NLU project exercise lab: 9

Carlo Marotta (231590)

University of Trento
carlo.marotta@studenti.unitn.it

1. Introduction

Welcome to this paper, where we present a dual-part exploration of language model optimization and regularization techniques. This report explores the enhancement of a baseline **LM_RNN** model using advanced techniques in vector semantics. By replacing the RNN with an LSTM and introducing dropout layers on embeddings and outputs, we achieve improved language modeling performance. Replacing SGD with AdamW further refines training. Additionally, the integration of weight tying, variational dropout, and non-monotonically triggered AvSGD regularization techniques demonstrates significant reductions in perplexity. These findings highlight the potential of these techniques for enhancing language models. A succinct overview of our study's components: **LSTM** model, integration of **dropout** layers and usage of **AdamW** optimizer as first part of model enhancement and optimization; the second part of advanced regularization techniques includes **weight tying**, **variational dropout** and **Non-monotonically Triggered AvSGD**.

2. Implementation details

Regarding the first two models, **LSTM** and **LSTM with dropout**, both employing the SGD optimizer, our investigations have shown that the best results were achieved when setting the `LEARNING_RATE` to 0.5. Additionally, for both models, the most favorable perplexity was observed when configuring the embedding and hidden dimensions to be 800 and 1000, respectively. In terms of the `TRAIN_BATCH_SIZE`, both models demonstrated their peak performance with a value of 128. It's worth noting that for the dropout model, setting the `PATIENCE` variable (for early stopping) to 10 was instrumental in allowing the training process to steadily improve, owing to its slower progress in terms of performance.

Subsequently, when we introduced the **AdamW** optimizer, it yielded intriguing outcomes. Exceptional performance was attained with a `LEARNING_RATE` of 0.001, while cases involving **weight tying** required a slight adjustment to 0.01. Other hyperparameters, such as `EMB_SIZE` at 600, `HID_SIZE` at 800, `WEIGHT_DECAY` at 0.01, and `EPSILON` at $1e-7$, remained consistent throughout the entire process.

In a subsequent experiment, we explored the incorporation of **variational dropout** into the model, and substantial improvements were observed when we adopted a `LEARNING_RATE` of 0.005 while keeping the other hyperparameters unchanged. Conversely, our final model, which utilized the **Non-monotonically Triggered AvSGD** optimizer based on a non-monotonic criterion that conservatively triggers averaging when the validation metric fails to improve for multiple cycles, necessitated the exploration of high learning rates (until reaching the value of 30). These high rates would dynamically update during the model training process to yield significant performance enhancements.

3. Results

For evaluating the model, I utilized the central `train` function, within which I initialized the model, optimizer, and loss functions. This function encompassed two important components: the training loop and the evaluation loop.

The `train_loop` oversaw the model training process, involving loss calculation, parameter updates via an optimizer, and iterative performance enhancement. Simultaneously, the `eval_loop` was responsible for computing both loss and perplexity, the metric that quantifies the uncertainty in the model's predictions.

An important detail to highlight is that, within the `train` function, when using **Non-monotonically Triggered AvSGD** for model training, the `get_seq_len` function was invoked. This function continuously adjusted the learning rate, making use of a non-monotonic condition to determine the averaging trigger, as opposed to relying on user-defined tuning.

To determine the most optimal model, I implemented an early stopping mechanism, typically configured with a `PATIENCE` parameter set to 5. This mechanism effectively mitigated the risk of overfitting, by decrementing each time the calculated perplexity exceeded its previous value, and when the patience level reached 0, the training iteration was halted.

3.1. Part 1

The provided data illustrates that, in all three scenarios, we achieved a satisfactory level of perplexity. Our objective was to obtain values below the 250 threshold, where lower values signify superior model performance. We accomplished a perplexity score of approximately 196 for the **LSTM** model using the SGD optimizer by simply fine-tuning the `LEARNING_RATE` in the initial model. In the second model, in which we have just incorporated **dropout**, led to an even better perplexity, reaching 186. Crucially, a substantial reduction in perplexity was observed when transitioning from the SGD optimizer to AdamW, coupled with an appropriate adjustment of the `LEARNING_RATE`. This combination yielded a perplexity score of around 174.

