Grundzüge einer visuellen Stilometrie

Laubrock, Jochen
laubrock@uni-potsdam.de
Universität Potsdam, Deutschland

Dubray, David
ddubray@uni-potsdam.de
Universität Potsdam, Deutschland

Was kennzeichnet visuellen Stil? Nachdem die digitalen Geisteswissenschaften stark durch textanalytische Verfahren aus der Computerlinguistik und verwandten Gebieten geprägt waren, sind in den letzten Jahren vermehrt Methoden zur Beschreibung visuellen Materials vorgeschlagen worden. Diese sollten insbesondere den Bildwissenschaften neue methodische Zugänge ermöglichen. Die Frage nach einer formalen Beschreibung visuellen Stils etwa hat die kunstgeschichtlichen Forschung seit ihrem Beginn umgetrieben (Wölfflin 1915). In der Form- und Strukturanalyse dominieren jedoch verbale Beschreibungen, eine quantitative Lösung jenseits deskriptiver Ansätze steht aus. Neuere Entwicklungen im Bereich des maschinellen Sehens (*Computer Vision*) lassen nun eine formale Beschreibung visuellen Stils greifbar werden. Diese basiert auf Repräsentationen in den tieferen Schichten sogenannter Convolutional Neural Networks (CNNs).

CNNs sind eine Klasse tiefer neuronaler Netze, die inspiriert von biologischen visuellen Systemen entwickelt wurden, um ingenieurwissenschaftliche Probleme wie z.B. Handschrifterkennung zu lösen (LeCun et al. 1989). Durch nur lokale Konnektivität sind CNNs deutlich sparsamer und effizienter als klassische „ *fully connected*“ neuronale Netze. CNNs lassen sich beschreiben als eine hierarchische Anordnung computationaler Einheiten, die visuelle Information in einem *Feedforward*-Prozess verarbeiten. Jede Schicht der Hierarchie lässt sich interpretieren als eine Menge von Filtern, die bestimmte Merkmale des Eingabebildes extrahieren. Die Filterkoeffizienten werden durch Anpassung an die Daten gelernt. Die Ausgabe einer Schicht besteht aus einer Menge sogenannter Merkmalskarten, welche unterschiedlich gefilterte Versionen des Eingabebildes sind. Filter auf höheren Schichten erhalten als Eingabe im Wesentlichen eine gewichtete Rekombination der Merkmalskarten niedrigerer Schichten. In den unteren Schichten sind die Repräsentationen relativ einfach und entdecken beispielsweise Kanten oder Farben. Repräsentationen mittlerer Schichten können z.B. texturartig sein, während höhere Schichten deutlich komplexer sind und z.B. Objektteile repräsentieren können. Die unterschiedlichen Repräsentationsebenen weisen eine starke Ähnlichkeit mit der hierarchischen Verarbeitung im für Objekterkennung zuständigen ventralen Pfad des menschlichen visuellen Systems auf (Yamins and DiCarlo 2016), weshalb CNNs auch aussichtsreiche Kandidaten für die nähere und quantitativ fundierte Untersuchung ungelöster Probleme der sogenannten *Mid-Level Vision* sind. Auch in diesem Bereich dominierten bis vor kurzem qualitative Beschreibungen wie z.B. gestaltpsychologische Ansätze.

Neural Style Transfer. Wie kodieren CNNs nun Stil? Leon Gatys, Alexander Ecker und Matthias Bethge haben die Methode des Style Transfer entwickelt, in der sie zeigen, dass Stil und Inhalt eines Bildes in CNNs zu einem gewissen Grad unabhängig voneinander repräsentiert werden. Am Beispiel des VGG-Netzwerkes (Simonyan and Ziisserman 2014) demonstrieren Gatys et al., dass stilistische Elemente auf niedrigeren Schichten und Bildinhalte auf höheren Schichten des Netzwerks kodiert werden. Der Stil eines Bildes A lässt sich deshalb prinzipiell auf den Inhalt eines Bildes B übertragen. Neuere Arbeiten haben diesen *Style Transfer* weiter optimiert (Sanakoyeu et al. 2018). Interessanterweise ist Stil in diesen Arbeiten von gegenständlichen und auch abstrakten Gemälden extrahiert worden, obwohl das zugrundeliegende VGG für die Klassifikation von Fotos vortrainiert war. Die gelernten Filter sind offensichtlich hinreichend generisch, um derartigen Transfer zu ermöglichen.

