Künstliche Neuronale Netze in computerbasierten Musikinterfaces

Aristotelis Hadjakos

Zentrum für Musik- und Filminformatik, Hochschule für Musik Detmold

1 Einleitung

Künstliche Neuronale Netze weisen, nach einer längeren Phase der Stagnation, in den letzten Jahren beachtliche Erfolge vor. Sie übertreffen konkurrierende Verfahren in der Bild- und Spracherkennung, liefern vielversprechende Resultate beim Verarbeiten von Texten (Übersetzen, Themenentdeckung, Sentimentanalyse) und werden in verschiedensten Bereichen in Wissenschaft und Wirtschaft eingesetzt (LeCun, Bengio & Hinton 2015). Diese Erfolge wurden nicht zuletzt durch algorithmische und architektonische Fortschritte ermöglicht, z.B. durch Rectified Linear Units (ReLU), Convolutional Networks, Residual Networks, Gated Recurrent Units (GRU), etc. Das WaveNet von van den Oord et al. (2016) ist ein tiefes Neuronales Netz, das aus einer Vielzahl von Faltungsschichten (convolutional layers) besteht und Audiosignale auf Samplebasis rekonstruiert. Es sagt dabei, gegeben die vorherigen Samples, das jeweils nächste Sample vorher. Je nach Trainingsdaten kann WaveNet Sprache oder auch Musik synthetisieren. Engel et al. (2017) haben mit einer Architektur, die WaveNet um einen Autoencoder erweitert, den neuronalen Synthesizer NSynth implementiert. Dabei wird NSynth mit Klängen verschiedener Instrumente trainiert und erlaubt dem Nutzer dann, zwischen den Instrumenten zu interpolieren. Dabei gelingt es NSynth eine Interpolation zu finden, die linear klingt („sounds linear“), wie es von Hoskinson, van den Doel & Fels (2003) auf der „Conference on New Interfaces for Musical Expression“ (NIME) gefordert wurde. Als Basis für weitere Forschungen folgt ein grober Überblick über bisherige, auf der NIME publizierte Arbeiten zu Neuronalen Netzen.

2 Erkennung & Mapping

Eine ganze Reihe von Arbeiten setzt Neuronale Netze zur Mustererkennung ein, z.B. für die Erkennung von Vokalen anhand von Bildern der Mundform (Poepel, Feitsch, Strobel & Geiger 2014), für die Erkennung von unterschiedlichen Trommelschlägen mit der Hand

Veröffentlicht durch die Gesellschaft für Informatik e. V. 2017 in
M. Burghardt, R. Wimmer, C. Wolff, C. Womser-Hacker (Hrsg.):
Mensch und Computer 2017 – Workshopband, 10.–13. September 2017, Regensburg.
 https://doi.org/10.18420/muc2017-ws15-0371

anhand von Audiosignalen eines Kontaktmikrofons (Jathal & Park 2016) oder für die Erkennung von Dirigiergesten (Lee, Grüll, Kiel & Borchers 2006). Die so ermittelten objektiven Klassifikationen oder Regressionen werden dann anhand eines Mappings auf die Parameter der Klangsynthese abgebildet, wobei beim Mapping wiederum große gestalterische Freiheit möglich ist. Die eingesetzten Neuronalen Netze sind üblicherweise relativ einfache Feed-Forward-Netze mit wenigen, voll-verbundenen Schichten. Des weiteren können Neuronale Netze für die Vorverarbeitung von Sensordaten genutzt werden, z.B. um Bewegungen des Nutzers außerhalb des Tracking-Bereichs vorherzusagen (Wu, Zhou, Rau, Zhang & Wright, 2017).

Alternativ kann ein Mapping auch direkt durch ein Neuronales Netz ausgedrückt werden. Dabei entfällt der Zwischenschritt der objektiven Klassifikation bzw. Regression. So bilden Hoskinson, van den Doel & Fels (2003) Graustufenbilder mit 16 x 16 Pixeln auf die Parameter einer Modalsynthese ab. Perez & Bonada (2010) nutzen die Daten eines elektromagnetischen Motion-Tracking-Systems, um mit einem Rohr und einem Bogen eine Geige nachzubilden. Das Mapping auf die Parameter einer additiven Synthese wird durch Neuronale Netze durchgeführt. Trainiert werden die Netze mit Aufnahmen beim Spiel einer echten Geige, wobei Winkel, Geschwindigkeit und Beschleunigung des Bogens sowie die gespielte Saite vom Sensorsystem erfasst wird. Die Ausgabe der additiven Synthese wird mit einer Impulsantwort gefaltet, die den Effekt eines Geigenkörpers nachbildet.

