ASR

AUTO SPEECH RECOGNITION

Saruj Chutapornpong

2110391 INDIVDUAL STUDY IN COMPUTER ENGINEERING

Table of Contents

ASR

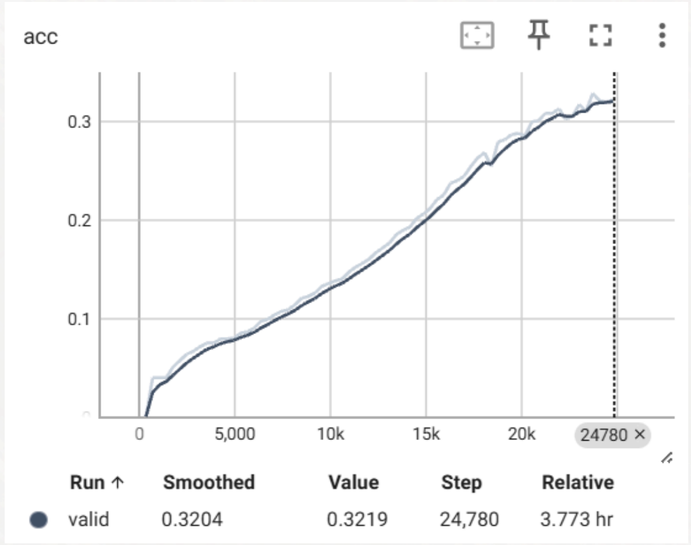

Speech as Computer Inputs
14 บทเรียน
Automatic Speech Recognition
EkapolC · หลักสูตร
ดูทั้งหลักสูตร
Levenshtein distance (edit distance)
• The minimum number of word/character edits (insertions, deletions or substitutions) required to change one sentence into the other
18 บทเรียน
Fall2023-SpeechRecognition&Understan...
WAVLab · หลักสูตร
ดูทั้งหลักสูตร

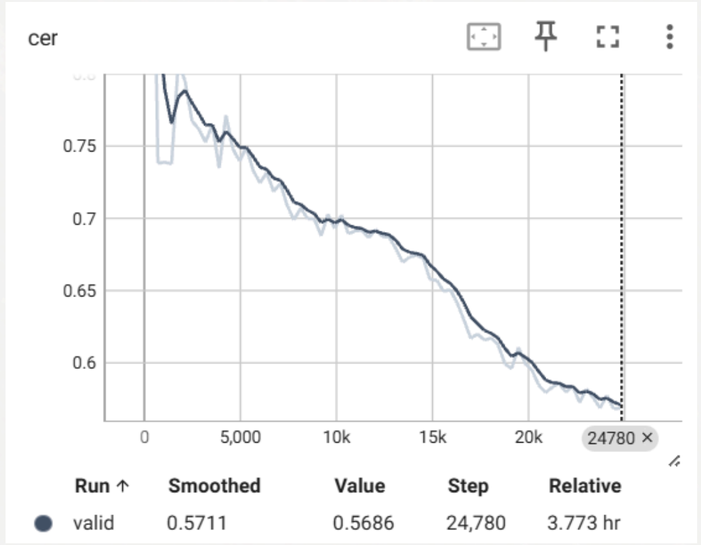
ESPnet

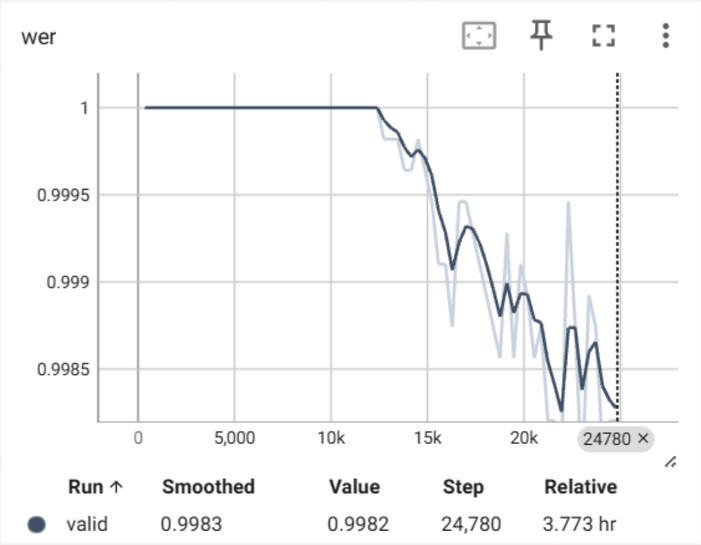

ESPnet
Tutorials
Demos
Recipes
Python API
Shell API
Search Ctrl K
ESPnet
End-to-end Speech Processing toolkit
ESPnet is the state-of-the-art toolkit that covers end-to-end speech recognition, text-to-speech, speech translation, speech enhancement, speaker diarization, spoken language understanding, and much more!
Get started with ESPnet!
Running inference on existing ESPnet models
pip install espnet espnet-
Fine-tuning ESPnet models
pip install espnet and use
Complete installation to fully reproduce ESPnet models

Sample demo for ESPnet-EZ[1]

Train an Automatic Speech Recognition (ASR) model using the Librispeech-100 dataset.







Train with 10 hours of Librispeech dataset

Result

1

Sentence:

LONG DURATION OTHER INDICATIONS OF LONG DURATION I THINK OF A REGION SOMEWHERE ABOVE THIS EARTH'S SURFACE IN WHICH GRAVITATION IS INOPERATIVE AND IS NOT GOVERNED BY THE SQUARE OF THE DISTANCE

Prediction:

LONGENTION ARE IN EDUCATIONATIONS OF CONSCIENCEATION I THINK A READING SOME MORE ABOUT THIS THIRDS IN WHICH TRATION AS IN ORDER AND IS NOT GATHER BY THIS SPIRIT OF THE BUSINESS

Result

2

Sentence:

