Fine-Tuning eines Whisper-Modells für die Sprache Obersorbisch

Zielsetzung des Projekts

Das Ziel dieses Projekts bestand darin, ein Whisper-Modell für die Sprache Obersorbisch (HSB) zu fine-tunen, um effektiv somit die Wortfehlerrate (WER) zu optimieren, um damit eine praxistaugliche und idealerweise auch weltweit führende Lösung zu finden. Das Modell wird anschließend der Öffentlichkeit zur Verfügung gestellt. Die einzelnen Schritte sowie relevante, zu berücksichtigende Aspekte, um eine ideale Performance zu erzielen, sind in diesem Dokument festgehalten.

Über Whisper

Whisper ist eine State-of-the-Art Lösung , um automatisiert Transkriptionen (Speech-To-Text) anhand von Audio-Dateien zu erstellen.

Whisper-Modelle und Ressourcenanforderungen

Whisper ist in verschiedenen Modellgrößen verfügbar. Diese Varianten wurden je nach Einsatzzweck und Ausgangssituation konzipiert. Hier gilt es die ideale Balance zwischen Genauigkeit und Rechenaufwand zu finden. Dabei gilt grundsätzlich:

- **Tiny**: Kleinste Whisper Modellvariante, ideal für ressourcenschwache Umgebungen.
- **Base**: Höhere Genauigkeit als *Tiny*, jedoch weiterhin ressourcenschonend.
- **Small**: Genauigkeit, die sich für den Produktionsbetrieb bereits eignet. Guter Kompromiss zwischen Modellgröße und Leistung.
- **Medium**: Hohe Genauigkeit für anspruchsvollere Anwendungen.
- **Large**: Leistungsstärkstes Modell mit der höchsten Präzision.

Size	Layers	Width	Heads	Parameters	English-only	Multilingual
tiny	4	384	6	39 M	✓	✓
base	6	512	8	74 M	✓	✓
small	12	768	12	244 M	✓	✓
medium	24	1024	16	769 M	✓	✓
large	32	1280	20	1550 M	x	✓
large-v2	32	1280	20	1550 M	x	✓
large-v3	32	1280	20	1550 M	x	✓

Für diesen Projekteinsatz scheiden *tiny* und *base* infolge der marginalen Architekturtiefe aus. Eine Generalisierung anhand der vorliegenden Trainingsdaten ist unwahrscheinlich.

Für den ersten technischen Durchstich und zur Sicherstellung des korrekten technischen Ablauf wurde deshalb das *small*-Modell verwendet, da sich dies infolge der noch immer geringen Modellkomplexität (Anzahl Parameter) für zügige Iterationen und erste Experimente eignet.

Hinweis: Sollte das Modell für gewisse Projekteinsätze gewzungenermaßen auf ressourcenschwachen Endgeräten ausgeführt werden, sollte dennoch die Benutzung des base-Models in Betracht gezogen werden, sofern die Hardware-Anforderungen für größere Models nicht gegeben sind. Tiny- und Base-Modell eignen sich gut für den Einsatz auf CPUs.

Vorbereitung der Datensätze

Die Audio-Rohdaten werden auf eine Abtastrate von 16 kHz umgewandelt, um mit den vortrainierten Whisper-Modellen kompatibel zu sein.

Die Transkriptionen wurden stichprobenartig geprüft. Grundsätzlich gilt, dass eine Vielfalt in der Audioqualität dazu beitragen kann, dass das Modell die Sprache abstrahiert und tatsächlich lernt (und nicht nur eine bestimmte Sprechweise, Stimmfarbe etc.). Bei Transkriptionen verhält es sich gegensätzlich. Hier ist eine Vielfalt in der Qualität der Transkriptionen nicht vorteilhaft. Das Transkriptionsformat sollte deshalb einheitlich sein. Es muss sichergestellt werden, dass alle Transkriptionen in Kleinbuchstaben umgewandelt und sämtliche Satzzeichen entfernt bzw. einheitlich verwendet werden. Für zukünftige Optimierungen wäre es daher ratsam, von einem Native-Speaker die Transkriptionen zu prüfen und auf eine einheitliche Repräsentation zu achten.

Fine-Tuning

Fine-Tuning bedeutet - ganz allgemein gesprochen, ein bereits vortrainiertes Modell mit spezifischen Trainingsdaten zu trainieren, um dessen Leistung für eine bestimmte Aufgabe zu verbessern. Dazu werden die Modellparameter auf eine kleinere, domänenspezifische Datenmenge optimiert.

In diesem Fall erfolgte das Fine-Tuning auf Basis des bereitgestellten Obersorbischen-Datensatz.

Evaluation: Wortfehlerrate (WER)

Zur Bewertung der Modelle wurde die für gewöhnlich genutzte Word Error Rate (WER) als Metrik herangezogen.

Hyperparameter-Tuning

Lernrate

Die Lernrate ist die wesentlichste Stellschraube im Fine-Tuning-Prozess.

Wir orientieren uns hier an den Aussagen von J. W. Kim, der als Mitautor des Whisper-Papers grundsätzlich folgenden Vorschlag genannt hat: Beim fIne-Tuning sollte die Lernrate 40-mal kleiner sein als die Lernrate beim Pretraining.

