Analyse von Deep Learning Methoden für eine Orca Geräusch Erkennung

Nils Bohnhof, Jan-Ole Perschewski, Sebastian Stober

AI Lab, Institut für Intelligente Kooperierende Systeme, 39106 Magdeburg, Deutschland, Email: Nils-Bohnhof@web.de

Einleitung

Orcas sind hochintelligente Tiere mit einer komplexen Kommunikations- und Populationsstruktur. Aufgrund von Nahrungsmangel, Umweltverschmutzung, Schiffslärm und anderen Faktoren sind Orca-Arten, wie der Southern Resident Killer Whale, vom Aussterben bedroht. Um das zu verhindern werden Orcas unter dem Hauptaspekt der Kommunikation, Lokalisierung und sozialen Interaktion erforscht. Meeresbiologen haben in der Vergangenheit hunderte von Stunden an Unterwasseraufnahmen gehört, um mögliche Orcas zu entdecken. Das Ziel ist die Entwicklung und Analyse von robusten Methoden zur automatischen Erkennung von Orca-Rufen. Dazu werden tiefe neuronale Netze auf Audiodateien vom OrcaLab trainiert. Erstmalig werden Methoden zur Daten-Augmentierung, Regularisierung, Datenrepräsentationen und Modellarchitekturen auf diesem Datensatz untersucht. Zusätzlich wenden wir eine neuartige Methode der Datenvorverarbeitung an, um die gelabelten Daten besser zu nutzen. Außerdem werden teilüberwachte Methoden aus der Bilderkennungsdomäne auf die Audiodomäne übertragen.
Das beste Modell verbesserte den F1-Score des Basismodells von 0.69 auf 0.90 auf dem von OrcaLab bereitgestellten Testdatensatz. Auf dem Orca Activity Sub-Challenge-Datensatz der Interspeech ComParE Challenge 2019 erreichte das Modell auf dem Testdatensatz eine AUC von 0.903, welches einer Verbesserung des Basismodells um 0.037 entspricht.

Orca-Vokalisierung

Orcas nutzen ihre Rufe für die Kommunikation, Ortung und Unterscheidung von Objekten. Sie besitzen ein komplexes Kommunikationssystem. In Abbildung 1 sind die drei Kommunikationsarten in Spektrogrammen visuell abgebildet. Für die soziale Kommunikation verwenden Orcas Pfeif- und Impulsrufe. Pfeif-Rufe haben einen Frequenzbereich von 0.5 bis 40 kHz und dienen zur lokalen Kommunikation und zur Verhaltenskoordination [2]. Für die Gruppenzuordnung verwenden Orcas Impulsrufe, welche eine Frequenz von 0.5 bis 25kHz haben [2]. Zur Objekterkennung und Unterscheidung, benutzen Orcas Klick-Geräusche. Die Hauptfrequenz der Echoortung liegt bei 20 bis 30 kHz und bei 40 bis 60 kHz [2].

OrcaLab Datensatz

Das OrcaLab[1] wurde 1970 von Dr. Paul Spong gegründet mit dem Hauptsitz im Norden von Vancouver Island in Kanada. Im Rahmen des Projektes werden Orcas mithilfe eines Netzwerkes von Unterwasser-Mikrofonen erforscht. Getreu dem Motto: „Es ist möglich, die Wildnis zu studieren, ohne in das Leben oder den Lebensraum einzugreifen“[1] umfasst das Projekt zusätzlich Ziele wie

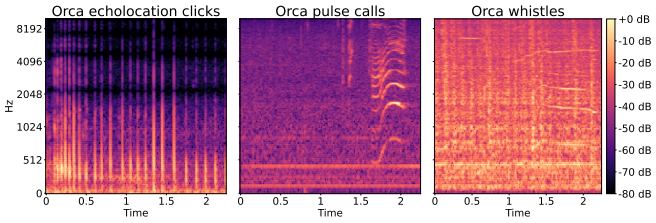

Abbildung 1: Spektrogramme der drei Hauptkommunikationsarten von Orcas. Links: Echoortungsklicken. Mitte: Impulsrufe. Rechts: Pfeif-Rufe.

