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

Chengyu Wang, Mengli Cheng, Xu Hu, Jun Huang

Alibaba Group

Contents

Introduction (Background)

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

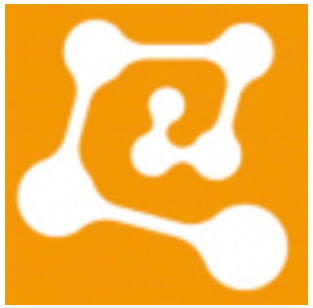
✓Large ASR models bring additional challenges:

- 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



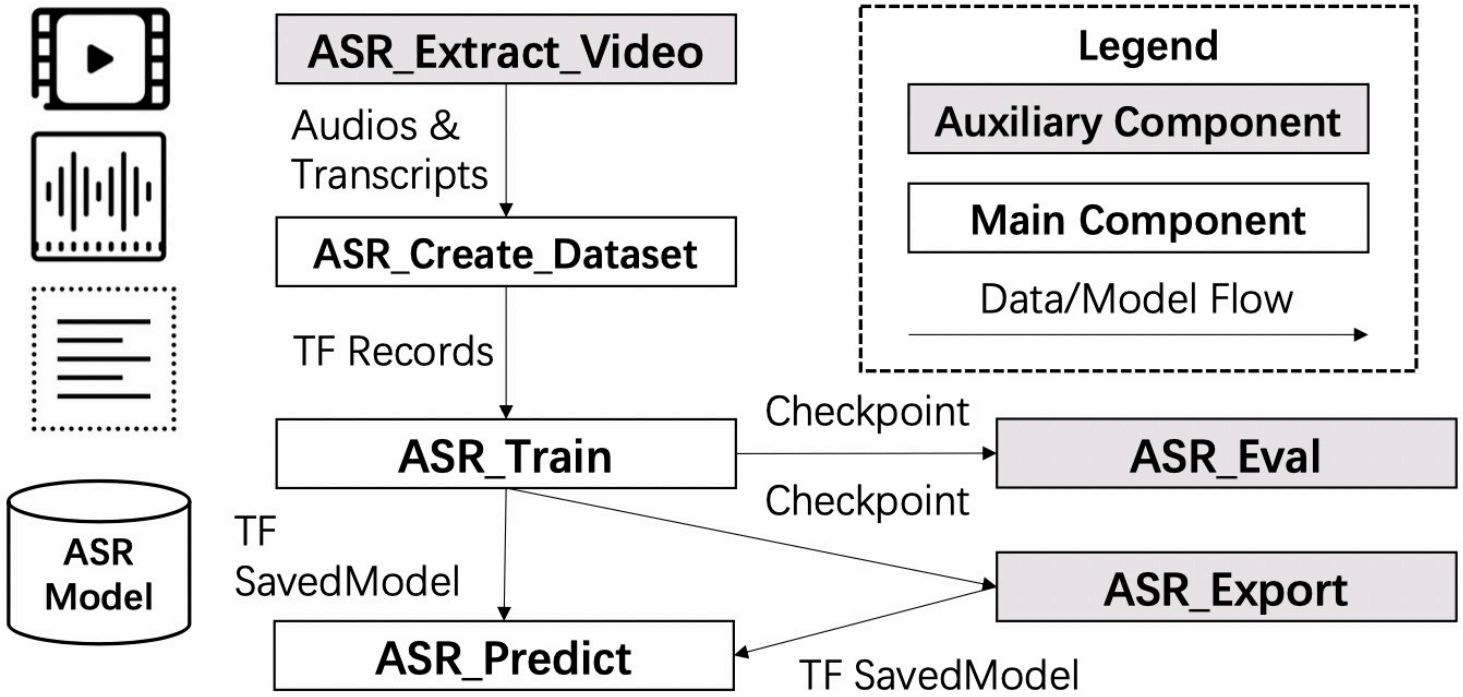
Machine Learning
Platform for AI

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



- Fast model inference

- Customized model evaluation and 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

KALDI ESPnet facebook research NVIDIA

Platform Description (User Interface)

✓Simple PAI commands (example for ASR_Train)

```
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)

```
"encoder": TransformerEncoder,
"encoder_params": {
    "encoder_layers": 12, "num_heads": 8...
  },
"decoder": JointCTCAttenDecoder,
"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

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

* 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

Alibaba Group
阿里巴巴集团

THANKS

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