Critical Machine Vision. Eine Perspektive für die Digital Humanities

Bell, Peter
peter.bell@fau.de
FAU Erlangen-Nürnberg, Deutschland

Offert, Fabian
fabian@zentralwerkstatt.org
FAU Erlangen-Nürnberg, Deutschland

Wir fragen nach neuen Spielräumen der Digital Humanities im Feld des maschinellen Lernens. Dazu dekonstruieren wir etablierte Computer-Vision-Modelle mit Methoden der Bildwissenschaft/Visual Studies.

Computer Vision, also der visuelle Zweig künstlicher Intelligenz, spielt eine immer wichtigere Rolle in Wirtschaft (z.B. Industrie 4.0, autonomes Fahren), Sozialem (Überwachung, Medizin) und Wissenschaft (vorrangig in den Natur-, Ingenieur-, und Lebenswissenschaften). Auch wenn erste Prototypen in den Digital Humanities, zum Beispiel in der digitalen Kunstgeschichte (Bell / Impett, 2019) und in den A/V-orientierten Digital Humanities (Arnold / Tilton, 2019a), entwickelt werden, wird die rasante Entwicklung des visuellen maschinellen Lernens in den Geisteswissenschaften noch relativ wenig reflektiert. Dies liegt teilweise daran, dass dessen Erforschung und kritische Reflektion eine fundierte Kenntnis der technischen Prozesse erfordert.

In besonderen Maße gilt dies für die Interpretierbarkeit von Computer-Vision-Modellen aus dem Bereich des Deep Learning, also des maschinellen Sehens mit komplexen neuronalen Netzwerken wie Convolutional Neural Networks (LeCun et. al. 1989, Krizhevsky et. al, 2012). Obwohl Interpretation als Eckpfeiler humanistischer Methoden und Theoriemodelle gilt, und obwohl Interpretable Machine Learning gegenwärtig in den Technikwissenschaften mit großer Aufmerksamkeit bedacht wird (Lipton, 2016, Selbst / Barocas, 2018, Mittelstadt et. al., 2019, Doshi-Velez und Kim, 2017, Gilpin et. al. 2018), ist das Problem der Interpretation, also der sinnschaffenden Analyse und Kritik von Computer-Vision-Modellen und Arbeitsabläufen weder in den Geisteswissenschaften noch in den Digital Humanities ausführlicher gewürdigt worden. Erste Ansätze finden sich z.B. in (Underwood, 2019) oder (Arnold / Tilton, 2019b). Dies ist insofern überraschend, als die Interpretation von Computer-Vision-Modellen eine Reihe von Fragen aufwirft, die Strukturähnlichkeiten zu Problemen in den Geisteswissenschaften im Allgemeinen, und in den Bildwissenschaften im Besonderen aufweisen. Dazu gehören das Problem der visuellen Mehrdeutigkeit, das epistemologische Problem der Verortung von Wissen, und das Problem des Verhältnisses von Form und Bedeutung.

Obwohl diese Aspekte sich im Bereich der Computer Vision als technische Probleme mit technischen Lösungsansätzen manifestieren (Olah et. al. 2017, 2018, Hohman et. al. 2018), bleibt ihre kritische Sprengkraft erhalten und erfordert eine nicht-technische Aufarbeitung. Beispielhaft ist hier die Andersartigkeit der maschinellen Wahrnehmung mit Convolutional Neural Networks zu nennen, die nachweisbar sehr viel mehr auf das Erkennen von Oberflächenbeschaffenheit aufbaut als auf das Erkennen von Formen (Geirhos et. al., 2019), und generell mit kaum wahrnehmbaren Bildbestandteilen operiert (Ilyas et. al., 2019). Wie beeinflusst diese andersartige Weltsicht die Aussagekraft von maschinellen Analysen in den Digital Humanities? Interpretierbarkeit könnte daher als ein grundsätzlich interdisziplinäres Problem angesehen werden, welches das Potenzial hat, Anstrengungen in der Informatik und den Digital Humanities zu verbinden und zu festigen.

