Bildgenerierung mit Künstlicher Intelligenz

Philipp Nöcker-Prior, 20. März 2023

Hochschule der Medien, Stuttgart, Deutschland
Professor Dr. Koch, Aktuelle Themen, Wintersemester 2022/23
Künstliche Intelligenz, Aktueller Stand und ihre Auswirkungen auf Ihre Zukunft

Mittels Künstlicher Intelligenz (KI) können scheinbar im Handumdrehen beeindruckende Bilder generiert werden. Durch Eingabe weniger Worte ist es möglich Bilder zu erzeugen, die eine gewünschte Szene darstellen, beschriebene Details enthalten oder in einem gewünschten Stil erstellt werden. Möglich machen das verschiedene deeplearning Netzwerke, die in einem geschickten Zusammenspiel miteinander Informationen austauschen und zu einem Text-Bild-Modell zur Bildgenerierung verschmelzen.

Text-Bild-Modell

Um aus Text in natürlicher Sprache Bilder zu generieren, muss der Eingabetext (Prompt) mit einem Sprachmodell verarbeitet und in eine Form gebracht werden, die der Bildgenerator verwenden kann.

Für die Bildgenerierung werden zwei wesentliche Komponenten benötigt: ein neuronales Netzwerk, das darauf trainiert wurde, ein Bild mit einem passenden Text zu verknüpfen, der das Bild beschreibt, sowie ein weiteres Netzwerk, das darauf spezialisiert ist, Bilder von Grund auf zu generieren. Die Intention hinter dieser Methode ist, das zweite neuronale Netzwerk so zu trainieren, dass es ein Bild erstellt, welches vom ersten Netzwerk als passende Antwort auf die gegebene Anfrage akzeptiert wird.

Zusammen beschreiben diese drei Hauptkomponenten ein Text-Bild-Modell, wie es bei den aktuellen KI-Modellen DALL-E 2, Midjourney und Stable Diffusion Anwendung findet. [1]

Bilder

Um zu verstehen, was die Herausforderungen bei der Bildgenerierung durch eine KI sind, muss zunächst einmal verstanden werden, was ein Bild ausmacht. Mithilfe von ChatGPT, einem auf KI basierenden Sprachmodell von OpenAI, lassen sich dazu Kernaussagen formulieren, die Darstellung, Bedeutung und Betrachter einbeziehen:

Ein Bild ist eine visuelle Darstellung von etwas, das mit einem Medium erstellt wurde, das Bilder aufnehmen oder produzieren kann, wie beispielsweise eine Kamera oder ein Pinsel. Ein Bild kann verschiedene Elemente enthalten, darunter Farben, Formen, Linien, Texturen und Muster. Diese Elemente können auf verschiedene Arten angeordnet werden, um eine bestimmte Botschaft oder Wirkung zu vermitteln.

Ein Bild kann auch verschiedene Ebenen der Bedeutung haben, die auf der Art des Bildes und den Erfahrungen und Überzeugungen des Betrachters basieren. Es kann eine direkte Darstellung von etwas sein, das in der realen Welt existiert, wie ein Porträt oder eine Landschaft. Es kann aber auch abstrakt sein und sich auf Ideen oder Emotionen beziehen, die durch Farben, Formen oder Linien ausgedrückt werden.

Insgesamt besteht ein Bild aus einer visuellen Darstellung von etwas, das auf verschiedene Arten angeordnet sein kann und verschiedene Bedeutungen haben kann, die auf der Perspektive des Betrachters basieren. [2]

Die Erstellung von Fotos, handgemalten Bildern oder Digital-Art ist ein – oft künstlerisch geprägtes – Handwerk. Neben der verwendeten Technik zur Herstellung, bestimmen auch Art, Inhalt und Komposition der Darstellungen die Wirkung eines Bildes für den Betrachter. Eine Zusammenfassung verschiedener Bilder ist in Genres oder auch anhand besonderer Merkmale, die ein Künstler in seine Werke einfließen lässt, möglich. Es lassen sich eine Vielzahl verschiedener Informationen über Werke sammeln. Dies trifft auch dann zu, wenn sie nicht mit dem Anspruch Kunst zu sein, entstanden sind. Es spielt keine Rolle, ob die Bilder dokumentarischer, künstlerischer oder emotionaler Intention entsprungen, noch ob sie professionell oder privat erstellt worden sind: Jedem Bild können Merkmale zugesprochen und nach diesen kategorisiert werden.

