Proyecto de Aprendizaje de Máquinas

Daniela Rodríguez Cepero
Elena Rodríguez Horta
Hasel Martí Blanco
Karel Camilo Manresa León
Karlos Alejandro Alfonso Rodríguez
Lázaro Alejandro Castro

Universidad de La Habana,
San Lázaro y L. Plaza de la Revolución, La Habana, Cuba
https://github.com/hbm99/bias-project-ML.git
http://www.uh.cu

Índice general

1. Introducción

Hace unos años, al hablar de modelos de inteligencia artificial (IA), prácticamente se asumía que nos regíamos a unos contextos muy concretos y exclusivos: grandes empresas tecnológicas, centros de investigación en la materia y entornos similares. Esta situación ha cambiado drásticamente al día de hoy.

En la actualidad esos modelos de IA están en nuestros teléfonos móviles, la smart TV de nuestro salón, hablando exclusivamente de dispositivos físicos. También están integrados en nuestro software de correo, en las redes sociales y muchos de los servicios telemáticos que usamos a diario. Fuera del ámbito personal, si no nos restringimos y abrimos el abanico a entornos empresariales e institucionales, los procesos en los que se recurre a modelos generados por algoritmos de aprendizaje automático son incontables.

El creciente impacto de los modelos de aprendizaje automático en la sociedad, así como la variedad de campos en los que son aplicados, ha llamado la atención de la comunidad hacia las consecuencias que estos algoritmos pueden traer para las personas y la sociedad de forma general. En particular, preocupan los efectos que puedan tener en áreas como las de la educación y formación, el empleo, los servicios, la legalidad y los sistemas judiciales. Se ha vuelto necesario que los sistemas de inteligencia artificial sean desarrollados con responsabilidad, teniendo en cuenta aspectos como la imparcialidad, seguridad y privacidad, comprensibilidad honestidad y ecuanimidad resultan esenciales.[16].

Algunos gobiernos usan los modelos de aprendizaje automático en la predicción de brotes virales y focos de crímenes. Los sesgos existentes en estos modelos tienen repercusiones trascendentales, lo cual es preocupante por el hecho de que pueden reforzarse o derivar en resultados inesperados[17].

El proceso de contratación se ha convertido en un proceso automatizado en muchos casos, desde la selección de currículums hasta la elección de los candidatos. Como consecuencia de la participación cada vez mayor de sistemas automatizados y de aprendizaje automático en esta tarea de gran repercusión para las personas, es necesario controlar cuándo una decisión es injusta [18].

De manera general, en la comunidad existen posiciones diferentes acerca del uso de los algoritmos de aprendizaje automático en la toma de decisiones. Algunos plantean que el uso de algoritmos en la toma de decisiones, en particular en la evaluación de riesgo en escenarios judiciales, debe ser detenido, pues estos modelos son fuentes de incertidumbres y alteraciones [19].

Considerando las definiciones de equidad presentes en la literatura, existen numerosos algoritmos de clasificación que tienen un buen desempeño. Sin embargo, debido a que estos se enfocan en el tiempo presente y dado que incorporan el sesgo humano, existe el riesgo y la propensión a repetir y aumentar dichos

sesgos [20].

Teniendo en cuenta los riesgos que supone la utilización de los modelos de aprendizaje automático en la toma de decisiones, es necesario enfrentar los sesgos y prejuicios que contienen para que no se mantengan y extiendan. Es importante implementar conceptos éticos fundamentales en la inteligencia artificial [21] y construir sistemas que se comprometan con los valores sociales [22]. Es decir, determinar cuán daino puede ser un sesgo y, por tanto, cuáles no pueden ser permitidos en los algoritmos. Es una tarea difícil pero sumamente necesaria por lo que requiere la atención de todos.

2. Sesgo

Durante los últimos años, con el creciente uso de este tipo de tecnologías, se han ido descubriendo múltiples sesgos de Machine Learning que nos deberían dar qué pensar. Si bien el aprendizaje automático ofrece una fuente de información muy valiosa y nos dota de herramientas de gran utilidad para comprender el mundo que nos rodea, los sesgos descubiertos podrían resultar contraproducentes para el interés general o beneficiar a algunos en detrimento de otros.

