Reproduce SALMONN

Speech **A**udio **L**anguage **M**usic **O**pen **N**eural **N**etwork

학생 원종찬, 최재원

Prof. Jae-Hong Lee

Environments

- author trained SALMONN using A100-SXM-80GB

- in my case, RTX5090 32GB * 4

🌈 How to train a model

For SALMONN-13B v1, you need to use the following dependencies:

1. Our environment: The python version is 3.9.17, and other required packages can be installed with the following command: `pip install -r requirements.txt`.
2. Download whisper large v2 to `whisper_path`.
3. Download Fine-tuned BEATs_iter3+ (AS2M) (cpt2) to `beats_path`.
4. Download vicuna 13B v1.1 to `llama_path`.
5. Running with `python3 train.py --cfg-path configs/config.yaml` in A100-SXM-80GB.

🌈 How to inference in CLI

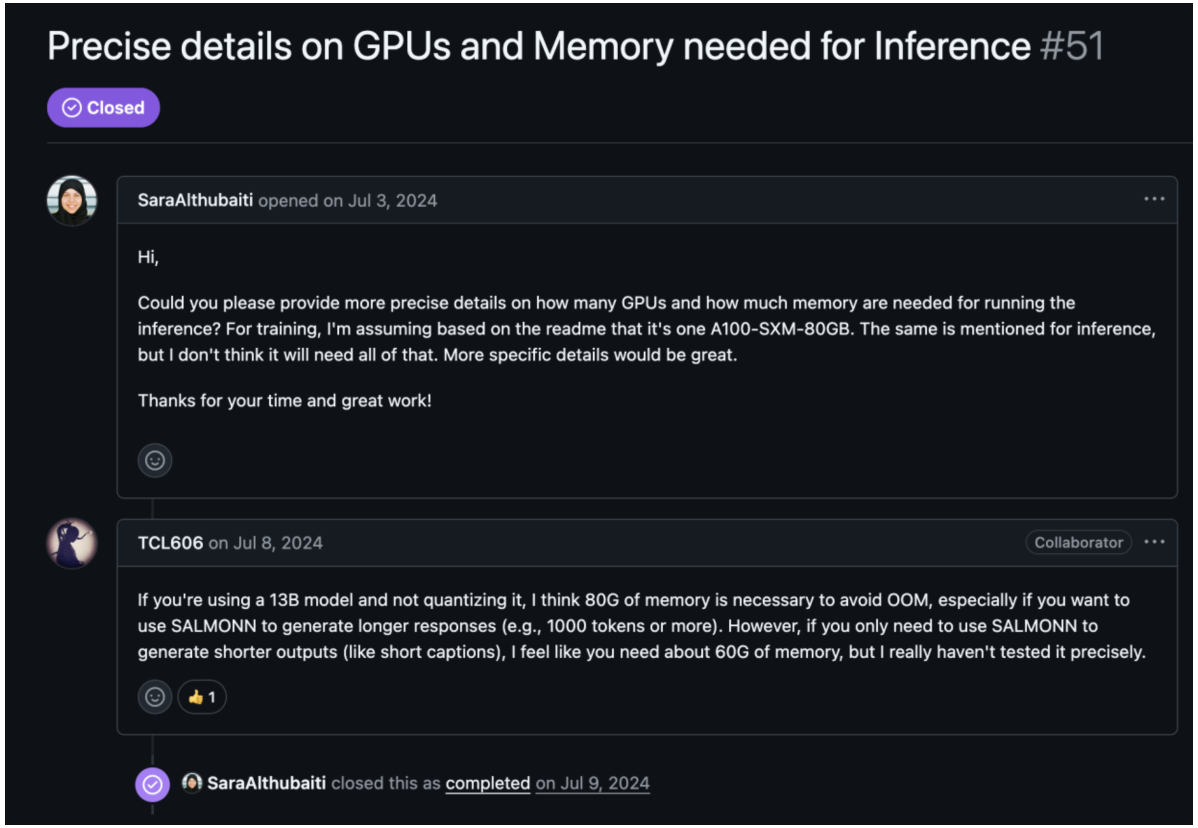
1. Same as **How to train a model: 1-4**.
2. Download salmonn v1 to `ckpt`.
3. Running with `python3 cli_inference.py --cfg-path configs/decode_config.yaml` in A100-SXM-80GB. Now you can input `wav_path` and `prompt`. Enjoy yourself !