3.2. Part 2

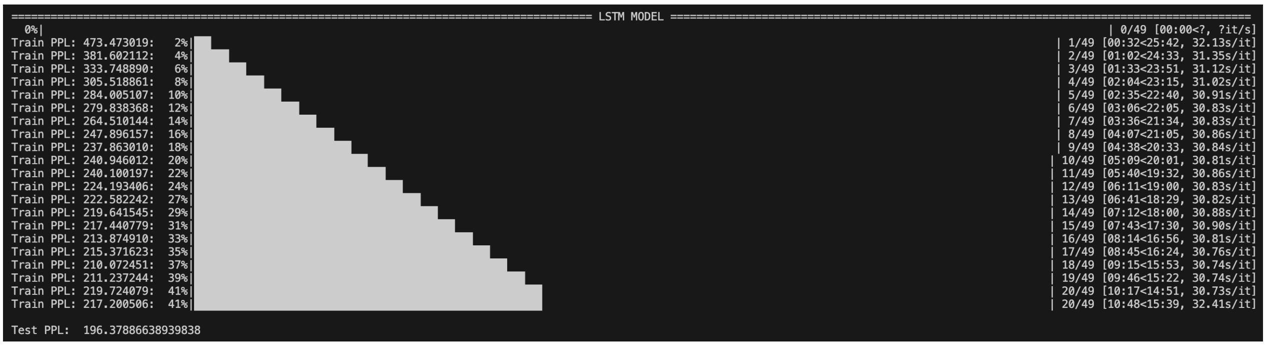
The application of advanced optimization techniques has led to a substantial reduction in perplexity values across these three study cases. When we implemented **Weight Tying** while retaining the same AdamW optimizer as in the previous model, we successfully reduced perplexity to below 170. The substitution of standard dropout with **variational dropout** yielded a perplexity value around 160. In the final model, where we adjusted the optimizer to utilize a **Non-monotonically Triggered AvSGD**, we achieved the lowest perplexity among all the models trained in this exercise. Once again, the results surpassed our initial expectations for this task.

4. Showing results .

Table 1: *Part 1*

MODEL	BATCH_SIZES	EMB/HID_SIZE	LR	CLIP	WD	EPS	**PPL**
LSTM	128 / 512 / 512	800 / 1000	0.5	10	-	-	196.37886
LSTM + Dropout	128 / 512 / 512	800 / 1000	0.5	10	-	-	186.43481
LSTM + Dropout + AdamW	128 / 512 / 512	600 / 800	0.001	10	0.01	1e-7	174.28317

