RUHR-UNIVERSITÄT BOCHUM

RUB

RUBIN

WISSENSCHAFTSMAGAZIN

SONDERAUSGABE

IT-SICHERHEIT

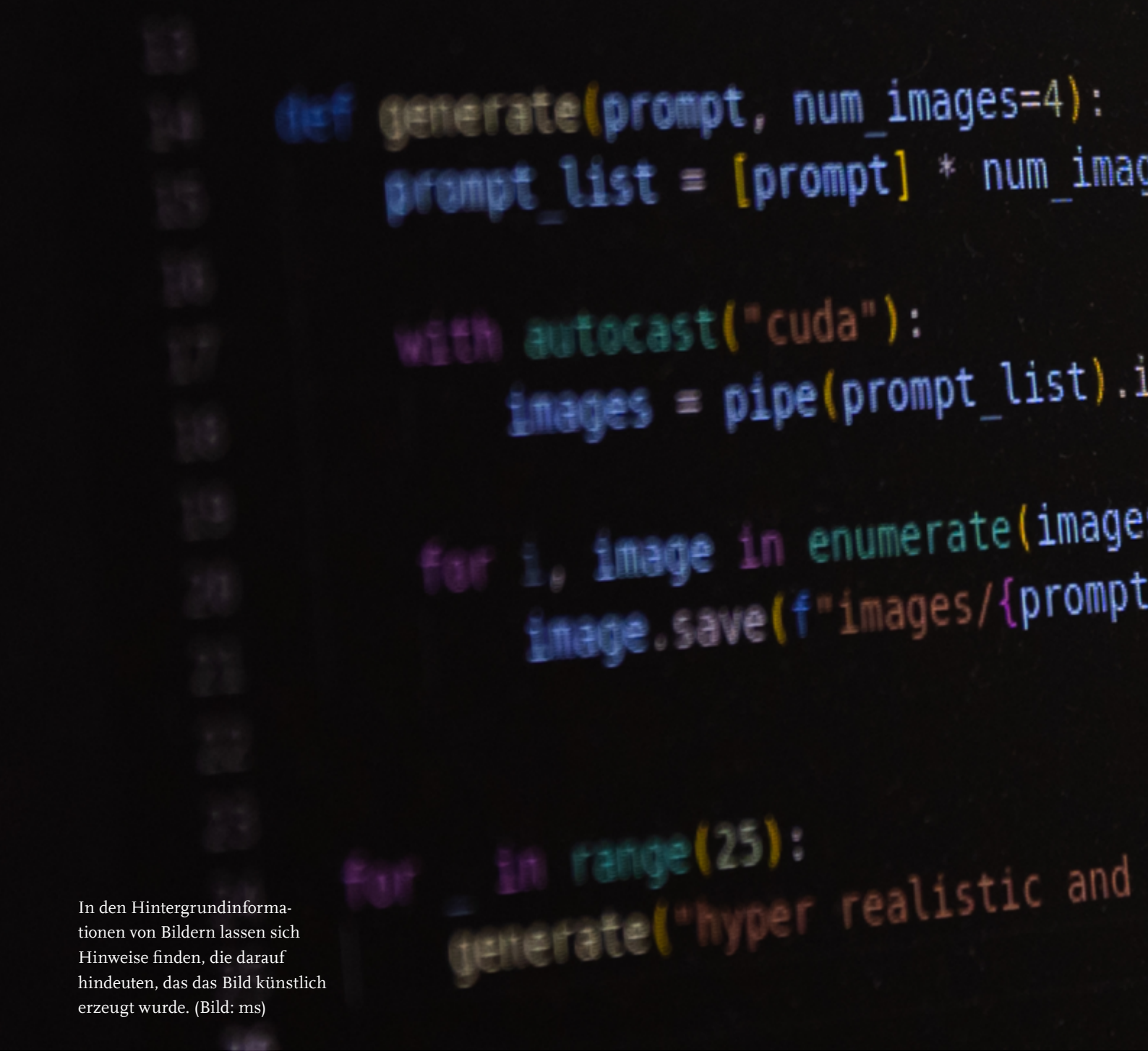
Wie sich künstlich erzeugte Bilder verraten

