

# 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]. Dieser Erfolg basiert nicht nur auf einer guten Verfügbarkeit der Lieder, dem günstigen Preis sondern auch automatische Musikempfehlungsdienste welche den Nutzern ein angenehmeres Konsumverhalten ermöglichen.

Obwohl diese in den letzten Jahren viel erforscht wurden, existieren noch Probleme, die bislang zu wenig in Musikempfehlungssystemen berücksichtigt wurden. Neben der großen Anzahl an verschiedenen Stile und Genre beeinflusst sowohl soziales- und geographisches Umfeld, sowie der aktuelle Gemütszustand die Vorliebe eines Hörers [3].

Nach Scheff [4] gibt es in der Musik Information Retrieval (MIR) vier Kategorien 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. [5] 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 [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.

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 (KF) bezeichnet. In der Studie von Vigliani und Fujinaga [6] 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%iger Genauigkeit erreicht werden.

Der weitere Verlauf der wissenschaftlichen Arbeit ist wie folgt organisiert. Im 2. Abschnitt werden verschiedenen Ansätzen in den jeweiligen Methodenbereichen vorgestellt. Im 3. Kapitel werden die erfolgreichsten Ansätze miteinander

der verglichen. Teil 4 zeigt  eigenes Experiment zu dem Thema. Abschnitt 5 schließt diese Arbeit ab und diskutiert zukünftige Forschungsrichtungen. //TO DO 


## 2. METHODEN ZUR MUSIKEMPFEHLUNG

Es gibt vier 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.[7].

### 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 [8]. Durch explizite oder implizite Rückmeldung an das Empfehlungssystem empfiehlt dieses neue Lieder, indem es Gemeinsamkeiten auf Basis der Bewertungen vergleicht [9].



Im diesen Verfahren wird der Ansatz verfolgt das 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 wird im vom System, Lieder vorgeschlagen welche Nutzer in Zeitraum davor nach diesem diesem Stück hörten. Dieses Verfahren geht davon aus das 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 analysiert um so über Ähnlichkeiten der Kundenpräferenzen mit derer anderer, diesen wiederum bessere Empfehlungen aussprechen zu können.  werden Lieder einem Musikstil zugeordnet und so zielgerichtet dem Nutzer nahegelegt.

Verschiedene Studien ([9][11]) zeigen das 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 Probleme die bei der Verwendung dieser Methode beachtet werden müssen. Beim Cold-Start Problem liegen noch keine Bewertungen für ein Lied vor, 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. [8] Neben dem Cold-Start Problem gibt es noch weitere Probleme die in [8] aufgeführt werden.

### 2.2 Merkmalsbasierter Filter

Als erstes wird nun ein genauerer Blick auf den inhaltsbezogenen Ansatz geworfen. Mittels diesem Verfahrens werden Nutz Musikstücke aufgrund aus Lied gewonnenen Informationen vorgeschlagen. Dies bedeutet im Detail dass aus den Musikstücken mittels verschiedener Metriken die Audio Signale eines Liedes analysiert werden.  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 verschiedene Filtertechniken miteinander verknüpft, wodurch ein besseres Empfehlungsergebnis erzielt werden soll. Meistens wird ein kollaborativer Filter mit einem anderem kombiniert.

Durch diese Kombination können Nachteile einer einzelnen Methode verschwinden. [12]

Burke [12] definiert unterschiedliche Arten von Hybrid-Filtern, die von verschiedenen Forschern benutzt wurden.

Als *gewichtet* wird eine hybride Methode bezeichnet, die alle Ergebnisse einzelner Empfehlungen zusammenfügt und daraus den Wert des empfohlenen Liedes errechnet. Durch unterschiedliche Gewichtung der Methoden kann das der Empfehlungsprozess 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 liefert. Dann wechselt das System den Filter und bekommt ein besseres Empfehlungsergebnis. Bei *gemischten* hybriden Empfehlungen werden unterschiedliche Techniken direkt miteinander vermischt. Dadurch kann für ein System mit inhaltsbasierten Filter das Cold-Start Problem vermieden werden. Als *Waterfall* Methode wird ein gestufter Ansatz bezeichnet, in dem das Ergebnis des ersten Filters als zusätzliche Eingabe des nächsten dient. Durch die Bewertung in der ersten Stufe, wird es möglich die Empfehlungen zu verfeinern.

