-Bootcamp-
How to get Started with HW1P2

Pavitra Kadiyala, Denis Musinguzi, Josh Lee
Overview

Dataset of Audio Recordings → Predict Phoneme labels
Workflow

1. Download dataset from kaggle
2. Load dataset into Dataloader
3. Set Hyperparameters (model architecture, lr, optimizer, scheduler, etc)
4. Train and test the model
5. Use best set of hyperparameters to train and submit to Kaggle

Track metrics and change hyperparameters to study the overall effects (e.g., increasing width of layers, changing activations)
Data

Raw Speech Signal (Speech waveform) → Short Time Fourier Transform → Melspectogram
Data

class AudioDataset(torch.utils.data.Dataset):
    
def __init__(self, root, phonemes = PHONEMES, context=0, partition= "train-clean-100"): # Feel free to add more arguments
        self.context = context
        self.phonemes = phonemes

        # TODO: MFCC directory - use partition to acces train/dev directories from kaggle data using root
        self.mfcc_dir = NotImplemented
        # TODO: Transcripts directory - use partition to acces train/dev directories from kaggle data using root
        self.transcript_dir = NotImplemented

        # TODO: List files in self.mfcc_dir using os.listdir in sorted order
        mfcc_names = NotImplemented
        # TODO: List files in self.transcript_dir using os.listdir in sorted order
        transcript_names = NotImplemented

        # Making sure that we have the same no. of mfcc and transcripts
        assert len(mfcc_names) == len(transcript_names)

        self.mfccs, self.transcripts = [], []

        # TODO: Iterate through mfccs and transcripts
        for i in range(len(mfcc_names)):
            # Load a single mfcc
            mfcc = NotImplemented
            # Do Cepstral Normalization of mfcc (explained in writeup)
            # Load the corresponding transcript
            transcript = NotImplemented # Remove [SOS] and [EOS] from the transcript
            # (Is there an efficient way to do this without traversing through the transcript?)
            # Note that SOS will always be in the starting and EOS at end, as the name suggests.
            # Append each mfcc to self.mfcc, transcript to self.transcript
            self.mfccs.append(mfcc)
            self.transcripts.append(transcript)
Dataset

# NOTE:
# Each mfcc is of shape T1 x 28, T2 x 28, ...
# Each transcript is of shape (T1+2) x 28, (T2+2) x 28 before removing [SOS] and [EOS]

# TODO: Concatenate all mfccs in self.mfccs such that
# the final shape is T x 28 (Where T = T1 + T2 + ...)
self.mfccs = NotImplemented

# TODO: Concatenate all transcripts in self.transcripts such that
# the final shape is (T,) meaning, each time step has one phoneme output
self.transcripts = NotImplemented
# Hint: Use numpy to concatenate

# Length of the dataset is now the length of concatenated mfccs/transcripts
self.length = len(self.mfccs)

# Take some time to think about what we have done.
# self.mfcc is an array of the format (Frames x Features).
# Our goal is to recognize phonemes of each frame
# From now, you will be knowing what context is.
# We can introduce context by padding zeros on top and bottom of self.mfcc
self.mfccs = NotImplemented # TODO

# The available phonemes in the transcript are of string data type
# But the neural network cannot predict strings as such.
# Hence, we map these phonemes to integers

# TODO: Map the phonemes to their corresponding list indexes in self.phonemes
self.transcripts = NotImplemented
# Now, if an element in self.transcript is 0, it means that it is 'SIL' (as per the above example)
Running Ablations

Types of Headache

Migraine

Hypertension

Stress

Tuning Hyperparameters
## Running Ablations

<table>
<thead>
<tr>
<th>Epoch</th>
<th>ctx</th>
<th>layers</th>
<th>activations</th>
<th>architecture</th>
<th>batchsize</th>
<th>dropout</th>
<th>BN</th>
<th>optimizer</th>
<th>scheduler</th>
<th>weight init</th>
<th>Regularization</th>
<th>Initial LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0</td>
<td>2</td>
<td>relu</td>
<td>Pyramid (max(1024, 10^D) → 128)</td>
<td>256</td>
<td>none</td>
<td>none</td>
<td>ADAM</td>
<td>steplr</td>
<td>gaussian</td>
<td>none</td>
<td>0.001</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>2</td>
<td>relu</td>
<td>Pyramid (max(1024, 10^D) → 128)</td>
<td>256</td>
<td>none</td>
<td>none</td>
<td>ADAM</td>
<td>steplr</td>
<td>gaussian</td>
<td>none</td>
<td>0.001</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>2</td>
<td>relu</td>
<td>Pyramid (max(1024, 10^D) → 128)</td>
<td>256</td>
<td>none</td>
<td>none</td>
<td>ADAM</td>
<td>steplr</td>
<td>gaussian</td>
<td>none</td>
<td>0.001</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>2</td>
<td>relu</td>
<td>Pyramid (max(1024, 10^D) → 128)</td>
<td>256</td>
<td>0.25 every layer</td>
<td>every layer preactivation</td>
<td>ADAM</td>
<td>reduce on plateau</td>
<td>xavier</td>
<td>none</td>
<td>0.001</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>4</td>
<td>splus</td>
<td>inverted pyramid (max(2D, 128) → 4D D→→ → 2048)</td>
<td>512</td>
<td>0.25 every layer</td>
<td>every layer preactivation</td>
<td>ADAM</td>
<td>reduce on plateau</td>
<td>xavier</td>
<td>none</td>
<td>0.001</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>4</td>
<td>splus</td>
<td>inverted pyramid (max(2D, 128) → 4D D→→ → 2048)</td>
<td>512</td>
<td>0.25 every layer</td>
<td>every layer preactivation</td>
<td>ADAM</td>
<td>reduce on plateau</td>
<td>xavier</td>
<td>none</td>
<td>0.001</td>
</tr>
</tbody>
</table>

| 5 | 16 | 4 | splus | inverted pyramid (max(2D, 128) → 4D D→→ → 2048) | 512 | 0.25 every layer | every layer preactivation | ADAM | reduce on plateau | xavier | none | 0.001 |
| 5 | 16 | 4 | splus | inverted pyramid (max(2D, 128) → 4D D→→ → 2048) | 512 | 0.25 every layer | every layer preactivation | ADAM | reduce on plateau | xavier | none | 0.001 |

---

Carnegie Mellon University
Running Ablations
Running Ablations

- Progressively build on your experiments
- Incorporate some domain knowledge
- Start with several simple architectures
High Cutoff Architecture

https://www.youtube.com/watch?v=dQw4w9WgXcQ
*wink*