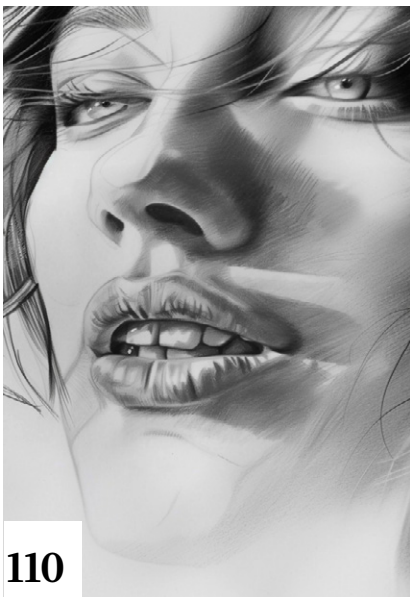
Inhalt

118

80

114



110

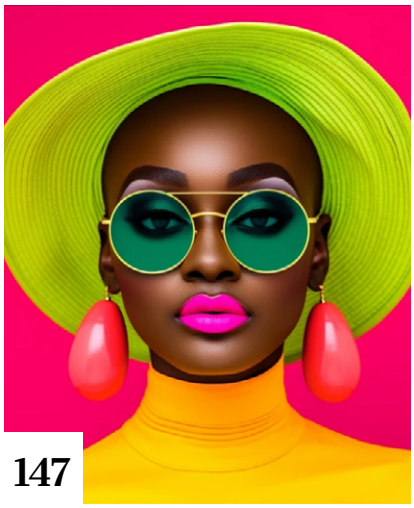
166


52


56


126


91


147


71

Die Gefühle sind echt, die Bilder nicht

Selbstvergessene Tänzer, Modebilder mit Waffenaccessoires, dicke Engel auf Reisen, reiche Kinder auf Sylt oder Männer mit Phallussymbolen – die Bilderwelt der Nina Puri ist vielschichtig. **Christoph Künne** hat mit ihr über ihre KI-Erfahrungen gesprochen.

In der schier unerschöpflichen KI-Bilderflut bei Instagram stechen Nina Puris ungewöhnliche Bildserien heraus. Ihre – auf den ersten Blick etwas bizarr wirkenden – Motive sind anders, scheinen dennoch alltäglich. Sie wollen nicht perfekt sein, trotzdem erscheinen sie auf eine eigenartige Weise wahr. Und das, obwohl sie spätestens beim zweiten Blick ihren künstlich-intelligenten Ursprung nie ganz verbergen. Um hinter das Geheimnis dieser Bilder zu kommen, haben wir uns mit ihrer Schöpferin in Hamburg getroffen.

DOCMA: Sie sind eine hochdekorierte Werbetexterin und mehrfache Bestseller-Autorin. Was hat Sie zur Bild-KI gebracht?
NINA PURI: Als studierte Grafikdesignerin habe ich einen engen Bezug zu grafischen Themen. Beruflich bin ich in der Werbung bereits früh in Richtung Text abgebogen, weil Texter am Berufsanfang stärker in die konzeptionellen Überlegungen eingebunden waren als Grafiker. Und ich fand Konzeption schon immer am spannendsten.

DOCMA: Und die Grafik ist dann hinten heruntergefallen?
NINA PURI: Nicht direkt. Alle meine Konzepte konnte ich auch mit einem Haufen Skizzen visualisieren. Diese Liebe zu Bildern, die mit Texten in Verbindung stehen, hat mich vermutlich auch zu den ersten KI-Experimenten angeregt. ▶







287

8819



DOCMA: Genau wissen Sie es nicht?
NINA PURI: Nein, ich kann mich nur gut daran erinnern, dass ich recht früh mit Dall·E experimentiert habe. Anfangs faszinierte mich dessen etwas holziger Stil. Aber nach ungefähr einem Monat hatte ich genug davon und bin zu Midjourney übergelaufen. Da waren die Bilder dann nicht mehr so schräg, aber dafür berücksichtigten sie mehr von meinen Textvorgaben.

DOCMA: Wenn man wie Sie schon beruflich ein so ausdifferenziertes Verhältnis zur deutschen Sprache hat, ist es dann nicht besonders schwierig, die Nuancen ins Englische zu übertragen?
NINA PURI: (lacht) Vielleicht ist es das, aber mir als gebürtiger Engländerin fällt das nicht so auf – auch wenn ich in England nur neun Jahre lang gelebt habe.