## 2.5 Bestehende Vorgehensweisen zur Problemlösung

### 3. NEURONALE NETZWERKE UND DEREN ANWENDUNG

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 [13] in Gesichtserkennung [14], Spracherkennung [15] und der inhaltsbasierten Musikempfehlung [3] mehr genutzt und erforscht.

Um diese unterschiedliche Funktionalität zu lernen, werden tiefen Neuronalen Netzwerke (DNN) mit drei verschiedenen Schichten 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 Unüberwachten Lernen (unsupervised learning) erhält das DNN verschiedene Eingaben und soll selbstständig Zusammenhänge zwischen diesen erkennen. Beim Verstärkten Lernen (reinforcement learning) befindet sich das DNN in einer ihm unbekannten Umgebung, die es zu erforschen gilt. Gewünschtes Verhalten wird belohnt, durch es lernt die richtigen Entscheidungen zu treffen [16].

Vor allem in den letzten Jahren hat sich das Convolutionale Neuronale Netzwerk (CNN) als das erfolgreichste 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.

#### 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 große Inputdaten in derselben Geschwindigkeit zu verarbeiten wie zuvor [17]. 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 klein (Höhe und Breite) aber erstreckt sich über dieselbe Tiefe der Eingabematrix. Durch die Iteration über jeden Punkt in der Eingabematrix erstellt die Faltungsschicht eine zweidimensionale Aktivie-

rungskarte anhand dieser erkennt die Schicht dann gewünschte Merkmale wieder [17]. Sei die Eingabematrix  $I$  eine  $7 \times 7 \times 3$  Matrix und  $K$  ein  $3 \times 3 \times 3$  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 Gellermann [18](328f). Die Faltung wird in Abbildung 1 dargestellt.

$$S(i, j) = (I * K)(i, j) \quad (1)$$

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

Gleichung (2) zeigt eigentlich Correlation, wird aber oft auch als Faltung bezeichnet [18](328)

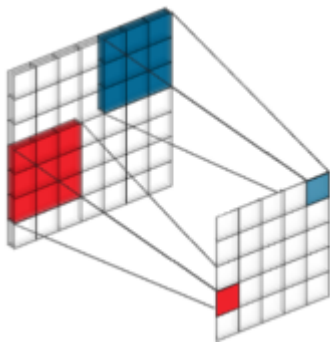


Abbildung 1: Faltung einer  $7 \times 7 \times 3$  Matrix mit einem  $3 \times 3 \times 3$  Filter und erzeugter Aktivierungskarte [19]

### 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 [17]. 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 [20]: eine Übergabe der größten Zahl in einem rechteckigen Umfeld. Weitere Methoden sind die

Durchschnittsberechnung des Umfeldes oder ein gewichteter Durchschnitt basierend auf der Entfernung eines zentralen Punktes [18](351).

### Vollständig verbundenen Schicht

Neuronen in einer vollständig verbundenen Schicht haben Verbindungen zu allen Neuronen der vorherigen Schicht. Ihre Aktivierung wird durch eine Matrixmultiplikation und einem Bias-Offset berechnet [17]. Die vollständig verbundene Schicht wird als Ausgangsschicht verwendet, um aus der Eingabematrix 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] 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, profiling, 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 Tampere, 2017.
- [8] Òscar Celma. *Music Recommendation and Discovery - The Long Tail, Long Tail, 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. Ekstrand, John T. Riedl, and Joseph A. Konstan. Collaborative filtering recommender systems. *Foundations and Trends in Human-Computer Interaction*, 4:175–243, 2011.
- [11] Robin Burke. Hybrid recommender systems: Survey and experiments. *User Model. User-Adapt. Interact.*, pages 331–370, 2002.
- [12] 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.
- [13] 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.
- [14] Changxing Ding and D. Tao. Robust face recognition via multimodal deep face representation. *Multimedia, IEEE Transactions on*, Volume 17:2049–2058, 2015.
- [15] 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.
- [16] Xinxi Wang, Ye Wang, David Hsu, and Ye Wang. Exploration in interactive personalized music recommendation: A reinforcement learning approach. *ACM Transactions on Multimedia Computing, Communications, and Applications*, Volume 11 Issue 1, 2014.
- [17] Andrej Karpathy. *Convolutional Neural Networks for Visual Recognition*. Stanford University, 2017. <https://github.com/cs231n/cs231n.github.io>.
- [18] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [19] Jonas Knupp. Einführung in deep learning – lstm und cnn. 2015.
- [20] Zhou Y. and Chellappa R. Computation of optical flow using a neural network. *IEEE International Conference*, 71–78, 1988.