🌈 How to launch a web demo

1. Same as **How to train a model: 1-4**.
2. Download salmonn v1 to `ckpt`.
3. Running with `python3 web_demo.py --cfg-path configs/decode_config.yaml` in A100-SXM-80GB.

Environments

- no detailed environments are provided



Data Preparation

🔥 News

- [2024-05-28] 🧳 We have released all the annotations (including 600k SQA/AQA data and 50k audio-based storytelling data) for the 3-stage training of SALMONN! Feel free to download them here!

Data Preparation


```
335M Sep 15 16:28 salmonn_stage1_data.json
696M Sep 15 16:32 salmonn_stage2_data.json
 43M Sep 15 16:32 salmonn_stage3_data.json
```

Data Preparation


```
(salmonn) jpong@hufs_5090_4ea:~/Workspace/jaeeewon/repr_salmonn$ python util/anns_inspector.py
===== salmonn_stage1_data.json =====
inspecting ./ann/salmonn_stage1_data.json...: 100%|| 1626555/1626555 [00:06<00:00, 261214.99it/s]
[asr] LibriSpeech | ready: 281241, not_ready: 0, total: 281241 | ready_rate: 100.00%
[asr] GigaSpeech | ready: 910140, not_ready: 0, total: 910140 | ready_rate: 100.00%
[audiocaption] AudioCaps | ready: 44211, not_ready: 4056, total: 48267 | ready_rate: 91.60%
[audiocaption] WavCaps | ready: 107310, not_ready: 260402, total: 367712 | ready_rate: 29.18%
[audiocaption] Clotho | ready: 14465, not_ready: 4730, total: 19195 | ready_rate: 75.36%
===== salmonn_stage2_data.json =====
inspecting ./ann/salmonn_stage2_data.json...: 100%|| 2390842/2390842 [00:09<00:00, 257238.94it/s]
[asr] LibriSpeech | ready: 281241, not_ready: 0, total: 281241 | ready_rate: 100.00%
[asr] GigaSpeech | ready: 200000, not_ready: 0, total: 200000 | ready_rate: 100.00%
[audiocaption] AudioCaps | ready: 44211, not_ready: 4056, total: 48267 | ready_rate: 91.60%
[audiocaption_v2] Clotho | ready: 14465, not_ready: 4730, total: 19195 | ready_rate: 75.36%
[translation_ec] CommonVoice | ready: 289354, not_ready: 0, total: 289354 | ready_rate: 100.00%
[phone_recognition] LibriSpeech | ready: 281239, not_ready: 0, total: 281239 | ready_rate: 100.00%
[emotion_recognition] IEMOCAP | ready: 0, not_ready: 4090, total: 4090 | ready_rate: 0.00%
[music_description] MusicCaps | ready: 0, not_ready: 2643, total: 2643 | ready_rate: 0.00%
[QA] MillionSongDatasetSpotify | ready: 0, not_ready: 48137, total: 48137 | ready_rate: 0.00%
[QA] MusicNet | ready: 0, not_ready: 320, total: 320 | ready_rate: 0.00%
[QA] LibriSpeech | ready: 281037, not_ready: 0, total: 281037 | ready_rate: 100.00%
[QA] AudioCaps | ready: 44211, not_ready: 4056, total: 48267 | ready_rate: 91.60%
[QA] WavCaps | ready: 107310, not_ready: 163092, total: 270402 | ready_rate: 39.69%
[speech_separation] LibriMix | ready: 0, not_ready: 64700, total: 64700 | ready_rate: 0.00%
[speaker_verification] Voxceleb1 | ready: 0, not_ready: 523411, total: 523411 | ready_rate: 0.00%
[gender_recognition] LibriSpeech | ready: 28539, not_ready: 0, total: 28539 | ready_rate: 100.00%
===== salmonn_stage3_data.json =====
inspecting ./ann/salmonn_stage3_data.json...: 100%|| 48272/48272 [00:00<00:00, 206153.68it/s]
[audio_story_telling] AudioCaps | ready: 44216, not_ready: 4056, total: 48272 | ready_rate: 91.60%
```

