



(12) **Offenlegungsschrift**

(21) Aktenzeichen: **10 2021 104 672.6**
 (22) Anmeldetag: **26.02.2021**
 (43) Offenlegungstag: **01.09.2022**

(51) Int Cl.: **G06V 30/194 (2022.01)**

<p>(71) Anmelder: Robert Bosch Gesellschaft mit beschränkter Haftung, 70469 Stuttgart, DE</p> <p>(72) Erfinder: Blaiotta, Claudia, 70178 Stuttgart, DE; Katiyar, Prateek, 72074 Tübingen, DE</p> <p>(56) Ermittelte Stand der Technik: US 2019 / 0 354 804 A1 BAO, J., et al.: CVAE-GAN: fine-grained image generation through asymmetric training. In: Proceedings of the IEEE international conference on computer vision. 2017. S. 2745-2754. (URL: https://openaccess.thecvf.com/content_ICCV_2017/papers/Bao_CVAE-GAN_Fine-Grained_Image_ICCV_2017_paper.pdf) BIAN, J., et al.: A novel and efficient CVAE-GAN-based approach with informative manifold for semi-supervised anomaly detection. In: IEEE</p>	<p>Access, 2019, 7. Jg., S. 88903-88916. doi: 10.1109/ACCESS.2019.2920251 HAM, H., Jun, T. J., Kim, D.: Unbalanced gans: Pre-training the generator of generative adversarial network using variational autoencoder. In: arXiv preprint arXiv:2002.02112, 2020, S. 1-9. (URL: https://arxiv.org/pdf/2002.02112) ZHANG, Y., et al.: A sequential resampling approach for imbalanced batch process fault detection in semiconductor manufacturing. In: Journal of Intelligent Manufacturing, 2020, S. 1-16. (inhaltsgleiche Version online abrufbar über https://www.researchgate.net/profile/Yi-Zhang-429/publication/346301963_A_sequential_resampling_approach_for_imbalanced_batch_process_fault_detection_in_semiconductor_manufacturing/links/5fbdc122458515b7976a0899/A-sequential-resampling-approach-for-imbalanced-batch-process-fault-detection-in-semiconductor-manufacturing.pdf)</p>
---	---

Rechercheantrag gemäß § 43 PatG ist gestellt.

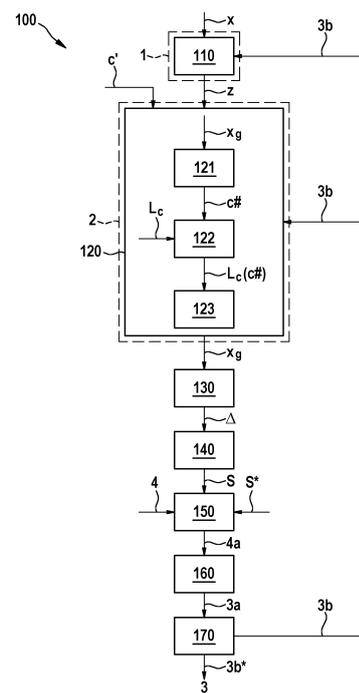
Die folgenden Angaben sind den vom Anmelder eingereichten Unterlagen entnommen.

(54) Bezeichnung: **Erzeugen von kontrafaktischen Bildern für die Auswertung von Bildklassifikatoren**

(57) Zusammenfassung: Ein Verfahren (100) zum Erzeugen eines kontrafaktischen Bildes x_g für ein Eingangsbild x , das ein gegebener Bildklassifikator (3) in eine Quellenklasse c klassifiziert, die aus mehreren verfügbaren Klassen einer gegebenen Klassifikation ausgewählt ist, das die folgenden Schritte umfasst:

- Abbilden (110) des Eingangsbildes x auf eine Darstellung z mit niedrigerer Dimension in einem latenten Raum mittels eines trainierten Codierernetzes (1); und Abbilden (120) zumindest einer Kombination der Darstellung z mit niedrigerer Dimension und einer Angabe einer Zielklasse $c' \neq c$, die aus den mehreren verfügbaren Klassen ausgewählt ist, auf ein kontrafaktisches Bild x_g , das in die Zielklasse c' durch den gegebenen Bildklassifikator (3) klassifiziert wird, mittels eines trainierten Generatornetzes (2).
- Ein Verfahren (200) zum Trainieren einer Kombination eines Codierernetzes (1) und eines Generatornetzes (2) für die Verwendung im Verfahren (100).

Eine Vorrichtung (10), um aus einem Quellbild x , das ein Teil einer Quellendomäne von Bildern ist, ein Zielbild x_g , das ein anderes Muster aus derselben Quellendomäne ist, zu erzeugen.



Beschreibung

[0001] Die vorliegende Erfindung bezieht sich auf die Erzeugung von kontrafaktischen Bildern, die unter anderem für die Auswertung, auf welche Teile eines Bildes ein Bildklassifikator seine Entscheidung gestützt hat, verwendet werden können.

Hintergrund

[0002] Wenn Produkte in Massen produziert werden, ist es üblicherweise eine Anforderung, die Qualität der Produktion kontinuierlich zu überwachen. Das Ziel besteht darin, Qualitätsprobleme so früh wie möglich zu detektieren, um die Grundursache so bald wie möglich zu beheben und nicht zu viele Einheiten des Produkts als Abfall zu verlieren.

[0003] Die optische Untersuchung der Geometrie und/oder Oberfläche eines Produkts ist schnell und zerstörungsfrei. WO 2018/197 074 A1 offenbart eine Untersuchungsvorrichtung, in der ein Objekt einer breiten Vielfalt von Beleuchtungskonfigurationen ausgesetzt werden kann. In jeder Beleuchtungskonfiguration erfasst eine Kamera Bilder des Objekts. Aus diesen Bildern wird die Topographie des Objekts ausgewertet.

[0004] Bilder des Produkts können auch direkt auf eine oder mehrere Klassen einer gegebenen Klassifikation mittels eines Bildklassifikators auf der Basis von neuronalen Netzen abgebildet werden. Auf der Basis des Ergebnisses kann das Produkt einer oder mehreren vorbestimmten Qualitätsklassen zugewiesen werden. Im einfachsten Fall ist dies eine binäre Klassifikation der Form „OK“/ „nicht OK“ = „NOK“.

Offenbarung der Erfindung

[0005] Die Erfindung schafft ein Verfahren zum Erzeugen eines kontrafaktischen Bildes x_g für ein Eingangsbild x . Das Eingangsbild x wird in eine Quellenklasse c aus mehreren verfügbaren Klassen einer gegebenen Klassifikation durch einen Bildklassifikator klassifiziert. Der Name „kontrafaktisch“ des Bildes x_g wird so verstanden, dass er impliziert, dass dieses Bild in eine Zielklasse c' , die aus den mehreren verfügbaren Klassen ausgewählt ist, klassifiziert wird und diese Zielklasse c' von der Quellenklasse c verschieden ist.

