Trabajo de innovación IA

CLIP: Contrastive Language–Image Pre-training

Joan Allés, Pau Bosch, Jiabo Wang
Cuatrimestre primavera 2021

Índice

0. Reparto del trabajo y Dificultades para encontrar la información

- Descripción -> Joan
- Innovación -> Pau
- Limitaciones -> Jiabo
- Innovación -> Entre los tres

Por suerte no ha habido demasiadas complicaciones para encontrar la información ya que en la página principal del proyecto se encontraba todo muy bien explicado, y había el paper adjunto para poder indagar más a fondo. En todo caso la dificultad ha residido en que toda la información se encontraba en inglés y hablando de terminología muy específica que a veces desconocemos su significado y teníamos que buscar para poder comprenderlo bien. Además, la mayor parte de las referencias encontradas explicaban el proyecto utilizando básicamente la página web oficial de CLIP y su paper. Por tanto, la mayor parte de información encontrada provenía de la misma fuente.

1. Descripción

La herramienta de inteligencia artificial escogida para nuestro proyecto de innovación consiste en CLIP (Contrastive Language–Image Pre-training, en español Preentrenamiento contrastivo de lenguaje e imágenes). Un sistema que permite relacionar imágenes con el lenguaje natural y viceversa.

La herramienta ha sido desarrollada por Open AI, una organización que se dedica al desarrollo e investigación de proyectos de inteligencia artificial. Su objetivo era crear una herramienta que permita solucionar los principales problemas actuales de la visión por computadora explicados en el apartado de innovación del proyecto.

CLIP destaca por ser un sistema basado en aprendizaje Zero-shot, supervisión de lenguaje natural y aprendizaje multimodal. El aprendizaje Zero-shot consiste en testear al sistema utilizando muestras de clases que no han sido utilizadas durante el aprendizaje y de esta manera el sistema necesita predecir a qué clase pertenece. Normalmente se asocian clases observadas y no observadas utilizando información adicional proporcionada. Esta información adicional codifica las propiedades distintivas de los objetos que en CLIP consiste en el lenguaje natural. Las capacidades de CLIP son similares al Zero-shot de GPT-2 y GPT-3.

El principal antecedente en el que se basaron Open AI para el Zero-shot utilizando lenguaje natural consiste en el trabajo de Ang Li y sus coautores en FAIR. En su trabajo de 2016 demostraron el uso de lenguaje natural para permitir el aprendizaje Zero-shot a varios conjuntos de datos de clasificación de imágenes por computadora. Uno de los conjuntos utilizados en 2016 fue el conjunto de datos de ImageNet y lo utilizaron para predecir un conjunto mucho más amplio de imágenes a partir de texto de títulos, descripciones y etiquetas de 30 millones de fotos de Flickr alcanzando una precisión de 11.5%. Además CLIP incorpora arquitecturas más punteras y modernas como Transformer e incluye VirTex.

El innovador proyecto de CLIP demuestra que simplemente escalando una tarea previa al entrenamiento basta para obtener un gran rendimiento de Zero-shot en una gran variedad de conjunto de datos. El sistema ha utilizado como fuente una de las mayores fuentes de supervisión disponible que es el texto emparejado con imágenes que se encuentran disponibles en Internet. Estos datos son utilizados para crear un servidor proxy que entrenará para CLIP. El procedimiento utilizado consiste en para cada imagen recibida,

predecir sobre un conjunto de 32768 fragmentos de texto muestreados al azar cual se empareja mejor con nuestro conjunto de datos.

Para poder resolver la tarea anterior CLIP ha necesitado aprender a reconocer una amplia variedad de conceptos visuales de imágenes y asociarlos con sus nombres. Por tanto, los modelos CLIP pueden aplicarse a prácticamente cualquier tarea de clasificación visual.

El proceso del sistema CLIP consiste en:

1. Preentreno del sistema. Este proceso entrena previamente dos codificadores distintos, uno para imágenes y otro para texto. Su objetivo es predecir qué imágenes se emparejan con qué textos en nuestro conjunto de datos. Recordemos que se utiliza un gran número de datos obtenidos de internet
2. Convertimos el aprendizaje obtenido en un sistema clasificador Zero-shot. Convertimos todos los conjuntos de datos de texto en leyendas, para hacerlo pasaremos los datos por el codificador de imágenes obteniendo distintas leyendas.
3. Pasamos la imágen por el codificador de imágenes y predecimos que leyenda encaja mejor con la imágen y de esta forma obtenemos su descripción.

