GENERACIÓN DE DATOS TEXTUALES

CAMPO TÉCNICO

La presente divulgación se refiere a la generación de datos textuales que describen a un usuario de un dispositivo.

ANTECEDENTES

El reconocimiento temprano de la enfermedad es fundamental para el inicio oportuno del tratamiento. Varias enfermedades pueden manifestar síntomas físicos en el rostro de un paciente, y podría ser que los médicos astutos detectaran estos cambios sutiles como parte de su valoración.

Por ejemplo, el síndrome de Cushing ocurre cuando el organismo de un paciente se expone a altas concentraciones de la hormona cortisol durante un largo periodo. Algunos de los síntomas son un rostro redondeado, rojizo y lleno –con frecuencia denominado «cara de luna llena»– y estrías de color púrpura o rosa en la piel. El hipotiroidismo puede producir hinchazón facial, particularmente alrededor de los ojos, y la enfermedad de Parkinson tiene síntomas que incluyen una expresión similar a una máscara y una menor frecuencia de parpadeo.

La gestalt clínica es una herramienta diagnóstica que usa el análisis de los rasgos faciales de los pacientes para ayudar a detectar enfermedades. El valor de la gestalt clínica como herramienta diagnóstica se ha estudiado para el síndrome coronario agudo, la insuficiencia cardíaca, la neumonía, la Covid 19 y la sepsis. La gestalt clínica que proporcionaron y registraron los médicos fue comparable a los puntajes clínicos usados en la «decisión» y «exclusión» de pacientes que se presentaron en el departamento de emergencias con ciertos síntomas. De acuerdo con los impulsores del Tercer Consenso Internacional para la definición de sepsis y shock séptico (sepsis 3), se recomienda a los clínicos que, además de los criterios del síndrome de respuesta

inflamatoria sistémica (SIRS), usen la gestalt clínica en la selección, tratamiento y estratificación de riesgo de los pacientes con infección (véase “Deep Learning for Identification of Acute Illness and Facial Cues of Illness” de Castela Forte et al, Font. Med., 26 de julio de 2021 Sec. Translation Medicine volumen 8 – 2021, https://doi.org/10.3389/fmed.2021.661309).

Recientemente, se demostró que los modelos de aprendizaje profundo que se entrenaron con una variedad de mediciones clínicas predijeron o detectaron signos tempranos de enfermedades.

En el documento “Deep Learning for Identification of Acute Illness and Facial Cues of Illness”, los autores idearon un algoritmo de aprendizaje profundo para distinguir entre individuos sanos y enfermos graves sobre la base de un análisis de los rasgos faciales. Usaron un conjunto de datos sintético que se creó al aplicar maquillaje en los individuos sanos para simular los rasgos faciales característicos de la enfermedad aguda.

El conjunto de datos sintético se creó sobre las fotografías de veintiséis voluntarios, tomadas con una cámara de teléfono inteligente. Se seleccionaron e incluyeron dos fotografías de cada participante en el estudio: una sin maquillaje para representar el estado de control «sano» y la otra para representar el estado de «enfermo grave». Se usó un entorno estandarizado con fondo gris y luz del diodo fotoemisor (LED). Un fotógrafo profesional estandarizó el balance de blancos de todo el conjunto de fotografías mediante una herramienta muy conocida de edición de fotos.

Los llamados dispositivos «espejo inteligente» se pueden usar para vigilar los rasgos faciales, la afección cutánea y otros atributos físicos de una persona. La solicitud de patente estadounidense US 2018/253840 propone un «espejo inteligente» como un sistema de inteligencia artificial (IA) que usa las tecnologías de análisis de imágenes y reconocimiento de patrones, que combina una visualización y un sistema de aprendizaje

automático (ML). Entre múltiples características, el espejo propuesto puede vigilar, con diferentes sensores, los datos relacionados con la salud de un usuario, tales como signos vitales, calidad del sueño, actividad física, mediciones corporales, así como también estrés u otros estados de salud mental basados en la conducta e interacción del usuario. Otra solicitud de patente relacionada con un espejo inteligente (solicitud de patente coreana KR 20220052439) propone una medición sin contacto de la temperatura del usuario. El documento US 2020/342987 divulga un sistema para el intercambio de información que comprende un espejo configurado para captar la información facial de un usuario y visualizar la información de inferencia sobre el usuario. Un dispositivo de computación en espejo está configurado para recibir y procesar la información facial del usuario y producir la información de inferencia sobre el usuario. El espejo está acoplado con un *backend* basado en la nube y, juntos, están dispuestos en un marco de aprendizaje federado, con un sensor y/o pesos del modelo que se envían entre el espejo y el servidor basado en la nube.

El rendimiento de los sistemas de IA, particularmente aquellos basados en los procedimientos de aprendizaje profundo, depende considerablemente de la cantidad y calidad de los datos de entrenamiento disponibles. En el contexto del análisis de imágenes y videos, la abundancia del conjunto de datos puede influir de manera significativa en la precisión y eficacia de los modelos de IA. No obstante, con el desarrollo y mejora continua de las tecnologías de aprendizaje automático y video inteligente, ha habido un desarrollo paralelo en las tecnologías de anonimización de imágenes.

En las etapas tempranas de anonimización de imágenes, los procedimientos tales como enmascaramiento, ofuscación o pixelación se usaron típicamente para esconder información sensible en las imágenes. No obstante, alteran permanentemente la imagen y es posible que se pierda información importante. Por lo general, requieren una

identificación manual de las zonas para anonimizar, lo cual puede ser poco práctico para grandes conjuntos de datos que serían aptos como datos de prueba o entrenamiento en el aprendizaje automático.

La solicitud de patente WO 2023/060918 “Image anonymization method based on guidance of semantic and pose graphs” divulga un procedimiento para anonimizar toda la imagen al tiempo que se mantienen la información semántica original de la imagen y la información de pose de la(s) persona(s). En este procedimiento, los rostros humanos, los cuerpos y el fondo son reemplazados totalmente por mapas semánticos abstractos. El documento WO 2023/060918 sugiere que la información sobre la semántica, la postura de carácter y el movimiento de objetos de las imágenes puede seguir proporcionando una gran cantidad de datos de entrenamiento para los modelos de IA.

SUMARIO

Sigue habiendo un conflicto entre la cantidad y calidad de los datos de entrenamiento disponibles para los modelos de IA, y la necesidad de anonimización de las imágenes para aumentar la seguridad. A pesar de la gestalt clínica o análisis de los rasgos faciales que se usa cada vez más para construir modelos de aprendizaje profundo, el desarrollo se enlentece por la falta de conjuntos de datos que sean suficientemente grandes y la calidad de las imágenes.

El documento WO 2023/060918 no resuelve el conflicto antes mencionado porque puede proporcionar datos de entrenamiento usables solamente para el desarrollo y optimización de la detección de humanoides, la detección de movimiento y otros modelos de algoritmo de IA que no tienen grandes requisitos sobre los rostros humanos.

El documento “Deep Learning for Identification of Acute Illness and Facial Cues of Illness” presenta varias limitaciones: un conjunto de datos de entrenamiento pequeño; podría ser que las características de la enfermedad simulada no representaran adecuadamente el espectro de

las presentaciones de enfermedad aguda, la población de estudio era predominantemente caucásica, lo que limita la generalizabilidad del modelo a otros orígenes étnicos; falta de pruebas reales; y las fotografías de pacientes reales presentarían desafíos como condiciones de alumbrado discrepantes; problemas éticos y de seguridad.

De acuerdo con el documento US 2018/253840, el espejo inteligente recopila una gran cantidad de datos personales; pero esto podría ocasionar problemas de seguridad significativos y un potencial mal uso de datos. El documento US 2018/253840 no se centra en el algoritmo de análisis facial para distinguir entre individuos sanos y enfermos graves sobre la base de los rasgos faciales.

Por lo tanto, sigue habiendo una serie de problemas con la generación de datos aptos a partir de las imágenes de un sujeto.

Ciertos aspectos de la divulgación y sus formas de realización pueden proporcionar soluciones a estos u otros desafíos.

Se proponen un dispositivo y un procedimiento capaces de recopilar un gran conjunto de datos sensibles aptos para poner a prueba o entrenar modelos de IA al tiempo que se aseguran los datos de un usuario. Los datos sensibles pueden incluir datos relacionados con los rasgos faciales, la voz y los rasgos corporales de un usuario, que pueden relacionarse con, pero no se limitan a, la vigilancia de la salud y la enfermedad.