Auf Grundlage dieser Erkenntnis kann die Erstellung eines Datensatzes erfolgen, der Bildern eine kurze und prägnante Textbeschreibung zuordnet. Anhand dieses Datensatzes kann das neuronale Netz in einem späteren Schritt trainiert werden.

Sprache

Eine weitere Herausforderung auf dem Weg vom Text zum Bild ist die Fähigkeit eines Systems, Eingaben in natürlicher Sprache verstehen und verarbeiten zu können.

KI versteht Sprache mithilfe von sogenannten Natural Language Processing (NLP) Technologien. NLP-Technologien ermöglichen es Computern, menschliche Sprache zu analysieren, zu verstehen und darauf zu reagieren.

Das Verständnis von Sprache durch KI beginnt mit der Verarbeitung von Rohdaten in Form von geschriebenem Text. Anschließend wird diese Information in Einheiten wie Wörter oder Sätze aufgeteilt, die dann weiter analysiert werden. KI-Modelle können semantische und syntaktische Beziehungen zwischen diesen Einheiten erkennen, um die Bedeutung und den Kontext der Sprache besser zu verstehen.

Es gibt verschiedene Techniken und Algorithmen, die in NLP-Modellen verwendet werden, um Sprache zu analysieren und zu verstehen. Dazu gehören beispielsweise die Aufteilung von Text in Einheiten wie Wörter oder Sätze, die Bestimmung der syntaktischen Rolle jedes Wortes in einem Satz und die Erkennung von spezifischen Entitäten wie Personen, Orte oder Organisationen. Aber auch die Bestimmung der positiven oder negativen Stimmung eines Textes oder die automatische Übersetzung von Texten von einer Sprache in eine andere. [3]

Insgesamt ermöglichen NLP-Technologien es KI-Modellen, menschliche Sprache besser zu verstehen und darauf zu reagieren. Die sog. Transformer wandeln die Texteingabe in verarbeitbare Informationen. In Kombination mit den Informationen, die KI aus Datensätzen mit Bildern und beschreibenden Texten ziehen kann, ist es möglich relevante Informationen aus Texten zu gewinnen, die für die Bildgenerierung genutzt werden können. [4]

Diffusions-Modell

Bildgenerierung kann durch verschiedenartig aufgebaute neuronale Netze realisiert werden. Zu den einfacheren Verfahren gehören Autoencoder, Denoising-Autoencoder (DAE) und Variational Autoencoder (VAE). Zu den komplexeren Verfahren zählen Generative Adversarial Networks (GAN) und Diffusions-Modelle.

Autoencoder sind neuronale Netze, die darauf trainiert sind, die zugrunde liegende Struktur eines Bilddatensatzes zu erlernen. Sie können verwendet werden, um neue Bilder zu erzeugen, indem sie ein Bild in eine niedriger dimensionale Darstellung mit weniger verfügbaren Informationen kodieren und diese dann wieder in ein äquivalentes Bild dekodieren. [5]

GANs bestehen aus zwei neuronalen Netzen, einem Generator und einem Diskriminator, die so trainiert werden, dass sie gegeneinander antreten. Der Generator versucht,

Bilder zu erzeugen, die dem Diskriminator vorgaukeln, dass sie echt sind, während der Diskriminator versucht, korrekt zu erkennen, ob die Bilder echt oder erzeugt sind. Diese Architektur weist jedoch Schwächen in kritischen Bereichen des Trainings auf, die das Framework stagnieren lassen. Ggü. GANs können Diffusions-Modelle bessere Ergebnisse bei Bildbeispielen liefern, sind aber gleichzeitig auch in anderen Punkten limitiert. [6]