Drei harte Nüsse für Quantencomputer

Start-up: Fit für die neue Mobilfunkgeneration

#33 Jahrgang

Sonderausgabe 2023



In den Hintergrundinformationen von Bildern lassen sich Hinweise finden, die darauf hindeuten, das das Bild künstlich erzeugt wurde. (Bild: ms)

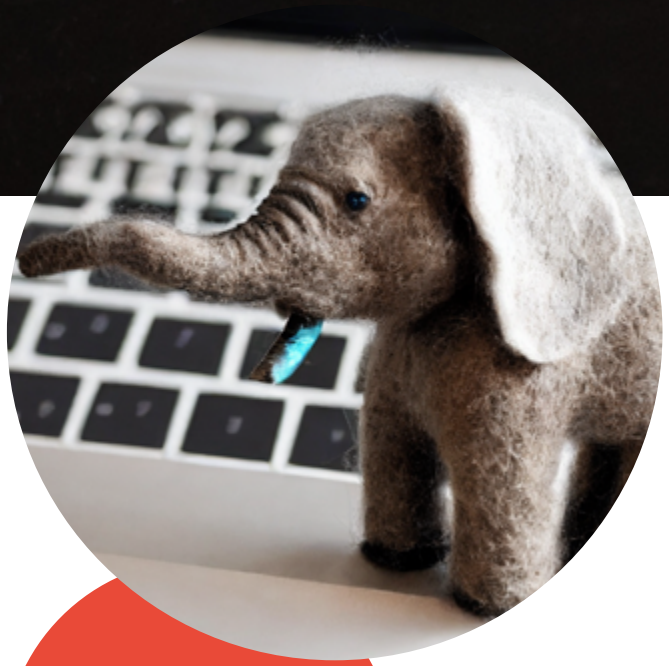
Menschen haben oft keine Chance, künstlich erzeugte Bilder, Audios oder Videos von echten zu unterscheiden. Deswegen arbeiten Forschende am Horst-Görtz-Institut für IT-Sicherheit an einer automatisierten Erkennung.

Wladimir Putin steht hinter einem Rednerpult und wendet sich an die USA: Man habe durchaus die Möglichkeit, die Demokratie Amerikas zu schädigen – doch habe man das gar nicht nötig. Die USA würden das schon selbst erledigen. Die Gesellschaft sei bereits gespalten. Das Video sieht aus wie echt – ist es aber nicht. Youtube ist voll von solchen Videos, die mal besser, mal schlechter gemacht sind. „Es ist schon noch viel Arbeit, aber wer will, der schafft es, zum Beispiel das Gesicht einer berühmten Persönlichkeit so gekonnt auf einen anderen Körper zu montieren, dass man es auf den ersten Blick nicht bemerkt“, sagt Jonas Ricker.

Er hat sich für seine Doktorarbeit, die er an der Fakultät für Informatik schreibt, auf gefakte Bilder spezialisiert. Im Mittelpunkt seiner Arbeit stehen allerdings nicht Videos,

Deepfake

WIE SICH **KÜNSTLICH ERZEUGTE BILDER** VERRATEN



Sieht aus wie echt: Dieser Wollelefant ist durch Text-zu-Bild-Generierung entstanden. (Bild: Hugging Face)

sondern Fotos. Er kann auf Anhieb mehrere Links aus dem Ärmel schütteln, unter denen man zum Beispiel Bilder von Personen anschauen kann, die nicht existieren, oder raten kann, ob das Bild einer gezeigten Person echt ist oder nicht. Die gefakten Bilder werden mithilfe von Deep Learning, einer Methode des maschinellen Lernens erzeugt – daher die Bezeichnung „Deepfake“. „Bei älteren Verfahren kann man manchmal sehen, dass es Auffälligkeiten bei der Symmetrie gibt“, zeigt er. „Zum Beispiel sind verschieden aussehende Ohrringe verräterisch oder asymmetrische Brillengläser. Aber die Methoden werden immer besser, und Studien haben belegt, dass Menschen bei der Unterscheidung echter und gefälschter Bilder eher schlecht sind.“