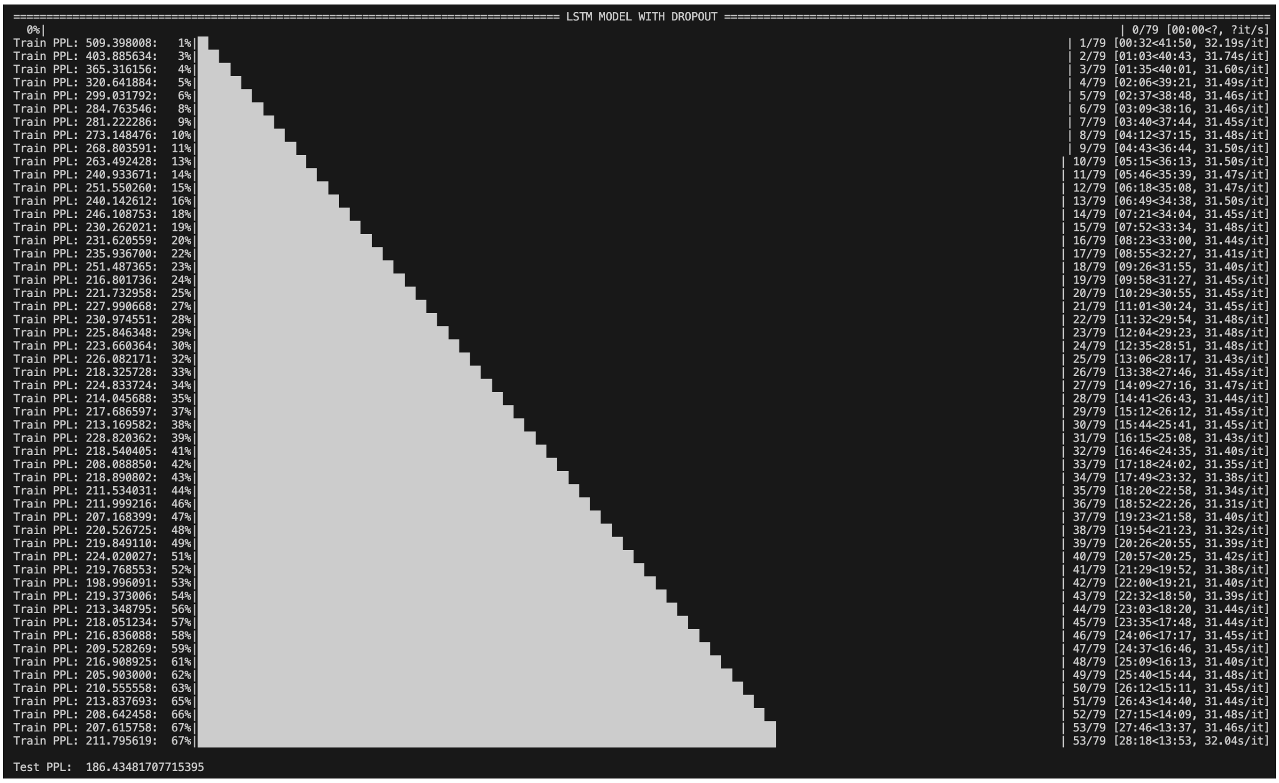
MODEL	BATCH_SIZES	EMB/HID_SIZE	LR	CLIP	WD	EPS	BPTT	**PPL**
Weight Tying	256 / 512 / 512	600 / 800	0.01	5	0.01	1e-7	-	164.58303
Variational Dropout	256 / 512 / 512	600 / 800	0.005	5	0.01	1e-7	-	159.06867
NT AvSGD	256 / 512 / 512	600 / 800	30	5	-	-	70	154.77544



```
=================================================== LSTM MODEL ===================================================
  0%|                                                                          | 0/49 [00:00<?, ?it/s]
Train PPL: 473.473019:   2%|                                           | 1/49 [00:32<25:42, 32.13s/it]
Train PPL: 381.602112:   4%|                                           | 2/49 [01:02<24:33, 31.35s/it]
Train PPL: 333.748890:   6%|                                           | 3/49 [01:33<23:51, 31.12s/it]
Train PPL: 305.518861:   8%|                                           | 4/49 [02:04<23:15, 31.02s/it]
Train PPL: 284.005107:  10%|                                           | 5/49 [02:35<22:40, 30.91s/it]
Train PPL: 279.838368:  12%|                                           | 6/49 [03:06<22:05, 30.83s/it]
Train PPL: 264.510144:  14%|                                           | 7/49 [03:36<21:34, 30.83s/it]
Train PPL: 247.896157:  16%|                                           | 8/49 [04:07<21:05, 30.86s/it]
Train PPL: 237.863010:  18%|                                           | 9/49 [04:38<20:33, 30.84s/it]
Train PPL: 240.946012:  20%|                                          | 10/49 [05:09<20:01, 30.81s/it]
Train PPL: 240.100197:  22%|                                          | 11/49 [05:40<19:32, 30.86s/it]
Train PPL: 224.193406:  24%|                                          | 12/49 [06:11<19:00, 30.83s/it]
Train PPL: 222.582242:  27%|                                          | 13/49 [06:41<18:29, 30.82s/it]
Train PPL: 219.641545:  29%|                                          | 14/49 [07:12<18:00, 30.88s/it]
Train PPL: 217.440779:  31%|                                          | 15/49 [07:43<17:30, 30.90s/it]
Train PPL: 213.874910:  33%|                                          | 16/49 [08:14<16:56, 30.81s/it]
Train PPL: 215.371623:  35%|                                          | 17/49 [08:45<16:24, 30.76s/it]
Train PPL: 210.072451:  37%|                                          | 18/49 [09:15<15:53, 30.74s/it]
Train PPL: 211.237244:  39%|                                          | 19/49 [09:46<15:22, 30.74s/it]
Train PPL: 219.724079:  41%|                                          | 20/49 [10:17<14:51, 30.73s/it]
Train PPL: 217.200506:  41%|                                          | 20/49 [10:48<15:39, 32.41s/it]

Test PPL:  196.37886638939838
```