Natur- und Tierschutz, Erhaltung und Rehabilitation der Orca Spezies und Schutz der Lebensräume dieser Tiere. Das OrcaLab umfasst verschiedenste Projekte wie das Orcasound und AI4Orcas Projekt [1]. Orcasound hat das Ziel, die Unterwasser-Mikrofone weiter auszubauen und der Entwicklung von neuen bioakustischen Lösungen für die Aufnahme von Orca Geräuschen. Orcasound stellt eine große Menge von Audioaufnahmen zur Verfügung, welche teilweise durch Meeresbiologen gelabelt wurden. Neben gelabelten Daten, stehen eine große Datenmenge an ungelabelten Audioaufnahmen zur Verfügung. Das AI4Orcas Projekt nutzt die Aufnahmen des Orcasound Projektes unter anderem für die Erkennung von Orca Vokalisierungen. Im Rahmen dieses Projektes wurde ein Datensatz erstellt für das Trainieren und Testen von Modellen, welche das Ziel haben, die Orca Geräusche zeitlich zu erkennen. Der Trainingsdatensatz umfasst Aufnahmen von 1958 bis 2020 mit einer Länge von 17 Stunden, welche in 2863 Audiodateien aufgeteilt sind. Zusätzlich sind 5199 annotierte Events von Orca Geräuschen im Datensatz enthalten, welche auch den Ort und das Datum der Aufnahme umfassen. Der Testdatensatz umfasst 112 Audiodateien mit einer Länge von 40 Minuten und stammt aus den Jahren 2019 und 2020. Die Audioaufnahmen besitzen eine Abtastrate von 20kHz.
Das AI4Orcas Projekt beinhaltet ein Modell für die automatische Erkennung von Orca Geräuschen mithilfe von Deep Learning Methoden. Dieses Modell wird als Referenzmodell mit einem F1-Score von 0.4 verwendet.

Orca Activity Sub-Challenge-Datensatz

Der zweite Datensatz welcher in dieser Arbeit verwendet wird, ist der Orca Activity Sub-Challenge Datensatz, welcher im Rahmen der Interspeech ComParE Challenge 2019 genutzt wurde [3]. Im Datensatz sind Orca Aufnahmen enthalten, welche von 2017 bis 2018 in Northern British Columbia aufgenommen wurden. Insgesamt besitzt der Datensatz eine Länge von ca. 4 Stunden und eine Abtastrate von 44.1 kHz. Als Referenz enthält die Challenge ein Basismodell, welches auf dem bereitgestellten Testdatensatz einen AUC von 0.866 erreicht [3]. Als

Metrik für diesen Datensatz wird die Fläche unter der Kurve (AUC) verwendet.

Datenvorverarbeitung

Alle annotierten Audio Dateien werden in ein standardisiertes Format umgewandelt. Dabei wird Mehrkanal-Audio auf einen Audio-Kanal reduziert und in eine einheitliche Abtastrate konvertiert. Die Abtastrate wird durch die Datensätze festgelegt, welches 20kHz für Orcasound Datensatz und 44.1kHz für die Orca Activity Sub-Challenge [3] beträgt. Audio im 16bit Integer Format wird in eine normalisierte float Repräsentation konvertiert.

Die Audiodateien werden in Fenster mit einer wählbaren Fenstergröße aufgeteilt. Jedes Fenster besitzt eine feste Klassenzugehörigkeit, welche das Problem der Audioevent Erkennung auf eine binäre Klassifikation reduziert. Fenster, welche der Klasse 1 angehören, enthalten annotierte Orca Geräusche und die 0 repräsentiert die Klasse, welche keine Orca Geräusche enthält.

Die Datenvorverarbeitung wurde vom AI4Orcas Projekt übernommen und mit verschiedenen Methoden erweitert. Durch die nicht überlappende Fenstereinteilungen gingen Teile der Orca Geräusche in diesem Schritt verloren. Um den Anteil der ungenutzten Daten zu reduzieren, wurde die Datenvorverarbeitung auf die Nutzung von überlappenden Fenstern erweitert. Die Größe der Überlappung kann festgelegt werden und bestimmt maßgeblich die Anzahl der resultierenden Fenster. Außerdem kann festgelegt werden, ob die Überlappung nur für Fenster der Orca Klasse erfolgen soll oder für alle Fenster.