Ein Verfahren hinter der Erzeugung solcher Bilder nennt sich GAN für Generative Adversarial Networks. „Im Grunde ▸

sind solche Netzwerke immer zweigeteilt: Ein Teil generiert das Bild, ein anderer, der sogenannte Diskriminator, entscheidet, ob das generierte Bild echt aussieht oder nicht“, erklärt Jonas Ricker. „Man kann sich das so vorstellen, als wäre der eine Teil ein Geldfälscher, der andere Teil die Polizei, die gefälschte von echten Banknoten unterscheiden muss.“ Diese Entscheidung trifft die Künstliche Intelligenz auf der Basis vieler echter Bilder, die als Lerndatensatz einfließen. Am Anfang erzeugt der Generator einfach zufällig irgendwelche Pixel. Im Verlauf lernt er durch die Rückmeldung des Diskriminators immer mehr, worauf es ankommt. Auch der Diskriminator wird immer besser darin, die Bilder des Generators von echten zu unterscheiden. Generator und Diskriminator trainieren sich quasi gegenseitig, was schlussendlich zu täuschend echten Bildern führt.

In einem 2020 veröffentlichten Artikel beschreibt sein ehemaliger Kollege Joel Frank eine Möglichkeit, wie man gefälschten Bildern auf die Spur kommen kann. Der Schlüssel liegt in den sogenannten Frequenzen. „Es ist schwierig zu erklären, was Frequenzen bei Bildern sind“, so der Forscher. Am besten kann man sich Frequenzen als Hell-Dunkel-Unterschiede vorstellen. In Gesichtern von Menschen sind niedrige Frequenzen häufig. Hohe Frequenzen können zum Beispiel bei Haaren vorkommen. Die hohen Frequenzen werden unbewusster wahrgenommen. Ein Bild, bei dem hohe Frequenzen verändert wurden, sieht für uns daher fast genauso aus wie das originale Bild. Die Technik lässt sich aber nicht so leicht blenden: „Bei hohen Frequenzen gibt es bei GAN-erzeugten Bildern charakteristische Abweichungen von echten Fotos“, erklärt Jonas Ricker. Die hohen Frequenzen kommen bei künstlich erzeugten Bildern übermäßig häufig vor. Das lässt sich nachvollziehen, und die Bilder lassen sich anhand dessen von echten Fotos unterscheiden.

Jonas Ricker beschäftigt sich zurzeit mit einer anderen Klasse von Modellen zur Bilderzeugung, den sogenannten Diffusion Models. Während GANs schon 2014 vorgestellt

Künstliche Intelligenzen sind in der Lage, Bilder zu erzeugen, die Menschen nicht von Fotografien unterscheiden können. (Bilder: ms)

> LETZTLICH WIRD JEDES BILD VERDÄCHTIG UND AUCH VERNEINBAR, SOGAR BILDER ALS BEWEISE VOR GERICHT.
>
> Jonas Ricker

wurden, werden diese erst seit etwa drei Jahren erforscht, mit herausragenden Ergebnissen. „Das grundlegende Prinzip von Diffusion Models klingt zunächst verwunderlich“, so Ricker: „Ein echtes Bild wird Schritt für Schritt zerstört, indem zufälliges Rauschen hinzugefügt wird – daher der Name. Nach einigen hundert Schritten sind keine Bildinformationen mehr vorhanden, das Bild ist vollständig verrauscht. Das Ziel des Modells ist nun, diesen Prozess umzukehren, um das ursprüngliche Bild zu rekonstruieren – was ein schwieriges Problem darstellt.“

Der Schlüssel liegt darin, das Bild nicht direkt vorherzusagen, sondern wie beim Verrauschen Schritt für Schritt vorzugehen. Mit einer ausreichend großen Anzahl an Trainingsdaten kann das Modell lernen, ein verrauschtes Bild ein kleines bisschen weniger verrauscht zu machen. Durch die wiederholte Anwendung lassen sich dann aus zufälligem Rauschen komplett neue Bilder erzeugen. „Ein Schwachpunkt dieser Methode ist die lange Laufzeit aufgrund der mehreren hundert Schritte“, schränkt Jonas Ricker ein. „Allerdings wurden schon Techniken zur Optimierung vorgestellt, und die Forschung macht ständig Fortschritte.“