Oftmals ist das Mapping aber nicht durch das Beispiel eines traditionellen Instruments vorgegeben, sondern frei. Zudem kann es gewünscht sein, das Mapping improvisatorisch während einer weiterlaufenden Performance zu ändern. Frameworks für die musikalische Nutzung von Machine Learning, wie der Wekinator von Fiebrink, Trueman & Cook (2009), greifen solche Anforderungen auf. Auch für die graphischen Sprachen Pure Data und Max/MSP gibt es Frameworks mit denen Neuronale Netze eingebunden werden können (Steiner 2006; Cont, Coduys & Henry 2004; Bevilacqua, Müller & Schnell 2005). Alternativ bietet sich immer an, die Echtzeit-Daten per UDP/OSC an Machine-Learning-Umgebungen wie Tensorflow oder Theano zu senden und die Ergebnisse auf entsprechendem Weg wieder zurück.

3 Architekturen

Neben den einfachen Feed-Forward-Netzen finden sich zwei Netzwerkarchitekturen vermehrt in den Tagungsbänden der NIME: Time Delay Neural Networks (TDNN) und Echo State Networks (ESN). Bei einem TDNN handelt es sich um ein Feed-Forward-Netz, das in weiten Teilen mit einem Convolutional Network vergleichbar ist. Die Eingabe umfasst aktuelle und alte Werte. Ein Neuron der ersten Schicht bezieht sich dabei jeweils auf einen begrenzten Zeitbereich, wobei sich die „zeitlich benachbarten“ Neuronen dieselben Gewichte teilen. Mit TDNNs lassen sich z.B. vorgegebene Handbewegungen voneinander unterscheiden (Modler, Myatt & Saup 2003; Modler, Myatt & Saup 2008).

Echo State Networks (ESN) sind eine Art von Recurrent Neural Network (RNN). Dabei werden die rekurrierten Verbindungen mit Zufallszahlen initialisiert, die so gewählt sind, dass sich der Betrag der Aktivierungen ungefähr gleich groß bleibt. Nur die letzte lineare Schicht wird trainiert. Die verborgenen Schichten fungieren dabei als neuronale Oszillatoren. Daher eignen sich ESNs besonders für die Synthese von Signalen mit zeitlich repetitiver Struktur. ESNs wurden u.a. eingesetzt, um aus Messdaten eines elektrisch leitenden Schaumstoffs musikalisch nutzbare Controller-Daten zu erzeugen (Kiefer 2010). Darüber hinaus gibt es für Pure Data Unterstützung für ESNs (Kiefer 2014).

4 Neue Richtungen

Fried & Fiebrink (2013) erstellen cross-modale Mappings, z.B. zwischen Bildern und Klängen, indem sie zwei Autoencoder aufeinander abbilden. Ein Autoencoder ist ein Neuronales Netz, das mittels eins unüberwachten Verfahrens lernt, die Daten möglichst verlustfrei in einen niedrigdimensionalen Raum zu projizieren. Die inneren Knoten des Netzes erlernen dabei abstrakte Eigenschaften zu erkennen, wie z.B. das Vorhandensein von Kanten oder anderen Formen in einem Bild. Die Abbildung der inneren Knoten erfolgt bei Fried & Fiebrink (2013) entweder direkt durch eine explizite Abbildungsvorschrift oder durch ein drittes Neuronales Netzwerk. Dabei wird diese Abbildung beliebig gewählt. Bemerkenswerterweise bleibt das Mapping, z.B. zwischen Bewegungen mit der Maus und einer FM Synthese, trotzdem nachvollziehbar (https://youtu.be/CkdaZFl5sLY).