DOCMA: Als ich Ihre Bilder zum ersten Mal sah, habe ich durch die Themenbearbeitung mit jeweils mehreren Motiven erwartet, es bei Ihnen mit einer Künstlerin zu tun zu haben.
NINA PURI: Das Gestalten in Serie ist vermutlich dem Kampagnen-Denken der Werbung geschuldet. Dort beweist unter anderem die mögliche Vielfalt von Variationen die Tragfähigkeit eines Konzepts. Da gibt es sicher eine Ähnlichkeit mit den für Künstler typischen Werkzyklen.

DOCMA: Wie würden Sie Ihre eigenen Arbeiten beschreiben?
NINA PURI: Mir liegt nicht daran, technisch bis ins letzte Detail ausgefeilte Bilder zu produzieren. In meinen Motiven entsteht die Wirkung eher durch die sinnhafte Verknüpfung von Bild und Text. Leider gehen die Texte zu den Bildern bei der Präsentation auf Instagram oft unter.

DOCMA: Was ist Ihre inhaltliche Zielsetzung?
NINA PURI: In der KI wimmelt es von Stereotypen – zum Beispiel, was Frauen- und Männerbilder angeht. Es gibt haufenweise Abbildungen von Frauen, die nichts machen außer da zu ein und gut auszusehen. Und Männer, die starke Macher sind. Ich möchte aber echte Wesen zeigen. Authentische menschliche Gefühle in all ihrer Vielfalt, Tragik und Komik aufgreifen und erzeugen, nicht nur Abziehbilder reproduzieren. Ich frage mich immer: Was ist das Wahre an einem bestimmten Thema? Was ist das wirkliche Gefühl, das hinter dem Bild steht? Kann man das erkennen und woran muss ich noch arbeiten, damit es klarer herauskommt?

DOCMA: Gibt es Hürden, auf die Sie beim Prompten regelmäßig stoßen?
NINA PURI: Die visuelle Perfektion, die bei statischen Motiven leicht erreichbar ist, verschwindet, wenn die Figuren auf den Bildern etwas Bestimmtes tun sollen. Und zwar immer mehr, je ungewöhnlicher die Kombination von Figur und Handlung ist. Nehmen wir zum Beispiel einen Marathonläufer, der strickt oder Kaffee trinkt. Auch eine Frau die Autoreifen wechselt, macht mehr Probleme, als eine Frau mit Katze auf einem Bett. Ohnehin ist es schwer, der KI halbwegs realistische Frauenfiguren abzuringen, ohne dass sie daraus harsche, übergewichtige oder alte Witzfiguren macht. Ich habe für mich die Formel entdeckt: Je weniger Klischees, desto schlechter die Details.

DOCMA: Sehen Sie im Prompten für sich selbst als Grafik-affine Texterin einen neuen Geschäftszweig?
NINA PURI: Beruflich bekomme ich erste Anfragen. Inzwischen laufen diese über meine Repräsentanz yesweprompt.de. Aber in der Praxis erweist es sich noch als etwas schwierig, wenn die Kunden und Kundinnen ganz konkrete, festgelegte Vorstellungen haben, denn so klare und präzise Umsetzungen wie mit Fotografie, Photoshop und CGI kann man mit KI bisher nicht erreichen. Manches lässt sich heute einfach noch nicht zuverlässig prompten, Looks sind möglich, aber nicht unbedingt in jedem Detail vorhersehbar. Es ist eher so ein wenig wie früher, als die Vorgaben bei Kampagnen weniger präzise waren und die Freiheit der Ausführenden größer.

Foto: Ralf Nolting

NINA PURI

... ist freie Texterin, Kreativdirektorin, Autorin, Dozentin, Diplom-Grafikdesignerin, Mitglied im Art Directors Club Deutschland und arbeitet seit vielen Jahren für Agenturen, Verlage und Direktkunden. Sie hat viele Peise gewonnen – unter anderem Cannes Lions, D&AD, The New York Festivals und ADC Germany, Bislang sind sieben Bücher von ihr zu sehr unterschiedlichen Themen erscheinen, darunter einige Bestseller.

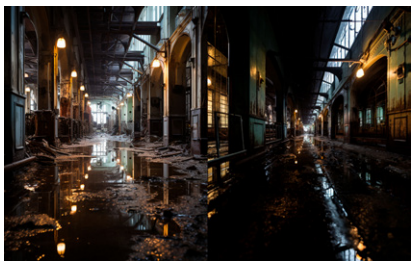
Mehr Infos:
www.ninapuri.de

Wer KI-Jobs beauftragt, muss also viel mehr experimentelle Bandbreite zulassen. Das birgt die Chance auf kreativeres Arbeiten.

