Abschlussbericht

29.03.2025

Generelles

Meine Aufgabe war es neue Trainingsdaten für Paule aus dem Common Voice Datensatz (cv-corpus-18.0-2024-06-14) zu holen.

Spezifischer : mnt/Restricted/Corpora/CommonVoiceVTL/corpus_as_df_mp_folder_{language} | en oder de

Ich bin mit dem Trainieren der Modelle nicht fertig geworden aber es existiert ein Repo für die Synthese mit einigem an Dokumentation , man sollte es also ohne Probleme für neue Daten im ähnlichen Stil benutzen können

> https://github.com/quantling/create_vtl_corpus

Daten

Für sowohl Deutsch als auch Englisch wurden Daten generiert und für die Deutschen daten mit ca. 80/10/10 gesplittet. Dabei sollte jedes Wort auf jeden Fall im Testset vorkommen, was bedeutet, das dies unter Umständen schwerer ist als Validation und Training. Bei den englischen Daten könnte Speicherplatz ein Problem werden, da diese noch nicht gesplittet ist

Ferner wurden sie auch bereinigt, aber es können weitere Fehler vorkommen

Der Code für das Training und das Splitten befindet sich hier

> https://github.com/JoJoBarthold2/training

Training

Ich habe ferner versucht ein training für den Embedder, das Inverse und das Forward Modell zu bauen.
Diese scheinen einigermaßen zu funktionieren,zumindest geht der Loss für den deutschen Minicorpus runter.

Trainingszeit für die ganzen Trainingsdaten dauert sehr lange ca. 13 min pro Datensatz bei Forward Modell, schneller beim Embedder.

Ich habe wenn ich mir unsicher war einfach angaben des Skripts von Paul übernommen, dass auch in dem obigen training Repo ist.

TODOs

Man sollte also das Training laufen lassen, schauen ob es funktioniert und dann die Skripte anpassen und verbessern.

Das geschieht gerade für alle Modelle. Wahrscheinlich dauert dies für 200 Epochen ein halbes Jahr, also wirklich lange. Jede Epoch wird gespeichert. Es gibt die Möglichkeit Optimizer und Model wieder zu laden.

Der letzte Schritt wäre dann das Trainieren der GANs und im Endeffekt das Zusammenfügen der Komponenten um das Gesamtmodell für Paule zu validieren.

Kontakt

Obwohl ich zu diesem Zeitpunkt (wahrscheinlich) nicht mehr für Quantling arbeite kann man mich bei Fragen gerne über Github oder meine Website kontaktieren.

gez. Valentin Schmidt

http://valentinschmidt.eu