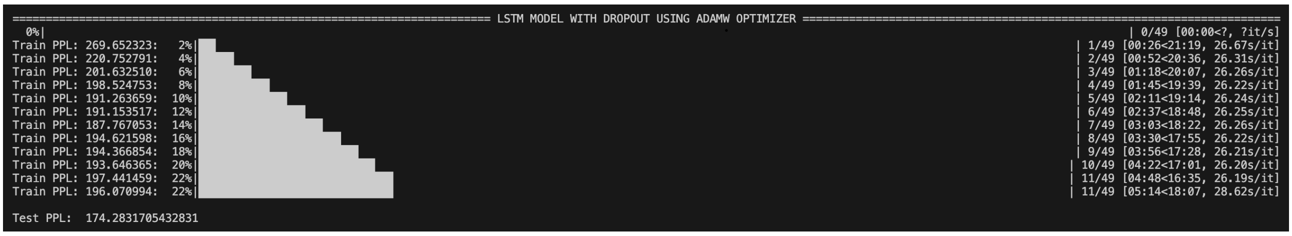
Figure 1: *LSTM model*



```
============================================= LSTM MODEL WITH DROPOUT =============================================
  0%|                                                                          | 0/79 [00:00<?, ?it/s]
Train PPL: 509.398008:   1%|                                           | 1/79 [00:32<41:50, 32.19s/it]
Train PPL: 403.885634:   3%|                                           | 2/79 [01:03<40:43, 31.74s/it]
Train PPL: 365.316156:   4%|                                           | 3/79 [01:35<40:01, 31.60s/it]
Train PPL: 320.641884:   5%|                                           | 4/79 [02:06<39:21, 31.49s/it]
Train PPL: 299.031792:   6%|                                           | 5/79 [02:37<38:48, 31.46s/it]
Train PPL: 284.763546:   8%|                                           | 6/79 [03:09<38:16, 31.46s/it]
Train PPL: 281.222286:   9%|                                           | 7/79 [03:40<37:44, 31.45s/it]
Train PPL: 273.148476:  10%|                                           | 8/79 [04:12<37:15, 31.48s/it]
Train PPL: 268.803591:  11%|                                           | 9/79 [04:43<36:44, 31.50s/it]
Train PPL: 263.492428:  13%|                                          | 10/79 [05:15<36:13, 31.50s/it]
Train PPL: 240.933671:  14%|                                          | 11/79 [05:46<35:39, 31.47s/it]
Train PPL: 251.550260:  15%|                                          | 12/79 [06:18<35:08, 31.47s/it]
Train PPL: 240.142612:  16%|                                          | 13/79 [06:49<34:38, 31.50s/it]
Train PPL: 246.108753:  18%|                                          | 14/79 [07:21<34:04, 31.45s/it]
Train PPL: 230.262021:  19%|                                          | 15/79 [07:52<33:34, 31.48s/it]
Train PPL: 231.620559:  20%|                                          | 16/79 [08:23<33:00, 31.44s/it]
Train PPL: 235.936700:  22%|                                          | 17/79 [08:55<32:27, 31.41s/it]
Train PPL: 251.487365:  23%|                                          | 18/79 [09:26<31:55, 31.40s/it]
Train PPL: 216.801736:  24%|                                          | 19/79 [09:58<31:27, 31.45s/it]
Train PPL: 221.732958:  25%|                                          | 20/79 [10:29<30:55, 31.45s/it]
Train PPL: 227.990668:  27%|                                          | 21/79 [11:01<30:24, 31.45s/it]
Train PPL: 230.974551:  28%|                                          | 22/79 [11:32<29:54, 31.48s/it]
Train PPL: 225.846348:  29%|                                          | 23/79 [12:04<29:23, 31.48s/it]
Train PPL: 223.660364:  30%|                                          | 24/79 [12:35<28:51, 31.48s/it]
Train PPL: 226.082171:  32%|                                          | 25/79 [13:06<28:17, 31.43s/it]
Train PPL: 218.325728:  33%|                                          | 26/79 [13:38<27:46, 31.45s/it]
Train PPL: 224.833724:  34%|                                          | 27/79 [14:09<27:16, 31.47s/it]
Train PPL: 214.045688:  35%|                                          | 28/79 [14:41<26:43, 31.44s/it]
Train PPL: 217.686597:  37%|                                          | 29/79 [15:12<26:12, 31.45s/it]
Train PPL: 213.169582:  38%|                                          | 30/79 [15:44<25:41, 31.45s/it]
Train PPL: 228.820362:  39%|                                          | 31/79 [16:15<25:08, 31.43s/it]
Train PPL: 218.540405:  41%|                                          | 32/79 [16:46<24:35, 31.40s/it]
Train PPL: 208.088850:  42%|                                          | 33/79 [17:18<24:02, 31.35s/it]
Train PPL: 218.890802:  43%|                                          | 34/79 [17:49<23:32, 31.38s/it]
Train PPL: 211.534031:  44%|                                          | 35/79 [18:20<22:58, 31.34s/it]
Train PPL: 211.999216:  46%|                                          | 36/79 [18:52<22:26, 31.31s/it]
Train PPL: 207.168399:  47%|                                          | 37/79 [19:23<21:58, 31.40s/it]
Train PPL: 220.526725:  48%|                                          | 38/79 [19:54<21:23, 31.32s/it]
Train PPL: 219.849110:  49%|                                          | 39/79 [20:26<20:55, 31.39s/it]
Train PPL: 224.020027:  51%|                                          | 40/79 [20:57<20:25, 31.42s/it]
Train PPL: 219.768553:  52%|                                          | 41/79 [21:29<19:52, 31.38s/it]
Train PPL: 198.996091:  53%|                                          | 42/79 [22:00<19:21, 31.40s/it]
Train PPL: 219.373006:  54%|                                          | 43/79 [22:32<18:50, 31.39s/it]
Train PPL: 213.348795:  56%|                                          | 44/79 [23:03<18:20, 31.44s/it]
Train PPL: 218.051234:  57%|                                          | 45/79 [23:35<17:48, 31.44s/it]
Train PPL: 216.836088:  58%|                                          | 46/79 [24:06<17:17, 31.45s/it]
Train PPL: 209.528269:  59%|                                          | 47/79 [24:37<16:46, 31.45s/it]
Train PPL: 216.908925:  61%|                                          | 48/79 [25:09<16:13, 31.40s/it]
Train PPL: 205.903000:  62%|                                          | 49/79 [25:40<15:44, 31.48s/it]
Train PPL: 210.555558:  63%|                                          | 50/79 [26:12<15:11, 31.45s/it]
Train PPL: 213.837693:  65%|                                          | 51/79 [26:43<14:40, 31.44s/it]
Train PPL: 208.642458:  66%|                                          | 52/79 [27:15<14:09, 31.48s/it]
Train PPL: 207.615758:  67%|                                          | 53/79 [27:46<13:37, 31.46s/it]
Train PPL: 211.795619:  67%|                                          | 53/79 [28:18<13:53, 32.04s/it]

Test PPL:  186.43481707715395
```