DOCMA: Lassen Sie uns zum Abschluss noch einen Blick in die Glaskugel werfen: Welche Zukunft prognostizieren Sie den Kreativen?
NINA PURI: Ersetzt werden vor allem die 08/15-Sachen und damit die Kreativen, die sich vor allem als Auftragnehmer sehen. Diejenigen, die eigene Ideen haben, können in der KI einen Sparringspartner finden. Einen, der allzeit bereit ist und beim Ping-Pong zwischen der ersten Idee und der finalen Auswahl aus dem Berg der entstandenen Variationen hilft. Meist sagen die Bilder-Feeds von KI-Kreativen ähnlich viel über ihre Persönlichkeit aus wie Playlists oder ein Bücherregal. Die einen (re-)produzieren Klischees ohne eigene Note, die anderen entwickeln aus den Möglichkeiten einen eigenen Stil. Gewinnen werden am Ende die, deren Arbeiten man länger folgen will.

DOCMA: Vielen Dank für das offene Gespräch. ■

VIDEO 1 (9 min) *Bilder per KI auf der Seite nightmare.ai kostenlos vergrößern*



VIDEO 2 (9 min) *Neue Funktionen in Midjourney 5.2: Variationsstärke und Zoom out*

VIDEO 3 (13 min) *Alte Bilder mit der neuen Version variieren.*

VIDEO 4 (14 min) *Promptverkürzung und Prompt-Analyse mit dem »/shorten«-Befehl*

VIDEO 5 (10 min) *Midjourney-Bilder mit der generativen Füllung in Photoshop retuschieren*

DIE VIDEOS UND WEITERE BILDER FINDEN SIE UNTER
www.docma.info/22842

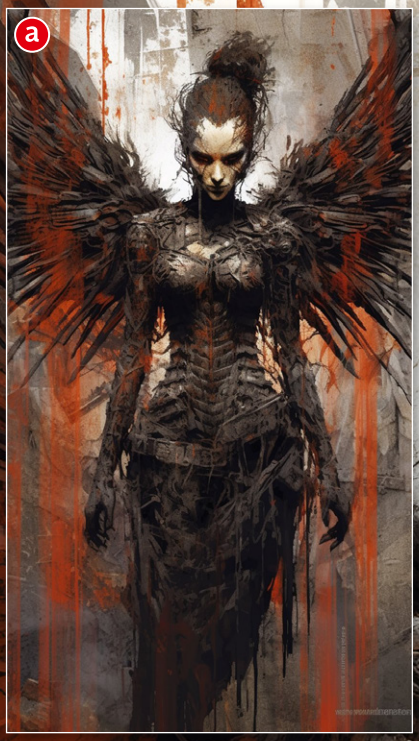
+ Video

MIDJOURNEY 5.2 UND TIPPS & TRICKS

Per KI von Text zu Bild

Die künstliche Intelligenz Midjourney erzeugt Bilder anhand von Texteingaben. **Peter Braunschmid** zeigt in fünf Videos, was es in der Midjourney-Version 5.2 Neues gibt und wie Sie die neuen Funktionen in Ihren Workflow integrieren können.

Mit keiner anderen KI erhalten Sie so einfach ansehnliche Ergebnisse wie mit Midjourney. Die Version 5.2 legt die Messlatte wieder ein Stück höher. Die neuen Funktionen »*Vary*«, »*Zoom out*« und »*Shorten*« helfen beim Optimieren der Bildergebnisse durch kreatives und analytisches Vorgehen. In seinen Videos bringt Peter Braunschmid Sie auf den neuesten Stand seiner Arbeitsweise mit Midjourney. In Video 1 stellt er – als Alternative zum kommerziellen Branchenführer Topaz Gigapixel AI – einen Online-Dienst vor, mit dem Sie Ihre Bilder KI-basiert vergrößern können. In Video 5 erfahren Sie, wie Sie mit Midjourney erzeugte Bilder in der Photoshop Beta retuschieren. In diesem Fall wiederum mithilfe von künstlicher Intelligenz – dank der generativen Füllung. ▶



c

Das in Midjourney generierte Bild ***(a)*** *ist hier mit einer rechnerischen Auflösung von circa 390 ppi abgedruckt. Auf Seitengröße skaliert, hätte es eine wahrnehmbar schlechtere Auflösung von etwa 100 ppi – siehe Ausschnitt* ***(b)****.*

Die über die Seite nightmare.ai hochskalierte Variante ***(c)*** *hat im Layout-Programm in der vollen Seitengröße der DOCMA noch eine effektive Auflösung von circa 970 ppi – also mehr als genug Potenzial für jede gewünschte Anwendung.*