[0006] Im Verlauf des Verfahrens wird das Eingangsbild x auf eine Darstellung z mit niedrigerer Dimension in einem latenten Raum mittels eines trainierten Codierernetzes abgebildet. Dieser latente Raum kann beispielsweise durch ein Training des Codierernetzes in Verbindung mit dem Training eines Decodierers, der das ursprüngliche Eingangsbild x rekonstruieren soll, definiert sein. Eine solche Codierer-Decodierer-Struktur wird „Autocodierer“

genannt. Ein Typ von Autocodierern, die vorteilhaft im Zusammenhang mit der vorliegenden Erfindung verwendet werden können, ist der variantenreiche Autocodierer, VAE. Der latente Raum kann als Untermannigfaltigkeit eines kartesischen Raums beliebiger Dimensionalität verstanden werden. Aufgrund dieser Eigenschaft ist es typischerweise leichter, ein Muster z von diesem latenten Raum zu erhalten, indem ein Muster in dem Raum von Eingangsbildern x genommen wird und dieses in das Muster z mittels des Codierernetzes transformiert wird, als es ist, direkt ein Muster z zu finden, das zu dem latenten Raum gehört. Da die Dimensionalität der Darstellung z niedriger ist als die Dimensionalität des Eingangsbildes x (und seiner Rekonstruktion), werden, wenn die Darstellung berechnet wird und dann das ursprüngliche Eingangsbild rekonstruiert wird, die Informationen durch einen „Engpass“ gezwungen. Dieser Engpass zwingt den Codierer, nur die Informationen in die Darstellung z zu codieren, die für die Rekonstruktion des ursprünglichen Bildes x am wichtigsten sind.

[0007] Mittels eines trainierten Generatornetzes wird zumindest eine Kombination der Darstellung z mit niedrigerer Dimension und einer Angabe einer Zielklasse $c \neq c'$, die aus den mehreren verfügbaren Klassen ausgewählt ist, auf das gesuchte kontrafaktische Bild x_g abgebildet. Hier bedeutet „zumindest“, dass das Generatornetz weitere Eingaben wie z. B. ein Rauschmuster, das willkürlich aus irgendeiner Verteilung bezogen wird, empfangen kann.

[0008] Es ist bekannt, dass Generatornetze wie z. B. Generatorteile von generativen gegnerischen Netzen, GANs, Bilder erzeugen, die zu einer gewünschten Domäne gehören und daher im Zusammenhang mit anderen Bildern, die zu dieser Domäne gehören, „realistisch“ sind. Diese Bilder können an zusätzlichen gewünschten Eigenschaften, wie z. B. einer Klasse, zu der das erzeugte Bild gehören soll, weiter „aufbereitet“ werden. Durch Kombinieren des Generatornetzes mit dem Codierernetz und Einspeisen der Darstellung z mit niedrigerer Dimension in das Generatornetz kann das resultierende kontrafaktische Bild x_g spezifisch so erzeugt werden, dass es zum ursprünglichen Eingangsbild so ähnlich wie möglich ist. Das heißt, die wichtigsten Informationen vom ursprünglichen Eingangsbild x , das in der Darstellung z mit niedrigerer Dimension verdichtet wird, werden im kontrafaktischen Bild x_g bewahrt.

[0009] In einer besonders vorteilhaften Ausführungsform wird das kontrafaktische Bild x_g mit dem Eingangsbild x verglichen. Die Bereiche, in denen sich das kontrafaktische Bild x_g vom Eingangsbild x unterscheidet, werden als Bereiche bestimmt, die in Bezug auf die Klassengrenze zwischen der Quellenklasse c und der Zielklasse c' signifikant sind. Dann kann bestimmt werden, ob diese Bereiche tatsäch-

lich Informationen enthalten, die für die vorliegende Anwendung relevant sind.

[0010] Eine Klassifikationsbewertung, die durch einen Bildklassifikator während der optischen Untersuchung von Produkten ausgegeben wird, ist beispielsweise nicht sehr glaubwürdig, wenn sich herausstellt, dass die Entscheidung auf einigen Merkmalen im Hintergrund des Bildes basiert, die mit dem Produkt selbst nichts zu tun haben. Wenn das Ändern einiger Merkmale in einer Schaufensterauslage verursacht, dass eine Verkehrssituation auf der Straße vor diesem Schaufenster anders klassifiziert wird, besteht ebenso ein begründeter Zweifel, ob dieser Klassifikator tatsächlich passend ist, um Verkehrssituationen korrekt zu beurteilen. Ein solches Verhalten eines Bildklassifikators ist etwas ähnlich zum Kaufen eines Produkts A anstelle eines ähnlichen Produkts B aus dem einzigen Grund, dass der Vertriebsmitarbeiter für das Produkt B größer ist als der Vertriebsmitarbeiter für das Produkt A.

[0011] Insbesondere für die Beurteilung von Verkehrssituationen ist es auch sehr wichtig zu wissen, auf welche Teile eines Objekts ein Bildklassifikator seine Entscheidung stützt, um die Anwesenheit dieses Objekts zu erkennen. In vielen Verkehrssituationen werden Objekte teilweise durch andere Objekte verdeckt, müssen jedoch trotzdem erkannt werden. Wenn beispielsweise ein oktogonales Stoppschild durch Schnee bedeckt ist, wird erwartet, dass ein Fahrer und auch ein automatisiertes Fahrzeug es immer noch erkennt und sich dementsprechend verhält. Die rechtzeitige Detektion von Fußgängern, die gleich die Bahn des Fahrzeugs kreuzen, hängt auch kritisch davon ab, welche Teile des Fußgängers sichtbar sein müssen, damit der Fußgänger erkannt wird. Viele gefährliche Situationen entstehen, wenn ein Fußgänger ganz oder teilweise durch geparkte Autos oder andere Hindernisse verdeckt ist und plötzlich in den Weg eines Fahrzeugs schreitet. Wenn der Bildklassifikator benötigt, dass zumindest der Kopf und die zwei Beine sichtbar sind, bevor er den Fußgänger erkennt, und die Beine für eine lange Zeit verdeckt sind, kann der Fußgänger nicht erkannt werden, bis es zu spät ist, um eine Kollision abzuwenden. Wenn dagegen das Erscheinen des Kopfs, des Torso oder eines Arms bereits für die Detektion des Fußgängers ausreicht, wird viel mehr wertvolle Zeit zum Abwenden der Kollision gewonnen.

[0012] Daher wird in einer weiteren besonders vorteilhaften Ausführungsform durch eine gegebene Metrik bestimmt, wie gut die Bereiche, die in Bezug auf eine Klassengrenze signifikant sind, mit gegebenen Bereichen des Eingangsbildes x , die Merkmale des Eingangsbildes enthalten, die als auffällig erachtet werden, in Übereinstimmung sind. Eine Bewertung, die dem durch die gegebene Metrik ausgegebenen Ergebnis entspricht, wird dem Bildklassifikator

zugeschrieben. Wenn beispielsweise in Massen produzierte Produkte optisch untersucht werden, liegen das Produkt und die Kamera immer in derselben räumlichen Anordnung in Bezug aufeinander. Daher ist im Voraus bekannt, welche Teile des erfassten Bildes zum Produkt gehören. Ebenso ist es in Verkehrssituationen bekannt, dass der Himmel keine anderen Fahrzeuge enthält, auf die ein Auto reagieren muss.

[0013] Insbesondere kann im Verwendungsfall der optischen Untersuchung die Metrik messen, ob die Bereiche, die mit Bezug auf die Klassifikation eines Produkts in die Klasse „nicht OK = NOK“ oder irgendeine andere nicht optimale Klasse signifikant sind, Bereichen mit konkreten Defekten oder Mängeln entsprechen. Dies bringt die automatisierte optische Untersuchung mit der manuellen Untersuchung in Einklang, bei der erwartet wird, dass ein menschlicher Qualitätsprüfer, wenn er gefragt wird, warum ein konkretes Produkt verworfen werden soll, auf spezielle Defekte oder Mängel zeigt.