El sesgo en aprendizaje automático, también conocido como sesgo de modelo, es la diferencia entre el valor medio predicho por el modelo y el valor medio real, aparece cuando un modelo produce resultados erróneos de forma sistemática. La aparición de estos es debida a que los modelos son desarrollados por personas. Las cuales tiene tienen preferencias que transfieren a los modelos. Tanto sean conscientes como inconscientes. Muchas veces estas pueden pasar desapercibidos hasta que los modelos se ponen en producción.

En los procesos de toma de decisiones el término sesgo tiene generalmente connotaciones negativas. No es deseable que un proceso automático lo tenga de ningún tipo. La palabra sesgo procede de sesgar, un verbo que hace referencia a torcer o atravesar algo hacia uno de sus lados. Por lo que una decisión sesgada, que se tuerce en algún sentido, no es deseable. Los modelos de aprendizaje automático (machine learnig) no están exentos de este problema, ya que son desarrollados por personas. Así que es importante conocer qué es el sesgo en aprendizaje automático y cómo se puede minimizar su aparición.

Debido a todo esto se ha dirigido el estudio de muchos investigadores hacia la detección de sesgos en sistemas existentes por ejemplo existe un estudio que analiza y evalúa los sesgos existentes en los algoritmos y conjuntos de datos de análisis facial automatizados (tales como IARPA Janus Benchmark-A(IJB-A) y Adience [23]). En él se plantea la relevancia que puede tener en el análisis de los datos los subgrupos que se consideran, que pueden constituir la diferencia entre encontrar sesgos o no.

Varios sistemas de NLP usan los vectores (embeddings) de palabras, que son una forma de representarlas y capturar sus asociaciones y semántica. Estos vectores son construidos a partir de los contextos en que aparecen las palabras en los textos. Varios estudios se han enfocado en el reconocimiento de sesgos en estos vectores. Uno de ellos reconoce sesgos presentes en vectores generados por modelos como GloVe [29], y Word2Vec [30], entrenado con los artículos de Google News [31].

Los prejuicios y sesgos existentes en la sociedad han sido detectados también en cuerpos documentales como Wikipedia. Esto se refleja en las diferencias entre las formas en las que las mujeres y los hombres son descritos [26]. Anuncios generados por modelos de aprendizaje automático [27], as como los motores de

búsqueda en la Web [28] han sido detectados como fuentes de sesgos, por ejemplo, de género. Otros ejemplos de sistemas en los que se han detectado sesgos son los chatbots, los sistemas de reconocimiento facial y de voz , y algunos sistemas empleados en concursos de belleza.

s. COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) es un software usado en las cortes de Estados Unidos para determinar la probabilidad de que una persona reincurra en un crimen. Se encontró que el software estaba sesgado hacia los afroamericanos [24]. Esto quiere decir que, dados dos individuos con el mismo perfil, bastaba con que se diferenciaran únicamente en que uno fuera afroamericano y el otro no para asociarle un mayor riesgo al primero. COMPAS es usado por los jueces para decidir si liberar a un prisionero o mantenerlo en la cárcel. Se ha mostrado que COMPAS no es más preciso que otros sistemas más simples que tienen el mismo objetivo, ni siquiera toma decisiones más imparciales que las que pueden tomar personas con ningún conocimiento judicial [25]

2.1. Tipos de Sesgos

Los investigadores han identificado tres categorías principales de sesgo en la IA:

- **Sesgos Técnicos:** Se relaciona con el sesgo que empeora el prejuicio preexistente causado por uno de los procesos de decisión internos del algoritmo. Es decir son los sesgos provenientes del algoritmo.