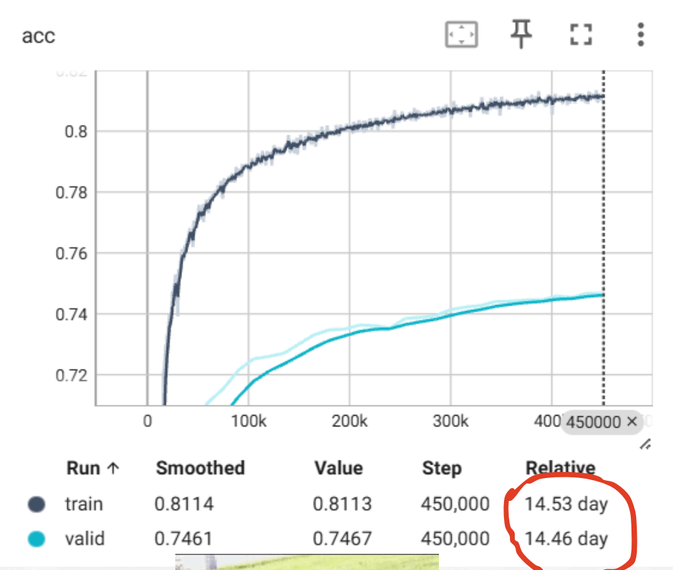
ONLY BECAUSE I COULD NOT HIRE A HORSE TO GO OR MY FELLOW WAS RUN AWAY THAT WAS TO ATTEND ME WAS RIDICULOUS SINCE AT THE TIME I HAD MY HEALTH AND LIMBS AND OTHER SERVANTS AND MIGHT

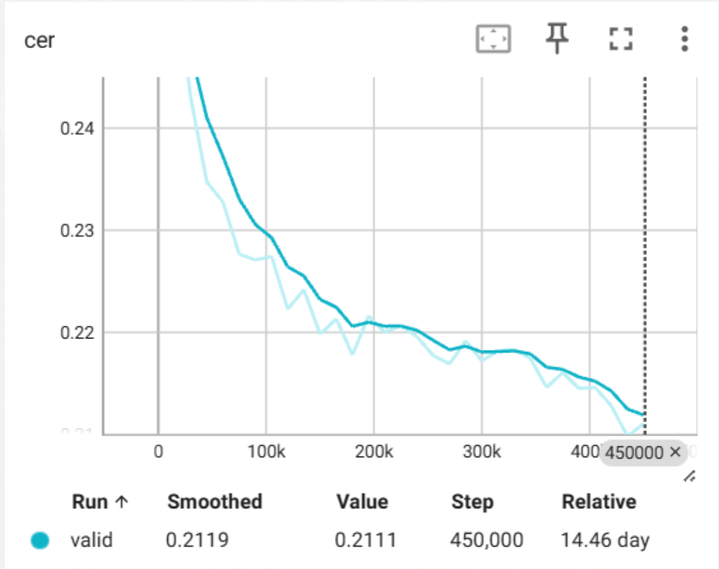
Prediction:

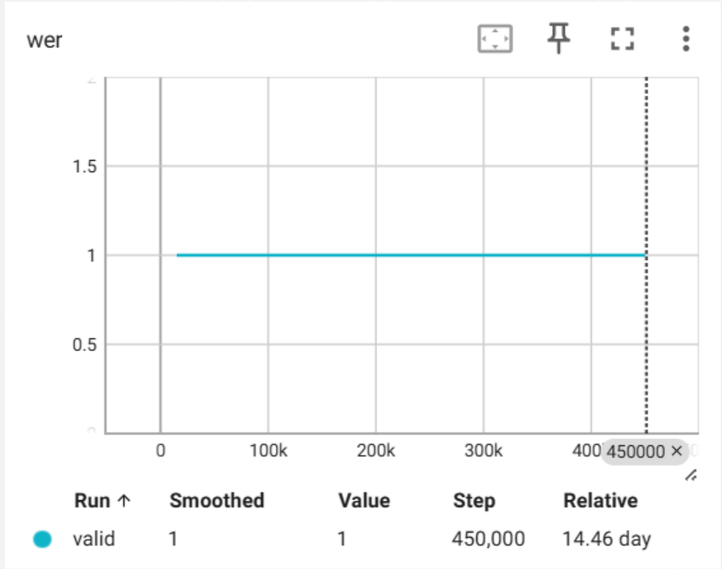
ONLY BECAUSE I COULD NOT HIGHER A FIRST TO GO OR MY FELT IT WAS ON HER WAY THAT WAS TO A MINUTE ME WAS TO THE TIME I HAD MY HELP AND ARMS AND OTHER SOLDIERSANCE IN MY

OWSM finetuning with custom dataset

Finetuning an OWSM model for ASR task with ESPnetEZ







Train with 40 hours of Korat dialect dataset

Result

1

Sentence:

ป้า ขาย เข่าหอมมะลิ ทุ่งกุ ลา ร่อง ไห่ ห่อ ละ หนึ่งหมื่น สามพัน บาท ค่ะ

Prediction:

เบำคายข้อรมริด ทุงกราร่างหาย , เหริงเมื่องสำเพรมายขา

2

Sentence:

มีสินค่าทั่งหม๊ด สิ้บเอ็ด กิโล ค่ะ

Prediction:

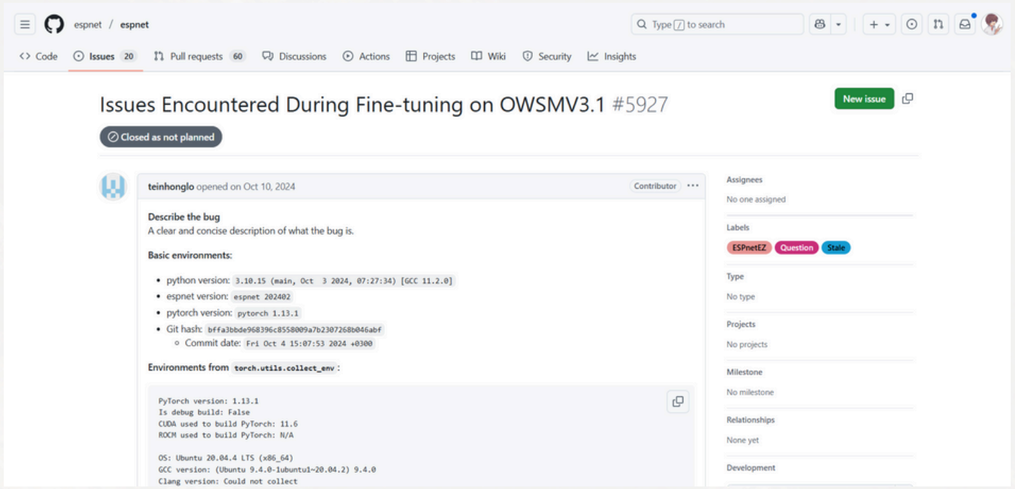
เมื่อสิมเสิมสิมเชื่องมดสิดเคิรู้ค

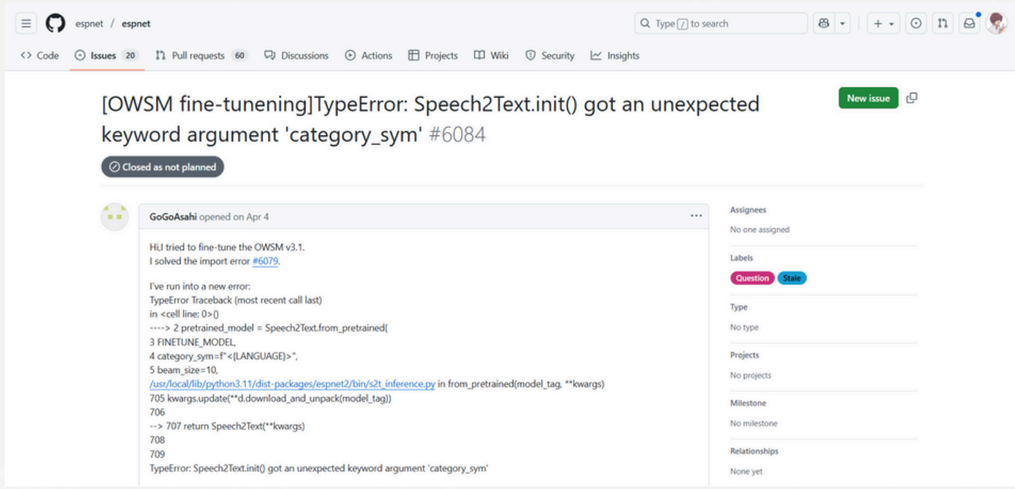
PROBLEM

Problem

1. Limited Hardware

2. Too many fault in the provided script making following along difficult





Limited Hardware

```
torch.OutOfMemoryError: CUDA out of memory. Tried to allocate 7161.40 GiB. GPU 0 has a total capacity of 4.00 GiB of whi
ch 1.66 GiB is free. Including non-PyTorch memory, this process has 17179869184.00 GiB memory in use. Of the allocated m
emory 451.71 MiB is allocated by PyTorch, and 1.08 GiB is reserved by PyTorch but unallocated. If reserved but unallocat
ed memory is large try setting PYTORCH_CUDA_ALLOC_CONF=expandable_segments:True to avoid fragmentation. See documentatio
n for Memory Management (https://pytorch.org/docs/stable/notes/cuda.html#environment-variables)
```

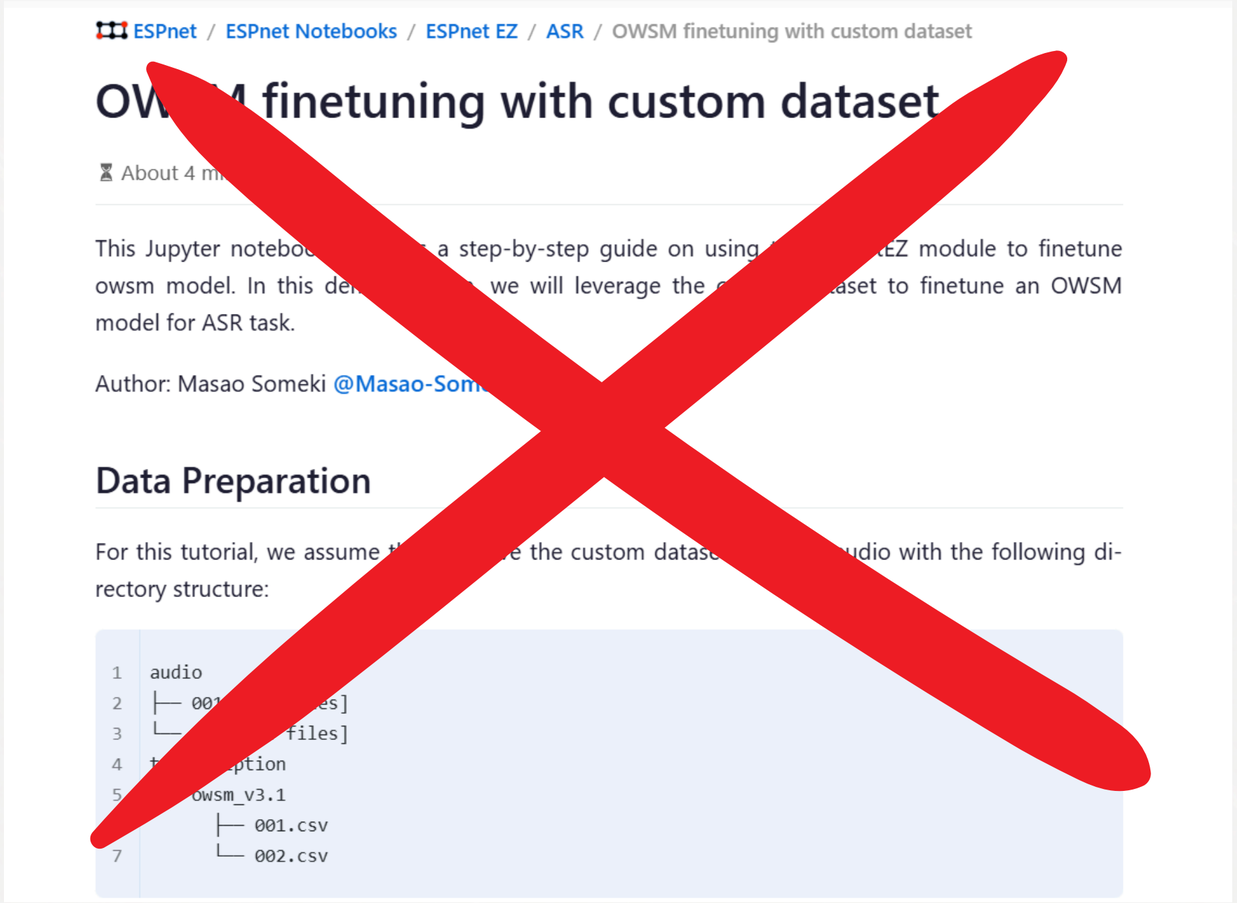
- As we pre-train larger models, full fine-tuning, which retrains all model parameters, becomes less feasible.
- OWSM has 100M parameters, though possible to train, but it will take a very long time.

Introducing LoRA (Hu et al., 2022)[2]

- Low-Rank Adaptation, or LoRA, freezes the pretrained model weights and injects trainable rank decomposition matrices into each layer of the Transformer architecture, greatly reducing the number of trainable parameters for downstream tasks.
- LoRA can reduce the number of trainable parameters by 10,000 times and the GPU memory requirement by 3 times.
- LoRA performs on-par or better than finetuning in model quality on RoBERTa, DeBERTa, GPT-2, and GPT-3, despite having fewer trainable parameters, a higher training throughput, and, unlike adapters, no additional inference latency.