Zuletzt erregten Diffusion Models durch die sogenannte Text-zu-Bild-Generierung großes Aufsehen. Damit lassen sich Bilder auf Basis einer Texteingabe erzeugen, mit erstaunlichem Detailgrad. Trainiert werden diese Modelle mithilfe unzähliger Bild-Text-Paare aus dem Internet. Sowohl diese Datensammlung als auch das eigentliche Training ist extrem rechen- und damit kostenintensiv. Bis vor kurzem waren daher nur große Unternehmen wie Google (Imagen) und OpenAI (DALL-E 2) imstande, diese Modelle in hoher Qualität zu trainieren – und die halten die Modelle weitestgehend unter Verschluss.

Mit „Stable Diffusion“ gibt es jedoch nun ein frei zugängliches Modell, welches im Prinzip jeder selbst nutzen kann, vorausgesetzt der eigene Computer verfügt über genug Leistung. Die Anforderungen sind jedoch moderat, zudem gibt ▸

es inzwischen auch Webseiten, auf denen man sich Bilder zu eigenen Texten erstellen lassen kann.

Das Diffusion Model wird von einer Organisation vorangetrieben, die dank einer Spende über entsprechende Mittel und Rechenleistung verfügt. „Es ist schon jetzt sehr gut in der Erzeugung täuschend echter Bilder und wird sich künftig noch verbessern“, ist Jonas Ricker sicher. Das macht es noch schwieriger, echte Bilder von so erzeugten zu unterscheiden. Mittels Frequenzen klappt das hier schon mal weniger gut als bei GAN-Bildern. „Es gibt den Ansatz, die Reflexionen von Licht in den Augen für die Unterscheidung heranzuziehen – das klappt immerhin bei Bildern von Personen“, so Jonas Ricker. Er testet aktuell verschiedene Ansätze, die es erlauben, durch das Modell erzeugte Bilder von echten Fotos zu unterscheiden.

Ein universeller Detektor, der für alle möglichen GAN-Bilder funktioniert, funktioniert für diese Art von Bildern zum Beispiel eigentlich nicht – es sei denn, man stellt ihn durch ein gewisses Finetuning besser ein. Damit ist gemeint, dass man dem Detektor, der als Lernmaterial sehr viele echte und gefälschte Bilder mitsamt der dazugehörigen Information „echt“ oder „falsch“ zur Verfügung gestellt bekommt, zusätzliche Trainingsdaten gibt, um die Detektion für die neuen Daten zu optimieren. So kann er lernen, die mittels Diffusion Model erzeugten Bilder korrekt zu unterscheiden. Wie er das allerdings macht – das ist unklar.

Wichtig ist die Unterscheidung echter und gefälschter Bilder nicht nur, um Fake News zu enttarnen, die zum Beispiel als Video daherkommen, sondern auch, um Fake-Profile in Social Media dingfest zu machen. Sie werden in großem Stil eingesetzt, um zum Beispiel die öffentliche Meinung politisch zu beeinflussen. „Im Exzellenzcluster CASA geht es genau darum: großskalige Angreifer wie Staaten oder Geheimdienste zu enttarnen, die über die Mittel verfügen, mittels Deepfakes Propaganda zu machen“, so Jonas Ricker.

Die Erkennung gefälschter Fotos hat auch strafrechtliche Relevanz, etwa wenn es um unfreiwillige Pornografie geht, bei der Gesichter von Personen auf die Körper von anderen montiert werden. „Ganz allgemein führt die Masse künstlich erzeugter Bilder zu einem Schwund an Vertrauen, auch in seriöse Medien“, so Jonas Ricker. „Letztlich wird jedes Bild dadurch verdächtig und auch verneinbar, sogar Bilder als Beweise vor Gericht.“

Auch wenn Ricker daran arbeitet, dass gefälschte Bilder automatisch erkennbar werden, schätzt er, dass es letztlich auf etwas anderes hinauslaufen wird: „Ich glaube, am Ende wird es darum gehen, echte Bilder zu zertifizieren“, mutmaßt er. „Das könnte man sich zum Beispiel mit kryptografischen Methoden vorstellen, die schon in der Kamera des Fotografen eingebaut sein müssten und jedes echte Bild unzweifelhaft überprüfbar macht.“

md

> „DIE MASSE KÜNSTLICH ERZEUGTER BILDER FÜHRT ZU EINEM SCHWUND AN VERTRAUEN, AUCH IN SERIÖSE MEDIEN.“
>
> Jonas Ricker

Quiz

WELCHE **PERSONEN SIND ECHT?**

Jeweils eines der beiden Gesichter in jeder Gegenüberstellung ist echt, das andere ist mittels maschinellem Lernen erzeugt. Welche Bilder zeigen reale Fotos?
Die Lösungen finden sich auf Seite 62. Das Material stammt von der Seite whichfaceisreal.com.

