

Deepfakes: Technische Hintergründe und Trends

Philipp Ladwig

05.12.2024 / 5 Minuten zu lesen

Deepfakes sind seit 2017 im Gespräch. Das erste Mal tauchten sie auf der Plattform Reddit auf. Doch in welchem Kontext? Und was sind Deepfakes eigentlich?



Welches Gesicht zeigt uns die KI? Über Deepfakes werden realitätsgetreue Abbildungen von Menschen erzeugt oder sogar neue Gesichter kreiert. Die Maske ist ebenfalls KI generiert. | Illustration: www.leitwerk.com (© bpb)

Im Text „Was ist KI und welche Formen von KI gibt es?“ [[Link: https://www.bpb.de/lernen/bewegt-bild-und-politische-bildung/555997/was-ist-ki-und-welche-formen-von-ki-gibt-es/](https://www.bpb.de/lernen/bewegt-bild-und-politische-bildung/555997/was-ist-ki-und-welche-formen-von-ki-gibt-es/)] wurde die hohe Leistungsfähigkeit von künstlichen neuronalen Netzen beschrieben. Die Generierung von Deepfakes wird heute in der Regel ebenfalls durch solche neuronalen Netze realisiert. Ein Deepfake beschreibt realistisch aussehende bzw. klingende, aber gefälschte Bilder, Videos oder Audioaufnahmen. Dabei werden bestehende audio-visuelle Medieninhalte so manipuliert, dass Personen Dinge sagen oder tun, die sie nie gesagt oder getan haben. Diese Technologie wird entweder zur Täuschung oder für betrügerische Absichten eingesetzt, kann aber auch mit guten Absichten zu Unterhaltungs-, Kunst- oder Bildungszwecken verwendet werden.

Der Ursprung von Deepfakes liegt auf der Plattform Reddit. Ende 2017 wurden dort von einem Nutzer mit dem Pseudonym „Deepfakes“ manipulierte pornografische Videos mit den Gesichtern berühmter Schauspielerinnen^[1] veröffentlicht. Im Jahr 2020 wurde eine Open-Source-Software^[2] veröffentlicht, mit der Deepfake-Videos in hoher Qualität kostenlos produziert

werden können. Mittlerweile gibt es mehrere Unternehmen, die die Produktion von Video- und auch Audio-Deepfakes gegen Bezahlung anbieten. Die meisten dieser professionellen Anbieter haben Sicherheitsrichtlinien, nach denen nicht jede Person, jede Stimme oder jedes Bild ohne Zustimmung des Urhebers synthetisiert werden kann, aber dieser Zustimmungsmechanismus hat Lücken. Mit den Fortschritten in der Entwicklung neuronaler Netze ist die Synthese von Audio-, Bild- oder Videoinhalten in einen qualitativen Bereich vorgedrungen, in dem es sowohl für Menschen als auch für Algorithmen zunehmend schwerer bis unmöglich ist, diese Inhalte zu identifizieren.

Die Erstellung von Trainingsdatensätzen ist denkbar einfach, da man nur wenige Minuten bis hin zu wenigen Sekunden Video bzw. Audio benötigt, um glaubhafte Deepfakes produzieren zu können. Die Erzeugung eines Video-Deepfakes ist auch mit nur einem Bild möglich, allerdings ist die schlechte Qualität dieser Deepfakes oft auch für das ungeübte Auge erkennbar. In der Regel verbessert sich die resultierende Qualität mit einem größeren Trainingsdatensatz, bis eine Unterscheidung zu echten Videos nicht mehr möglich ist. Fünf Minuten Video oder Audio reichen oft aus, wobei die Tendenz besteht, dass zukünftig Algorithmen mit zunehmend weniger Daten glaubhafte Resultate liefern.

Das Steuern von Deepfakes kann auf unterschiedliche Weise erfolgen. Stimmen können durch Verfahren generiert werden, die Text in eine authentische Audiospur umwandeln, oder durch eine direkte Stimmenkonvertierung. Dabei kann der Autor seine Stimme aufnehmen, welche anschließend in die Zielstimme umgewandelt wird. Bei Video-Deepfakes wird in der Regel nur das Gesicht der Zielperson in einem bereits bestehenden Video ersetzt. Das bedeutet, dass der untere Teil der Stirn, der Augen, Nase, Mund und Kinn neu animiert werden. Oft auch nur der Bereich um den Mund und das Kinn, damit die Lippensynchronität mit einem Audio Deepfake gewährleistet ist. Dabei orientieren sich die synthetisierten Mundbewegungen an dem Inhalt des Audio Deepfakes. Die Bereiche der Haare, Ohren, Hals und damit auch die gesamte Kopfposition und Rotation werden oft nicht neu animiert. Alternativ kann der Video-Deepfake auch durch das Gesicht einer anderen Person gesteuert werden, so dass die Zielperson im Deepfake die Gesichtsausdrücke einer anderen Person repliziert. Deepfakes auf Basis von Bildern werden in der Regel nur mit Texteingabe, so genannten Prompts, realisiert.

Ein praktisches Beispiel wie ein Video Deepfake entsteht

Stellen wir uns einen Algorithmus vor, der den Gesichtsausdruck einer Person in einem Bild in 100 Werte zwischen 0 und 1 kodieren kann. Wenn die Person in einem Bild beispielsweise lacht, könnte der neunte Wert 0,95 sein. Die anderen Werte könnten Informationen über die Augenbrauen, die Blickrichtung und andere Merkmale eines Gesichtsausdrucks enthalten.

