Eine qualitative Untersuchung der Generalisierungsverhaltens von CNNs zur Instrumentenerkennung

Roman B. Gebhardt[1,2], Athanasios Lykartsis[1], Stefan Weinzierl[1]

[1] *TU Berlin, Fachgebiet Audiokommunikation, Straße des 17. Juni 135, 10623 Berlin, Deutschland*

[2] *Cyanite, elceedee UG, Gneisenaustr. 44/45, 10961 Berlin, Deutschland*

Email: roman@cyanite.ai, athanasios.lykartsis@tu-berlin.de

Abstract

Künstliche neuronale Netze (ANNs) haben sich im Bereich des maschinellen Lernens für Audiodaten als erfolgreichstes Werkzeug mit hoher Klassifikationsrate etabliert [1]. Ein bedeutender Nachteil besteht aus wissenschaftlicher Sicht jedoch in der schweren Interpretierbarkeit des von ANNs tatsächlich gelernten Inhalts [2, 3]. Um dieses Problem anzugehen untersuchen wir in dieser Arbeit den Lern- und Generalisierungsprozess eines Convolutional Neural Networks (CNNs) für Multi-Label Instrumentenerkennung in den Hidden Layers des Netzwerks. Wir betrachten die unterschiedlichen Aktivierungen aller Layers durch unterschiedliche Instrumentenklassen um nachzuvollziehen, ab welcher Tiefe das Netzwerk in der Lage ist, zwei von der gleichen Klasse stammenden Stimuli als ähnlich zu erkennen. Wir wiederholen das Experiment mit den gleichen Stimuli für ein auf die Erkennung von vier Emotionen trainiertes CNNs. Dabei bestätigen sich einerseits viele unserer Betrachtungen zum Generalisierungsprozess, gleichzeitig lassen die Ergebnisse darauf schließen, dass das auf Emotionserkennung trainierte Netzwerk in der Lage ist, instrumententypische Patterns zu lernen.

Einleitung

Ziel des Beitrags ist es, qualitative Aussagen zu dem Verhalten des CNNs zu treffen und die Reduktionsprozesse, die in den verschiedenen Layers stattfinden, besser zu verstehen. Zu diesem Zweck betrachten wir durch ein CNN erlernte Feature-Maps, wie sie in Abbildung 1 zu sehen sind. Eine Feature-Map ist eine Repräsentation der Aktivierungen auf einem einzelnen Filter des CNN. Der Vergleich dieser für unterschiedliche Stimuli derselben Instrumentenklasse hilft uns nachzuvollziehen, ab welcher Tiefe das Netzwerk in der Lage ist, zwei von der gleichen Klasse stammenden Inputs als ähnlich zu erkennen. Die in dieser Arbeit betrachteten CNNs treffen auf Grundlage einer Spektrogrammdarstellung des Audiosignals Aussagen über die Wahrscheinlichkeit für die Zugehörigkeit des Stimulus zu einer Klasse, wie etwa des Instruments bzw. einer evozierten Emotion (im Folgenden als *Mood* bezeichnet). Abbildung 1 zeigt sowohl die Spektrogramme zweier Stimuli der Klasse *bass*, sowie die dadurch erzeugten Feature Maps des auf Instrumentenklassifizierung trainierten Netzwerks. Wie ersichtlich wird, weichen die Spektrogramme der beiden Stimuli stark voneinander ab. Mit steigender Zahl durchlaufener Layers gleichen sich die Feature Maps des Filters immer weiter an. So gleicht die Feature Map in der ersten Layer optisch noch stark dem Spektrogramm, während lokale Informationen ab dem dritten Layer kaum noch zu erkennen sind. Die Feature Map der vierten Convolutional Layer beider Stimuli weist kaum noch Unterschiede auf.

