Hund

Gassi gehen im Latent Space

Lasse Scherffig und Thomas Hawranke

> »Thanks to Deep Dream, we now know that machines see things through a kind of fractal prism that puts doggy faces everywhere.«[1]

Einleitung

Im Jahr 2015 erreicht der aktuelle Hype um ›Künstliche Intelligenz"[2] erstmals die breite Öffentlichkeit. Sein Gesicht sind Hundeschnauzen – zu sehen auf Bildern und in Videos, gemorpht in Wolken, Architektur und andere Szenen. Auch der wissenschaftliche Diskurs des Feldes rekurriert immer wieder auf Tiere: Ihre Bilder dienen als Trainingsmaterial für zahllose ›neuronale Netze‹, sie werden genutzt, um die Grenzen von Systemen zu diskutieren[3] und sie werden zur technischen Komponente solcher Systeme, wenn etwa ihr Fell als Bildrauschen in Denoising-

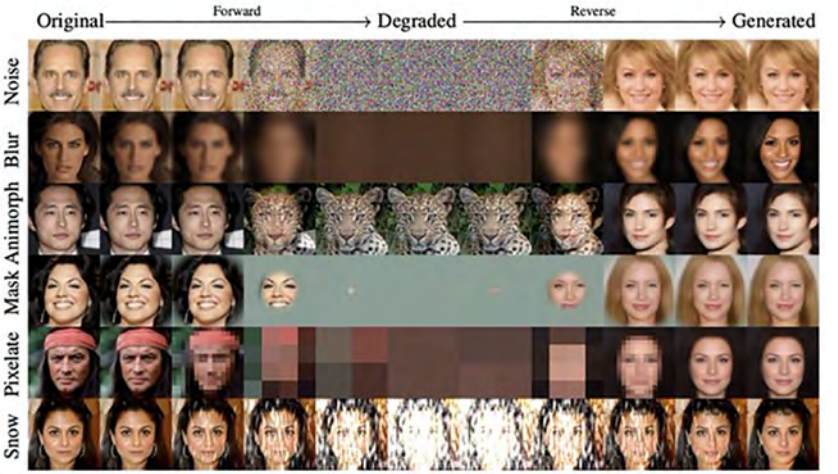
1 Connor, Michael (2015): »Why is Deep Dream turning the world into a doggy monster hellscape?«, in: Rhizome. Online unter: https://rhizome.org/editorial/2015/jul/10/deep-dream-doggy-monster/ (letzter Zugriff: 26.03.2024).

2 Um auf die oft problematischen (sprachlichen) Analogien zwischen statistischen Verfahren und lebenden Systemen hinzuweisen, setzen wir die entsprechenden Fachbegriffe aus der Welt ›Künstlicher Intelligenz‹ in einfache Anführungszeichen und verweisen auf Diskussionen wie Hunger, Francis (2023): Unhype Artificial ›Intelligence‹! A proposal to replace the deceiving terminology of AI. Online unter: https://zenodo.org/records/7524493 (letzter Zugriff: 08.03.2024).

3 »There is no human that could do what current image-generation AI does […] On the other hand, a human that has seen one million horses knows that horses have 4 legs.« Chollet, François (25.09.2022, 2:46), Twitter: https://twitter.com/fchollet/status/1573836241875120128 (letzter Zugriff: 28.03.2024).

Architekturen[4] zum Einsatz kommt (siehe Abb. 1). Die Gründe hierfür sind vielfältig und reichen von der schlichten Verfügbarkeit der Bilder bis zur Tatsache, dass die Geschichte ›neuronaler Netze‹ eng mit der Geschichte des Klassifizierens und der Taxonomisierung von Tieren verbunden ist. Sie sollen im Folgenden näher beleuchtet werden.

Abbildung 1: Tiere als Rauschen



Dieser Beitrag startet mit einer Einführung in das Feld heutiger ›generativer Künstlicher Intelligenz‹ und dessen Ausgangspunkt in Google's DeepDream, die uns immer wieder zum Yorkshire Terrier und anderen Hunden führt. Der virtuelle Yorkshire Terrier fungiert hier als metaphorischer und methodischer Zugang zu den Funktionsweisen dieser Systeme und deren Grenzen. Wir nehmen den Yorki an die Leine und gehen mit ihm an die frische Luft.

Träumende Netze

Netze

Sogenannte ›künstliche neuronale Netze‹ sind seit den 1940er-Jahren Bestandteil der Forschung in Kybernetik und Informatik. Motiviert in (stark vereinfachter) Ana-

4 Denoising-Architekturen bezeichnen hier statistische Modelle, mit deren Hilfe sich Bilder von Rauschen befreien lassen, die aber auch, wie wir unten erläutern, genutzt werden können, um neue Bilder zu erzeugen.

logie zu biologischen neuronalen Netzen sind sie doch eindeutig in der Signalverarbeitung der Kybernetik verortet. Sie stehen damit in der Tradition von Norbert Wieners Gründungsschrift »Cybernetics: or Control and Communication in the Animal and the Machine«[5] und fungieren lange als ›subsymbolische‹ Alternative zur ›symbolischen Künstlichen Intelligenz‹.

Die Entwicklung dieser Technologien verläuft in verschiedenen Phasen. Grundsätzliche Funktionsweisen und Anwendungsfelder sind dabei aber bemerkenswert stabil geblieben: Das »Perceptron«, das erste ›neuronale‹ Modell, das im heutigen Sinne ›lernen‹ konnte, wurde in der Bilderkennung eingesetzt und funktionierte als Klassifikator.[6] Es war in der Lage, visuelle Eingabedaten diskreten Klassen zuzuordnen und damit die Zugehörigkeit von Instanzen zu diesen Klassen zu ›erkennen‹.

Auch die grundsätzlichen Topologien ›neuronaler‹ statistischer Modelle basieren bis heute auf Ideen aus der Zeit der Kybernetik. Mehr noch: Das ›Lernen‹ dieser Systeme basiert weiter im Wesentlichen auf der Perzeptron-Lernregel.[7] Nach jahrzehntelanger Weiterentwicklung dieser Prinzipien, sind ›künstliche neuronale Netze‹ damit weiterhin mathematische Optimierungsverfahren, die auf Fehlerreduktion basieren. Der heute ›loss‹ genannte Abstand zwischen Ausgabegröße und Zielgröße schreibt das kybernetische Erbe des Feldes fort, für das die Verwendung solchen negativen Feedbacks konstituierend ist.[8]

Vor diesem Hintergrund wundert es nicht, dass entscheidende Entwicklungen auf dem Weg zur heutigen ›Künstlichen Intelligenz‹ den Bereich der Klassifikation betreffen. Jährlich stattfindende Wettbewerbe dokumentieren hier eine beständige Auseinandersetzung mit Klassifikationsaufgaben, einer der wichtigsten dieser Wettbewerbe war dabei lange der »ImageNet Large Scale Visual Recognition Challenge« (ILSVRC). Hier und in anderen Wettbewerben spielten ›neuronale‹ Architekturen für eine lange Zeit allerdings eine untergeordnete Rolle. Dies änderte sich spätestens, als im Jahr 2012 erstmals ein Vertreter einer neuen Art dieser Systeme den ILSVRC gewann: AlexNet war nicht nur ein ›tiefes neuronales Netz‹ und damit ein Vertreter des Deep-Learning, sondern auch ein Convolutional Neural Network (CNN).[9]

---

5 Wiener, Norbert (2013): Cybernetics: or Control and Communication in the Animal and the Machine (1948), Cambridge, Mass.: MIT Press.

6 Vgl. Rosenblatt, Frank (1958): »The perceptron: a probabilistic model for information storage and organization in the brain«, in: Psychological Review 65, S. 386–408.

7 Vgl. Offert, Fabian/Bell, Peter (2021): »Perceptual bias and technical metapictures: critical machine vision as a humanities challenge«, in: AI & SOCIETY 36, S. 1133–1144, hier S. 1134.

8 Vgl. Galison, Peter (1994): »The Ontology of the Enemy: Norbert Wiener and the Cybernetic Vision«, in: Critical Inquiry 21, S. 228–266.

9 Vgl. Mitchell, Melanie (2019): Artificial intelligence. A guide for thinking humans, New York: Picador, S. 85.

Bis heute klassifizieren ›neuronale Netze‹, indem sie numerische Eingaben durch eine Reihe mathematischer Operationen in eine ebenfalls numerische Ausgabe transformieren. Dabei werden Eingabevektoren mit Gewichten multipliziert, während das Ergebnis dieser Multiplikationen in ›Neuronen‹ genannten Elementen aufsummiert wird. Diese Summe dient einer Aktivierungsfunktion als Argument, die die Ausgabe des jeweiligen Elements bestimmt. Elemente und Gewichte sind dabei in Schichten organisiert und Eingaben durchlaufen das System Schicht für Schicht.

Im ›maschinellen Lernen‹ (oder ›Training‹) von Klassifikatoren werden die Gewichte zunächst zufällig gewählt, womit die Daten ebenfalls zufällig klassifiziert werden. Diese Klassifikation lässt sich nun mit einer gegebenen vergleichen, die für die Trainingsdaten bekannt sein muss (›überwachtes Lernen‹). Auf Grundlage der Differenz von berechneter und gewünschter Ausgabe werden dann die Gewichte des Systems angepasst, bis irgendwann praktisch alle Trainingsdaten korrekt klassifiziert werden.