Unter dem Begriff “Critical Machine Vision” möchten wir in den Digital Humanities daher einen Bereich etablieren, in dem die Digital Humanities nicht nur digitale Methoden auf geisteswissenschaftliche Gegenstände anwenden, sondern umgekehrt die informatischen Werkzeuge mit Methoden der Digital Humanities und der Geisteswissenschaften analysieren. Critical Machine Vision stellt drei zentrale Fragen: (1) Was und wie wird von mit Hilfe von Computer Vision gelernt, (2) welche Stereotypen und Vorurteile werden in diesem Lernprozess affirmiert oder erzeugt, und (3) wie können diese Verzerrungen durch neuartige Formen von Bilddatensätzen und Annotationsmethoden gemindert werden, und so Ansätze aus dem Forschungsbereich Fairness, Accountability, and Transparency of Machine Learning, kurz FAT-ML, (vgl. Friedler et. al., 2019, Suresh / Guttag, 2019) für Bilddatensätze neu gedacht werden. Wir befassen uns insbesondere mit der kritischen Analyse der wichtigen Bilddatensätze ImageNet (Deng et. al., 2009) bzw. der ILSVRC2012-Auswahl von ImageNet (Russakovsky et. al., 2015) und COCO (Lin et. al., 2014), mit denen Convolutional Neural Networks trainiert und evaluiert werden.

ImageNet ist ein umfangreicher digitaler Bilddatensatz, der die automatische Klassifizierung von Bildern in Bezug auf die abgebildeten Objekte ermöglichen soll (Object Recognition). Er besteht aus über vierzehn Millionen Bildern in über 21.000 Kategorien. Wir konzentrieren uns in unserer Analyse weniger auf aus unserer Sicht eher unkritische Klassifizierungen (z. B. Hunderassen oder Fahrzeugtypen), sondern auf streitbare Zuschreibungen: die Kennzeichnung der Menschen, ihre Assoziation mit sozialen Gruppen und menschlichen Interaktionen. Diesem kritischen Bereich von ImageNet entsprechen ähnlichen Kategorien in der Bilddatenbank COCO (Common Objects in Context), die mit ihrem Fokus auf "Common Objects" den Alltag und dementsprechend auch viele

Menschen und menschliche Interaktionen einbezieht. Im Gegensatz zu ImageNet hat COCO weniger Kategorien, aber mehr Instanzen pro Kategorie. Auf diese Weise können detaillierte Objektmodelle erlernt werden, die eine präzise 2D-Lokalisierung ermöglichen. Am relevantesten für den vorgeschlagenen Beitrag ist jedoch die Tatsache, dass die alltäglichen Szenen und Objekte in COCO hauptsächlich aus westlichen, bürgerlichen Kontexten des 21. Jahrhunderts stammen, also nur einen begrenzten Ausschnitt von Welt bieten, der wiederum von einer ebenfalls nicht repräsentativen Gruppe von Menschen annotiert wurde.

Beide Bilddatensätze werden mit Methoden der Informatik und der Bildwissenschaft untersucht, aber eben ganz bewusst als Teil der Digital Humanities. Unser Beitrag liegt also nicht nur im neuartigen Ansatz der granularen, technisch fundierten, Dekonstruktion und konstruktiven Umgestaltung von digitalen Bilddatensätzen, sondern auch in der Transdisziplinarität unter dem Dach der Digital Humanities Computer Vision/Informatik und Bildwissenschaften zu verbinden. Wir verändern damit die Blickrichtung der Digital Humanities. Sahen wir bisher mit den Werkzeugen der Computer Vision auf geisteswissenschaftliche Gegenstände, schauen wir jetzt mit geisteswissenschaftlichen Werkzeugen auf die Methoden der Computer Vision. Dabei ist dieses Verhältnis allerdings mehrfach gebrochen, denn wir nutzen dabei wiederum digitale Werkzeuge wie z.B. Convolutional Neural Networks und Generative Adversarial Networks (Goodfellow et. al., 2014), oder Werkzeuge aus dem Bereich der Visual Analytics wie Summit (Hohman et. al., 2019) und schauen auf geistes- und sozialwissenschaftliche Gegenstände (z.B. Gender, Race, Habitus und Diskurs). Die Öffnung der Black Box ist somit ein Ergebnis der konsequenten gegenseitigen Ergänzung von geisteswissenschaftlich-kritischen Werkzeugen und der Nutzbarmachung experimenteller informatischer Werkzeuge aus dem Bereich des maschinellen Lernens.