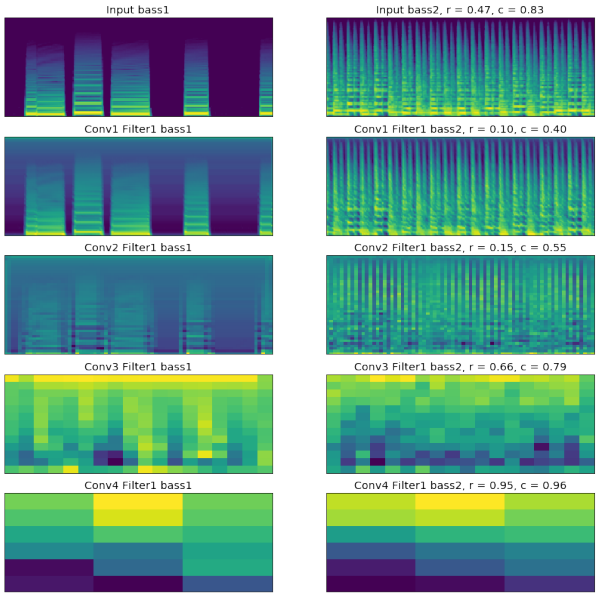

Abbildung 1: Einzelne Feature-Maps für zwei Bass-Spur-Stimuli. Von oben nach unten sind die Spektrogramme (Input-Layer) und die erste Feature Map der Convolutional Layers 1-4 dargestellt.

Mithilfe dreier Ähnlichkeitsmaße ermöglichen wir eine Quantifizierung der Ähnlichkeit zwischen den Feature-Maps von Stimuli der gleichen Instrumentenklasse über die Layers des Netzwerks hinweg. Die Ähnlichkeiten je nach Instrumentenklasse sollen auch darüber Aufschluss geben, für welche Instrumente eine Generalisierung besser gelingt.

Methode

Im Folgenden beschreiben wir zunächst die Struktur des genutzten Datensatzes, den wir für das Training des Netzwerks zur Instrumentenklassifizeriung sowie unser erstes Experiment genutzt haben. Die Architektur der beiden genutzten Netzwerke sowie die Wahl unserer Ähnlichkeitsmaße beschreiben wir in den weiteren beiden Unterkapiteln.

Datensatz

Zum Training des Netzwerks sowie zur Evaluation unseres ersten Experiments nutzen wir den *musdb18* Datensatz [4] für Source Separation, welcher aus isolierten 150 Musikstücken mit isolierten Stems der Instrumententypen *drums*, *bass*, *vocals* und *rest* besteht. Dabei stellen letztere eine Zusammenfassung aus allen restlichen Instrumenten wie etwa Gitarren oder Piano dar. Wir benutzen den Train-Split des Datensatzes mit 100 Tracks zum Trainieren unseres Netzwerkes. Dabei unterteilen wir alle Tracks in 10-sekündige Abschnitte und beziehen nur die Abschnitte mit ein, bei denen jede der vier Instrumentengruppen mindestens in 5s aktiv ist. Für diese Abschnitte erzeugen wir alle möglichen Kombinationen an Instrumenten, bei denen zumindest eines der Instrumente aktiv ist, was zu insgesamt 15 verschiedenen Audio-Signalen pro Segment führt. Durch dieses Vorgehen erzeugen wir 24908 Trainings-Samples für einen Multi-Learning-Task. Den Test-Split des Datensatzes benutzen wir für das eigentliche Experiment. Hier separieren wir nach dem oben beschriebenen Schema die Audio-Signale mit jeweils exklusiv einem einzelnen der vier Klassen. Alle Aktivierungen der einzelnen Layers, die durch Stimuli der gleichen Klasse erzeugt werden, sollen im Experiment jeweils paarweise verglichen werden. Die Daten zur Emotionsklassifizierung stammen von einem privaten Datensatz der Firma Cyanite [5]. Es handelt sich hierbei um 10000 Tracks, die jeweils ein Label aus den vier Quadranten der Valence-Arousal-Ebene [6] besitzen.

Netzwerkstruktur

Tabelle 1 zeigt die genutzte Netzwerkarchitektur beider Classifier, bestehend aus 4 Convolutional Layers, die jeweils von einer Max-Pooling Layer gefolgt werden und anschließend zweier Fully Connected (Dense) Layers. Diese Struktur entspricht einem State-Of-The-Art Netzwerk, inspiriert von Choi [7].

