Deep Learning für visuelle Medien: Annotation, Training, Analyse

Howanitz, Gernot
gernot.howanitz@uni-passau.de
Universität Passau, Deutschland

Radisch, Erik
e.radisch@gmx.at
Sächsische Akademie der Wissenschaften zu Leipzig

Workshop-Konzept

Zahlreiche Projekte der eHumanities fokussieren auf die Verarbeitung von Information, die in Textform codiert sind. Andere Modalitäten der Informationsübermittlung und -übertragung, wie beispielsweise visuelle Medien, bleiben in den DH häufig außen vor. Ein Grund dafür ist das breite Spektrum an etablierten Verfahren, das für solche Fragestellungen zur Verfügung steht. Ist man bei der Analyse von Texten in der Zwischenzeit so weit, sich den Inhalten und Kontexten durch automatisierte computergestützte Analyseverfahren zu nähern, verweilt man bei anderen Modalitäten wie Bildern allzu oft auf einer Ebene, wo verglichen mit der Textanalyse eher Buchstaben gezählt werden. Aus der Perspektive der Kulturwissenschaften ergibt sich hier ein *desideratum*; schließlich widmen sich diese der (menschlichen) Kultur in ihrer ganzen Bandbreite und decken kulturelle Äußerungen im weitesten Sinne ab, die unterschiedlichste Modalitäten, wie beispielsweise physische Artefakte und performative Handlungen, mit einschließen. Zwar ist es eingeschränkt möglich, kulturelle Phänomene zu transkribieren, also in textuelle Form zu bringen, was aber kaum automatisierbar ist und Informationsverluste birgt. Native Ansätze, welche auf jeweils spezifische Eigenschaften der zu untersuchenden Modalität eingehen, erscheinen deshalb vielversprechend. Neueste Ansätze der Computer Vision bieten hier ein großes Potential, denn sie ermöglichen es, sich auch Inhalten jenseits des Textes automatisiert zu nähern, was Möglichkeiten für computergestützte multimodale Analysen eröffnet.

In jüngster Zeit halten nun Methoden der Computer Vision langsam Einzug in die Digital Humanities (Tilton/ Arnold 2018). Dabei wird allerdings das volle Potential des Deep Learning nicht ausgeschöpft. Zwar finden Neuronale Netze zur Bilderkennung Anwendung in den Digital Humanities, diese beschränken sich aber oft auf die Nutzung vortrainierter Netze. Hier ergeben sich potentielle Probleme, werden diese Netze doch in der Regel auf einige wenige etablierte Bildkorpora wie etwa Microsofts COCO-Dataset (Common Objects in Context, http://cocodataset.org) trainiert. Dabei handelt es sich um Bildmaterial, das vorwiegend aus dem Nordamerika des 21. Jahrhunderts stammt. Für viele kulturwissenschaftliche Fragestellungen, die auf andere Zeiträume und/oder andere Kulturkreise abzielen, ergibt sich daraus ein Bias, der die Ergebnisse verfälschen kann. In solchen Fällen ist dann ein selbst durchgeführtes, zielgerichtetes Training notwendig, das auf die spezifische Fragestellung abgestimmt ist. Ein solches Training neuronaler Netze ist jedoch keineswegs trivial, sondern erfordert eine Menge Vorarbeit, Wissen um grundlegende Trainingsstrategien und vor allem auch Erfahrung im Tweaken der Parameter.

Der Workshop soll grundlegende Kenntnisse zur Anwendung von State-of-the-Art-Algorithmen der Computer Vision in den Digital Humanities vermitteln. Er baut auf den Erfahrungen auf, die die beiden Workshopleiter im Rahmen ihrer Tätigkeit am Passau Center for eHumanities (PACE) sammeln konnten. Die Ergebnisse wurden auch in den letzten zwei DHd-Konferenzen in Vorträgen vorgestellt (Bermeitinger et al. 2017; Decker et al. 2018). In der dreijährigen Projektlaufzeit wurde ein reicher Schatz an Erfahrungen im Umgang mit Neuronalen Netzen gesammelt und eine Reihe von einfachen und klar strukturierten Workflows entwickelt, die nun mit einer interessierten Öffentlichkeit geteilt werden sollen. Eine Hoffnung ist, dass der Workshop Anstoß für Projekte gibt, die visuelle Medien quantitativ erfassen wollen und gleichzeitig die vorgestellten Methoden einer kritischen Evaluation unterziehen und weiterentwickeln.