[0014] Die Bewertung, die dem Bildklassifikator zugeschrieben wird, kann als Rückmeldung verwendet werden, um den Bildklassifikator zu verbessern. In einer weiteren vorteilhaften Ausführungsform werden daher Parameter, die das Verhalten des Bildklassifikators charakterisieren, optimiert, so dass, wenn die Berechnung des kontrafaktischen Bildes x_g und die anschließende Auswertung dieses kontrafaktischen Bildes x_g wiederholt werden und die Bewertung für den Bildklassifikator erneut berechnet wird, diese Bewertung sich wahrscheinlich verbessert.

[0015] Ein Verwendungsfall dessen ist ein anwendungsspezifisches weiteres Training eines Bildklassifikators, der vorher in einer allgemeineren Weise vortrainiert wurde. Ein Bildklassifikator kann beispielsweise allgemein trainiert werden, um bestimmte Defekte oder Mängel zu detektieren, aber in einer konkreten Anwendung kann er weiter trainiert werden, um auf die richtigen Stellen zu blicken, die zum tatsächlichen Produkt und nicht zum Hintergrund gehören.

[0016] In einer weiteren vorteilhaften Ausführungsform kann im Verlauf der Abbildung der Darstellung z mit niedrigerer Dimension und der Zielklasse c' auf das kontrafaktische Bild x_g das kontrafaktische Bild x_g auf eine Klassifikationsbewertung $c\#$ durch einen gegebenen Bildklassifikator abgebildet werden. Dieser Bildklassifikator kann derselbe sein, der bereits das Eingangsbild x in die Klasse c klassifiziert hat, aber dies ist nicht erforderlich. Mittels einer gegebenen Klassifikationsverlustfunktion wird bestimmt, wie gut die Klassifikationsbewertung $c\#$ mit einer Klassifikation des kontrafaktischen Bildes x_g in die Zielklasse c' in Übereinstimmung ist. Mindestens eine Eingabe in das Generatormetz wird optimiert, so dass die erneute Berechnung des kontrafaktischen

Bildes x_g auf der Basis der geänderten Eingabe wahrscheinlich bewirkt, dass sich der Wert der Klassifikationsverlustfunktion verbessert. In dieser Weise kann die Erzeugung des kontrafaktischen Bildes x_g feinabgestimmt werden, so dass es eindeutiger in die Zielklasse c' klassifiziert wird.

[0017] Eine oder mehrere Komponenten der Darstellung z mit niedrigerer Dimension oder eines Rauschmusters, das zusätzlich in den Generator eingespeist wird, können beispielsweise geändert werden.

[0018] Wie vorher erörtert, kann das Eingangsbild x ein Bild eines hergestellten Produkts sein, das im Verlauf der optischen Untersuchung des Produkts erfasst wurde, und die Klassen der gegebenen Klassifikation stellen Qualitätsstufen für das Produkt dar. In einem anderen Beispiel kann das Eingangsbild x ein Bild einer Verkehrssituation sein und die Klassen der gegebenen Klassifikation können Objekte darstellen, die für die Interpretation der Verkehrssituation relevant sind.

[0019] Die Erfindung schafft auch ein Verfahren für das Training einer Kombination eines Codierernetzes und eines Generatornetzes für die Verwendung im vorher beschriebenen Verfahren.

[0020] Im Verlauf dieses Trainingsverfahrens wird ein Decodierernetz bereitgestellt. Dieses Decodierernetz ist dazu konfiguriert, eine Darstellung z mit niedrigerer Dimension in einem latenten Raum, die vom Codierernetz erhalten wurde, in ein rekonstruiertes Bild x_d in der Domäne von ursprünglichen Eingangsbildern x abzubilden. Parameter, die das Verhalten der Codierer- und Decodierernetze charakterisieren, werden dann mit dem Ziel optimiert, dass das rekonstruierte Bild x_d dem ursprünglichen Eingangsbild x entspricht, aus dem die Darstellung z erhalten wurde. Wie vorher erörtert, bilden dann das Codierernetz und das Decodierernetz einen Autocodierer mit einem Informationsengpass, der zu einer Konzentration der wichtigsten Bildmerkmale in der Darstellung z führt.

[0021] Ferner wird ein Diskriminatornetz geschaffen. Das Diskriminatornetz ist dazu konfiguriert zu unterscheiden, ob ein Bild von der Domäne von ursprünglichen Eingangsbildern x oder von der Domäne von erzeugten Bildern x_f stammt. Ein Bildklassifikator wird auch bereitgestellt. Dieser Bildklassifikator ist dazu konfiguriert, das ursprüngliche Eingangsbild x und das erzeugte Bild x_f auf eine oder mehrere Klassen der gegebenen Klassifikation abzubilden. Die erzeugten Bilder werden auch „gefälschte“ Bilder genannt.

[0022] Das Generatornetz und das Diskriminatornetz werden gegnerisch trainiert. Das Training kann

beispielsweise zwischen dem Training des Generatornetzes und dem Training des Diskriminatornetzes abwechseln.

[0023] Parameter, die das Verhalten des Generatornetzes charakterisieren, werden mit den Zielen optimiert, dass

- die Genauigkeit, mit der das Diskriminatornetz zwischen ursprünglichen Eingangsbildern x und erzeugten Bildern x_f unterscheidet, abnimmt, und
- der Bildklassifikator gefälschte (d. h. erzeugte) Bilder x_f auf ihre gegebene Zielklasse c abbildet.

[0024] Andererseits werden Parameter, die das Verhalten des Diskriminators charakterisieren, mit dem Ziel optimiert, dass die Genauigkeit, mit der das Diskriminatornetz zwischen ursprünglichen Eingangsbildern x und gefälschten Bildern x_f unterscheidet, zunimmt.

[0025] Dieses gegnerische Training kann beispielsweise durch Optimieren einer Verlustfunktion bewerkstelligt werden, die Folgendes umfasst

- einen gegnerischen Verlust, der die Genauigkeit misst, mit der das Diskriminatornetz zwischen ursprünglichen Eingangsbildern x und erzeugten Bildern x_f unterscheidet, und
- einen Klassifikationsverlust, der misst, wie gut der Bildklassifikator erzeugte Bilder x_f auf ihre gegebene Zielklasse c' abbildet. Dieser Klassifikationsverlust kann beispielsweise ein Kreuzentropieverlust sein.

[0026] Das Gesamtziel des gegnerischen Trainings besteht darin, diese kombinierte Verlustfunktion zu minimieren. Hier hängt der Klassifikationsverlust nur von den Parametern ab, die das Verhalten des Generatornetzes charakterisieren. Aber der gegnerische Verlust hängt zusätzlich von den Parametern ab, die das Verhalten des Diskriminatornetzes charakterisieren. Die Generatorparameter können optimiert werden, um den gegnerischen Verlust zu minimieren, und die Diskriminatorparameter können gleichzeitig optimiert werden, um den gegnerischen Verlust zu maximieren, oder umgekehrt.

[0027] Nachdem das gegnerische Training vollendet wurde, kann die Kombination des Codierernetzes und des Generatornetzes im Verfahren zum Erzeugen eines kontrafaktischen Bildes x_g , das vorstehend beschrieben ist, verwendet werden, um ein erzeugtes Bild x_f zu produzieren, das als gesuchtes kontrafaktisches Bild x_g dienen kann.

[0028] Das gegnerische Training kann nach dem Training des Autocodierers durchgeführt werden. Das heißt, nachdem der Autocodierer trainiert

wurde, können die Parameter des Codierernetzes und des Decodierernetzes fest bleiben. Aber das Training des Autocodierers und das gegnerische Training können auch zu einer einzelnen Trainingsprozedur kombiniert werden.