Eine weitere Ursache für die nicht optimale Nutzung der Orca Geräusche ist die große Anzahl an überlappenden Annotationen, welche man in eine Annotation zusammenfassen kann. Überlappende Annotation treten auf, wenn ein annotiertes Orca Geräusch e_1 nach dem Orca Geräusch e_2 anfängt und das Geräusch e_2 nach e_1 aufhört. Sei die Startzeit als t_s und die Endzeit als t_e definiert, dann tritt der Fall auf, wenn $t_s(e_1) < t_s(e_2)$ und $t_e(e_2) > t_s(e_1)$ gilt. In diesem Fall können beide Annotationen zu einer Annotation mit Startzeit $t_s(e_1)$ und Endzeit $t_e(e_2)$ zusammengefasst werden. Durch die Zusammenfassung können die Daten noch besser ausgenutzt werden, besonders bei der Nutzung von überlappenden Fenstern.

Die resultierenden Fenster werden durch eine Kurzzeit-Fourier-Transformation vom Zeitbereich in den Frequenzbereich transformiert. Anschließend werden verschiedene Repräsentationen untersucht. Dazu zählen: Mel Spektrogramm, Lineare Frequenz Kompression und Mel Frequency Cepstral Coefficients. Das AI4Orcas Projekt nutzte bis dahin ausschließlich Mel Spektrogramme.

Daten Augmentierung

Die Augmentierung der Daten werden in zwei verschiedene Ansätze geteilt. Als erstes werden Augmentierungsmethoden auf den Audio Dateien verwendet. Der zweite Ansatz ist die Augmentierung der resultierenden Spektrogramme.

Für die Augmentierung auf den Audio Dateien werden drei verschiedene Methoden verwendet, welche zufällig die folgenden Eigenschaften verändern: Abspielgeschwindigkeit, Tonhöhe, zeitliche Verschiebung.

Als zweite Augmentierung auf den Spektrogrammen wird die Methode der zufälligen Maskierung von Bereichen des Spektrogrammes aus dem AI4Orcas Projekt übernommen, welche den Namen SpecAug [8] besitzt. Dabei werden zufällige Zeit- und Frequenzblöcke abgedeckt, sodass das Modell mit weniger Informationen umgehen muss. Zusätzlich wird die Kombination von Augmentierungsmethoden auf Audio Dateien und Spektrogrammen genutzt.

Modelle

Für die Modellierung wurden verschiedene Architekturen explorativ untersucht, welche sich eignen für die Klassifikation von den resultierenden Spektrogrammen. Dabei sollten die Modelle im Idealfall die typischen visuellen Muster erkennen, die Orka Geräusche in Spektrogrammen aufweisen. Mit der Annahme, dass diese Aufgabe mit Methoden aus der Bilderkennung gelöst werden kann, wurden verschiedene Modelle aus diesem Bereich evaluiert.

Das AI4Orcas Repository benutzte Transferlernen auf einem VGGish Modell. Wie im F1-Score ersichtlich wurde, hatte das Basismodell nicht die ausreichende Komplexität um die visuellen Muster im Spektrogramm zu erlernen, welche repräsentativ für die Klassen sind. Daher wurde, angelegt an die Arbeit von [7], die Nutzung von ResNet untersucht. Dabei wurden vier verschiedene Architekturen des ResNets verwendet. Dazu zählen: ResNet18, ResNet34, ResNet50 und das ResNet101.

Überwachtes vs. Teilüberwachtes Lernen

Aus dem Orcasound Projekt enstehen große Mengen an Unterwasseraufnahmen, welche teilweise von Meeresbiologen gelabelt wurden. Jedoch enthält der größte Anteil der Daten keine Annotationen, wodurch die Daten im Bezug auf die Erkennung von Orca Geräuschen bisher ungenutzt blieben. Im Rahmen dieser Arbeit werden sowohl Methoden aus dem überwachten als auch dem teilüberwachten Lernen verwendet. Das überwachte Lernen nutzt als Fehlerfunktion die binäre Kreuzentropie. Für das teilüberwachte Lernen werden Methoden aus dem Bildbereich an den Audiobereich adaptiert. Die Methoden, die im Rahmen dieser Arbeit verwendet werden, sind: MixMatch (MM)[4] und der Mutual Exclusivity Loss (MEL)[6].