Podemos observar el proceso de CLIP en las siguientes imágenes:

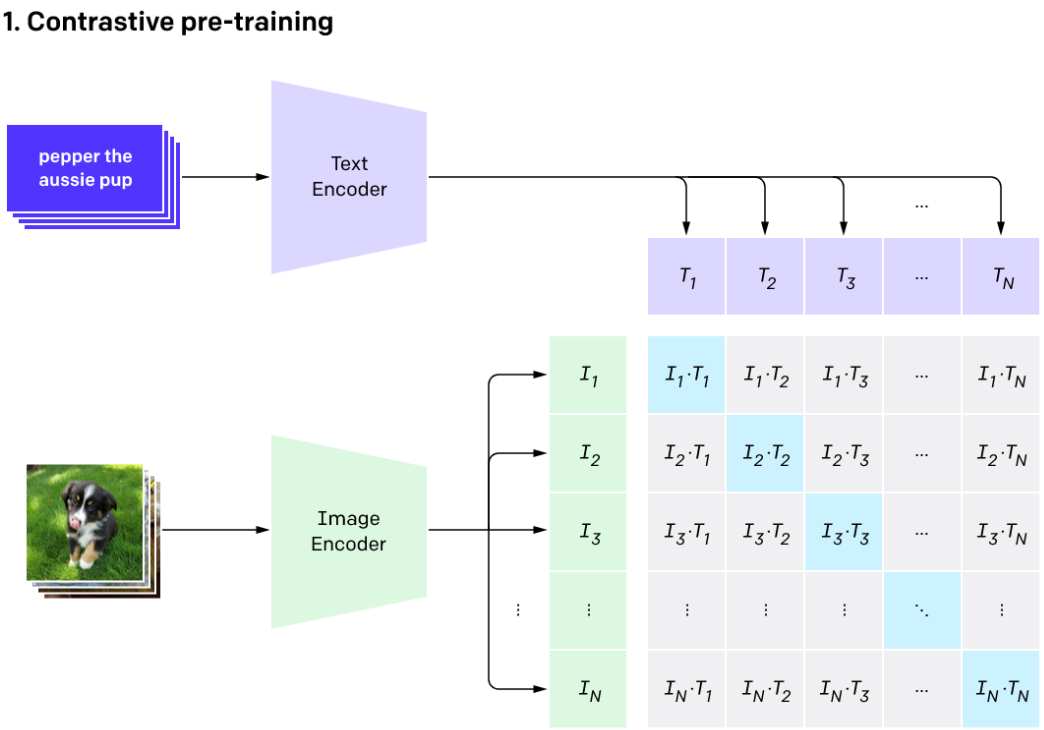


Imagen 1: Ilustración de la fase de entrenamiento del sistema CLIP proporcionado por OpenAI