[0029] In einer weiteren vorteilhaften Ausführungsform werden die Parameter, die das Verhalten des Codierernetzes charakterisieren, zusätzlich auf das Ziel hin optimiert, dass gegenseitige Informationen zwischen dem ursprünglichen Eingangsbild x und seiner Darstellung z maximiert werden. Dies stellt sicher, dass die Darstellung z einige wichtige Attribute des Eingangsbildes bewahrt, so dass sie auf das erzeugte Bild x_f übertragen werden können. Dies vermeidet Situationen, in denen das Codierernetz und das Decodierernetz eine sehr gute Rekonstruktion des ursprünglichen Eingangsbildes x zu dem Preis bewerkstelligen, dass die Darstellung z eine geringe oder keine sichtbare Korrelation mit dem Eingangsbild x aufweist.

[0030] Alternativ oder in Kombination werden die Parameter, die das Verhalten des Generatornetzes charakterisieren, zusätzlich auf das Ziel hin optimiert, dass gegenseitige Informationen zwischen dem gefälschten Bild x_f und der Darstellung z maximiert werden. Dies bewirkt, dass Attribute vom ursprünglichen Eingangsbild x , die auf die Darstellung z übertragen wurden, sich zum erzeugten Bild x_f weiter bewegen. Insbesondere können solche Attribute Merkmale sein, die nicht durch die Klasse erfasst werden, zu der das Bild gehört. Beispiele dessen sind Bildstilattribute wie z. B. Liniendicken oder Farben, die vielmehr über alle Klassen vorhanden sind als dass sie an spezielle Klassen gebunden sind.

[0031] In einer weiteren vorteilhaften Ausführungsform umfasst der Bildklassifikator ein trainierbares Netz, das dazu konfiguriert ist, Eingangsbilder x und erzeugte Bilder x_f auf eine Kombination einer Darstellung z mit niedrigerer Dimension im latenten Raum und einer Klassifikationsbewertung $c\#$ abzubilden. Parameter, die das Verhalten dieses trainierbaren Netzes charakterisieren, werden mit den Zielen optimiert, dass:

- der Bildklassifikator ein ursprüngliches Eingangsbild x auf eine Kombination von: einer Darstellung z , die der Darstellung z entspricht, die durch das Codierernetz erzeugt wird, und einer Klassifikationsbewertung $c\#$, die mit einer Ground-Truth-Kennung c des Eingangsbildes x konsistent ist, abbildet; und
- der Bildklassifikator ein erzeugtes Bild x_f auf eine Kombination von: einer Darstellung z , die der Darstellung z entspricht, aus der das erzeugte Bild x_f erzeugt wurde, und einer Klassifikationsbewertung $c\#$, die mit der Zielklasse c'

konsistent ist, für die das erzeugte Bild x_f erzeugt wurde, abbildet.

[0032] In dieser Weise schafft der Bildklassifikator nicht nur eine Rückmeldung über diese Klasse des erzeugten Bildes x_f . Vielmehr dient er auch dazu, die Selbstkonsistenz im latenten Raum von Darstellungen z zu überwachen.

[0033] Allgemeiner schafft die Erfindung auch eine Vorrichtung, um aus einem Quellenbild x , das ein Teil einer Quellenverteilung von Bildern (wie z. B. Bildern, die eine bestimmte Art von Szenerie zeigen), ist, ein Zielbild x_g , das ein anderes Muster aus derselben Quellenverteilung ist, zu erzeugen. Diese Vorrichtung umfasst Folgendes:

- ein trainiertes Codierernetz, das dazu konfiguriert ist, das Quellenbild x in eine Darstellung z mit niedrigerer Dimension in einem latenten Raum abzubilden; und
- ein trainiertes Generatornetz, das dazu konfiguriert ist, zumindest die Darstellung z mit niedrigerer Dimension auf das Zielbild x_g abzubilden.

[0034] Wie vorher erörtert, bewirkt im Vergleich zur Verwendung nur eines trainierten Generatornetzes die Kombination mit dem trainierten Codierernetz, dass das Zielbild x_g nicht nur in derselben Domäne wie das Quellenbild x liegt, sondern auch eine semantische Ähnlichkeit zu diesem Quellenbild x aufweist.

[0035] In einer vorteilhaften Ausführungsform ist das Generatornetz dazu konfiguriert, eine Kombination der Darstellung z mit niedrigerer Dimension und einer Zielklasse c' auf ein Zielbild x_g abzubilden, das ein Mitglied der Zielklasse c' ist. Die Zielklasse c' ist eine von mehreren Klassen, die in einer vorbestimmten Klassifikation von Bildern in der Quellendomäne enthalten sind. Insbesondere kann diese Zielklasse c' von einer Quellenklasse c verschieden sein, zu der das Quellenbild x gehört. In dieser Weise kann das Zielbild x_g dann zu einem kontrafaktischen Bild in der Hinsicht werden, dass es zum Quellenbild x ähnlich ist, während es gleichzeitig zu einer anderen Klasse gehört.

[0036] Die vorstehend beschriebenen Verfahren können ganz oder teilweise computerimplementiert sein und folglich in Software verkörpert sein. Die Erfindung bezieht sich daher auch auf ein Computerprogramm mit maschinenlesbaren Befehlen, die, wenn sie durch einen oder mehrere Computer ausgeführt werden, bewirken, dass der eine oder die mehreren Computer ein vorstehend beschriebenes Verfahren durchführen. In dieser Hinsicht sind Steuereinheiten für Fahrzeuge und andere eingebettete Systeme, die einen ausführbaren Programm-

code abarbeiten können, ebenso als Computer zu verstehen. Ein nichttransitorisches Speichermedium und/oder ein Download-Produkt kann das Computerprogramm umfassen. Ein Download-Produkt ist ein elektronisches Produkt, das online verkauft und über ein Netz für die unmittelbare Erfüllung übertragen werden kann. Ein oder mehrere Computer können mit dem Computerprogramm und/oder mit dem nichttransitorischen Speichermedium und/oder Download-Produkt ausgestattet sein.

[0037] Im Folgenden werden die Erfindung und ihre bevorzugten Ausführungsformen unter Verwendung von Figuren, ohne irgendeine Absicht, den Schutzbereich der Erfindung zu begrenzen, dargestellt.

[0038] In den Figuren zeigen:

Fig. 1 eine beispielhafte Ausführungsform des Verfahrens 100 zum Erzeugen eines kontrafaktischen Bildes x_g ;

Fig. 2 eine beispielhafte Ausführungsform des Verfahrens 200 zum Trainieren einer Kombination eines Codierernetzes 1 und eines Generatortetzes 2;

Fig. 3 eine beispielhafte Ausführungsform einer Konfiguration zum Durchführen des Verfahrens 200.

[0039] **Fig. 1** ist ein schematischer Ablaufplan einer Ausführungsform des Verfahrens 100 zum Erzeugen eines kontrafaktischen Bildes x_g für ein Eingangsbild x , das zu einer Quellenklasse c einer gegebenen Klassifikation gemäß einem gegebenen Bildklassifikator 3 gehört. In Schritt 110 bildet ein trainiertes Codierernetz 1 ein Eingangsbild x auf eine Darstellung z mit niedrigerer Dimension in einem latenten Raum ab. In Schritt 120 bildet ein trainiertes Generatortnetz eine Kombination dieser Darstellung z und einer Angabe einer Zielklasse $c' \neq c$ auf ein kontrafaktisches Bild x_g ab, das durch den gegebenen Bildklassifikator 3 in die Zielklasse c' klassifiziert wird.