Während das einfachste Perzeptron nur eine einzige Schicht besitzt, die ein einziges Element enthält, folgt in moderneren Architekturen auf die Eingabeschicht mindestens eine Zwischenschicht vor der Ausgabe, der sogenannte ›Hidden Layer‹.

Mit dem Erfolg von AlexNet wurde das Potential von Deep-Learning-Architekturen mit mehr als einer Zwischenschicht offensichtlich. Neben leichten Veränderungen in der Lernregel und Aktivierungsfunktion hat dies vor allem Gründe, die ihrem Ursprung außerhalb des Feldes haben: »It turns out that the recent success of deep learning is due less to new breakthroughs in AI than to the availability of huge amounts of data (thank you, internet!) and very fast parallel computer hardware.«[10]

Das Problem mit großen Datenmengen ist aber, dass deren effiziente Verarbeitung eine Vorverarbeitung voraussetzt, die die Größe einzelner Datenpunkte reduziert. Insbesondere in der Bilderkennung kamen hier früh Filter zum Einsatz, die Bilddaten glätten, schärfen oder Kanten darin erkennen, um bereits vor dem ›Lernen‹ relevante Eigenschaften der Daten zu extrahieren. Solches »feature engineering« bedeutet »applying hardcoded (nonlearned) transformations to the data before it goes into the model« und setzt ein gutes Verständnis der betreffenden Problemstellung und Daten voraus.[11]

Die zentrale Idee von »Convolutional Neural Networks« ist eben diese Vorverarbeitung ›mitzulernen‹.[12] CNNs optimieren ihre Vorverarbeitungsfilter im Sinne ihrer Aufgabe. In tiefen Netzen können diese Filter hierarchisch aufeinander aufbauen, vom ›Erkennen‹ lokaler Strukturen (wie Kanten und deren Orientierung) zu

10 Mitchell: Artificial intelligence, S. 89.

11 Chollet, François (2018): Deep learning with Python, Shelter Island, NY: Manning Publications Co, S. 102–103.

12 Zur Einführung in CNNs vgl. Mitchell: Artificial intelligence, S. 70–79.

komplexeren Strukturen. »This allows convnets to efficiently learn increasingly complex and abstract visual concepts […]«[13] Aber wie sehen diese ›Konzepte‹ aus?

DeepDream

> »A concrete example would be that of a ›cat‹ and ›dog‹ classifier. Hypothetically, we can train a standard CNN architecture on a large dataset of cat images and dog images. […] But, how are the concepts ›dog‹ and ›cat‹ encoded in the system—a system that clearly somehow ›knows‹ what these concepts are?«[14]

Diese Frage stellen sich auch die Entwickler*innen beim Einsatz von CNNs. Die Systeme gelten als schwer zu verstehen, subsymbolische Verfahren haben ihren Namen schließlich genau aus dem Umstand, dass ihre Repräsentationen zwar technisch funktionieren, aber nicht ohne weiteres »extrasymbolisch«, im Sinne eines Für-etwas-Stehens, interpretiert werden können.[15] »Interpretability« und »explainability« dieser Modelle sind nicht umsonst aktuelle Forschungsfelder.[16] Das Team, das bei Google an den Inception-Netzen arbeitet, die zur Teilnahme am ILSVRC trainiert werden, stellt dementsprechend fest: »One of the challenges of neural networks is understanding what exactly goes on at each layer.«[17]

Unter Rückgriff auf vorangegangene Arbeiten[18] entwickelt das Team schließlich das Verfahren, das unter dem Namen DeepDream berühmt werden wird. Es ist ein Spezialfall von »feature visualization«, der Visualisierung der Eigenschaften, die die Filter eines CNNs aus den Daten extrahieren. Es zielt darauf ab, Bilder zu erzeugen, auf die die Filter in den verschiedenen Schichten eines Modells optimal ansprechen und so zu zeigen, auf was zu ›erkennen‹ sie eingestellt sind.

Technisch gesehen ist auch Feature Visualization ein Optimierungsproblem: Statt aber hier den Fehler der Ausgabe eines Netzes zu minimieren, um eine korrekte Klassifizierung zu erreichen, wird die Eingabe optimiert, um die Aktivierung von Teilen des Netzes zu maximieren. Das Verfahren startet mit einem Bild aus zufälligem Rauschen und verändert die Pixel des Bildes dann im Sinne einer möglichst starken Aktivierung. Aus dem Gradientenabstieg maschinellen Lernens wird ein

---

13 Chollet: Deep learning with Python, S. 123.

14 Offert/Bell: Perceptual bias and technical metapictures, S. 1136.

15 Zu »intrasymbolischer« und »extrasymbolischer« Bedeutung vgl. Krämer, Sybille (1988): Symbolische Maschinen. Die Idee der Formalisierung in geschichtlichem Abriss, Darmstadt: Wissenschaftliche Buchgesellschaft, S. 60.

16 Vgl. Offert/Bell: Perceptual bias and technical metapictures, S. 1135–1136.

17 Mordvintsev, Alexander/Olah, Christopher/Tyka, Mike (2015): »Inceptionism: Going Deeper into Neural Networks«, in: Google Research Blog. Online unter: https://blog.research.google/2015/06/inceptionism-going-deeper-into-neural.html (letzter Zugriff: 27.03.2024).

18 Einen Überblick über Methoden liefert Chollet: Deep learning with Python, S. 160–177.

Gradientenaufstieg auf den Bilddaten selbst, der den Akt der Klassifikation durch ein ›neuronales Netz‹ ins Generative umkehrt.[19]

Das Verfahren erzeugt Bilder, die zeigen, was die einzelnen Teile eines Netzes nach dem Training ›erwarten‹, um klassifizieren zu können. Dabei lässt sich üblicherweise leicht erkennen, dass die Filter mit zunehmender Tiefe des Netzes auf zunehmend komplexere Textureigenschaften eingestellt sind: Während das anfangs einfache Kanten sein mögen, sind dies später komplexe Texturen.[20]

In der Folge entfaltet DeepDream seine Wirkung vor allem als ästhetisches Phänomen. Das Team bei Google verändert das Visualisierungsverfahren vor allem dahingehend, dass statt reinem Rauschens ein beliebiges Bild als Eingabe verwendet wird, womit die Optimierung der Bilddaten nun zu dessen ›irgendwie künstlerischen‹ Veränderung führt: »the resulting effects latch on to preexisting visual patterns, distorting elements of the image in a somewhat artistic fashion.«[21]

Das Ergebnis schlägt ein als »internet sensation«[22], mit Bildern, die der Medienkünstler und *early adopter* Memo Akten als »trippy, surreal, abstract, psychedelic, painterly, rich in detail«[23] beschreibt. Als »Nebenprodukt des Versuchs, die Wahrnehmungslogik der CNNs begreifbar zu machen« ist DeepDream damit Ausgangspunkt heutiger ›generativer Künstlicher Intelligenz"[24]

---

19 Vgl. ebd., S. 280.

20 Weil diese Methoden, die im Modell repräsentierten ›visuellen Konzepte‹ erfahrbar machen, suggerieren sie deren leichte Verständlichkeit. Ganz so einfach ist die Lage allerdings nicht. Wie Offert und Bell ausführlich gezeigt haben, ist eine reine Feature Visualization weniger interpretierbar, als eine, die in Richtung einer Bildlichkeit gelenkt wurde, die unseren Wahrnehmungskonventionen entspricht. Ohne dieses Lenken sind die entstehenden Bilder im Wesentlichen »noise and nonsensical high-frequency patterns« (Olah et al.). Das liegt unter anderem daran, dass die Repräsentation visueller Konzepte hier gerade nicht in Form unserer Sehgewohnheiten stattfindet. Da jedes Element eines Netzes an der ›Erkennung‹ aller Kategorien beteiligt ist, sind diese Kategorien zwangsläufig zu einem gewissen Grad auf alle Elemente verteilt. Je lesbarer eine Visualisierung wird, desto weniger nah ist sie an der »non-humanness« (Offert und Bell) mit der Konzepte in CNNs tatsächlich kodiert sind. Interpretierbare Feature Visualizations sind also als Illustrationen zu verstehen, als Sichtbarmachen des Nicht-Visuellen. Vgl. hierzu Offert/Bell: Perceptual bias and technical metapictures sowie Olah, Chris/Mordvintsev, Alexander/Schubert, Ludwig (2017): »Feature Visualization«, in: Distill 2. Online unter: https://distill.pub/2017/feature-visualization/ (letzter Zugriff: 14.05.2024).

21 Chollet: Deep learning with Python, S. 280.

22 Ebd., S. 280.

23 Akten, Memo (2015): »#Deepdream is blowing my mind«, in: Medium. Online unter: https://memoakten.medium.com/deepdream-is-blowing-my-mind-6a2c8669c698 (letzter Zugriff: 27.03.2024).

