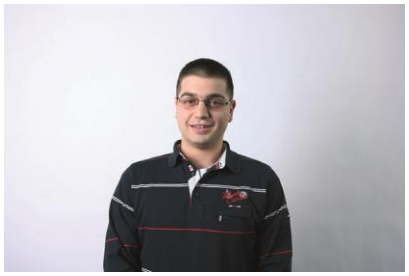
Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

InIT Institut für angewandte
Informationstechnologie

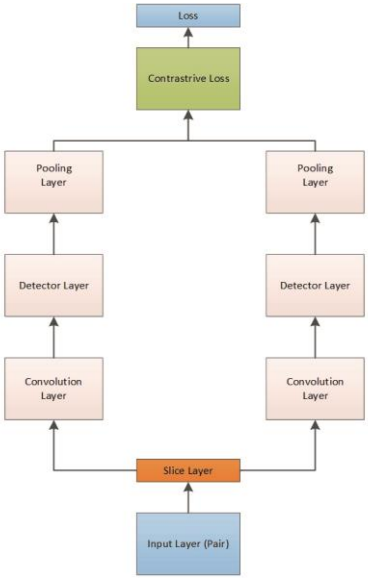
Deep Learning für Automatische Stimmerkennung



Diplomand
Gabriel Eyyi

Dozierende
Oliver Dürr
Thilo Stadelmann

Die vorliegende Arbeit beschäftigt sich mit der wissenschaftlichen Problemstellung des Speaker Clustering: dem Erkennen der Sprecheridentität für jeden Zeitpunkt einer Sprachaufzeichnung. Dabei sind weder die Anzahl der Sprecher noch die Anzahl der Aussagen pro Sprecher oder deren Reihenfolge im Voraus bekannt. Herkömmliche Verfahren sind bis heute nicht in der Lage, Sprecher unter diesen Bedingungen automatisch zu erkennen. Vorangegangene Forschungsarbeiten zeigten jedoch, dass der zeitliche Kontext eines Audiosignals, der wichtige sprecherspezifische Informationen enthält, in herkömmlichen Sprechererkennungssystemen vernachlässigt wird.
Ziel dieser Arbeit ist es, Speaker Clustering durch Ausnutzung dieses zeitlichen Kontexts mithilfe von Deep Learning Verfahren zu verbessern. Verschiedene Ansätze aus angrenzenden Forschungsbereichen werden hinsichtlich der Integration des zeitlichen Kontexts und der Unterscheidung von Sprechern evaluiert. Als vielversprechend erweisen sich Convolutional Neural Networks (CNN), die explizit lokale Zusammenhänge in den Daten erkennen, wie etwa die zeitliche Ausdehnung von Klang auf der horizontalen Achse eines Spektrogramms. Zusätzlich ergeben die Untersuchungen, dass sich eine siamesische Architektur des neuronalen Netzes zum Erlernen eines Distanzmasses zwischen zwei Sprachsegmenten hinsichtlich der Ähnlichkeit der beiden Sprecher anbietet. Mithilfe dieses Distanzmasses soll das Speaker Clustering realisiert werden.
Für die Evaluation des Konzepts werden systematische Experimente in drei Versuchsreihen auf dem TIMIT-Datensatz durchgeführt. Die erste Versuchsreihe dient zur Vorbereitung. In der zweiten Versuchsreihe wird das Konzept implementiert und evaluiert. In der dritten Versuchsreihe wird das experimentelle Setup vereinfacht, um das Verhalten des CNNs zu analysieren. In dieser dritten Versuchsreihe ist dem Netzwerk die Anzahl und Identität der Sprecher vorgegeben.
Die Ergebnisse aus Versuchsreihe 3, in denen 79.5% der Sprecher korrekt identifiziert werden konnten, legen nahe, dass für diese Anwendung eines CNN zu wenige Daten verfügbar waren. Dies führte zu Overfitting-Problemen, die einen direkten Einsatz der Verfahren als Distanzmass zwischen Sprachsegmenten momentan unmöglich machen.
Für die weitere Forschung liefert dies wertvolle Hinweise: Es sind Verfahren für Audiodaten zu entwickeln, die einen bestehenden Datensatz durch inhaltserhaltende Transformationen künstlich erweitern können.



Die Darstellung zeigt eine siamesische Architektur, bestehend aus zwei Convolutional Neural Networks.