Una forma de realización del dispositivo propuesto es una superficie (p. ej. una superficie reflectante, tal como un espejo), que puede usarse en el hogar, en un espacio público o en un establecimiento de asistencia sanitaria. El dispositivo puede ser portátil o montado en la pared / montable en la pared. El dispositivo (p. ej. una superficie, un espejo, etcétera) puede comprender una o más cámaras y, opcionalmente, otros sensores, alguna forma de tecnología de la comunicación, tal como wifi o Bluetooth, un sistema de iluminación/alumbrado para iluminar al usuario y una o más máquinas/procesadores de ML. La(s) cámara(s) puede(n) obtener una(s) imagen(es) del usuario en frente del espejo/dispositivo. El

dispositivo usa la(s) cámara(s) para reunir datos personales y extrae los datos textuales de las imágenes, de modo tal que los individuos no pueden reconocerse a partir de los datos textuales emitidos por el dispositivo.

El procedimiento propuesto puede usar uno o más algoritmos de ML/IA para analizar las imágenes obtenidas mediante la una o más cámaras y los sensores opcionales. Los algoritmos de ML/IA pueden ser o incluir modelos de detección de objetos, modelos de aprendizaje profundo, modelos de lenguaje, modelos de lenguaje multimodales, grandes modelos de lenguaje (LLM) y/o grandes modelos de lenguaje (LLM) multimodales. Se pueden usar algoritmos de visión mediante computadora para identificar rasgos faciales claves, tales como los ojos, la nariz, la boca, la piel, las cejas, la boca y el rabillo de los ojos, y otros detalles más finos como los iris, el tono de piel, las variaciones en la textura de la piel, la pigmentación, los círculos debajo de los ojos, etc. El dispositivo usa un algoritmo/sistema de IA embebido de imagen a texto para traducir los datos recopilados por contenido textual. El audio capturado por la(s) cámara(s) u otros micrófonos puede ser registrado/almacenado, particularmente el audio relacionado con la tos, la respiración, etcétera, y un LLM multimodal puede traducir la información representada en esa captura de audio, y las imágenes, por datos textuales.

De este modo, la entrada en el sistema/procedimiento puede ser un conjunto de imágenes faciales y/o corporales de uno o más usuarios, registradas/obtenidas por la(s) cámara(s) que está(n) embebida(s) en, o de lo contrario forma(n) parte de, el dispositivo. La salida del sistema/procedimiento puede ser una traducción detallada de la(s) imagen(es) por el contenido de descripción textual que incluye o se relaciona con el rostro, el cuerpo, la piel, los ojos, el cabello, etc., del usuario.

La(s) imagen(es) original(es) de la(s) cámara(s) se

destruirá(n)/suprimirá(n) de forma permanente en el dispositivo luego de crear, procesar y guardar el contenido textual, y sin que la(s) imagen(es) se envíe(n) al servidor. Esto posibilitará que la seguridad de los datos del usuario esté completamente protegida. A continuación, el contenido textual puede enviarse a un servidor, por ejemplo a un servidor que gestiona una base de datos del entrenamiento de IA global para que se use en el entrenamiento del modelo de ML/IA, para detectar los estados de salud.

Ciertas formas de realización pueden proporcionar una o más de las ventajas técnicas siguientes. Las técnicas descritas en la presente memoria permiten reunir un único conjunto de datos de una manera segura. Los datos sensibles sobre el usuario pueden recopilarse y compartirse al tiempo que se respeta la seguridad de los datos del usuario. El conjunto de datos proporciona datos reales para la vigilancia de la salud. Las formas de realización estipulan que las características de la enfermedad representarán de manera adecuada el espectro amplio de la enfermedad aguda y no habrá ninguna limitación en la generalizabilidad o aplicabilidad de un modelo subsiguientemente entrenado a otros orígenes étnicos, lo cual significa que el rendimiento y precisión de cualquier sistema de vigilancia o diagnóstico que funcione sobre la base de dichos datos se mejorará de manera significativa. Ciertas formas de realización pueden reducir el efecto de iluminación en los datos. Las técnicas, y/o un sistema que usa un modelo entrenado sobre la base del conjunto de datos recopilados, también pueden posibilitar la descarga de decisiones diagnósticas de los hospitales y/o de médicos con experiencia. Para ciertos tipos de enfermedad, es posible detectar signos tempranos de advertencia sin que sea necesario concurrir regularmente al hospital.

De acuerdo con un primer aspecto, se da a conocer un procedimiento implementado por computadora para operar un dispositivo. El procedimiento en el dispositivo comprende obtener una o más imágenes de un usuario del dispositivo; usar un modelo de lenguaje para

generar datos textuales que describen al usuario mostrado en la una o más imágenes; y enviar los datos textuales generados a un servidor.

De acuerdo con un segundo aspecto, se da a conocer un dispositivo que está configurado para obtener una o más imágenes de un usuario del dispositivo; usar un modelo de lenguaje para generar datos textuales que describen al usuario mostrado en la una o más imágenes; y enviar los datos textuales generados a un servidor.

De acuerdo con un tercer aspecto, se da a conocer un dispositivo que comprende un procesador y una memoria, conteniendo dicha memoria instrucciones ejecutables mediante dicho procesador, de manera que dicho dispositivo es operativo para obtener una o más imágenes de un usuario del dispositivo; usar un modelo de lenguaje para generar datos textuales que describen al usuario mostrado en la una o más imágenes; y enviar los datos textuales generados a un servidor.

De acuerdo con un cuarto aspecto, se da a conocer un producto de programa informático que comprende un medio legible por computadora que tiene un código legible por computadora incorporado en el mismo, estando el código legible por computadora configurado de modo tal que, cuando se ejecuta mediante una computadora o procesador adecuado, la computadora o procesador hace que se realice el procedimiento de acuerdo con el primer aspecto o cualquier forma de realización del mismo.

De acuerdo con un quinto aspecto, se da a conocer un sistema de vigilancia de la salud que comprende: un dispositivo según el segundo o tercer aspecto, o cualesquier formas de realización de los mismos; y un servidor configurado u operativo para recibir los datos textuales del dispositivo.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

Algunas de las formas de realización contempladas en la presente memoria se describirán a continuación de manera más completa con referencia a los dibujos adjuntos, en los que:

la Fig. 1 ilustra un dispositivo como un espejo, de acuerdo con una forma de realización, y los ejemplos de entradas y salidas de la conversión de imagen a texto, de acuerdo con las formas de realización de las técnicas descritas en la presente memoria;

la Fig. 2 es un diagrama de bloques simplificado de un dispositivo, de conformidad con una o más formas de realización;

la Fig. 3 es un diagrama de bloques que ilustra un generador de datos textuales ejemplificador; y

la Fig. 4 es un diagrama de flujo que ilustra un procedimiento, de conformidad con algunas formas de realización.

DESCRIPCIÓN DETALLADA

Algunas de las formas de realización contempladas en la presente memoria se describirán a continuación de manera más completa con referencia a los dibujos adjuntos. Las formas de realización se proporcionan a modo de ejemplo para transmitir el alcance del objeto a los expertos en la técnica.

La Fig. 1 ilustra un dispositivo 102 como una superficie, tal como un espejo, de acuerdo con una forma de realización, junto con los ejemplos de entradas 104 y salidas 106 de un generador de datos textuales 108. El generador de datos textuales 108 es, o comprende, un modelo de lenguaje, tal como un gran modelo de lenguaje (LLM), y, en formas de realización en particular, es un modelo de lenguaje multimodal (p. ej. un LLM multimodal) que puede generar datos textuales a partir de entradas no textuales, tales como video, imágenes, señales de audio, etc., o una combinación de los mismos. Como se analiza en más detalle a continuación con referencia a la Fig. 3, el generador de datos textuales 108 puede incluir módulos o funciones que pueden procesar entradas no textuales, p. ej., para codificar y/o «percibir» el contenido de las entradas no textuales, para posibilitar que el modelo de lenguaje genere datos textuales para esas entradas. Alternativamente, por ejemplo para un

modelo de lenguaje multimodal, estos módulos o funciones pueden integrarse en, o de lo contrario formar parte de, el modelo de lenguaje. En la explicación siguiente, las expresiones «modelo de lenguaje» y «LLM» se usan generalmente de manera intercambiable. Los datos textuales generados mediante el modelo de lenguaje se envían a un servidor. El servidor puede usar los datos textuales para cualquier propósito adecuado, tal como para entrenar y/o poner a prueba un modelo de ML, establecer un sistema basado en reglas o un sistema experto, y/o para la vigilancia y/o visualización directa por un profesional sanitario, por ejemplo. El modelo de ML, el sistema basado en reglas o el sistema experto puede ser para cualquier propósito en particular, p. ej. para diagnosticar un estado de salud, una enfermedad, un padecimiento, etc., pero se entenderá que el propósito del modelo o sistema de ML no tiene relevancia en el procedimiento de recopilación y generación de los datos textuales, como se describe en la presente memoria.