Unreliable Guide

- Take too long
- Use better guide that have similar setting
- Let’s try following the paper of this corpus



Transfer-based Curriculum Learning

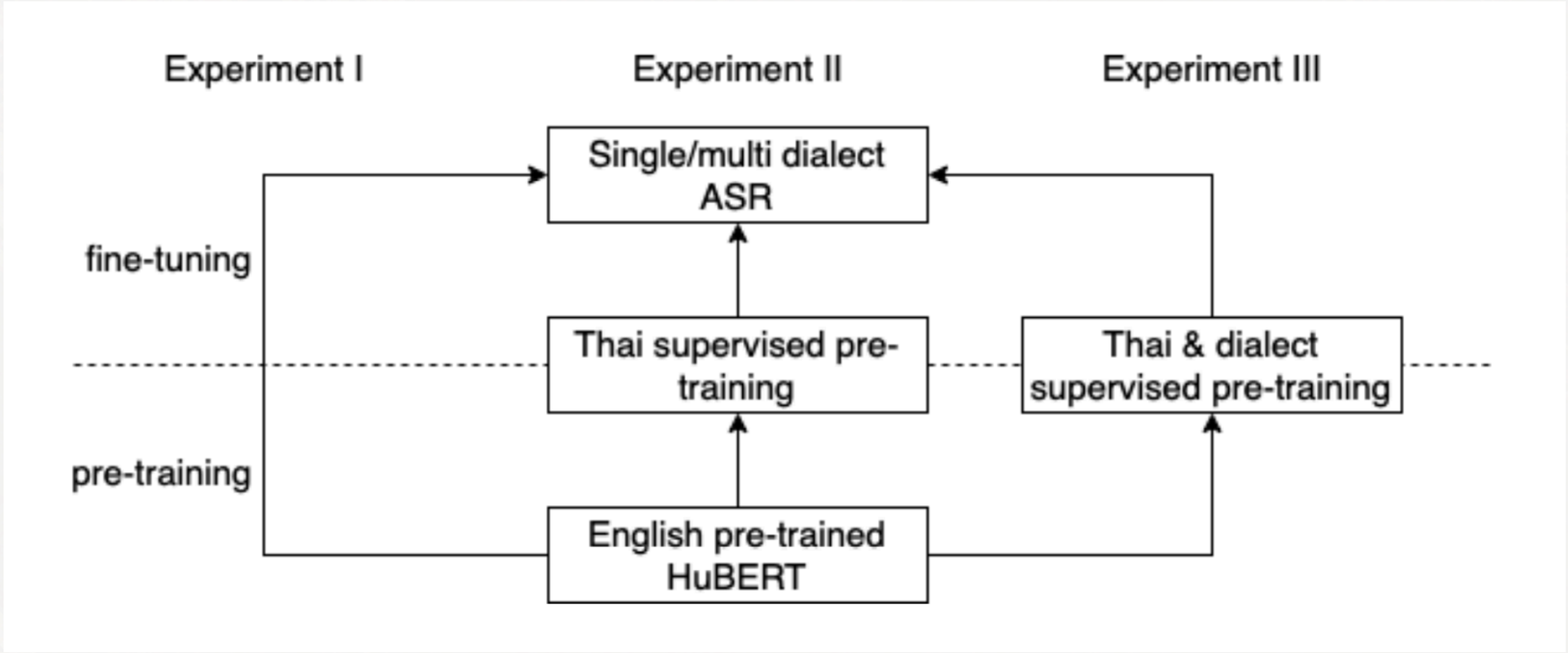
(Chuangsuwanich et al., 2023)[3]

- Thai dialectal languages often have a high degree of similarity in speech to the main dialect.
- Chuangsuwanich, et al. demonstrated transfer learning helps the most if the target language is more similar to the source language.
- Curriculum learning (CL) is an approach for ordering data to optimize training efficiency, inspired by the “starting small” concept. When used correctly, CL can reduce training steps and improve accuracy.

Setup

Plan

Reproduce an experiment



Only Thai supervised pre-traning model available, so we'll go with experiment II

Objective

Beat or match the official result (Experiment III) with inferior setup

CATEGORY	Count	Micro CER	Macro CER	Micro WER	Macro WER
All	1080	11.12%	14.30%	16.49%	20.70%
ECOM	698	8.71%	9.06%	11.80%	12.60%
SURV	382	20.35%	23.88%	31.00%	35.49%

*Test on validation set using PyThaiNLP

Execution

Attempt 1

Use the same setup as OWSM but change to Thai supervised pre-traning

CATEGORY	Count	Micro CER	Macro CER	Micro WER	Macro WER
All	1080	23.32%	25.68%	44.70%	47.75%
ECOM	698	22.00%	21.76%	44.47%	46.34%
SURV	382	28.29%	32.86%	45.39%	50.34%

Result

1

Sentence:

ป้า ขาย เข่าหอมมะลิ ทุ่งกุ ลา ร่อง ไห่ ห่อ ละ หนึ่งหมื่น สามพัน บาท ค่ะ

Prediction:

ป้า ขาย ข้าว หอมมะลิ ทุ่งกุลา ร้องไห ห่อ ละ หนึ่ง หมื่น สาม พัน บาท ค่ะ

2

Sentence:

มีสินค่าทั่งหม๊ด หนึ่งพันสามร่อย ซอง ครับ

Prediction:

มี สินค้า ทั้งหม้ด หนึ่ง พัน สาม ร้อย ซอง ครับ

The result improve immensely

เบำคายข้อรมริด ทุงกราร่างหาย , เหริงเมื่องสำเพรมายขา ➡ ป้า ขาย ข้าว หอมมะลิ ทุ่งกุลา ร้องไห ห่อ ละ หนึ่ง หมื่น สาม พัน บาท ค่ะ

Result

1

Sentence:

ป้า ขาย **เข่า**หอมมะลิ ทุ่งกุ ลา **ร่อง ไห่** ห่อ ละ หนึ่งหมื่น สามพัน บาท ค่ะ

Prediction:

ป้า ขาย **ข้าว** หอมมะลิ ทุ่งกุลา **ร้องไห** ห่อ ละ หนึ่ง หมื่น สาม พัน บาท ค่ะ

2

Sentence:

มี**สินค่า**ทั่งหม๊ด หนึ่งพันสาม**ร่อย** ซอง ครับ

Prediction:

มี **สินค้า** ทั้งหม้ด หนึ่ง พัน สาม **ร้อย** ซอง ครับ

But there’s a problem

Central thai word is correct. Dialect on the other hand, not so much.