  El sesgo emergente ocurre como resultado del uso y la interacción con usuarios reales. Este sesgo surge como resultado del cambio en la población, los valores culturales o el conocimiento social, generalmente algún tiempo después de la finalización del diseo. Este tipo de sesgo es más probable que se observe en las interfaces de usuario, ya que las interfaces tienden a reflejar las capacidades, características y hábitos de los posibles usuarios por diseo . Este tipo de sesgo se puede dividir en más subtipos, como se analiza en detalle en la Referencia.[35]

  Los sesgos populares se ponen de manifiesto por ejemplo cuando los artículos que son más populares tienden a estar más expuestos. Sin embargo, las métricas de popularidad están sujetas a manipulación, por ejemplo, por reseñas falsas o bots sociales . Por ejemplo, este tipo de sesgo se puede ver en los motores de búsqueda o en los sistemas de recomendación donde los objetos populares se presentarían ms al público. Pero esta presentación puede no ser el resultado de una buena calidad; en cambio, puede deberse a otros factores sesgados.
  [36]
- **Preexistentes:** Se refiere al sesgo ya presente en los datos utilizados para entrenar el modelo de IA.

El sesgo histórico es el sesgo y los problemas sociotécnicos ya existentes en el mundo y puede filtrarse desde el proceso de generación de datos incluso con un muestreo y una selección de características perfectos [37]. Un ejemplo de este tipo de sesgo se puede encontrar en un resultado de búsqueda de imágenes de 2018 donde la búsqueda de directoras ejecutivas resultó en última instancia en menos imágenes de directoras ejecutivas debido al hecho de que solo el 5 % de las directoras ejecutivas de Fortune 500 eran mujeres, lo que provocaría que los resultados de la búsqueda estar sesgado hacia los directores ejecutivos masculinos [37]. Estos resultados de búsqueda, por supuesto, reflejaban la realidad, pero vale la pena considerar si los algoritmos de búsqueda deben o no reflejar esta realidad.

El sesgo social ocurre cuando las acciones de otros afectan nuestro juicio [39]. Un ejemplo de este tipo de sesgo puede ser un caso en el que queremos calificar o revisar un ítem con una puntuación baja, pero al estar influenciados por otras calificaciones altas, cambiamos nuestra puntuación pensando que tal vez estamos siendo demasiado duros.

- **Sesgo Emergente:** Se produce debido a la interacción con uno o más usuarios.

  El sesgo de comportamiento surge del diferente comportamiento de los usuarios entre plataformas, contextos o diferentes conjuntos de datos [38]. Un ejemplo de este tipo de sesgo se puede observar en la Referencia [104], donde los autores muestran cḿo las diferencias en las representaciones de emoji entre plataformas pueden resultar en diferentes reacciones y comportamientos de las personas y, a veces, incluso conducen a errores en la comunicación.

  El sesgo de representación surge de cḿo tomamos muestras de una población durante el proceso de recopilación de datos [37]. Las muestras no representativas carecen de la diversidad de la población, con subgrupos faltantes y otras anomalías. La falta de diversidad geográfica en conjuntos de datos como ImageNet da como resultado un sesgo demostrable hacia las culturas occidentales.

2.2. Métricas y equidad

Las métricas para estimar la quidad de los modelos pueden agruparse según si son métricas para medir sesgos en grupos o en individuos.

La mayoría de medidas de equidad dependen de diferentes métricas, de modo que comenzaremos por definirlas. Cuando trabajamos con un clasificador binario, tanto la clase predicha por el algoritmo como la real pueden tomar dos valores: positivo y negativo. Empecemos ahora explicando las posibles relaciones entre el resultado predicho y el real:

- **Verdadero positivo (TP):** Cuando el resultado predicho y el real pertenecen a la clase positiva.
- **Verdadero negativo (TN):** Cuando el resultado predicho y el real pertenecen a la clase negativa.
- **Falso positivo (FP):** Cuando el resultado predicho es positivo pero el real pertenece a la clase negativa.
- **Falso negativo (FN):** Cuando el resultado predicho es negativo pero el real pertenece a la clase positiva.

Estas relaciones pueden ser representadas fácilmente con una matriz de confusión, una tabla que describe la precisión de un modelo de clasificación. En esta matriz, las columnas y las filas representan instancias de las clases predichas y reales, respectivamente.

