University of Applied Sciences and Arts Northwestern Switzerland
School of Computer Science

noxenum

member of
swissuniversities

SwissText - Whisper Finetune

Fine-tuning Whisper Large-v2 for Swiss-German

Vincenzo Timmel
10/11/2025

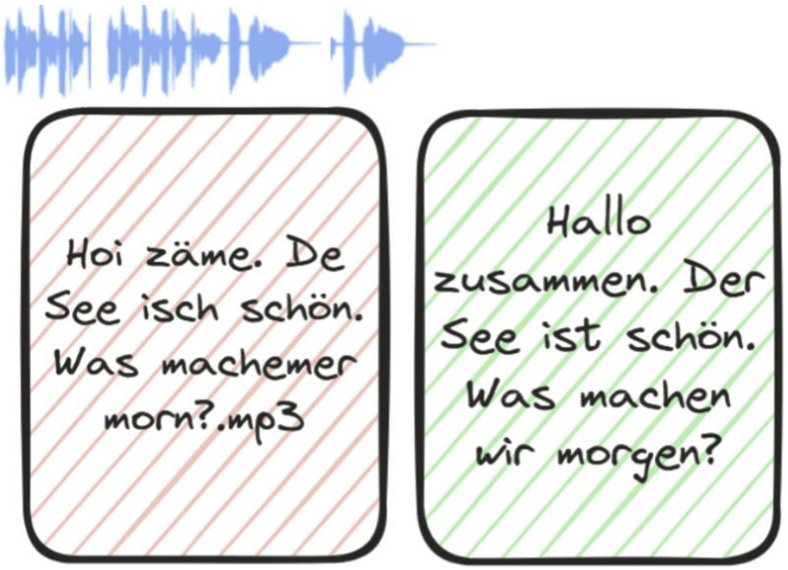
About Me

- BSc Data Science, FHNW
- Working in applied research at the University of Applied Sciences and Arts Northwestern Switzerland, in the NLP-team of Manfred Vogel and Daniel Perrouchoud.
- Focus is Natural Language Processing, with occasional projects in astronomy – but always within the scope of data science.

https://stt4sg.fhnw.ch

Swiss German Speech-to-Text (STT)

- Swiss German Speech to German text is also a **translation**, not only **transcription**.
- Training requires **paired** Swiss German Speech with Standard German text.
- Data is rare.
- Data is copyrighted and/or not shared
- Swiss-German comprises ≈ 5 million speakers - unlocking STT opens accessibility and business applications.