Stil von Illustratoren. In Vorarbeiten haben wir gezeigt, dass sich CNN-Repräsentationen auch zur Beschreibung grafischer Literatur wie Comics, Graphic Novels etc. eignen (Laubrock, Hohenstein and Kümmerer 2018). *Welche Repräsentationen sind nun aber charakteristisch für den Stil von Illustratoren?* Dieser Frage gehen wir in der vorliegenden Untersuchung nach. Wir nutzen dazu das Xception-Netzwerk (Chollet 2016), das deutlich effizienter ist als VGG und bei weniger Parametern typischerweise eine bessere Klassifikationsleistung erbringt. Als „Signatur“ eines Zeichners extrahieren wir das Muster der mittleren Antwortstärke über verschiedene Filter. Diese benutzen wir als Prädiktor für eine Illustrator-Klassifikation. Experimentell variieren wir dabei, aus welchen Ebenen des Netzwerks Filter zu Klassifikation genutzt werden. Die Güte der Klassifikation als Funktion der benutzten Filter dient zur Abschätzung dafür, wie relevant auf einer bestimmten Hierarchieebene repräsentierte Merkmale für den Individualstil sind. Zusätzlich berechnen wir eine Ähnlichkeitsmatrix basierend auf den CNN-Aktivierungen als Grundlage für eine bildbasierte Suche.

Material

Als Material verwenden wir zwei Sammlungen grafischer Literatur: (a) das Graphic Narrative Corpus (GNC; Dunst, Hartel and Laubrock 2009) und (b) Manga109 (Matsui et al. 2017). Das GNC ist eine kuratierte Sammlung über 200 zeitgenössischer Graphic Novels aus den Jahren

1979 bis 2017 mit einem Gesamtumfang von mehr als 50.000 Seiten. Der GNC beinhaltet Werke verschiedener Genres (z.B. Autobiographie, New Journalism, Crime, Superhelden). Manga109 besteht aus 109 Manga-Bänden (mehr als 20.000 Seiten), die zwischen 1970 und 2010 im Handel erhältlich waren und 2017 der Wissenschaft zur Verfügung gestellt wurden. Die Korpora wurden durch zufällige geschichtete Stichprobenziehung in ein Trainings- und ein Testcorpus unterteilt.

Methode

Der CNN-Teil eines auf dem ImageNet-Datensatz (Deng et al. 2009) vortrainierten Xception-Netzwerk wurde benutzt, um Illustratoren in den beiden Korpora zu klassifizieren.

Bildsignatur. Für jedes Bild wurden zunächst Merkmalskarten auf verschiedenen ausgewählten Schichten des Xception-Netzes berechnet. Für die Merkmalskarten pro Filter wurde dann die mittlere Aktivierung (*global average pooling*) berechnet. Ein Vektor der mittleren Aktivierung über eine Menge von Filtern wurde als Signatur des Bildes gespeichert. Dies resultiert in einer recht kompakten Repräsentation mit einem Kompressionsverhältnis im Vergleich zum Ausgangsbild von ca. 1:800 für frühe bzw. ca. 1:100 für späte Schichten und 1:21 bei Verwendung aller Filter.

Klassifikation. Zur Klassifikation trainierten wir ein einfaches *fully-connected* neuronales Netz mit einer verdeckten Schicht von 1024 Einheiten. Diese erhielten als Input die Bildsignatur (*average-pooled feature maps*, s.o.) und als Output die Illustratoren. Das Trainings-Set enthielt 90% der Seiten eines jeden Buches, zufällig bestimmte 10% der Seiten pro Comicband wurden nicht während des Trainings präsentiert, sondern als Test-Set zur Seite gelegt zur Bewertung der Klassifikationsleistung des trainierten Netzes.

Läsionsexperimente. Weil wir uns dafür interessierten, welche Art von Merkmal am charakteristischsten für den Stil eines Illustrators ist, haben wir das CNN auf fortschreitend niedrigeren Ebenen läsioniert und die Klassifikationsleistung mit dem vollen, auf allen Schichten basierenden Modell verglichen.

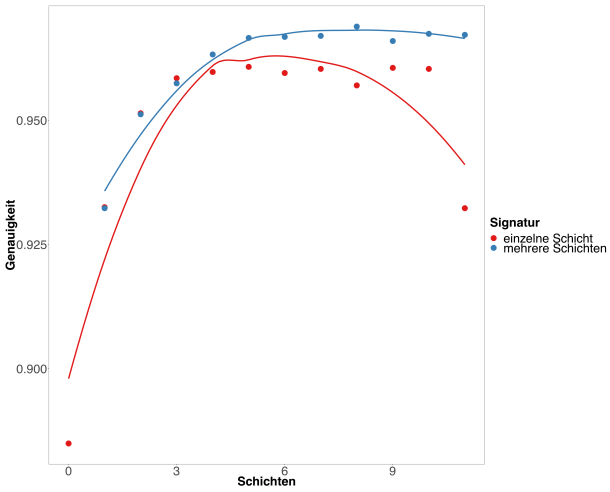
Zusätzlich haben wir Klassifikationen basierend auf dem Output einzelner Schichten berechnet. Insgesamt vergleichen wir also die Klassifikation unter Berücksichtigung einzelner Schichten 0, 1, ..., k vs. mit der bei Berücksichtigung aller Schichten von 0 bis k. Der Merkmalsvektor wurde im letzteren Fall durch einfache Verkettung der Signaturen gebildet.