Der Workshop *Deep Learning für visuelle Medien* intendiert in mehrere der in PACE erarbeiteten Workflows einzuführen , die es erlauben, Neuronale Netze für visuelle Medien auf bestimmte Fragestellungen der Geisteswissenschaften anzuwenden, für eigene Fragestellungen zu adaptieren bzw. zu trainieren und die Ergebnisse zu analysieren. Damit soll die Grundlage gelegt werden, Forschern selbst das Training und die Anwendung Neuronaler Netze sowie die Analyse deren Ergebnisse zu ermöglichen. Im Zentrum des Workshops stehen drei Neuronale Netze, die über verschiedene Features verfügen.

Das von Facebook Artificial Intelligence Research Group (FAIR) entwickelte Framework *Detectron* (Girshick et al. 2018) kombiniert verschiedene neuronale Netze und ermöglicht ein breites Nutzungsspektrum. Dieses leistungsstarke Framework erlaubt nicht nur das Trainieren der Objekterkennung, sondern kann ebenfalls eine Reihe wichtiger Keypoints des menschlichen Körpers (z.B.: Kopf, Schultern, Ellenbogen, Knie, usw.) erkennen, die wiederum wichtige Rückschlüsse auf die Haltung der Personen zulassen. OpenPose (Zhe et al. 2017), das ebenfalls im Rahmen des Workshops vorgestellt wird, befasst sich ebenfalls mit diesen Keypoints. Im Gegensatz zu Detectron kann OpenPose auch einzelne Finger erkennen. Anders ausgedrückt liefert dieses Netz deutlich mehr Informationen zurück. Das dritte Neuronale Netz, auf das eingegangen werden wird, ist OpenFace von Tadas

Baltrusaitiš (Baltrusaitiš et al. 2018). Dieses mächtige Neuronale Netz kann nicht nur Gesichter erkennen, sondern auch deren dreidimensionale Ausrichtung errechnen und eine ganze Reihe von Keypoints im Gesicht erkennen. Diese Keypoints lassen ebenfalls Rückschlüsse auf sogenannte Facial Expression Units (Decker et al. 2019), welche genutzt werden können, um Aussagen über die Emotionen machen zu können, die eine Person zeigt.

Im Rahmen des Workshops werden sowohl Installation als auch Setup und erste Schritte mit diesen Frameworks thematisiert. Darüber hinaus geht es im Rahmen dieses Workshops auch darum, Netze zielgerichtet für die eigene Fragestellung zu trainieren. Wie ein Netz erfolgreich mit verhältnismäßig kleinen Korpora trainiert werden kann, wird im Kurs vermittelt werden. Auch die Evaluierung von Trainingsergebnissen wird diskutiert. In einem letzten Schritt soll auch in die Arbeit mit den extrahierten Features eingeführt und verschiedene Analysemöglichkeiten vermittelt werden.

Programm

Vor dem Workshop:
Vernetzung der Teilnehmerinnen und Teilnehmer über Github (https://github.com/passau-centre-for-ehumanities/visual_media), Identifizierung von gemeinsamen Forschungsinteressen, Gruppenbildung, falls notwending Hilfestellung zur Installation grundlegender Software (Jupyter Notebooks), um beim Workshop selbst möglichst wenig Zeit zu verlieren.

9:00-9:30 *Kick-Off und Kennenlernrunde, Abfragen der Erwartungen*

9:30-10:30 *Allgemeine Einführung in Deep Learning*
Zum Auftakt des Workshops wird eine allgemeine kurz gehaltene Einführung zu künstlichen neuronalen Netzen und Deep Learning-Algorithmen im Allgemeinen gegeben, um ein Verständnis der Funktionsweise von Detectron zu entwickeln.