24 Vgl. Offert, Fabian (2023): »KI-basierte Verfahren in der bildenden Kunst«, in: Stephanie Catani (Hg.), Handbuch Künstliche Intelligenz und die Künste, De Gruyter, S. 202–216, hier S. 203–206.

Neben ihrem psychedelischen Charakter zeigen die Bilder vor allem Tierhaftes: Was sich in die existierenden Muster der eingegeben Bilder einklinkt, sind Texturen von Fell, Federn, Augen und Schnauzen. Sie verwandeln die Ausgangsbilder in eine »doggy monster hellscape«[25] mit Spezies wie der berüchtigten ›Puppy Slug‹-Chimäre aus Hund und Schnecke (siehe Abb. 2).

Abbildung 2: Puppy Slug

So werden Hunde zum Gesicht des ersten Hypes um ›generative Künstliche Intelligenz‹. Der Grund für die Sichtbarkeit von Tierlichkeit in diesem Feld liegt aber tiefer, als dass Hunde ein dankbares Bildmotiv und damit ein »happy medium«[26] für das sehr technische Problem der Klassifikation von Bildern darstellen. Er verweist vielmehr auf den Kern ›generative Künstliche Intelligenz‹ als Dispositiv, der sich am besten am Zusammenhang von DeepDream mit dem ISLVRC ablesen lässt.

Der ISLVRC wurde 2010 als jährlicher Benchmark für die Klassifikation von Bildern gegründet. Sein Datensatz basiert auf dem ImageNet-Datensatz. ImageNet wiederum ist Ergebnis des Versuchs, eine Ontologie anzubieten, die als standardisierte Grundlage für die Entwicklung von Algorithmen zur Bilderkennung fungieren kann.[27]

25 Connor: Why is Deep Dream turning the world into a doggy monster hellscape?

26 Ebd.

27 Vgl. Deng, Jia/Dong, Wei/Socher, Richard/Li, Li-Jia/Li, Kai/Fei-Fei, Li (2009): »ImageNet: A large-scale hierarchical image database«, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, S. 248–255, hier S. 248.

In der Praxis ist diese Ontologie eine Datenbank, die Bilddateien mit Schlagworten verknüpft, wobei die Schlagworte eindeutig bezeichnen sollen, was auf den Bildern zu sehen ist. Das Projekt steht damit vor einem grundsätzlichen Problem, das sich durch die Geschichte des Klassifizierens zieht: Klassifikationen benötigen eine standardisierte Terminologie und sind von ihr nicht eindeutig zu trennen.[28] Die Antwort von ImageNet ist pragmatisch und greift auf vorgefertigte Standardisierungen zurück: ImageNet verwendet WordNet, eine Datenbank von Wörtern und ihren semantischen Beziehungen, die Sprache hierarchisch vom Speziellen zum Allgemeinen organisiert (zum Beispiel vom Yorkshire Terrier zum Hund, zum Tier, zum Organismus). Sie fasst Wörter in »synsets« genannte Gruppen gleicher Bedeutung (»cognitive synonyms«) zusammen und kann damit die Aufgabe der Disambiguierung für ImageNet übernehmen.[29]

Das Projekt steht aber weiter vor dem Problem, dass Bilder und deren sprachliche Beschreibung alles andere als eindeutig zusammenhängen.[30] Aber auch die Aufgabe der Zuordnung von Bildern zu Synsets wird im Projekt pragmatisch abgegeben. Nachdem Millionen Bilder als potentielle Vertreter der ImageNet-Kategorien automatisiert zusammengetragen wurden (»thank you, internet!«), wird das Klassifizieren der Bilder zur Aufgabe einer großen Zahl ›Clickworker‹. Akquiriert über die Plattform Amazon Mechanical Turk, werden sie bezahlt, die Bilder auf ihre Zugehörigkeit zu bestimmten Synsets zu überprüfen (siehe Abb. 3). So werden zwölf Millionen Bilder in 21000 Kategorien eingeordnet. Das Projekt beschäftigt damit 20.000-30.000 Arbeiter*innen im Jahr und ist zeitweise der größte akademische Arbeitgeber auf der Plattform.[31] Das bedeutet auch, dass die Klassifizierungen hier in der Regel nicht von Expert*innen für irgendeines der betroffenen Themengebiete vorgenommen werden, sondern von schlecht bezahlten Lai*innen.

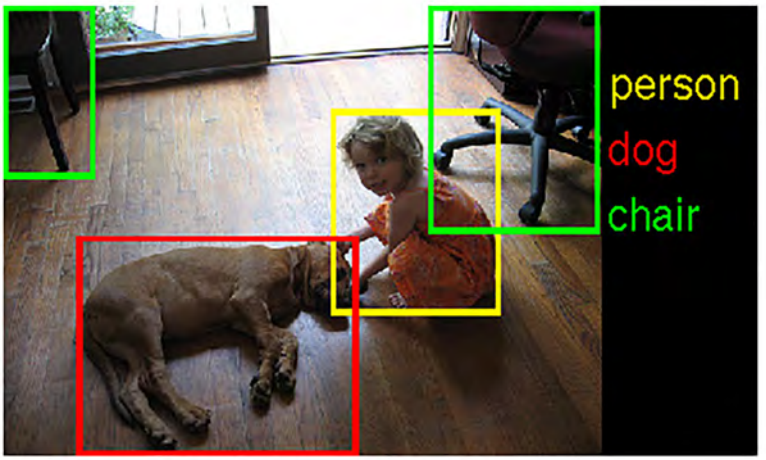
28 Vgl. Ritvo, Harriet (1998): The platypus and the mermaid, and other figments of the classifying imagination, Cambridge, Mass.: Harvard University Press, S. 51–58.

29 Vgl. Russakovsky, Olga/Deng, Jia/Su, Hao/Krause, Jonathan/Satheesh, Sanjeev/Ma, Sean/Huang, Zhiheng/Karpathy, Andrej/Khosla, Aditya/Bernstein, Michael/Berg, Alexander C./Fei-Fei, Li (2015): »ImageNet Large Scale Visual Recognition Challenge«, in: International Journal of Computer Vision 115, S. 211–252.

30 Für eine Diskussion im Kontext von Bilderkennung vgl. Crawford, Kate/Paglen, Trevor (2021): »Excavating AI: the politics of images in machine learning training sets«, in: AI & SOCIETY 36, S. 1105–1116.

31 Vgl. Markoff, John (2012): »Seeking a Better Way to Find Web Images«, in: The New York Times vom 19.11.2012. Online unter: https://www.nytimes.com/2012/11/20/science/for-web-images-creating-new-technology-to-seek-and-find.html (letzter Zugriff: 27.03.2024).

Abbildung 3: Klassifikation auf Amazon MTurk



Auf Grundlage von ImageNet kann so 2010 der ILSVRC starten.[32] Der Wettbewerb verwendet eine Teilmenge von ImageNet, die zunächst aus 1,4 Millionen Bildern in 1.000 Kategorien besteht. Die 1.000 Klassen[33] werden anfangs zufällig ausgewählt. Diese zufällige Zusammensetzung wird aber wiederholt geändert, insbesondere im Jahr 2012: »In ILSVRC2012, 90 synsets were replaced with categories corresponding to dog breeds […].«[34] Die Kategorien werden aus dem Datensatz »Stanford Dogs«[35] übernommen, der interessanterweise, und entgegen seiner Beschreibung, eine Tierart enthält, die kein domestizierter Hund ist – der »African wild dog«.

Die Änderung sollte dazu dienen, evaluieren zu können, ob die teilnehmenden Bilderkennungstechnologien zu »fine-grained object classification«[36] in der Lage sind – ob sie also die Unterschiede zwischen verschiedenen Hunderassen[37] genau-

32 Vgl. Russakovsky et al.: ImageNet Large Scale Visual Recognition Challenge.

33 Die Begriffe ›Klasse‹ und ›Kategorie‹ werden im Diskurs um ›maschinelles Lernen‹ synonym verwendet.

34 Ebd., S. 215.

35 Khosla, Aditya/Jayadevaprakash, Nityananda/Yao, Bangpeng/Fei-Fei, Li (2011): Stanford Dogs Dataset. Online unter: http://vision.stanford.edu/aditya86/ImageNetDogs/ (letzter Zugriff: 27.03.2024).

36 Russakovsky et al.: ImageNet Large Scale Visual Recognition Challenge, S. 215.

37 Wie wir später sehen werden, ist der englische Begriff ›dog breed‹ weitaus präziser als der deutsche Begriff ›Hunderasse‹.

so gut erkennen können, wie etwa die zwischen Hunden und Katzen, Autos oder Hanteln.

Bereits im ImageNet-Datensatz sind Hunde überproportional häufig vertreten, mit dem Yorkshire-Terrier als häufigstes Bild.[38] Mit der Anpassung des ILSVRC werden auch hier Hunde zu den meist vertretenen Motiven auf Bildern. Nicht nur finden sich jetzt unter den 1.000 Kategorien 120 (bzw. 119) Hunderassen. Von den 1.281.167 Bildern in den Trainingsdaten sind nun 150.473 (oder knapp 12 %) als Hunde klassifiziert.[39] Konsequenterweise bilden die ›Netze‹, die auf dem Datensatz trainiert werden, eigene ›Yorkshire-Terrier-Neurone‹ aus (siehe Abb. 4). Ziel des ILSVRC ist seitdem gewissermaßen das ›Erkennen‹ von Hunden.

