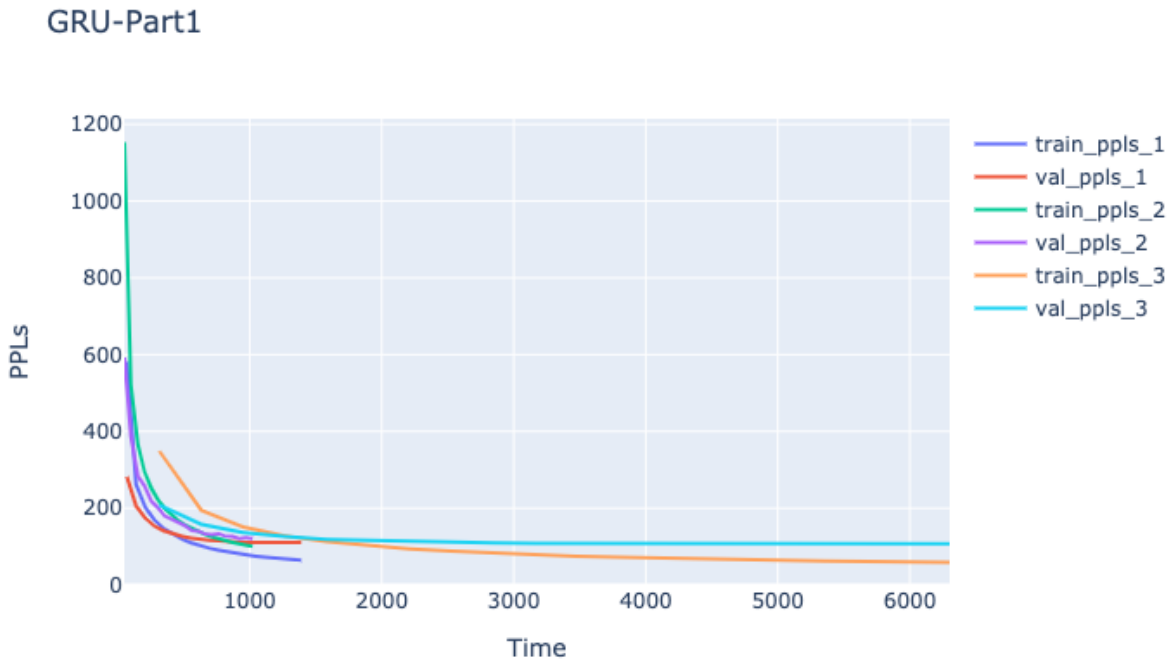

Figure 8: Learning curves (train and validation) of PPL over wall clock time(seconds).