Utilizando estas relaciones, podemos definir múltiples métricas que podemos usar después para medir la equidad de un algoritmo:

- **Valor predicho positivo (PPV):** la fracción de casos positivos que han sido predichos correctamente de entre todas las predicciones positivas. Con frecuencia, se denomina como precisión, y representa la probabilidad de que una predicción positiva sea correcta. Viene dada por la siguiente fórmula:
  $PPV = P(actual = +|prediction = +) = \frac{TP}{TP + FP}$
- **Tasa de descubrimiento de falsos (FDR):** la fracción de predicciones positivas que eran en realidad negativas de entre todas las predicciones positivas. Representa la probabilidad de que una predicción positiva sea errónea, y viene dada por la siguiente fórmula:
  $FDR = P(actual = -|prediction = +) = \frac{FP}{TP + FP}$
- **Valor predicho negativo (NPV):** la fracción de casos negativos que han sido predichos correctamente de entre todas las predicciones negativas. Representa la probabilidad de que una predicción negativa sea correcta, y viene dada por la siguiente fórmula:
  $NPV = P(actual = -|prediction = -) = \frac{TN}{TN + FN}$
- **Tasa de omisin de falsos (FOR):** la fracción de predicciones negativas que eran en realidad positivas de entre todas las predicciones negativas. Representa la probabilidad de que una predicción negativa sea errónea, y viene dada por la siguiente fórmula:
  $FOR = P(actual = +|prediction = -) = \frac{FN}{TN + FN}$

- **Tasa de verdaderos positivos (TPR):** la fracción de casos positivos que han sido predichos correctamente de entre todos los casos positivos. Con frecuencia, se denomina como exhaustividad, y representa la probabilidad de que los sujetos positivos sean clasificados correctamente como tales. Viene dada por la fórmula:
$TPR = P(prediction = +|actual = +) = \frac{TP}{TP + FN}$

- **Tasa de falsos negativos (FNR):** la fracción de casos positivos que han sido predichos de forma errónea como negativos de entre todos los casos positivos. Representa la probabilidad de que los sujetos positivos sean clasificados erróneamente como negativos, y viene dada por la fórmula:
$FNR = P(prediction = -|actual = +) = \frac{FN}{TP + FN}$

- **Tasa de verdaderos negativos (TNR):** la fracción de casos negativos que han sido predichos correctamente de entre todos los casos negativos. Representa la probabilidad de que los sujetos negativos sean clasificados correctamente como tales, y viene dada por la fórmula:
$TNR = P(prediction = -|actual = -) = \frac{TN}{TN + FP}$

- **Tasa de falsos positivos (FPR):** la fracción de casos negativos que han sido predichos de forma errónea como positivos de entre todos los casos negativos. Representa la probabilidad de que los sujetos negativos sean clasificados erróneamente como positivos, y viene dada por la fórmula:
$FPR = P(prediction = +|actual = -) = \frac{FP}{TN + FP}$

Entre las métricas de grupo más usadas se encuentran:

- **Statistical o Demographic Parity** [32] Un algoritmo predictor Y cumple con Demographic Parity con respecto a un atributo A con valores en el conjunto 0,1 si se cumple: $P(Y|A = 0) = P(Y|A = 1)$. Esto implica que el valor del atributo protegido no influye en el resultado del algoritmo Y.

- **Equal Opportunity** [32] [33] Dado un predictor binario Y con respecto a un atributo A y salida Y , esta métrica se cumple cuando la probabilidad de que una persona de la clase 1 sea correctamente asignada sea la misma para los grupos definidos por A = 1 y A = 0, es decir, $P(Y = 1|A = 1, Y = 1) = P(Y = 1|A = 0, Y = 1)$.

- **Equalized Odds** [33] Un predictor Y satisface esta métrica con respecto a un atributo protegido A, si A y el predictor son independientes dado Y , es decir, si $P(Y = 1|A = 1, Y = y) = P(Y = 1|A = 0, Y = y)$, y $\in 0, 1$. Implica que los grupos protegidos y los que no lo son deben tener razones iguales de falsos positivos y verdaderos positivos.

Otras métricas para analizar y detectar sesgos, dirigidas a los individuos, [34] son:

- **Fairness Through Awareness:** Individuos similares según alguna métrica definida deben obtener resultados similares.
- **Fairness Through Unawareness:** Un algoritmo es imparcial siempre que no utilice al atributo protegido como decisor.
- **Conterfactual fairness:** Una decisión es imparcial hacia un individuo si es la misma en el contexto real y en un contexto hipotético en el que el individuo pertenece a otro grupo.

