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

2. Bestehende Ansätze zur Problemlösung

2.1 Inhaltsbasierter Filter

2.2 Kontextbasierter Filter

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 unbekannter Umgebung, die es zu erforschen gilt. Gewünschtes Verhalten wird belohnt, wodurch es lernt die richtigen Entscheidungen zu treffen [5].
Vorallem in den letzen Jahren hat sich das Convolutional Neuronales 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 Inhaltbasierten Musikempfehlung miteinander verglichen.

3.1 Aufbau eines Convolutional Neuronalen Netzes

Im Unterschied zu reglären DNN verwedet das CNN Neuronen, die drei Dimensionale angeordnete 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 erstreck sich über die selbe Tiefe der Ein-

gangsmatrix. 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 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 [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 bezeichent [7](328)

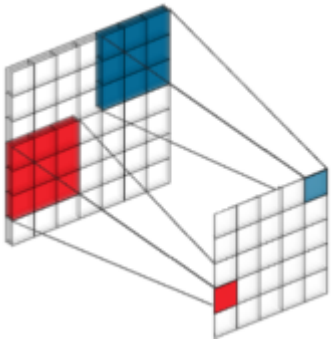

Abbildung 1: Faltung eine 7x7x3 Matrix mit einem 3x3x3 Filter und erzeugter Aktivierungskarte [8]

Verbindungsschicht

Üblicherweise wird eine Verbindungsschicht zwischen zwei Faltungschichten eingefügt. Seine Funktionbesteht darin, schrittweise die Größe der Darstellung zu reduzieren, um die Anzahl der Parameter und dadurch die Berechnung des gesammt Netwerkes zu verringern [6]. Sie ersetzt die Ausgabe eines Netzes an einem bestimmten Punkt durch eine statistische Zusammenfassung der nahegelegenen Ausgängen. Zum Beispiel Max Pooling [9] übergibt die größte Zahl in einem rechteckigen Umfeld, Durchschnittsberechnung einer rechteckigen Nachbarschaft oder ein gewichteter Durschnitt basiered 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 Matrixmultipliklation und einem Bias-Offset berechnet [6]. 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 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] Xinixi 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.