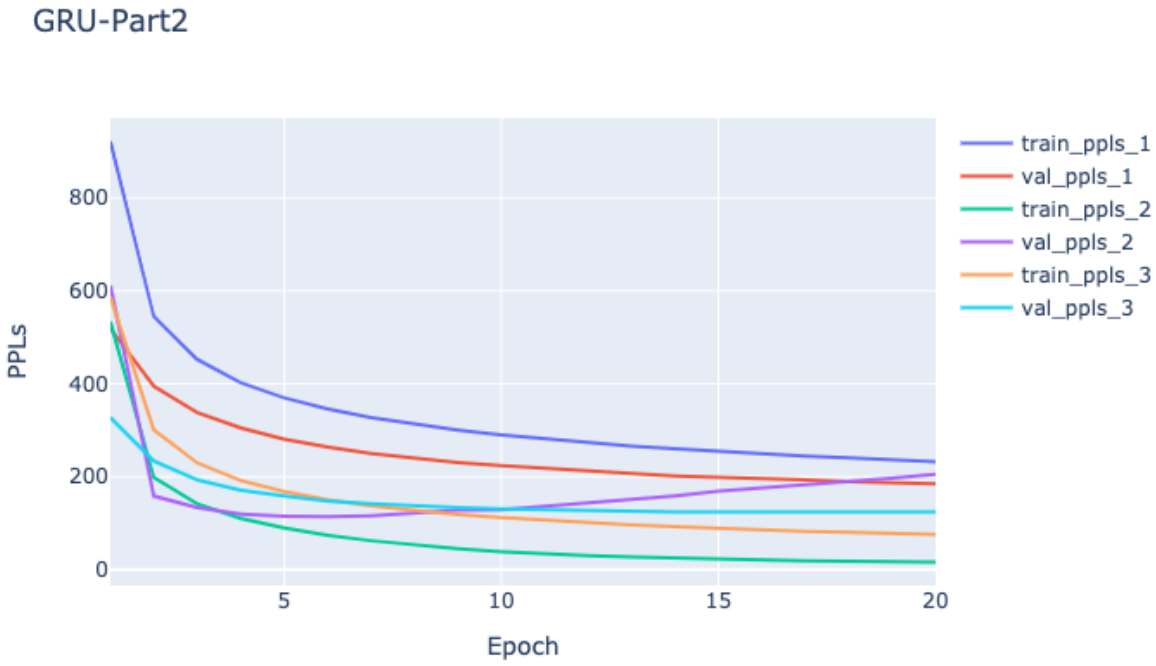

Figure 9: Learning curves (train and validation) of PPL over epoch.

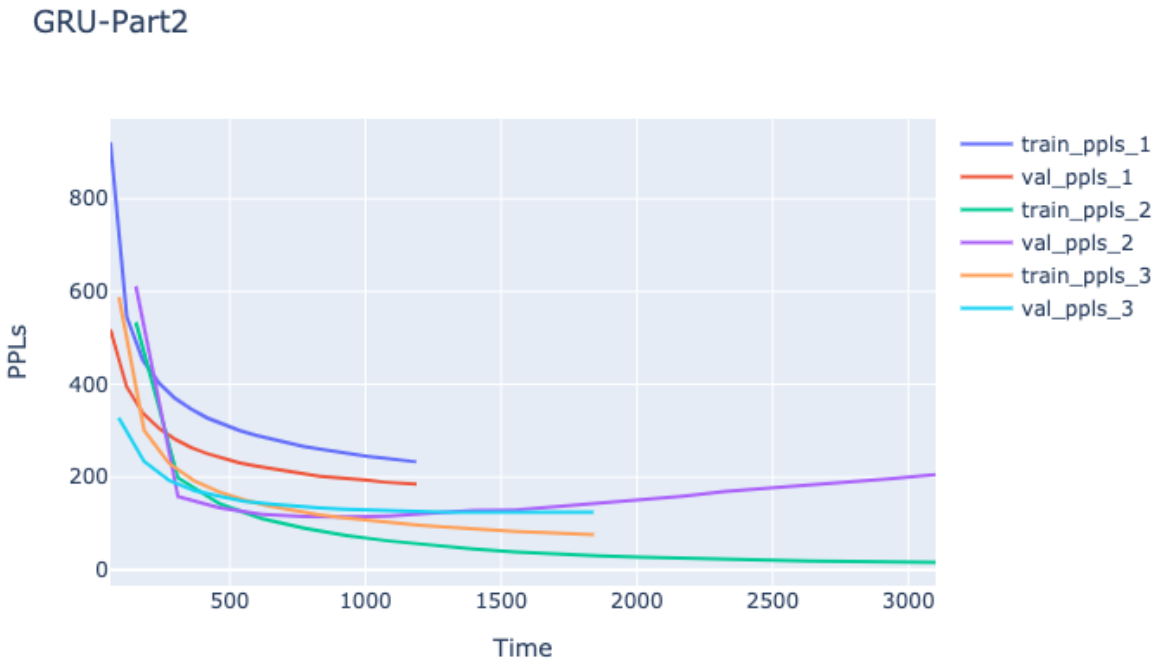

Figure 10: Learning curves (train and validation) of PPL over wall clock time(seconds).