En cuanto a las métricas para el análisis y detección de sesgos, existe la posibilidad de que al evaluar un modelo con varias métricas, una de ellas detecte sesgo mientras que las otras no. Excepto en casos especiales restringidos, satisfacer algunas de las métricas a la misma vez resulta imposible.

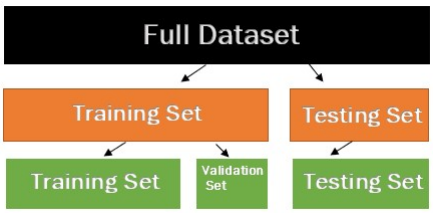
Es importante analizar el contexto y la aplicación de cada definición de equidad. No todas las definiciones se ajustan a todos los contextos y pueden tener resultados negativos si no son las adecuadas. De igual forma, analizar el tipo de sesgo es importante a la hora de tratar de lograr la equidad. La modelación y mediciń temporal es importante en la evaluación de los criterios de equidad e introduce un nuevo rango de retos en este sentido.

3. DataSet

Dataset: Se define como **una colección de datos que una computadora trata como una sola unidad** . Esto significa que un conjunto de datos contiene una gran cantidad de datos separados, pero se puede usar para entrenar un algoritmo con el objetivo de encontrar patrones predecibles dentro de todo el conjunto de datos.

Los datos son un componente esencial de cualquier modelo de IA y, básicamente, la única razón del aumento en la popularidad del aprendizaje automático que presenciamos hoy.

Generalmente un dataset se divide en varias partes, lo cual es necesario para verificar qué tan bien fue el entrenamiento del modelo. Para este propósito, un conjunto de datos de prueba generalmente se separa de los datos. A continuación, un conjunto de datos de validación, aunque no es estrictamente crucial, es bastante útil para evitar entrenar su algoritmo en el mismo tipo de datos y hacer predicciones sesgadas.



4. Algoritmo

Los algoritmos de aprendizaje automático son fragmentos de código que ayudan a los usuarios a explorar y analizar conjuntos de datos complejos y a buscar significado en ellos. Cada algoritmo es un conjunto finito de instrucciones paso a paso inequívocas que puede seguir una máquina para lograr un determinado objetivo. En un modelo de aprendizaje automático, el objetivo es establecer o detectar patrones que los usuarios puedan usar para hacer predicciones o clasificar informacin.

4.1. Técnicas Algorítmicas

Los algoritmos de aprendizaje automático se agrupan en técnicas de aprendizaje automático que se usan par el aprendizaje supervisado, no supervisado y por refuerzo.

- **Aprendizaje supervisado**
  En el aprendizaje supervisado, la máquina se enseña con ejemplos. De este modo, el operador proporciona al algoritmo de aprendizaje automático un conjunto de datos conocidos que incluye las entradas y salidas deseadas, y el algoritmo debe encontrar un método para determinar cómo llegar a esas entradas y salidas.
  Mientras el operador conoce las respuestas correctas al problema, el algoritmo identifica patrones en los datos, aprende de las observaciones y hace predicciones. El algoritmo realiza predicciones y es corregido por el operador, y este proceso sigue hasta que el algoritmo alcanza un alto nivel de precisión y rendimiento.

  Esta técnica es útil cuando sabes cómo será el resultado. Por ejemplo, imagina que proporcionas un conjunto de información que incluye la población de una serie de ciudades por año durante los últimos 100 años y quieres saber cuál será la población de una ciudad específica dentro de cuatro aos. El resultado utiliza etiquetas que ya existen en el conjunto de datos: población, ciudad y año.

  Dentro de esta técnica algunos de los algoritmos más utilizados son:

- Algoritmos de Clasificación.
- Algoritmos de Regresión.

