Inhaltsbasierte Musikempfehlung mit Convolutional Neuronalen Netzwerken

Weidhas Philipp
Matr.nr: 123456
philipp.weidhas@st.oth-regensburg.de

Wildgruber Markus
Matr.nr: 123456
markus.wildgruber@stud.oth-regensburg.de

Zusammenfassung

Hier kommt die Zusammenfassung...

1. Einleitung

Im ersten Halbjahr des Jahres 2017 wurden 62% der Einnahmen der Amerikanischen Musikindustrie durch Streaming Plattformen (wie Spotify, Apple Music, Pandora etc.) erzielt. Im Vergleich zu Vorjahr erhöhten sich dadurch die Einnahmen um 48% auf 2.5$ Milliarde [1][2]. Es zeigt sich, dass automatisierte Empfehlungssysteme weit verbreitet sind.
Obwohl diese in den letzten Jahren viel erforscht wurden, existieren noch Probleme, die bislang zu wenig in Musikempfehlungssystemen berücksichtigt wurden. Neben der Schwierigkeit der große Anzahl an verschiedenen Stile und Genres, beeinflussen sowohl soziales - und geographisches Umfeld, sowie der aktuelle Gemütszustandes die Vorliebe eines Hörers. [3]

Nach Schedl [4] gibt es in der Musik Information Retrieval(MIR) vier Kategorien die einen Einfluß auf die Wahrnehmung von ähnlicher Musik haben. Musikinhalt bezieht sich auf alles, dass aus dem Signal selbst herausgefiltert werden kann. Dazu zählen Aspekte wie der Rhythmus, die Melodie, die Harmonie oder die Stimmung eines Stücks.[5]
Als Musikkontext versteht man alle Aspekte, die nicht aus dem Audiosignal abgeleitet werden, sondern Informationen die über ein Musikstück bekannt sind. Die sogenannten Metadaten wie der Titel eines Lieds, das Genre, Name des Künstlers oder das Erscheinungsjahr [5].
Die Benutzereigenschaften beziehen sie auf Persönlichkeitsmerkmale, wie Geschmack, musikalisches Wissen und Erfahrung oder den demographischen Hintergrund [5].
Im Unterschied dazu steht der Benutzerkontext, der sich auf die aktuelle Situation des Hörers bezieht. Dabei wird er durch seine Umgebung, seiner Stimmung oder der aktuellen Aktivität beeinflußt [5].
Bislang werden Informationen über den Hörer durch ein Benutzerprofil repräsentiert. Das Profil enthält oftmals nur wenig Hintergrund Informationen des Hörers und beschränkt sich auf Lieder, die ein Benutzer angehört und bewertet hat [3]. Das Nutzen dieser Daten um Musikvorschläge zu machen wird als Kollaboratives Filtern (CF) bezeichnet. In der Studie von Vigliensoni und Fujinaga [6] zeigt sich ein deutlicher Unterschied zwischen herkömmlichen Benutzerprofilen und das Einfügen von Zusatzinformationen. Durch das Hinzufügen der Feature demographischen Hintergrund und Entdeckergeist des Hörers konnte im Vergleich zu einem herkömmlichen Profil eine 12% besser Genauigkeit erreicht werden.

Der weiter Verlauf der wissenschaftlichen Arbeit ist wie folgt organisiert. Im 2. Abschnitt werden verschieden historische Ansätze in den jeweiligen Methodenbereichen vorgestellt. Im 3. Kapitel werden die erfolgreichsten Ansätze miteinander verglichen. Teil 4 zeigt ein eige-

nes Experiment zu dem Thema. Abschnitt 5 schließt dieses Arbeit ab und diskutiert zukünftige Forschungsrichtungen.

2. Methoden zur Musikempfehlung

Es gibt vier verschiedene Methoden, die in Musikempfehlungssystemen verwendet werden: kollaboratives -, inhaltsbasierte -, kontextbasiertes Filtern und die hybride Methode [7].

