

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...

ten vorgeschlagen werden die einen gleichen oder sehr ähnlichen Inhalt bieten.

1. EINLEITUNG

2. BESTEHENDE ANSÄTZE ZUR PROBLEMLÖSUNG

Zur Lösung des in der Einleitung erläuterten Problems, gibt es bereits mehrere erprobte und in verschiedenster Weise angewandte Problemlösungsansätze. Diese Ansätze werden in diesem Kapitel nun eingehend erläutert. Im genauen werden aktuell 3 verschiedene Verfahren angewandt, diese sind ein inhaltsbezogener Ansatz, ein Kontextbasierender Ansatz und ein Hybrides Verfahren das beide Verfahren in sich vereint.

2.1 Inhaltsbasierter Filter

TODO: Inhaltsbasierend, Inhaltsbezogen ist meiner Meinung nach der Falsche Ausdruck dafür! :TODO Als erstes wird nun ein genauerer Blick auf den inhaltsbezogenen Ansatz geworfen. Mittels diesem Verfahrens werden Nutz 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 Liedes, die Frequenz oder Rythmus zu erhalten. Auf Grund dieser Informationen können Stücke dem Konsumenten

2.2 Kontextbasierter Filter

Das zweite Verfahren welches nun eingehender betrachtet wird, ist ein kontextbasiertes Verfahren. 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. So werden Lieder einem Musikstil zugeordnet und so zielgerichtet dem Nutzer nahegelegt.

2.3 Hybrider Ansatz

3. CONVOLUTIONAL NEURONALE NETZWERKE

Nachdem Alex Krizhevsky mit seinem Team den ImageNet ILSVRC-2012 Kontest mit Hilfe eines tiefen Neuronalen Netzwerks (DNN) gewann. Wurden DNNs auch in anderen Bereichen neben der Bildklassifizierung [1] in Gesichtserkennung [2], Spracherkennung [3] und der Inhaltsbasierten Musikempfehlung [4] 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 Unüberwachten 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 unbekannten Umgebung, die es zu erforschen gilt. Gewünschtes Verhalten wird belohnt, wodurch es lernt die richtigen Entscheidungen zu treffen [5].

Vor allem in den letzten Jahren hat sich das Convolutional Neuronale Netzwerk (CNN) als das erfolgsversprechendste 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 angeordnet sind. Durch diese Anordnung ist es möglich größere Inputdaten in der selben

Geschwindigkeit zu verarbeiten wie zuvor [6]. 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 erstreckt sich über die selbe Tiefe der Eingabematrix. Durch die Iteration über jeden Punkt in der Eingabematrix erstellt die Faltungsschicht eine zweidimensionale Aktivierungskarte. Anhand dieser erkennt die Schicht dann gewünschte Merkmale wieder [6].

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 Goodfellow [7](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 Cross-Correlation wird aber oft auch als Faltung bezeichnet [7](328)

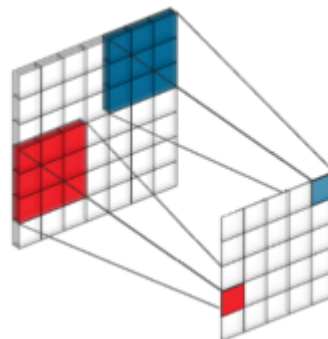


Abbildung 1: Faltung eine $7 \times 7 \times 3$ Matrix mit einem $3 \times 3 \times 3$ Filter und erzeugter Aktivierungskarte [8]

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 [6]. Sie ersetzt die Ausgabe eines Netzes an einem bestimmten Punkt durch eine statistische Zusammenfassung der nahegelegenen Ausgänge. Zum Beispiel Max Pooling [9] Übergabe der größte Zahl in einem rechteckigen Umfeld, Durchschnittsberechnung des Umfeldes oder ein gewichteter Durchschnitt basierend auf die Entfernung eines zentralen Punktes [7](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 [6]. Die vollständig verbundenen Schicht wird als Ausgangsschicht 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 UND AUSBLICK

LITERATUR

[1] 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.

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

[3] 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.

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

[5] 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.

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

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

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

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