Training & Validation & Test Data

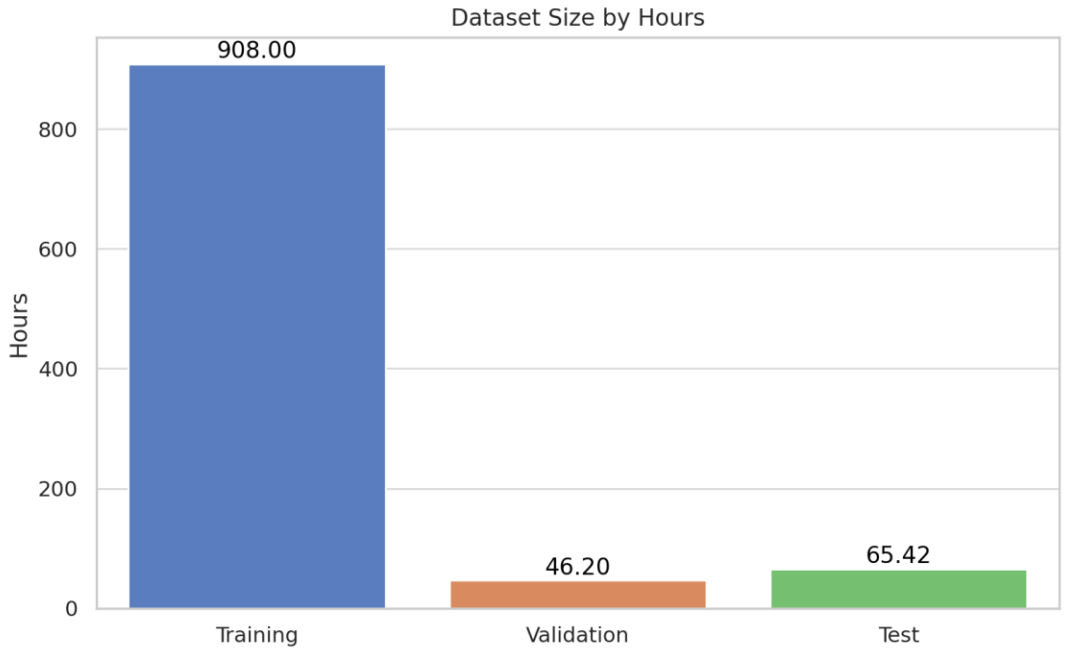


Table 3: Overview of the datasets used for training, validation, and testing, including totals per split.

Name (Variant)	Split	Hours	# Speakers
SDS-200 (Clean)	Train	50	1,799
STT4SG-350 (All)	Train	276	219
SPC (0.9 IOU)	Train	176	194
SRG (PL)	Train	406	–
Total		908	> 2,212
SDS-200 (Clean)	Val	5.2	288
STT4SG-350 (All)	Val	21	219
SRG (PL)	Val	20	–
Total		46.2	> 507
SDS-200 (Clean)	Test	5.2	281
STT4SG-350 (All)	Test	34	56
SPC (0.9 IOU)	Test	6	26
SRG (SWISS TXT)	Test	20	–
Dataset-A	Test	0.22	2
Total		65.42	> 365

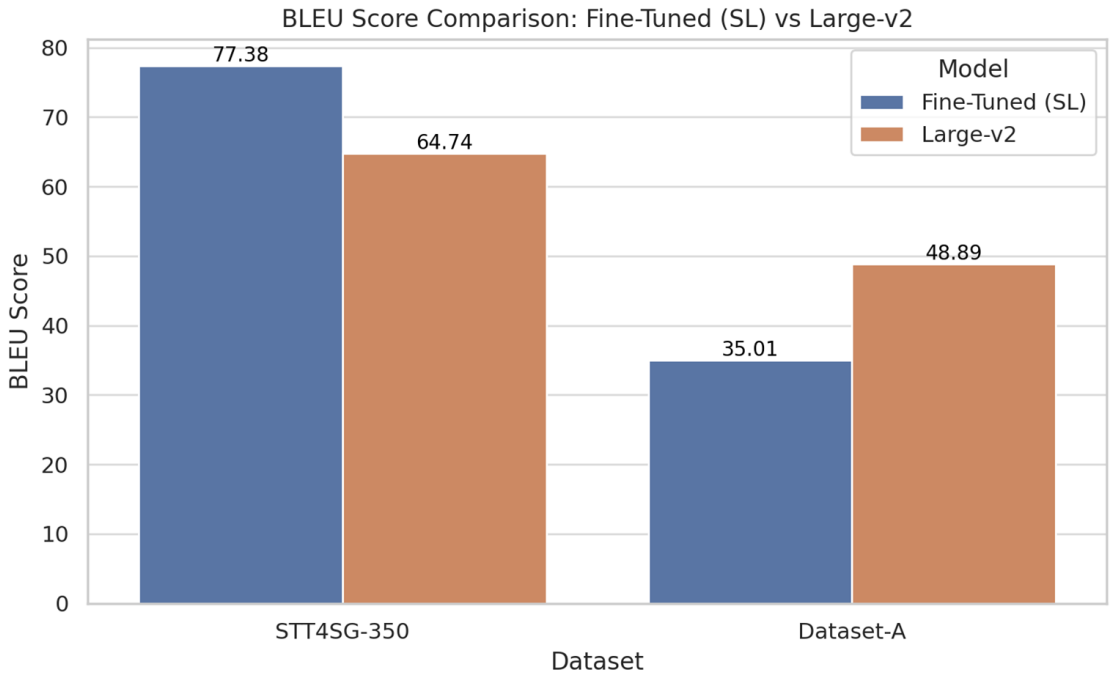
Long-Form Data

- Many datasets are at sentence-level.
- “Real-life” Speech-to-Text is mostly at long-form.
  - **Can we create long-form from sentence-level?**