2.1 Kollaboratives Filtern

Kollaboratives Filtern prognostiziert Vorlieben eines Hörers, indem es aus unterschiedlichen Benutzer-Lied Verhältnissen lernt. Es basiert auf der Annahme, dass Verhalten und Bewertungen andere Nutzer auf eine vernünftige Vorhersage für den aktiven Benutzer schließen lassen [8]. Durch explizite oder implizite Rückmeldung an das Empfehlungssystem empfiehlt dieses neue Lieder, indem es Gemeinsamkeiten auf Basis der Bewertungen vergleicht [9].
Eine Vorhersage, ob ein Lied einem Benutzer vorgeschlagen werden soll, kann auf zwei verschiedene Arten erfolgen. In der ersten Methode werden Ähnlichkeiten von Bewertungsmustern verglichen. Lieder werden als ähnlich erachtet wenn sie von den selben Benutzern positiv bewertet wurden. Die zweite Methode berechnet ihre Vorhersage in dem sie ähnliche Profile zu einem Bestimmten sucht und deren Positiv bewerteten Lieder als Vorschlag benutzt. [10]
Verschiedene Studien ([9][11]) zeigen das CF alternative Methoden in der Genauigkeit übertrifft, weshalb es nicht nur im Bereich der Musikempfehlung als die Erfolgreichste gilt.
Trotz der Popularität des CF gibt Probleme die bei der Verwendung der Methode beachtet werden müssen. Beim Cold-Start Problem liegen noch keine Bewertungen für ein Lied vor, wodurch es auch nicht vorgeschlagen werden kann. Das selbe Problem gibt es bei einem neuen Benutzer, diesem kann keine guter Vorschlag gemacht werden, da es an Information mangelt welche Art von Musik ihm gefällt. [8]
Neben dem Cold-Start Problem gibt es noch weitere Probleme die in [8] aufgeführt werden.

2.2 Inhaltsbasierter Filter

2.3 Kontext-basierter Filter

2.4 Hybrider Methode

3. Bestehende Ansätze zur Problemlösung

4. Convolutional Neuronale Netzwerke

Nachdem Alex Krizhevsky mit seinem Team den ImageNet ILSVRc-2012 Contest mit Hilfe eines tiefen Neuronalen Netzwerks (DNN) gewann. Wurden DNNs auch in anderen Bereichen neben der Bildklassifizierung [12] in Gesichtserkennung [13], Spracherkennung [14] und der Inhaltsbasierten Musikempfehlung [3] mehr genutzt und erforscht.
Um diese unterschiedliche Funktionalität zu lernen, werden DNN mit drei verschiedenen Arten trainiert. Dem überwachten Lernen (supervised learning) bei dem das DNN eine Eingabe erhält, dessen Ausgabe bekannt ist. Durch das Vergleichen der Netzwerkausgabe mit der Erwarteten, kann das DNN dementsprechend konfiguriert werden. Beim Unbewachten Lernen (unsupervised learning) erhält das DNN verschiedene Eingaben und soll selbständig zusammenhänge zwischen diesen erkennen. Beim bestärkten Lernen (reinforcement learning) befindet sich das DNN in einer ihm unbekannter Umgebung, die es zu erforschen gilt. Gewünschtes Verhalten wird belohnt, wodurch es lernt die richtigen Entscheidungen zu treffen [15].
Vor allem in den letzten Jahren hat sich das Convolutional Neuronales Netzwerk (CNN) als das erfolgversprechendste DNN erwiesen.
Im folgenden Absatz wird eine Übersicht über

den Aufbau, das Training und die Besonderheiten eines CNNs dargelegt. Anschließend werden verschiedene Ansätze der Inhaltsbasierten Musikempfehlung miteinander verglichen.

4.1 Aufbau eines Convolutional Neuronalen Netzes

Im Unterschied zu regulären DNN verwendet das CNN Neuronen, die drei Dimensionale angeordnete sind. Durch diese Anordnung ist es möglich größere Inputdaten in der selben Geschwindigkeit zu verarbeiten wie zuvor [16]. Um eine CNN Architektur zu erstellen werden drei Haupttypen von Schichten verwendet: Faltungs- (convolutional layer), Vereinigungs- (pooling layer) und einer vollständig verbundenen Schicht (fully-connected layer).

Faltungsschicht

Jede Faltungsschicht besteht aus einem oder mehreren lernfähigen Filtern. Jeder dieser Filter ist räumlich kleiner (Höhe und Breite) aber erstreck sich über die selbe Tiefe der Eingangsmatrix. Durch die Iteration über jeden Punkt in der Eingabematrix erstellt die Faltungsschicht eine zweidimensionale Aktivierungskarte. Anhand dieser erkennt die Schicht dann gewünschte Merkmale wieder [16].
Sei die Eingabematrix I eine 7x7x3 Matrix und K ein 3x3x3 Filter. So wird in der Ausgabematrix S die Stelle (i,j) durch die Gleichung (1) berechnet. Eine genauere Herleitung der Gleichung findet der Leser u. a. bei Goodfellow [17](328f). Die Faltung wird in Abbildung 1 dargestellt.

$$S(i,j) = (I * K)(i,j) \tag{1}$$

$$(I * K)(i,j) = \sum_m \sum_n I(i+m, j+n) K(m,n) \tag{2}$$

Gleichung (2) zeigt eigentlich Cross-Correlation wird aber oft auch als Faltung bezeichnet [17](328)

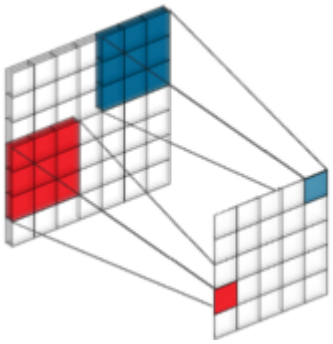

Abbildung 1: Faltung eine 7x7x3 Matrix mit einem 3x3x3 Filter und erzeugter Aktivierungskarte [18]