What happened?

The model wasn’t trained properly.

Maybe LoRA rank is too small?

What happened?

The model wasn’t trained properly.

Maybe LoRA rank is too small? ❌

Not really.

In fact, it doesn’t even matter.

What happened?

The model wasn’t trained properly.

> Dettmers et al. (2023) test LoRA rank = { 8, 16, 32, 64, 128, 256} and find that LoRA rank is unrelated to final performance if LoRA is used on all layers
>
> (Dettmers et al., 2023)[4]

What happened?

The model wasn’t trained properly.

> Dettmers et al. (2023) test LoRA rank = {8, 16, 32, 64, 128, 256} and find that LoRA rank is unrelated to final performance if LoRA is used on all layers
>
> (Dettmers et al., 2023)[4]

The authors discovered that **Where** you put the adapters matters much more than How **Big** they are.

Fix

Change LoRA targert

```
LORA_TARGET = ["w_1", "w_2", "merge_proj"]
```

```
LORA_TARGET = [
    "fc1", "fc2",           # Encoder Feed-Forward
    "w_1", "w_2",           # Decoder Feed-Forward
    "q_proj", "v_proj",     # Encoder Attention
    "ctc_lo"                # CTC Classification Head
]
```

Attempt 2

Change LoRA target

CATEGORY	Count	Micro CER	Macro CER	Micro WER	Macro WER
All	1080	13.20%	17.01%	18.99%	23.36%
ECOM	698	10.05%	10.17%	13.23%	13.72%
SURV	382	25.22%	29.51%	36.81%	40.96%

CATEGORY	Count	Micro CER	Macro CER	Micro WER	Macro WER
All	1080	23.32%	25.68%	44.70%	47.75%
ECOM	698	22.00%	21.76%	44.47%	46.34%
SURV	382	28.29%	32.86%	45.39%	50.34%

OLD

CATEGORY	Count	Micro CER	Macro CER	Micro WER	Macro WER
All	1080	13.20%	17.01%	18.99%	23.36%
ECOM	698	10.05%	10.17%	13.23%	13.72%
SURV	382	25.22%	29.51%	36.81%	40.96%

NEW

Result

1

Sentence:

ป้า ขาย เข่าหอมมะลิ ทุ่งกุ ลา ร่อง ไห่ ห่อ ละ หนึ่งหมื่น สามพัน บาท ค่ะ

Prediction:

ป้า ขาย เข่าหอมมะลิ ทุ่งกุ ลา ร่อง ไห่ ห่อ ละ หนึ่งหมื่น สามพัน บาท ค่ะ

2

Sentence:

มีสินค่าทั่งหม๊ด หนึ่งพันสามร่อย ซอง ครับ

Prediction:

มีสินค่าทั่งหม๊ด หนึ่งพันสามร่อย ซอง ครับ

Dialect word and intonation are more accurate.

What’s next?

Closer to the official result, but didn’t quite beat them yet

What’s next?

Closer to the official result, but didn’t quite beat them yet

New Technique?

Data Scarcity?

Data Augmentation

- Introducing Speed Perturbation, Speed Perturbation is a data augmentation technique that alters the speed of the speech signal by scaling it up or down.
- Gauthier et al. (2016) demonstrated that speed perturbation is particularly effective for low-resource languages with limited training data. (Gauthier et al., 2016)[5]
- Following the findings of Ko et al. (2015), We will employ speed perturbation (0.9x, 1.0x, 1.1x) to augment the dataset to mimic Vocal Tract Length Perturbation (VTLP). (Ko et al., 2015)[6]

Attempt 3

Apply speed perturbation

CATEGORY	Count	Micro CER	Macro CER	Micro WER	Macro WER
All	1080	13.18%	16.72%	19.72%	24.16%
ECOM	698	10.24%	10.45%	14.04%	14.61%
SURV	382	24.42%	28.16%	37.27%	41.61%

CATEGORY	Count	Micro CER	Macro CER	Micro WER	Macro WER
All	1080	13.20%	17.01%	18.99%	23.36%
ECOM	698	10.05%	10.17%	13.23%	13.72%
SURV	382	25.22%	29.51%	36.81%	40.96%

OLD

CATEGORY	Count	Micro CER	Macro CER	Micro WER	Macro WER
All	1080	13.18%	16.72%	19.72%	24.16%
ECOM	698	10.24%	10.45%	14.04%	14.61%
SURV	382	24.42%	28.16%	37.27%	41.61%

Speed Perturbation

Result

1. Accuracy is about the same.

2. **But,** traning time is significantly faster (from 20.72 hours to 5.26 hours, almost 4x), the model converge faster which was also imply by Ko et al. (2015).

Although this method doesn’t improve accuracy but it reduced training time by 4x. So I decided to keep this implementation and find another method.

Other techniques result in failure

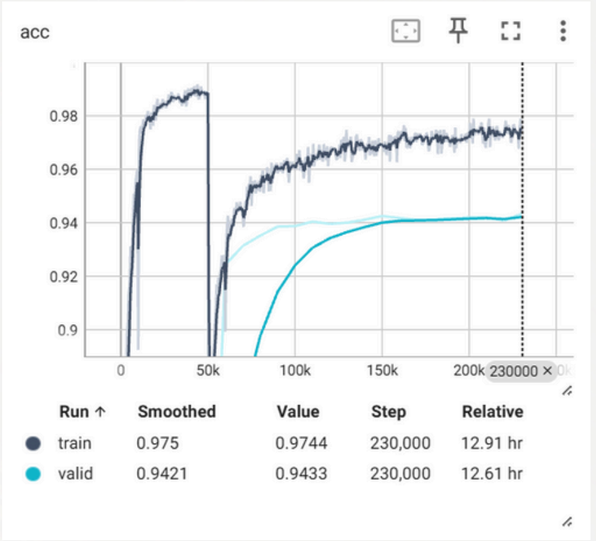
Attempt 4 (Intermediate CTC) (Jaesong Lee and Shinji Watanabe, 2021)[7]

- Intermediate CTC is overkill at this level

CATEGORY	Count	Micro CER	Macro CER	Micro WER	Macro WER
All	1080	13.12%	16.46%	18.83%	23.01%
ECOM	698	10.44%	10.59%	13.65%	14.28%
SURV	382	23.32%	27.19%	34.82%	38.96%