Data Preparation

- dataset not distinguished

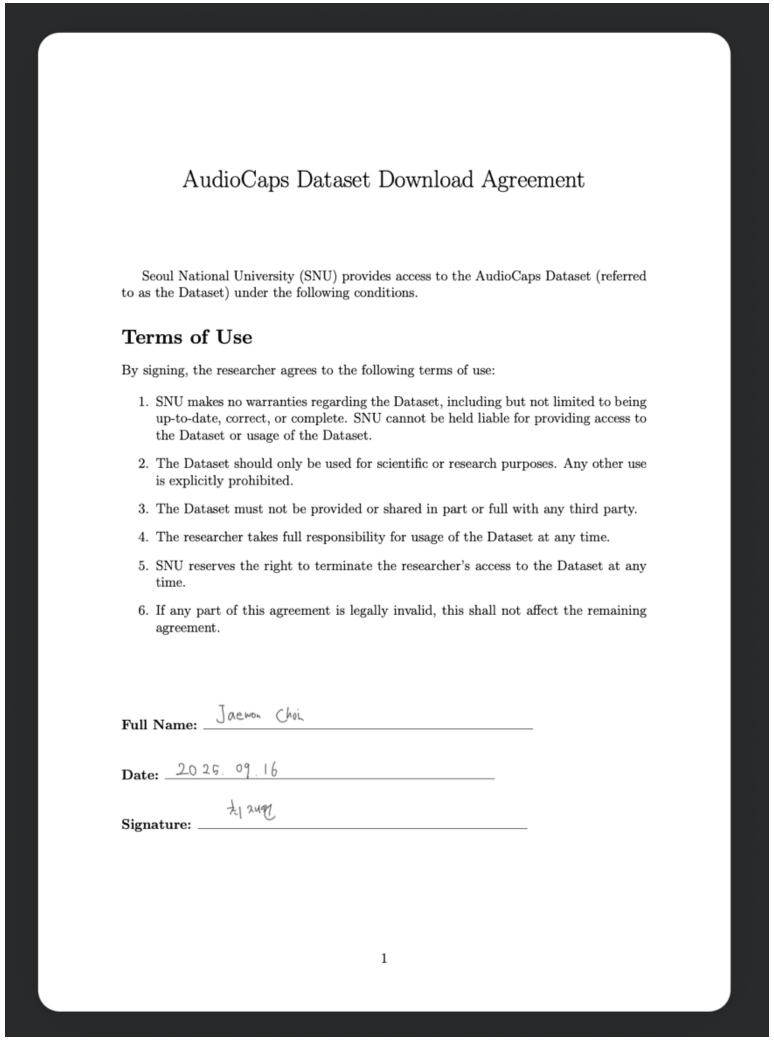
```python
# split into train, valid, test
ds_rate = {"train": 0.99, "valid": 0.005, "test": 0.005}
ds = {"train": [], "valid": [], "test": []}

for d in store_list:
    rand = random.random()
    if rand < ds_rate["train"]:
        ds["train"].append(d)
    elif rand < ds_rate["train"] + ds_rate["valid"]:
        ds["valid"].append(d)
    else:
        ds["test"].append(d)

for dset in ds:
    save_path = f"./ann/{train_set}_{dset}_ensured.json"
    with open(save_path, "w") as f:
        json.dump({"annotation": ds[dset]}, f)
```

Data Preparation

- no response



AudioCaps Dataset Download Agreement

Seoul National University (SNU) provides access to the AudioCaps Dataset (referred to as the Dataset) under the following conditions.