Como se observa, el modelo de lenguaje puede un LLM. Los modelos actualmente disponibles suelen tener un tamaño de varios gigabytes. Por ejemplo, el modelo GPT-3 de OpenAI con 175 mil millones de parámetros tiene un tamaño de alrededor de 350 GB. No obstante, recientemente han surgido LLM ligeros con un muy buen rendimiento, que pueden ejecutar la inferencia mediante un hardware que es económico y no demasiado pesado (<600 USD). Un ejemplo de un LLM ligero es Alpaca de la Universidad de Stanford (7B), en el que, con solo 52 mil demostraciones siguiendo las instrucciones, estuvo cerca de los modelos de OpenAI.

Para las técnicas descritas en la presente memoria, puede ser necesario un modelo multimodal basado en GPT con <100 mil imágenes e informes correspondientes de los especialistas médicos. A continuación, la red resultante podría destilarse en un modelo más pequeño. El GPT-4 es multimodal y tiene 1,76 billones de parámetros. Si se supone que una red destilada tiene una proporción similar a la de

Davinci de OpenAI frente a Alpaca, puede lograrse un tamaño de red de aproximadamente 70B. Ya hay versiones tales como MiniGPT-4, como se describe en “Minigpt-4: Enhancing Vision-Language Understanding with Advanced Large Language Models” de Zhu et al., https://arxiv.org/pdf/2304.10592.pdf. Puede considerarse el uso de modelos con barreras de seguridad para fines médicos, para alinearse con reglas y limitaciones predefinidos, por ejemplo para evitar consejos que no sean médicos.

El dispositivo 102 puede ser un tipo de espejo inteligente, que puede usarse en el hogar de una persona o colocarse en hospitales/clínicas y/u otros espacios públicos.

Si bien las formas de realización se describen en la presente memoria con referencia al usuario que es una persona, se entenderá que el usuario 110 puede ser, en cambio, un animal, tal como una mascota, por ejemplo un perro o un gato.

La Fig. 2 es un diagrama de bloques simplificado del dispositivo 102, de conformidad con una o más formas de realización. El dispositivo 102 comprende una circuitería (o lógica) de procesamiento 201. Se entenderá que el dispositivo 102 puede comprender una o más máquinas virtuales que ejecutan diferentes programas y/o procedimientos. La circuitería de procesamiento 201 controla el funcionamiento del dispositivo 102 para implementar los procedimientos descritos en la presente memoria y, en particular, usa el modelo de lenguaje / LLM para generar datos textuales para el usuario 110 a partir de las imágenes de entrada 112/114. La circuitería de procesamiento 201 puede comprender uno o más procesadores, unidades de procesamiento, procesadores o módulos de múltiples núcleos que están configurados o programados para controlar el dispositivo 102 de la manera descrita en la presente memoria. En implementaciones en particular, la circuitería de procesamiento 201 puede comprender una pluralidad de módulos de software y/o hardware que están, cada uno, configurados para realizar, o son para realizar, pasos

individuales o múltiples del procedimiento descrito en la presente memoria, en relación con el dispositivo 102.

El dispositivo 102 también comprende una interfaz de comunicaciones 202. La interfaz de comunicaciones 202 se usa para posibilitar las comunicaciones con otros dispositivos, teléfonos inteligentes del usuario, computadoras, servidores remotos, etc. La interfaz de comunicaciones 202 puede estar configurada para transmitir hacia y/o recibir desde otros dispositivos, teléfonos inteligentes, computadoras o servidores, solicitudes, acuses de recibo, información, datos, señales o similares. Por ejemplo, la interfaz de comunicaciones 202 puede usarse para transmitir la salida 106 del texto a un servidor que maneja y/o almacena los datos para usar en el entrenamiento de un modelo del ML/IA. La interfaz de comunicaciones 202 puede usar cualquier tecnología de comunicación adecuada, por ejemplo wifi, Bluetooth, etcétera.

El dispositivo 102 puede comprender una memoria 203. En algunas formas de realización, la memoria 203 puede estar configurada para almacenar un código de programa que puede ejecutarse mediante la circuitería de procesamiento 201 para realizar los procedimientos descritos en la presente memoria, en relación con el dispositivo 102. Alternativa o adicionalmente, la memoria 203 puede estar configurada para almacenar cualquier dato de entrada 104 (p. ej. imágenes, datos de imágenes, mediciones de sensores), datos requeridos por el modelo de lenguaje / LLM 108 (que incluye el mismo modelo de lenguaje / LLM), y cualquier dato de salida 106 (p. ej. descripción textual de cualquier parte de los datos de entrada 104). La circuitería de procesamiento 201 puede estar configurada para controlar la memoria 203, para almacenar dicha información en la misma.

El dispositivo 102 comprende, o está asociado a, uno o más sensores de imágenes 207 (p. ej. las cámaras que se usan para obtener/generar una o más imágenes de un usuario 110 que está en

frente del dispositivo espejo 102. El uno o más sensores de imágenes 207 pueden ser integrales al dispositivo 102 o pueden estar separados del dispositivo 102 y proporcionar las imágenes obtenidas al dispositivo 102 / a la circuitería de procesamiento 201 para que se procesen. El uno o más sensores de imágenes 207 pueden registrar una transmisión de video que comprende una serie de imágenes. El uno o más sensores de imágenes 207 pueden incluir un sensor de imágenes 207 de color rojo, verde y azul (RGB) para obtener imágenes basadas en la luz visible. Alternativa o adicionalmente, el uno o más sensores de imágenes 207 pueden incluir un sensor de imágenes 207 infrarrojo (IR) para obtener imágenes mediante la luz IR. El uno o más sensores de imágenes 207 deberían proporcionar imágenes de una calidad adecuada (p. ej. alta resolución, alta frecuencia de trama, foco, etc.) para posibilitar una extracción de rasgos faciales eficaz y precisa. Algunas formas de realización pueden incluir múltiples sensores de imágenes 207 dispuestos para obtener imágenes del usuario 110 desde múltiples ángulos, lo cual posibilita que se obtengan informaciones/imágenes tridimensionales (3D) del usuario 110.

La Fig. 2 muestra una o más fuentes de luz 209 que se usan para iluminar al usuario 110 cuando se obtienen imágenes mediante el (los) sensor(es) de imágenes 207. La presencia y/o el uso de la una o más fuentes de luz 209 es opcional. Preferentemente, la una o más fuentes de luz 209 están dispuestas o son controlables para iluminar de manera uniforme al usuario 110, p. ej. para minimizar las sombras, y permitir que se obtengan imágenes precisas en el color / consistentes en el color. De este modo, la(s) fuente(s) de luz 209 debería(n) generar una luz adecuada y puede(n) ser, por ejemplo, lámparas de luz de día natural, luces LED (p. ej. como luces con tiras de luz ubicadas en los laterales, los bordes superiores y/o inferiores del espejo 102), luces de tocador montadas en cualquier lateral del espejo 102, al nivel del ojo para eliminar las sombras en el rostro, retroiluminación que proporciona iluminación

difusa desde atrás de la superficie del espejo para una distribución de luz suave y uniforme, etc. La(s) fuente(s) de luz 209 puede(n) formar parte del dispositivo 102 o puede(n) estar separada(s) del dispositivo 102. En cualquier caso, la circuitería de procesamiento 201 del dispositivo 102 puede controlar el funcionamiento de la(s) fuente(s) de luz 209.

La Fig. 2 muestra uno o más de «otros» sensores 211 opcionales que se usan para obtener mediciones de una o más características del usuario 110 y/o del entorno alrededor del usuario 110. Cualquier sensor 211 en particular puede formar parte del dispositivo 102 o puede estar separado del dispositivo 102. En cualquier caso, la circuitería de procesamiento 201 del dispositivo 102 puede recibir mediciones o una señal de medición del (de los) sensor(es) 211.

Otro sensor 211 / Un sensor 211 adicional puede ser un micrófono que se usa para registrar los sonidos asociados al usuario 110, tales como el sonido de la tos y/o la respiración de un usuario. Los micrófonos también pueden usarse para captar otra entrada de audio, lo cual permite que los usuarios 110 interactúen con el espejo 102 mediante comandos de voz, etcétera.

Otro sensor 211 / Un sensor 211 adicional puede ser un termómetro, p. ej. un termómetro IR, que se usa para medir la temperatura corporal del usuario 110. El termómetro puede medir la temperatura de manera remota (p. ej. sin que el usuario 110 tenga que tocar el dispositivo 102) o puede ser necesario que el usuario 110 tenga que tocar una parte del dispositivo 102 sensible a la temperatura con la mano u otra parte del cuerpo.

Otro sensor 211 / Un sensor 211 adicional puede ser un sensor de pesos, p. ej. balanzas, balanzas inteligentes, etcétera, que se usa para medir el peso del usuario.

