HW3 - Symbolic-domain Music Generation

Name: 譚至斌
ID: d12942015

Novelty

- Speed up inference with k-v cache
- Inference with sink-attention-like pattern

Methodology

- The methodology can be described by these steps:

    1. Convert all midi data to REMI tokens
    2. Train a decoder-only transformer on the dataset
    3. Inference from scratch

Methodology (cont.)

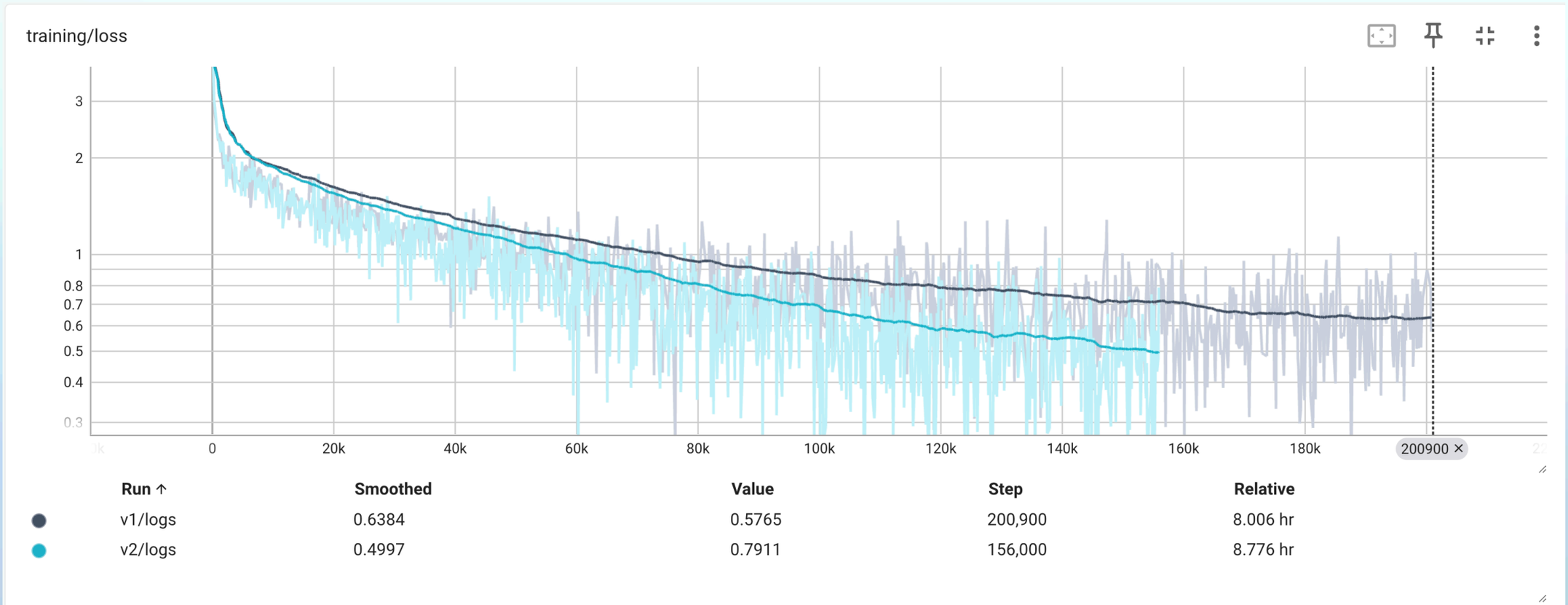
- To enable cache mechanism for Transformer, the positional encoding should be chosen carefully
  - Rotary Position Embedding is used due to its linear property
- Sink-attention claims that Transformer often attends to first few tokens
  - I keeps the ‘bos’ token at the start of the input sequence in the sliding window on inference

Experiment Setup

- The models are built with 2 types of configurations. Both have:
  - 512-dim latent embeddings
  - 8 heads
  - 2048-dim dense layers.
- The difference is one of them (v1) has 12 transformer layers, but the other (v2) has only 8 layers.

Result

- v1 took ~200000 steps to converge
- v2 didn’t converge even it had been update for ~150000



Run ↑	Smoothed	Value	Step	Relative
v1/logs	0.6384	0.5765	200,900	8.006 hr
v2/logs	0.4997	0.7911	156,000	8.776 hr