Attempt 5 (Curriculum Learning)

- Split data based on how hard they are (calculated from Frequency and Levenshtein Distance between dialect and central)



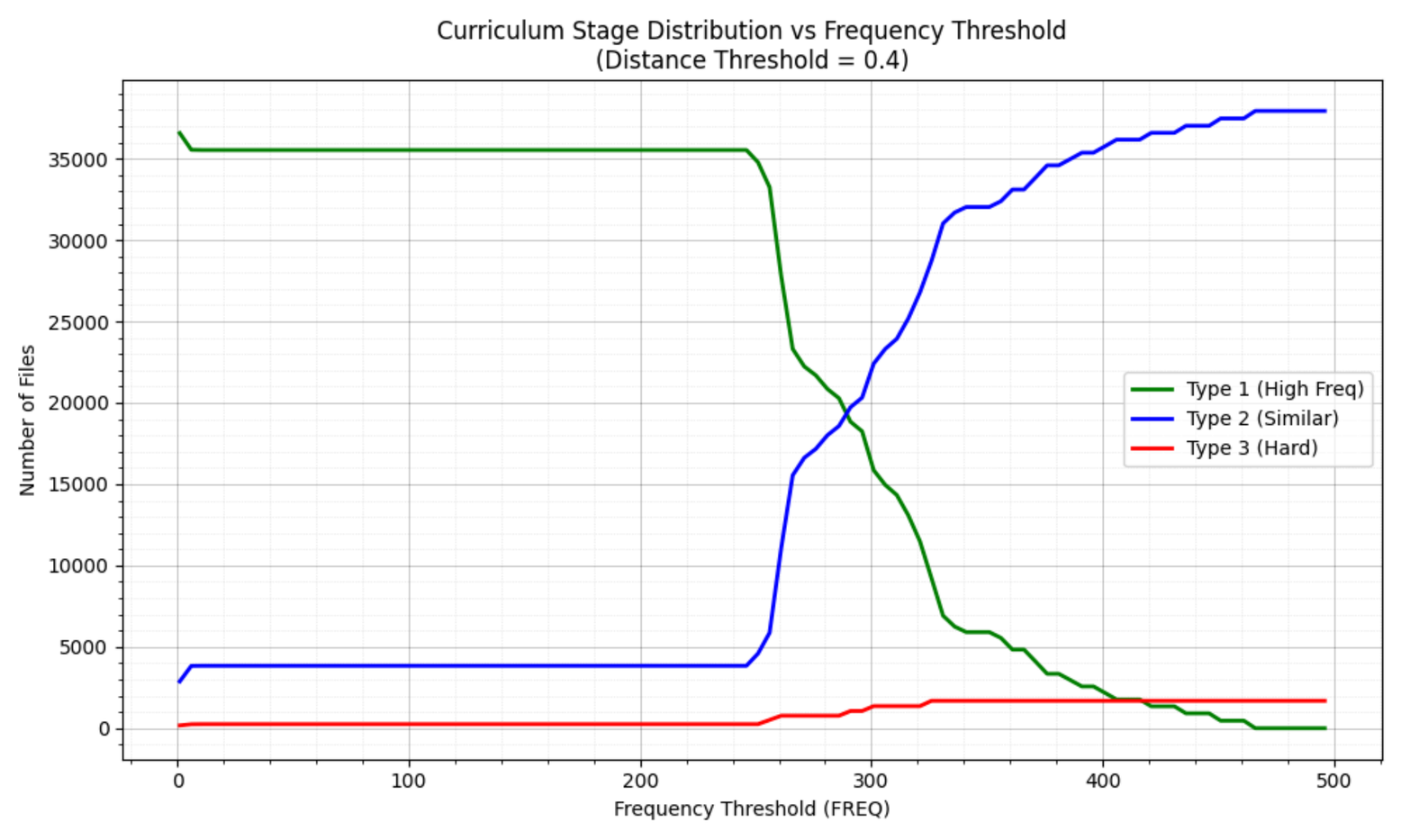
Let’s review all the result

Figured it out what to do next

- The result stuck at the same place.
- All of them has high SURV error rate.
- In attempt 5, while spliting data, I noticed that dataset has a lot of repetetive word especially ECOM. So i try plotting out the dataset.