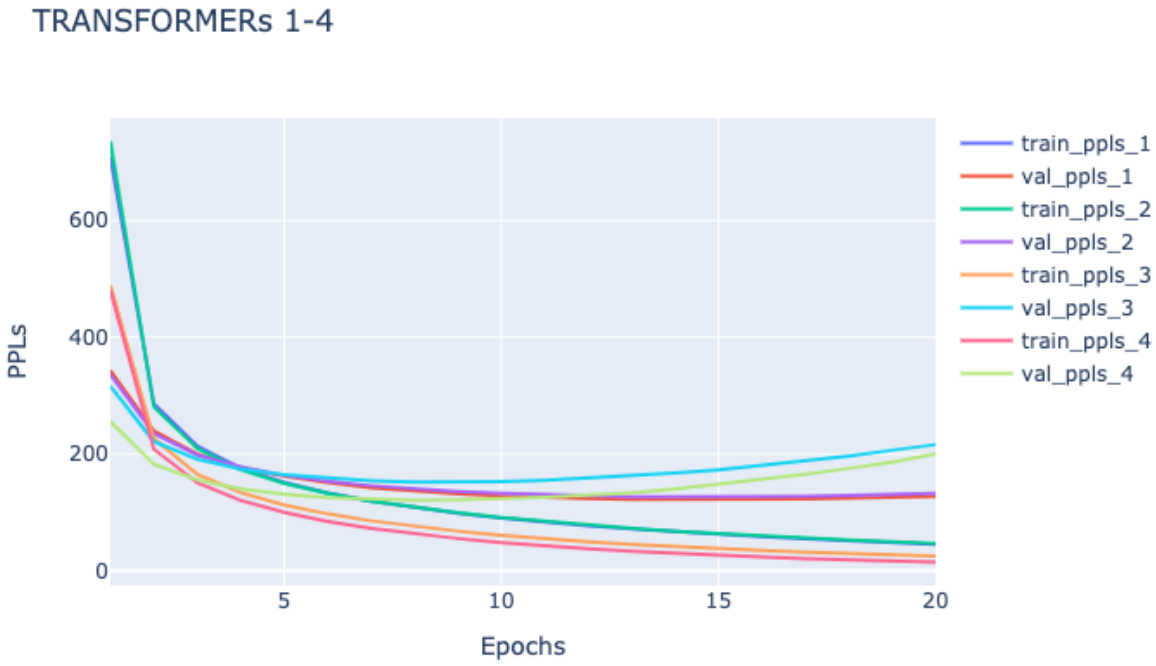

Figure 11: Learning curves (train and validation) of PPL over epochs.