$Mel-spectrograminput: 80 \times 300 \times 1$
$Conv3 \times 3 \times 32$
$MP(2,4)(output: 40 \times 75 \times 32)$
$Conv3 \times 3 \times 64$
$MP(3,4)(output: 13 \times 18 \times 64)$
$Conv3 \times 3 \times 64$
$MP(2,5)(output: 6 \times 3 \times 64)$
$Conv3 \times 3 \times 128$
$MP(2,2)(output: 3 \times 1 \times 128)$
$Dense(output: 128 \times 1)$
$Dense(output: 65 \times 1)$
$Output(4 \times 1), sigmoid$

Tabelle 1: Netzwerkstruktur des CNN zur Instrumentenklassifizierung.

Die Input-Ebene entspricht dem log-ampitude Mel-Spektrogramm mit 80 Mel-Bändern des 10s Audio-Snippets wie oben beschrieben. Wir nutzen *lReLU* (leaky Rectified Linear Unit) Aktivierungsfunktionen sowie Batch Normalization nach jedem Layer. Zusätzlich nutzen wir Dropout von 0.2 sowie l2-Regularization nach jeder Dense Layer um Overfitting vorzubeugen. Aufgrund des Multi-Task-Learnings nutzen wir für die Output-Layer die Sigmoid-Funktion.

Die Netzwerkstruktur des Netzwerks zur Emotionserkennung gleicht dem oben Beschriebenen mit der Ausnahme, dass es sich hier um einen Single-Label-Classifier handelt und wir deshalb die Softmax-Funktion für die Output-Layer nutzen.

Ähnlichkeitsmaße

Zur Untersuchung des Generalisierungsverhaltens der oben beschriebenen Netzwerke vergleichen wir die Feature-Maps aller Layers, die durch Stimulipaare der gleichen Klasse erzeugt werden. Nach oben beschriebenem Verfahren erhalten wir 47 10-sekündige Ausschnitte unterschiedlicher Tracks aus dem Test-Set, welche gleichzeitig alle Instrumente enthalten. Die Anzahl aller möglichen Paarkombinationen beläuft sich damit auf $\frac{47 \times (47-1)}{2} = 1081$. Zum Vergleich stellen wir vereinfachend alle Feature-Maps in einem Layer in einem eindimensonalen Vektor dar. In Paarvergleichen werden diese Vektoren nun mithilfe

- der Pearson-Korrelation

$$r_{u,v} = \frac{\sum V_i V_i - n\bar{U}\bar{V}}{(n-1)s_U s_V},$$

- der Kosinus-Ähnlichkeit

$$cos(u,v) = \frac{U \cdot V}{||U|| \cdot ||V||},$$

- und dem Mean-Squared-Error

$$MSE_{u,v} = \frac{1}{n}\sum_{i=1}^{n}(U-V)^2$$

verglichen. V und U stehen dabei für die entsprechenden Vektoren, n die Länge des Vektors, s die Standardabweichung. Schließlich wird pro Layer der Mittelwert der Ähnlichkeitsmaße aller Kombinationen gebildet, um auf ein globales Ähnlichkeitsmaß zu schließen. Um einen Vergleich der MSE zu den Aktivierungen der Feature-Maps zu gewährleisten, normieren wir den MSE für jede Paarkombination auf deren jeweiligen Maximalwert. Während der Trainingsphase lernen einzelne Neuronen gewisse Muster im Input-Signal zu erkennen, die mit zunehmender Tiefe des Netzwerks und damit steigender Nonlinearität komplexerer Struktur sein können. Unter der Annahme, dass die Darstellung eines Instrumentenklangs im Spektrogramm eine komplexe Sruktur darstellt, erwarten wir mit zunehmender Tiefe des Spektrogramms eine steigende Ähnlichkeit der Aktivierungen in den Layern des Netzwerks für Stimulipaare derselben isolierten Instrumentengruppe. Dies wäre in einem Anstieg der Werte der Korrelation und der Kosinus-Ähnlichkeit sowie einem Abfall des Fehlers (MSE) mit steigender Tiefe des Netzwerks abzulesen.