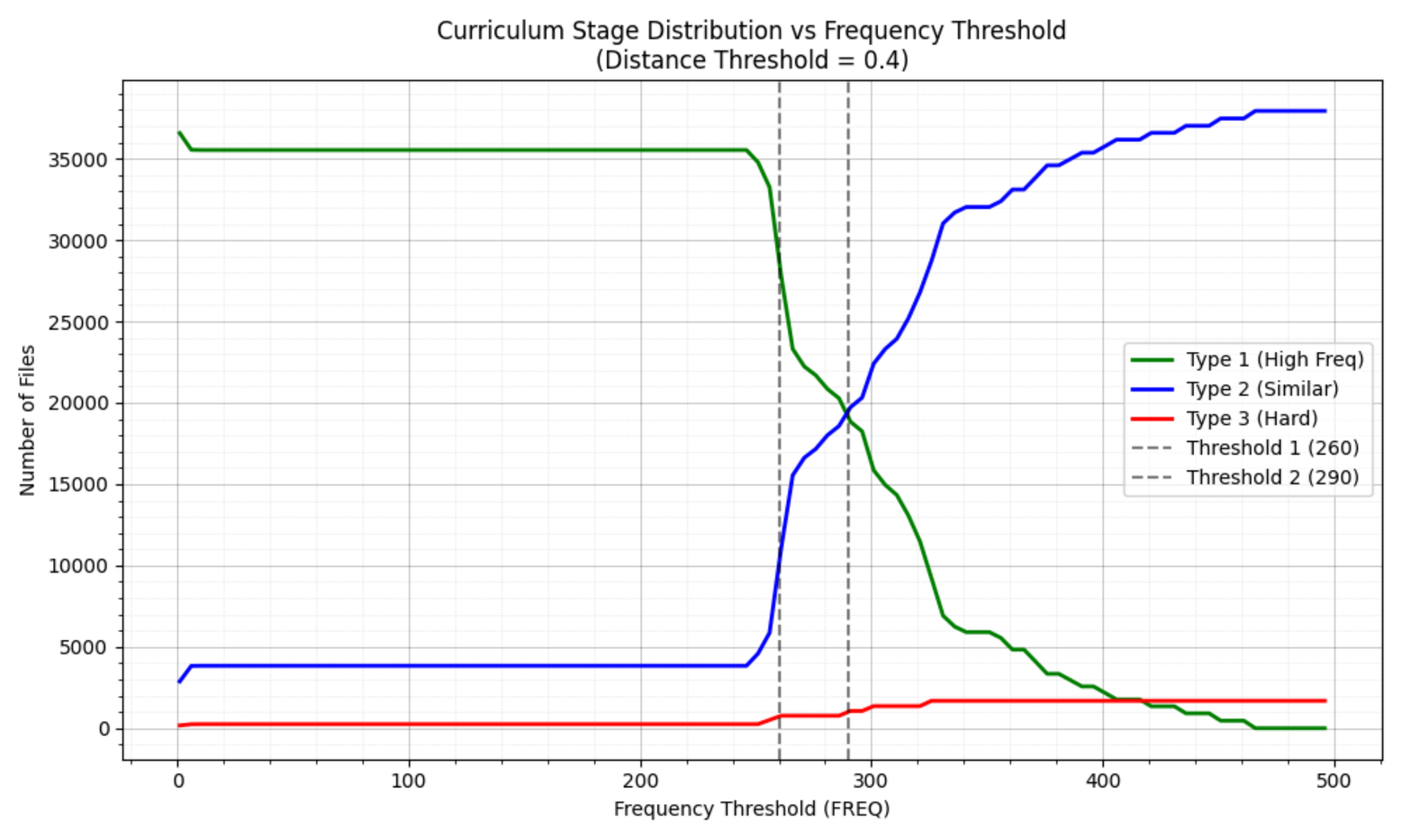
Data Distribution

Data range on their repetitiveness



Data Distribution

From observation, there are two points worth testing



Why?

At FREQ = 260

1. Statistics (Threshold=260):

   Type 1 (High Freq): 29110 files (frequency > 260) -> SURV: 6636, ECOM: 22474
   Type 2 (Similar): 10006 files (distance < 0.4) -> SURV: 6661, ECOM: 3345
   Type 3 (Hard): 508 files (distance >= 0.4) -> SURV: 347, ECOM: 161

2. We analyzed the frequency distribution of sentences in the training corpus. FREQ=260 was identified as the statistical 'elbow' or cutoff point that separated the vast majority of highly repetitive command-based sentences (ECOM) from the more varied natural speech (SURV).

Attempt 6

Resampling[8] at FREQ = 260 (downsampling type 1 and upsampling type 3)

CATEGORY	Count	Micro CER	Macro CER	Micro WER	Macro WER
All	1080	12.88%	15.99%	18.75%	22.68%
ECOM	698	10.44%	10.70%	13.95%	14.62%
SURV	382	22.17%	25.67%	33.58%	37.39%

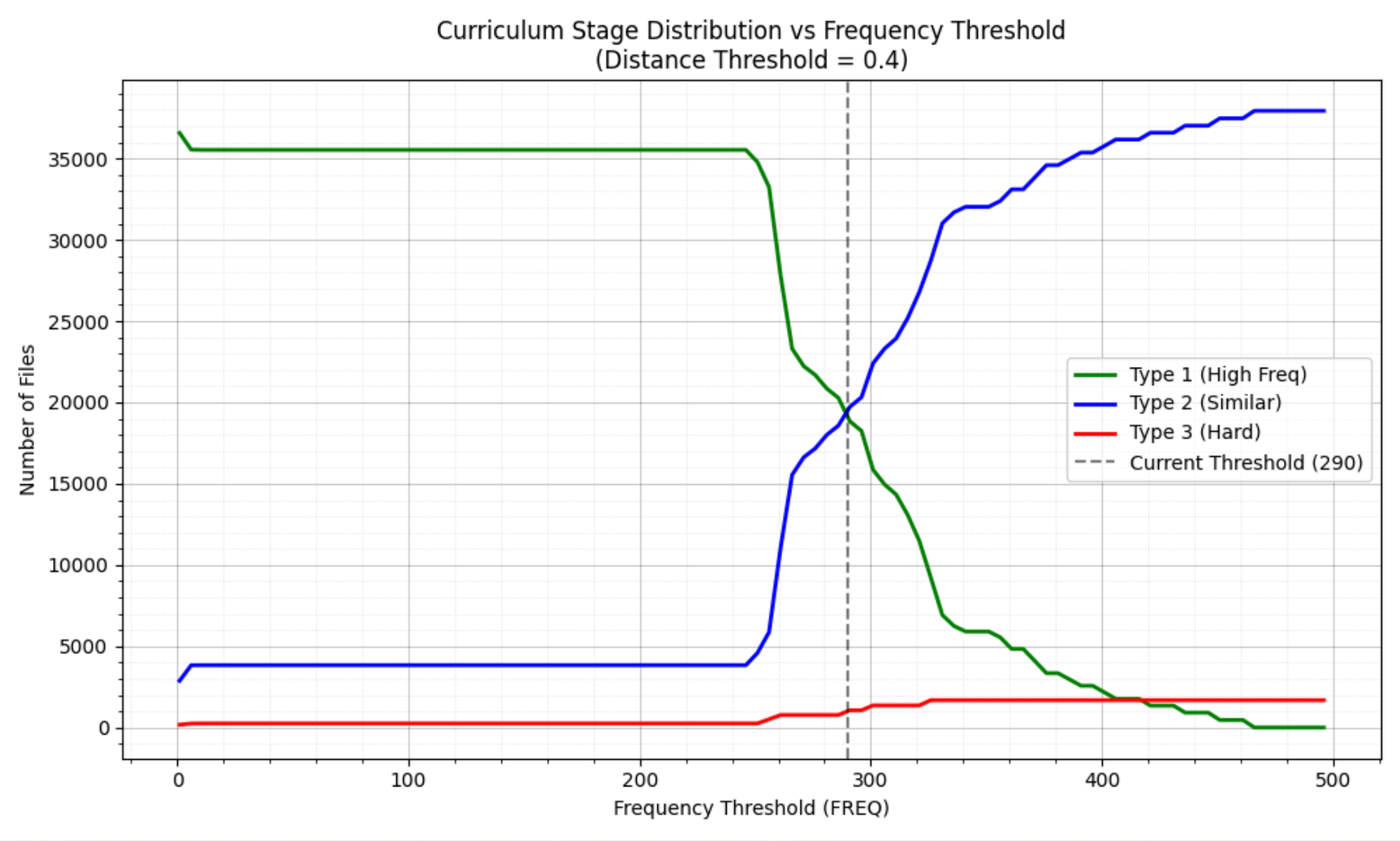
Result

1. Accuracy improved a little bit.

2. Trade ECOM accuracy for SURV accuracy.

Data Distribution

Next is FREQ = 290



Why?