Wie Sie Bilder am besten hochskalieren, erfahren Sie in Video 1*.*

*Eine neue Funktion von Midjourney 5.2 namens »Zoom out« erweitert ein generiertes Bild (**a**) an den Bildrändern, als würden Sie mit einem Zoom-Objektiv herauszoomen (**b**). Damit lässt sich virtuell um das 1,1- bis 2-Fache herauszoomen. In KI-Sprech handelt es sich dabei eigentlich um schlichtes Outpainting – bei dem Midjourney aber Können beweist. Es gibt jedoch ein Problem dabei: Midjourney vergrößert ein Bild dabei nicht, sondern verkleinert es und strickt erst dann die neuen Randbereiche an. Die quadratische Standardbildgröße von 1024 × 1024 Pixeln wird also nicht überschritten.* ***Mehr dazu erfahren Sie in Video 2.***

*Aus einem zuvor mit einem eigenen Prompt generierten Bild (**c**) wurde mit »Describe« von Midjourney ein Prompt (**d**) erzeugt. Dieser ließ sich mithilfe der neuen »Shorten«-Funktion analysieren (**f**), um aus den resultierenden Prompts erneut Bilder wie das nebenstehende (**e**) zu generieren. (og)* ■

Glühende Objekte

Um glühende Objekte mit einer Kamera aufzunehmen, ist zumeist ein großer Aufwand bei der Lichtgestaltung oder alternativ viel Montagearbeit in Photoshop nötig. KI kann die Arbeitswege erheblich verkürzen.

Im Prinzip eignet sich alles für das Themenfeld glühende Objekte, was nicht sonderlich lichtdicht ist: Pflanzen, Weichtiere, Gläser oder leichte Stoffe. Alternativ kann man auch Lichtundurchlässiges mit Schwarzlichtfarben zum Leuchten bringen.

Fangen wir gleich einmal mit einer Übung an, die sich so nur sehr schwer in der Fotografie umsetzen lässt: Glühende Quallen. Am besten in der Ästthetik einer Studiofotografie.

Ups, das war wohl etwas zu viel des Guten. Es sollte ja nur eine Qualle im Dunklen aufs Bild kommen.

Prompt ►
studio photography, glowing jellyfisch

Also noch einmal. Und nun bleiben wir fotografisch vielleicht besser dort, wo Quallen unter realen Bedingungen auch zu finden sind. Das ist dann deutlich näher am Thema.

◀ *Prompt*
underwater photography, glowing jellyfish

Wenn man sich einfach auf Fische im Plural und ohne erweiterte Spezifikation konzentriert, dann ist das Ergebnis ebenfalls recht ansehnlich. ▶

◀ *Prompt*
black and white drawing

Gehen wir nun an Land: Ein besonders beliebtes fotografisches Motiv glühender Objekte sind Pilze, weil man sie vergleichsweise simpel mit einer kleinen Lampe, ein paar Langzeitbelichtungen und Photoshop umsetzen kann. Einfacher und auch handwerklich weniger herausfordernd ist natürlich ein Textprompt. Vor allem, wenn der illuminierte Kunstwerke erzeugt, die wenig mit dem zu tun haben, was die meisten Pilzbeleuchter umsetzen können.

Prompt ►
nature photography, strong glowing mushroom by night

Mit wenigen kleinen Textänderungen werden aus unseren Waldpilzen interessante urbane Mutationen. Aber die haben mit dem Natur-Thema natürlich nichts zu tun. Interessant sehen sie trotzdem aus.

Prompt ►
photography, neon glowing mushroom in asian cityscape

Wendet man das innere Leuchten auf Blumen an, entstehen leicht Motive mit „Das-kann-man-sofort-an-die-Wand-hängen“-Charakter. Um die Plakativität der Motive zu steigern, hilft (mal wieder) die Begriff „art photography“.

◀ *Prompt*
art photography, glowing flower

Eine visuell oft interessantere Leuchtbefehl-Alternative ist der Begriff „bioluminiszierend“, die wir hier zum Abschluss mal auf einen Blumenstrauß anwenden wollen. ■

◀ *Prompt*
art photography, bouquet of bioluminiscent flowers

Bild: Christoph Künne mit Midjourney

DIE FUNKTIONSWEISE VON TEXT-ZU-BILD-SYSTEMEN

Wo kommen die Bilder her?

Generative Systeme wie Midjourney, Stable Diffusion und Firefly sind ein offenbar unerschöpflicher Quell immer neuer Bilder. Michael J. Hußmann erklärt, wie sie entstehen.

