EasyASR: A Distributed Machine Learning Platform for End-to-end Automatic Speech Recognition

Chengyu Wang, Mengli Cheng, Xu Hu, Jun Huang

Alibaba Group
Contents

1 Introduction

2 Platform Description

3 Conclusion
Introduction (Background)

- Deep neural network based ASR models have large performance gain.

- Large ASR models bring additional challenges:
  - Require abundant labeled training data for learning large models (labor-intensive, financially expensive)
  - Need an efficient distributed, computing framework for model training and serving at scale
Introduction (EasyASR)

EasyASR: a distributed machine learning platform to address both challenges.

- Support weakly supervised extraction of wave-transcript pairs and training data augmentation
- Built upon the Machine Learning Platform for AI (PAI) of Alibaba Cloud for efficient distributed model learning and inference
- Achieve state-of-the-art results for Mandarin speech recognition
Platform Description (Function Design)

- Extract audio-transcript pairs from massive video data without labeling*
- Extract features of in TFRecord
- Enlarge training sets via data augmentation
- Train/fine-tune ASR models on distributed GPU clusters
- Support automatic evaluation and model export

* Refer to the paper “Weakly Supervised Construction of ASR Systems with Massive Video Data” arXiv 2020
Platform Description (System Design)

✓ Key elements in EasyASR to support efficient distributed learning and inference
  • PAI TensorFlow: deeply optimized in communication, thread, memory allocation and I/O
  • PAISoar: significantly speeds up the training process distributed across multiple workers and GPUs

✓ Comparison against other frameworks
  • Examples: Kaldi, OpenSeq2Seq, ESPNet, wav2letter++, etc.
  • EasyASR: integrates our ASR library with PAI for efficient distributed learning
Platform Description (User Interface)

✔ Simple PAI commands (example for ASR_Train)

```bash
PAI -name ASR_Train -Dfinetune=false
-Dconfig='your_path/model_config'
-Dexport='your_path/model_export_dir'
-Dcluster='{"worker": {"count": 4, "cpu": 2000, "gpu": 800, "memory": 100000}}';
```

Computational resources on PAI cluster

✔ Model configuration (example for our transformer model)

```json
"encoder": TransformerEncoder,
"encoder_params": {
   "encoder_layers": 12, "num_heads": 8...
 },
"decoder": JointCTCAattenDecoder,
"decoder_params": {
   "attn_decoder": TransformerDecoder,
   "attn_decoder_params": {
      "hidden_layers": 6,"num_heads": 8...
   },
   "ctc_decoder": CTCDecoder,
   "ctc_decoder_params": {...},
 },
"loss": MultiTaskCTCEntropyLoss,
"loss_params": {
   "seq_loss_params": {...},
   "ctc_loss_params": {...},
   "lambda_value": 0.30,
}
```
Platform Description (Performance)

✓ State-of-the-art results for Mandarin speech recognition

<table>
<thead>
<tr>
<th>Model</th>
<th>ST_CMDS</th>
<th>AISHELL-1</th>
<th>AISHELL-2</th>
<th>AIDATANG</th>
<th>MagicData</th>
<th>HKUST</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDNN [12]</td>
<td>-</td>
<td>8.7</td>
<td>-</td>
<td>7.2</td>
<td>-</td>
<td>32.7</td>
</tr>
<tr>
<td>Chain-Model [13]</td>
<td>-</td>
<td>7.5</td>
<td>-</td>
<td>5.6</td>
<td>-</td>
<td>28.1</td>
</tr>
<tr>
<td>MS-Attn [18]</td>
<td>-</td>
<td>-</td>
<td>8.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SpeechBERT [32]</td>
<td>-</td>
<td>7.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>21.0</td>
</tr>
<tr>
<td>SAN-M [19]</td>
<td>-</td>
<td>6.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>wav2letter (w/o. WSP)</td>
<td>4.5</td>
<td>11.7</td>
<td>12.5</td>
<td>12.9</td>
<td>7.4</td>
<td>35.7</td>
</tr>
<tr>
<td>wav2letter (w. WSP)</td>
<td>2.4</td>
<td>7.1</td>
<td>10.0</td>
<td>9.2</td>
<td>6.7</td>
<td>29.3</td>
</tr>
<tr>
<td>Speech Transformer (w/o. WSP)</td>
<td>4.4</td>
<td>6.7</td>
<td>7.4</td>
<td>7.8</td>
<td>3.6</td>
<td>23.5</td>
</tr>
<tr>
<td>Speech Transformer (w. WSP)</td>
<td>2.1</td>
<td>5.9</td>
<td>5.9</td>
<td>4.9</td>
<td>3.3</td>
<td>20.0</td>
</tr>
</tbody>
</table>

* Refer to the paper “Weakly Supervised Construction of ASR Systems with Massive Video Data” arXiv 2020
Conclusion

✓ EasyASR: a distributed machine learning platform for end-to-end ASR models
  • Efficient model learning and inference across multiple workers and GPUs
  • Simple user interface (PAI commands)
  • State-of-the-art performance for Mandarin speech recognition

✓ Future work
  • Developing EasyASR to support more state-of-the-art ASR models
  • Making EasyASR publicly available
THANKS

-------- Q&A Section --------