[0040] Insbesondere kann das endgültige kontrafaktische Bild x_g durch einen Optimierungsprozess erhalten werden. Gemäß dem Block 121 kann ein anfängliches kontrafaktisches Bild x_g auf eine Klassifikationsbewertung $c\#$ durch den gegebenen Bildklassifikator 3 abgebildet werden. Mittels einer gegebenen Klassifikationsverlustfunktion L_c kann dann gemäß dem Block 122 bestimmt werden, wie gut die Klassifikationsbewertung $c\#$ mit einer Klassifikation des kontrafaktischen Bildes x_g in die Zielklasse c' in Übereinstimmung ist. Gemäß dem Block 123 kann dann mindestens eine Eingabe in das Generatortnetz 2 derart optimiert werden, dass eine erneute Berechnung des kontrafaktischen Bildes x_g auf der Basis der geänderten Eingabe wahrscheinlich bewirkt, dass sich der Wert der Klassifikationsverlustfunktion L_c verbessert.

[0041] In Schritt 130 wird das kontrafaktische Bild x_g mit dem Eingangsbild x verglichen. In Schritt 140 werden Bereiche Δ , in denen das kontrafaktische Bild x_g sich vom Eingangsbild x unterscheidet, als Bereiche S bestimmt, die mit Bezug auf die Klassengrenze zwischen der Quellenklasse c und der Zielklasse c' signifikant sind.

[0042] In Schritt 150 wird bestimmt, wie gut diese Bereiche S mit gegebenen Bereichen S^* des Eingangsbildes x , die Merkmale des Eingangsbildes x enthalten, die als auffällig erachtet werden, in Übereinstimmung sind. Eine Bewertung 3a, die dem Ergebnis 4a entspricht, das durch die gegebene Metrik 4 ausgegeben wird, wird dem Bildklassifikator 3 in Schritt 160 zugeschrieben.

[0043] In Schritt 170 werden Parameter (3b), die das Verhalten des gegebenen Bildklassifikators 3 charakterisieren, optimiert, so dass, wenn die Berechnung des kontrafaktischen Bildes x_g in Schritt 120 und die anschließende Auswertung dieses kontrafaktischen Bildes x_g in Schritten 130 bis 160 wiederholt werden, die Bewertung 3a des Bildklassifikators 3 sich wahrscheinlich verbessert. Der schließlich optimierte Zustand der Parameter 3b wird mit dem Bezugszeichen 3b* bezeichnet.

[0044] **Fig. 2** ist ein schematischer Ablaufplan einer Ausführungsform des Verfahrens 200 zum Trainieren einer Kombination eines Codierernetzes 1 und eines Generatortetzes 2 für die Verwendung im vorher beschriebenen Verfahren 100.

[0045] In Schritt 210 wird ein Decodierernetz 5 bereitgestellt. Dieses Decodierernetz 5 ist dazu konfiguriert, eine Darstellung z mit niedrigerer Dimension in einem latenten Raum, die vom Codierernetz 1 erhalten wurde, auf ein rekonstruiertes Bild x_d in der Domäne von ursprünglichen Eingangsbildern x abzubilden. In Schritt 220 werden Parameter 1a, 5a, die das Verhalten der Netze des Codierers 1 und des Decodierers 5 charakterisieren, mit dem Ziel optimiert, dass das rekonstruierte Bild x_d dem ursprünglichen Eingangsbild x entspricht, von dem die Darstellung z erhalten wurde. In dieser Weise werden der Codierer 1 und der Decodierer 5 trainiert, so dass sie zu einem (variantenreichen) Autocodierer (V)AE werden. Die schließlich optimierten Zustände der Parameter 1a und 5a werden mit den Bezugszeichen 1a* bzw. 5a* bezeichnet. Gemäß dem Block 221 kann ein zusätzliches Ziel der Optimierung darin bestehen, dass gegenseitige Informationen zwischen dem ursprünglichen Eingangsbild x und seiner Darstellung z maximiert werden.

[0046] In Schritt 230 wird ein Diskriminatortnetz 6 bereitgestellt. Dieses Diskriminatortnetz 6 ist dazu konfiguriert zu unterscheiden, ob ein Bild von der Domäne von ursprünglichen Eingangsbildern x oder

von der Domäne von gefälschten Bildern x_f stammt. Ein Bildklassifikator 3 wird auch bereitgestellt. Dieser Bildklassifikator 3 ist dazu konfiguriert, das ursprüngliche Eingangsbild x und das gefälschte Bild x_f auf eine oder mehrere Klassen der gegebenen Klassifikation abzubilden.

[0047] Dieser Bildklassifikator 3 kann fest sein und als solcher verwendet werden. Aber in der in **Fig. 2** gezeigten Ausführungsform wird der Bildklassifikator 3 ebenso trainiert. Gemäß dem Block 241 kann der Bildklassifikator 3 ein trainierbares Netz umfassen, das dazu konfiguriert ist, Eingangsbilder x und erzeugte Bilder x_f auf eine Kombination einer Darstellung z mit niedrigerer Dimension im latenten Raum und einer Klassifikationsbewertung $c\#$ abzubilden. Gemäß dem Block 242 werden Parameter 3b, die das Verhalten dieses trainierbaren Netzes charakterisieren, mit den Zielen optimiert, dass:

- der Bildklassifikator 3 ein ursprüngliches Eingangsbild x auf eine Kombination von: einer Darstellung z , die der Darstellung z entspricht, die durch das Codierernetz 1 erzeugt wird, und einer Klassifikationsbewertung $c\#$, die mit einer Ground-Truth-Kennung c des Eingangsbildes x konsistent ist, abbildet; und
- der Bildklassifikator 3 ein gefälschtes Bild x_f auf eine Kombination von:

einer Darstellung z , die der Darstellung z entspricht, von der das gefälschte Bild x_f erzeugt wurde, und einer Klassifikationsbewertung $c\#$,

die mit der Zielklasse c konsistent ist, für die das Bild x_f erzeugt wurde, abbildet.

[0048] Der schließlich optimierte Zustand der Parameter 3b wird mit dem Bezugszeichen 3b* bezeichnet.

[0049] In Schritt 250 werden Parameter 2a, die das Verhalten des Generatornetzes 2 charakterisieren, mit den Zielen optimiert, dass:

- die Genauigkeit, mit der das Diskriminatornetz 5 zwischen ursprünglichen Eingangsbildern x und gefälschten Bildern x_f unterscheidet, abnimmt, und
- der Bildklassifikator 3 erzeugte Bilder x_f auf ihre gegebene Zielklasse c abbildet.

[0050] Gemäß dem Block 251 kann ein weiteres Optimierungsziel darin bestehen, dass gegenseitige Informationen zwischen dem Bild x_f und der Darstellung z maximiert werden. Der schließlich optimierte Zustand der Parameter 2a wird mit dem Bezugszeichen 2a* bezeichnet.

[0051] Gleichzeitig in Schritt 260 Parameter 6a, die das Verhalten des Diskriminatornetzes 6 charakteri-

sieren, mit dem Ziel, dass die Genauigkeit, mit der das Diskriminatornetz 6 zwischen ursprünglichen Eingangsbildern x und erzeugten Bildern x_f unterscheidet, zunimmt. Der schließlich optimierte Zustand der Parameter 6a wird mit dem Bezugszeichen 6a* bezeichnet.

[0052] Das heißt, das Training 250 des Generatornetzes 2 und das Training 260 des Diskriminatornetzes 6 werden in gegnerischer Weise durchgeführt.