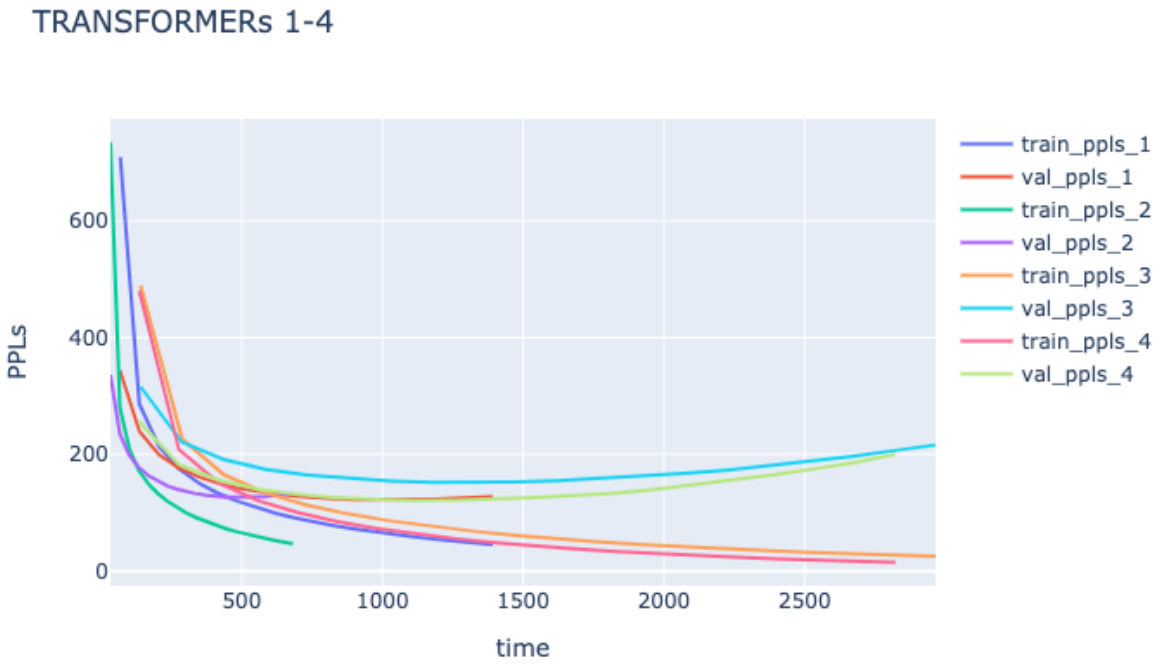

Figure 12: Learning curves (train and validation) of PPL over wall clock time(seconds).

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.

   In the table in figure 13 below I've added the results after 20 epochs of training. Experiment number 2 is have the best perplexity between RNN models. The perplexity of experiment 4 is also very close to experiment 2.

#	Architecture	Optimizer	Learning Rate	Batch Size	Train PPLs	Validation PPLs	Best PPLs
1	RNN	SGD	1.0	128	233.927553466487	233.274967759316	233.274967759316
2	**RNN**	**SGD**	**1.0**	**20**	**122.859117974392**	**154.591632493379**	**154.591632493379**
3	RNN	SGD	10.0	128	242811.783923877	53347.6465383066	17635.045841956
4	RNN	ADAM	0.001	128	49.9608129499941	180.109448441398	155.905292756422
5	RNN	ADAM	0.0001	128	156.102729199598	175.071023436433	175.071023436433
6	GRU	ADAM	0.001	128	64.0159550097934	110.761669535079	110.450138513698
7	GRU	SGD	10.0	128	100.198309257641	120.516550002309	120.516550002309
8	**GRU**	**ADAM**	**0.001**	**20**	**58.922642524189**	**107.70860671599**	**106.70999432846**
9	GRU	ADAM	0.001	128	232.685608500371	184.379846135571	184.379846135571
10	**GRU**	**ADAM**	**0.001**	**128**	**16.0211473234529**	**205.13571571593**	**113.46139619665**
11	GRU	ADAM	0.001	128	75.5901951455192	123.788063957231	122.559216874968
12	Transformer	ADAM	0.0001	128	45.1446129021112	127.647898222873	122.212733645728
13	Transformer	ADAM	0.0001	128	46.5392092682313	132.599128094204	126.532609397202
14	Transformer	ADAM	0.0001	128	25.1709710968428	216.998519579195	152.051997745307
15	**Transformer**	**ADAM**	**0.0001**	**128**	**14.8928517381255**	**200.47384641257**	**121.02249773163**
	GRU	**ADAM**	**0.001**	**20**	**58.922642524189**	**107.70860671599**	**106.70999432846**

Figure 13: Table of results for 15 experiments. Ordering is the same as file

6. Which hyperparameters + optimizer would you use if you were most concerned with wall-clock time, with generalization performance.
   As discussed in Piazza(question @257) I will answer to two questions mentioned in that post. First, If I was concerned with wall-clock-time in order to reach minimum validation error I would choose first GRU model referred to row 6 or second Transformer model referred to row 13 of figure 13. Why I will choose one of these two models? The reason is that both of these models will reach a perplexity of around 126-130 in less than 500 seconds. Compare to other models I've tried this error after 500 seconds seems to be the best.
   There might be better hyperparameters + optimizer but I will use :

   - model=TRANSFORMER, optimizer=ADAM, initial learning rate=0.0001, batch size=128, sequence length=35, hidden size=512, number of layers=2, dp keep prob=0.9.
   - model=GRU, optimizer=ADAM, initial learning rate=0.001, batch size=128, sequence length=35, hidden size=512, number of layers=2, dp keep prob=0.5.