Wenn Text-zu-Bild-Systeme noch die abseitigsten Prompts halbwegs passend visualisieren, fühlt man sich an einen Zaubertrick erinnert: Ein Illusionist zieht ein Kaninchen nach dem anderen aus seinem Zylinder, in dem unmöglich genug Platz für all die Tiere gewesen sein kann. Wo nehmen Midjourney & Co. ihre Bilder her? Diese KI-Systeme haben nicht einmal einen Speicher, in dem sie enthalten gewesen sein könnten.

Die generative Magie basiert auf künstlichen neuronalen Netzen, also einer stark vereinfachten Simulation eines Nervensystems. Die simulierten Nervenzellen (Neuronen) haben Eingänge, über die sie mit den Ausgängen anderer Neuronen verbunden sind. Wenn die Summe der Werte an den Eingängen einen Schwellwert überschreitet, „feuert“ das Neuron und gibt einen Wert an seinem Ausgang aus, der wiederum mit den Eingängen weiterer Neuronen verknüpft ist. Die neuronalen Netze der KI bestehen heutzutage aus Millionen solcher simulierten Neuronen und einer noch viel größeren Zahl von Verbindungen zwischen ihnen. Generative KI-Systeme enthalten typischerweise mehrere Schichten von Neuronen: Eine Input-Schicht, in die ein Prompt-Text eingespeist wird, weitere Schichten für die eigentliche Verarbeitung und schließlich eine Output-Schicht, die das Bildergebnis erzeugt.

Was das neuronale Netz tut, wird durch seine Verschaltung bestimmt, also vor allem durch die Stärke der Verbindungen zwischen den Neuronen. Diese Variablen bilden gewissermaßen die Software der KI und werden als sogenanntes *Modell* gespeichert. Schaut man sich die Größe der Modelldateien an, dann sind sie viel kleiner als der Korpus aus Milliarden von Bildern, mit dem die KI für ihre Aufgabe trainiert worden war, aber auch kleiner als der Strom von Bildern, der daraus generiert werden kann. Wie bei den aus dem Hut gezogenen Kaninchen muss ein Trick hinter der generativen KI stecken, aber wie funktioniert er?

Der entscheidende Punkt ist, wie man Bilder im Computer repräsentiert. Wir kennen vor allem Bitmaps, also Bilder, die aus Zeilen und Reihen von Pixeln bestehen. Für die Zwecke der generativen KI möchte man die Bilder dagegen so repräsentieren, dass ähnliche Bilder auch als ähnliche Folgen von Bits und Bytes gespeichert werden. Mit RGB-Bitmaps, wie wir sie in Photoshop bearbeiten, funktionierte das nicht: Wenn wir mehrere Fotos einer Katze vergleichen, die sie von vorne, von hinten, von der Seite und mit unterschiedlichen Körperhaltungen zeigen, dann sind sich die RGB-Bilder nicht besonders ähnlich. Die Pixel für Pixel berechnete Differenz kann auch mal größer als die zwischen dem Bild der Katze und dem eines Hundes ähnlicher Fellfarbe sein. Statt durch die Farbe der einzelnen Pixel kann man die Bilder aber ▶

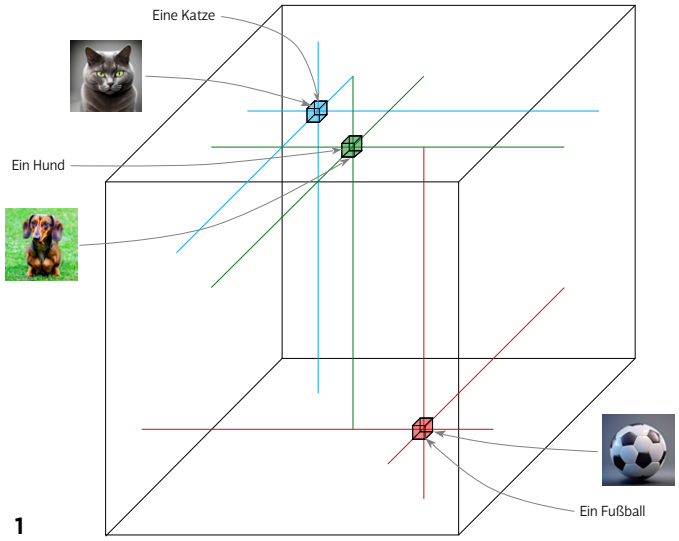
durch ihre Merkmale beschreiben. Für eine ausreichend treffgenaue Beschreibung sind Tausende, wenn nicht Millionen von Merkmalen nötig, aber für ein neuronales Netz ist es keine Herausforderung, Bilder darin umzurechnen. Für diese Aufgabe muss es lediglich mit Milliarden von Bildern trainiert werden. Das Netz liefert dann zu jedem Bild die Werte seiner Merkmale, also Koordinaten in einem vieldimensionalen Merkmalsraum. Die Merkmale können beispielsweise *Alter*, *Geschlecht*, *Farbe* und *Aggregatzustand* (sofern jeweils anwendbar) sein, aber auch Eigenschaften, die das neuronale Netz im Training selbst entdeckt hat.