10:30-11:00 *Kaffeepause*

11:00-12:30 *Einführung in Detectron und OpenPose, Praktische Erfahrungen*
Im Anschluss daran wird der generelle Aufbau von Detectron in die Funktionsweise der wichtigsten Bestandteile des Frameworks vorgestellt. Neben dem Trainingsaufbau wird hier aufgezeigt, wie man Standardmodelle lädt und auf visuelle Medien anwenden kann. Insbesondere geht es darum, Detectron und OpenPose zu nutzen, um Personen und deren Haltung in Bildern erkennen (Abb. 1)

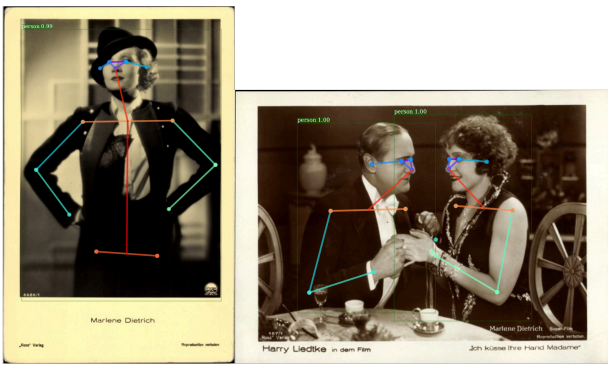

Abbildung 1: Posenerkennung mit Detectron

Als drittes Standbein soll im Kurs des Weiteren in die Anwendung von OpenFace eingeführt werden, mit dessen Hilfe es möglich ist, Keypoints von Gesichtern auszulesen, die dafür verwendet werden, um sogenannte Action Units abschätzen zu können. Action Units sind Basisbestandteile von menschlichen Emotionsausdrücken (Abb. 2)

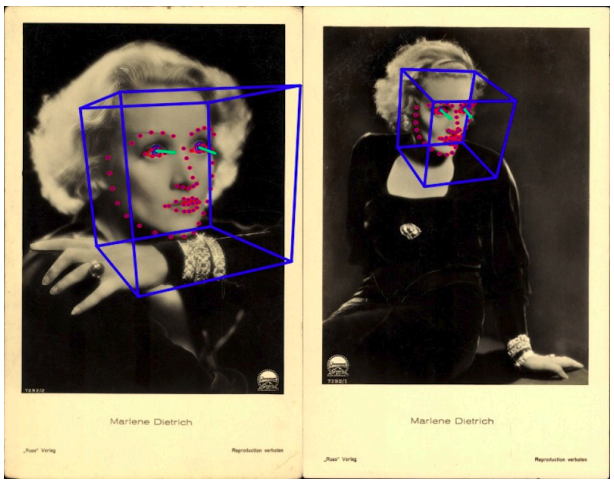

Abbildung 2: OpenFace blickt Marlene Dietrich ins Gesicht: Lage im Raum (blau), Keypoints (rot), berechnete Blickrichtung (grün)

12:30-14:00 *Mittagspause*

14:00-16:00 *Detectron Trainieren*
Ein wichtiger Bestandteil des Workshops wird es sein, in ein im Rahmen des Passau Center for eHumanities entwickelten Workflow zum Trainieren von Detectron einzuführen.

Es wird konkret an einfachen Beispielen vermittelt, wie man die einzelnen Bestandteile des Workflows installieren und auf die individuelle Forschungsfrage hin anwenden kann. Es wird neben der Vermittlung des Workflows ebenfalls großen Wert darauf gelegt, den Kursteilnehmern zu vermitteln, welche typischen Fehler beim Trainingsaufbau zu vermeiden sind. Die Teilnehmer sollen am Ende des Workshops dazu in der Lage sein:

- selbstständig Trainingskorpora für Detectron erzeugen zu können
- ein grundlegendes Verständnis dafür entwickelt haben, auf welche Parameter zu achten sind, um auch mit kleinen Bildkorpora trainieren zu können
- den Trainingsprozess mit den eigenen Daten initiiren zu können

Abbildung 3: Beispiel der Annotation. Als Werkzeug wird Labelme genutzt (https://github.com/wkentaro/labelme)

Abbildung 4: Beispiel der Anwendung von Detectron, trainiert auf ukrainische Symbole

16:00-16:30 *Kaffeepause*

16:30-18:00 *Vorstellung von Analysetechniken für die produzierten Ergebnisse*

Im letzten Teil des Kurses werden Analysetechniken vorgestellt, die es ermöglichen, die Produzierten daten nach interessanten Mustern zu explorieren bzw. deren Inhalte zu analysieren (Abb. 5, 6).