Result

tag	H	GS	ED
dataset_cache_bd_4	2.5207	0.7789	6.5399
v1_00150000_top-p_0.95_0.85_sink	2.1453	0.8793	7.0275
v1_00150000_top-p_0.95_0.95_sink	2.4171	0.8316	6.7367
v1_00150000_top-p_0.95_0.9_sink	2.3908	0.8301	6.7614
v1_00150000_top-p_0.9_0.85_sink	2.2634	0.9058	6.8974
v1_00150000_top-p_0.9_0.95_sink	2.268	0.8371	6.7717
v1_00150000_top-p_0.9_0.9_sink	2.2726	0.8513	6.8623
v1_00150000_top-p_1.05_0.85_sink	2.4071	0.8482	6.8361
v1_00150000_top-p_1.05_0.95_sink	2.4577	0.8017	6.8011
v1_00150000_top-p_1.05_0.9_sink	2.4074	0.8323	6.8667
v1_00150000_top-p_1.0_0.85_sink	2.2647	0.8373	6.8187
v1_00150000_top-p_1.0_0.95_sink	2.401	0.8181	6.8199
v1_00150000_top-p_1.0_0.9_sink	2.3629	0.7955	6.9031
v1_00150000_top-p_1.1_0.85_sink	2.4591	0.8129	6.8414
v1_00150000_top-p_1.1_0.95_sink	2.4138	0.7892	6.7087
v1_00150000_top-p_1.1_0.9_sink	2.3098	0.8159	6.7633
v1_00200000_top-p_0.95_0.85_sink	2.3153	0.8402	6.9499
v1_00200000_top-p_0.95_0.95_sink	2.4392	0.8219	6.8349
v1_00200000_top-p_0.95_0.9_sink	2.3316	0.8159	6.7586
v1_00200000_top-p_0.9_0.85_sink	2.3851	0.8977	6.8829
v1_00200000_top-p_0.9_0.95_sink	2.353	0.8225	6.7778
v1_00200000_top-p_0.9_0.9_sink	2.4437	0.8444	6.8581
v1_00200000_top-p_1.05_0.85_sink	2.2468	0.8816	6.8994
v1_00200000_top-p_1.05_0.95_sink	2.4413	0.8006	6.7414
v1_00200000_top-p_1.05_0.9_sink	2.2414	0.8372	6.7892
v1_00200000_top-p_1.0_0.85_sink	2.4541	0.8471	6.8391
v1_00200000_top-p_1.0_0.95_sink	2.4007	0.7968	6.7148
v1_00200000_top-p_1.0_0.9_sink	2.3934	0.8496	6.7891
v1_00200000_top-p_1.1_0.85_sink	2.3107	0.8123	7.3758
v1_00200000_top-p_1.1_0.95_sink	2.3992	0.827	6.7361
v1_00200000_top-p_1.1_0.9_sink	2.4012	0.8258	6.9008

tag	H	GS	ED
dataset_cache_bd_4	2.5207	0.7789	6.5399
v2_00100000_top-p_0.95_0.85_sink	2.3119	0.8827	6.9466
v2_00100000_top-p_0.95_0.95_sink	2.3981	0.8547	6.8309
v2_00100000_top-p_0.95_0.9_sink	2.2617	0.8602	6.8026
v2_00100000_top-p_0.9_0.85_sink	2.1737	0.8946	6.9782
v2_00100000_top-p_0.9_0.95_sink	2.3703	0.8246	6.7807
v2_00100000_top-p_0.9_0.9_sink	2.3076	0.8396	6.8616
v2_00100000_top-p_1.05_0.85_sink	2.4415	0.8244	6.8225
v2_00100000_top-p_1.05_0.95_sink	2.3809	0.8138	6.7446
v2_00100000_top-p_1.05_0.9_sink	2.4036	0.8454	6.7733
v2_00100000_top-p_1.0_0.85_sink	2.3148	0.8578	6.9164
v2_00100000_top-p_1.0_0.95_sink	2.4285	0.7894	6.8019
v2_00100000_top-p_1.0_0.9_sink	2.2499	0.868	6.8084
v2_00100000_top-p_1.1_0.85_sink	2.242	0.8713	6.933
v2_00100000_top-p_1.1_0.95_sink	2.4588	0.8514	6.788
v2_00100000_top-p_1.1_0.9_sink	2.4219	0.7626	6.8042
v2_00150000_top-p_0.95_0.85_sink	2.1862	0.8558	6.9605
v2_00150000_top-p_0.95_0.95_sink	2.3219	0.8027	6.9036
v2_00150000_top-p_0.95_0.9_sink	2.2802	0.8437	6.8821
v2_00150000_top-p_0.9_0.85_sink	2.391	0.9033	7.2008
v2_00150000_top-p_0.9_0.95_sink	2.3889	0.844	6.8119
v2_00150000_top-p_0.9_0.9_sink	2.3696	0.8401	6.8455
v2_00150000_top-p_1.05_0.85_sink	2.3626	0.8291	6.9299
v2_00150000_top-p_1.05_0.95_sink	2.3835	0.8205	6.7757
v2_00150000_top-p_1.05_0.9_sink	2.307	0.8126	6.9814
v2_00150000_top-p_1.0_0.85_sink	2.3435	0.8477	6.7973
v2_00150000_top-p_1.0_0.95_sink	2.4267	0.7903	6.8518
v2_00150000_top-p_1.0_0.9_sink	2.431	0.8601	6.8534
v2_00150000_top-p_1.1_0.85_sink	2.3967	0.832	6.939
v2_00150000_top-p_1.1_0.95_sink	2.458	0.8228	6.8469
v2_00150000_top-p_1.1_0.9_sink	2.3808	0.8737	6.7967

Findings

- I was surprised that even though the dataset is small (~1700 pop song), the small model (with 8 layers) couldn't overfit this dataset (the loss stopped decreasing).

Appendix

- source code: https://github.com/tanchihpin0517/NTU/tree/master/deepmir/hw3
- model: https://www.dropbox.com/scl/fi/k26tsdzdgpq7jjcb3sevs/model_00150000?rlkey=fn3v0ypdlkwrsy961ke3c44ug&dl=0