1 a ◯

1 b ◯

2 a ◯

2 b ◯

3 a ◯

3 b ◯

4 a ◯

4 b ◯

5 a ◯

5 b ◯

6 a ◯

6 b ◯

REDAKTIONSSCHLUSS

Die Hasen im CASA Universe sind aufgeschreckt: Der scheinbar gut gesicherte Zugang zum Karottenvorrat von Hase Mark wurde gehackt und alle Wintervorräte geraubt. Die mutige Häsin Betty macht sich daraufhin auf die Suche nach Unterstützung im nahegelegenen CASA Hub C – einem geheimnisvollen Ort, der Lösungen für digitale Sicherheit bereithalten soll. So beginnt das Abenteuer von Häsin Betty, der Protagonistin des ersten Wissenschaftscomics des Exzellenzclusters CASA. Gemeinsam mit Betty lernen die Leserinnen und Leser bei ihrem Streifzug durch den Research Hub die Forschungsschwerpunkte und Herausforderungen kennen, mit denen sich die Wissenschaftlerinnen und Wissenschaftler im Forschungsbereich Hub C „Sichere Systeme" tagtäglich beschäftigen. Wie Sie alle Comics der Reihe kostenlos lesen können, erfahren Sie unter:

↗ *casa.rub.de/outreach/wissenschaftscomics*

Auflösung
DEEPFAKE-QUIZ
Folgende Gesichter sind echt:
1a, 2a, 3b, 4a, 5b, 6a

IMPRESSUM

HERAUSGEBER: Exzellenzcluster CASA und Horst-Görtz-Institut für IT-Sicherheit der Ruhr-Universität Bochum in Verbindung mit dem Dezernat Hochschulkommunikation der Ruhr-Universität Bochum (Hubert Hundt, v.i.S.d.P.)

REDAKTIONSANSCHRIFT: Dezernat Hochschulkommunikation, Redaktion Rubin, Ruhr-Universität Bochum, 44780 Bochum, Tel.: 0234/32-25228, rubin@rub.de, news.rub.de/rubin

REDAKTION: Dr. Julia Weiler (jwe, Redaktionsleitung); Meike Drießen (md); Lisa Bischoff (lb)

FOTOGRAFIE: Michael Schwettmann (ms), Dammstr. 6, 44892 Bochum, Tel.: 0177/3443543, info@michaelschwettmann.de, www.michaelschwettmann.de

COVER: Sashkin – stock.adobe.com

BILDNACHWEISE INHALTSVERZEICHNIS: Michael Schwettmann

GRAFIK, ILLUSTRATION, LAYOUT UND SATZ:
Agentur für Markenkommunikation, Ruhr-Universität Bochum,
www.einrichtungen.rub.de/de/agentur-fuer-markenkommunikation

DRUCK: LD Medienhaus GmbH & Co. KG, Van-Delden-Str. 6-8, 48683 Ahaus, Tel.: 0231/90592000, info@ld-medienhaus.de, www.ld-medienhaus.de

AUFLAGE: 4.700

BEZUG: Die reguläre Ausgabe von Rubin erscheint zweimal jährlich und ist erhältlich im Dezernat Hochschulkommunikation der Ruhr-Universität Bochum. Das Heft kann kostenlos abonniert werden unter news.rub.de/rubin. Das Abonnement kann per E-Mail an rubin@rub.de gekündigt werden. Die Sonderausgabe 2023 ist erhältlich beim Horst-Görtz-Institut für IT-Sicherheit. Interessierte können sich per E-Mail an hgi-presse@rub.de melden.

ISSN: 0942-6639