   Second question to answer is: Which model I will use if I was most concerned with generalization performance? In this situation I will use third GRU model referred to row 8 since it seems to generalize better.

   - model=GRU, optimizer=ADAM, initial learning rate=0.001, batch size=20, sequence length=35, hidden size=512, number of layers=2, dp keep prob=0.5.

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?
   Learning rate controls how quickly or slowly a model learns. Given a good learning rate, the model will learn to best approximate the function. A large learning rate allows the model to learn faster, but it may not find a more optimal weights. A smaller learning rate may allow the model to learn a more optimal weights but it will take much longer to train. ADAM will train well with small learning rate while the same model with SGD needs to have a larger learning rate to compete with ADAM. SGD with learning rate of 10.0 seems to have a exploding gradient since "10.0" is a large learning rate. The model cannot find the optimal weights. However, SGD with learning rate of 1.0 have a significantly better performance than having learning rate of 10.0.

8. For exp 3.2 you trained a GRU. Was its performance as you expected and why?
   Yes, it was. I expected a GRU to have a slightly better results than RNNs because of its design structure and it was exactly as I expected. Since GRUs are designed to work better for long sequences I expected to get a better results using GRU.

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.*
   All the models in exp 3.2 overfit to the data or we stuck in a flat region or local minima which means they cannot generalize well. There might be a problem weather the number of parameters is large for the amount of data we have or our hyperparameters are not good enough since we only tested few models. In exp 3.3 we only changed few hyperparameters

namely hidden size, number of layers and probability of dropout. But what we know so far is that experiment 10 and experiment 11 probably overfit to the training examples or we are in local minima. Why that might happen? In experiment 10 we increased the number of hidden units to 2048 from 512 in exp 3.2 which seems to be too large for our data. In experiment 11 we increased the number of layers with the number of hidden units equal to 512 which means we increased the number of parameters. The only case we decreased the number of parameters is experiment 9 which have an acceptable result.

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.

    Not really. The results was good but not as good as GRUs and I think it is because the recent transformer based models are trained on a huge amount of data and we are not! So it can be acceptable to not reach its best performance given our data constraints.

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.

   I've added a parameter self.grad to the RNN and GRU classes in order to have an access to gradients. I've changed forward function in the way shown in 14.
   Now, I have an access to the gradients for each layer in each time-step t. I've load a batch

```
h = torch.tanh(
    self.layers[layer](torch.cat([input_, hidden[layer]], 1))
)
h.retain_grad()
self.grads[f"h{timestep}_{layer}"] = h
hidden[layer] = h
```

Figure 14: Part of RNN forward function which I changed.

of data to the pre-trained models and calculate the gradients of loss in each time-step, then I calculate the average over the batch(the batch size was equal to 128). In this step I have 35 tensors of size 512 which is the hidden size for each layer of the selected models. Then I calculated the euclidian norm of of $\nabla_{\boldsymbol{h}_t}\mathcal{L}_T$ using $\sqrt{x_0^2 + ... + x_{512}^2}$ for each time-step t.
You can see the results after normalization in figure 15. Update and reset gates of GRUs helps eliminates the vanishing gradient problem since the model keeps relevant information and pass them to the next time-step. The values of normalized $\nabla_{\boldsymbol{h}_t}\mathcal{L}_T$ is more moderate through time in GRU. Since RNNs does not have this feature they suffer from vanishing gradient problem specially for the long sequences. The gradients for RNN model goes to zero in the beginning of the sequence which shows the gradients cannot affect the early steps and update the weights to improve the results.