[0053] **Fig. 3** zeigt eine beispielhafte Konfiguration, die verwendet werden kann, um das vorstehend beschriebene Trainingsverfahren 200 durchzuführen. Das Codierernetz 1 und das Decodierernetz 5 bilden einen (variantenreichen) Autocodierer (V)AE. Das Generatornetz 2 und das Diskriminatornetz 6 bilden ein generatives gegnerisches Netz, GAN. Das Diskriminatornetz 6 gibt für jedes in diesen eingegebene Bild eine Klassifikation aus, ob dieses Bild sich in der Domäne von ursprünglichen Eingangsbildern x befindet oder ob es ein gefälschtes Bild x_f ist. In der in **Fig. 3** gezeigten Ausführungsform wird der Bildklassifikator 3 verwendet, um nicht nur eine Klassifikationsbewertung $c\#$ vorherzusagen, sondern auch eine Darstellung z mit niedrigerer Dimension, die dem ursprünglichen Eingangsbild x bzw. dem erzeugten Bild x_f entspricht. Daher kann er als Teil des GAN betrachtet werden. Das Codierernetz 1 und das Generatornetz 2 bilden die Vorrichtung 10, um aus einem Quellenbild x , das ein Teil einer Quellenverteilung von Bildern ist, ein Zielbild x_g , das ein anderes Muster aus derselben Quellenverteilung ist, zu erzeugen.

ZITATE ENTHALTEN IN DER BESCHREIBUNG

Zitierte Patentliteratur

- WO 2018/197074 A1 [0003]

Patentansprüche

1. Verfahren (100) zum Erzeugen eines kontrafaktischen Bildes x_g für ein Eingangsbild x , das ein gegebener Bildklassifikator (3) in eine Quellenklasse c klassifiziert, die aus mehreren verfügbaren Klassen einer gegebenen Klassifikation ausgewählt ist, das die Folgenden Schritte umfasst:

- Abbilden (110) des Eingangsbildes x in eine Darstellung z mit niedrigerer Dimension in einem latenten Raum mittels eines trainierten Codierernetzes (1); und
- Abbilden (120) mindestens einer Kombination der Darstellung z mit niedrigerer Dimension und einer Angabe einer Zielklasse $c' \neq c$, die aus den mehreren verfügbaren Klassen ausgewählt ist, auf ein kontrafaktisches Bild x_g , das in die Zielklasse c' durch den gegebenen Bildklassifikator (3) klassifiziert wird, mittels eines trainierten Generatornetzes (2).

2. Verfahren (100) nach Anspruch 1, das ferner die Folgenden Schritte umfasst:

- Vergleichen (130) des kontrafaktischen Bildes x_g mit dem Eingangsbild x ; und
- Bestimmen (140) von Bereichen (Δ), in denen das kontrafaktische Bild x_g sich von dem Eingangsbild x unterscheidet, als Bereiche (S), die mit Bezug auf die Klassengrenze zwischen der Quellenklasse c und der Zielklasse c' signifikant sind.

3. Verfahren (100) nach Anspruch 2, das ferner die folgenden Schritte umfasst:

- Bestimmen (150) mittels einer gegebenen Metrik (4), wie gut die Bereiche (S), die mit Bezug auf die Klassengrenze signifikant sind, mit gegebenen Bereichen (S^*) des Eingangsbildes x , die Merkmale des Eingangsbildes x enthalten, die als auffällig erachtet werden, in Übereinstimmung sind; und
- Zuschreiben (160) einer Bewertung (3a), die dem Ergebnis (4a) entspricht, das durch die gegebene Metrik (4) ausgegeben wird, zum Bildklassifikator (3).

4. Verfahren (100) nach Anspruch 3, das ferner Folgendes umfasst:

Optimieren (170) von Parametern (3b), die das Verhalten des gegebenen Bildklassifikators (3) charakterisieren, so dass, wenn die vorherigen Schritte (120-160) des Verfahrens (100) wiederholt werden, die Bewertung (3a) des Bildklassifikators (3) sich wahrscheinlich verbessert.

5. Verfahren (100) nach einem der Ansprüche 1 bis 4, wobei die Abbildung (120) der Darstellung z mit niedrigerer Dimension und der Zielklasse c' auf das kontrafaktische Bild x_g ferner Folgendes umfasst:

- Abbilden (121) des kontrafaktischen Bildes x_g auf eine Klassifikationsbewertung $c\#$ mittels des gegebenen Bildklassifikators (3);

- Bestimmen (122), wie gut die Klassifikationsbewertung $c\#$ mit einer Klassifikation des kontrafaktischen Bildes x_g in die Zielklasse c' in Übereinstimmung ist, mittels einer gegebenen Klassifikationsverlustfunktion (L_c); und

- Optimieren (123) mindestens einer Eingabe in das Generatornetz (2), so dass die erneute Berechnung des kontrafaktischen Bildes x_g auf der Basis der geänderten Eingabe wahrscheinlich bewirkt, dass sich der Wert der Klassifikationsverlustfunktion (L_c) verbessert.

6. Verfahren (100) nach einem der Ansprüche 1 bis 5, wobei das Eingangsbild x ein Bild eines hergestellten Produkts ist, das im Verlauf einer optischen Untersuchung des Produkts erfasst wurde, und die Klassen der gegebenen Klassifikation Qualitätsstufen für das Produkt darstellen.

7. Verfahren (100) nach einem der Ansprüche 1 bis 5, wobei das Eingangsbild x ein Bild einer Verkehrssituation ist und die Klassen der gegebenen Klassifikation Objekte darstellen, die für die Interpretation der Verkehrssituation relevant sind.

8. Verfahren (200) zum Trainieren einer Kombination eines Codierernetzes (1) und eines Generatornetzes (2) für die Verwendung im Verfahren (100) nach einem der Ansprüche 1 bis 5, das die folgenden Schritte umfasst:

- Bereitstellen (210) eines Decodierernetzes (5), das dazu konfiguriert ist, eine Darstellung z mit einer niedrigeren Dimension in einem latenten Raum, die vom Codierernetz (1) erhalten wurde, auf ein rekonstruiertes Bild x_d in der Domäne von ursprünglichen Eingangsbildern x abzubilden;
- Optimieren (220) von Parametern (1a, 5a), die das Verhalten der Netze des Codierers (1) und des Decodierers (5) charakterisieren, mit dem Ziel, dass das rekonstruierte Bild x_d dem ursprünglichen Eingangsbild x entspricht, von dem die Darstellung z erhalten wurde;
- Bereitstellen (230) eines Diskriminatornetzes (6), das dazu konfiguriert ist zu unterscheiden, ob ein Bild von der Domäne von ursprünglichen Eingangsbildern x oder von der Domäne von erzeugten Bildern x_f stammt;
- Bereitstellen (240) eines Bildklassifikators (3), der dazu konfiguriert ist, das ursprüngliche Eingangsbild x und das erzeugte Bild x_f auf eine oder mehrere Klassen der gegebenen Klassifikation abzubilden;
- Optimieren (250) von Parametern (2a), die das Verhalten des Generatornetzes (2) charakterisieren, mit den Zielen, dass
 - die Genauigkeit, mit der das Diskriminatornetz (5) zwischen ursprünglichen Eingangsbildern x und erzeugten Bildern x_f unterscheidet, abnimmt, und
 - der Bildklassifikator (3) erzeugte Bilder x_f auf ihre gegebene Zielklasse c abbildet; und
- Optimieren (260) von Parametern (6a), die das

Verhalten des Diskriminatornetzes (6) charakterisieren, mit dem Ziel, dass die Genauigkeit, mit der das Diskriminatornetz (6) zwischen ursprünglichen Eingangsbildern x und erzeugten Bildern x_f unterscheidet, zunimmt.