Terms of Use

By signing, the researcher agrees to the following terms of use:

1. SNU makes no warranties regarding the Dataset, including but not limited to being up-to-date, correct, or complete. SNU cannot be held liable for providing access to the Dataset or usage of the Dataset.
2. The Dataset should only be used for scientific or research purposes. Any other use is explicitly prohibited.
3. The Dataset must not be provided or shared in part or full with any third party.
4. The researcher takes full responsibility for usage of the Dataset at any time.
5. SNU reserves the right to terminate the researcher's access to the Dataset at any time.
6. If any part of this agreement is legally invalid, this shall not affect the remaining agreement.

Full Name: Jaewon Choi

Date: 2025. 09. 16

Signature: 최재원

1

Data Preparation

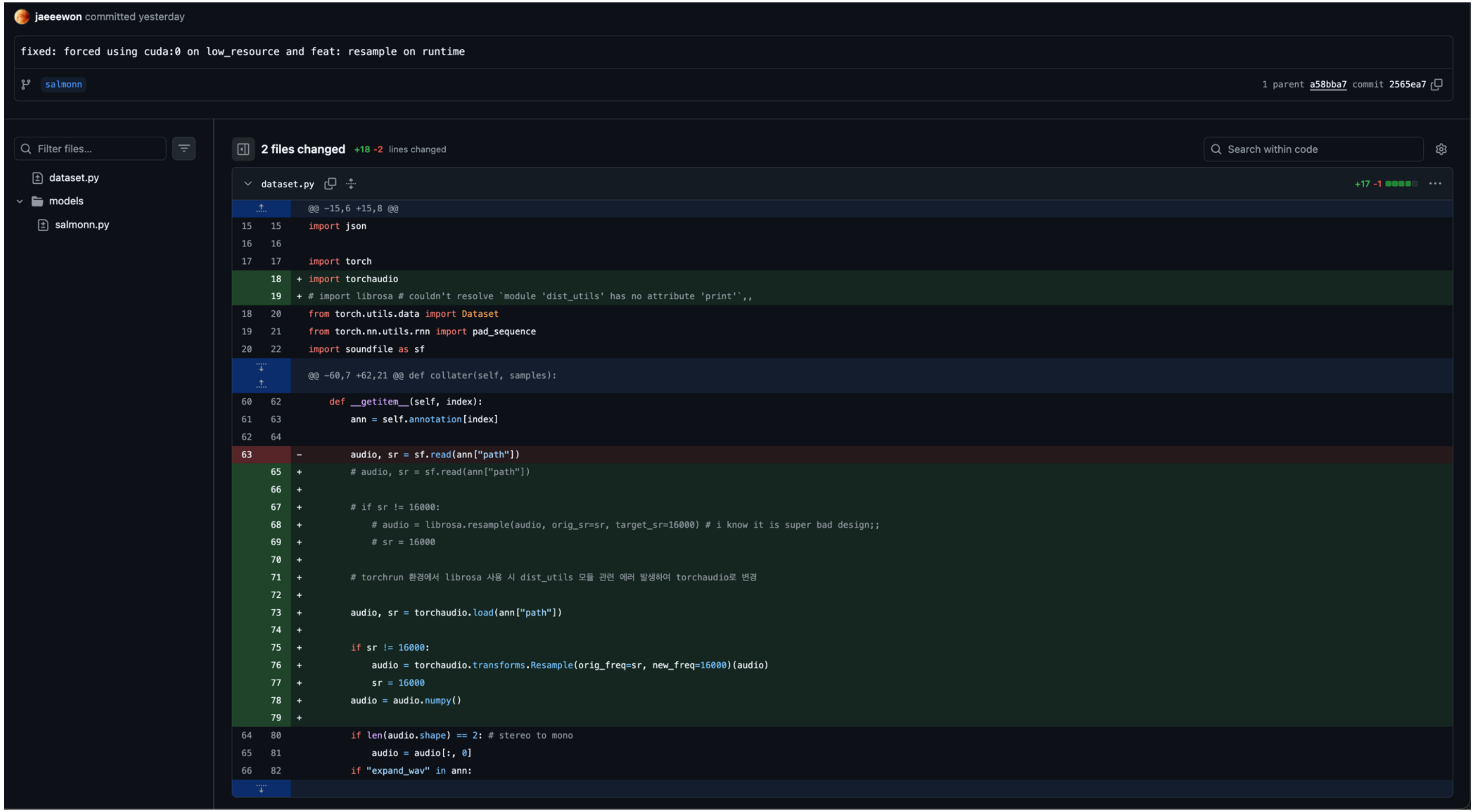
- whisper encoder force dataset to be sampled with 16k sr
- but the raw data differs between each set
- resampling is not affordable
    - resample on runtime!

```
(salmonn) jpong@hufs_5090_4ea:~/Workspace/jaeeewon/repr_salmonn$ python util/anns_inspector.py
===== salmonn_stage1_data.json =====
inspecting ./ann/salmonn_stage1_data.json...:   0%|
===== [asr] LibriSpeech | /LibriSpeech/train-clean-100/103/1240/103-1240-0000.flac =====
file path: /LibriSpeech/train-clean-100/103/1240
sampling rate: 16000
channels: 1
format: FLAC (Free Lossless Audio Codec) (FLAC)
duration: 14.09s
inspecting ./ann/salmonn_stage1_data.json...:  17%|
===== [asr] GigaSpeech | /GigaSpeech/YOU0000000315_S0000660.wav =====
file path: /GigaSpeech
sampling rate: 16000
channels: 1
format: WAV (Microsoft) (WAV)
duration: 3.18s
inspecting ./ann/salmonn_stage1_data.json...:  73%|
===== [audiocaption] AudioCaps | /AudioCaps/train/r1nicOVtvkQ.wav =====
file path: /AudioCaps/train
sampling rate: 24000
channels: 1
