Richtlinie zum „Software-Sprint“

DLR PT- Berlin
Eing.am:
01. Okt. 2019
Eingangsnr.: 5952 / BF

YV – Your Voice

Schlussbericht

Zuwendungsempfänger:

Thomas Werkmeister

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS18S55 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Kurze Darstellung der Aufgabenstellung und Motivation

Meine Motivation mit dem Projekt war auf neue Entwicklungen in der Sprachsynthese mittels neuronaler Netze hinzuweisen, die es ermöglichen natürliche und emotionale Sprache zu erzeugen und dabei auch die Stimmen von Personen nachempfinden können. Diese Technologie hat viele sinnvolle Anwendungen, aber bietet auch viel Raum für Missbrauch.

Die wichtigstens Meilensteine waren:

- Reproduktion der Ergebnisse aus den wissenschaftlichen Papieren von Google und der Tacotron Architektur für die Deutsche Sprache
- Verfeinerung der Ergebnisse bezüglich Natürlichkeit, Ausdrucksvermögen (Emotionen), und Anzahl der Sprecher
- Veröffentlichung einer Website, die die besagte Technik erläutert und ausprobierbar macht

Beitrag des Projektes zu den Zielen der Förderinitative „Software-Sprint“

Die Zielgruppe meines Projektes ist die an Technologie interessierte Allgemeinheit, Berufsgruppen die mit Missbrauch der Technologie in Kontakt kommen könnten, sowie Entscheider. Das Projekt YourVoice sollte sie auf die möglichen Konsequenzen der Technologie hinweisen und gleichzeitig technisches Wissen vermitteln, sodass sinnvolle Entscheidungen getroffen werden können.

Ausführliche Darstellung der Ergebnisse

Leider hat sich mein Projekt als weitaus forschungslastiger, komplexer und schwieriger herausgestellt als anfänglich gedacht. Ich habe es auch bis heute, Ende September, nicht geschafft den ersten Meilenstein zu erreichen und die Ergebnisse für Deutsch zu reproduzieren. Das in dem Kontext, dass ich weit über die veranschlagten Stunden gearbeitet habe und außerdem noch Rückenwind durch eine Kooperation mit Mozilla und der Telekom bekommen habe, die im Austausch von Fachwissen bestand. Das Projektziel ist nicht unmöglich, und ich habe auch große Fortschritte gemacht und diese auch zur offenen Implementierung von Mozilla beigetragen. Dennoch ist das Ziel sehr hoch gesteckt gewesen, da nicht alle Faktoren vollständig kontrollierbar waren. Vor allem hat sich die Datenqualität der bisher frei verfügbaren Datensätze für die Deutsche Sprache als ungenügend herausgestellt.

Zielgruppe, Nutzen und mögliche Weiterentwicklungen

Leider gibt es bis heute keine Ergebnisse im Sinne der ursprünglichen Zielstellung. Dennoch wurden Schritte in die richtige Richtung übernommen. So habe ich viele Fehler in der öffentlichen Implementierung der neuronalen Sprachsynthese von Mozilla behoben, diese für die Verwendung von Sprachen außer Englisch verbessert, die Verwendung von mehreren Sprechern für das Training implementiert, die Lerngeschwindigkeit bis zur Konvergenz stark gesteigert, und neue Lösungsansätze für besonders fragile Teile des neuronalen Netzwerks entwickelt. Darüber hinaus habe ich hunderte Experimente durchgeführt, um die mangelnde Datenqualität durch bessere Hyperparameter und Architektur des Netzwerks auszugleichen. Diese haben aber doch noch nicht den ersehnten Erfolg gebracht.

Auch zu diesem Zeitpunkt, Ende September, bin ich noch mit dem Projekt beschäftigt und unter anderem daran weitere Datensätze für die deutsche Sprache öffentlich verfügbar zu machen oder möglicherweise, finanziert durch Dritte, erstellen zu lassen.