MixMatch[4] ist ein Ansatz, der Labels mit einer geringen Entropie vorhersagt für augmentierte Daten ohne Labels. Anschließend werden ungelabelte Daten und deren pseudo Labels mit gelabelten Daten vermischt. Die Autoren haben MixMatch als teilüberwachten Bilderkennungsalgorithmus vorgestellt. Mit der Annahme, dass Spektrogramme wie Bilder behandelt werden können, wurde MixMatch in die Domäne für die Nutzung auf Spektrogrammen adaptiert. Dazu wurden die Augmentierungsmethoden mit den vorgestellten ersetzt. Die Augmentierung wird sowohl auf Audio Daten als auch auf Spektrogrammen ausgeübt. Zusätzlich wird die unüberwachte L2 Fehlerfunktion mit der Kreuzentropie verglichen, welche auch in der Arbeit von [5] genutzt wurde.

Als zweite teilüberwachte Lernmethode wurde der Mutual Exclusivity Loss [6] verwendet. Dabei wurde ein zusätzlicher Term in die Fehlerfunktion hinzugefügt, welcher für die gelabelten und ungelabelten Daten verwendet wird. Dieser Term hat die Eigenschaft, dass sich Vorhersagen, die sich für jede Klasse gegenseitig ausschließen, belohnt werden. Somit werden Vorhersagen, die nah an Entscheidungsgrenzen liegen, bestraft.

Experimente

Das Ziel dieser Arbeit ist die Analyse von verschiedenen Deep Learning Methoden, welche geeignet sind für eine Orca Geräusch Erkennung. Daher werden in verschiedenen Experimenten die folgenden Hauptfragen untersucht:

- Welche Methoden können für die Orca Geräusch Erkennung verwendet werden?
- Wie kann die Orca Geräusch Erkennung durch die zusätzliche Nutzung von ungelabelten Daten verbessert werden?

Dabei werden die vorgestellten Parameter und Methoden nacheinander auf dem OrcaLab Datensatz untersucht. Diese sind: Fenstergröße in Sekunden, Nutzung von überlappenden Fenstern und der Zusammenfassung von Events, FFT-Größe und Hop Größe der schnellen Fourier Transformation, Frequenz Repräsentation, Daten Augmentierung und Regularisierung, Modellarchitekturen. Außerdem werden die teilüberwachte Methoden MM und MEL auf dem OrcaLab Datensatz untersucht. Anschließend wird ein Modell mit den besten Parameter vom OrcaLab Datensatz für die Orca Activity Sub-Challenge [3] verwendet.

Ergebnisse

Um die Methoden zu untersuchen, werden nacheinander Parameter und Methoden auf dem OrcaLab Datensatz untersucht. Dabei wird ein Parameter untersucht und alle anderen fixiert. Die initialen Parameter werden vom OrcaLab Basismodell übernommen.
Der erste untersuchte Parameter ist die Fenstergröße, wobei Fenster der Größe 1.5, 1.75, 2, 2.25 und 2.45 Sekunden miteinander verglichen wurden. Die Fensterlängen von 2.25 und 1.75 schnitten am besten ab, mit einem F1-Score von 0.84 und einem Recall von 0.75. Nur in der Precision unterscheiden sich die zwei Parameter, da die Fensterlänge von 2.25 Sekunden einen Precision Score von 0.96 und 1.75 Sekunden von 0.94 erzielt. Anschließend wurden die Nutzung von überlappenden Fenstern untersucht. Die Fenster nutzen eine Überlappung von einem viertel bzw. halben Fenster. Außerdem wurde die Nutzung der Zusammenführung der Events und die Nutzung von überlappenden Fenster nur für die Orca Geräusch Klasse untersucht. Alle Methoden weisen einen geringeren F1-Score, hohe Precision aber niedrigen Recall im Gegensatz zum Modell mit nicht überlappender Fenstereinteilung. Die Nutzung einer Überlappung von einem viertel des Fensters in Kombination mit dem Zusammenführen der Events resultierte in einem F1-Score von 0.77, welches 0.07 schlechter ist als ohne der Nutzung von überlappenden Fenstern. Für die FFT-Größe wurden 1024, 2048 und 4096 Samples mit einer Hop Größe von 256, 441 und 512 Samples verwendet. Die FFT-Größe von 1024 mit einer Hop Size von 256 schnitten mit einem F1-Score von 0.88 am besten am vor 4096 und 512. Für die Repräsentation im Frequenzbereich, erzielte das Mel Spektrogramm mit 256 Mel Filtern und einem F1-Score von 0.89 die besten Resultate. Die lineare Frequenzrepräsentation fiel durch einen hohen Precision von 0.99 aber geringeren Recall von 0.69 im Gegensatz zum Mel Spektrogramm. MFCC schneidet mit einem F1-Score von 0.83 besser als die lineare FreqenzRepräsentation und schlechter als das Mel Spektrogramm. Die Augmentierung verbessert die Qualität des Modells bezüglich des F1-Scores um 0.01. Dabei erzielten sowohl SpecAug [8] als auch die Nutzung von zufälligen Abspielgeschwindigkeiten, Tonhöhen, zeitlichen Verschiebungen und Rauschen als Augmentierung einen F1-Score von 0.90. Dadurch, dass der Recall von SpecAug höher ist, wird SpecAug verwendet. Bedingt durch die Gefahr von Overfitting werden verschiedene Regularisierungs Methoden verwendet. Dazu zählt die Nutzung des AdamW Optimierers, Weight decay und Focal Loss. Außerdem wird die Methode Stochastic Depth [9] untersucht, welche zur Regularisierung für die ResNet Architektur eingesetzt wird. Die Regularisierungs Methoden erzielen keine Verbesserung im Gegensatz zu keiner Regularisierung. Stochastic Depth [9] führte zu einer höheren Precision aber niedrigeren Recall im Vergleich zum Modell, welches keine Regularisierung nutzt. Bei der Gegenüberstellung der Modelle zeigte sich, dass alle Modelle sehr ähnliche F1-Scores erzielten, unabhängig von der Größe des ResNets. Daher wird das kleinste Modell verwendet, welches einen F1-Score von 0.90 erzielte.
Die besten gefundenen Parameter sind in der Tabelle 1 abgebildet. Anschließend wird der F1-Score und Recall für dieses Modell auf dem Training-, Validierungs- und Testdatensatz in Tabelle 2 gezeigt.
Um zu testen, ob diese Parameter auch für einen anderen Datensatz geeignet sind, wurde ein Modell auf dem Trainingsdatensatz der Orca Activity Sub-Challenge-Datensatz der Interspeech ComParE Challenge 2019 [3] trainiert. Das Modell erzielte eine AUC auf dem Testdaten von 0.903 und verbesserte somit den AUC des Basismodells um 0.037.

