[IIAI30003] Digital Speech Processing

Homework 5

Po-Chuan, Chen

Student ID: 311511052

`present90308.ee11@nycu.edu.tw`

National Yang Ming Chiao Tung University

December 17, 2023

Contents

1 Build vits2_pytorch

```
git clone https://github.com/p0p4k/vits2_pytorch
```

Listing 1: Clone vits2

```
pip install -r requirement
sudo apt-get install espeak
```

Listing 2: Install requirements

2 Data Pre-processing

```
ln -s /path/to/LJSpeech-1.1/wavs DUMMY1
```

Listing 3: Create softlink

```
cd monotonic_align
python setup.py build_ext --inplace
```

Listing 4: Build Cython-version Monotonoic Alignment Search

Important! We need to do the same process with our own dataset.

3 Train Model

```
python train.py -c configs/vits2_ljs_nosdp.json -m ljs_base
```

Listing 5: Training

```
python export_onnx.py --model-path="G_64000.pth" --config-path="config.
    json" --output="vits2.onnx"
```

Listing 6: Export trained models to onnx

```
python infer_onnx.py --model="vits2.onnx" --config-path="config.json"
    --output-wav-path="output.wav" --text="huan1 ging5 siu1 khuann3
    kong1 si7 tai5 gi2 tai5, gua2 si7 tsu2 poo3 ong5 sio2 bing5"
```

Listing 7: Inference

4 Other thing

We will face an issue about

```
AttributeError: 'HParams' object has no attribute 'duration_discriminator_type'
```

Listing 8: Issue

We need add Add `"duration_discriminator_type" :"dur_disc_2"`, in the model from `configs/vits2_ljs_nosdp.json`

About LJ speech example

```
{
"train": {
    "log_interval": 200,
    "eval_interval": 1000,
    "seed": 1234,
    "epochs": 20000,
    "learning_rate": 2e-4,
    "betas": [0.8, 0.99],
    "eps": 1e-9,
    "batch_size": 64,
    "fp16_run": false,
    "lr_decay": 0.999875,
    "segment_size": 8192,
    "init_lr_ratio": 1,
    "warmup_epochs": 0,
    "c_mel": 45,
    "c_kl": 1.0
},
"data": {
    "use_mel_posterior_encoder": true,
    "training_files":"filelists/ljs_audio_text_train_filelist.txt.cleaned",
    "validation_files":"filelists/ljs_audio_text_val_filelist.txt.cleaned",
    "text_cleaners":["english_cleaners2"],
    "max_wav_value": 32768.0,
    "sampling_rate": 22050,
    "filter_length": 1024,
    "hop_length": 256,
    "win_length": 1024,
    "n_mel_channels": 80,
    "mel_fmin": 0.0,
    "mel_fmax": null,
    "add_blank": false,
    "n_speakers": 0,
    "cleaned_text": true
},
"model": {
    "use_mel_posterior_encoder": true,
    "use_transformer_flows": true,
    "transformer_flow_type": "pre_conv",
    "use_spk_conditioned_encoder": false,
    "use_noise_scaled_mas": true,
    "use_duration_discriminator": true,
    "inter_channels": 192,
    "hidden_channels": 192,
    "filter_channels": 768,
    "n_heads": 2,
    "n_layers": 6,
    "kernel_size": 3,
    "p_dropout": 0.1,
    "resblock": "1",
    "resblock_kernel_sizes": [3,7,11],
    "resblock_dilation_sizes": [[1,3,5], [1,3,5], [1,3,5]],
    "upsample_rates": [8,8,2,2],
    "upsample_initial_channel": 512,
    "upsample_kernel_sizes": [16,16,4,4],
    "n_layers_q": 3,
    "use_spectral_norm": false,
    "use_sdp": false
}
}
```

Listing 9: config

Change the cleaner to `"text_cleaners":[basic_cleaner"],`

Also

```
"training_files":"filelists/ljs_audio_text_train_filelist.txt.cleaned",
"validation_files":"filelists/ljs_audio_text_val_filelist.txt.cleaned",
```

Listing 10: Train data and valid data

need to be change to `SuiSiann-Dataset`, and it needs to be split into train and valid.

After training about 3000 epochs, we can get good voice from the output wave file.