Verbindungsschicht

Üblicherweise wird eine Verbindungsschicht zwischen zwei Faltungsschichten eingefügt. Seine Funktion besteht darin, schrittweise die Größe der Darstellung zu reduzieren, um die Anzahl der Parameter und dadurch die Berechnung des gesamt Netzwerkes zu verringern [16]. Sie ersetzt die Ausgabe eines Netzes an einem bestimmten Punkt durch eine statistische Zusammenfassung der nahegelegenen Ausgängen. Zum Beispiel Max Pooling [19] Übergabe der größte Zahl in einem rechteckigen Umfeld, Durchschnittsberechnung des Umfeldes oder ein gewichteter Durchschnitt basierend auf die Entfernung eines zentralen Punktes [17](355).

Vollständig verbundenen Schicht

Neuronen in einer vollständig verbundenen Schicht haben Verbindungen zu allen Knoten der vorherigen Schicht. Ihre Aktivierung wird durch eine Matrixmultiplikation und einem Bias-Offset berechnet [16]. Die vollständig verbundenen Schicht wird als Ausgabeschicht verwendet um aus der Eingangsmatrix einen Vektor zu erzeugen.

Training

4.2 Vergleich verschiedener Ansätze

5. Experiment

5.1 Aufbau

5.2 Ergebnis

6. Vergleich mit Stand der Forschung und Ausblick

Literatur

[1] Joshua P. Friedlander. News and notes on 2017 mid-year riaa revenue statistics. *RIAA*, 2017.

[2] Dan Rys. *U.S. Music Industry's Revenue Growth Accelerates As Paid Streaming Subscriptions Rise 50 Percent*. Billboard, 2017. `https://www.billboard.com/articles/business/7972868/us-music-industry-revenue-growth-accelerates-paid-streaming-50-percent`.

[3] Aäron van den Oord, Sander Dieleman, and Benjamin Schrauwen. Deep content-based music recommendation. *Advances in Neural Information Processing Systems 26*, 2013.

[4] Markus Schedl, Arthur Flexer, and Julián Urbano. *The neglected user in music information retrieval research*, volume 36. Springer, 2013.

[5] Peter Knees and Markus Schedl. *Music Similarity and Retrieval*, volume 41. Springer, 2016.

[6] Gabriel Vigliensoni and Ichiro Fujinaga. Automatic music recommendation systems: Do demographic, pro

ling, and contextual features improve their performance? *Proceedings of the 17th International Society for Music Information Retrieval Conference*, pages 94–100, 2016.

[7] Juuso Kaitila. A content-based music recommender system. Master thesis, University of Tmpere, 2017.

[8] Òscar Celma. *Music Recommendation and Discovery - The Long Tail, Long Fail, and Long Play in the Digital Music Space*. Springer, 2010.

[9] B. McFee, T. Bertin-Mahieux, D. P. W. Ellis, and G. R. G. Lanckriet. The million song dataset challenge. *21st International Conference Companion on World Wide Web*, pages 909–916, 2012.

[10] Michael D. Ekstrandand, John T. Riedl, and Joseph A. Konstan. Collaborative filtering recommender systems. *Foundations and Trends in Human-Computer Interaction*, 4:175–243, 2011.

[11] Luke Barrington, Reid Oda, and Gert Lanckriet. Smarter than genius? human evaluation of music recommender systems. *Proceedings of the 10th International Society for Music Information Retrieval Conference*, pages 357–362, 2009.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, pages 1097–1105, 2012.

[13] Changxing Ding and D. Tao. Robust face recognition via multimodal deep face representation. *Multimedia, IEEE Transactions on*, Volume 17:2049–2058, 2015.

[14] Alex Graves, Abdel-Tahman Mohamed, and Geoffrey E. Hinton. Speech recognition with deep recurrent neural networks. *Acoustics, Speech and Signal Processing, IEEE International Conference on*, pages 6645 – 6649, 2013.

[15] Xinixi Wang, Ye Wang, David Hsu, and Ye Wang. Exploration in interactive personalized music recommendation: A reinfor-

cement learning approach. *ACM Transactions on Multimedia Computing, Communications, and Applications*, Volume 11 Issue 1, 2014.

[16] Andrej Karpathy. *Convolutional Neural Networks for Visual Recognition*. Stanford University, 2017. `https://github.com/cs231n/cs231n.github.io`.

[17] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[18] Jonas Knupp. Einführung in deep learning – lstm und cnn. 2015.

[19] Zhou Y. and Chellappa R. Computation of optical flow using a neural network. *IEEE International Conference*, 71–78, 1988.