Ähnlichkeitsmatrix. Die Ähnlichkeiten der Merkmalsvektoren aller Bilder wurden mittels euklidischer Distanz berechnet. Basierend auf dieser Matrix wurde eine Ähnlichkeitssuche implementiert.

Semantische Segmentierung. Mit Hilfe von CNN-Repräsentationen lassen sich auch sehr gut einzelne Bildelemente identifizieren. Zur Detektion von Sprechblasen trainieren wir ein Fully-Convolutional neuronales Netz nach der U-Net-Architektur (Ronneberger et al. 2015) auf 750 annotierte Comicseiten. In dieser Architektur wird neben einem Enkodier- auch ein Dekodierpfad benutzt, in dem die recht abstrakten, semantiknahen Repräsentationen höherer Ebenen mit Kopien der Information niedrigerer Ebenen verrechnet wird, um Ortsinformation zu rekonstruieren. Auch hier haben wir beim Enkodierpfad wieder mit vortrainierten Repräsentationen begonnen und nur ein Feintuning vorgenommen.

Ergebnisse

Abbildung 1 zeigt die Genauigkeit der Klassifikation als Funktion der zugrundeliegenden Merkmale. Insgesamt lassen sich die Seiten aufgrund rein visueller Analyse sehr gut ihren Urhebern zuordnen. Man erkennt am Abfall der Kurve für Merkmale aus einzelnen Schichten, dass für die Illustrator-Klassifikation die Repräsentationen mittlerer Ebenen am entscheidendsten sind. Die stilistische Signatur einer Graphic Novel basiert scheinbar eher auf Merkmalen mittlerer Komplexität wie Schraffuren, Texturen oder Schwüngen als auf höher integrierten Merkmalen wie Objektteilen oder spezifischen Motiven.

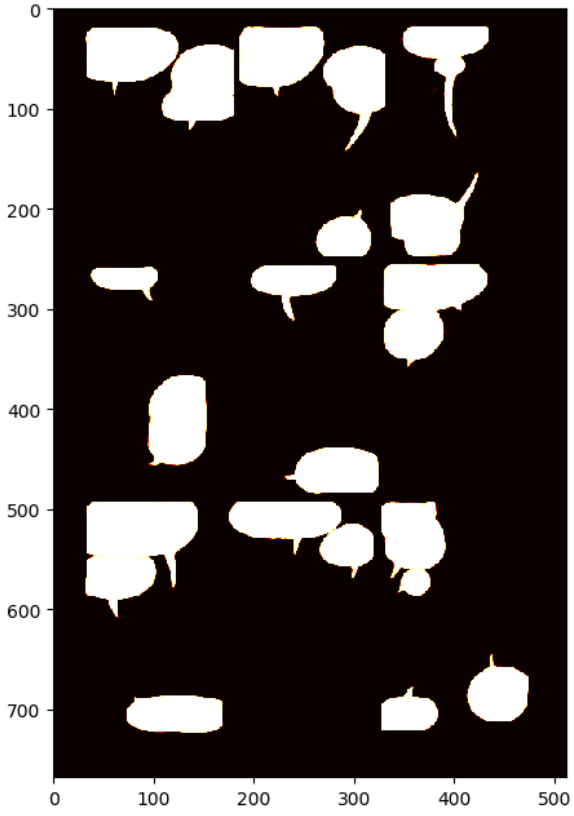


Genauigkeit der Klassifikation. Die x-Achse gibt an, welche Netzwerkschichten zur Berechnung der Signatur herangezogen wurden (siehe Text für Details). Die Linienfarbe unterscheidet, ob die Klassifikation auf Signaturen einzelner Netzwerkschichten k basiert (rot) oder auf Signaturen der Schichten von 0 bis k, k#{0, ..., 11} (blau).