format: WAV (Microsoft) (WAV)
duration: 6.67s
inspecting ./ann/salmonn_stage1_data.json...:  75%|
===== [audiocaption] WavCaps | /WavCaps/AudioSet_SL/YbJgb7tyh6Uk.flac =====
file path: /WavCaps/AudioSet_SL
sampling rate: 32000
channels: 1
format: FLAC (Free Lossless Audio Codec) (FLAC)
duration: 9.25s
inspecting ./ann/salmonn_stage1_data.json...:  98%|
===== [audiocaption] Clotho | /Clotho/train/Distorted AM Radio noise.wav =====
file path: /Clotho/train
sampling rate: 44100
channels: 1
format: WAV (Microsoft) (WAV)
duration: 26.16s
```

```
(base) user@hufs:~$ df -h | grep nvme
/dev/nvme0n1p2  3.6T  3.1T  386G  90% /
```

Data Preparation

- resample on runtime!



Data Preparation

- invalid media crashed learning repeatedly
  - check media integrity!

1 file changed +60 -7 lines changed

Top Search within code

util/anns_inspector.py +60 -7

```
+    def _check_media_integrity(
+        self, ann: dict, cfg: dict, store_dict: dict, store_list: list
+    ):
+        task = ann.get("task")
+        # {"path": "/LibriSpeech/train-clean-100/path_to_audio.flac", "text": "recognized", "task": "asr"}
+        path = ann.get("path")
+        prefix = cfg.get("prefix")
+        splt_path = path.split("/")
+        path_dir = "/".join(splt_path[:-1])
+        dataset = splt_path[1]  # LibriSpeech
+
+        if dataset == "CommonVoice":
+            path = path.replace(".wav", ".mp3")
+        # CommonVoice 데이터셋은 .mp3로 구성됐으므로 오직 빠진 파일이 있는지 체크하는 용도로만 사용함
+
+        file_name = splt_path[-1]  # path_to_audio.flac
+        real_path = f"{prefix}{path}"
+
+        key = "ready" if os.path.exists(real_path) else "not_ready"
+
+        if key == "ready":
+            try:
+                sf.read(real_path)
+            except Exception as e:
+                store_list.append(real_path)
+                print(f"failed to load audio for {real_path}: {e}")
+
```