- **Aprendizaje sin supervisión**
  Aquí, el algoritmo de aprendizaje automático estudia los datos para identificar patrones. No hay una clave de respuesta o un operador humano para proporcionar instrucción. En cambio, la máquina determina las correlaciones y las relaciones mediante el análisis de los datos disponibles.
  En un proceso de aprendizaje no supervisado, se deja que el algoritmo de aprendizaje automático interprete grandes conjuntos de datos y dirija esos datos en consecuencia. Así, el algoritmo intenta organizar esos datos de alguna manera para describir su estructura. Esto podría significar la necesidad de agrupar los datos en grupos u organizarlos de manera que se vean más organizados.
  A medida que evalúa más datos, su capacidad para tomar decisiones sobre los mismos mejora gradualmente y se vuelve más refinada. Esta técnica es útil cuando no sabes cómo será el resultado.
  Por ejemplo, imagina que proporcionas datos de clientes y quieres crear segmentos de clientes a los que les gustan productos similares. Los datos que proporcionas no están etiquetados y las etiquetas de los resultados se generan en función de las similitudes detectadas entre los puntos de datos.

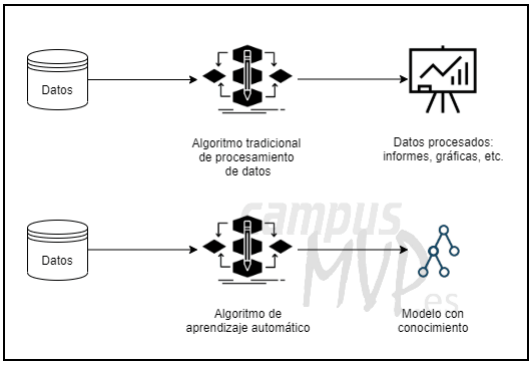
  Dentro de esta técnica algunos de los algoritmos más utilizados son:
  - Algoritmos de Clustering.
  - Algoritmos de Reducción de dimensionalidad.

- **Aprendizaje por refuerzo**
  El aprendizaje por refuerzo se centra en los procesos de aprendizajes reglamentados, en los que se proporcionan algoritmos de aprendizaje automáticos con un conjunto de acciones, parámetros y valores finales.
  Al definir las reglas, el algoritmo de aprendizaje automático intenta explorar diferentes opciones y posibilidades, monitorizando y evaluando cada resultado para determinar cuál es el óptimo.
  En consecuencia, este sistema ensea la máquina a través del proceso de ensayo y error. Aprende de experiencias pasadas y comienza a adaptar su enfoque en respuesta a la situación para lograr el mejor resultado posible. Es una buena técnica para usarla en sistemas automatizados que tienen que tomar muchas decisiones pequeas sin la intervención humana.
  Por ejemplo, imagina que estás diseando un automóvil autónomo y quieres asegurarse de que respeta la ley y mantiene la seguridad de los pasajeros. A medida que el coche adquiere experiencia y un historial de refuerzo, aprende a permanecer en su carril, a respetar el límite de velocidad y a frenar cuando hay peatones.

5. Modelo

Un modelo de aprendizaje automático es un archivo, software o aplicación de computadora que se ha entrenado para juzgar y reconocer determinados tipos de patrones. Puede entrenar un modelo con un conjunto de datos, y proporcionarle un algoritmo que puede usar para aprender de esos datso y así lograr averiguar y obtener información. Una vez entrenado el modelo, puedes usarlo para desglosar los datos que no ha visto antes y realizar predicciones sobre estos.

Todos los modelos de aprendizaje automático se clasifican en supervisados, no supervisados y aprendizaje reforzado.



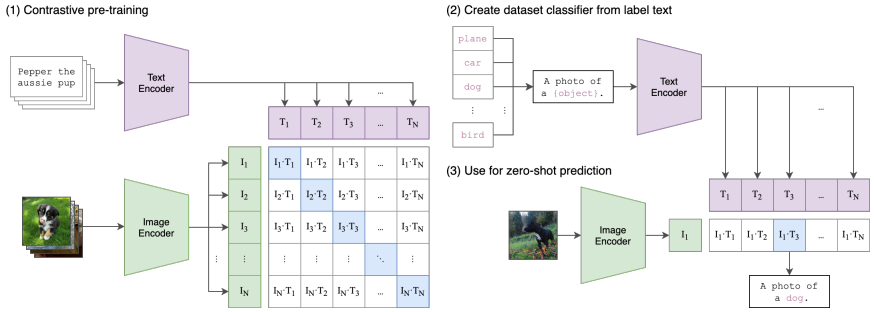
[10]