Otro sensor 211 / Un sensor 211 adicional puede ser un monitor de presión arterial para medir la presión arterial del usuario.

Otro sensor 211 / Un sensor 211 adicional puede ser un sensor que

puede obtener información de la frecuencia cardíaca, tal como la frecuencia cardíaca y/o la variabilidad de la frecuencia cardíaca del usuario.

Otro sensor 211 / Un sensor 211 adicional puede ser un sensor que puede medir la frecuencia respiratoria del usuario.

Otro sensor 211 / Un sensor 211 adicional puede ser uno o más sensores de luz para medir la intensidad de la luz en el entorno alrededor del dispositivo 102 (que también se denomina luz ambiente o grado de luz ambiente). El dispositivo 102 puede usar una medición de la luz ambiente para: ajustar el brillo de la luz generada por la(s) fuente(s) de luz 209 para proporcionar un grado de iluminación apropiado; ajustar la sensibilidad de la luz del (de los) sensor(es) de imágenes 207; y/o ajustar el brillo de cualesquier componentes de visualización del dispositivo 102.

Otro sensor 211 / Un sensor 211 adicional puede ser uno o más sensores de proximidad que pueden detectar la presencia de un usuario 110 en frente del espejo 102 y activar la visualización del espejo y/o ajustar su brillo cuando el usuario 110 se acerca o se aleja del espejo 102.

Otro sensor 211 / Un sensor 211 adicional puede ser uno o más sensores de movimiento, tales como los sensores IR y/o ultrasónicos que pueden detectar movimiento en la cercanía del espejo 102. Las señales de mediciones del (de los) sensor(es) de movimiento pueden usarse para desencadenar acciones específicas, tales como encender la visualización del espejo 102 o activar ciertas características cuando se detecta movimiento.

Otro sensor 211 / Un sensor 211 adicional puede ser uno o más sensores táctiles que permiten que el usuario interactúe con la interfaz de control del espejo y/o controle diversas funciones.

Como se observó anteriormente, el dispositivo 102 puede ser, o comprender, una superficie. La superficie puede ser una superficie por la que el usuario puede pasar cuando realiza su rutina diaria, tal como

vidrieras, paredes del elevador, baños públicos, cuartos de baño, etc. La superficie puede ser una superficie reflectante, p. ej. un espejo. En algunas formas de realización, el espejo comprende una superficie espejada tradicional para reflejar la luz. El espejo también puede incluir una pantalla de visualización separada para exponer información al usuario 110, o los elementos de visualización pueden estar integrados en la superficie reflectante. En otras formas de realización, en vez de una superficie reflectante, el espejo puede comprender una pantalla de visualización que se usa para visualizar una transmisión de video del usuario 110, que se obtiene mediante el uno o más sensores de imágenes 207.

El dispositivo 102 puede formar parte de un sistema, tal como un sistema de vigilancia de la salud, que también comprende un servidor para recibir los datos textuales del dispositivo 102. Como se observó anteriormente, el servidor puede usar los datos textuales para cualquier propósito adecuado, tal como para entrenar y/o poner a prueba un modelo de ML, establecer un sistema basado en reglas o un sistema experto, y/o para la vigilancia y/o visualización directa por un profesional sanitario. El sistema de vigilancia de la salud puede generar emisiones de la vigilancia de la salud mediante el modelo de ML. Por ejemplo, en función del tipo de modelo de ML usado, una emisión de la vigilancia de la salud puede ser cualquiera de: una indicación de un estado de salud / padecimiento / enfermedad del usuario 110; una indicación respecto de si el usuario 110 debería buscar consejo médico de un profesional sanitario; asesoramiento sobre ejercicios/actividades/dieta que sean adecuados para cualesquier estados de salud detectados del usuario 110; etcétera.

La Fig. 1 muestra varias entradas 104 ejemplificadoras relacionadas con el usuario 110 que se obtienen mediante el dispositivo 102. En este ejemplo, las entradas 104 incluyen una primera imagen o primer conjunto de imágenes 112 que se obtienen en un primer momento mediante un sensor de imágenes 207. La primera imagen o primer conjunto de

imágenes 112 son de un usuario 110 que es sano o relativamente sano. Las entradas 104 también muestran una segunda imagen o segundo conjunto de imágenes 114 que se obtienen mediante el sensor de imágenes 207. La segunda imagen o segundo conjunto de imágenes 114 son de un usuario 110 que es enfermizo / está enfermo. Las entradas 104 también incluyen una medición de audio/sonido que se obtiene mediante un micrófono. La medición de audio/sonido 116 puede haberse obtenido al mismo tiempo o en un tiempo similar a la segunda imagen o segundo conjunto de imágenes 114.

Las entradas 104 se proporcionan al LLM 108, opcionalmente luego de que una o más de las entradas 104 se han procesado previamente, como se describe a continuación, y el LLM 108 genera datos textuales que describen al usuario 110. Los datos textuales generados pueden describir el rostro del usuario 110, la voz/respiración del usuario y/o pueden referirse a los aspectos de salud del usuario.

La Fig. 1 muestra tres salidas 106 textuales ejemplificadoras para cada una de las entradas 104. Si se ingresó la primera imagen 112, el LLM puede generar datos textuales 118 que comprenden el texto: «la piel tiene un color natural; los labios no presentan una coloración significativa; los ojos son brillantes y claros; la nariz no presenta un enrojecimiento persistente; la boca es de simetría normal». Si se ingresó la segunda imagen 114, el LLM puede generar datos textuales 120 que comprenden el texto: «tono de piel más pálido; labios pálidos; enrojecimiento alrededor de los ojos; ojos hundidos; enrojecimiento alrededor del ala de la nariz; boca caída». Para la señal de audio 116, el LLM genera datos textuales 122 que comprenden el texto seleccionado de entre: «tos seca/húmeda/ferina; voz ronca, débil o nasal; respiración rápida, superficial; estornudo y aspiración». Se entenderá que, cuando la señal de audio 116 y una de las imágenes / conjuntos de imágenes 112/114 se ingresa en el LLM, el LLM puede proporcionar una salida de datos textuales sobre la base de ambos tipos de entrada 104.

Las imágenes 112/114 ingresadas en el LLM 108 pueden ser estáticas o dinámicas (p. ej. una transmisión de video / secuencia de imágenes). El LLM puede recibir directamente la(s) imagen(es) 112/114 y generar los datos textuales a partir de la(s) imagen(es) 112/114. Alternativamente, la(s) imagen(es) 112/114 pueden procesarse previamente y la emisión del procesamiento previo puede proporcionarse al LLM para usar en la generación de los datos textuales. En este caso, la emisión del procesamiento previo y la(s) imagen(es) 112/114 original(es) pueden proporcionarse al LLM, y ambas pueden usarse para generar los datos textuales. Por ejemplo, como se describe con referencia a la Fig. 3 a continuación, se puede usar un módulo de codificación y/o módulo de percepción para codificar la(s) imagen(es) y/o percibir el contenido de la(s) imagen(es), generando el LLM los datos textuales sobre la base de la emisión del módulo de codificación y/o módulo de percepción. Alternativamente, el procesamiento previo de la(s) imagen(es) 112/114 puede comprender cualquiera de:

- Un algoritmo de detección de objetos que se usa para identificar los objetos (es decir, parte(s) del cuerpo) en la(s) imagen(es). El algoritmo de detección de objetos puede proporcionar una caja delimitadora alrededor del (de los) objeto(s) detectado(s).
- Un algoritmo de segmentación semántica que se puede usar para segmentar la imagen en información a nivel de píxel.
- Un algoritmo de clasificación facial, tal como el clasificador facial cascada de Haar, que puede usarse para identificar toda la zona del rostro en una imagen.
- Un clasificador de rasgos faciales, p. ej. un detector de puntos de referencia faciales, que se usa para identificar rasgos faciales, tales como el rabillo del ojo, la punta de la nariz, la comisura de la boca, etc. Los ejemplos de clasificadores de rasgos faciales adecuados incluyen los modelos de FaceNet o VGGFace. Este tipo de clasificador/modelo también se conoce como «Detección de puntos

de referencia faciales», siendo un ejemplo FPENet. El artículo "Recombinator networks: Learning coarse-to-fine feature aggregation" de Honari, S., Yosinski, J., Vincent, P., & Pal, C., en Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (págs. 5743-5752) 2016 proporciona un modelo de detección de puntos de referencia faciales de ejemplo.

En algunas formas de realización, la(s) imagen(es) 112/114 puede(n) procesarse para extraer información biométrica para el usuario. La información biométrica puede referirse a los rasgos faciales o a la forma del rostro. La información biométrica puede usarse como un identificador único para el usuario en los sistemas biométricos.

