Training wav2vec on Multiple Languages From Scratch

Heting Gao, Mahir Morshed, Shuju Shi, Liming Wang, Junkai Wu
Introduction

• Large amount of parallel speech-text data not available in most languages
• wav2vec: a new paradigm of training an ASR system by splitting the training process into two stages:
  • Self-supervised pretraining (only unlabeled audio is required)
  • Low-resource finetuning (small amount of parallel speech-text data is required)
Introduction

• Model Architecture
  • Multi-layer convolutional feature encoder
  • Quantization module to discretize the features into codewords in a codebook
  • Transformer encoder to output contextual representation of each frame
• Pretraining to predict the codeword of the current frame using content representation
  • Contrastive loss with negative sampling
  • Codebook diversity loss
• Finetuning
  • Additional linear layers and CTC loss
  • Freeze the pretrained wav2vec and finetune only the linear layers
Introduction

• Previous works:
  • wav2vec 2.0¹ only trained the model on English
  • XLSR-53² trained a multilingual wav2vec on a 53-language dataset
    • Multilingual wav2vec has a better cross-lingual performance
  • VoxPopuli³ released a large-scale multilingual speech corpus
    • 23 languages of the European Union
      • of which 16 partially have transcriptions
    • Pretrained wav2vec models provided

• Objectives:
  • How does wav2vec work with other languages (such as Asian languages)?
  • How does a wav2vec pretrained on different languages affect performance?

Datasets considered
(chosen languages highlighted)

- Multilingual LibriVox
  - **English (en)** (36000 h), German (1691 h), Dutch (1264 h), French (897 h), Spanish (735 h), Italian (220 h), Portuguese (136 h), Polish (91 h)
- LibriLight for **English (ft-en)** (10h used)
- LaboroTVSpeech for **Japanese (ja)** (2000 h)
- Corpus of Spontaneous **Japanese (ft-ja)** (661 h)
- Babel (200 h per language)
  - Bengali, Vietnamese, Zulu, Amharic, Javanese, Georgian, Cantonese, Lao
- GlobalPhone (20 h per language)
  - Czech, French, **Mandarin (ft-zh)**, Thai, German, Portuguese, Turkish, **Bulgarian (ft-bg)**, Croatian, Spanish, Polish
- OpenSLR (all but the last two from Google)
  - Javanese (295 h), Sundanese (332 h), Sinhala (225), Bengali (215 h), Nepali (154 h), Korean (51 h), Kazakh (332 h)
- Europarl-ST
  - English (637 h), French (176 h), German (153 h), Italian (181 h), Spanish (116 h), Portuguese (82 h), Polish (151 h), Romanian (108 h), Dutch (38 h)
- United Nations Proceedings Speech (~1000h per language)
  - English, **Mandarin (zh)**, **Standard Arabic (ar)**, French, **Russian (ru)**, Spanish
- Common Voice
  - Kinyarwanda (2000 h), Esperanto (1300 h), Catalan (1200 h), Belarusian (1000 h)
- VoxPopuli (2.7k to 24.1k h)
  - English, German, French, Spanish, Polish, Italian, Romanian, Hungarian, Czech, Dutch, Finnish, Croatian, Slovak, Slovene, Estonian, Lithuanian, Portuguese, **Bulgarian (bg)**, Greek, Latvian, Maltese, Swedish, Danish
- GALE Broadcast News Datasets for **Standard Arabic (ft-ar)** (~120 h for Phase 3 Part 2)
- **Russian (ft-ru)** LibriSpeech (~100 h)
Experiment (English Baseline)

- Hardware
  - NCSA's HAL cluster
    - 4 x 16 GB NVIDIA V100 GPU
- wav2vec model settings
  - wav2vec base model has max token per batch: 1M (originally was 1.4M)
  - XLSR-53 model has max token per batch: 600k (originally was 1.28M)
  - Update frequency is 16 to simulate 64 GPU training (2 weeks)
- English baseline
  - Better performance (unit error rate) when finetuned on English
  - Slightly worse performance when finetuned on Bulgarian