Ergebnisse

Abbildung 2 zeigt die mittleren Ähnlichkeitsmaße der Aktivierungen der Layers des Instrumenten-Classifiers. In der ersten Layer ist zunächst (vor allem bei der Kosinusähnlichkeit) ein Abfall zu betrachten. Dies ist vermutlich darauf zurückzuführen, dass diese, wie aus der Bilderkennung bekannt vertikale und horizontale Kanten detektiert (Edge Detection), was im musikalischem Sinne einem Onset-Detektor entspricht [8]. Da diese nicht als instrumentenspezifisch anzusehen sind, ist ein Abfall der Ähnlichkeit so durchaus schlüssig zu erklären. Wie zu erwarten zeigen im Weiteren sowohl Korrelation als auch Kosinusähnlichkeit das postulierte Verhalten bis zur dritten Convolutional Layer. Eine Ausnahme stellt die *bass*-Klasse dar, die bereits in der Input Layer einen großen Wert an Ähnlichkeit aufweist.

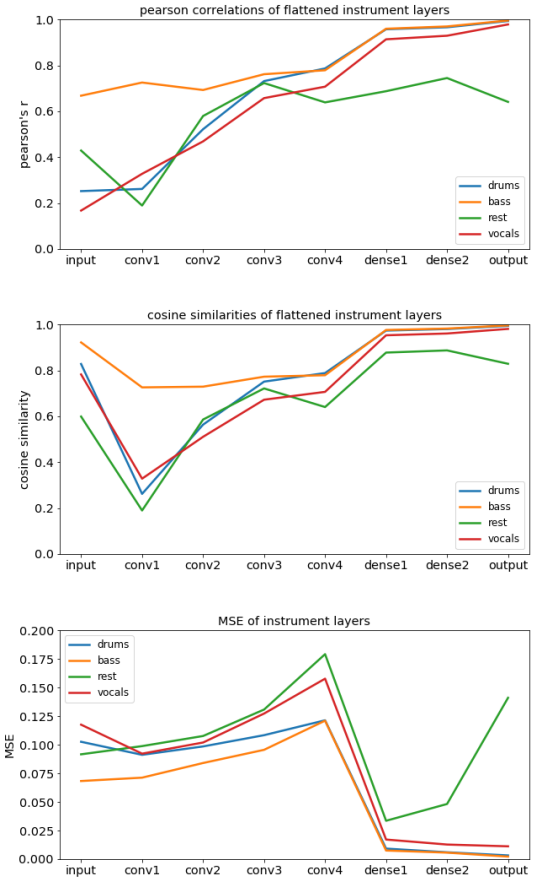

Abbildung 2: Ähnlichkeitsmaße der Layers des Instrumenten-Classifiers für die vier isolierten Instrumentenklassen.

Diese Tatsache ist auf die Frequenzverteilung der Bass-Spur zurückzuführen, die sich auf den unteren Frequenzbereich beschränkt. Der Vergleich von Abbildung 1 mit Abbildung 3, welche dieselben Feature Maps für die Drum-Spur der jeweiligen Test-Stimuli zeigt, verdeutlicht diese Tatsache. Zwischen dritter und vierter Conv-Layer wird ein Plateau erreicht. Der Übergang zur Dense Layer zeigt dann erneut eine Zunahme an Ähnlichkeit. In dieser Tiefe des Netzwerks sticht hier die *rest*-Klasse heraus. Deren Ähnlichkeit liegt hier niedriger als die der anderen Instrumentenklassen, die sich bis zum Output-Layer an den Maximalwert von 1 annähern. Dies ist mit der in der Beschreibung des Datensatzes genannten Tatsache zu erklären, dass sich diese Klasse aus unterschiedlichen Instrumenten zusammensetzt und diese dementsprechend auch schwerer zu generalisieren ist. Zwischen beiden Dense Layers ist der Anstieg der Ähnlichkeit nur noch gering.

Die mittlere quadratische Abweichung zeigt ein etwas überraschenderes Verhalten. Der grundsätzliche Verlauf ist der erwartete: Der Fehler ist am Output kleiner als am Input. Allerdings lässt sich beobachten, dass die mittlere quadratische Abweichung für alle Instrumentengruppen in den Convolutional Layers steigt, bevor sie in den Dense Layers wieder abnimmt. Dies ließe sich auf die Tatsache zurückzuführen, dass die MSE die nicht auf den Betrag normierte Differenz der Vektoren bzw. Feature Maps berücksichtigt. Dies führt dazu, dass der Winkel zwischen den Vektoren kleiner wird, deren tatsächliche Differenz aber größer. Diese Tendenz endet bei den Dense Layers, in denen nur noch nur noch ein sehr geringer Fehler besteht. Es ist anzumerken, dass sich die Unterschiede zwischen den Fehlern in den Conv-Layers in einem verhältnismäßig kleinen Rahmen bewegen.