6. Modelo Clip de OpenAI

CLIP (Contrastive Language-Image Pre-Training) es un modelo desarrollado por investigadores de OpenAI a principios de 2021, parece más un sistema de reconocimiento de imágenes. Su diferencia radica en que no ha aprendido a reconocer imágenes a partir de ejemplos etiquetados en los conjuntos de datos seleccionados, como la mayoría de los modelos existentes, sino a partir de las imágenes y sus descripciones publicadas en internet. CLIP aprende qué hay en una imagen en función de una descripción en vez de una etiqueta de una palabra como *gato*.

CLIP fue entrenado con la orden de predecir cuál de las 32.768 descripciones de una selección aleatoria era la correcta para una imagen determinada. Para lograrlo, CLIP aprendió a vincular una amplia variedad de objetos con sus nombres y las palabras que los describen. Esto le permite identificar objetos en imágenes que no pertenecen a su conjunto de entrenamiento.

CLIP entrena previamente un codificador de imágenes y un codificador de texto para predecir qué imágenes se emparejaron con qué textos en nuestro conjunto de datos. Luego usamos este comportamiento para convertir CLIP en un clasificador de tiro cero ($zero - shot$). Convertimos todas las clases de un conjunto de datos en leyendas como **una foto de un perro** y predecimos la clase de la leyenda que CLIP estima que se empareja mejor con una imagen dada.



[12]

La mayoría de los sistemas de reconocimiento de imágenes se entrenan para identificar ciertos tipos de objetos, como rostros en los vídeos de vigilancia o edificios en las imágenes de satélite. Al igual que GPT-3, CLIP puede generalizar entre tareas sin necesidad de entrenamiento adicional. También es menos propenso que otros modelos de reconocimiento de imágenes de última generación a dejarse engañar por algunos ejemplos contradictorios, alterados sutilmente de distintas formas que suelen confundir los algoritmos, aunque las personas no noten la diferencia.

Referencias

1. https://bytespider.eu/3-tipos-de-sesgo-en-los-modelos-de-ia-y-como-podemos-abordarlos/
2. https://empresas.blogthinkbig.com/que-es-machine-bias-los-sesgos-en/
3. https://www.analyticslane.com/2019/04/24/que-es-el-sesgo-en-aprendizaje-automatico/
4. https://labelyourdata.com/articles/what-is-dataset-in-machine-learning
5. https://keepler.io/es/2021/03/la-dicotomia-sesgo-varianza-en-modelos-de-machine-learning/
6. https://azure.microsoft.com/es-es/resources/cloud-computing-dictionary/what-are-machine-learning-algorithms
7. https://www.apd.es/algoritmos-del-machine-learning/
8. https://learn.microsoft.com/es-es/windows/ai/windows-ml/what-is-a-machine-learning-model
9. https://geekflare.com/es/machine-learning-models/
10. https://www.campusmvp.es/recursos/post/que-peligro-implican-los-sesgos-en-los-modelos-de-inteligencia-artificial.aspx
11. https://www.technologyreview.es//s/13061/dalle-y-clip-dan-un-paso-mas-hacia-el-futuro-de-la-inteligencia-artificial
12. https://openai.com/research/clip
13. https://impulsatek.com/clip-claramente-explicado-que-es-y-como-funciona/
14. Bias And Unfairness in Machine Learning Models: A Sistematic Literature Review
15. Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. ACM Comput. Surv. 54, 6, Article 115 (July 2021), 35 pages
16. Tommaso Di Noia y col. Recommender systems under European AI regulations. En: Communications of the ACM 65.4 (2022), págs. 69-73 (vid. pág)
17. Yogesh K. Dwivedi y col. Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. En: International Journal of Information Management 57 (2021), pág. 101994 (vid. pág. 5).
18. Candice Schumann y col. We need fairness and explainability in algorithmic hiring. En: International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS). 2020 (vid. pág. 6).
19. Pascal D Konig y Georg Wenzelburger. When politicization stops algorithms in criminal justice. En: The British Journal of Criminology 61.3 (2021), págs. 832-851 (vid. pág. 6).
20. Benjamin Paaben y col. Dynamic fairness-Breaking vicious cycles in automatic decision making. En: arXiv preprint arXiv:1902.00375 (2019) (vid. págs. 6, 11, 12)
21. Claudio Feijóo y col. Harnessing artificial intelligence (AI) to increase wellbeing for all: The case for a new technology diplomacy. En: Telecommunications Policy 44.6 (2020), pág. 101988 (vid. pág. 6).
22. Julia Stoyanovich, Bill Howe y HV. Jagadish. Responsible data management. En: Proceedings of the VLDB Endowment 13.12 (2020) (vid. págs. 6, 9).
23. Joy Buolamwini y Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. En: Conference on fairness, accountability and transparency. PMLR. 2018, págs. 77-91 (vid. pág. 7).
24. Lauren Kirchner Jeff Larson Surya Mattu y Julia Angwin. Machine Bias ProPublica. https://www.propublica.org/article/machine-bias-riskassessments-in-criminal-sentencing. Mayo de 2016 (vid. pág. 7)