In den erwähnten aktuellen Text-Bild-Modellen finden Diffusions-Modelle Anwendung. Diese sind eine Art von generativen Modellen, die für die Bilderzeugung verwendet werden können. Sie bestehen aus mehreren Stufen, von denen jede für das Hinzufügen weiterer Details zum Bild verantwortlich ist. Diffusion vereint DAEs und VAEs in einem Modell.

Ein DAE ist eine Art Autoencoder, der trainiert wird, um Rauschen aus einem Bild zu entfernen. Im Zusammenhang mit einem Diffusionsmodell zur Bilderzeugung kann ein DAE als erste Stufe verwendet werden, in der er trainiert wird, um Rauschen aus der anfänglichen Zufallsrauscheingabe zu entfernen und ein Bild mit niedriger Auflösung zu erzeugen. Dieses niedrig aufgelöste Bild wird dann an die nachfolgenden Diffusionsschritte weitergegeben, wo es schrittweise verfeinert wird, um ein hoch aufgelöstes Bild zu erzeugen.

VAEs sind eine Art von Autoencodern, die trainiert werden, um die zugrunde liegende Wahrscheinlichkeitsverteilung eines Bilddatensatzes zu lernen. Sie können verwendet werden, um neue Bilder durch Stichproben aus der erlernten Wahrscheinlichkeitsverteilung zu erzeugen. Im Zusammenhang mit einem Diffusionsmodell zur Bilderzeugung kann ein VAE darauf trainiert werden, ein Bild mit niedriger Auflösung auszuwerten und die Wahrscheinlichkeit zu berechnen, dass ein Pixel eine bestimmte Farbe hat. Die Engstelle des VAEs ist dabei stets das Bild der letzten Iterationsstufe.

Sowohl DAEs als auch VAEs können verwendet werden, um die Qualität der erzeugten Bilder in einem Diffusionsmodell zu verbessern. Denoising-Autoencoder können helfen, das Rauschen aus dem anfänglichen niedrig aufgelösten Bild zu entfernen, während VAEs helfen können, die zugrunde liegende Wahrscheinlichkeitsverteilung der Trainingsbilder zu erlernen, was den Realismus der erzeugten Bilder verbessern kann.

Die erste Stufe des Diffusions-Models ist in der Regel ein einfacher Zufallsrauschgenerator, der ein Bild mit geringer Auflösung erzeugt. Die nachfolgenden Stufen, die so genannten Diffusionsschritte, werden so trainiert, dass sie dem Bild mehr Details hinzufügen, indem sie lernen, die Informationen der vorherigen Stufe zu "verbreiten". Bei jedem Diffusionsschritt wird ein neuronales Netz, ein so genannter Diffusionstransformator, auf das Bild angewendet, um mehr Details hinzuzufügen. Der Diffusionstransformator ist darauf trainiert, die Muster und Merkmale echter Bilder zu lernen, und nutzt dieses Wissen, um dem erzeugten Bild realistische Details hinzuzufügen. Das Ergebnis ist im Idealfall ein hochauflösendes Bild, das realen Bildern ähnlich ist.

Im Training wird diesem Netz ein Set an Bildern zugeführt, welches immer weiter verrauscht wird. Das Ergebnis ist eine automatische Bildgenerierung aus beliebigem Rauschen. Jedoch kann auf diesen Prozess zunächst keinen Einfluss genommen werden. Die Funktionalität der sog. Guided-Diffusion bekommt das Modell, indem eine Anbindung des Diffusions-Modells an den Transformer stattfindet. [7]