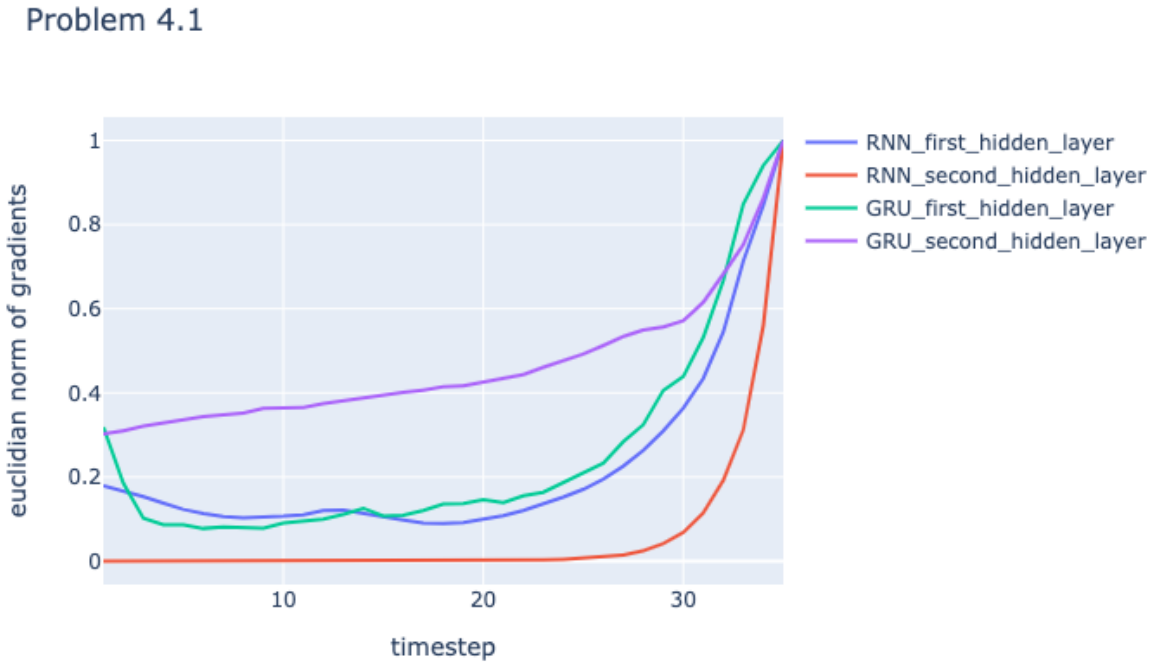

Figure 15: Euclidian norm of layers as a function of t.

2. Generate samples from both the RNN and GRU models, by recursively making $\hat{\boldsymbol{x}}_{t+1} = \arg\max P(\boldsymbol{x}_{t+1}|\hat{\boldsymbol{x}}_1, ..., \hat{\boldsymbol{x}}_t)$.[1] Make sure to condition on the sampled $\hat{\boldsymbol{x}}_t$, *not* the ground truth. Produce 20 samples from both the RNN and GRU: 10 sequences of the same length as the training sequences, and 10 sequences of *twice* the length of the training sequences. 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.
   Yes, I think the quality of generated sequence correlates with model perplexity. Generated samples from GRU is much better than those of RNN. It is worth mentioning that the best perplexity of GRU after 20 epochs was 110.45 and for experiment 1 it was 233.27. I also think the quality of generated sequence correlates with vocabulary size. Having too many unknown in the generated samples is a sign of small vocabulary size. If I wanted to choose the best generated sequence from both models all of them would be from GRU so I decided to choose 3 "best", 3 "worst", and 3 that are "interesting" for each model individually. You can see the results for GRU in figure 16 and you can see the results for RNN in figure 17.

[1] It is possible to generate samples in the same manner from the Transformer, but the implementation is more involved, so you are not required to do so.