Abbildung 4: Yorkshire Terrier Neuron im OpenAI Microscope

Text to image

DeepDream ist mittlerweile Geschichte. Die Techniken, mit denen generative Modelle Text und Bilder erzeugen, haben sich seit 2015 stark verändert. Aktuelle Transformer- oder Diffusion-Architekturen bekommen weitaus mehr Aufmerksamkeit als die Puppy Slugs von damals. Im Folgenden wollen wir aber zeigen, dass die alten und neuen Verfahren ›generativer KI‹ im Klassifizieren und Standardisieren einen gemeinsamen Moment besitzen – und dass sich dieser Moment durch den Blick auf die Klassifizierung und Standardisierung von Hunderassen verstehen lässt.

Die entscheidende Entwicklung hinter den neuen Verfahren zur Bilderzeugung ist die Veröffentlichung von CLIP 2021: »Contrastive Language-Image Pre-training« steht hier für die Zusammenführung verschiedener Ansätze aus den Feldern der

38 Vgl. Thoma, Martin (2016): »Answer to: What is the distribution of categories in imagenet training set (ILSVRC2012)«, in: Stackexchange. Online unter: https://datascience.stackexchange.com/questions/11777/what-is-the-distribution-of-categories-in-imagenet-training-set-ilsvrc2012/13452#13452 (letzter Zugriff: 28.03.2024).

39 Ermittelt durch den algorithmischen Abgleich der Kategorien aus dem Datensatz mit ihrer Position in der Wordnet-Hierarchie, die bei Hunden unterhalb des Synsets für domestizierte Hunde liegen muss.

Bild- und Textverarbeitung.[40] Eine wichtige Rolle spielt dabei das Ersetzen der Überwachung des ›Trainings‹: vom Lernen aus vorgefertigten Beispielen hin zu »natural language supervision«. Die Bestimmung dessen, was wahr im Sinne des Trainings ist, erfolgt also nicht länger durch die Klassifikation von Bildern, sondern durch die Paarung von Bildern mit ihrer ›natürlichsprachlichen‹ Beschreibung.

Weiter stützt sich CLIP auf etablierte Deep-Learning-Methoden, die ihre Trainingsdaten in einen Vektorraum abbilden, der zwar hochdimensional ist, aber weniger Dimensionen besitzt als die Daten selbst. Dieser »Latent Space« genannte Raum bildet gewissermaßen ein statistisches Abbild dessen, was einem Datensatz als stabile Muster gemein ist und kann genutzt werden um Instanzen (also z.B. einzelne Bilder) zu generieren, die als plausible Vertreter der Trainingsdaten erscheinen – obwohl sie selbst nicht darin enthalten sind.

Der Latent Space wird durch die Parameter der Netze definiert, die Daten auf ihre Vektorrepräsentation (die auch ›embedding‹ genannt wird) abbilden oder diese Repräsentation wieder in den Typ der Trainingsdaten überführen (Kodierung/Dekodierung). Er ist dabei typischerweise signifikant kleiner als die Trainingsdaten selbst und kann damit auch als komprimierte Repräsentation dieser Daten angesehen werden. Vor allem aber ist er durch ihre Gemeinsamkeiten im Sinne der Aufgabe strukturiert.

Bei der Anwendung relativ einfacher generative Modelle (z.B. Autoencoder[41]) auf Datensätze handgeschriebener Ziffern, die mit den Klassen 0–9 annotiert wurden, bildet der entstehende Latent Space beispielsweise einen Raum, in dem alle möglichen Schreibweisen der ›0‹ ein räumlich abgegrenztes Cluster kreieren, das wiederum größere Nähe zu Ziffern wie ›6‹, ›8‹ oder ›9‹ besitzt, als etwa zur ›1‹. Der Begriff der Nähe ist hier wörtlich zu verstehen, als geometrischer Abstand von Vektoren. Die Clusterbildung ist dabei zu gleichen Teilen Folge der visuellen Struktur der Trainingsdaten, wie auch ihrer Annotierung in Klassen.

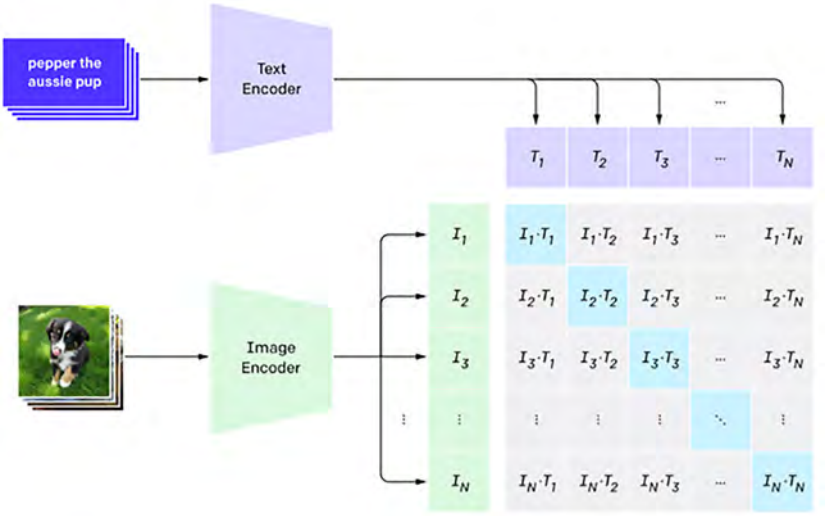
Da CLIP auf Bilddaten mit Überwachung aus Bildbeschreibungen trainiert wird, bildet es nun einen gemeinsamen Latent Space für Texte und Bilder (siehe Abb. 5): »CLIP learns a multi-modal embedding space by jointly training an image encoder and text encoder to maximize the cosine similarity of the image and text embeddings […].«[42] Die »cosine similarity« bezeichnet hier ein geometrisches Maß des Abstandes zweier Winkel.

40 Radford, Alec/Kim, Jong W./Hallacy, Chris/Ramesh, Aditya/Goh, Gabriel/Agarwal, Sandhini/Sastry, Girish/Askell, Amanda/Mishkin, Pamela/Clark, Jack/Krueger, Gretchen/Sutskever, Ilya (2021): »Learning Transferable Visual Models From Natural Language Supervision«, in: arXiv. Online unter: https://arxiv.org/abs/2103.00020 (letzter Zugriff: 28.03.2024).

41 Zur Einführung vgl. Chollet: Deep learning with Python, S. 296–304.

42 Radford et al.: Learning Transferable Visual Models From Natural Language Supervision, S. 4.

Abbildung 5: Multimodaler Latent Space in Clip (mit Hund)



Dieser ›multimodale‹ Latent Space bildet damit Text und Bilder auf das gleiche Format ab und kann genutzt werden, um ihre ›Ähnlichkeit‹ zu bestimmen. Die Verknüpfung von Text und Bild hat zur Folge, dass man nun Bilder (durch Dekodierung) erzeugen und diese Erzeugung durch Texte lenken kann (was gemeinhin als ›Konditionierung‹ bezeichnet wird).[43] Sobald nämlich die Berechnung einer Ähnlichkeit von Text und Bild (als Kosinus-Ähnlichkeit von Vektoren) möglich ist, lässt sich die Erzeugung von Bildern dahingehend optimieren, dass sie möglichst einer gegebenen textuellen Beschreibung entsprechen. Diese Beschreibung wird in diesem Kontext als ›Prompt‹ bezeichnet.

Zero-shot Classification

Obwohl multimodale Modelle derzeit vor allem im Bereich der generativen Verfahren Anwendung finden, lassen sich diese nicht eindeutig von Verfahren zur Klassifizierung trennen. Als mit CLIP das Lernen vorgegebener Klassen durch ein sprachüberwachtes Training ersetzt wird, geschieht dies weiterhin um Klassifikationsaufgaben bewältigen zu können, unter anderem bei der »fine-grained object

43 Dall-e, das zu den ersten ›multimodalen‹ Systemen gehört, verwendet nicht direkt CLIP, wird aber teilweise vom selben Team unter Verwendung ähnlicher Prinzipien entwickelt. Andere Systeme verwenden CLIP oder CLIP-Derivate direkt.

classification«[44] wegen der bereits der ILSVRC zum Hundeerkennungswettbewerb gemacht worden war. Anders als in anderen Systemen, bedeutet Klassifizieren hier »zero-shot classification« – das Zuordnen zu Kategorien, die das System zuvor nicht ›gesehen‹ hat. Hierzu werden die Kategorien eines Datensatzes als textuelle Beschreibungen eines Bildinhaltes aufgefasst, womit CLIP nun aus allen Kategorien des Datensatzes den bestimmt, der am wahrscheinlichsten zutrifft.[45] Dieser Prozess kommt dabei nicht ganz ohne »Prompt Engineering« aus, weil die Datensätze, mit denen CLIP getestet wird, nicht frei vom Problem einer fehlenden Nomenklatur sind und ImageNet, zum Beispiel, den Namen ›Boxer‹ sowohl für Hunde, als auch für sporttreibende Menschen verwendet. Generell hilft es der zero-shot classification auch, wenn statt eines Klassenlabels ein ganzer Satz verwendet wird: »A photo of a {label}«, womöglich mit dem Zusatz »a type of pet.«[46]

Das Training mit sprachlichen Beschreibungen von Bildern verspricht eine größere »generality and usability«[47], weil die entstehenden Modelle hier nicht auf eine zuvor definierte Taxonomie angewiesen sind, sondern auf das implizite taxonomische Wissen ›natürlicher‹ Sprache zurückgreifen können. Zugleich geht es hier aber auch um das Problem der Arbeit: »Learning directly from raw text about images […] leverages a much broader source of supervision.«[48] Die Deep-Learning-Revolution ist wiederholt als Projekt in der Tradition des Taylorismus beschrieben worden.[49] Dass es für Firmen wie OpenAI sinnvoll ist, das (wenn auch schlecht) bezahlte manuelle Klassifizieren durch ein automatisches Scraping des Internets zu ersetzen, das freundlicherweise (thank you internet!) voll von Bildern und deren Beschreibungen ist, liegt nahe.

