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]. Dieser Erfolg basiert nicht nur auf einer guten Verfügbarkeit der Lieder und einem günstigen Preis sondern auch auf automatischen Musikempfehlungsdiensten, welche dem Nutzer ein angenehmeres Konsumverhalten ermöglichen. Obwohl Empfehlungsdienste in den letzten Jahren viel erforscht wurden, ist das Problem der Musikempfehlung sehr komplex. Neben einer große Anzahl an verschiedenen Stile und Genres, beeinflussen sowohl soziales- und geographisches Umfeld, sowie der aktuelle Gemütszustandes die Vorliebe eines Hörers. [2]
TO DO

In der Musik Information Retrieval(MIR) gibt es vier Kategorien [3] die einen Einfluß auf die Wahrnehmung von ähnlicher Musik haben.
Musikmerkmale sind Eigenschaften, welche aus dem Audiosignal eines Liedes extrahiert werden. Dazu zählen Aspekte wie der Rhythmus, die Melodie, die Harmonie oder die Stimmung eines Stücks.
Als *Musikkontext* versteht man alle Aspekte, die nicht aus dem Audiosignal abgeleitet werden, sondern Informationen die über ein Musikstück bekannt sind. Beispielsweise Metadaten wie der Titel eines Lieds, das Genre, Name des Künstlers oder das Erscheinungsjahr.
Die *Benutzereigenschaften* beziehen sie auf Persönlichkeitsmerkmale, wie Geschmack, musikalisches Wissen und Erfahrung oder den demographischen Hintergrund.
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. [4]

Bislang werden Informationen über den Hörer durch ein Benutzerprofil repräsentiert. Das Profil enthält nur wenig Hintergrund Informationen des Hörers und beschränkt sich auf Lieder, die ein Benutzer angehört und bewertet hat [2]. Das Nutzen dieser Daten, um Musikvorschläge abzugeben wird als kollaboratives Filtern (KF) bezeichnet. In der Studie von Vigliensoni und Fujinaga [5] zeigt sich ein deutlicher Unterschied zwischen herkömmlichen Benutzerprofilen und das Einfügen von Zusatzinformationen. Durch das Hinzufügen der Features demographischen Hintergrund und Entdeckergeist des Hörers konnte im Vergleich zu einem herkömmlichen Profil eine 12% besser Genauigkeit erreicht werden.

Der weitere Verlauf der wissenschaftlichen

Arbeit ist wie folgt organisiert. Im 2. Abschnitt werden verschieden Ansätze in den jeweiligen Methodenbereichen vorgestellt. Im 3. Kapitel werden die erfolgreichsten Ansätze miteinander verglichen. Teil 4 zeigt ein eigenes Experiment zu dem Thema. Abschnitt 5 schließt diese Arbeit ab und diskutiert zukünftige Forschungsrichtungen. //TO DO

2. Methoden zur Musikempfehlung

Es gibt verschiedene Methoden, die in Musikempfehlungssystemen verwendet werden: kollaboratives -, merkmalsbasiertes -, kontextbasiertes Filtern und die hybride Methode. Diese werden genutzt, um Informationen aus der in der Einleitung genannten Eigenschaften zu gewinnen und diese für Empfehlungen an den Nutzer zu verarbeiten. [6]

2.1 Kollaborativer Filter

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 [7]. Durch explizite (Bewertungen eines Nutzers) und implizite (Beobachten des Konsumverhalten) Rückmeldung eines Hörers an das Empfehlungssystem empfiehlt dieses neue Lieder, indem es Gemeinsamkeiten auf Basis der Bewertungen vergleicht [8].
Im diesen Verfahren wird der Ansatz verfolgt, dass Lieder einem Nutzer auf Grundlage von Nutzungsverhalten anderer Anwender der gleichen Plattform vorgeschlagen werden. In der praktischen Umsetzung bedeutet dies: hört ein Anwender ein bestimmtes Musikstück, werden ihm von der Empfehlungsplattform, Lieder vorgeschlagen welche Nutzer in Zeitraum zuvor nach diesem Stück hörten. Dieses Verfahren geht davon aus, dass durch die Verbindung der Lieder durch vorhergehende Aufrufe eine gute Aussage darüber getroffen werden kann wie gut diese Stücke zusammen passen. Werden Lieder häufig nacheinander gehört, wird diese Verbindung höher bewertet und die Empfehlung häufiger ausgesprochen. Auch wird das Verhalten und der Musikgeschmack des Kunden selbst durch ein System analysiert, um so über Ähnlichkeiten der Kundenpräferenzen mit derer anderer, diesen wiederum bessere Empfehlungen aussprechen zu können. So werden Lieder einem Musikstil zugeordnet und so zielgerichtet dem Nutzer nahegelegt.
Verschiedene Studien ([8][9]) zeigen, dass KF alternative Methoden in der Genauigkeit übertrifft, weshalb es nicht nur im Bereich der Musikempfehlung als die erfolgreichste gilt.
Trotz der Popularität des KF gibt es Probleme, die bei der Verwendung dieser Methode beachtet werden müssen. Das Cold-Start Problem besteht darin, dass noch keine Bewertungen für ein Lied vorliegen, wodurch es auch nicht vorgeschlagen werden kann. Dasselbe Problem gibt es bei einem neuen Benutzer: diesem kann kein guter Vorschlag gemacht werden, da es an Information mangelt welche Art von Musik ihm gefällt. Neben dem Cold-Start Problem gibt es noch weitere Probleme. [7]