Angenommen, wir haben eine Reihe von Bildern des amtierenden Bundeskanzlers Olaf Scholz während einer Rede und für jedes Bild haben wir diese kodierten Gesichtsausdrücke. Der Trainingsdatensatz besteht also aus Einzelbildern aus dem Video, und zu jedem Bild sind die entsprechenden 100 Gesichtsparemeter gespeichert. Beim Training eines neuronalen Netzes für Deepfakes erhält das Netz ein Eingangssignal (das sind die 100 Werte für die Gesichtsausdrücke), woraufhin es eine Vorhersage generiert, in unserem Fall ein Bild von Olaf Scholz. Dieses generierte Bild wird nun mit dem entsprechenden echten Referenzbild aus dem Trainingsdatensatz verglichen. Dies ist das Bild, dem die 100 zuvor kodierten Gesichtsparemeter zugeordnet wurden.

Zu Beginn des Trainings erzeugt das Netz zufällige Bilder, die keine erkennbare Struktur aufweisen. Mit der Zeit lernt das Netz jedoch die Zusammenhänge zwischen den Eingangsdaten (den Gesichtsparemetern) und dem echten Referenzbild. Dieses Lernen wird durch die Fehlerfunktion ermöglicht. Die Fehlerfunktion für Video-Deepfakes gibt einen Zahlenwert wieder, der den Unterschied zwischen zwei Bildern ausdrückt. Gibt man z.B. zwei identische Bilder in die Fehlerfunktion ein, so erhält man den Wert 0, d.h. es ist kein Fehler zu messen. Je unterschiedlicher die beiden Bilder sind, desto größer ist der Fehlerwert. Dieser Fehler, bezogen auf das Referenzbild aus dem Trainingsdatensatz, wird dem Netz für jedes von ihm erzeugte Bild mitgeteilt. Das Ziel des neuronalen Netzes während des Trainings ist es, diesen Fehlerwert zwischen dem echten Referenzbild und dem von ihm erzeugten Bild zu minimieren. Dieser Prozess kann mit dem Spiel „Topfschlagen“ verglichen werden: Je weiter man vom Topf entfernt ist, desto lauter wird „ganz kalt“ gerufen (großer Fehlerwert). Je näher man dem Topf kommt, desto lauter hört man „wärmer“, „sehr heiß“ (Fehlerwert klein), bis man schließlich den Topf findet (das erzeugte Bild mit dem Referenzbild nahezu identisch). Während des Trainings durchläuft das neuronale Netz viele Millionen Durchläufe dieses „Topfschlagens“.

Durch diesen kontinuierlichen Prozess lernt das neuronale Netz, immer realistischere Bilder zu erzeugen, die den vorgegebenen 100 Gesichtsparametern entsprechen. Das Ergebnis ist, dass das neuronale Netz einen Zusammenhang zwischen den 100 Gesichtsparametern und den Ausgabebildern gelernt hat. Ist der neunte Wert 0,95 so lacht Olaf Scholz nun auch im synthetisierten Bild. Im Anschluss kann eine weitere KI die 100 Gesichtsparameter entsprechend auf einen Audio-Deepfake lippensynchron animieren, so dass die Mundbewegung zum Gesagten passt. Die einzelnen Bilder werden zu einem Video zusammengefügt, sind nun täuschend echt und spiegeln die Mimik und Gesichtszüge der „trainierten“ Person exakt wider.

Technische Entwicklung und Trends

Wie oben bereits erwähnt, zeichnet sich der Trend ab, dass die Qualität der Systeme stetig zunimmt. Hinzu kommt, dass unabhängige Entwicklerinnen und Entwickler, die oft nur unter einem Pseudonym bekannt sind, Systeme ohne Sicherheitsbarrieren veröffentlichen. Das bedeutet, dass qualitativ hochwertige Bilder, Videos oder Stimmen tendenziell leichter ohne Sicherheitsabfragen erzeugt werden können. Ein weiterer Trend ist die Verkürzung der Generierungszeit von Audio- und Videoinhalten bis hin zur Echtzeitfähigkeit. So wird es immer einfacher, Telefongespräche oder Videokonferenzen mit Deepfakes zu realisieren.

Fußnoten



[1] In diesen Videos wurden die Gesichter berühmter Schauspielerinnen mit denen von Pornodarstellerinnen getauscht. Mehr unter Kap. 2.2 [Nicht-einvernehmliche pornographische Deepfakes](#).

[2] ↗ <https://de.wikipedia.org/wiki/DeepFaceLab> Die Software ist mittlerweile gesperrt und nicht mehr verfügbar, aber ähnliche Open-Source-Softwares sind bereits auf dem Markt und werden stetig weiterentwickelt.



Dieser Text ist unter der Creative Commons Lizenz "[CC BY-NC-ND 4.0 - Namensnennung - Nicht kommerziell - Keine Bearbeitungen 4.0 International](https://creativecommons.org/licenses/by-nc-nd/4.0/)" [[Link: https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de](https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de)] veröffentlicht.
Urheberrechtliche Angaben zu Bildern / Grafiken / Videos finden sich direkt bei den Abbildungen.
[\[Link: https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de\]](https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de)
[Sie wollen einen Inhalt von bpb.de nutzen? \[Link: https://www.bpb.de/die-bpb/faq/184955/nutzungsrechte/\]](https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de)