(Stable) Diffusion

Viele dieser Entwicklungen kommen aktuell in der Software Stable Diffusion zusammen. Das Projekt ist unter anderem deswegen interessant, weil es einerseits dem Stand der Technik entspricht, andererseits als Open-Source-Projekt relativ leicht zugänglich ist.

Um Prompts in Vektorrepräsentationen zu überführen, verwendet Stable Diffusion ursprünglich CLIP sowie später andere ›multimodale‹ Modelle. Das Gesamtsystem ist eine Assemblage unterschiedlicher Komponenten, die teilweise unabhängig

44 Radford et al.: Learning Transferable Visual Models From Natural Language Supervision, S. 1.

45 Vgl. ebd., S. 6.

46 Ebd., S. 7.

47 Ebd., S. 1.

48 Ebd.

49 Für eine umfassende Diskussion vgl. Pasquinelli, Matteo (2023): The eye of the master. A social history of artificial intelligence, London, New York: Verso.

voneinander trainiert werden.[50] Die gelenkte Erzeugung von Bildern erfolgt bei Stable Diffusion aber gerade nicht auf Bilddaten, sondern in einem Latent Space, der abstrakte und damit kleinere, weniger rechenaufwendige Datensätze enthält. Erst eine Dekodierung erzeugt hier die eigentlichen Bilder. In diesem Latent Space werden Inhalte durch »diffusion« erzeugt: Dazu wird ein statistisches Modell verwendet, das darauf trainiert wurde, verrauschte Bilder von diesem Rauschen zu befreien – ein Rauschen das, wie oben gesehen, durchaus auch die Felltextur eines Tieres sein kann. Die schlichte Idee, ein solches Modell nach dem Training auf *reines Rauschen* anzuwenden, führt dazu, dass hier Bilder gewissermaßen aus dem Nichts entstehen. Wenn dieser Prozess gelenkt wird, können die entstehenden Bilder auch einem bestimmten Text entsprechen (weiter natürlich im Sinne der Kosinus-Ähnlichkeit von Vektoren).

Das Training des Modells, das die Diffusion ausführt, geschieht auf LAION-5B, oder, je nach Version, auf Teilmengen davon.[51] LAION-5B besteht dabei aus etwa 5 Milliarden Bild-Text-Paaren, die automatisiert aus dem Internet extrahiert wurden. Es umfasst damit einen »cultural snapshot«[52], der einen algorithmisch bestimmten Teil des Internets und der hier repräsentierten Text- und Bildkulturen umfasst – einschließlich der darin vertretenen Stereotypen und Vorurteile[53], die in solchen Modellen nicht nur abgebildet werden, sondern sich mit Zunahme ihrer Größe sogar verschärfen.[54]

---

50 Einen Überblick geben Rombach, Robin/Blattmann, Andreas/Lorenz, Dominik/Esser, Patrick/Ommer, Bjorn (2022): »High-Resolution Image Synthesis with Latent Diffusion Models«, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, S. 10674–10685.

51 Vgl. Schuhmann, Christoph/Beaumont, Romain/Vencu, Richard/Gordon, Cade/Wightman, Ross/Cherti, Mehdi/Coombes, Theo/Katta, Aarush/Mullis, Clayton/Wortsman, Mitchell/Schramowski, Patrick/Kundurthy, Srivatsa/Crowson, Katherine/Schmidt, Ludwig/Kaczmarczyk, Robert/Jitsev, Jenia (2022): »LAION-5B: An open large-scale dataset for training next generation image-text models«, in: S. Koyejo/S. Mohamed/A. Agarwal u.a. (Hg.), Advances in Neural Information Processing Systems, Curran Associates, Inc, S. 25278–25294.

52 Cetinic, Eva (2022): »The Myth of Culturally Agnostic AI Models«, in: arXiv. Online unter: https://arxiv.org/abs/2211.15271 (letzter Zugriff: 28.03.2024).

53 Vgl. Bianchi, Federico/Kalluri, Pratyusha/Durmus, Esin/Ladhak, Faisal/Cheng, Myra/Nozza, Debora/Hashimoto, Tatsunori/Jurafsky, Dan/Zou, James/Caliskan, Aylin (2023): »Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale«, in: 2023 ACM Conference on Fairness, Accountability, and Transparency, New York, NY, USA: ACM, S. 1493–1504.

54 Vgl. Birhane, Abeba/Prabhu, Vinay/Han, Sang/Boddeti, Vishnu N. (2023): »On Hate Scaling Laws For Data-Swamps«, in: arXiv. Online unter: https://arxiv.org/abs/2306.13141 (letzter Zugriff: 28.03.2024).

Gassi gehen

> »Who can resist the charms of a Yorkshire Terrier? What could shake the blues from your lonely evening more readily than a blue and tan toy terrier? It would appear that almost anyone who's inclined to own a Yorkshire Terrier should do so! There are so many gigantic advantages wrapped up in this smallest of dog breeds.«[55]

Wie oben gesehen, können dem Yorkshire Terrier auch Systeme wie Google's Inception nicht widerstehen und widmen ihm sogar ein eigenes ›Neuron‹. Wie alle Hunde benötigt er aber auch Auslauf.[56] Wir nehmen ihn dementsprechend an die Leine, um mit ihm Gassi zu gehen. Gassi gehen fungiert hier zugleich als Gedankenexperiment und Experiment am Material, als metaphorischer und methodischer Zugang zu den Latent Spaces ›generativer Künstlicher Intelligenz‹ im Zeichen des Hundes, der an der »internet sensation« von DeepDream nicht ganz unbeteiligt war. Gassi gehen bedeutet hier weiter, dass wir nicht mit kurzer Leine spazieren, sondern die Leine so lang lassen, dass der Yorki den eigenen Instinkten nachgehen kann. Das bedeutet, dass auch wir uns bewegen müssen: »The balance of power, in this case, can swing like a seesaw as first the human and then the animal gains the upper hand. Each, alternately, ›walks‹ the other.«[57] Der Yorki ist ein Jagdhund. Wir hoffen, dass er etwas findet.

Nach dem Informatiker und Computerkünstler der ersten Generation Frieder Nake ist eine zentrale Eigenschaft generativer Computergrafik, dass sie die Erzeugung von Systemen oder »Klassen« von Bildern statt einzelner Exemplare betrifft.[58] Es geht ihr um die Eingrenzung von ästhetischen Möglichkeitsräumen auf bestimmte Teilmengen durch die Festlegung von Wahrscheinlichkeitsverteilungen.[59] Dies verbindet sie mit heutigen Praktiken ›generativer Künstlicher Intelligenz‹, die die Möglichkeitsräume anhand der »inhärenten Wahrscheinlichkeitsverteilungen eines (Bild-)Datensatzes« eingrenzen.[60]

---

55 Dog Fancy Magazine (2009): Yorkshire terrier. (Smart Owner's Guide, Kennel Club Books Interactive Series), Freehold, NJ: Kennel Club Books, S. 4.

56 Vgl. Dog Fancy Magazine: Yorkshire terrier, S. 155.

57 Ingold, Tim/Vergunst, Jo Lee (Hg.): Ways of walking. Ethnography and practice on foot (= Anthropological studies of creativity and perception), London, New York: Routledge, Taylor & Francis Group 2016, S. 12.

58 Vgl. Nake, Frieder/Grabowski, Susanne (2021): »Zwei Weisen, das Computerbild zu betrachten. Ansicht des Analogen und des Digitalen«, in: Jan Distelmeyer/Sophie Ehrmanntraut/Boris Müller (Hg.), Algorithmen & Zeichen, Berlin: Kulturverlag Kadmos, S. 36–62, hier S. 51.

59 Vgl. Nake, Frieder (2021): »Über eine generative Ästhetik«, in: Jan Distelmeyer/Sophie Ehrmanntraut/Boris Müller (Hg.), Algorithmen & Zeichen, Berlin: Kulturverlag Kadmos, S. 178–184, hier S. 178.

