Exposé - YALR

Richard Bihlmeier
richard.bihlmeier@hs-duesseldorf.de
Hochschule Düsseldorf

Jannis Bollien
jannis.bollien@study.hs-duesseldorf.de
Hochschule Düsseldorf

12. Oktober 2025

Projektbeschreibung

Im Rahmen des Projekts soll ein Programm entwickelt werden, das stumme Videoaufnahmen von Personen in Text überführt. Zu diesem Zweck wird ein bereits bestehendes neuronales Netz **AV-HuBERT (Audio-Visual Hidden Unit BERT)** verwendet. Das Modell ist darauf spezialisiert, audiovisuelle Informationen zu verarbeiten und kann auch ohne Tonspur visuelle Bewegungen und Mundformen analysieren. Das Ziel des Projekts besteht darin, das Netz so zu implementieren, dass aus stummen Videosequenzen automatisch eine sinnvolle textuelle Darstellung generiert werden kann.

Motivation

Die Realisierung dieses Projekts ist aus mehreren Gründen von Interesse. Ein System, das stumme Videoaufnahmen in Text überträgt, kann in verschiedenen Anwendungsfeldern von Nutzen sein. Für Menschen mit auditiven oder sprachlichen Beeinträchtigungen könnte diese Technologie einen barrierefreien Zugang zu visuellen Inhalten ermöglichen. Da nicht jedes Medium automatisch mit Untertiteln ausgestattet ist, sind betroffene Personen häufig von bestimmten Informations- und Unterhaltungsangeboten ausgeschlossen. Das vorgesehene Tool dient der Unterstützung durch die Auswertung von Aufnahmen, in welchen der Mund der sprechenden Person sichtbar ist, mit dem Ziel der Textgenerierung.

Darüber hinaus ließe sich das Verfahren auch zur Analyse historischer, stummer Videoaufnahmen einsetzen, um deren Inhalte besser verständlich und dokumentierbar zu machen.

Des Weiteren ermöglicht das Projekt eine kritische Auseinandersetzung mit den gesellschaftlichen und ethischen Implikationen solcher Technologien. Die automatisierte “Entzifferung” visueller Informationen löst Bedenken in Bezug auf Datenschutz, Überwachung und den Schutz der Privatsphäre aus, insbesondere in politischen oder sicherheitsrelevanten Kontexten. Das Projekt zielt darauf ab, das technische Potenzial solcher Anwendungen zu demonstrieren und gleichzeitig die potenziellen Risiken und Grenzen zu beleuchten.

KI-Con Präsentation

Im Rahmen der KI-Con wird eine funktionsfähige Software präsentiert, die es ermöglicht, aus stummen Videoaufnahmen gesprochene Worte in kurzer Zeit, idealerweise in Echtzeit, zu identifizieren und in Text zu konvertieren. Die Evaluierung der Leistungsfähigkeit des **AV-HuBERT-Modells** erfolgt durch einen Vergleich mit den im Original-Paper veröffentlichten

Ergebnissen, wobei die Grundlage das bestehende **AV-HuBERT-Modell** bildet. Darüber hinaus werden erste Tests mit dem **GLips-Datensatz** präsentiert, um die Effektivität des Modells auf deutschsprachigem Material zu demonstrieren. Optional wird, je nach verbleibender Projektlaufzeit, auch die Erweiterung des Systems durch Transfer Learning und ein anschließendes Text-to-Speech-Modul vorgestellt.

Technische Umsetzung

Zu Beginn des Projekts wird das bestehende **AV-HuBERT-Modell** in einer Testumgebung eingerichtet und anhand von Beispielvideos überprüft, um die grundsätzliche Lauffähigkeit des Netzwerks und die Generierung sinnvoller Ergebnisse sicherzustellen.

Im nächsten Schritt wird der vollständige **Oxford-BBC Lip Reading Sentences 2 (LRS2)** Datensatz eingesetzt, um die im ursprünglichen **AV-HuBERT-Paper** beschriebenen Performance-Werte nachzuvollziehen. Das Ziel besteht darin, die Reproduzierbarkeit der publizierten Ergebnisse zu prüfen und potenzielle Abweichungen zu analysieren. Darüber hinaus soll evaluiert werden, inwiefern das Modell in seiner aktuellen, englisch trainierten Form auch auf deutschsprachige Daten anwendbar ist. Zu diesem Zweck wird das GLips-German Lipreading Dataset herangezogen, um erste Vergleichsergebnisse zwischen englischem und deutschem Sprachmaterial zu erlangen.

Nach Abschluss dieser Phase erfolgt die Entwicklung einer Softwareanwendung, die eine Anwendung des Modells auf einen Kamerastream ermöglicht. Dies kann entweder in nahezu Echtzeit oder mit einer kurzen Verzögerung erfolgen.

Für den Fall, dass im weiteren Projektverlauf noch ausreichend Zeit zur Verfügung steht, ist als nächster Schritt ein Transfer Learning geplant, um das Modell gezielt an deutschsprachige Daten anzupassen. Im Falle einer verbleibenden Projektzeit ist die Implementierung eines Text-to-Speech-Moduls zu erwägen, welches die erkannten Inhalte automatisch vertont.

Interaktive Demonstation

Die Resultate werden in Form einer Live-Demonstration präsentiert. Im Rahmen dessen werden kurze Videosequenzen präsentiert, aus welchen das Programm in Echtzeit oder mit minimaler Verzögerung den gesprochenen Text erkennt. Darüber hinaus ist vorgesehen, dass die Zuschauer*innen selbst kurze Sätze einsprechen oder hinter einer Scheibe gefilmt werden können, um das System unmittelbar zu testen und die erzeugten Textausgaben in Echtzeit zu beobachten.

Ergebnisse der Literaturrecherche

Als Ausgangspunkt für die eigene Entwicklung konnten bereits mehrere relevante Quellen identifiziert werden:

- https://github.com/facebookresearch/av_hubert?tab=readme-ov-file
- https://www.fdr.uni-hamburg.de/record/10048
- https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs2.html