En algunas formas de realización, los datos textuales generados deberían asociarse a un usuario 110 en particular del dispositivo 102. Por ejemplo, múltiples usuarios 110 pueden usar el dispositivo 102 (p. ej. diferentes miembros de una familia), y es útil garantizar que los datos textuales generados están asociados a los datos textuales previamente generados para el mismo usuario 110. Un enfoque consiste en crear una base de datos o galería de rostros que contenga plantillas faciales para cada uno de los diferentes usuarios 110 que pueden usar el dispositivo 102. Esta base de datos puede servir como referencia para la circuitería de procesamiento 201, para comparar la(s) imagen(es) recién adquirida(s) y encontrar un rostro correspondiente. De este modo, cuando se comparte el contenido textual generado (p. ej. se envía a un servidor en el que se usa como datos de entrenamiento para un modelo de ML), los datos textuales solo se asocian a los datos textuales previos para ese usuario 110 y no se mezclan con los datos textuales para los otros usuarios.

En algunas formas de realización, resulta provechoso si los datos textuales comprenden o están acompañados por la información del perfil de población para el usuario, puesto que puede proporcionar un contexto útil para los datos textuales. La información del perfil de población es la

información que no permitiría identificar al usuario, pero proporciona información general sobre la demografía o antecedentes del usuario. Por ejemplo, la información del perfil de población puede comprender cualquier información sobre la edad, el sexo, el origen étnico, etc., del usuario. La información del perfil de población para un usuario puede proporcionar una indicación de los aspectos de su apariencia visual. Los factores tales como la edad, el sexo, la altura y el origen étnico pueden desempeñar una función significativa respecto de cómo se percibe visualmente a una persona. La provisión de información sobre estas características básicas en o con los datos textuales posibilita explicar las características de la población cuando se entrena el modelo de ML, al sistema basado en reglas o al sistema experto usando los datos textuales y/o para posibilitar que el modelo de ML, el sistema basado en reglas o el sistema experto proporcione una emisión que sea más relevante para el perfil de población de un paciente.

Como se observó anteriormente, para generar datos textuales, se usa el LLM. El LLM puede ser un LLM multimodal, lo cual significa que puede procesar imágenes, y uno o más de otros tipos de datos de entrada (p. ej. una señal de audio) las procesan juntas para generar los datos textuales. El LLM puede evaluar las diferentes modalidades de entrada por separado (p. ej. lo cual genera datos textuales que comprenden salidas de datos textuales 118, 120 y 122 en la Fig. 1). Alternativamente, el LLM puede evaluar las diferentes modalidades de entrada en conjunto y generar datos textuales a partir de correlaciones e interacciones entre las modalidades de entrada.

Para el LLM o LLM multimodal, puede ser necesario convertir o codificar los datos de entrada, p. ej. imágenes, señales de audio, mediciones de sensores, señales de mediciones de los sensores, etc., en un formato que el LLM pueda entender. Para las imágenes, se pueden usar técnicas como las redes neuronales convolucionales (CNN) para extraer rasgos visuales de las imágenes. De manera similar, para las

señales de audio, se pueden emplear procedimientos como una CNN adaptada para los datos de sonido, una red recurrente, tal como una memoria a corto plazo de larga duración (LSTM), una representación de espectrograma o inserciones de audio, para representar los datos de audio. Cada codificador/modelo puede extraer rasgos / vectores de rasgos de las modalidades respectivas (imágenes, audio, etcétera). En algunos casos, los rasgos visuales y de audio extraídos (y los rasgos de otras modalidades, si están presentes) se pueden combinar (p. ej. concatenar o unir) o fusionar juntos para crear una representación/vector unificado que represente todas las modalidades. Este vector de rasgos unidos puede procesarse además usando capas completamente conectadas. A continuación, la representación fusionada puede ingresarse en el LLM. El LLM procesa la(s) representación(es) de entrada y determina las descripciones y frases textuales que corresponden a los rasgos visuales y de audio en la representación fusionada.

El LLM puede basarse en un LLM disponible en el mercado, con algunas modificaciones (p. ej. los llamados pasamanos o barreras de seguridad) y/o entrenamiento extra para posibilitar que el LLM genere datos textuales apropiados, p. ej. que se refieren a los aspectos de salud de los usuarios. Para el entrenamiento, pueden usarse funciones de pérdida personalizadas, particularmente si el conjunto de datos de entrenamiento no es equilibrado. El LLM puede entrenarse para mapear las entradas (p. ej. imágenes, sonidos) con las salidas (p. ej. descripciones textuales) usando un conjunto de datos etiquetado. En el caso de un aprendizaje supervisado, la validación en terreno debería ser idealmente supervisada por profesionales sanitarios con experiencia relevante. Debe entenderse que los datos de entrenamiento usados para entrenar el LLM, para generar los datos textuales apropiados, no deben generarse de la manera descrita en la presente divulgación.

Se entenderá que la calidad o rendimiento del LLM en la descripción de detalles faciales o sombras de color / tono de piel, etcétera, depende

en gran medida de sus datos de entrenamiento y de las técnicas algorítmicas empleadas.

La circuitería de procesamiento 201 / memoria 203 puede almacenar los pesos y/u otros parámetros del LLM. La inferencia (es decir, la evaluación de las imágenes y/u otras modalidades de entrada) mediante el LLM se realiza en el dispositivo 102. Los datos textuales emitidos por el LLM se almacenan en el dispositivo 102 y/o se envían a un servidor remoto. La(s) imagen(es) original(es) y/u otras mediciones de sensores que se usan para generar los datos textuales se suprimen.

El diagrama de bloques de la Fig. 3 muestra una implementación ejemplificadora de un generador de datos textuales 108 que puede procesar entradas multimodales para generar datos textuales. El generador de datos textuales 108 proporciona un modelo multimodal unificado que se vale de un modelo de lenguaje 140 (p. ej. un LLM) como catalizador para interpretar y correlacionar diferentes modalidades de sensor.

La Fig. 3 muestra entradas 104 multimodales en el generador de datos textuales 108 que incluyen imágenes en bruto 112, grabaciones de sonido 116 y, opcionalmente, entradas de sensores 150, 152 extras, como se describió anteriormente.

El generador de datos textuales 108 puede comprender un conjunto de módulos de codificación que se usan para codificar los datos de entrada 104 respectivos. Es decir, un módulo de codificación de imágenes 154 codifica las imágenes de entrada 112, un módulo de codificación de sonidos 156 codifica el sonido de entrada 116 y los bloques de codificación de sensores 158, 160 codifican, respectivamente, las entradas de sensores 150, 152. Por lo general, los módulos de codificación pueden determinar un vector de espacio de inserción o latente a partir de los datos de entrada.

El generador de datos textuales 108 también puede comprender un conjunto de módulos de percepción que se usan para percibir los datos de

entrada respectivos. Es decir, un módulo de percepción de imágenes 162 puede percibir las imágenes 112 codificadas, un módulo de percepción de sonidos 164 puede percibir el sonido 116 codificado y los bloques de percepción de sensores 166, 168 pueden percibir, respectivamente, las entradas de sensores 150, 152 codificadas. Por lo general, los módulos de percepción pueden proyectar las inserciones de la entrada en un espacio semántico del modelo de lenguaje.

La emisión de los módulos de percepción 162, 164, 166, 168 se proporciona al modelo de lenguaje 140 que genera una respuesta de lenguaje 106 adecuada.

Como se muestra en la Fig. 3, se codifican las diferentes modalidades de entrada, respectivamente, en las que los codificadores 154, 156, 158, 160 suelen entrenarse previamente. Los modelos actuales del estado de la técnica para los codificadores (y video) de visión incluyen, p. ej., EVA-ViT-G (como se describe en “EVA: Exploring the Limits of Masked Visual Representation Learning at Scale”, de Fang et al., CVPR 2023) y, para audio, BEAT (como se describe en “BEATs: Audio Pre-Training with Acoustic Tokenizers”, de Chen et al. ICML, 2023). Otras modalidades requerirán diferentes módulos codificadores. Independientemente del codificador (y la modalidad), el codificador transforma la entrada en bruto en una inserción de longitud variable. Típicamente, esto puede realizarse usando una red neuronal profunda. Muchas de las redes actuales del estado de la técnica para diversas tareas de detección y reconocimiento se construyen sobre un trabajo previo. Un ejemplo es ResNet (como se analizó en “Deep residual learning for image recognition” de He et al., CVPR 2016) que se ha usado como *eje troncal* (que significa la parte de la red que extrae el rasgo) para las redes de localización y segmentación semántica, p. ej. Mask R-CNN (como se analizó en “Mask R-CNN" de He et al., CVPR 2017). Cabe observar que el eje troncal no es la emisión final de la red, sino más bien una emisión intermedia de una capa que codifica información relevante.