Le Groux, Manzolli & Verschure (2007) nutzen ein bereits trainiertes Neuronales Netz, um die Bewegungen eines Avatars bildbasiert zu erfassen und damit Klänge zu steuern. Analoge Ansätze, bei denen bereits trainierte Netze genutzt werden, könnten aufgrund ihrer Verfügbarkeit für NIME-Entwickler heute von besonderem Interesse sein.

Nur relativ wenige Arbeiten nutzen die generativen Möglichkeiten Neuronaler Netze aus. Vogl & Knees (2017) modellieren den Beat von elektronischer Tanzmusik mittels einer Restricted Boltzmann Machine, einem Neuronalen Netz das unüberwacht trainiert wird. Nach dem Training werden neue Beat Tracks per Gibbs-Sampling erzeugt.

Schließlich gibt es auch noch Arbeiten wie das Artificial Analog Neural Network von Stearns (2009) oder das „Fragmented Orchestra“ (Jones., Hodgson, Grant, Matthias, Outram & Ryan 2009) bei denen Neuronale Netze weniger aufgrund ihrer technischen Möglichkeiten, sondern vielmehr als künstlerische Idee eingesetzt werden.

Literaturverzeichnis

Bevilacqua, F., Müller, R., Schnell, N. (2005): MnM: a Max/MSP mapping toolbox, NIME, S. 85-88.

Cont, A., Coduys, T., Henry, C. (2004): Real-time Gesture Mapping in Pd Environment using Neural Networks, NIME, S. 39-42.

Engel, J., Resnick, C., Roberts, A., Dieleman, S., Eck, D., Simonyan, K., & Norouzi, M. (2017). Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. arXiv preprint arXiv:1704.01279.

Fiebrink, R., Trueman, D., Cook, P. R. (2009): A Meta-Instrument for Interactive, On-the-fly Machine Learning, NIME, S. 280-285.

Fried, O., Fiebrink, R. (2013): Cross-modal Sound Mapping Using Deep Learning, NIME, S. 531-534.

Hoskinson, R., van den Doel, K., Fels, S. (2003): Real-time Adaptive Control of Modal Synthesis, NIME, S. 99-103.

Jathal, K., Park, T.-H. (2016): The HandSolo: A Hand Drum Controller for Natural Rhythm Entry and Production, NIME, S. 218-223.

Jones., D., Hodgson, T., Grant, J., Matthias, J., Outram, N., Ryan, N. (2009): The Fragmented Orchestra, NIME, S. 297-302.

Kiefer, C. (2010): A Malleable Interface for Sonic Exploration, NIME, S. 291-296.

Kiefer, C. (2014): Musical Instrument Mapping Design with Echo State Networks, NIME, S. 293-298.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

Lee, E., Grüll, I., Kiel, H., Borchers, J. (2006): conga: A Framework for Adaptive Conducting Gesture Analysis, NIME, S. 260-265.

Le Groux, S., Manzolli, J., Verschure, P. F.M. J. (2007): VR-RoBoser: Real-Time Adaptive Sonification of Virtual Environments Based on Avatar Behavior, NIME, S. 371-374.

Modler, P., Myatt, T., Saup, M. (2003): An Experimental Set of Hand Gestures for Expressive Control of Musical Parameters in Realtime, NIME, S. 146-150.

Modler, P., Myatt, T. (2008): Video Based Recognition of Hand Gestures by Neural Networks for the Control of Sound and Music, NIME, S. 358-359.

Perez, A., Bonada, J. (2010): The Bowed Tube: a Virtual Violin, NIME, S. 229-232.

Poepel, C., Feitsch, J., Strobel, M., Geiger, C. (2014): Design and Evaluation of a Gesture Controlled Singing Voice Installation, NIME, S. 359-362.

Steiner, H.-C. (2006): Towards a catalog and software library of mapping methods, NIME, S. 106-109.

Stearns, P. (2009): AANN: Artificial Analog Neural Network, NIME, S. 341.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. CoRR abs/1609.03499.

Vogl, R., Knees, P. (2017): An Intelligent Drum Machine for Electronic Dance Music Production and Performance, NIME, S. 251-255.

Wu, J. C., Zhou, Y., Rau, M., Zhang, Y., Wright, M. J. (2017): Towards Robust Tracking with an Unreliable Motion Sensor Using Machine Learning, NIME, S. 42-47.