Motivation for Long-Form

- Fine-tuned on **sentence-level** makes the model **better** on sentence-level data but **worse** on real data.
- It's vital to evaluate the model on data (or as similar as possible) you will pass to the model during inference.
- Whisper is already trained on Swiss-German data.

  - **Key idea:** generate synthetic Long-Form Data from Sentences!



Sentence-Level Model failures

Input Audio
...
Ich zeig der, wo de Bartli de Moscht holt.
...
Whisper Large-v2
...
[00:00:08] Ich zeige dir, wo Bartli den Most holt. [00:00:11]
...
Whisper Large-v2 (fine-tuned on sentences)
...
Ich zeige dir, wo es die Bartli in den Most holt.
...

- Generally, **lower** quality on long-form data than untouched Whisper Large-v2
- On **very** long sentences, it starts to transcribe only “.”
- Segmentation via time-stamp prediction works rarely, especially on very long audios.
- **Partially** fixable with VAD-preprocessing but lower quality remains.

```
1
00:00:04,787 --> 00:00:20,469
Ich begrüsse Sie zurück im Ratssaal zu unserer ersten
Abendsession in dieser neuen
Legislatur.....
```

Long-Form Data Creation

- "**Stitch**" together sentences to create a long-form.
- Idea **easy**, details very **difficult** to get right.
- Our experience, the more **difficult** the audio, the better the fine-tuned Whisper becomes for "real-life" applications.