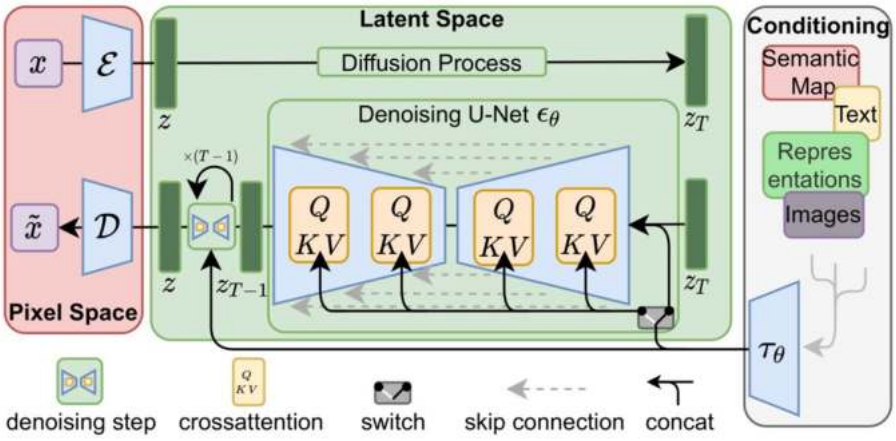

Abbildung 1: Guided Diffusion Framework [8]

Guided Diffusion

Um die Möglichkeit zu erlangen den Diffusions-Prozess steuern zu können, muss ein neuronales Netz in die Architektur des Text-Bild-Modells integriert werden, welches in der Lage ist Bildinhalte zu erkennen und mit textueller Beschreibung zu vergleichen. CLIP (Contrastive Language-Image Pre-Training) ist ein neuronales Netzwerk von OpenAI. Das CLIP-Modell besteht aus einem bereits trainierten Transformer-Sprachmodell sowie einem Bilderkennungsnetzwerk. Beide Komponenten geben eine Vektorrepräsentation aus, die die Bedeutung des Satzes oder des Bildes kodieren soll und deren Länge identisch ist. [5]

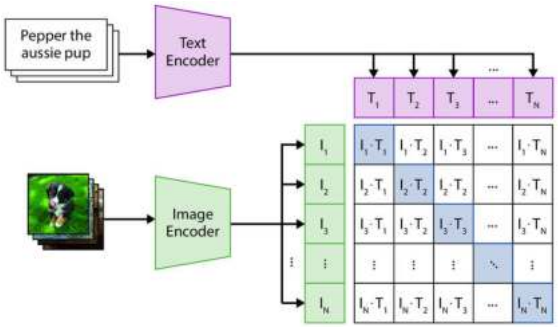

Abbildung 2: Vergleichsmatrix des CLIP [9]

Die Wortbedeutung muss mit dem entstehenden Bildinhalt abgeglichen werden und so ein geführter Diffusionsprozess ermöglicht werden. Durch dieses Bindeglied bekommt das Text-Bild-Modell die Möglichkeit die verarbeiteten Texteingaben in einen Abgleichungsprozess einfließen zu lassen, der den Diffusionsprozess steuern kann.

Dabei wird eine Anfangsbildfolge, die als "Guide" dient, verwendet, um das Modell beim Generieren des endgültigen Bildes zu lenken. Das Modell berechnet dann iterativ die Wahrscheinlichkeitsverteilung für jeden Pixel im Bild und generiert das endgültige Ergebnis durch Abtasten dieser Verteilungen.