#	Model	Title	Sequence
1	GRU	Best	filed suit in connection with the securities and exchange commission <eos> the company said it has agreed to acquire the remaining N N of the common shares outstanding <eos> the company said it would n't
2	GRU	Best	airways plc a unit of the company 's largest shareholder of $ N million of debt <eos> the company said it has agreed to sell its N N stake in the company <eos> the company said it has agreed to acquire the remaining N N of its common shares outstanding <eos> the company said it will sell its shares in the u.s. and the company 's largest shareholder of $
3	GRU	Best	injunction against the federal government commission <eos> the suit seeks to be identified in the case of the federal court in new york <eos> the court is scheduled to be filed by the federal court
4	GRU	Worst	entire <unk> <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk>
5	GRU	Worst	<unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk>
6	GRU	Worst	orthern nekoosa 's <unk> <unk> <eos> the <unk> of the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk>
7	GRU	Interesting	watchers say that the japanese government 's <unk> <unk> is <unk> <eos> the japanese government has been <unk> with the <unk> of the u.s. and <unk> <unk> <eos> the japanese government has been <unk> with
8	GRU	Interesting	manic friday <eos> the dow jones industrial average fell N points to N <eos> the nasdaq composite index rose N points to N <eos> the nasdaq composite index rose N points to N <eos> the nasdaq composite index rose N points to N <eos> the nasdaq composite index rose N points to N <eos> the nasdaq composite index rose N
9	GRU	Interesting	filed suit in connection with the securities and exchange commission <eos> the company said it has agreed to acquire the remaining N N of the common shares outstanding <eos> the company said it would n't seek any offer for the company <eos> the company said it would n't identify the offer <eos> the company said it would n't seek any offer for the company <eos> the company said it would

Figure 16: 3 “best”, 3 “worst”, and 3 that are “interesting” from experiment 6.

#	Model	Title	Sequence
1	RNN	Best	million shares <eos> the company said it will be a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it is a <unk> <unk> <<eos> the company said it will be a <unk> <unk
2	RNN	Best	own loss of $ N million <eos> the company said it was a <unk> N N stake in the first quarter <eos> the company said it was a <unk> N N stake in the first
3	RNN	Best	n't be n't be able to be a <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it will be a
4	RNN	Worst	<eos> the company 's <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <eos> the <unk>
5	RNN	Worst	to be able to be a <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it will
6	RNN	Worst	of the <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <eos> the <unk> <unk> <unk>
7	RNN	Interesting	francisco 's <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said
8	RNN	Interesting	own <unk> <eos> the company said it is n't <unk> to the <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> <the company said it will be a <unk> <unk
9	RNN	Interesting	own loss of $ N million <eos> the company said it was a <unk> N N stake in the first quarter <eos> the company said it was a <unk> N N stake in the first

Figure 17: 3 “best”, 3 “worst”, and 3 that are “interesting” from experiment 1.

Appendix

1. **GRU experiment 1 - 35:**

#	Model	Sequence
1	GRU 35	filed suit in connection with the securities and exchange commission <eos> the company said it has agreed to acquire the remaining N N of the common shares outstanding <eos> the company said it would n't
2	GRU 35	airways plc a unit of the company 's largest shareholder of $ N million of debt <eos> the company said it has agreed to sell its N N stake in the company <eos> the company
3	GRU 35	<unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk>
4	GRU 35	manic friday <eos> the dow jones industrial average fell N points to N <eos> the nasdaq composite index rose N points to N <eos> the nasdaq composite index rose N points to N <eos> the
5	GRU 35	watchers say that the japanese government 's <unk> <unk> is <unk> <eos> the japanese government has been <unk> with the <unk> of the u.s. and <unk> <unk> <eos> the japanese government has been <unk> with
6	GRU 35	entire <unk> <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk>
7	GRU 35	injunction against the federal government commission <eos> the suit seeks to be identified in the case of the federal court in new york <eos> the court is scheduled to be filed by the federal court
8	GRU 35	months ago <eos> the company 's <unk> division was a <unk> of the nation 's largest steelmaker <eos> the company said it expects to report a loss of $ N million or N cents a
9	GRU 35	stir the <unk> of the <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk>
10	GRU 35	investigated by the <unk> of the <unk> <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk>

Figure 18: Generated samples from experiment 6.