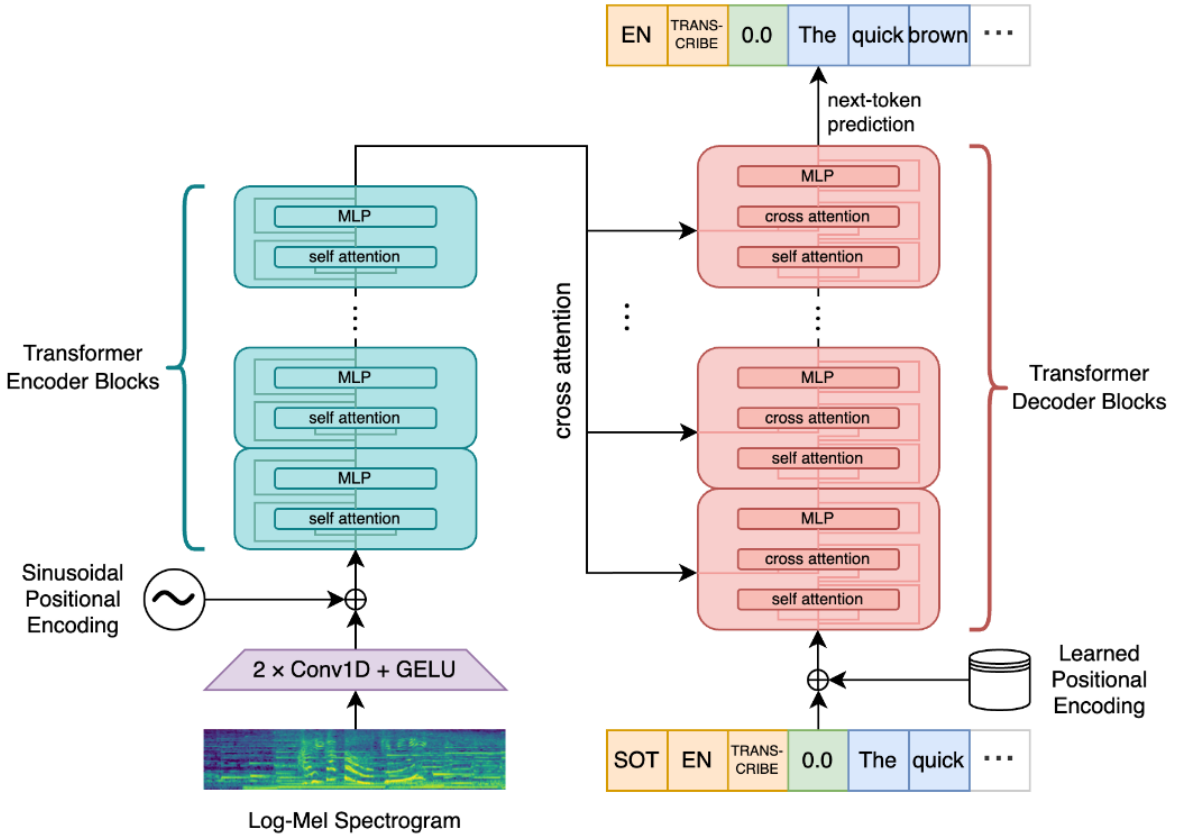
Out-of-Distribution Data

- Out-of-distribution data is **not** unseen data!
- Sentence-level data is very different from meetings, phone calls, patient-doctor discussions, …
- Out-of-distribution data is data created under different circumstances.
- Meetings vs. Meteo Schweiz (Discussion vs. Monolog).
    - **Rarely (never?) do people want to transcribe sentence-by-sentence, like we have in sentence-level datasets.**

Whisper

- **Key idea**: treat speech to text transcription as a classical text-to-text machine translation task but add audio via cross attention!
  - leverage Encoder-decoder Transformer with audio input
  - use weakly supervised audio-text data from the internet (English: 680,000 h, 117,000 h in 96 other languages)

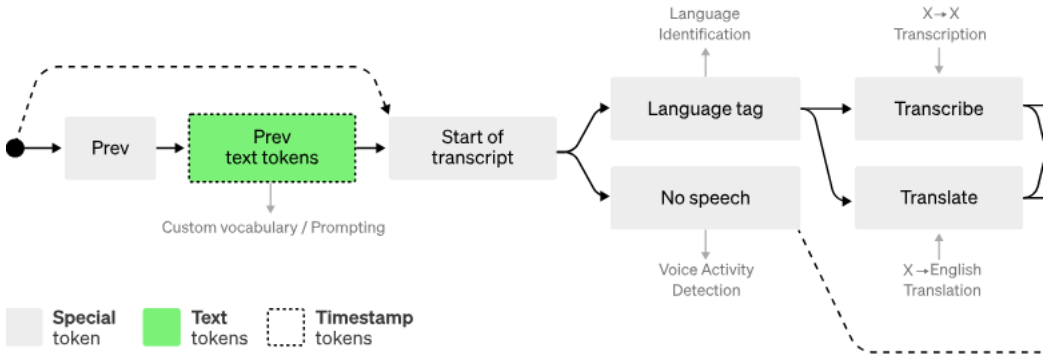
→ Robust out-of-the box, open-source solution for STT.



https://openai.com/index/whisper/

Preprocessed Training Data for Whisper

- **Text**:
  - <|0.32|> Der See ist schön. <|7.26|><|12.94|> Was machen wir morgen?.<|13.24|><|14.26|>
- **Prev:**
  - <|2.50|> Hallo zusammen.<|5.20|><|8.32|>
- **Language**:
  - De
- **Path:**
  - <name>.mp3.



<|startofprev|> Hallo zusammen. <|startoftranscript|><|de|><|transcribe|><|0.32|> Der See ist schön. <|7.26|><|12.94|> Was machen wir morgen? <|13.24|><|14.26|>

https://openai.com/index/whisper/

https://github.com/i4Ds/whisper-prep

Ready to train?