Der Name für die so gefundenen Koordinaten lautet *Embedding*, aber den vergessen Sie am besten gleich wieder. In der KI werden alltäglich gebrauchte Wörter wie *Embedding* (*Einbettung*) oder *Attention* (*Aufmerksamkeit*) als Fachbegriffe in einer ganz anderen als der vertrauten Bedeutung verwendet. Außerhalb der Wissenschaft verwirrt das nur, weshalb ich diese Wörter hier vermeide.

Es macht übrigens nichts, wenn Sie sich einen vieldimensionalen Raum nicht anschaulich vorstellen können – das kann vermutlich niemand. Wichtig zu wissen ist nur, dass ähnliche Bilder ähnliche Merkmale haben und daher im Merkmalsraum benachbart sind. Sie liegen sich darin um so näher, je mehr sie sich ähneln. Alle Bilder derselben Katze finden sich in einem eng begrenzten Gebiet, umgeben von einer Zone mit Bildern anderer Katzen, während die Hundebilder in einer anderen, weiter entfernten Zone liegen. Bilder von Buschwindröschen, Einfamilienhäusern und Fußbällen liegen wiederum ganz woanders. Da das neuronale Netz im Training lernen musste, zu verallgemeinern, funktioniert das meist auch mit Bildern, die es noch nie zuvor gesehen hat.

Die Bilder sind zwar nach ihrer Umwandlung in Koordinaten im Merkmalsraum verschwunden, aber diese Form der Repräsentation erlaubt bereits einige nützliche Anwendungen. Inhaltsbasierte Vergleiche zweier Bilder sind trivial: Man ermittelt ihre Koordinaten und berechnet deren Abstand, der ein Maßstab für ihre Ähnlichkeit ist. Anhand seiner Koordinaten kann man auch leicht herausfinden, ob ein Bild eine Katze, einen Hund, ein Buschwindröschen oder einen Fußball enthält, da man die Bereiche des Merkmalsraums kennt, in dem Bilder solcher Motive liegen.

Noch interessanter wird diese Art der Repräsentation, wenn man zum Training Bilder mit einem deren Inhalt beschreibenden Text verwendet. Dann kann man neben einem neuronalen Netz für die Bilder ein zweites für die Texte trainieren – und zwar so, dass den Texten dieselben oder zumindest sehr ähnliche Koordinaten im Merkmalsraum wie den beschriebenen Bildern zugeordnet werden **[1]**. Auch das neuronale Netz für die Texte muss dabei lernen, zu



1

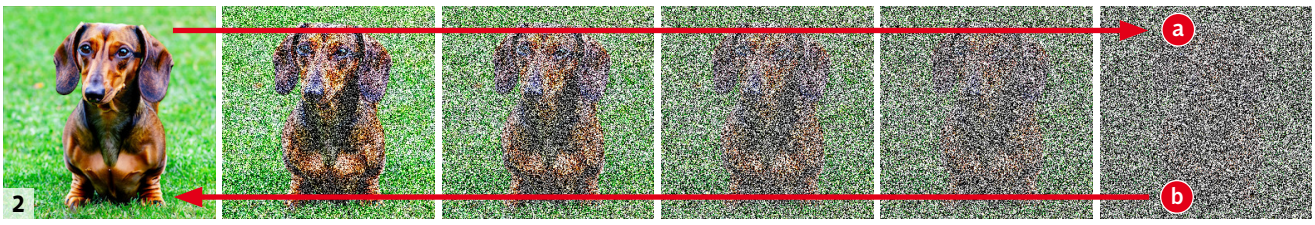
Neuronale Netze können darauf trainiert werden, Bilder und ihre Beschreibungen an denselben Koordinaten eines multidimensionalen Merkmalsraumes zu platzieren.

verallgemeinern, wird also auch für solche Texte passende Koordinaten finden, die im Trainingskorpus nicht vorkamen.

Man könnte denken, dass sich damit bereits ein Text-zu-Bild-System verwirklichen ließe, da man die zu einem Prompt zugehörigen Koordinaten berechnen kann, die gleichzeitig die Koordinaten eines so beschriebenen Bildes sind. Umgekehrt ließe sich zu einem Bild eine passende Bildbeschreibung finden. Aber ganz so einfach ist es nicht.