Ergebnis

Guided Diffusion hat sich als sehr leistungsfähig erwiesen, da es in der Lage ist, realistisch wirkende Bilder mit hoher Auflösung zu erzeugen. Es hat auch den Vorteil, dass es sehr flexibel ist und in der Lage ist, eine Vielzahl von Bildgenerierungsaufgaben zu bewältigen, wie z.B. die Generierung von Bildbeschreibungen oder die Entfernung von Störungen aus Bildern. Das Ergebnis der Bildgenerierung ist abhängig von der Qualität des Sprachmodells und der Verknüpfung von Diffusions-Prozess und dem Text-Bild-Vergleich. Dabei spielt die Bedeutungsebene eine entscheidende Rolle, da Sprachmodelle natürliche Sprache zwar analysieren, aber nicht im engeren Sinne verstehen. Mit Stable Diffusion 2.0, OpenCLIP und ChatGPT-4 sind bereits zu Beginn des Jahres 2023 erneut performantere und größere KI-Systeme vorgestellt worden, die diese Lücke versuchen zu schließen. Eine Weiterentwicklung dieser Komponenten wird zukünftig noch stärkere Text-Bild-Anwendungen ermöglichen. Folgend werden für zwei Prompts die Ergebnisse von DALL-E 2, Midjourney und Stable Diffusion Web gezeigt:

Prompt 1

„robot eating a green cable, digital art“

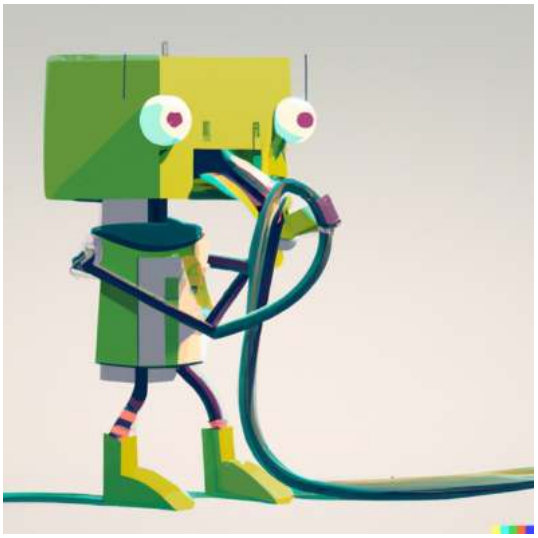

Abbildung 3: Ein Ergebnis von DALL-E 2 zu Prompt 1

Abbildung 4: Ein Ergebnis von Midjourney zu Prompt 1

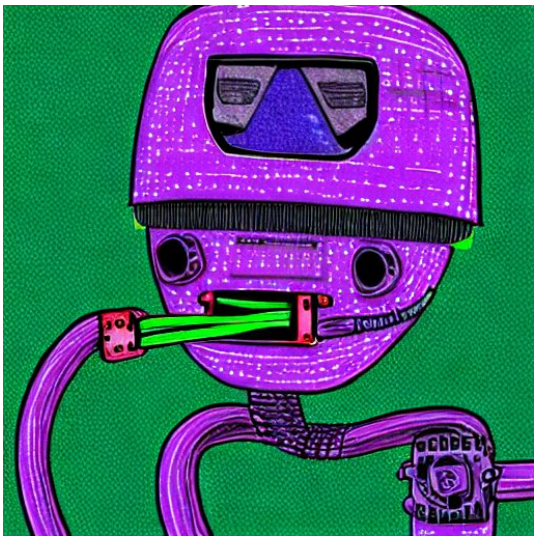

Abbildung 5: Ein Ergebnis von Stable Diffusion Web zu Prompt 1

Prompt 2

„ girl standing in a forest, shining light, 4k photography “

Abbildung 6: Ein Ergebnis von DALL-E 2 zu Prompt 2

Abbildung 7: Ein Ergebnis von Midjourney zu Prompt 2

Abbildung 8: Ein Ergebnis von Stable Diffusion Web zu Prompt 2

Schlussbetrachtung

Anhand der dargestellten zufälligen Ergebnisse der drei gewählten Text-Bild-Modelle ist bereits ersichtlich, dass sich die Qualität der erzeugten Bilder zwischen den jeweiligen KIs abhängig vom formulierten Prompt stark unterscheidet. Mit einer größeren Anzahl von Vergleichsbildern der Bildgeneratoren ließe sich dieser Eindruck weiter manifestieren.