A continuación, se pasa a la *cabeza* de la red para la tarea específica, p. ej. reconocimiento de la caja delimitadora (clasificación y regresión) y predicción de máscara, como en el ejemplo de Mask R-CNN. El artículo "Proper Reuse of Image Classification Features Improves Object Detection" de Vasconcelos et al., CVPR 2022 muestra que congelar el eje troncal y reentrenar solamente la cabeza puede mejorar muchos modelos de detección diferentes.

Los módulos perceptores 162, 164, 166, 168 abarcan los codificadores 154, 156, 158, 160 y el modelo de lenguaje 140. Dichos módulos perceptores se han usado con éxito en, por ejemplo, Flamingo (como se describe en "Flamingo: a Visual Language Model for Few-Shot Learning" de Alayrac et al., NeurIPS 2022) y BLIP-2 (como se describe en "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models" de Li et al., 2023), en los que se denominan Perceiver Resampler (remuestreador perceptor) y Q-Former (transformador de consultas), respectivamente. Estos son específicos del modo; no obstante, el propósito del módulo perceptor 162, 164, 166, 168 es proyectar las inserciones de los diferentes codificadores de sensores 154, 156, 158, 160 en el espacio semántico del modelo de lenguaje 140, es decir, convertir los datos no basados en texto en una representación del lenguaje. La emisión es típicamente bastante menor que el tamaño de los rasgos congelados del codificador y, por ende, crea una arquitectura de cuello de botella para forzar a que las consultas extraigan la información sensorial que sea más relevante para el texto.

Con el fin de alinear las dimensiones de estas inserciones cuasi lingüísticas y la dimensión de la inserción del modelo de lenguaje 140, se puede usar una capa completamente conectada para proyectar las inserciones de consultas emitidas en la misma dimensión que la inserción de texto del modelo de lenguaje 140. La proyección puede ser lineal, p. ej. CLIP (como se describe en "Learning Transferable Visual Models From Natural Language Supervision" de Radford et al., ICML 2021), o no lineal,

p. ej. AMDIM (como se describe en “Learning representations by maximizing mutual information across views” de Bachman et al. en Advances in Neural Information Processing Systems, págs. 15535-15545, 2019) y SimCLR (como se describe en “A simple framework for contrastive learning of visual representations" de Chen et al. ICML 2020). A continuación, las inserciones de consultas proyectadas se pueden añadir al inicio de las inserciones del texto de entrada, de manera similar a BLIP-2. Se analizan conceptos similares en “X-LLM: Bootstrapping Advanced Large Language Models by Treating Multi-Modalities as Foreign Languages” de Chen et al., 2023 (https://arxiv.org/abs/2305.04160) y “ChatBridge: Bridging Modalities with Large Language Model as a Language Catalyst” de Zhao et al., 2023 (https://arxiv.org/pdf/2305.16103.pdf).

A continuación, se proporcionan detalles adicionales sobre el entrenamiento del (gran) modelo de lenguaje 140 para fines médicos. Existen muchas variaciones respecto de cómo entrenar la red. Debido a la alta complejidad del modelo, es habitual congelar las partes del modelo y solo ajustar ciertas partes de la red, por ejemplo como se describe en los artículos mencionados anteriormente de Alayrac et al. (2022), Li et al. (2023), Chen et al. (2023) y Zhao et al. (2023), y también en los artículos “Grounding Language Models to Images for Multimodal Inputs and Outputs” de Koh et al., ICML 2023 y “LLM-CXR: Instruction-Finetuned LLM for CXR Image Understanding and Generation” de Lee et al., 2023 (https://arxiv.org/abs/2305.11490). Los perceptores 162, 164, 166, 168 y sus testigos de consulta entrenables pueden entrenarse al tiempo que los codificadores 154, 156, 158, 160 y el modelo de lenguaje 140 se congelan durante todo el procedimiento de entrenamiento, por ejemplo debido a recursos computacionales limitados.

Con el fin de añadir testigos de imagen a un modelo de lenguaje 140 que está preentrenado solamente con texto, los testigos extrasensoriales (imagen, sonido, etcétera) deberían ponerse en el mismo espacio de

inserción que los testigos de texto. Esto se ha tratado de manera diferente en distintos marcos. Por ejemplo, DALL-E (como se describe en "Zero-Shot Text-to-Image Generation" de Ramesh et al., ICML 2021) trata las imágenes como secuencias de testigos discretos usando un modelo de texto a imagen basado en un transformador. Flamingo, mencionado anteriormente, propone un modelo de lenguaje visual para la generación de texto. El artículo de Koh et al. propone una forma de fundamentar los modelos de lenguaje preentrenados solamente con texto en el dominio visual, lo cual les posibilita procesar arbitrariamente los datos entrelazados de imagen y texto, y generar texto entrelazado con las imágenes recuperadas.

El artículo antes mencionado de Lee et al. muestra que la tokenización de las radiografías de tórax (CXR) que preservan la información clínica mejora el rendimiento tanto en la tarea de informe a CXR como en la tarea de CXR a informe, al utilizar el trabajo de Koh et al. y al aplicar las técnicas de ajuste. Estos son procedimientos estándares de ajuste para los LLM y no se realizan cambios además de la expansión de la tabla de inserción de testigo, y, como tal, se podría usar la plantilla usada por Alpaca (como se analizó en "Stanford Alpaca: An Instruction-following LLaMA model" de Taori et al., 2023, https://github.com/tatsu-lab/stanford_alpaca). En este caso, la generación de datos textuales, como se propone en la presente divulgación, puede consistir en la entrada de datos sensoriales tokenizados en el modelo de lenguaje 140, y los datos textuales emitidos consisten en un informe correspondiente de especialistas médicos basado en dicha entrada de datos.

El espejo/dispositivo 102 puede tener una funcionalidad extra a la descrita anteriormente. Por ejemplo, la funcionalidad de cámara y visualización del espejo puede usarse para conectar a los pacientes con los profesionales sanitarios, a los fines de realizar consultas a distancia. El LLM puede proporcionar nuevos datos textuales para el usuario 110 durante esta consulta, y/o el (los) sensor(es) 211 en o asociado(s) al

espejo/dispositivo 102 puede(n) usarse para proporcionar datos de salud nuevos o en vivo para el usuario 110.

Si bien la descripción anterior indica que los datos textuales se envían a un servidor remoto, por ejemplo para usar como datos de entrenamiento para un modelo de ML, un sistema basado en reglas o un sistema experto, se entenderá que puede ser que los datos textuales no se envíen tan pronto como se generan. En cambio, los datos textuales pueden generarse en una serie de instancias diferentes (p. ej. en el transcurso de varios días o semanas) y almacenarse en el dispositivo 102. Los datos textuales almacenados pueden enviarse al servidor remoto en el momento adecuado, p. ej. cuando se ha hecho un diagnóstico para el usuario 110, si se detecta que la salud del usuario 110 se ha deteriorado, una vez que se ha almacenado una cantidad adecuada de datos textuales, periódicamente, etcétera.

Una vez que los datos textuales para un usuario 110 se han enviado al servidor remoto, es posible que se generen otros datos textuales relacionados con el mismo usuario 110 a partir de otra(s) imagen(es), etcétera, obtenidos en los días o semanas subsiguientes. Para mejorar el uso de los datos textuales en el servidor remoto, puede ser útil que los nuevos datos textuales se asocien a los datos textuales para el mismo usuario 110 que se compartieron previamente con el servidor.

En algunos casos, se puede evitar esta necesidad de asociación mediante el dispositivo 102 que almacena y envía un conjunto completo de datos textuales para el usuario 110 cada vez que los datos textuales se envían al servidor. Es decir, los nuevos datos textuales se pueden enviar al servidor, junto con los datos textuales previos que ya se han enviado al servidor. En algunos casos, el conjunto de datos textuales enviado al servidor puede cubrir un periodo en el que el usuario 110 pasó de estar sano a estar enfermo, y luego se recuperó.

Si la asociación entre los nuevos datos textuales y los datos textuales previamente enviados se requiere en el servidor, el dispositivo

102 y el servidor pueden usar una única clave privada, p. ej. un ID del algoritmo de hash seguro (SHA-ID), para transferir los datos textuales de manera segura. El uso de este ID, junto con los datos textuales anonimizados, se consideraría difícil de rastrear hasta el usuario 110. Podría usarse una solución similar si el usuario 110 desea compartir los nuevos datos textuales y/o los datos textuales acumulados del dispositivo 102 con un profesional médico.

Para un usuario 110 con signos tempranos de una enfermedad, a medida que la enfermedad del usuario evoluciona, se pueden registrar notas (p. ej. contenidos de un archivo de registro, además de los datos textuales generados) en el dispositivo 102. En algún momento, los síntomas son evidentes y el paciente puede recibir un diagnóstico de un profesional sanitario, y este diagnóstico puede añadirse al archivo de registro. El archivo de registro puede enviarse al servidor junto con, o como parte de, los datos textuales.