Aus einem Text können wir zwar Koordinaten berechnen, die gleichzeitig die Koordinaten der gewünschten Bilder sind. Diese Bilder müssen aber erst noch aus den Koordinaten berechnet werden, während wir bislang nur ein Verfahren für den umgekehrten Weg haben, also zur Berechnung von Koordinaten aus Bildern. Dazu ist ein drittes neuronales Netz nötig, das zur Bildgenerierung ein Diffusionsverfahren anwendet. Nach dieser Methode der Diffusion ist Stable Diffusion benannt, aber sie liegt auch anderen Systemen wie Midjourney und Firefly zugrunde.

Das Diffusionsverfahren soll aus dem Nichts beliebige Bilder erzeugen, und „Nichts“ heißt in diesem Fall: Rauschen. Um ein neuronales Netz darauf zu trainieren, im Rauschen Bilddetails zu erkennen, geht man zunächst den umgekehrten Weg und fügt Bildern schrittweise immer mehr Rauschen hinzu, bis sie nur noch aus Pixeln mit zufälligen Farb- und Helligkeitswerten bestehen. Das neuronale Netz muss dann lernen, die verrauschten Bilder wieder zu entrauschen, also jeweils das Bild aus dem Schritt davor zu rekonstruieren, das noch weniger Rauschen enthielt **[2]**.



2

Illustrationen auf dieser Seite: Michael J. Hußmann

*Nachdem man den Bildern im Trainingskorpus schrittweise immer mehr Rauschen hinzugefügt hat (**a**), kann man ein Diffusion-Modell darauf trainieren, verrauschte Bilder in der umgekehrten Richtung Schritt für Schritt zu entrauschen (**b**).*

Mit dem austrainierten Diffusions-Netz kann man dann aus reinem Rauschen ein Bild berechnen, in dem sich bereits erste Motivdetails erahnen lassen. Diesem wird erneut ein wenig Rauschen hinzugefügt, woraufhin der Vorgang wiederholt wird. Schritt für Schritt treten die Motive klarer hervor, und am Ende erhält man ein rauschfreies, detailreiches Ergebnis. Auf diesem Wege entstehen allerdings völlig beliebige Bilder, während ein Text-zu-Bild-System ja ganz bestimmte Bilder generieren soll – nämlich solche, die zu einem vorgegebenen Prompt passen. Dieser muss also die Bildgenerierung in die gewünschte Richtung steuern.

Beim Trainieren des Diffusions-Netzes gehört deshalb neben dem verrauschten Bild auch dessen Beschreibung zum Input – oder vielmehr deren Koordinaten im Merkmalsraum. Das Netz soll lernen, das verrauschte Bild nicht unbedingt im Sinne der vorgegebenen weniger verrauschten Version zu verbessern, sondern nur, wenn ein entsprechender Text als Prompt vorliegt. In der Praxis zeigt sich, dass dies als Steuerung der Bildgenerierung noch nicht ausreicht, aber hier hilft ein Kniff, den Einfluss eines Prompts zu vergrößern. Dazu lässt man das neuronale Netz zwei entrauschte Bilder berechnen, eines mit Berücksichtigung des Prompt und eines ohne. Die Differenz beider Bilder zeigt dann an, in welche Richtung der Prompt die Generierung steuert, und wenn man diese Differenz mit einem Faktor multipliziert und dem ohne Prompt generierten Bild hinzufügt, wird die steuernde Wirkung des Prompt verstärkt. Dieser Faktor ist der *Guidance*-Parameter, den man in manchen Text-zu-Bild-Systemen frei wählen kann.

Das Verfahren, mit dem der Prompt die Bildgenerierung anleitet, ist ein wichtiger Faktor, der die Qualität der Ergebnisse bestimmt. Daher experimentieren die Hersteller generativer KI-Systeme mit verschiedenen Guidance-Strategien, um die typischen Fehler solcher Systeme (siehe Promptologie 1 ab Seite 44) künftig zu vermeiden.

Prinzipiell wäre es auf diesem Wege auch möglich, die Erzeugung überzähliger Gliedmaßen zu vermeiden, aber abgesehen davon, dass sich die neuronalen Netze mit dem Zählen von Elementen schwer tun, müssten dazu stark verrauschte Bilder analysiert werden, wenn die Generierung rechtzeitig in die richtige Bahn gelenkt werden soll. Die entscheidenden Details zeichnen sich in dieser frühen Phase allerdings noch kaum im Rauschen ab.