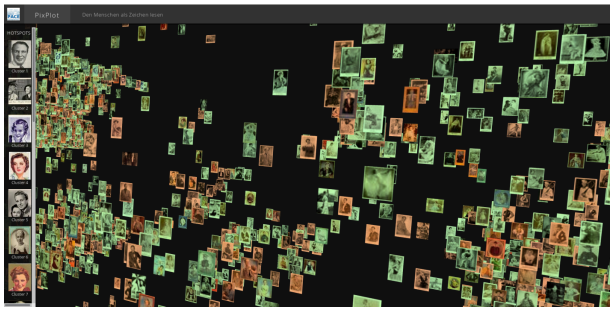

Abbildung 5: Beispiel einer skalierbaren, dreidimensionalen Visualisierung eines Clusterings mittels einer modifizierten Version von Pixplot. Die Bilder werden anhand von Metadateninformationen zusätzlich eingefärbt.

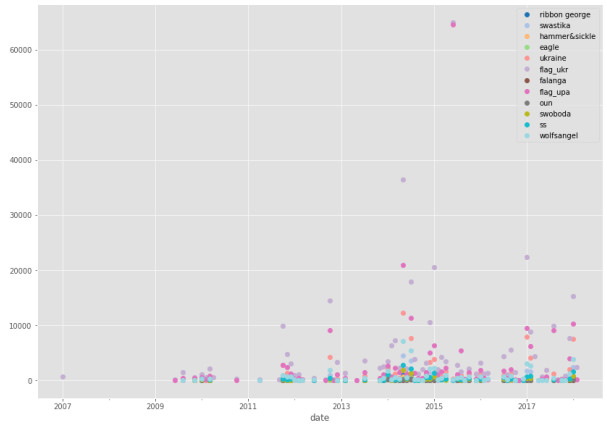

Abbildung 6: Beispiel für eine Analyse der Ergebnisse. Hier konkret: die Verteilung von Symbolen in Youtube-Videos über die Zeit.

18:00-18:30 *Schlussrunde, Workshop-Evaluation*

Dieses Programm ist hoffentlich geeignet, als Inspiration und erste Einführung in das Thema Deep Learning für visuelle Medien zu dienen. Eine Nachbereitung und weitere Vernetzung über Github ist ausdrücklich erwünscht, um eine weitere Begleitung der Projekte zu garantieren.

Zusätzliche Angaben

- Benötigte technische Ausstattung: Beamer, WLAN-Zugang, ausreichend Steckdosen für die Laptops der Teilnehmerinnen und Teilnehmer
- Zahl der möglichen Teilnehmer: 20

Bibliographie

Bermeitinger, B. / Howanitz, G. / Radisch, E. (2018): Contextualizing Bandera: Ein Distant Watching Ansatz,

in Kritik der digitalen Vernunft Konferenzabstracts. Köln, 26.02-02.03.2018. http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf [abgerufen am 01.05.2018].

Baltrusaitiš, T. / Zadeh, A. / Chong Lim, Y., Morency, L.-P. (2018): "OpenFace 2.0: Facial Behavior Analysis Toolkit," 2018 13th IEEE International Conference on. IEEE, 2018, S. 59-66.

Decker, J.-O. / Howanitz, G. / Radisch, E. / Rehbein, M. (2019): Den Menschen als Zeichen lesen. Quantitative Lesarten körperlicher Zeichenhaftigkeit in visuellen Medien. In: Dhd 2019, Digital Humanities: multimodal & multimedial, Konferenzabstracts, Frankfurt am Main 2019, S. 106-109. https://zenodo.org/record/2596095#.XLBUNUNS-V4

Girshick, R. / Radosavovic, I. / Gkioxari, G. / Dollár, P. / He, K. (2018): Detectron. https://github.com/facebookresearch/detectron . [Letzter Zugriff 25. 09. 2018]

Zhe C. / Tomas S. / Shih-En W. / Yaser Sh. (2017): Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In: CVPR.