60 Offert: »KI-basierte Verfahren in der bildenden Kunst«, hier S. 211.

Die Möglichkeitsräume ›generativer KI‹ haben wir oben als ihre Latent Spaces kennengelernt. Wenn ›multimodale‹ Modelle eine Zäsur markieren, dann vor allem im Zugriff auf diese Räume. Da der Latent Space lediglich als mathematisches Konstrukt vorliegt (und *latent* – also verborgen – ist), erfolgt der Zugriff mathematisch, indem Vektoren aus dem Latent Space (oft zufällig) ausgewählt werden, um die dazugehörigen Bilder zu erhalten. Dies ist eine Form der »Erkundung« und es wundert nicht, dass die wichtigste Systematisierung dieser Erkundung eine Form räumlicher Navigation wird: der »latent space walk«[61]. Dabei wird ein Vektor im Latent Space variiert, um zu benachbarten Kodierungen zu gelangen. Ins Bildliche dekodiert erscheinen diese Walks als kontinuierliche Bewegung durch die Klasse der Bilder, die von einem generativen System ermöglicht wird.

Mit CLIP haben multimodale Modelle diesen mathematischen Zugriff durch einen text-basierten abgelöst: »prompt engineering, also die gezielte Komposition von Texteingaben zur Erzeugung ganz bestimmter Bildinhalte, ersetzt die formalen Experimente« der Zeit vor 2021.[62] Macht sich der Walk die geometrische und damit räumliche Struktur des Latent Space als Vektorraum zunutze, kann man überspitzt behaupten, der Zugriff auf neuere Modelle wäre ›semantisch‹[63].

Allerdings entsteht auch der Latent Space ›multimodaler‹ Modelle durch die strukturierte Abbildung großer Datenmengen auf einen Vektorraum. Damit stellt sich einerseits die Frage, ob der Zugriff via Prompt wirklich semantisch ist und ob Kosinus-Ähnlichkeit eine semantische Ähnlichkeit implizieren kann. Andererseits ist der geometrische Zugriff natürlich weiterhin möglich. Auch wenn gelten mag, »Latent spaces können nun gezielt befragt, statt bloß passiv durchschritten werden«[64], interessieren wir uns gerade für das passive Durchschreiten. Diese Art der Erkundung mag vielleicht nicht zu den gewünschten Entdeckungen führen, weil hochdimensionale Räume die Tendenz besitzen, ›unlesbar‹ zu werden.[65] Wir unternehmen sie dennoch, mit dem Yorkshire Terrier an der Leine.

61 Offert, Fabian (2023): »KI-Kunst als Skulptur«, in: Richard Groß/Rita Jordan (Hg.), KI-Realitäten, Bielefeld: transcript, S. 273–286, hier S. 280.

62 Ebd., S. 284.

63 Vgl. Offert, Fabian (in Vorbereitung): »Five Theses on the End of AI Art«, in: KITeGG (Hg.), Un/learn AI. Approaching AI in aesthetic practices, Mainz: HS Mainz.

64 Offert: KI-Kunst als Skulptur, S. 285.

65 Vgl. Offert, Fabian (2017): Intuition and Epistemology of High-Dimensional Vector Space. Online unter: https://zentralwerkstatt.org/blog/vsm (letzter Zugriff: 28.03.2024).

Gassi gehen mit dem Yorkshire Terrier

> »A man starts from a point O and walks I yards in a straight line; he then turns through any angle whatever and walks another I yards in a second straight line. He repeats this process n times.«[66]

Dieser zufällige Spaziergang ist seit 1905 unter dem Namen »random walk« bekannt. Ursprünglich formuliert als stochastisches Problem, sind Random Walks heute in der generativen Grafik ein beliebtes Mittel zur Erzeugung interessanter Strukturen. Das Verfahren eignet sich zur stichprobenartigen Erkundung eines unbekannten Raumes, weil es keine der möglichen Richtungen bevorzugt und nach unendlich langer Zeit auch wirklich überall vorbeikommen wird.

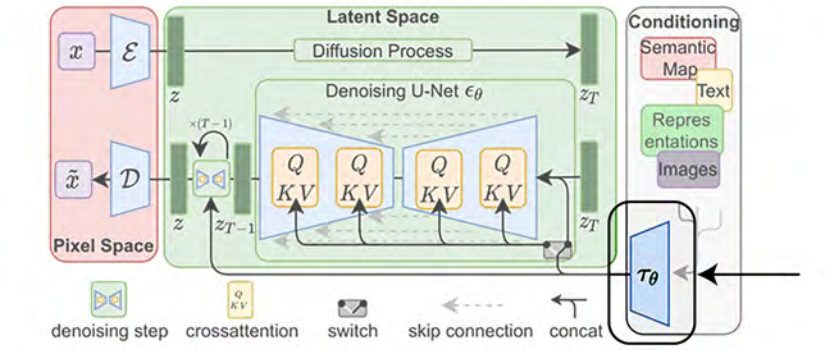
Random Walks mit dem Yorkshire Terrier lassen sich zum Beispiel in den Latent Spaces von Stable Diffusion durchführen. Zum einen im ›multimodalen‹ Raum von CLIP, zum anderen in dem Raum, in dem die Diffusion stattfindet, die die Bilderzeugung ermöglicht.

Das Gassi Gehen im CLIP-Raum (siehe Abb. 6a) startet wie eine Zero-Shot-Klassifikation mit dem Prompt »A photo of a Yorkshire Terrier«. Aufgrund der Funktionsweise von CLIP wird dieser in seine linguistischen Bestandteile (›token‹) zerlegt und dann in ein Embedding übersetzt, wobei jeder der 77 möglichen Token 768 Dimensionen besitzt. Dieses Embedding dient der ›sematischen‹ Lenkung der Bildgenerierung, die im Latent Space der Diffusion stattfindet und mit zufälligem Rauschen beginnt. Das Ergebnis ist ein Bild, das deterministisch aus dem gewählten Rauschen und dem Embedding des Prompts erzeugt wird.

Nun lässt sich das Embedding der Texteingabe rechnerisch manipulieren, indem seine Komponenten um kleine Zahlenwerte verändert werden. Wir bewegen uns also mit kleinen, zufälligen Schritten durch den Raum, den die möglichen Text-Bild-Kombinationen von CLIP erzeugen und erkunden diese ›Landschaft‹. Jeder Schritt wird dabei von einem Bild begleitet. Die resultierende Serie von Bildern lässt sich als Animation des Gassi gehens betrachten (siehe Abb. 6b). Sie startet mit typischen Darstellungen von Yorkshire Terriern, wie sie die Trainingsdaten von Stable Diffusion bestimmen. Sie bewegt sich von dort aus durch verschiedene Darstellungsformen, wobei die Bilder zunehmend abstrakter, bisweilen auch »trippy, surreal, psychedelic, painterly« werden. Die Walks enden nach einer festgesetzten Anzahl Schritte, oft mit Bildern, die keine Yorkshire Terrier mehr zeigen, wenig gegenständlich und insgesamt von Rauschen geprägt sind.

66 Pearson, Karl (1905): »The Problem of the Random Walk«, in: Nature 72, S. 294.

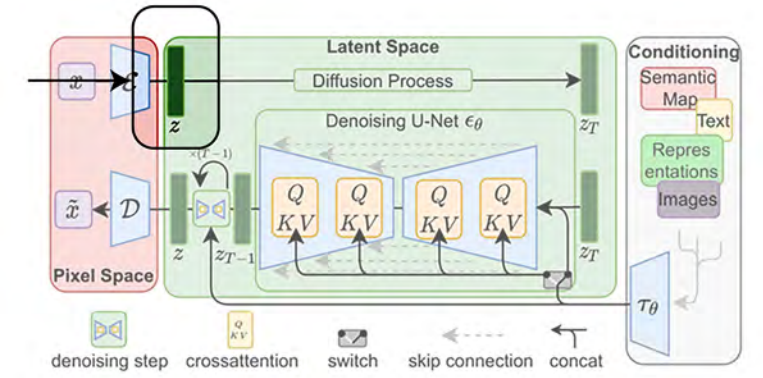
Wegstrecke 1: Zufälliges Gassi Gehen im Latent Space von CLIP
Abbildung 6a: Versuch 1, Position im Latent Space – Grafik: Diagramm mit Highlight, wo wir gerade sind; Abbildung 6b: Versuch 1, Bildfolge



Begeben wir uns mit den Yorkshire Terrier in den Latent Space, in dem Stable Diffusion die gelenkte Erzeugung von Bildern betreibt, ist der Start jedes Spaziergangs das Rauschen, von dem aus die Diffusion ihren Anfang nimmt (siehe Abb. 7a). Dieser Raum hat 30x64x64x4 Dimensionen und lässt sich nach dem gleichen Verfahren durchschreiten. Anders als im CLIP-Raum nehmen wir den Yorkshire Terrier hier enger an die Leine und bewegen uns auf kreisförmigen Pfaden. Die beschriebenen Kreise stellen dabei als loopbare Animationen stichprobenartige Erkundungen der möglichen Bildwelten dar, die ein einziger Prompt ermöglicht (Siehe Abb. 7b). Hier bleiben die Bilder im Bereich einer erwartbaren Bildlichkeit, ihre Motive verändern sich kontinuierlich von einer Darstellung eines Yorkshire Terrier zur