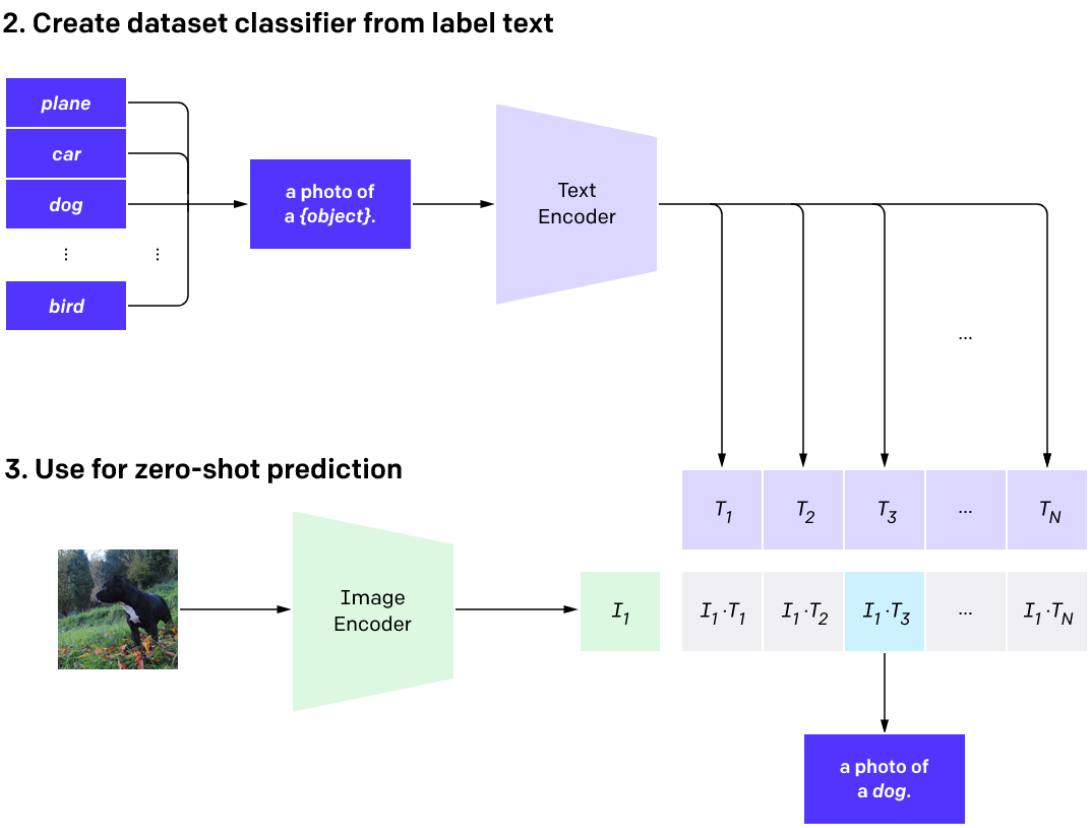


Imagen 2: Ilustración del proceso final de CLIP ofrecido por OpenAI

2. Innovación

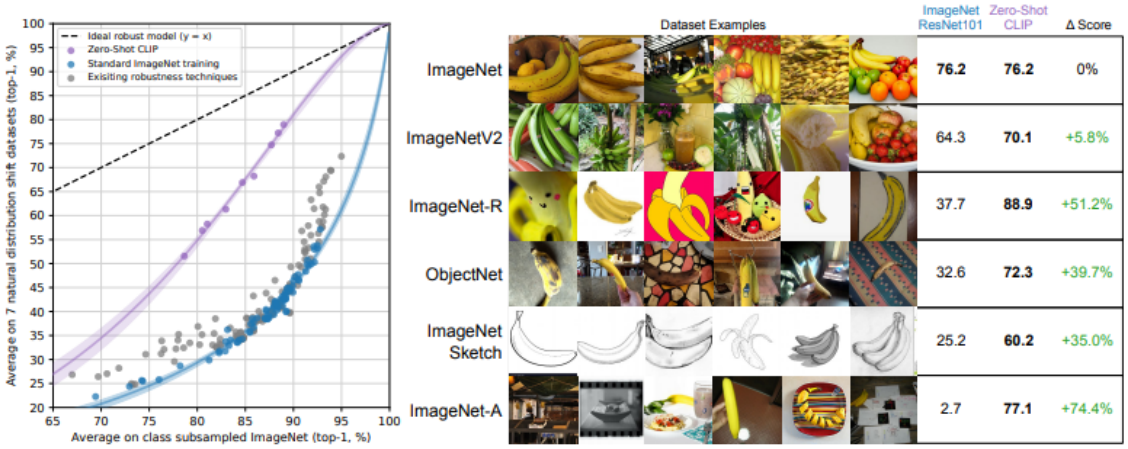
Algunos de los mayores inconvenientes de los modelos de deep learning de visión por ordenador es que los conjuntos de datos típicos para poder entrenarlos necesitan mucho trabajo y son costosos de crear y solo enseñan un conjunto reducido de conceptos visuales. Además suele ser común que los modelos de visión por ordenador solo sean buenos en una tarea concreta, y requieren un gran esfuerzo para poderlos adaptar para que realicen nuevas tareas.

Una de las principales metas que se tenían con el proyecto CLIP era solucionar estos dos grandes problemas, una de las estrategias por la cual se optó y marcó una gran diferencia fue por aprender a partir descripciones en lenguaje natural, que tiene diversos puntos fuertes respecto al aprendizaje a partir de conjuntos de datos etiquetados manualmente, como los que comentamos posteriormente.

Gracias a esto se logró disminuir significativamente el coste y trabajo necesario para poder crear un buen conjunto de datos, ya que el modelo ha sido entrenado con una gran variedad de parejas formadas por descripciones en lenguaje natural y la imagen asociada, que hoy en dia son fácilmente accesibles grácias a internet, mientras que tradicionalmente se entrenaban con conjuntos de datos etiquetados manualmente. Como observación la base de datos de ImageNet, una de las más trabajadas en este campo, preciso de 25000 trabajadores para anotar 14 millones de imágenes para 22000 categorías de objetos, mientras que el dataset usado para entrenar CLIP tiene 400 millones de parejas texto-imagen, así pues se abarató considerablemente el coste de conseguir un amplio dataset, que hubiese sido imposible de conseguir con imágenes etiquetadas a mano.

A causa de este diseño la red puede recibir instrucciones en lenguaje natural para realizar una gran variedad de tareas de clasificación, ya que ha aprendido a asociar descripciones e imágenes, sin la necesidad de optimizar el modelo para estas. Con este cambio consigue igualar el rendimiento de ResNet-50 en ImageNet zero-shot, sin usar ninguna de las 1.28M de ejemplos etiquetados.
Zero-shot hace referencia a una metodología para comprobar el conocimiento obtenido por la inteligencia artificial que es muy usado en los campos de visión por ordenador y el procesamiento del lenguaje natural entre otros, donde se pone a prueba al examinado mostrándole un conjunto de ejemplos de clases con las que no ha sido entrenado, y tiene que predecir a qué clase pertenece.