Tabelle 1: In der Tabelle sind die besten Hyperparameter bezüglich des F1-Scores des Testdatensatzes vom OrcaLab aufgelistet. Die Nutzung von Regularisierungs Methoden, überlappenden Fenstern und Zusammenfassung der Event Liste sorgten für keine Verbesserung der Vorhersagequalität.

Hyperparameter	Beste Werte
Fenstergröße in Sekunden	2.25
FFT-größe in Samples	1024
Hop größe in Samples	256
Best Freq. Repräsentation	Mel
Größe der Freq. Repräsentation	256
Augmentierung	SpecAug
Modell Architektur	ResNet18

Das nächste Experiment beschäftigt sich mit der Frage, ob die zusätzliche Nutzung von ungelabelten Daten und

Tabelle 2: Die Tabelle zeigt den F1-Score und Recall des Modells mit dem besten Ergebnis auf dem OrcaLab Datensatz bezüglich des F1-Scores.

OrcaLab Datensatz	F1-Score	Recall
Training	0.8811	0.8959
Validation	0.8783	0.8170
Test	0.9026	0.8501

somit von den teilüberwachten Lernmethoden MixMatch [4] und Mutual Exclusivity Loss [6] eine Verbesserung auf dem OrcaLab Datensatz herbeiführt. Die Ergebnisse beider Methoden sind in Tabelle 3 abgebildet. Es ist ersichtlich, dass die teilüberwachten Methoden einen sehr hohen Precision Wert erreicht, aber dafür einen niedrigen Recall und somit auch F1-Score im Vergleich zu dem überwachten Modell.
Bei der Nutzung von MixMatch erzielte die Nutzung von Augmentierungsmethoden auf dem Audio bessere Ergebnisse als die Spektrogramm Augmentierung. Der MEL [6] schneidet ebenfalls besser ab ohne die Nutzung von SpecAug.

Tabelle 3: F1-Score, Recall und Precision werden in der unteren Tabelle für die teilüberwachten Methoden gezeigt.