Kurze Darstellung der Arbeiten, die zu keiner Lösung geführt haben

Es hat sich herausgestellt, dass der Weg zur Lösung länger und beschwerlicher ist, als das ein sechsmonatiges Projekt zulässt. Was besonders viel Zeit gekostet hat, war die Verwendung von ungeeigneten Datensätzen, wie z.B. des Common Voice Datensatzes von Mozilla, der zwar für Spracherkennung gut geeignet ist, aber aufgrund der geringen Audioqualität und der geringen Menge an Material, wenig für die Sprachsynthese zu bieten hat. Aber auch andere offene Datensätze wie z.B. gelesene Audiobücher von librivox, die mehr Material von einem einzelnen Sprecher haben, haben noch nicht das gewünschte Ergebnis gebracht. Hier wirken sich Segmentierungsfehler, Atemgeräusche, und Versprecher negativ auf die Leistung des Systems aus. Für einen dieser Audiobuch Datensätze habe ich in knapp 4000 von 17000 Audiodateien Segmentierungsfehler mit einem halbautomatischen heuritischen Verfahren korrigiert, leider ohne direkt messbaren Einfluss auf die Resultate.

Natürlich zählen hierzu auch etliche Experimente, die bisher noch zu keiner Lösung geführt haben, mich aber mehr und mehr mit der Architektur vertraut gemacht haben. Die Architektur des neuronalen Netzes ist sehr komplex und besteht aus 5 Subnetzen, die sich je auf verschiedenste Weisen gestalten lassen. Insgesamt besteht das Netz aus mehr als 60 Schichten. Dazu kommen nochmal 40 bis 50 Hyperparameter für die Vorverarbeitung der Sprachsequenzen, Texte, des Lernverhaltens sowie der konkreten Ausprägung der verschiedenen Schichten. Darüberhinaus ist

durch die sequentielle Natur der Aufgabe, d.h. die schrittweise und nicht parallelisierte Erstellung der gesprochenen Sprache, ein einzelnes Training wenigstens 24 Stunden aber eher mehrere Tage dauert. Die Vorbereitung, Auswertung, Analyse von Fehlern, und Korrektur, Konzipierung neuer Experimente etc. hat äußerst viel Zeit beansprucht. Dennoch haben die Ergebnisse die gesteckten Ziele bisher nicht erreichen können.

Kurze Angabe von Präsentationsmöglichkeiten für mögliche Nutzer

Zu diesem Zeitpunkt gibt es hierfür das Repository von Mozilla und meinen persönlichen Fork auf dem ich sehr viele Experimente basiert habe.

https://github.com/mozilla/TTS
https://github.com/twerkmeister/TTS

Kurze Erläuterung zur Einhaltung der Arbeits- und Kostenplanung

Ich bin mit meiner Arbeitszeit deutlich über die veranschlagten Stunden gekommen. Mit fast 200 Stunden Mehraufwand als ursprünglich geplant. Diesen Mehraufwand war ich bereit zu leisten, da ich weiterhin vom Wert des Projektes überzeugt bin.

Kurze Darstellung von etwaigen Ergebnissen bei anderen Stellen

Hier ließe sich die Zusammenarbeit mit Mozilla und der Telekom anführen. Die vielen fachlichen Diskussionen haben unterstützend auf das Projekt gewirkt. Die Ziele meines Projektes standen dabei für mich an erster Stelle. Das habe ich offen kommuniziert und glücklicherweise haben sie sich auch mit denen der besagten Unternehmen gedeckt. Diese waren zwar nicht an der eigentlichen Aufbereitung und Erklärung der Ergebnisse interessiert, aber an der Verwendung an den technischen Zwischenergebnissen die bis dorthin erbracht werden müssen. Dazu zählen v.A. die bisher erarbeiteten technischen Verbesserungen der offenen Sprachsynthese Software. Durch die offene Entwicklung stehen diese Verbesserungen natürlich auch vielen anderen Entwicklern offen. Das Repository von Mozilla hat über eintausend Interessenten.