Ein aktuell verfolgter Ansatz beruht darauf, Bilder nach ihren ästhetischen Qualitäten von Menschen bewerten zu lassen und mit diesen Daten ein neuronales Netz zu trainieren, das die erlernten ästhetischen Maßstäbe an die Zwischenstufen zum fertigen Bild anlegt und so die Generierung steuert. Leider ist es nicht praktikabel, Menschen unmittelbar in den Trainingsprozess einzubinden, da dieser Milliarden von Durchläufen umfasst und mit menschlichen Eingriffen viele Jahre in Anspruch nähme.

Das vom Diffusion-Netz erzeugte Bild hat schließlich eine Auflösung von beispielsweise 64 × 64 oder 128 × 128 Pixeln, mit der man in der Praxis noch wenig anfangen könnte. Eine von vornherein höhere Auflösung würde derzeit noch zu viel Rechenleistung beanspruchen. Daher muss ein weiteres neuronales Netz das Bild erst noch auf eine sinnvoll verwendbare Größe hochskalieren – in der grundsätzlich gleichen Weise, wie man das mit einer Software wie Topaz Gigapixel AI erreichen würde. Der Prompt spielt dabei keine Rolle mehr, da nur noch feine Details hinzugefügt werden. ■

Lernen, aber richtig

Im Training soll ein neuronales Netz lernen, zu einem bestimmten Input einen vorgegebenen Output zu erzeugen. Dabei könnte sich das Netz einfach aus der Affäre ziehen, indem es alle im Training verwendeten Kombinationen von Input und Output auswendig lernt. Damit wäre allerdings nichts gewonnen, denn das Netz würde nur zu den Eingaben aus dem Trainingskorpus die richtigen Resultate liefern, bei allen anderen aber versagen. Damit es nach Abschluss des Trainings auch mit unbekannten Eingaben zurecht kommt, muss man die Größe des Netzes beschränken – Auswendiglernen erfordert viel Speicherplatz, während ein kleineres Netz nur dann zu einem Trainingserfolg kommt, wenn es aus den Beispielen zu generalisieren gelernt hat. Andererseits darf es aber auch nicht zu klein bemessen sein, weil es der Aufgabe dann nicht gewachsen wäre.

Tiefe neuronale Netze vorm Transformer-Typ, wie sie zur Berechnung der Koordinaten von Bildern oder Texten im Merkmalsraum eingesetzt werden, nutzen dasselbe Prinzip: Von ihrer Input-Schicht aus nimmt die Zahl der Neuronen in den weiteren Schichten zunächst ab, um bis zur Output-Schicht wieder zuzunehmen. Die künstlich gebildete Engstelle zwingt das Netz im Training dazu, eine kompakte Repräsentation der Eingaben zu finden – es lernt also, die Eingabedaten zu komprimieren – zum Beispiel als Koordinaten im Merkmalsraum.

Das ist doch gar kein Pinguin!

Das Training der Umwandlung von Bildern in Koordinaten führt erfahrungsgemäß nicht zu hundertprozentig perfekten Ergebnissen. In manchen Mustern **(b)**, in denen wir kein bestimmtes Objekt erkennen können, sieht das neuronale Netz Motive wie beispielsweise einen Pinguin oder eine Gitarre, und das mit vermeintlich ebenso großer Sicherheit wie bei einer korrekten Klassifizierung **(a)**.

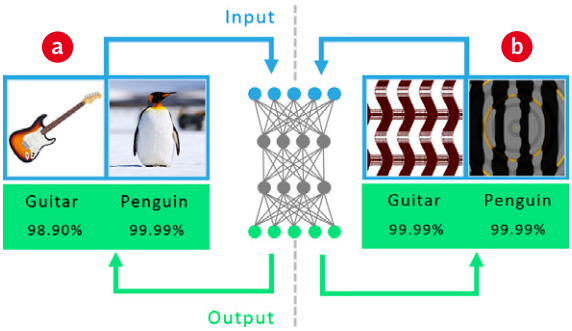


Illustration: Anh Nguyen Jason Yosinski und Jeff Clune

Bei Aufgaben wie einer Verschlagwortung von Fotos führt das zu Irritationen, etwa wenn eine Gesichtserkennung Gesichter oder gar bestimmte Personen zu erkennen meint, wo keine Menschen im Bild zu sehen sind. In Text-zu-Bild-Systemen fallen solche Fehlleistungen dagegen nicht auf, weil sie ihre Bilder mit einem Diffusion-Modell erzeugen, für den umgekehrten Weg von Koordinaten zu Bildern also ein anderes neuronales Netz als zur Bilderkennung verwenden.