Model	F1-Score	Recall	Precision
Überwachtes Modell	**0.90**	**0.85**	0.96
Spektrogramm MM	0.62	0.45	**0.99**
Audio MM	0.68	0.52	**0.99**
MEL	0.74	0.60	**0.99**
MEL + SpecAug	0.64	0.51	0.86

Zusammenfassung

In dieser Arbeit wurde das AI4Orcas Projekt um eine Vielzahl von Methoden angepasst. Diese sind: überlappende Fenster, Zusammenführung von Events, Audio Augmentierung, Regularisierung und Modell Architekturen. Ebenfalls wurden die teilüberwachten Methoden MixMatch und Mutual Exclusivity Loss an die Audio Domäne angepasst, um bisher ungenutzte ungelabelte Daten zu verwenden. Anschließend wurden die Parameter und Methoden für die Orca Activity Sub-Challenge vom Interspeech 2019 ComParE Challenge verwendet.
Die besten Parameter sind in der Tabelle 1 dargestellt. Das Modell erreicht einen F1-Score von 0.90 und verbessert den F1-Score des Basismodells vom AI4Orcas Projekt um 0.21. Das größte aufgetretende Problem ist die unbalanzierte Klassenverteilung, welche die Gefahr mit sich bringt, in lokalen Minima stecken zu bleiben. Somit erzielen Modelle, welche nur die häufigere Klasse vorhersagen, einen kleineren Fehler. Einige der untersuchten Methoden, schienen dieses Problem zu begünstigen. Dazu zählt die Nutzung von überlappenden Fenstern und die verwendeten Regularisierungs Methoden ausgenommen von Stochastic Depth.
Für die Nutzung von teilüberwachten Methoden, wurden ungelabelte Daten genutzt, welche vom OrcaLab aufgenommen wurden. Das Verwenden von der MM und MEL Methode, führte zu einer hohen Precision und niedrigen Recall. Dies bedeutet, dass die Hinzunahme von ungelabelten Daten vermehrt zur Vorhersage der Mehrheitsklasse führt. Das kann resultieren aus der unzureichenden Datenqualität der ungelabelten Daten, welche nicht zwingend die selbe Datenverteilung wie der OrcaLab Datensatz aufweisen muss.
Eine Fehleranalyse zeigte, dass knapp über ein Drittel der FN Klasse (114 Beispiele auf dem Testdatensatz) nicht eindeutig der Orca Klasse zuzuordnen sind. Die FP Klasse enthält 18 von 27 Fenstern welche Orca Geräusche enthalten, aber nicht der Orca Klasse zugeordnet sind. Dies zeigt, dass einige Fehler in den Annotationen vorhanden sind.
Zusätzlich zum OrcaLab Datensatz, verbesserte das Modell mit den Parametern aus Tabelle 1 das Basismodell von 0.866 auf 0.906. Dabei ist herauszuheben, dass sich der Datensatz sehr vom dem OrcaLab Datensatz unterscheidet. Der größte Unterschied befindet sich in der Abtastrate der gestellten Audiodateien von 44.1 kHz im Gegensatz zu 20kHz.

Literatur

[1] OrcaLab, URL: `https://orcalab.org`

[2] SeaWorld Parks and Entertainment, URL: `https://seaworld.org/animals/all-about/killer-whale/communication/`

[3] B. Schuller, A. Batliner, C. Bergler, F. Pokorny, J. Krajewski, M. Cychosz, R. Vollmann, S. Roelen, S. Schnieder, E. Bergelson, A. Cristia, A. Seidl, A. Warlaumont, L. Yankowitz, E. Noeth, S. Amiriparian, S. Hantke, and M. Schmitt. The interspeech 2019 computational paralinguistics challenge: Styrian dialects, continuous sleepiness, baby sounds and orca activity. pages 2378–2382, 09 2019.

[4] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. Raffel. Mixmatch: A holistic approach to semisupervised learning, 2019.

[5] L. Cances, E. Labbé, and T. Pellegrini. Improving deeplearning-based semi-supervised audio tagging with mixup, 2021.

[6] M. Sajjadi, M. Javanmardi, and T. Tasdizen. Mutual exclusivity loss for semi-supervised deep learning, 2016.

[7] C. Bergler, H. Schröter, R. Xi Cheng, V. Barth, M. Weber, E. Noeth, H. Hofer, and A. Maier. Orca-spot: An automatic killer whale sound detection toolkit using deep learning. Scientific Reports, 9, 12 2019

[8] D. S Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D Cubuk, and Q. V Le. Specaugment: A simple data augmentation method for automatic speech recognition. In Proc. Interspeech 2019, pages 2613–2617, 2019.

[9] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Weinberger. Deep networks with stochastic depth, 2016.