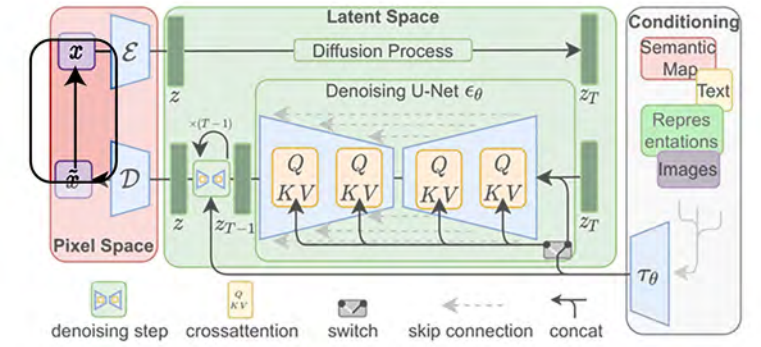
nächsten und zeigen dabei, wie hochgradig konventionalisiert diese Darstellungen sind – es ist zum Beispiel nie die Ober- oder Unterseite eines Hundes abgebildet. Weil CNNs grundsätzlich lokale Strukturen, wie Texturen oder lokale Formzusammenhänge besser abbilden als globale Zusammenhänge, zeigen auch diese Darstellungen überzeugendes Fell oder Schnauzen, scheitern aber häufig an der Anatomie der Tiere.[67]

Wegstrecke 2: Zirkuläres Gassi Gehen im Latent Space der Diffusion
Abbildung 7a: Versuch 2, Position im Latent Space – Grafik: Diagramm mit Highlight, wo wir gerade sind; Abbildung 7b: Versuch 2, Bildfolge



67 Vgl. Offert/Bell: Perceptual bias and technical metapictures, S. 1136.

*Wegstrecke 3: Gassi-Gehen mit Künstler*innen*
Abbildung 8a: Versuch 3, Position im Latent Space – Grafik: Diagramm mit Highlight, wo wir gerade sind; Abbildung 8b: Versuch 3, Bildfolge



Eine grundsätzliche Technik des Prompt-Engineering ist die Verwendung von Zusätzen, die die erzeugte Bildlichkeit prägen. In Foren, auf Discord-Servern und Websites sammelt und tauscht die Generative-AI-Community derartige Prompts wie Zaubersprüche, die bestimmte Stile aufrufen, zum Beispiel über die Namen von Künstler*innen. In pragmatischer Unkenntnis der Kunstgeschichte wird hier diesen Namen die Fähigkeit Landschaften oder Portraits darzustellen oder eine visuelle Verwandtschaft (wieder im Sinne der Kosinus-Ähnlichkeit) zu anderen Künstler*innen zugeschrieben. Aus diesen Listen suchen wir nach den Namen von Künstler*innen, deren generatives Potential in Stable Diffusion dem Yorkshire

Terrier Platz gibt, weil es zum Beispiel gerade keine Landschaften oder Portraits umfasst.

Ergänzt um diese Namen verändert sich so der Stil der Bilder, aber auch ihr Inhalt. Wir unterziehen die Bilder einem Animationsprozess, der jedes erzeugte Bild als Initialisierung des jeweils nächsten verwendet (siehe Abb. 8a). Gerade wenn die entsprechenden Kunstschaffenden nicht mit solchem Bildmaterial Teil der Trainingsdaten sind, wie es unseren Prompts entspricht, entfernen sich die Bilder schnell vom ursprünglichen Motiv (siehe Abb. 8b). »A photo of a Yorkshire Terrier by Francesca Woodman« wird so schnell zum Bild von Mensch-Tier-Chimären oder einfach von Menschen, im typischen schwarz-weiß-Stil der Fotografin, die schlicht und einfach keine Tiere fotografiert hat. Ausgehend vom Stil der Illustratorin Ida Rentoul Outhwaite landet der Spaziergang so schon nach wenigen Schritten beim Bild eines Kindes, das einen schwarzhaarigen Vogel in den Händen zu halten scheint.

Breeds, Brands und Design

> »The popular assumption is that latent spaces are ›filled‹ with images. This is false.«[68]

Wenn wir mit dem Yorkshire Terrier in den Latent Spaces von Stable Diffusion Gassi gehen, folgen wir der Prämisse, dass Latent Spaces keine Bildräume sind. Das ist einerseits der Fall, weil sie Kodierungen darstellen, die erst noch ins Bildliche dekodiert werden müssen. Vor allem aber, weil die Bilder, die sie scheinbar enthalten, Ergebnisse statistischer Verarbeitung sind und nur zeigen, was in den zugrundeliegenden Daten invariant ist. Sie sind Visualisierungen ihrer Trainingsdaten: »Every AI generated image is an infographic about the dataset.«[69] Für sie gilt, »the result is a ›mean image‹, a rendition of correlated averages«.[70]

François Chollet hat darauf hingewiesen, dass Deep Learning Datenpunkte in Strukturen überführt, die Abfragen ermöglichen, die zwischen diesen Punkten interpolieren.[71] Sie funktionieren damit wie Suchmaschinen oder Datenbanken, die

68 Offert: Five Theses on the End of AI Art.

69 Salvaggio, Eryk (2022): »How to Read an AI Image. The Datafication of a Kiss«, in: Cybernetic Forests. Online unter: https://cyberneticforests.substack.com/p/how-to-read-an-ai-image (letzter Zugriff: 26.03.2024).

70 Steyerl, Hito (2023): »Mean Images«, in: New Left Review 140/141, S. 82–97, hier S. 84.

71 Vgl. Chollet, François (26.8.2022, 15:15), Twitter: https://twitter.com/fchollet/status/1563153087514419206 (letzter Zugriff: 28.3.2024).

zugänglich machen, was zwischen den Daten steht. Andererseits lassen sich die Datenpunkte selbst praktisch nicht direkt abfragen – allenfalls dann, wenn einzelne Datenpunkte sehr oft in den Trainingsdaten dupliziert sind.[72] Es ist also nicht das Rauschen spezieller Bilder und einzelner Darstellungen, was hier abrufbar ist, sondern die stabilen Inseln im Meer dieses Rauschens, die durch die konventionalisierte Darstellung von immer wieder Demselben entstehen.

Die Bilder aus diesen Zwischenräumen repräsentieren vor allem die ihnen gemeinsame Wahrscheinlichkeitsverteilung. Als Ausdruck einer Klasse von Bildern, die die generative Computergrafik früher wie heute erzeugt, sind sie aber doppelt zu verstehen: Sie sind Variationen ästhetischer Phänomene, die einer zugrundeliegenden Wahrscheinlichkeitsverteilung entspringen, wie auch Gegenstände einer potentiellen Klassifikation, die dank der gemeinsamen Wahrscheinlichkeitsverteilung erst möglich ist und die in generativen Systemen seit DeepDream gewissermaßen rückwärts abläuft. Klassifizieren und Generieren wären damit zwei Seiten derselben Medaille.

Das Gassi gehen mit dem Yorkshire Terrier zeigt uns also nicht nur, dass bestimmte Kombinationen von Form und Inhalt schlicht und einfach nicht Teil des generativen Potentials von Stable Diffusion sind. Es zeigt auch die Konventionen, Kategorien oder Klassen der bildlichen Darstellung von Hunden, wie sie das westlich geprägte Internet bereithält. Wenn Generierung als Klassifikation rückwärts anzusehen ist, dann führt uns der Yorki durch die Welt einer Klassifikation als Yorkshire Terrier. Über das Klassifizieren von Tieren hält Roger Caillois fest:

> »Before classifying vertebrates as mammals, birds, batrachians, reptiles, or fish, they were grouped according to the number of feet they had. Horses were put into the same category as frogs and turtles. […] Nonetheless, it should be said that having four feet is an interesting feature as well, with certain specific and ineluctable consequences, which is almost eliminated as an object of study, though, by the new, improved taxonomy. Residual characteristics that have been legitimately disqualified surely give rise to remarkable relationships that are indubitably worth detecting and establishing. Even though they have been excluded, they are by no means insignificant. […] The universe is radiant. It supports any secant, median, chord, or bisectrix.«[73]

72 Vgl. Carlini, Nicholas/Hayes, Jamie/Nasr, Milad/Jagielski, Matthew/Sehwag, Vikash/Tramèr, Florian/Balle, Borja/Ippolito, Daphne/Wallace, Eric (2023): »Extracting Training Data from Diffusion Models«, in: arXiv. Online unter: http://arxiv.org/pdf/2301.13188v1 (letzter Zugriff: 28.03.2024); Somepalli, Gowthami/Singla, Vasu/Goldblum, Micah/Geiping, Jonas/Goldstein, Tom (2023): »Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models«, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, S. 6048–6058.

73 Caillois, Roger/Frank, Claudine/Naish, Camille (2003): The edge of surrealism. A Roger Caillois reader, Durham [N.C.], London: Duke University Press, S. 344–345.