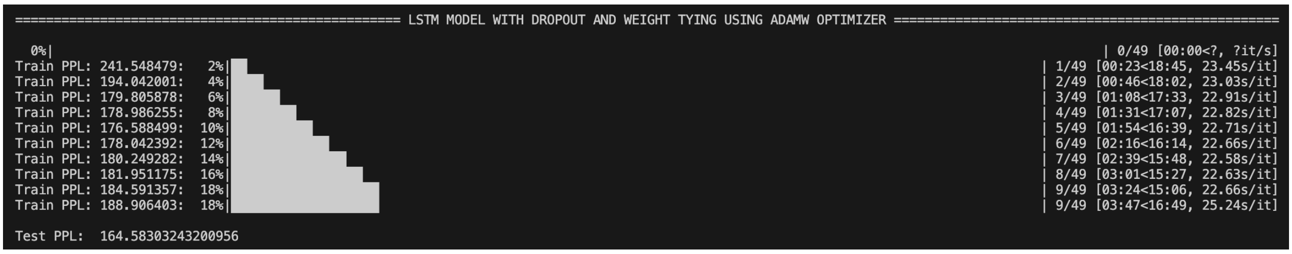
Figure 2: *LSTM model with dropout*



```
======================================== LSTM MODEL WITH DROPOUT USING ADAMW OPTIMIZER ========================================
  0%|                                                                  | 0/49 [00:00<?, ?it/s]
Train PPL: 269.652323:   2%|                                        | 1/49 [00:26<21:19, 26.67s/it]
Train PPL: 220.752791:   4%|                                        | 2/49 [00:52<20:36, 26.31s/it]
Train PPL: 201.632510:   6%|                                        | 3/49 [01:18<20:07, 26.26s/it]
Train PPL: 198.524753:   8%|                                        | 4/49 [01:45<19:39, 26.22s/it]
Train PPL: 191.263659:  10%|                                        | 5/49 [02:11<19:14, 26.24s/it]
Train PPL: 191.153517:  12%|                                        | 6/49 [02:37<18:48, 26.25s/it]
Train PPL: 187.767053:  14%|                                        | 7/49 [03:03<18:22, 26.26s/it]
Train PPL: 194.621598:  16%|                                        | 8/49 [03:30<17:55, 26.22s/it]
Train PPL: 194.366854:  18%|                                        | 9/49 [03:56<17:28, 26.21s/it]
Train PPL: 193.646365:  20%|                                       | 10/49 [04:22<17:01, 26.20s/it]
Train PPL: 197.441459:  22%|                                       | 11/49 [04:48<16:35, 26.19s/it]
Train PPL: 196.070994:  22%|                                       | 11/49 [05:14<18:07, 28.62s/it]

Test PPL:  174.2831705432831
```

Figure 3: *LSTM model with dropout using AdamW optimizer*

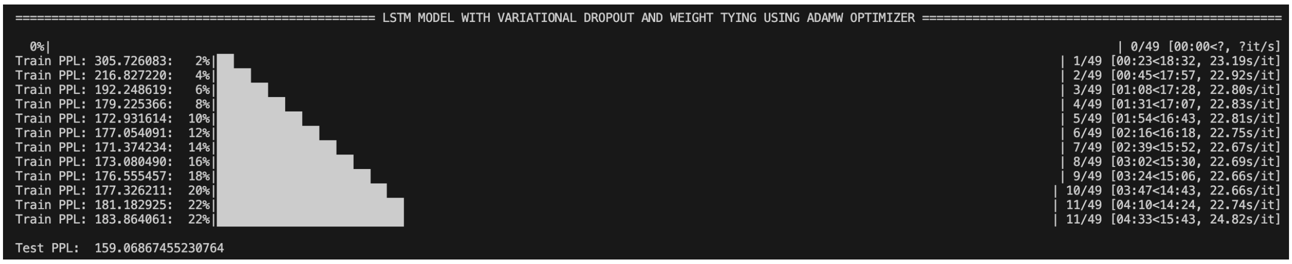


```
============================== LSTM MODEL WITH DROPOUT AND WEIGHT TYING USING ADAMW OPTIMIZER ==============================

  0%|                                                                  | 0/49 [00:00<?, ?it/s]
Train PPL: 241.548479:   2%|                                        | 1/49 [00:23<18:45, 23.45s/it]
Train PPL: 194.042001:   4%|                                        | 2/49 [00:46<18:02, 23.03s/it]
Train PPL: 179.805878:   6%|                                        | 3/49 [01:08<17:33, 22.91s/it]
Train PPL: 178.986255:   8%|                                        | 4/49 [01:31<17:07, 22.82s/it]
Train PPL: 176.588499:  10%|                                        | 5/49 [01:54<16:39, 22.71s/it]
Train PPL: 178.042392:  12%|                                        | 6/49 [02:16<16:14, 22.66s/it]
Train PPL: 180.249282:  14%|                                        | 7/49 [02:39<15:48, 22.58s/it]
Train PPL: 181.951175:  16%|                                        | 8/49 [03:01<15:27, 22.63s/it]
Train PPL: 184.591357:  18%|                                        | 9/49 [03:24<15:06, 22.66s/it]
Train PPL: 188.906403:  18%|                                        | 9/49 [03:47<16:49, 25.24s/it]

Test PPL:  164.58303243200956
```