9. Verfahren (200) nach Anspruch 8, wobei

- die Parameter (1a), die das Verhalten des Codierernetzes (1) charakterisieren, zusätzlich auf das Ziel hin optimiert werden (221), dass gegenseitige Informationen zwischen dem ursprünglichen Eingangsbild x und seiner Darstellung z maximiert werden; und/oder
- die Parameter (2a), die das Verhalten des Generatornetzes (2) charakterisieren, zusätzlich auf das Ziel hin optimiert werden (251), dass gegenseitige Informationen zwischen dem erzeugten Bild x_f und der Darstellung z maximiert werden.

10. Verfahren (200) nach einem der Ansprüche 8 bis 9, wobei

- der Bildklassifikator (3) ein trainierbares Netz umfasst (241), das dazu konfiguriert ist, Eingangsbilder x und erzeugte Bilder x_f auf eine Kombination einer Darstellung z mit niedrigerer Dimension im latenten Raum und einer Klassifikationsbewertung $c\#$ abzubilden; und
- Parameter (3b), die das Verhalten dieses trainierbaren Netzes charakterisieren, mit den Zielen optimiert werden (242), dass:
oder Bildklassifikator (3) ein ursprüngliches Eingangsbild x auf eine Kombination von: einer Darstellung z , die der Darstellung z entspricht, die durch das Codierernetz (1) erzeugt wird, und einer Klassifikationsbewertung $c\#$, die mit einer Ground-Truth-Kennung c des Eingangsbildes x konsistent ist, abbildet; und
oder Bildklassifikator (3) ein erzeugtes Bild x_f auf eine Kombination von: einer Darstellung z , die der Darstellung z entspricht, von der das erzeugte Bild x_f erzeugt wurde, und einer Klassifikationsbewertung $c\#$, die mit der Zielklasse c' konsistent ist, für die das erzeugte Bild x_f erzeugt wurde, abbildet.

11. Vorrichtung (10), um aus einem Quellenbild x , das ein Teil einer Quellenverteilung von Bildern ist, ein Zielbild x_g , das ein anderes Muster aus derselben Quellendomäne ist, zu erzeugen, die Folgendes umfasst:

- ein trainiertes Codierernetz (1), das dazu konfiguriert ist, das Quellenbild x in eine Darstellung z mit niedrigerer Dimension in einem latenten Raum abzubilden; und
- ein trainiertes Generatornetz (2), das dazu konfiguriert ist, zumindest die Darstellung z mit niedrigerer Dimension auf das Zielbild x_g abzubilden.

12. Vorrichtung (10) nach Anspruch 11, wobei das Generatornetz (2) dazu konfiguriert ist, eine Kombination der Darstellung z mit niedrigerer

Dimension und einer Zielklasse c' auf ein Zielbild x_g abzubilden, das ein Mitglied der Zielklasse c' ist, wobei die Zielklasse c' eine von mehreren Klassen ist, die in einer vorbestimmten Klassifikation von Bildern in der Quellendomäne enthalten sind.

13. Computerprogramm mit maschinenlesbaren Befehlen, die, wenn sie durch einen oder mehrere Computer ausgeführt werden, bewirken, dass der eine oder die mehreren Computer ein Verfahren (100, 200) nach einem der Ansprüche 1 bis 10 durchführen.

14. Nichttransistorisches Speichermedium mit dem Computerprogramm nach Anspruch 13.

15. Ein oder mehrere Computer mit dem Computerprogramm nach Anspruch 13 und/oder mit dem nichttransistorischen Speichermedium nach Anspruch 14.