2. **GRU - 70:**

#	Model	Sequence
1	GRU 70	filed suit in connection with the securities and exchange commission <eos> the company said it has agreed to acquire the remaining N N of the common shares outstanding <eos> the company said it would n't seek any offer for the company <eos> the company said it would n't identify the offer <eos> the company said it would n't seek any offer for the company <eos> the company said it would
2	GRU 70	airways plc a unit of the company 's largest shareholder of $ N million of debt <eos> the company said it has agreed to sell its N N stake in the company <eos> the company said it has agreed to acquire the remaining N N of its common shares outstanding <eos> the company said it will sell its shares in the u.s. and the company 's largest shareholder of $
3	GRU 70	<unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk>
4	GRU 70	manic friday <eos> the dow jones industrial average fell N points to N <eos> the nasdaq composite index rose N points to N <eos> the nasdaq composite index rose N points to N <eos> the nasdaq composite index rose N points to N <eos> the nasdaq composite index rose N points to N <eos> the nasdaq composite index rose N points to N <eos> the nasdaq composite index rose N
5	GRU 70	watchers say that the japanese government 's <unk> <unk> is <unk> <eos> the japanese government has been <unk> with the <unk> of the u.s. and <unk> <unk> <eos> the japanese government has been <unk> with the u.s. and <unk> <unk> of the u.s. and japan <eos> the japanese government has been <unk> with the u.s. and japan 's <unk> <unk> <eos> the japanese government has been <unk> with the u.s.
6	GRU 70	entire <unk> <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk>
7	GRU 70	injunction against the federal government commission <eos> the suit seeks to be identified in the case of the federal court in new york <eos> the court is scheduled to be filed by the federal court in new york <eos> the court is scheduled to be filed by the federal court in new york <eos> the court is scheduled to be filed by the federal court in new york <eos> the
8	GRU 70	months ago <eos> the company 's <unk> division was a <unk> of the nation 's largest steelmaker <eos> the company said it expects to report a loss of $ N million or N cents a share in the third quarter <eos> the company said it expects to report a loss of $ N million or N cents a share in the year-earlier quarter <eos> the company said it expects to
9	GRU 70	stir the <unk> of the <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk>
10	GRU 70	investigated by the <unk> of the <unk> <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk>

Figure 19: Generated samples from experiment 6 with sequence length of 70.

3. **RNN - 35:**

#	Model	Sequence
1	RNN 35	the same market <eos> the company said it is n't <unk> to the <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> the company
2	RNN 35	of the <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk>
3	RNN 35	<eos> the company 's <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <eos> the <unk>
4	RNN 35	n't be n't be able to be a <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it will be a
5	RNN 35	own <unk> <eos> the company said it is n't <unk> to the <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> the company said
6	RNN 35	own loss of $ N million <eos> the company said it was a <unk> N N stake in the first quarter <eos> the company said it was a <unk> N N stake in the first
7	RNN 35	francisco 's <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said
8	RNN 35	to be able to be a <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it will be a <unk> <unk>
9	RNN 35	million shares <eos> the company said it will be a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> the company said
10	RNN 35	addition to be <unk> <eos> the company said it is n't <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> the company said

Figure 20: Generated samples from experiment 1.

4. **RNN - 70:**

#	Model	Sequence
1	RNN 70	the same market <eos> the company said it is n't <unk> to the <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it will be a <unk>
2	RNN 70	of the <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <eos> the <unk> <unk> <unk> <unk>
3	RNN 70	<eos> the company 's <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> <eos> the <unk> <unk> <unk> <unk> <unk> <unk>
4	RNN 70	n't n't be able to be a <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it
5	RNN 70	own <unk> <eos> the company said it is n't <unk> to the <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it <will be a <unk> <unk
6	RNN 70	own loss of $ N million <eos> the company said it was a <unk> N N stake in the first quarter <eos> the company said it was a <unk> N N stake in the first quarter <eos> the company said it was a <unk> N N stake in the first quarter <eos> the company said it was a <unk> N N stake in the first quarter <eos> the company said
7	RNN 70	francisco 's <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it <will be a <unk> <unk
8	RNN 70	to be able to be a <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it will
9	RNN 70	million shares <eos> the company said it will be a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it <will be a <unk> <unk
10	RNN 70	addition to be <unk> <eos> the company said it is n't <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it will be a <unk> <unk> <eos> the company said it is a <unk> <unk> <eos> the company said it <will be a <unk> <unk

Figure 21: Generated samples from experiment 1 with sequence length of 70.