Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

InIT Institut für angewandte
Informationstechnologie

Recognizing birds by voice - the BirdCLEF 2018 challenge

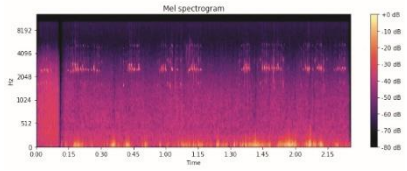
Diplomierende
Mario Marti
Lukas Müller

Dozierende
Thilo Stadelmann
Martin Braschler

Der internationale BirdCLEF 2018 Wettbewerb verfolgt das Ziel, automatische Klassifizierungssysteme für Vogelgesang zu verbessern. Letztlich könnte ein solches System für Ornithologen von Nutzen sein, bietet aber vor allem eine spannende Forschungskulisse für die Verbesserung von Audioerkennungsalgorithmen vor dem Hintergrund schwieriger akustischer Bedingungen: Es gibt einen grossen Trainingsdatensatz mit Audioaufzeichnungen unterschiedlicher Qualität.

Der Wettbewerb besteht aus zwei Teilaufgaben: Bei den 'Soundscapes' besteht das Ziel darin, alle hörbaren Vögel in Segmenten von 5 Sekunden Länge aus einer Aufnahme vorherzusagen, während die Teilnehmer bei 'monophonen Aufnahmen' den am besten hörbaren Vogel klassifizieren sollen.

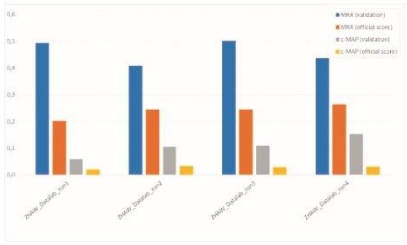
In den letzten Jahren waren Convolutional Neural Networks (CNNs) für die Vogelstimmerkennung im Rahmen des jährlichen BirdCLEF Wettbewerbes sehr beliebt. Während Recurrent Neural networks (RNNs) für viele ähnliche Aufgaben weit verbreitet sind, gelang es bislang niemandem, sie für BirdCLEF zu verwenden. Die jüngsten Arbeiten an der ZHAW zum Thema Speaker Clustering haben mit einem bidirektionalen LSTM, einer RNN Variante, vielversprechende Ergebnisse erbracht. In dieser Arbeit wollen wir untersuchen, ob ein LSTM für die gestellte Aufgabe geeignet ist und gleich gute oder gar bessere Ergebnisse erzielen kann als ein CNN.



Spektrogramm mit Vogelgesang und stillen Bereichen dazwischen, etwa von Minute 1:17 bis 1:30. Ebenso zeigt sich kurz vor Sekunde 15 ein Punkt, an dem jemand die Aufnahme bearbeitet hat.

Für das Training wurden die Audioaufzeichnungen in Melspektrogramme umgewandelt. Der Datensatz wurde mittels Downsampling auf diejenigen Samples verkleinert, die ein klares Signal enthielten. Das Netzwerk wurde mit Categorical Crossentropy als Loss-Funktion trainiert. Für beide Aufgaben haben wir vier Läufe eingereicht. Drei der Läufe benutzten unterschiedliche Netzwerkarchitekturen, wobei der vierte ein Ensemble von zwei der vorherigen Läufe war.

Bei der monophonen Aufgabe haben wir Mean Reciprocal Rank (MRR) Werte zwischen 0.2 und 0.26 erreicht. Auf den Soundscapes lagen die vom LSTM erreichten Classification Mean Average Precision (c-MAP) Werte zwischen 0.019 und 0.032. State of the Art Systeme erreichen im monophonen Task Werte von etwa 0.82 und für die Soundscapes einen Wert von fast 0.2.



Die Ergebnisse, die wir für die monophone Aufgabe (MRR) und die soundscape Aufgabe (c-MAP) erreicht haben. Bei beiden Aufgaben hat sich unser Ansatz auf einem lokalen Validierungsset besser bewährt als auf dem offiziellen Testset.

Die Frage, ob ein LSTM bei dieser Aufgabe ebenso gut funktioniert wie ein modernes CNN, bleibt nach der Arbeit weiter offen.