<table>
<thead>
<tr>
<th></th>
<th>Validation UER</th>
<th>Validation WER</th>
<th>Test UER</th>
</tr>
</thead>
<tbody>
<tr>
<td>official-en-ft-en</td>
<td></td>
<td>10.9</td>
<td></td>
</tr>
<tr>
<td>en-ft-en</td>
<td>3.51</td>
<td><strong>9.87</strong></td>
<td>2.97</td>
</tr>
<tr>
<td>official-en-ft-bg</td>
<td>3.34</td>
<td>17.37</td>
<td><strong>3.31</strong></td>
</tr>
<tr>
<td>en-ft-bg</td>
<td>3.48</td>
<td>17.68</td>
<td>3.47</td>
</tr>
</tbody>
</table>
Experiment (Mono- vs Cross- vs Multi-lingual)

- Three settings
  - Monolingual finetuning
  - Cross-lingual finetuning (English)
  - Multilingual finetuning (XLSR-53)
- Monolingual finetuning > Multilingual finetuning > Cross-lingual finetuning
  - English and Russian excepted

<table>
<thead>
<tr>
<th>Mono</th>
<th>Train Loss</th>
<th>Valid UER</th>
<th>Valid WER</th>
<th>Test UER</th>
<th>Cross</th>
<th>Valid UER</th>
<th>Valid WER</th>
<th>Test UER</th>
<th>Multi</th>
<th>Valid UER</th>
<th>Valid WER</th>
<th>Test UER</th>
</tr>
</thead>
<tbody>
<tr>
<td>en-ft-en</td>
<td>0.07</td>
<td>3.51</td>
<td>9.87</td>
<td>2.97</td>
<td>en-ft-en</td>
<td>3.51</td>
<td>9.87</td>
<td>2.97</td>
<td>xlsr-ft-en</td>
<td>1.91</td>
<td>6.58</td>
<td>1.91</td>
</tr>
<tr>
<td>bg-ft-bg</td>
<td>1.53</td>
<td>1.85</td>
<td>8.78</td>
<td>1.89</td>
<td>en-ft-bg</td>
<td>3.48</td>
<td>17.68</td>
<td>3.47</td>
<td>xlsr-ft-bg</td>
<td>2.67</td>
<td>13.79</td>
<td>2.70</td>
</tr>
<tr>
<td>zh-ft-zh</td>
<td>2.00</td>
<td>10.43</td>
<td>-</td>
<td>10.60</td>
<td>en-ft-zh</td>
<td>15.20</td>
<td>-</td>
<td>15.41</td>
<td>xlsr-ft-zh</td>
<td>14.15</td>
<td>-</td>
<td>14.56</td>
</tr>
<tr>
<td>ru-ft-ru</td>
<td>1.95</td>
<td>5.57</td>
<td>23.06</td>
<td>6.98</td>
<td>en-ft-ru</td>
<td>5.57</td>
<td>27.92</td>
<td>5.59</td>
<td>xlsr-ft-ru</td>
<td>5.84</td>
<td>28.21</td>
<td>4.84</td>
</tr>
<tr>
<td>ar-ft-ar</td>
<td>1.77</td>
<td>3.62</td>
<td>12.32</td>
<td>3.45</td>
<td>en-ft-ar</td>
<td>6.49</td>
<td>20.41</td>
<td>5.47</td>
<td>xlsr-ft-ar</td>
<td>4.67</td>
<td>17.56</td>
<td>4.58</td>
</tr>
</tbody>
</table>
Experiment (Grapheme vs IPA)

- Convert graphemes into International Phonetic Alphabet (IPA) using LanguageNet Grapheme-to-Phoneme Transducers (g2ps)
- Expected IPA transcripts to have lower error rates
  - Turns out not to be the case generally
    - Bulgarian and Mandarin have much lower error rates using graphemes
    - English, Russian and Arabic have slightly lower error rates using graphemes
    - Japanese has much lower error rate using IPA