Model Size	Max Learning Rate (paper)	Suggested fine-tuning Learning Rate (40x smaller)
tiny	1.5×10^{-3}	3.75×10^{-5}
base	1×10^{-3}	2.5×10^{-5}
small	5×10^{-4}	1.25×10^{-5}
medium	2.5×10^{-4}	6.25×10^{-6}
large	1.75×10^{-4}	4.375×10^{-6}
large-v2	2.0×10^{-4}	5×10^{-6}

Wir haben hierbei verschiedene Lernraten ausprobiert. Sowohl für das Medium, als auch das Large-Modell. Letztendlich können wir die vorgeschlagenen Lern-Raten als grundlegende Empfehlung bestätigen.

Warmup-Steps

Die warmup-Steps wurden methodisch vertestet. Hier zeigt sich, dass ein Wert, der etwa 10% der insgesamt zu durchführenden Steps entspricht, eine gute Performance erzielt.

Batch-Size

Angaben zur Batch-Size (Training als auch Evaluation) variieren lediglich von den zu Verfügung stehenden VRAM-Kapazitäten. Das Large-Model wurde hier beispielsweise mit einer Batch-Size von 8 trainiert.

FP16 VS BF16

BFloat16 bietet während des Trainings eine bessere Stabilität als FP16. Moderne Sprachmodelle - insbesondere gerade LLMs - werden in BFloat16 trainiert, da es einen Vorteil bzgl. der Stabilität bietet. Auf älteren GPUs scheidet BF16 dagegen aus. Aus diesem Grund ist grundsätzlich BF16 vorzuziehen, sofern es keine hardwarbedingte Einschränkung gibt.

Verwendete Hardware

Das Training wurde auf einer NVIDIA A40 GPU durchgeführt. Die Batch-Größe variierte je nach Modellgröße um die gegebenen Kapazitäten bestmöglich zu nutzen.

Exemplarische Durchführungen

Medium

Bei Experimenten rund um das Medium-Modell hat sich - wie angenommen - die Lernrate als die wesentlichste Stellschrauben herauskristallisiert. Die Lernrate 0.00000625 erbrachte hierbei die beste Performance beim Training.

Epoch	Training Loss	Validation Loss	Wer
1	0.425900	0.110605	11.780095
2	0.051300	0.087098	7.504583
3	0.017000	0.085249	6.488695

Wenn die Lernrate zu hoch eingestellt ist, kann es zum Overshoot kommen, bei dem es um den optimalen Punkt oszilliert oder sogar divergiert.

Andererseits kann eine zu niedrige Lernrate dazu führen, dass das Netzwerk sehr lange für die Konvergenz benötigt oder in lokalen Minima stecken bleibt, wodurch es das globale Minimum nicht erreichen kann. Ohne Experimente lässt sich der ideale Wert jedoch nicht im Vorfeld berechnen.

Large

Beim Large-Modell wurden verschiedene Experimente mit unterschiedlichen Konfigurationen durchgeführt. Auch hier hat sich die gleiche Lernrate wie die des Medium-Models als solide Konfiguration bewahrheitet.

Epoch	Training Loss	Validation Loss	Wer
1	0.203700	0.098389	9.011882
2	0.037800	0.083435	6.903162
3	0.012500	0.079834	5.990206
4	0.004300	0.077515	5.453062

Mit dem Large-Model ist damit - wie erwartbar - ein etwas besseres Ergebnis als mit dem Medium-Modell realisiert worden.

Um ein ressourcenschonendes Training des Models durchzuführen, ist der Einsatz von Deepspeed denkbar. DeepSpeed ist eine Optimierungslibrary, die sowohl verteiltes Training als auch Inferenz effizienter ermöglicht. Damit lässt sich ein Large-Model, welches je nach Konfiguration mehr als 40GB VRAM erfordert, auch auf leistungsschwächerer Hardware trainieren.

GGML

GGML ist eine C-Bibliothek, die eine effiziente Inferenz ermöglicht. Mit dessen Einsatz lassen sich größere Sprachmodelle ressourcenschonend einsetzen. GGML wird beispielsweise für die C/C++ Implementierung von Whisper eingesetzt (inoffizielle Portierung von OpenAIs Whisper).

Bei der Konvertierung des safetensor-Format zur GGML-Format erfolgt hierbei eine Quantisierung wodurch Leistungseinbußen in der Performancen, zugunsten eines niedrigen Ressourcenverbrauchs gelten.

Im Umgang mit huggingface empfiehlt es sich, das Modell im (Standard) safetensor-Format zu veröffentlichen und erst anwendungsspezifisch - sofern erforderlich - zu GGML zu konvertieren.

Ergebnisse und Fazit

Nach dem Fine-Tuning zeigte sich, dass das Large-Modell die besten Ergebnisse erzielte, jedoch auch höhere Rechenressourcen erforderte.

Das Projekt hat gezeigt, dass man mit einer relativ geringen Menge an domänenspezifischen Daten gute Leistungen erzielen kann. Dies unterstreicht das Potenzial von Fine-Tuning für ressourcenschwache Sprachen wie Obersorbisch.

Zudem muss man festhalten, dass man allgemein dem Menschen eine WER von circa 4% zuschreibt. Wenngleich der Drang zur Perfektion der maschinellen Transkription nachvollziehbar ist, so muss man an dieser Stelle auch darauf hinweisen, wie nah die bereits veröffentlichten Models den menschlichen Fähigkeiten nahekommen.