Figure 4: *LSTM model with dropout and weight tying using AdamW optimizer*

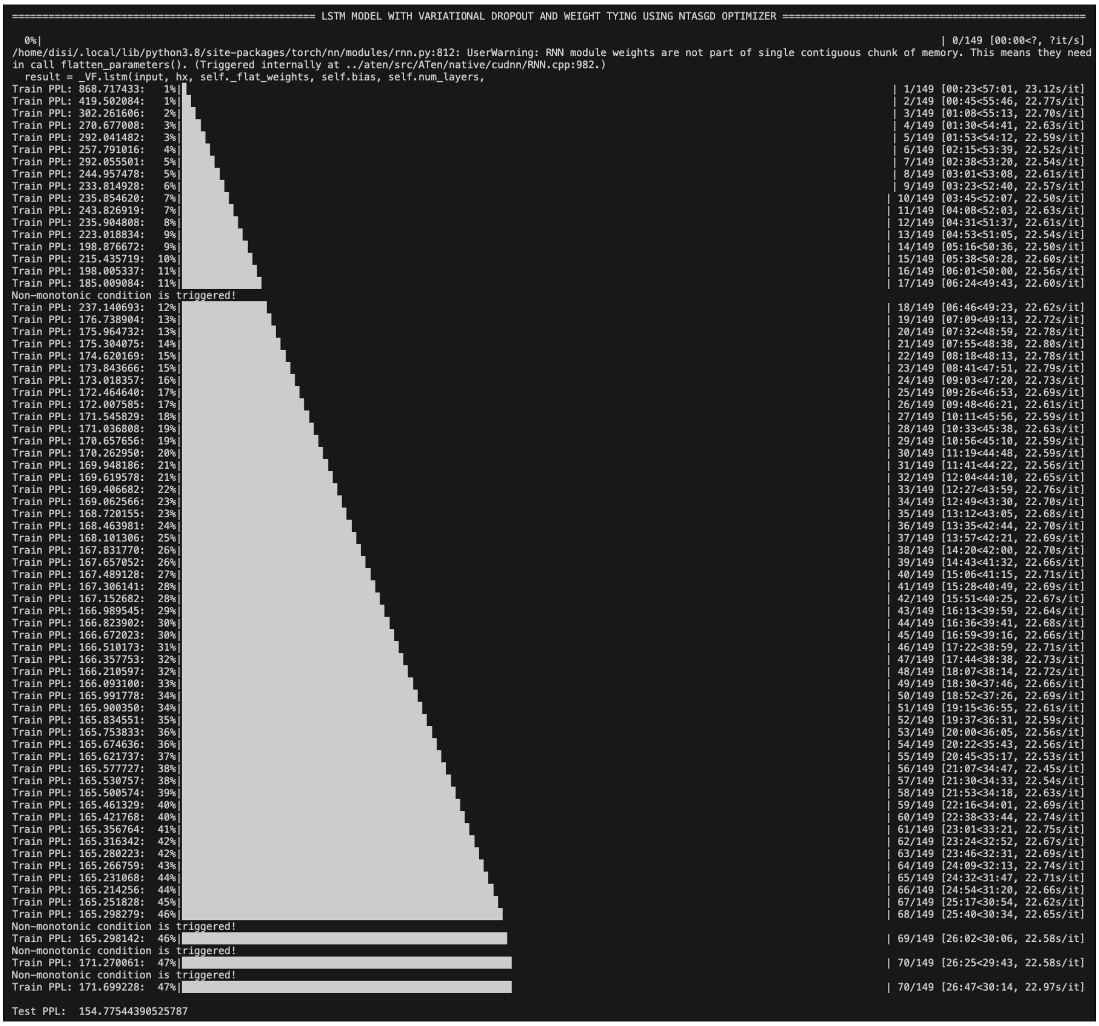


```
========================== LSTM MODEL WITH VARIATIONAL DROPOUT AND WEIGHT TYING USING ADAMW OPTIMIZER ==========================

  0%|                                                                  | 0/49 [00:00<?, ?it/s]
Train PPL: 305.726083:   2%|                                        | 1/49 [00:23<18:32, 23.19s/it]
Train PPL: 216.827220:   4%|                                        | 2/49 [00:45<17:57, 22.92s/it]
Train PPL: 192.248619:   6%|                                        | 3/49 [01:08<17:28, 22.80s/it]
Train PPL: 179.225366:   8%|                                        | 4/49 [01:31<17:07, 22.83s/it]
Train PPL: 172.931614:  10%|                                        | 5/49 [01:54<16:43, 22.81s/it]
Train PPL: 177.054091:  12%|                                        | 6/49 [02:16<16:18, 22.75s/it]
Train PPL: 171.374234:  14%|                                        | 7/49 [02:39<15:52, 22.67s/it]
Train PPL: 173.080490:  16%|                                        | 8/49 [03:02<15:30, 22.69s/it]
Train PPL: 176.555457:  18%|                                        | 9/49 [03:24<15:06, 22.66s/it]
Train PPL: 177.326211:  20%|                                       | 10/49 [03:47<14:43, 22.66s/it]
Train PPL: 181.182925:  22%|                                       | 11/49 [04:10<14:24, 22.74s/it]
Train PPL: 183.864061:  22%|                                       | 11/49 [04:33<15:43, 24.82s/it]

Test PPL:  159.06867455230764
```