25. Julia Dressel y Hany Farid. The accuracy, fairness, and limits of predicting recidivism. En: Science advances 4.1 (2018), eaao5580 (vid. págs. 1, 7).
26. Claudia Wagner y col. Its a mans Wikipedia? Assessing gender inequality in an online encyclopedia. En: Proceedings of the international AAAI conference on web and social media. Vol. 9. 1. 2015, págs. 454-463 (vid. pág. 7).
27. Latanya Sweeney. Discrimination in online ad delivery. En: Communications of the ACM 56.5 (2013), págs. 44-54 (vid. pág. 7).
28. Matthew Kay, Cynthia Matuszek y Sean A Munson. Unequal representation and gender stereotypes in image search results for occupations. En: Proceedings of the 33rd annual acm conference on human factors in computing systems. 2015, págs. 3819-3828 (vid. pág. 7).
29. Jeffrey Pennington, Richard Socher y Christopher D Manning. Glove: Global vectors for word representation. En: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014, págs. 1532-1543 (vid. pág. 8).
30. Tomas Mikolov y col. Efficient estimation of word representations in vector space. En: arXiv preprint arXiv:1301.3781 (2013) (vid. pág. 8).
31. Tolga Bolukbasi y col. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. En: Advances in neural information processing systems 29 (2016) (vid. págs. 8, 15).
32. Sahil Verma y Julia Rubin. Fairness definitions explained. En: 2018 ieee/acm international workshop on software fairness (fairware). IEEE. 2018, págs. 1-7 (vid. pág. 11).
33. Moritz Hardt, Eric Price y Nati Srebro. Equality of opportunity in supervised learning. En: Advances in neural information processing systems 29 (2016) (vid. pág. 11).
34. Matt J. Kusner y col. Counterfactual fairness. En: Advances in neural information processing systems 30 (2017) (vid. pág. 11).
35. Batya Friedman y Helen Nissenbaum. 1996. Sesgo en los sistemas informáticos. ACM Trans. información sist. 14, 3 (julio de 1996), 330347. DOI: https://doi.org/10.1145/230538.230561 [54] Anna Fry, Thomas J.
36. Azadeh Nematzadeh, Giovanni Luca Ciampaglia, Filippo Menczer y Alessandro Flammini. 2017. Cómo algorítmico el sesgo de popularidad obstaculiza o promueve la calidad. preimpresión de arXiv arXiv:1707.00574 (2017)
37. Harini Suresh y John V. Guttag. 2019. Un marco para comprender las consecuencias no deseadas del aprendizaje automtico
38. Alexandra Olteanu, Carlos Castillo, Fernando Díaz y Emre Kcman. 2019. Datos sociales: sesgos metodológicos trampas y límites éticos. Fronteras en Big Data 2 (2019), 13.
39. [9] Ricardo Baeza-Yates. 2018. Sesgo en la web. comn ACM 61, 6 (mayo de 2018), 5461. DOI: https://doi.org/10.1145/3209581