At FREQ = 290

1. Statistics (Threshold=290):

   Type 1 (High Freq): 19125 files (frequency > 290) -> SURV: 0, ECOM: 19125
   Type 2 (Similar): 19730 files (distance < 0.4) -> SURV: 13036, ECOM: 6694
   Type 3 (Hard): 769 files (distance >= 0.4) -> SURV: 608, ECOM: 161

2. This is the first time there's no SURV in Stage 1, making it easier to control the number of both files type. Also, the number of Repetitive files (Stage 1) and Unique files (Stage 2) converged to approximately equal counts (~19,000 each). This natural equilibrium allowed us to maximize the utilization of unique data without needing aggressive downsampling that might discard valuable linguistic variance, theoretically offering the most data-rich training set.

Attempt 7

Resampling at FREQ = 290 (downsampling type 1 and upsampling type 3)

Category	Count	Micro CER	Macro CER	Micro WER	Macro WER
All	1080	12.61%	15.48%	17.90%	21.46%
ECOM	698	10.30%	10.51%	13.37%	13.98%
SURV	382	21.42%	24.56%	31.92%	35.13%

All inference results here:
https://docs.google.com/spreadsheets/d/1VU6FD9N6SKrxXcYocIkukwitY6kjvad4Ey_OkMPn8Z0/edit?usp=sharing

Result

1. Both type accuracy improved from FREQ = 260 a little bit.

2. Doesn’t lose ECOM accuracy.

Conclusion


Right Arm of the Forbidden One
LoRA
[Spellcaster]
A forbidden right arm sealed by magic. Whosoever breaks this seal will know infinite power.
ATK/ 200 DEF/ 300
Exodia the Forbidden One
[Spellcaster / Effect]
If you have "Right Leg of the Forbidden One", "Left Leg of the Forbidden One", "Right Arm of the Forbidden One" and "Left Arm of the Forbidden One" in addition to this card in your hand, you win the Duel.
ATK/1000 DEF/1000
Left Arm of the Forbidden One
Resampling
[Spellcaster]
A forbidden left arm sealed by magic. Whosoever breaks this seal will know infinite power.
ATK/ 200 DEF/ 300
Right Leg of the Forbidden One
Curriculum Learning
[Spellcaster]
A forbidden right leg sealed by magic. Whosoever breaks this seal will know infinite power.
ATK/ 200 DEF/ 300
Left Leg of the Forbidden One
Speed Perturbation
[Spellcaster]
A forbidden left leg sealed by magic. Whosoever breaks this seal will know infinite power.
ATK/ 200 DEF/ 300

Final Result

CATEGORY	Count	Micro CER	Macro CER	Micro WER	Macro WER
All	1080	11.12%	14.30%	16.49%	20.70%
ECOM	698	8.71%	9.06%	11.80%	12.60%
SURV	382	20.35%	23.88%	31.00%	35.49%

About 1% off!

Official

VS

My Attempt

Category	Count	Micro CER	Macro CER	Micro WER	Macro WER
All	1080	12.61%	15.48%	17.90%	21.46%
ECOM	698	10.30%	10.51%	13.37%	13.98%
SURV	382	21.42%	24.56%	31.92%	35.13%

Conclusion

1. Despite not training full parameter, it is possible to archeive a model with a performance on far with full model.

2. Crucially, our model utilizes only 10% of the trainable parameters and was trained on a single consumer GPU (4GB VRAM), representing a massive reduction in computational resources. This demonstrates that Transfer-based Curriculum Learning and many other techniques can effectively bridge the performance gap for low-resource dialects where massive computing is unavailable."

What I have learned

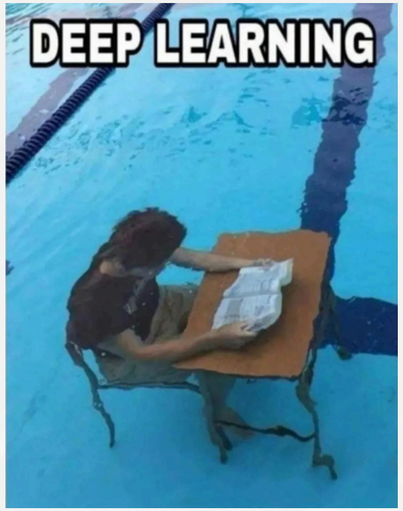
Everything so far

ASR
(Auto speech recognition)

- Speech recognition foundation
- Using modern tools like ESPnet2 to build a speech model including training and fine-tuning

Reserching

- Reserching relevant topic
- Literature Review & Synthesis
- Optimization of Deep Learning Architectures for Limited Hardware.



Conducting Experiment

- Experimental Design & Hypothesis Testing
- Data-Driven Decision Making
- Analyze dataset
- Diagnosed linguistic error pattern

The End

THANK YOU FOR LISTENING

Saruj Chutapornpong

2110391 INDIVDUAL STUDY IN COMPUTER ENGINEERING

References

[1] Someki et al., 2024. ESPnet-EZ: Python-only ESPnet for Easy Fine-tuning and Integration
https://arxiv.org/abs/2409.09506

[2] Hu et al., 2022. LoRA: Low-Rank Adaptation of Large Language Models
https://arxiv.org/abs/2106.09685

[3] Chuangsuwanich et al., 2023. Thai Dialect Corpus and Transfer-based Curriculum Learning Investigation for Dialect Automatic Speech Recognition
https://www.isca-archive.org/interspeech_2023/suwanbandit23_interspeech.html

[4] Dettmers et al., 2023. QLoRA: Efficient Finetuning of Quantized LLMs
https://arxiv.org/abs/2305.14314

References (Cont.)

[5] Gauthier et al., 2016. Speed Perturbation and Vowel Duration Modeling for ASR in Hausa and Wolof Languages
https://www.isca-archive.org/interspeech_2016/gauthier16b_interspeech.html

[6] Ko et al., 2015. Audio augmentation for speech recognition
https://www.isca-archive.org/interspeech_2015/ko15_interspeech.html

[7] Jaesong Lee and Shinji Watanabe, 2021. Intermediate Loss Regularization for CTC-based Speech Recognition
https://arxiv.org/abs/2102.03216

[8] Estabrooks et al., 2004. A Multiple Resampling Method for Learning from Imbalanced Data Sets
https://scispace.com/pdf/a-multiple-resampling-method-for-learning-from-imbalanced-41yrdsn9l0.pdf