At this point in time, we had:

- **1x A100** with 40GB of VRAM 👍
- A python script to fine-tune whisper the way OpenAI did 👍
- Logging of loss, metrics and checkpoints on Weights & Biases 👍
- Storing of data on Hugging Face. 🤗 👍
- Batch size of 1 👎
  - **Ready!? No, we only have a batch-size of 1.. A bigger batch size leads to VRAM OOM errors -> Results were not good.**

Calculations were performed at sciCORE (http://scicore.unibas.ch/) scientific computing core facility at University of Basel.

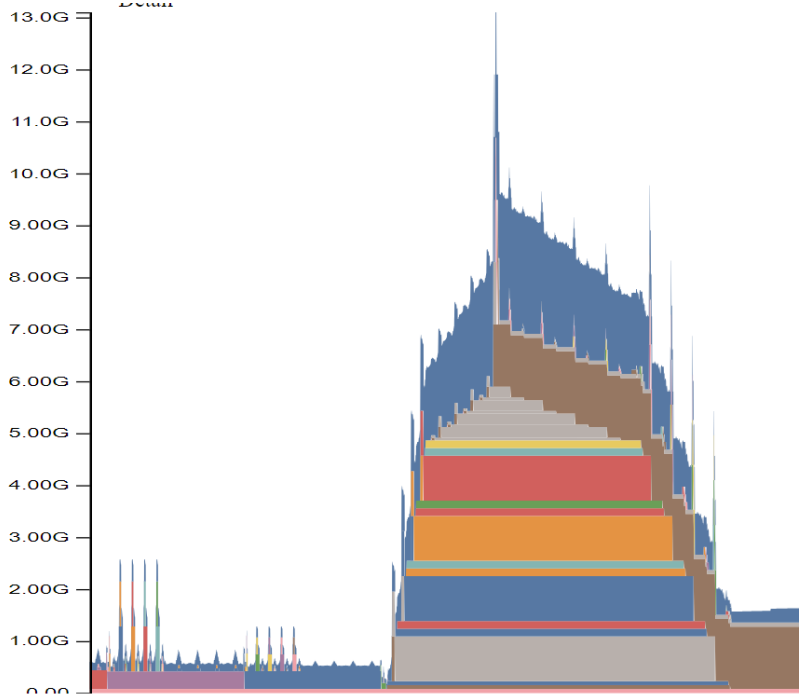
Reduction of VRAM Usage

Activation checkpointing

Usually, PyTorch's autograd stores every intermediate tensor for a faster backward pass. By using checkpointing, you can define which output is stored, requiring recalculation; trade computation for memory.