Es folgen 3 Seiten Zeichnungen

Anhängende Zeichnungen

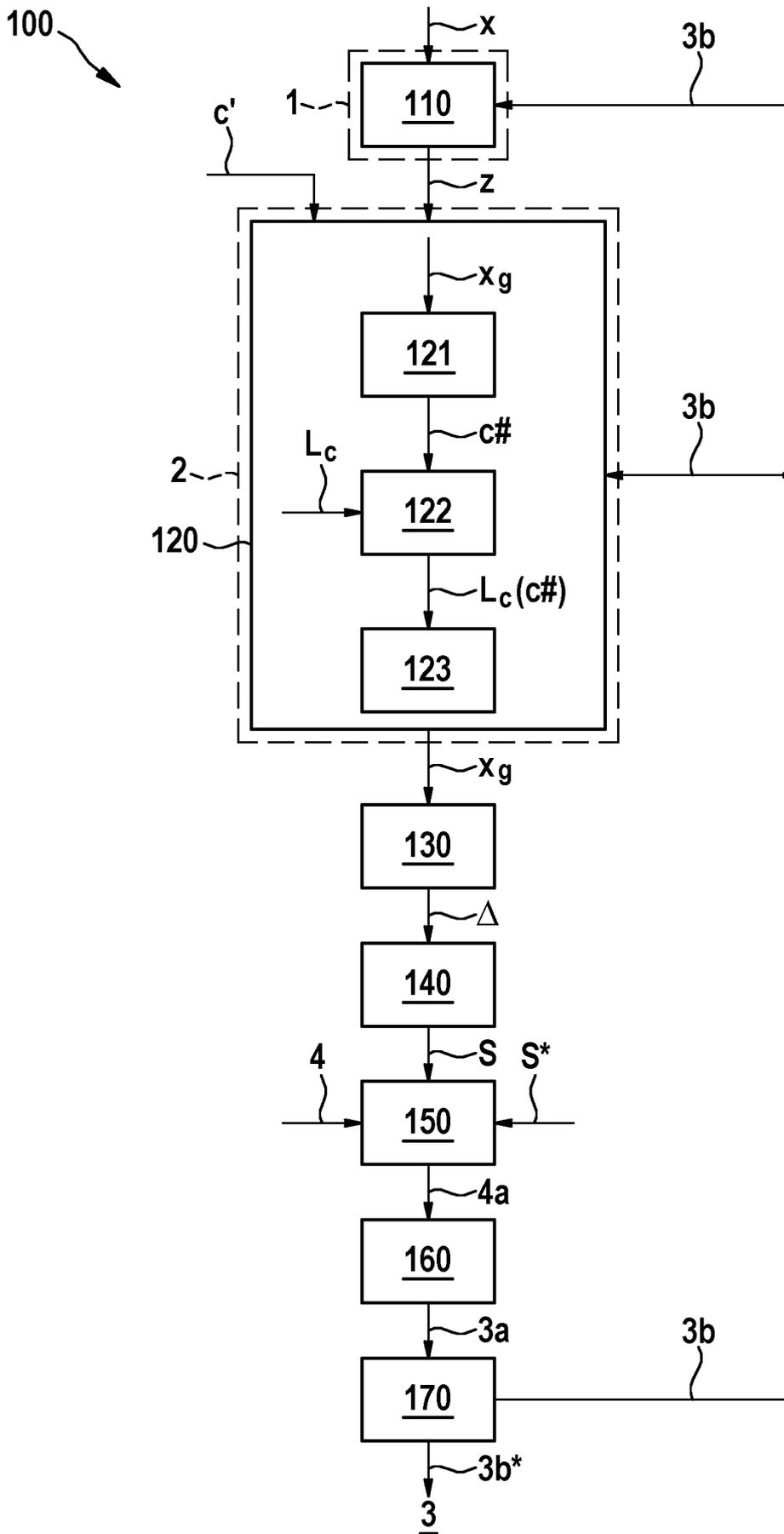


Fig. 1

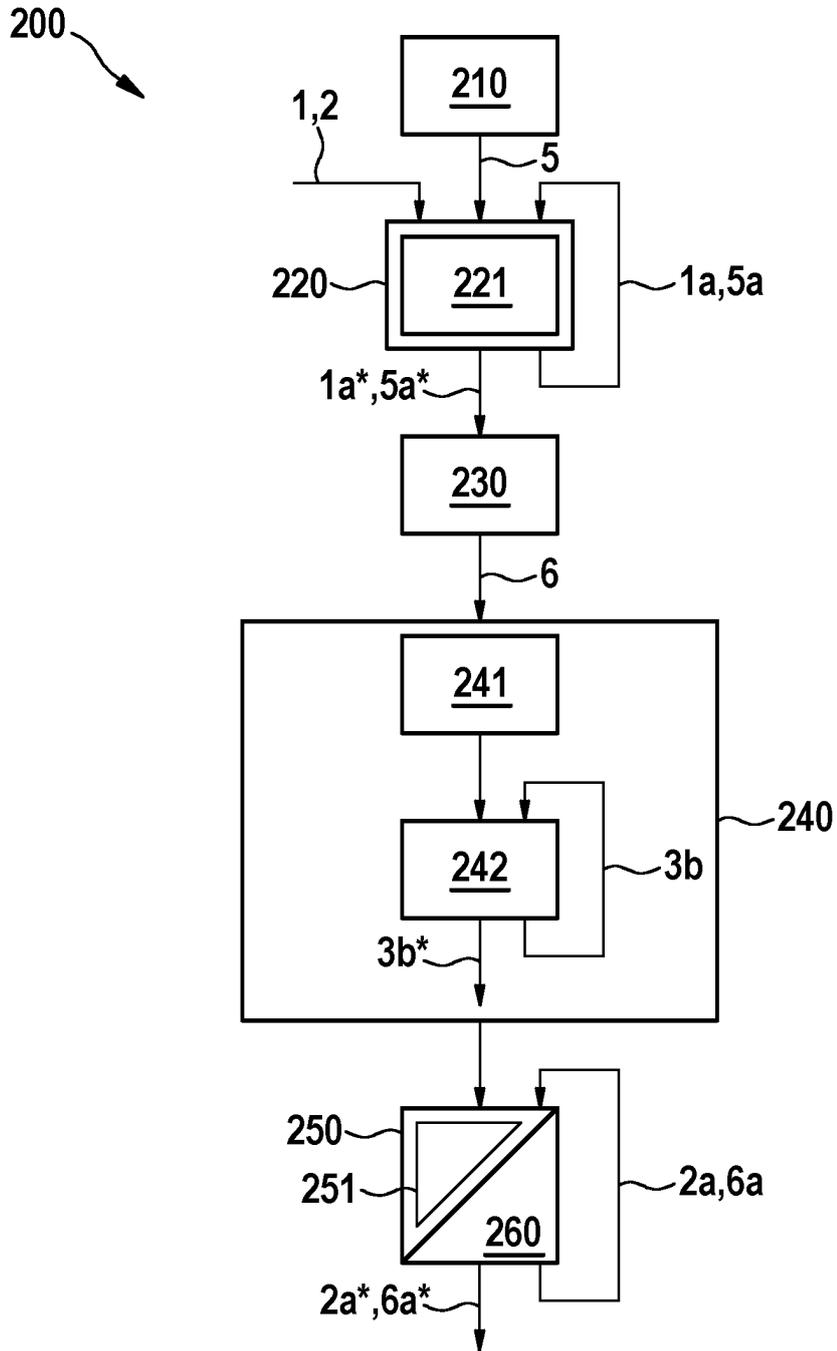


Fig. 2

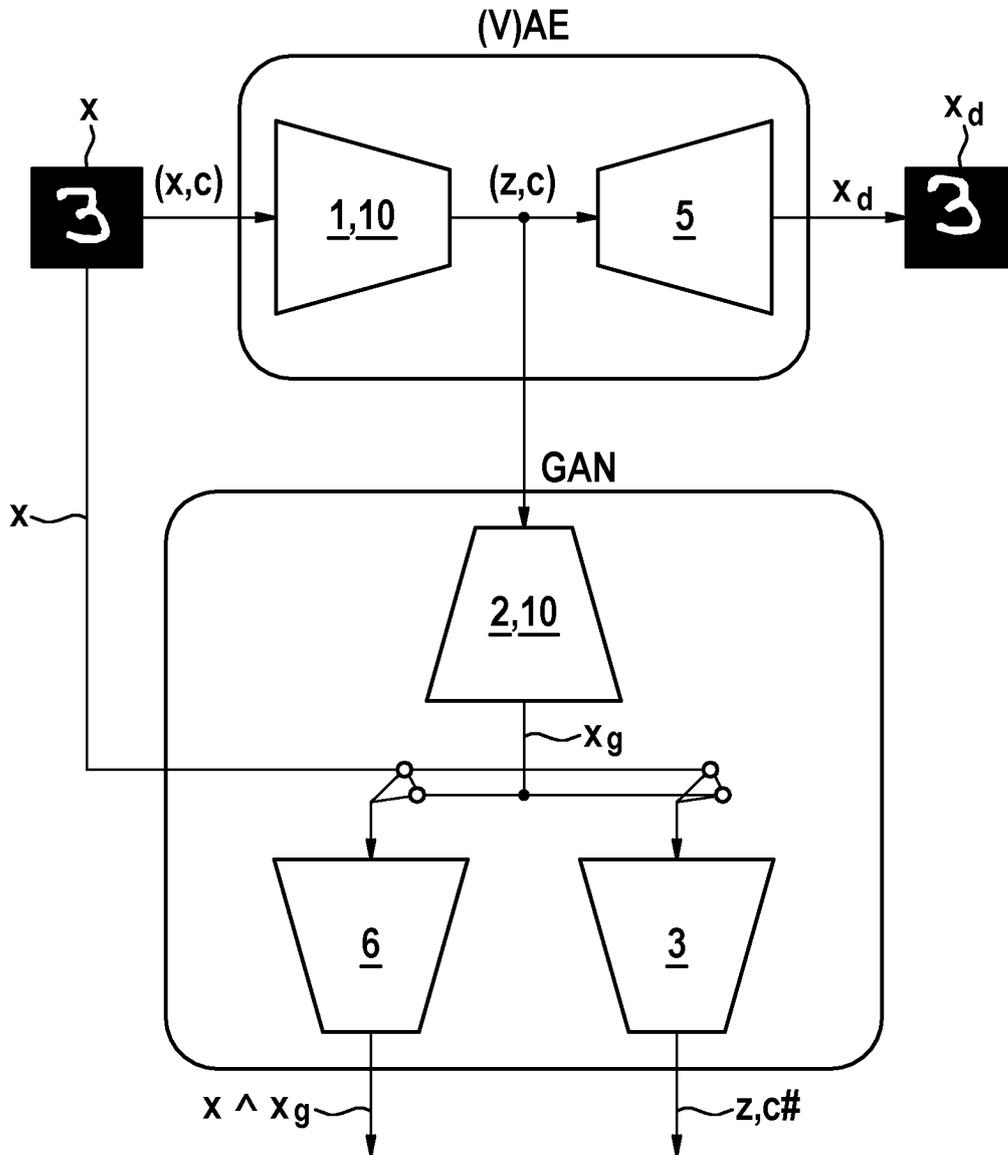


Fig. 3