Klassifikation ist also eine Dynamik des Ein- und Ausschlusses, die grundsätzlich kontingent ist. In ihr kommen die Eigenschaften der zu klassifizierenden Dinge genauso zum Ausdruck, wie die Eigenschaften des klassifizierenden Diskurses: »the classification of animals, like that of any group of significant objects, is apt to tell as much about the classifiers as about the classified«[74].

Es ist kein Zufall, dass der ILSVRC-Datensatz zahlreiche Hunderassen als Klassen enthält, aber keine Menschen.[75] Dies nicht nur, weil Hunde als »happy medium« einfach mehr Spaß versprechen oder weil das Einteilen von Hunden in Klassen mit sehr viel weniger politischen Problemen einhergehen mag. Vor allem ist es überhaupt möglich, Hunde einer »fine-grained object classification« zu unterziehen. Das liegt daran, dass ›Dog Breeds‹ im doppelten Sinne existieren: als Gegenstand von Klassifikation und Produkt einer vorangegangenen Standardisierung.

Die meisten ›Dog Breeds‹ entstanden in der viktorianischen Ära, als in Folge der Industrialisierung Hunde weniger als Arbeitstiere und mehr als Gegenstände von Wettkämpfen und Zucht eine Rolle zu spielen begannen.[76] ›Breed‹ ist hier gerade nicht als taxonomische Kategorie zu verstehen, vielmehr bezeichnet der Begriff »a division of a species with separate, distinctive characteristics that can be predicted genetically and reproduced consistently.«[77] Es geht also um äußerlich identifizierbare Eigenschaften, die sich durch Zucht herstellen und erhalten lassen. Als Unterscheidungskriterien sind sie nach Form, nicht nach Funktion differenziert und damit explizit ästhetisch.[78] Es kommt zu einer »predominance, even fetishization, of the notion of breed«, die mit einem »remodeling by selective breeding of dog varieties into standardized conformations« einhergeht.[79] ›Conformation‹ ist dabei als Verfahren zu verstehen, das ›breed‹ materialisiert und zu Gruppen vereinheitlicht.[80] Dabei werden Hunde derart re-modelliert, dass die Varianz innerhalb einzelner Klassen minimiert, die Abgrenzbarkeit der Klassen voneinander aber maximiert wird:

> »The difference between pre- and postbreed dogs can be compared to how colors appear in a rainbow versus on a modern paint sample card. The former has

74 Ritvo: The platypus and the mermaid, S. xii.

75 ImageNet verwendet dagegen zahlreiche Synsets, die Menschen beschreiben. Vgl. hierzu Crawford/Paglen: Excavating AI.

76 Vgl. Irvine, Leslie/Bekoff, Marc (2004): If you tame me. Understanding our connection with animals (= Animals, Culture, and Society), Philadelphia: Temple University Press, S. 54.

77 Ebd., S. 193.

78 Vgl. Worboys, Michael/Strange, Julie-Marie/Pemberton, Neil (2018): The Invention of the Modern Dog. Breed and Blood in Victorian Britain (= Animals, history, and culture), Baltimore: Johns Hopkins University Press, S. 8.

79 Ebd., S. 1.

80 Vgl. ebd., S. 2.

> distinct colors, but these vary in hue and shade, bleeding into one another where they meet. The latter consists of distinct, separate, and uniform blocks of color.«[81]

Im Zuge dessen wird die Wahrscheinlichkeitsverteilung des Erscheinungsbildes von Hunden weg von einer Normalverteilung in Richtung diskreter Klassen manipuliert.[82] Diese Standardisierung erfolgt nicht nur durch Vererbung und Selektion, gerade beim Yorkshire Terrier wird diesen Prozessen auch manuell nachgeholfen, durch die Wahl der Ernährung, aber auch durch das Stutzen von Frisur, Schwanz und Ohren.[83]

Wie auf dem Weg zur Deep-Learning-Revolution, ist das Format, das diese Entwicklung strukturiert, der Wettbewerb. Bei »dog shows« wird die Standardisierung der Breeds geprüft und fortgeschrieben. Sie erfolgt entlang quantifizierbarer Eigenschaften (genannt ›points‹), die sich aus ästhetischen Kategorien speisen, aber auch aus dem evolutionsbiologischen und sozio-politischen Denken der Zeit. »Breed and breeds were examples of order and hierarchy, and of tradition, even though much, if not all, were invented.«[84] Die quantifizierbaren Punkte der Unterscheidung sind dabei vor allem visuell und standardisierte Breeds funktionieren als »brands« oder »designs«.[85]

Resultat dieser Standardisierung ist, dass sich die quantifizierbaren Eigenschaften von Hunden in die Bildkultur ihrer Darstellung einschreibt und von hier aus in Datensätze wie den des ILSVRC gelangen. Zusammen mit der Diskretisierung der Breeds wird so die feinkörnige Klassifizierung von Hundebildern erst möglich. Ins Generative gekehrt visualisiert DeepDream dann die Klassifizierung von Hunden aus einer bestimmten Epoche der Entwicklung der westlichen Wissenschaft und Gesellschaft durch die Linse automatisch aus dem Internet extrahierter und manuell kategorisierter Bilder. Dabei gelangen dieselben Standards über Datensätze wie LAION auch in die heutigen ›multimodalen‹ Modelle. So liefert der Prompt »Yorkshire Terrier« auch hier immer eine ›Infografik‹ zu den möglichen Bildkompositionen, die mit dieser Marke verknüpft sind.

Wie beim Gassi gehen mit dem Yorki gesehen, rekombinieren generative Modelle dabei erfolgreich die Punkte der Hundezuchtwettbewerbe als lokale ästhetische Eigenschaften, verlieren dabei aber deren anatomischen Zusammenhang aus dem Blick. Das wundert nicht, sind die Punkte doch gegen die Bewegung der Tiere und ihre Darstellung im Bildraum invariant, während die Anatomie das nicht ist. Zugleich zeigt sich, dass im Latent Space dieser Systeme die Überschneidungen

81 Ebd.
82 Vgl. ebd., S. 8.
83 Vgl. ebd., S. 142–143.
84 Ebd., S. 7.
85 Ebd., S. 10–14.

möglicher Darstellungen von Yorkshire Terriern und möglicher Bilder einer Francesca Woodman oder Ida Rentoul Outhwaite sehr klein oder überhaupt nicht vorhanden sind. Ähnliches gilt für mögliche Bildkompositionen und Blickwinkel. Mit dem Yorki an der Leine treffen wir also auf das, was ein Optimierungsverfahren, das auf die Extraktion des Gemeinsamen, Häufigen und Standardisierten hinausläuft, nicht abbilden kann.

Fazit

Anders als DeepDream versuchen aktuelle generative Modelle nicht alles als Hundeschnauze darzustellen. Da sie aber auf den visuellen und textuellen Invarianten bestimmter Bild- und Textkulturen basieren, bringen sie selbstverständlich einen Bias für Bildinhalte und Kompositionen mit sich. Wie unser Gassi gehen zeigt, stehen sie darüber hinaus in der Tradition der CNNs und besitzen einen Fetisch für lokale Texturen und scheitern an globalen Zusammenhängen. Vor allem aber zeigen die Spaziergänge ihre Unfähigkeit, solche Inhalte mit anderen Inhalten oder Formen zu kombinieren, die in ihren Trainingsdaten selten sind. Der Latent Space scheint hier zwar »Orte« zu enthalten, die als Interpolation zwischen den Bildern der Künstler*innen und den Bildern von Yorkshire Terriern liegen – diese Orte sind oft aber instabil und es dominiert entweder das künstlerische Oeuvre oder die hochgradig konventionalisierten Darstellungen der Yorkshire Terrier (siehe Abb. 9).

Das generative Potential dieser Modelle erweist sich damit als ein Resampling kultureller Räume. Sie generieren Assemblagen der (Textur-)Eigenschaften konventionalisierter Darstellungen der (via Prompt) angefragten Bildsujets. Unser Gassi gehen zeigt eine derartige Konventionalisierung, deren Ursprung in der Standardisierung von Hunderassen im 19. Jahrhundert entlang (vorrangig) visueller Eigenschaften liegt. Es visualisiert die Wahrscheinlichkeitsverteilung möglicher Darstellungen von Yorkshire Terriern, die als Möglichkeit der Klassifikation von Hunderassen hier ins Generative gekehrt wird. Dabei ist es kein Zufall, dass ausgerechnet Bilder von Hunden die Geschichte ›generativer Künstlicher Intelligenz‹ permanent begleiten, denn hier kommt eine hochgradig standardisierte Bildkultur mit einer ebenso standardisierten Klassifikation und Nomenklatur zusammen, womit exemplarisch gezeigt wird, was den Kern generativer Verfahren ausmacht. Generiert werden kann, was bekannt ist und ›erkannt‹ werden kann – und selbst dabei wird das Fell eines Yorkshire Terriers meist besser getroffen als dessen Anatomie.

> »Those of us who have been lucky enough to share our lives with a Yorkie know there's a lot more to this breed than meets the eye.«[86]

86 Dog Fancy Magazine: Yorkshire terrier, S. 4.

Abbildung 9: Yorkies take cuteness to a whole other level.