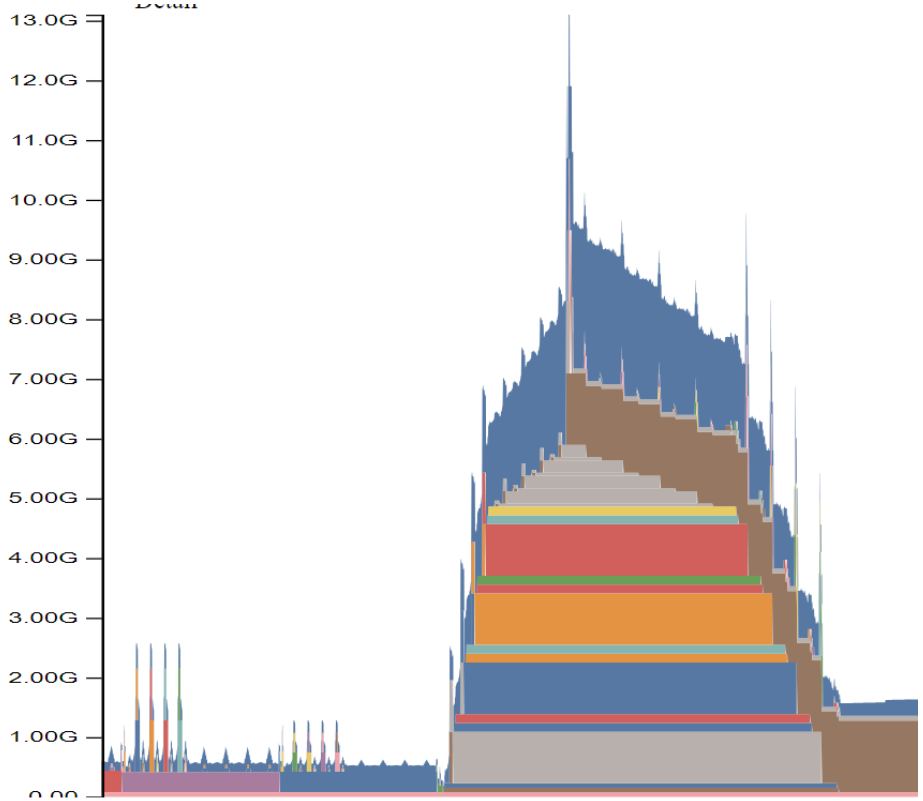
8Bit Optimizer

AdamW stores two states for each parameter in a model. Going from float32 to float8 is a saving of around **9GB** with Whisper Large-v2

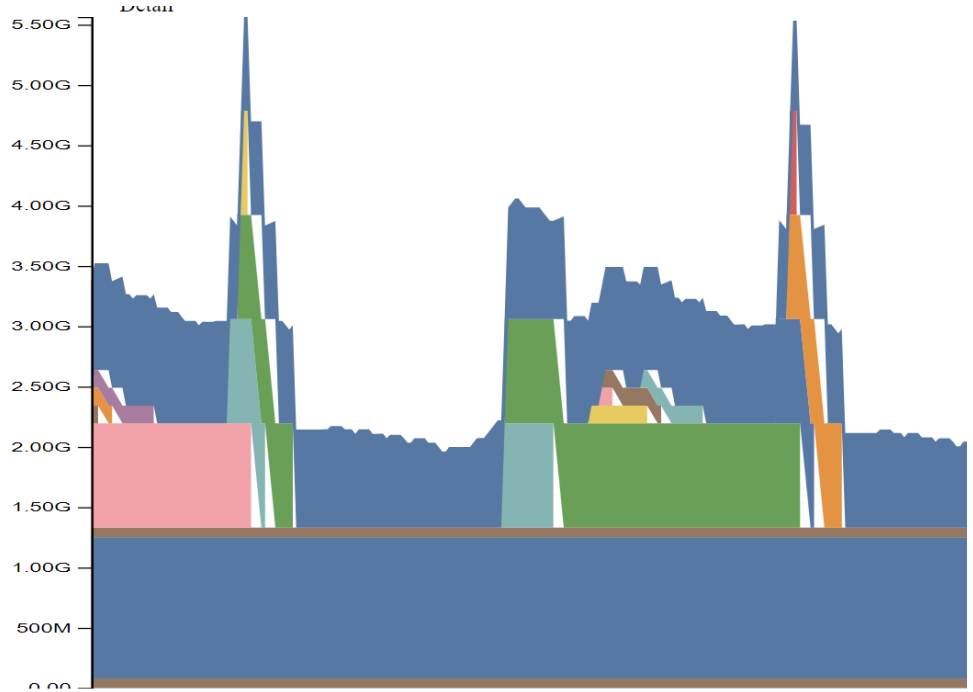


VRAM usage over time when fine-tuning Whisper.

Reduction of VRAM Usage



VRAM usage before the changes



VRAM usage after the changes

Ready to train?