Basierend auf den Merkmalsvektoren haben wir eine bildbasierte Ähnlichkeitssuche implementiert. Nach Eingabe eines Suchbildes werden beispielsweise die 10 ähnlichsten Bilder ausgegeben. Die Untersuchung der Klassifikationsfehler ist interessant, sie zeigt beispielsweise, dass unterschiedliche Werke eines Autors

zusammen gruppiert werden. Verwechslungen treten eher innerhalb von als zwischen Genres auf. Selbst historische Entwicklungen lassen sich abbilden: In „750 Years in Paris“ illustriert Vincent Mahé die Entwicklung eines Häuserblocks in Paris von 1265 bis 2015. Die Bildsuche mit einer „frühen“ Seite liefert Bilder aus der frühen Zeit, ebenso liefert die Bildsuche mit einer „späten“ Seite Bilder aus einer späteren Epoche.

Bei der semantischen Segmentation von Sprechblasen haben wir ein hervorragendes Ergebnis erzielt. Der F1-Score auf dem Testset betrug 0.935. Auch Elemente wie ein geschwungener Hinweisstrich / Dorn und an den Rändern offene Sprechblasen konnten sehr gut segmentiert werden. Abbildung 2 zeigt ein Beispiel einer Seite, auf der alle Sprechblasen korrekt detektiert und sehr gut segmentiert wurden.



Beispiel für Sprechblasen-Segmentation.

Diskussion

Wir haben verschiedene Sammlungen grafischer Literatur mit CNNs beschrieben und den Beitrag interner CNN-Repräsentationen unterschiedlicher Schichten zur Klassifikation von Zeichenstilen untersucht. Unsere Ergebnisse zeigen, dass der Individualstil eines Zeichners eher durch Merkmale mittlerer als durch solche höherer Komplexität charakterisiert ist. Allgemein haben CNN-basierte Repräsentationen das Potenzial, eine formale Beschreibung stilistischer Merkmale abzubilden. Sie sind deshalb aussichtsreiche Kandidaten für eine quantitative Fundierung bildwissenschaftlicher Form- und Strukturanalyse.

Bibliographie

Chollet, F. (2016): *Xception: Deep learning with depthwise separable convolutions*. In: CoRR: abs/1610.02357.

Chu, W. / Wu, Y. (2018): *“Image style classification based on learnt deep correlation features.”* In: IEEE Transactions on Multimedia 20(9): 2491–2502.

Deng, J. / Dong, W. / Socher, R. / Li, L.-J. / Li, K. / Fei-Fei, L. (2009): *„ImageNet: A Large-Scale Hierarchical Image Database.”* In: 2009 IEEE Conference on Computer Vision and Pattern Recognition: 248–255.

Dunst, A. / Hartel, R. / Laubrock, J. (2017): *„The Graphic Narrative Corpus (GNC): Design, annotation, and analysis for the Digital Humanities.”* In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) 03: 15–20.

Laubrock, J. / Hohenstein, S. / Kümmerer, M. (2018): *„Attention to comics: Cognitive processing during the reading of graphic literature.”* In: Dunst, A., Laubrock, J., and Wildfeuer, J., editors, Empirical Comics Research: Digital, Multimodal, and Cognitive Methods, ch.12: 239–263. Routledge, New York.

Matsui, Y. / Ito, K. / Aramaki, Y. / Fujimoto, A. / Ogawa, T. / Yamasaki, T. / Aizawa, K. (2017): *“Sketch-based manga retrieval using Manga109 dataset.”* In: Multimedia Tools and Applications 76(20): 21811–21838.

Gatys, L. A. / Ecker, A. S. / Bethge, M. (2016): *“Image style transfer using convolutional neural networks.”* In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016: 2414–2423.

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.”* In: Neural Computation 1(4): 541–551.

Ronneberger, O. / Fischer, P. / Brox, T. (2015): *„U-Net: Convolutional Networks for Biomedical Image Segmentation.”* In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, LNCS, Vol.9351: 234--241.

Sanakoyeu, A. / Kotovenko, D. / Lang, S., / Ommer, B. (2018): *“A Style-Aware Content Loss for Real-time HD Style Transfer.”* In: arXiv preprint arXiv:1807.10201.

Wölfflin, H. (1915): *Kunstgeschichtliche Grundbegriffe: das Problem der Stilentwickelung in der neueren Kunst.* München: Bruckmann.

Yamins, D. L. K. / DiCarlo, J. J. (2016): *“Using goal-driven deep learning models to understand sensory cortex”*. In: Nature Neuroscience 19(3):356–365.

Simonyan, K. / Zisserman, A. (2014): *"Very deep convolutional networks for large-scale image recognition."* In: arXiv preprint arXiv:1409.1556