Experiment Preparation

- run 4 GPU
  - deepspeed – 😭
    - interferes autocast
  - torchrun – implemented
    - **D**istributed **D**ata **P**arallel

```python
# model
self._model = model
self._model.to(self.device)
if self.use_distributed:
    self.model = DDP(
        self._model, device_ids=[self.config.config.run.gpu]
    )
else:
    self.model = self._model
```

Experiment Settings – attempt1

cute typo

experimental settings

- randomly split `salmonn_stage1_data.json` into train, validation and test set with 80:10:10 ratio
- use smaller speech model `whisper-large-v2` → `whisper-medium`
- use smaller llm `vicuna-13b-v1.1` → `vicuna-7b-v1.1`
- load llm in 8bit for low resource
- use torchrun for distributed learning

Experiment Logs – attempt1

train | 2nd epoch | crashed

```
Train: data epoch: [1]  [ 285/3000]  eta: 0:22:16  lr: 0.000030  loss: 2.7421  time: 0.4907  data: 0.0000  max mem:
[rank3]: Traceback (most recent call last):
[rank3]:   File "/home/jpong/Workspace/jaeeewon/repr_salmonn/salmonn/train.py", line 91, in <module>
[rank3]:     main()
[rank3]:   File "/home/jpong/Workspace/jaeeewon/repr_salmonn/salmonn/train.py", line 87, in main
[rank3]:     runner.train()
[rank3]:   File "/home/jpong/Workspace/jaeeewon/repr_salmonn/salmonn/runner.py", line 276, in train
[rank3]:     train_stats = self.train_epoch(cur_epoch)
[rank3]:   File "/home/jpong/Workspace/jaeeewon/repr_salmonn/salmonn/runner.py", line 116, in train_epoch
[rank3]:     samples = next(self.train_loader)
[rank3]:   File "/home/jpong/Workspace/jaeeewon/repr_salmonn/salmonn/utils.py", line 121, in __next__
[rank3]:     data = next(self.iter_loader)
[rank3]:   File "/home/jpong/miniconda3/envs/salmonn/lib/python3.9/site-packages/torch/utils/data/dataloader.py", li
[rank3]:     data = self._next_data()
[rank3]:   File "/home/jpong/miniconda3/envs/salmonn/lib/python3.9/site-packages/torch/utils/data/dataloader.py", li
[rank3]:     return self._process_data(data, worker_id)
[rank3]:   File "/home/jpong/miniconda3/envs/salmonn/lib/python3.9/site-packages/torch/utils/data/dataloader.py", li
[rank3]:     data.reraise()
[rank3]:   File "/home/jpong/miniconda3/envs/salmonn/lib/python3.9/site-packages/torch/_utils.py", line 769, in rera
[rank3]:     raise exception
[rank3]: soundfile.LibsndfileError: <exception str() failed>
```

Experiment Result – attempt1

result

first epoch

train | 1st epoch | completed

```
{"train_lr": "0.000", "train_loss": "3.291"}
```

eval | 1st epoch | completed

```
{"valid_loss": 3.066974401473999, "valid_agg_metrics": 0.3930814266204834, "valid_best_epoch": 0}
```

Experiment…

- https://github.com/jaeeewon/repr_salmonn/tree/master/configs

reproduce
SALMONN

학생 원종찬, 최재원
Language & AI융합전공

2025. 09. 22. 3 PM.
한국외대 교수회관 401호