At this point in time, we had:

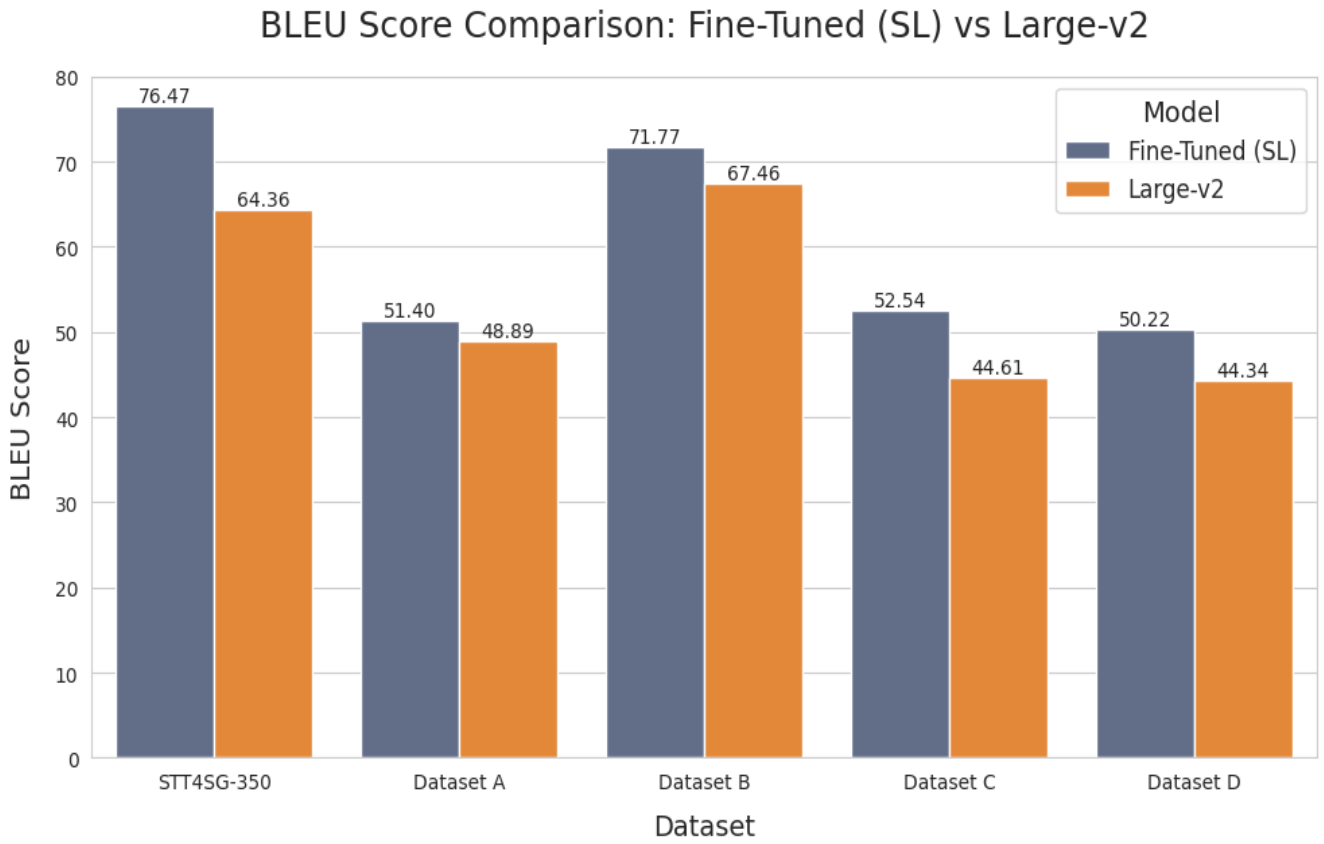
- **1x A100** with 40GB of VRAM 👍
- A python script to fine-tune Whisper the way OpenAI did (probably) 👍
- Logging of loss, metrics and checkpoints on Weights & Biases 👍
- Storing of data on Hugging Face. 🤗 👍
- Batch size of 16 👍
- Artificial Batch Size of 256 with gradient accumulation of 16 👍
- Data augmentation methods (SpecAugment, TimeWarp and Stochastic Depth) 👍

Calculations were performed at sciCORE (http://scicore.unibas.ch/) scientific computing core facility at University of Basel.