Figure 5: *LSTM model with variational dropout and weight tying using AdamW optimizer*



```
================================================== LSTM MODEL WITH VARIATIONAL DROPOUT AND WEIGHT TYING USING NTASGD OPTIMIZER ==================================================

  0%|                                                                  | 0/149 [00:00<?, ?it/s]
/home/disi/.local/lib/python3.8/site-packages/torch/nn/modules/rnn.py:812: UserWarning: RNN module weights are not part of single contiguous chunk of memory. This means they need
in call flatten_parameters(). (Triggered internally at ../aten/src/ATen/native/cudnn/RNN.cpp:982.)
  result = _VF.lstm(input, hx, self._flat_weights, self.bias, self.num_layers,
Train PPL: 868.717433:   1%|          | 1/149 [00:23<57:01, 23.12s/it]
Train PPL: 419.502084:   1%|          | 2/149 [00:45<55:46, 22.77s/it]
Train PPL: 302.261606:   2%|          | 3/149 [01:08<55:13, 22.70s/it]
Train PPL: 270.677008:   3%|          | 4/149 [01:30<54:41, 22.63s/it]
Train PPL: 292.041482:   3%|          | 5/149 [01:53<54:12, 22.59s/it]
Train PPL: 257.791016:   4%|          | 6/149 [02:15<53:39, 22.52s/it]
Train PPL: 292.055501:   5%|          | 7/149 [02:38<53:20, 22.54s/it]
Train PPL: 244.957478:   5%|          | 8/149 [03:01<53:08, 22.61s/it]
Train PPL: 233.814928:   6%|          | 9/149 [03:23<52:40, 22.57s/it]
Train PPL: 235.854620:   7%|         | 10/149 [03:45<52:07, 22.50s/it]
Train PPL: 243.826919:   7%|         | 11/149 [04:08<52:03, 22.63s/it]
Train PPL: 235.904808:   8%|         | 12/149 [04:31<51:37, 22.61s/it]
Train PPL: 223.018834:   9%|         | 13/149 [04:53<51:05, 22.54s/it]
Train PPL: 198.876672:   9%|         | 14/149 [05:16<50:36, 22.50s/it]
Train PPL: 215.435719:  10%|         | 15/149 [05:38<50:28, 22.60s/it]
Train PPL: 198.005337:  11%|         | 16/149 [06:01<50:00, 22.56s/it]
Train PPL: 185.009084:  11%|         | 17/149 [06:24<49:43, 22.60s/it]
Non-monotonic condition is triggered!
Train PPL: 237.140693:  12%|         | 18/149 [06:46<49:23, 22.62s/it]
Train PPL: 176.738904:  13%|         | 19/149 [07:09<49:13, 22.72s/it]
Train PPL: 175.964732:  13%|         | 20/149 [07:32<48:59, 22.78s/it]
Train PPL: 175.304075:  14%|         | 21/149 [07:55<48:38, 22.80s/it]
Train PPL: 174.620169:  15%|         | 22/149 [08:18<48:13, 22.78s/it]
Train PPL: 173.843666:  15%|         | 23/149 [08:41<47:51, 22.79s/it]
Train PPL: 173.018357:  16%|         | 24/149 [09:03<47:20, 22.73s/it]
Train PPL: 172.464640:  17%|         | 25/149 [09:26<46:53, 22.69s/it]
Train PPL: 172.007585:  17%|         | 26/149 [09:48<46:21, 22.61s/it]
Train PPL: 171.545829:  18%|         | 27/149 [10:11<45:56, 22.59s/it]
Train PPL: 171.036808:  19%|         | 28/149 [10:33<45:38, 22.63s/it]
Train PPL: 170.657656:  19%|         | 29/149 [10:56<45:10, 22.59s/it]
Train PPL: 170.262950:  20%|         | 30/149 [11:19<44:48, 22.59s/it]
Train PPL: 169.948186:  21%|         | 31/149 [11:41<44:22, 22.56s/it]
Train PPL: 169.619578:  21%|         | 32/149 [12:04<44:10, 22.65s/it]