Für Prompt 1 lässt sich feststellen, dass der gewünschte Stil „digital art“ von jedem Bildgenerator unterschiedlich interpretiert wurde und sich die Ergebnisse stark unterscheiden. Inhaltlich lassen sich jedoch Parallelen erkennen. Die geforderte Darstellung eines Roboters, der ein grünes Kabel frisst („robot eating a green cable“) wird in allen Ergebnissen umgesetzt. Mit einer Anpassung des Prompts ließen sich dahingehend noch exaktere Ergebnisse erzielen. Das sog. Prompt Engineering, also die Entwicklung von Texteingaben für die Erzielung bestimmter Ergebnisse, muss jedoch für jedes Bild-Text-Modell separat erfolgen, da sie unter-schiedliche Stärken aufweisen und dementsprechend jeweils anders optimiert werden müssen.

Bei den Ergebnissen für Prompt 2 ist auffällig, dass sowohl DALL-E 2, als auch Stable Diffusion Web die Umsetzung eines fotorealistischen Bildes erbringen. Durch den Zusatz „4k photography“ wird den Bildgeneratoren vorgegeben ein hochauflösendes Foto zu generieren. Lediglich Midjourney erstellt auch hier ein Bild, welches in das Genre Digital Art einzuordnen ist.

Beeinflusst werden die Ergebnisse maßgeblich von der gewählten Text-Bild-Modellarchitektur, sowie dem verwendeten Trainingsdatensatz. Eine Weiterentwicklung der Datenbasis ermöglicht über die o.g. technischen Möglichkeiten hinaus eine Verbesserung der Bildqualität und der Genauigkeit solcher Bildgeneratoren.

Literatur

[1] W. D. Heaven, „Wie KI-Bild- und Textgeneratoren die Kreativ-Branche umkrempeln,“ MIT Technology Review, 2. März 2023. [Online]. Available: https://www.heise.de/hintergrund/Wie-KI-Bild-und-Textgeneratoren-die-Kreativ-Branche-umkrempeln-7489270.html?seite=all. [Zugriff am 10. März 2023].

[2] „ChatGPT, Prompt: "Was macht ein Bild aus?"," OpenAI, [Online]. Available: https://chat.openai.com/chat. [Zugriff am 5. - 18. März 2023].

[3] D. W. Otter, J. R. Medina und J. K. Kalita, „A Survey of the Usages of Deep Learning in Natural Language Processing,“ *IEEE transactions on neural networks and learning systems,* Bd. 32, Nr. 2, pp. 604-624, 2020.

[4] P. Merkert, „KI: So funktionieren künstliche Sprachsysteme vom Typ "Transformer".,“ c't Magazin, 11. Mai 2022. [Online]. Available: https://www.heise.de/hintergrund/KI-So-funktionieren-kuenstliche-Sprachsysteme-vom-Typ-Transformer-7077832.html?seite=all. [Zugriff am 1. Dezember 2022].

[5] P. Merkert, „KI-Bildgeneratoren: Diese Technik Steckt dahinter.,“ c't Magazin, 18. November 2022. [Online]. Available: https://www.heise.de/hintergrund/KI-Bildgeneratoren-Diese-Technik-steckt-dahinter-7341800.html?seite=all. [Zugriff am 1. Dezember 2022].

[6] „ChatGPT, Prompt: "Tell me how GAN based image generators work.",“ OpenAI, [Online]. Available: https://chat.openai.com/chat. [Zugriff am 5. - 18. März 2023].

[7] P. Dhariwal und A. Nichol, „Diffusion Models Beat GANs on Image Synthesis,“ *Advances in Neural Information Processing Systems,* Nr. 34, pp. 8780 - 8794, 2021.

[8] R. Rombach, et al., „High-Resolution Image Synthesis With Latent Diffusion Models,“ *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 10684-10695, 2022.

[9] A. Redford, et al., „Learning Transferable Visual Models From Natural Language Supervision,“ Proceedings of the 38th International Conference on Machine Learning, pp. 8748-8763, 2021.