Results

- Our fine-tuned Whisper **surpasses** untouched Whisper Large-v2 on **all** Swiss-German datasets we evaluated it on.
- Our model **shines** on a Swiss named-entity; it's not transcribing "Brugg" as "Brücke" anymore, but as "Brugg".

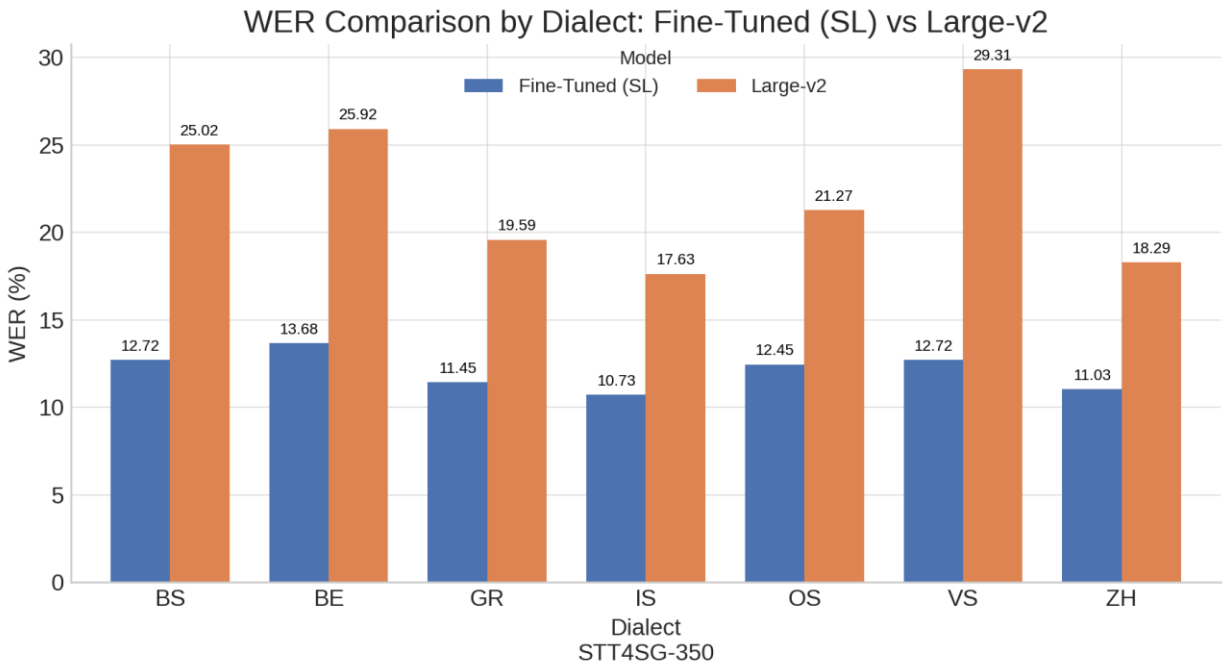
github.com/i4Ds/whisper-finetune



Datasets A-D are closed-source but generally around an hour of a Swiss-German discussion, such as a meeting.

Results per Dialect

- When evaluating **per** dialect, our fine-tuned model shows much **less** variance between dialects, while untouched Whisper Large-v2 has troubles with **Valais, Bern and Basel.**



github.com/i4Ds/whisper-finetune

Training Run Loss & Metric Curve

- Generally, the training loss drops strongly when an epoch is finished.
- The model with the **lowest** validation loss is not necessarily the **best** model.
- Validation loss can **increase**, while the word error rate goes **down**.

Which model should we keep?

Questions?