2.2 Merkmalsbasierter Filter

Als erstes wird nun ein genauerer Blick auf den inhaltsbezogenen Ansatz geworfen. Mittels diesem Verfahrens werden Nutze Musikstücke aufgrund aus Lieder gewonnener Informationen vorgeschlagen. Dies bedeutet im Detail dass aus den Musikstücken mittels verschiedenster Metriken die Audio Signale eines Liedes analysiert werden um Erkenntnisse über die Stimmung eines Musikstücks, die Frequenz oder Rhythmus zu erhalten. Auf Grund dieser Informationen können Stücke dem Konsumenten vorgeschlagen werden die einen gleichen oder sehr ähnlichen Inhalt bieten.

2.3 Kontextbasierter Filter

2.4 Hybride Methode

Bei hybriden Methoden werden kollaborative, merkmalsbasierte und kontextbasierter Filter miteinander verknüpft, wodurch ein besseres Empfehlungsergebnis mit weniger Nachteilen der einzelnen Methode zu erzielen. Meistens wird ein kollaborativer Filter mit einem der beiden anderem kombiniert.
Als *gewichtet* wird eine hybride Methode bezeichnet, bei der Empfehlungswerte der einzelnen Methoden durch eine Linearkombination zusammengerechnet wird. Das Ergebnis der Linearkombination stellt den Empfehlungswertes eines Liedes dar. Durch unterschiedliche Gewichtung der Methoden kann das Empfehlungsergebnis optimiert werden. Der *wechselnde* Ansatz benutzt ein bestimmtes Kriterium anhand dessen es die Methode zur Vorschlagsbestimmung wechselt. Dies kann beispielsweise dann der Fall sein, wenn der erste Filter kein zuverlässiges Ergebnis (semantische Unterschiede) liefert. Dann wechselt das System den Filter und kann ein besseres Empfehlungsergebnis bekommen. Bei *gemischten* hybriden Empfehlungen werden unterschiedliche Techniken direkt miteinander vermischt. Dadurch kann für ein System mit inhaltsbasierten Filter das Cold-Start Problem vermieden werden.
// TODO Hybride Methoden können einige Nachteile von kollaborativen Filtern entfernen. Allerdings stehen auch sie vor dem Neuen BenutzerProblem. Dennoch sind hybride Methoden sehr beliebt, da Information über einen neuen Benutzer schnell herausgefunden werden (ersten Musikstücke) oder bereits vorhanden sind. [10]

3. Convolutional Neuronale Netzwerke für Audiosignale

CNN sind durch das biologische Sehen inspiriert und konnten den ersten großen Erfolg im Bereich der Bildklassifizierung [11] verzeichnen. Trotzdem werden CNN auch in verschiedenen Audiobereich, wie der Spracherkennung [12] und der MIR mehr genutzt und erforscht. In der MIR nutzen die ersten Forschungen CNNs, um die Aufgabe der Musikgenre-Klassifizierung [13] zu untersuchen.

3.1 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 derselben Geschwindigkeit zu verarbeiten wie zuvor [14]. 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 [14].
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 [15](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 [15](328)

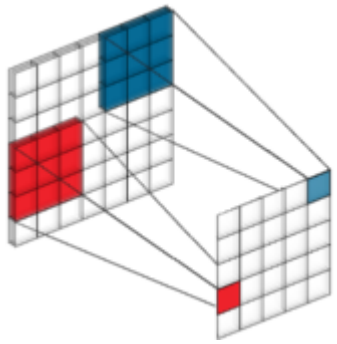

Abbildung 1: Faltung eine 7x7x3 Matrix mit einem 3x3x3 Filter und erzeugter Aktivierungskarte [16]

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 gesamten Netzwerkes zu verringern [14]. Sie ersetzt die Ausgabe eines Netzes an einem bestimmten Punkt durch eine statistische Zusammenfassung von nebeneinander liegenden Ausgängen. Verschiedene Ansätze dafür sind Max Pooling, definiert nach Zhou [17]: eine Übergabe der größten Zahl in einem rechteckigen Umfeld. Weitere Methoden sind die Durchschnittsberechnung des Umfeldes oder ein gewichteter Durchschnitt basierend auf die Entfernung eines zentralen Punktes [15](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 [14]. Die vollständig verbundenen Schicht wird als Ausgabeschicht verwendet um aus der Eingangsmatrix einen Vektor zu erzeugen.

Training

3.2 Vergleich verschiedener Ansätze

4. Experiment

4.1 Aufbau

4.2 Ergebnis

5. Vergleich mit Stand der Forschung

6. Diskussion der zukünftigen Forschungstrends

Literatur

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

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

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

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

[5] 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.

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

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

[8] 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.

[9] 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.

[10] Robin Burke. Hybrid recommender systems: Survey and experiments. *User Model. User-Adapt. Interact*, pages 331–370, 2002.

[11] 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.

[12] 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.

[13] Honglak Lee, Yan Largman, Peter Pham, and Andrew Y. Ng. Unsupervised feature learning for audio classi

cation using convolutional deep belief networks. *Advances in Neural Information Processing Systems*, 2009.

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

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

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

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