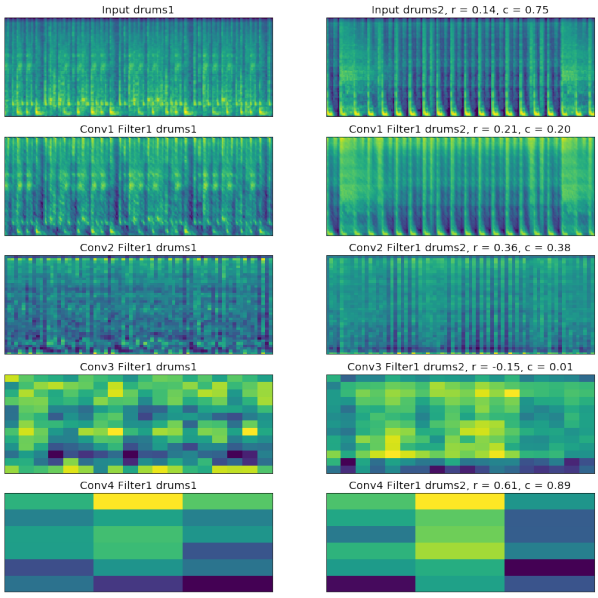

Abbildung 3: Einzelne Feature-Maps für zwei Drums-Spur-Stimuli

Beim Vergleich der Ähnlichkeitsmaße der Feature-Maps des auf Mood-Erkennung trainierten Netzwerks ist ein anderes Verhalten zu betrachten (Abbildung 4). Hier wird vor allem in Hinblick auf die Kosinusähnlichkeit eher eine Stagnation ersichtlich, nachdem in der ersten Conv-Layer bis auf die *rest*-Klasse die Ähnlichkeit ab der Input-Layer zunimmt. Überraschend ist hier, dass der in Abbildung 2 charakteristische Abfall an Ähnlichkeit in Layer 1 ausbleibt. Dies legt die Vermutung nahe, dass vom Mood-Netzwerk offenbar weniger Information bezüglich des Onset gelernt wird als dies beim Instrumenten-Classifier der Fall ist. Weiter ist zu vermuten, dass zwar ein-

fache Muster, die für die jeweilige Instrumentenklasse charakteristisch sind, erkannt werden, dann aber keine weitere Generalisierung bezüglich der Instrumentenklasse stattfindet. In der Outputlayer sehen wir einen starken Abfall der Ähnlichkeit und ein Anwachsen des MSE. Dies folgt der Intuition, da mit der Information über die Zugehörigkeit eines Stimulus zu einer Instrumentenklassen keine Aussage über die evozierte Emotion eines Audiosignals gegeben werden können sollte. Eine Ausnahme stellt hier allerdings die *bass*-Klasse dar, deren Ähnlichkeit in der Output-Layer überraschend hoch ist. Ursache dessen kann nur die Zuordnung dieser Klasse zu meist ein und derselben Mood sein. Dies muss als Fehlverhalten des Classifiers interpretiert werden, welches etwa auf Overfitting, zu wenig Trainingsdaten oder einer zu geringen Tiefe des Netzwerks für die komplexe Aufgaben wie Emotionserkennung zurückzuführen wäre. Schließlich sehen wir auch hier einen starken Abfall der Ähnlichkeit der *rest*-Klasse im Vergleich zu den anderen Instrumentenklassen. Aufgrund dieser Tatsache ist tatsächlich zu vermuten, dass eine Generalisierung bis zu einem gewissen Grad stattfindet.