En algunas formas de realización, las técnicas de aprendizaje federado (FL) pueden aplicarse al entrenamiento o ajuste (actualización) del generador de datos textuales 108, o al entrenamiento o ajuste de elementos del generador de datos textuales 108 (p. ej. los codificadores o perceptores). El aprendizaje federado es un enfoque de aprendizaje automático descentralizado en el que un modelo (p. ej. un modelo de lenguaje) se entrena en múltiples dispositivos de borde (en este caso dispositivos/espejos 102) o servidores descentralizados. Al usar datos de múltiples clientes, un modelo de lenguaje puede aprender de un intervalo diverso de ejemplos, y aumenta su entendimiento y rendimiento general, lo cual produce modelos de lenguaje / LLM más potentes y versátiles. La aplicación de técnicas de FL en diversos aspectos de los modelos de lenguaje, que incluyen (pre)entrenamiento y ajuste, ofrece numerosas ventajas.

El artículo “Federated Large Language Model: A Position Paper” de Chen et al., 2023 (https://arxiv.org/pdf/2307.08925.pdf) analiza el

concepto de un LLM federado que integra los procedimientos de FL y comprende el preentrenamiento del LLM federado y el ajuste del LLM federado. En comparación con los procedimientos tradicionales de entrenamiento del LLM que dependen exclusivamente de los conjuntos de datos públicos centralizados, el preentrenamiento del LLM de aprendizaje federado incorpora una combinación de fuentes de datos públicos centralizados y datos privados descentralizados. Esta integración de diversas fuentes de datos sirve a los propósitos de aumentar la generalización del modelo, posibilita el acceso a un espectro más amplio de conocimiento médico y facilita la escalabilidad futura del modelo y la expansión en términos de tamaño y alcance.

Luego del preentrenamiento, los LLM con frecuencia pueden ajustarse en tareas o dominios específicos. De este modo, el ajuste del modelo de lenguaje federado puede mejorar la generalización en las tareas y los dominios. El enfoque de ajuste convencional depende de las instituciones individuales que usan sus conjuntos de datos protegidos, lo cual suele limitar la colaboración y enfrentar desafíos relacionados con la disponibilidad y generalización de datos. Por el contrario, el ajuste del modelo de lenguaje federado (p. ej. LLM) promueve la colaboración entre múltiples instituciones, considera los requisitos de tareas específicas y emplea datos de diversos clientes para un entrenamiento multitarea colectivo. El aprendizaje federado y los LLM no son mutuamente exclusivos, pero se refuerzan y complementan entre sí.

Un enfoque basado en FL se puede usar para ajustar un modelo de lenguaje implementado (es decir, el modelo de lenguaje 140 en el generador de datos textuales 108). En este caso, en vez de que sea necesario enviar datos en bruto (p. ej. imágenes, audio, etcétera) a un servidor central para entrenar/actualizar el modelo de lenguaje / generador de datos textuales 108, las actualizaciones del generador de datos textuales 108 se computan localmente (es decir, en un espejo 102), y solamente estas actualizaciones se comparten y se agregan en un

servidor central para mejorar el modelo de lenguaje global / generador de datos textuales 108. Esto minimiza la necesidad de transferir grandes volúmenes de datos a un servidor central, y así se reduce el uso de ancho de banda y los costos asociados.

Si bien el aprendizaje federado proporciona varios beneficios, puede haber problemas respecto de la privacidad del usuario en otros casos de uso de FL. No obstante, en el presente caso, los datos son cadenas de texto impersonal (datos textuales) almacenadas en el espejo 102 y, por ende, la privacidad de un usuario se preserva incluso si esos datos están comprometidos. En particular, cada espejo/dispositivo 102 puede computar de manera independiente las actualizaciones del modelo de lenguaje 140 que usa localmente sus conjuntos de datos. Estas actualizaciones representan los cambios incrementales necesarios para mejorar el modelo de lenguaje global. A continuación, las actualizaciones se transmiten a un servidor central (que puede ser el mismo o un servidor diferente del servidor que recibe los datos textuales 106 y los usa para algún otro propósito). El servidor central agrega estas actualizaciones para crear un modelo de lenguaje 140 global mejorado. Este modelo de lenguaje 140 global encapsula el conocimiento colectivo y los hallazgos de todos los espejos/dispositivos 102 que están participando en el aprendizaje federado. Luego de que el servidor central agrega las actualizaciones y crea un modelo de lenguaje 140 global mejorado, este modelo de lenguaje 140 global actualizado se envía de regreso a los espejos/dispositivos 102 participantes. A continuación, cada espejo/dispositivo 102 recibe el modelo de lenguaje 140 global actualizado y además ajusta su modelo 140 local sobre la base de las mejoras en el modelo global.

Una investigación reciente sugiere abordar uno de los principales problemas de seguridad sobre la privacidad de datos al ajustar los modelos de lenguaje preentrenados en un cliente con una comunicación mínima al servidor central (p. ej. solo compartir los pesos del modelo).

Esta investigación se encuentra en “FedPETuning: When Federated Learning Meets the Parameter-Efficient Tuning Methods of Pre-trained Language Models” de Zhang et al., Findings of the Association for Computational Linguistics: ACL, 2023 (https://arxiv.org/pdf/2212.10025.pdf).

El dispositivo/espejo 102 dependerá principalmente de los datos privados; no obstante, esto no excluye los conjuntos de datos públicos. El artículo “Can Public Large Language Models Help Private Cross-device Federated Learning?” de Wang et al. 2023 (https://arxiv.org/pdf/2305.12132.pdf) propone una forma de mejorar el entrenamiento del LLM en el dispositivo usando tokenizadores públicos con destilación antes del aprendizaje federado privado entre dispositivos. Este enfoque proporciona mejoras significativas y demostró una gran precisión de aprendizaje privado. Este enfoque puede usarse para mejorar el aprendizaje federado privado con los LLM públicos, lo cual podría además mejorar el rendimiento de las técnicas descritas en la presente memoria. El artículo “FedJudge: Federated legal large language model” de Yue et al., 2023 (https://arxiv.org/pdf/2309.08173.pdf) demuestra que es factible considerar un enfoque de LLM federado para una tarea específica (aquí usan inteligencia jurídica como un caso de uso).

La Fig. 4 es un diagrama de flujo que ilustra un procedimiento para operar un dispositivo/espejo 102, de acuerdo con diversas formas de realización. El dispositivo/espejo 102 puede realizar el procedimiento en respuesta a ejecutar un código legible por computadora, adecuadamente formulado. El código legible por computadora puede incorporarse o almacenarse en un medio legible por computadora, tal como un chip de memoria, un disco óptico u otro medio de almacenamiento. El medio legible por computadora puede formar parte de un producto de programa informático.

En el paso 401, el dispositivo 102 obtiene una o más imágenes de

un usuario 110 del dispositivo 102. El paso 401 puede comprender el dispositivo 102 que controla un sensor de imágenes 207 para generar la una o más imágenes. Alternativamente, el paso 401 puede comprender el dispositivo 102 / la circuitería de procesamiento 201 que recibe la una o más imágenes de un sensor de imágenes 207 separado o que recupera las imágenes previamente generadas y almacenadas de una ubicación de almacenamiento (p. ej. la memoria 203).

El paso 401 también puede comprender controlar una o más fuentes de luz 209 para iluminar al usuario 110 y/o el entorno alrededor del usuario 110 cuando la(s) imagen(es) se genera(n) u obtiene(n) mediante el sensor de imágenes 207. En particular, la(s) fuente(s) de luz 209 puede(n) controlarse de modo tal que las condiciones de alumbrado en el usuario 110 y/o en el entorno sean consistentes con las condiciones de alumbrado cuando se generaron una o más imágenes previas. De esta forma, se puede reducir la variabilidad en la apariencia del usuario en las imágenes debido a las diferentes condiciones de alumbrado. Por ejemplo, la luz ambiente en el entorno alrededor del dispositivo 102 puede ser diferente a la mañana y a la noche, lo cual produce diferencias aparentes en las imágenes obtenidas en esos momentos, y un ajuste adecuado de la(s) fuente(s) de luz 209 puede proporcionar un alumbrado consistente cuando se obtienen las imágenes.

En el paso 403, el dispositivo 102 usa un modelo de lenguaje, por ejemplo un LLM, para generar datos textuales que describen al usuario 110 mostrado en la una o más imágenes. Los datos textuales pueden describir el rostro del usuario 110 y/o referirse a los aspectos de salud del usuario 110. El modelo de lenguaje puede ser un modelo de lenguaje multimodal que puede generar los datos textuales al menos a partir de las imágenes. Es decir, el modelo de lenguaje no está restringido a recibir entradas basadas en texto.