Eine der großen Herausforderungen der Computer Vision ist die Vielfalt und Heterogenität der realen Bildwelt, die sich mit technischen Mitteln nur schwer erfassen lässt. Während sich Computer Vision in der Vergangenheit auf ausgefeilte algorithmische Ansätze zur Erkennung von Merkmalen in Bildern konzentrierte, gelang es der jetzigen Generation des maschinellen Lernens diese weit zu übertreffen, indem komplexe (d.h. „tiefe“) Convolutional Neural Networks verwendet wurden, die auf großen Bilddatensätzen trainiert wurden. Mit der Einführung solcher Datensätze in den Computer-Vision-Prozess entsteht jedoch ein für die Schnittstelle von Computer Vision und maschinellem Lernen spezifisches Problem: Wie lässt sich die Vielfalt und Heterogenität der realen visuellen Welt in einer Reihe von Bildern – begrenzter Größe – darstellen? Historisch gesehen hängt dieses Problem mit dem allgemeinen erkenntnistheoretischen Problem zusammen, Taxonomien des Bestehenden zu erschaffen. Symbolische Taxonomien und Kategorien dienen hier der Ordnung konkreter Repräsentanten in Form von annotierten Fotografien. Unsere kritische Analyse setzt somit an sämtlichen Punkten des Prozesses an: die Ordnung der Taxonomien, die Auswahl und Art der Abbildungen, der Vorgang der Annotation bis hin zu den algorithmischen Details des Trainings.

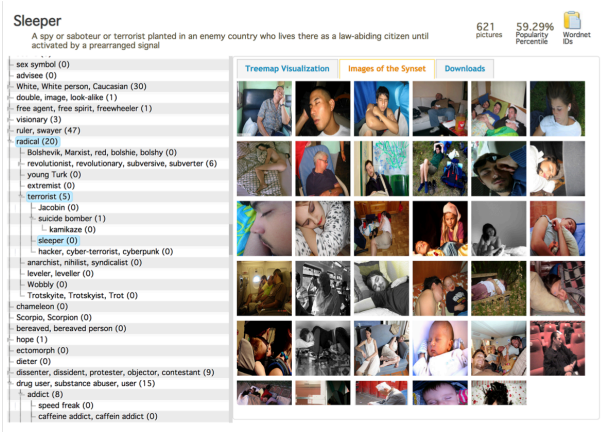

Abbildung 1: Die Kategorie “terrorist” in ImageNet enthält die Unterkategorie “sleeper” im Sinne verdeckter Terroristen (links). ImageNet illustriert den Begriff mit Bildern schlafender Menschen.

In jedem Schritt zeigen sich dabei Verzerrungen aufgrund von subjektiven Einschätzungen und (hegemonialen) Diskursen, die der demographischen Struktur der Akteur*innen (Fotograph*innen, Datenkurator*innen und Annotierenden) geschuldet sind. Diese Vorurteile lassen sich direkt an den Trainingsdaten, Kategorien und Strukturen ablesen, beispielsweise durch die von den Annotierenden erstellten Bildbeschreibungen (Captions) oder die vorgegebenen Objektklassen in COCO. Die in Teilen ungewollt komischen Bildbeschreibungen und US-amerikanisch geprägten Kategorien müssen immer vor dem Hintergrund betrachtet werden, dass Sie zum Training, Validieren und Testen von KI-Systemen verwendet werden. Eine anhand derart belasteter Datensätze trainierte KI wird zu einer Gefahr in jenem Moment, in dem die spezifische, eingeschränkte, und vorurteilsbelastete “Weltsicht” der KI auf Situationen der realen visuellen Welt trifft, und sich Mängel in den Datensätzen in undurchsichtige (da vielfach mediierte) und potenziell diskriminierende Fehlentscheidungen übersetzen.

Uns geht es jedoch nicht ausschließlich um eine Kritik der bestehenden Computer-Vision-Methoden, sondern um die Entwicklung und Erprobung neuer Verfahren in denen existierende Vorurteile durch bildwissenschaftliche Beschreibungs- und Ordnungsmodelle reduziert werden. Dabei stellt sich auch die Frage, wie sich eine größere Diversität der Bilddaten und letztlich des künstlichen Sehens nicht nur über eine räumlich größere Diversität, sondern auch eine zeitliche Diversität erreichen lässt. Zu welchem Maß spielen historische Bildwelten, das kulturelle Erbe, eine Rolle für unsere gegenwärtige Wahrnehmung,

in welchem Maß muss sie Berücksichtigung finden? Welche Veränderungen ergeben sich durch die Annotation von Expert*innen oder eine bessere Vorbereitung der Crowdworker? Die Analyse und die methodischen Ansätze zur Veränderung von Modellen und Prozessen zeigen wir anhand von wenigen Fallbeispielen (wie Abb. 1).