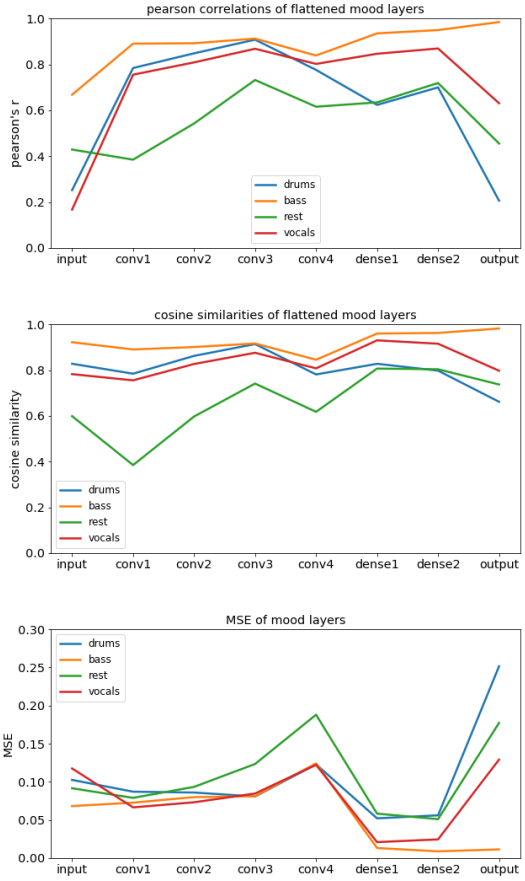

Abbildung 4: Ähnlichkeitsmaße der Layers des Mood-Classifiers für die vier isolierten Instrumentenklassen.

Fazit

In diesem Beitrag haben wir eine Interpretation über das Lernverhaltens der Hidden Layers eines CNN für die Problemstellung der Instrumentenerkennung präsentiert. Dazu haben wir die Feature Maps in den Layers visualisiert und Ähnlichkeitsmaße zwischen den Feature Maps von Stimuli derselben Instrumentengruppe ermittelt. Die Ergebnisse für einen Instrumenten-Classifier zeigen die erwartete Tatsache, dass mit zunehmender Tiefe ein Generalisierungsprozess stattfindet, der nach der dritten Conv-Layer ein Plateau erreicht und dann in den Dense-Layers noch einen Anstieg erreicht. Es ließe sich mutmaßen, dass das Netzwerk bereits nach 3 Layers in der Lage ist, die Komplexität der Aufgabe abzubilden. Interessant wäre für weitere Forschung, ob die Wegnahme einer Conv-Layer ähnlich gute Ergebnisse bei einer geringeren Anzahl an Parametern und somit weniger Rechenaufwand erzielen könnte. Selbiges gilt auch für die potentielle Wegnahme einer Dense Layer. Für zusätzliche qualitative Analyse haben wir das Experiment mit einem Netzwerk für Mood-Erkennung wiederholt. Auch wenn wir zu bedenken geben, dass dieses Netzwerk eine leicht abweichende Architektur besitzt, können wir das Generalisierungsverhalten fragmenthaft feststellen und beschreiben. Ein besseres Verständnis von Feature-Maps ist zusätzlich von großer Bedeutung für Audio-Matching (Shazam [9]) und Musikähnlichkeit[1] (Cyanite [5]).

Literatur

[1] Goodfellow, I.: Deep Learning. MIT Press,, Cambridge, MA, USA, 2016

[2] Zhang C., Bengio S., Hardt M., Recht B. & Vinyal O.: Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530, 2016

[3] Doshi-Velez F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608, 2017

[4] Rafii, Z., Liutkus, A., Stöter, F. R., Mimilakis, S. I., & Bittner, R.: MUSDB18-a corpus for music separation. 10.5281/zenodo.3338373, 2019

[5] Cyanite, URL: `https://cyanite.ai/`

[6] Russell, J..: A circumplex model of affect. Journal of Personality and Social Psychology, 39(6):1161, 1980

[7] Choi, K., Fazekas, G., & Sandler, M.: Automatic tagging using deep convolutional neural network. arXiv preprint arXiv:1606.00298.., 2016

[8] Choi, K., Fazekas, G., & Sandler, M.: Explaining deep convolutional neural networks on music classification. arXiv preprint arXiv:1607.02444., 2016

[9] Shazam, URL: `https://www.shazam.com/`

[1] `https://search.cyanite.ai/`