Train PPL: 169.406682:  22%|         | 33/149 [12:27<43:59, 22.76s/it]
Train PPL: 169.062566:  23%|         | 34/149 [12:49<43:30, 22.70s/it]
Train PPL: 168.720155:  23%|         | 35/149 [13:12<43:05, 22.68s/it]
Train PPL: 168.463981:  24%|         | 36/149 [13:35<42:44, 22.70s/it]
Train PPL: 168.101306:  25%|         | 37/149 [13:57<42:21, 22.69s/it]
Train PPL: 167.831770:  26%|         | 38/149 [14:20<42:00, 22.70s/it]
Train PPL: 167.657052:  26%|         | 39/149 [14:43<41:32, 22.66s/it]
Train PPL: 167.489128:  27%|         | 40/149 [15:06<41:15, 22.71s/it]
Train PPL: 167.306141:  28%|         | 41/149 [15:28<40:49, 22.69s/it]
Train PPL: 167.152682:  28%|         | 42/149 [15:51<40:25, 22.67s/it]
Train PPL: 166.989545:  29%|         | 43/149 [16:13<39:59, 22.64s/it]
Train PPL: 166.823902:  30%|         | 44/149 [16:36<39:41, 22.68s/it]
Train PPL: 166.672023:  30%|         | 45/149 [16:59<39:16, 22.66s/it]
Train PPL: 166.510173:  31%|         | 46/149 [17:22<38:59, 22.71s/it]
Train PPL: 166.357753:  32%|         | 47/149 [17:44<38:38, 22.73s/it]
Train PPL: 166.210597:  32%|         | 48/149 [18:07<38:14, 22.72s/it]
Train PPL: 166.093100:  33%|         | 49/149 [18:30<37:46, 22.66s/it]
Train PPL: 165.991778:  34%|         | 50/149 [18:52<37:26, 22.69s/it]
Train PPL: 165.900350:  34%|         | 51/149 [19:15<36:55, 22.61s/it]
Train PPL: 165.834551:  35%|         | 52/149 [19:37<36:31, 22.59s/it]
Train PPL: 165.753833:  36%|         | 53/149 [20:00<36:05, 22.56s/it]
Train PPL: 165.674636:  36%|         | 54/149 [20:22<35:43, 22.56s/it]
Train PPL: 165.621737:  37%|         | 55/149 [20:45<35:17, 22.53s/it]
Train PPL: 165.577727:  38%|         | 56/149 [21:07<34:47, 22.45s/it]
Train PPL: 165.530757:  38%|         | 57/149 [21:30<34:33, 22.54s/it]
Train PPL: 165.500574:  39%|         | 58/149 [21:53<34:18, 22.63s/it]
Train PPL: 165.461329:  40%|         | 59/149 [22:16<34:01, 22.69s/it]
Train PPL: 165.421768:  40%|         | 60/149 [22:38<33:44, 22.74s/it]
Train PPL: 165.356764:  41%|         | 61/149 [23:01<33:21, 22.75s/it]
Train PPL: 165.316342:  42%|         | 62/149 [23:24<32:52, 22.67s/it]
Train PPL: 165.280223:  42%|         | 63/149 [23:46<32:31, 22.69s/it]
Train PPL: 165.266759:  43%|         | 64/149 [24:09<32:13, 22.74s/it]
Train PPL: 165.231068:  44%|         | 65/149 [24:32<31:47, 22.71s/it]
Train PPL: 165.214256:  44%|         | 66/149 [24:54<31:20, 22.66s/it]
Train PPL: 165.251828:  45%|         | 67/149 [25:17<30:54, 22.62s/it]
Train PPL: 165.298279:  46%|         | 68/149 [25:40<30:34, 22.65s/it]
Non-monotonic condition is triggered!
Train PPL: 165.298142:  46%|         | 69/149 [26:02<30:06, 22.58s/it]
Non-monotonic condition is triggered!
Train PPL: 171.270061:  47%|         | 70/149 [26:25<29:43, 22.58s/it]
Non-monotonic condition is triggered!
Train PPL: 171.699228:  47%|         | 70/149 [26:47<30:14, 22.97s/it]

Test PPL:  154.77544390525787
```

Figure 6: *LSTM model with variational dropout and weight tying using NTAvSGD optimizer*