<table>
<thead>
<tr>
<th>Mono</th>
<th>Test UER</th>
<th>Test UER IPA</th>
<th>Cross</th>
<th>Test UER</th>
<th>Test UER IPA</th>
<th>Multi</th>
<th>Test UER</th>
<th>Test UER IPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>en-ft-en</td>
<td>2.97</td>
<td>3.19</td>
<td>en-ft-en</td>
<td>2.97</td>
<td>3.20</td>
<td>xlsr-ft-en</td>
<td>1.91</td>
<td>2.29</td>
</tr>
<tr>
<td>bg-ft-bg</td>
<td>1.89</td>
<td>11.28</td>
<td>en-ft-bg</td>
<td>3.47</td>
<td>14.86</td>
<td>xlsr-ft-bg</td>
<td>2.70</td>
<td>17.38</td>
</tr>
<tr>
<td>zh-ft-zh</td>
<td>10.60</td>
<td>15.03</td>
<td>en-ft-zh</td>
<td>15.41</td>
<td>18.05</td>
<td>xlsr-ft-zh</td>
<td>14.56</td>
<td>16.10</td>
</tr>
<tr>
<td>ru-ft-ru</td>
<td>6.98</td>
<td>7.87</td>
<td>en-ft-ru</td>
<td>5.59</td>
<td>6.10</td>
<td>xlsr-ft-ru</td>
<td>4.84</td>
<td>5.26</td>
</tr>
<tr>
<td>ar-ft-ar</td>
<td>3.45</td>
<td>4.20</td>
<td>en-ft-ar</td>
<td>5.47</td>
<td>6.55</td>
<td>xlsr-ft-ar</td>
<td>4.58</td>
<td>5.24</td>
</tr>
<tr>
<td>jp-ft-jp</td>
<td>9.91</td>
<td><strong>3.32</strong></td>
<td>en-ft-jp</td>
<td>16.33</td>
<td><strong>4.98</strong></td>
<td>xlsr-ft-jp</td>
<td>14.35</td>
<td><strong>4.51</strong></td>
</tr>
</tbody>
</table>
Experiment (Mandarin)

- Extra experiments on Mandarin Chinese
  - Chinese characters
  - IPA with or without tone
  - Pinyin with or without tone
- Converting to Pinyin phonemes can greatly reduce the error rate
- g2ps probably contain errors when transducing from Chinese characters to IPA
- wav2vec can capture tone information very well

<table>
<thead>
<tr>
<th>zh-ft-zh</th>
<th>Test UER</th>
<th>en-ft-zh</th>
<th>Test UER</th>
<th>xlsr-ft-zh</th>
<th>Test UER</th>
</tr>
</thead>
<tbody>
<tr>
<td>char</td>
<td>10.60</td>
<td>char</td>
<td>15.41</td>
<td>char</td>
<td>14.56</td>
</tr>
<tr>
<td>IPA</td>
<td>15.03</td>
<td>IPA</td>
<td>18.05</td>
<td>IPA</td>
<td>16.10</td>
</tr>
<tr>
<td>IPA w/o tone</td>
<td>14.56</td>
<td>IPA w/o tone</td>
<td>16.94</td>
<td>IPA w/o tone</td>
<td>15.44</td>
</tr>
<tr>
<td>Pinyin</td>
<td>2.96</td>
<td>Pinyin</td>
<td>3.80</td>
<td>Pinyin</td>
<td>4.02</td>
</tr>
<tr>
<td>Pinyin w/o tone</td>
<td>2.31</td>
<td>Pinyin w/o tone</td>
<td>2.97</td>
<td>Pinyin w/o tone</td>
<td>2.79</td>
</tr>
</tbody>
</table>
Experiment (Japanese Kana)

- Extra experiments on Japanese Kana
- Use wav2vec pretrained on Japanese, English, Mandarin Chinese, Spanish and XLSR-53
- Finetune on Japanese
  - Cross-lingually, English > Spanish > Mandarin Chinese
  - Multilingual XLSR-53 is better than the cross-lingual models.

<table>
<thead>
<tr>
<th>Kana</th>
<th>Test UER</th>
</tr>
</thead>
<tbody>
<tr>
<td>jp-ft-jp</td>
<td>5.08</td>
</tr>
<tr>
<td>en-ft-jp</td>
<td>8.18</td>
</tr>
<tr>
<td>es-ft-jp</td>
<td>8.60</td>
</tr>
<tr>
<td>zh-ft-jp</td>
<td>9.66</td>
</tr>
<tr>
<td>xlsr-ft-jp</td>
<td>7.16</td>
</tr>
</tbody>
</table>