Los datos textuales se generan de modo tal que no es posible identificar al usuario 110 específico a partir de los mismos datos textuales.

De este modo, la seguridad de los datos del usuario está protegida en el caso de que un tercero tenga acceso a los datos textuales (p. ej. de conformidad con el paso 305 a continuación).

En el paso 405, el dispositivo 102 envía los datos textuales generados a un servidor. Como los datos textuales se generan de modo tal que no es posible identificar al usuario 110 específico a partir de los mismos datos textuales, el envío de los datos textuales al servidor no viola la seguridad de los datos del usuario. Para finalizar, cabe observar que la(s) imagen(es) usada(s) por el modelo de lenguaje para generar los datos textuales no se envía(n) al servidor. La(s) imagen(es) obtenida(s) puede(n) suprimirse mediante el dispositivo 102, p. ej. suprimirse de la memoria 203, luego de que se usen para generar los datos textuales. El dispositivo 102 puede proporcionar un mensaje de confirmación al usuario 110 para indicar que las imágenes se han suprimido.

En algunas formas de realización, antes de usar el modelo de lenguaje para generar los datos textuales, o como parte del funcionamiento del modelo de lenguaje en la generación de los datos textuales, la(s) imagen(es) y otros tipos de datos de entrada pueden procesarse usando uno o más de los algoritmos o modelos siguientes. Por ejemplo, como se describe con referencia a la Fig. 3 anteriormente, puede usarse un módulo de codificación respectivo y/o un módulo de percepción respectivo para codificar los datos de entrada y/o percibir el contenido de los datos de entrada, generando el modelo de lenguaje 140 los datos textuales sobre la base de la emisión codificada/percibida.

En otro ejemplo, puede usarse un algoritmo de detección de objetos para detectar una parte del cuerpo (p. ej. el rostro) en la(s) imagen(es). Como otro ejemplo, puede usarse un clasificador facial para identificar un rostro del usuario 110 en la(s) imagen(es). Puede usarse un clasificador de rasgos faciales para identificar los rasgos faciales del rostro del usuario 110 (p. ej. ojos, nariz, boca, cejas, etc.). Puede usarse un algoritmo de segmentación semántica para segmentar la(s) imagen(es) en información

a nivel de píxel.

La información generada mediante estos algoritmos y/o modelos se ingresa en el modelo de lenguaje, y el modelo de lenguaje genera los datos textuales sobre la base de la(s) imagen(es) y la información de entrada.

En algunas formas de realización, el modelo de lenguaje genera los datos textuales sobre la base de la(s) imagen(es) y las mediciones/informaciones de uno o más sensores extras. Cualquiera de estos sensores 211 puede formar parte del dispositivo 102 o puede estar separado del dispositivo 102.

Los sensores 211 pueden incluir cualquiera de un micrófono, un termómetro, un sensor de frecuencia cardíaca, un sensor de frecuencia respiratoria, un sensor de presión arterial y balanzas. De este modo, las mediciones / señales de mediciones pueden comprender cualquiera de grabación de audio del usuario 110, temperatura del usuario 110, información de la frecuencia cardíaca para el usuario 110, frecuencia respiratoria del usuario 110, presión arterial del usuario 110 y peso del usuario 110.

De este modo, el dispositivo 102 obtiene mediciones relacionadas con el usuario 110 a partir de uno o más sensores 211 asociados con el dispositivo 102. Este paso puede comprender el dispositivo 102 que controla el (los) sensor(es) 211 para realizar mediciones y generar las mediciones de sensores / señales de mediciones de los sensores respectivas. Alternativamente, este paso puede comprender el dispositivo 102 / la circuitería de procesamiento 201 que recibe las mediciones / señales de mediciones de los sensores 211 separados o que recupera las mediciones de sensores previamente generadas y almacenadas de una ubicación de almacenamiento (p. ej. la memoria 203). Las mediciones de sensores / señales de mediciones son usadas por el modelo de lenguaje 108 en la generación de los datos textuales, en el paso 403. En algunas formas de realización, cualquiera o la totalidad de las mediciones de

sensores / señales de mediciones se pueden procesar previamente, de modo que tengan un formato adecuado para ser ingresadas en el modelo de lenguaje. De forma alternativa, un modelo de lenguaje multimodal puede recibir directamente las mediciones de sensores / señales de mediciones.

El dispositivo 102 puede almacenar un registro de datos para el usuario 110 (p. ej. en la memoria 203). Este registro de datos puede incluir datos textuales previamente generados para el usuario (es decir, generados a partir de imágenes previas) y/u otra información sobre el usuario, tal como información biométrica que el dispositivo 102 puede usar para identificar al usuario 110. El dispositivo 102 puede procesar la(s) imagen(es) para extraer información biométrica para el usuario 110 (p. ej. biometría relacionada con la forma/estructura del rostro) y puede usar la información biométrica extraída para identificar un registro de datos para el usuario 110 que está almacenado en el dispositivo 102. Los datos textuales generados en el paso 403 pueden añadirse al registro de datos almacenado.

En algunas formas de realización, la información del perfil de población para el usuario 110 se asocia a los datos textuales generados y se envía al servidor. Esta información del perfil de población puede proporcionar un contexto útil para los datos textuales en el servidor, p. ej. el tipo de persona para el que se generaron los datos textuales. La información del perfil de población puede haber sido ingresada manualmente en el dispositivo 102 por el usuario 110, o el dispositivo 102 puede derivar la información del perfil de población de la(s) imagen(es) del usuario 110. La información del perfil de población puede comprender cualquiera de edad, sexo, altura y origen étnico.

Si bien los dispositivos informáticos descritos en la presente memoria pueden incluir la combinación ilustrada de los componentes de hardware, otras formas de realización pueden comprender dispositivos informáticos con diferentes combinaciones de componentes. Se debe

entender que estos dispositivos informáticos pueden comprender cualquier combinación apta de hardware y/o software necesaria para realizar las tareas, las características, las funciones y los procedimientos divulgados en la presente memoria. Las operaciones de determinación, cálculo, obtención u operaciones similares descritas en la presente memoria pueden realizarse mediante la circuitería de procesamiento, que puede procesar la información, por ejemplo, al convertir la información obtenida en otra información, al comparar la información obtenida o información convertida con la información almacenada en el dispositivo, y/o al realizar una o más operaciones sobre la base de la información obtenida o información convertida, y como resultado de dicho procesamiento, al tomar una determinación. Asimismo, si bien los componentes se representan como recuadros individuales ubicados dentro de un recuadro más grande, o anidados dentro de múltiples recuadros, en la práctica, los dispositivos informáticos pueden comprender múltiples componentes físicos diferentes que constituyen un solo componente ilustrado, y la funcionalidad puede dividirse entre los componentes separados. Por ejemplo, una interfaz de comunicación puede estar configurada para incluir cualquiera de los componentes descritos en la presente memoria, y/o la funcionalidad de los componentes puede dividirse entre la circuitería de procesamiento y la interfaz de comunicación. En otro ejemplo, las funciones no computacionalmente intensivas de cualquiera de dichos componentes pueden implementarse en el software o firmware, y las funciones computacionalmente intensivas pueden implementarse en el hardware.

En ciertas formas de realización, una parte o la totalidad de la funcionalidad descrita en la presente memoria puede proporcionarse mediante la circuitería de procesamiento ejecutando las instrucciones almacenadas en la memoria, que, en ciertas formas de realización, puede ser un producto de programa informático como un medio de almacenamiento legible por computadora, no transitorio. En formas de

realización alternativas, una parte o la totalidad de la funcionalidad puede proporcionarse mediante la circuitería de procesamiento sin ejecutar las instrucciones almacenadas en un medio de almacenamiento legible por dispositivo, separado o discreto, tal como de una manera cableada. En cualquiera de aquellas formas de realización en particular, ya sea ejecutando instrucciones almacenadas en un medio de almacenamiento legible por computadora, no transitorio, o no, la circuitería de procesamiento puede estar configurada para realizar la funcionalidad descrita.

Lo anterior ilustra meramente los principios de la divulgación. Diversas modificaciones y alteraciones en las formas de realización descritas se pondrán de manifiesto para los expertos en la técnica, en vista de las enseñanzas de la presente memoria. De este modo, se entenderá que los expertos en la técnica puedan idear numerosos sistemas, disposiciones y procedimientos que, aunque no se describan o muestren explícitamente en la presente memoria, incorporen los principios de la divulgación y se incluyan así dentro del alcance de la divulgación. Pueden usarse diversas formas de realización ejemplificadoras juntas, unas con otras, así como también de manera intercambiable entre sí, como deben entenderlo los expertos en la técnica.