Unser Beitrag untersucht diese Fragen mit den kombinierten Mitteln der Computer Vision und der Bildwissenschaft, mit dem Ziel, diesen interdisziplinären Ansatz als neue Forschungsrichtung innerhalb der Digital Humanities vorzuschlagen, damit die Digital Humanities als kritischen Partner der Informatik neu zu etablieren, und ihre Spielräume somit signifikant zu erweitern.

Bibliographie

Arnold, T. / Tilton, L. (2019a): Distant viewing: Analyzing large visual corpora. Digital Scholarship in the Humanities.

Arnold, T. / Tilton, L. (2019b): Depth in Deep Learning: Knowledgeable, Layered, and Impenetrable.

Bell, P. / Impett, L. (2019): Ikonographie und Interaktion. Computergestützte Analyse von Posen in Bildern der Heilsgeschichte. Das Mittelalter 24, 31–53.

Deng, J. / Dong, W. / Socher, R. / Li, L. / Li, K. / Fei-Fei, L. (2009): Imagenet: A large-scale hierarchical image database, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference, 248–255.

Doshi-Velez, F. / Kim, B. (2017): Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Friedler, S.A / Scheidegger, C. / Venkatasubramanian, S. / Choudhary, S., Hamilton, E.P. / Roth, D. (2019): A comparative study of fairness-enhancing interventions in machine learning, in: ACM Conference on Fairness, Accountability, and Transparency (FAT*).

Geirhos, R. / Rubisch, P. / Michaelis, C. / Bethge, M. / Wichmann, F.A. / Brendel, W. (2019): ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231.

Gilpin, L.H. / Bau, D. / Yuan, B.Z. / Bajwa, A. / Specter, M. / Kagal, L. (2018): Explaining explanations: An overview of interpretability of machine learning, in: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). IEEE: 80–89.

Goodfellow, I. / Pouget-Abadie, J. / Mirza, M. / Xu, B. / Warde-Farley, D. / Ozair, S. / Courville, A. / Bengio, Y. (2014): Generative adversarial nets, in: Advances in Neural Information Processing Systems: 2672–2680.

Hohman, F.M. / Kahng, M. / Pienta, R. / Chau, D.H. (2018): Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. IEEE Transactions on Visualization and Computer Graphics.

Hohman, F. / Park, H. / Robinson, C. / Chau, D.H. (2019): Summit: Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations. arXiv preprint arXiv:1904.02323.

Ilyas, A. / Santurkar, S. / Tsipras, D. / Engstrom, L. / Tran, B. / Madry, A. (2019): Adversarial examples are not bugs, they are features. arXiv preprint arXiv:1905.02175.

Krizhevsky, A. / Sutskever, I. / Hinton, G.E. (2012): Imagenet Classification with Deep Convolutional Neural Networks, in: Advances in Neural Information Processing Systems: 1097–1105.

LeCun, Y. / Boser, B. / Denker, J.S. / Henderson, D. / Howard, R.E. / Hubbard, W. / Jackel, L.D. (1989): Backpropagation applied to handwritten zip code recognition. Neural computation 1: 541–551.

Lin, T.-Y. / Maire, M. / Belongie, S. / Hays, J., Perona, P. / Ramanan, D. / Dollár, P. / Zitnick, C.L. (2014): Microsoft COCO: Common objects in context, in: European Conference on Computer Vision. Springer: 740–755.

Lipton, Z.C. (2016): The mythos of model interpretability, in: 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY.

Mittelstadt, B. / Russel, C. / Wachter, S. (2019): Explaining Explanations in AI, in: ACM Conference on Fairness, Accountability, and Transparency (FAT*).

Olah, C. / Mordvintsev, A. / Schubert, L. (2017): Feature visualization. Distill. https://doi.org/10.23915/distill.00007 [letzter Zugriff 26 September 2019]

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., Mordvintsev, A. (2018). The building blocks of interpretability. Distill. https://doi.org/10.23915/distill.00010 [letzter Zugriff 26 September 2019]

Russakovsky, O. / Deng, J. / Su, H. / Krause, J. / Satheesh, S. / Ma, S. / Huang, Z. / Karpathy, A. / Khosla, A. / Bernstein, M. et al. (2015): Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115: 211–252.

Selbst, A.D. / Barocas, S. (2018): The intuitive appeal of explainable machines. Fordham Law Review 87.

Suresh, H. / Guttag, J.V. (2019): A Framework for Understanding Unintended Consequences of Machine Learning. arXiv preprint arXiv:1901.10002.

Underwood, T. (2019): Distant Horizons: Digital Evidence and Literary Change. University of Chicago Press.