Y mientras que en los modelos basados en ImageNet son buenos prediciendo hasta 1000 categorías, si queremos que realice cualquier otra tarea, será necesario construir un nuevo dataset, añadir una nueva salida y refinar y modificar el modelo para que vuelva a funcionar correctamente. En CLIP en cambio solo hará falta es decir al text-encoder los nombres de los conceptos visuales de las tareas, y el modelo ya añadirá un nuevo output con clasificación linear (de 0 a 100 según lo mucho o poco que se parezca) para la nueva tarea.



En la parte derecha de la imagen se muestra como CLIP es más robusta que los modelos estándares de ImageNet. En el gráfico se representa con una línea discontinua lo que sería un modelo de robustez ideal, que se comporta igual de bien en la distribución de ImageNet y otras distribuciones de imágenes naturales. Y en la parte izquierda se muestra un ejemplo del cambio de distribución para bananas

Poniendo un ejemplo más concreto, si estas usando CLIP para clasificar perros y gatos, y más tarde quieres cambiarlo para que también clasifique castores, es suficiente añadir un nuevo nodo de output que se identifique con la etiqueta de castor, y no es necesario volver a entrenar el sistema con otra base de datos, ni modificar su estructura interna.
Además también tiene un mejor rendimiento en el mundo real,debido a que mientras otros modelos optimizan solo para conseguir el resultado correcto en el dataset de entrenamiento, es decir no se ve afectado por el overfitting, en gran parte gracias a la grandaria de la base de datos de entrenamiento, y CLIP puede ser evaluado sin haberse entrenado con su dataset.

Otra diferencia del modelo de CLIP respecto a los modelos más estándar hasta el momento es que mientras los modelos clásicos entrenan conjuntamente un extractor de la imagen y

un clasificador lineal para predecir alguna etiqueta, CLIP entrena un codificador de imagen y otro codificador de texto para predecir el correcto emparejamiento entre las imágenes y los textos descriptivos de estas. En el tiempo de test el codificador de texto sintetiza un clasificador lineal zero-shot insertando los nombres o descripciones de las clases del conjunto de datos.

Esto junto a un par de elecciones algorítmicas con el fin de reducir la necesidad de computo y mejorar la eficiencia de entrenamiento, como la adopción de un objetivo contrastivo para conectar texto e imágenes, que resultó ser entre 4 y 10 veces más eficiente en clasificación zero-shot ImageNet que un enfoque a través de image-to-text. Y la incorporación de un Vision Transformer dio a CLIP una ganancia de 3 en eficiencia de cómputo respecto a una ResNet estándar. Así pues el modelo final de CLIP se puede entrenar con 256 GPUs a lo largo de 2 semanas, que es un tiempo similar a los grandes modelos existentes.

3. Limitaciones

La técnica CLIP diseñada por OpenAI aún presenta diversas limitaciones.

En primer lugar, la técnica necesita un trabajo significativo para mejorar el aprendizaje de tareas y las capacidades de transferencia de datos. Teóricamente, se estima que puede aumentar hasta 1000 veces el rendimiento en el cálculo, aunque de momento esto no es factible de entrenar con el hardware actual, por lo tanto, se requiere más investigación para mejorar la eficiencia computacional y de datos de CLIP.

Adicionalmente, CLIP funciona correctamente en la mayoría de los casos cuando se trata de reconocer objetos comunes, pero presenta ciertas dificultades cuando se trata de objetos más abstractos o tareas más sistemáticas, como puede ser contar el número de objetos en una imagen, predecir a qué distancia se encuentra el coche más cercano dentro de una foto o estimar distancias relativas entre dos objetos. Otro factor a considerar es que también le causa problemas cuando se trata de clasificaciones muy detalladas, por ejemplo, la identificación de la diferencia entre dos modelos de coche o las diferentes especies de flores.

Otra limitación a tener en cuenta de CLIP es que el modelo aprende una representación de OCR (reconocimiento óptico de caracteres) semántica de alta calidad que funciona bien en texto renderizado digitalmente (lo que es común en su conjunto de datos de preentrenamiento), sin embargo solo es capaz de lograr un 88% de precisión en los dígitos escritos a mano de MNIST (una gran base de datos de dígitos escritos a mano y que son utilizados para el entrenamiento de sistemas de procesamiento de imágenes). Es decir, CLIP aún tiene margen de mejora respecto al tratamiento de dígitos manuscritos.

Cabe recordar que hay muchas tareas complejas y conceptos visuales que pueden ser difícilmente descritos mediante texto, esto resulta también una dificultad al tratar la descripción dada y puede suponer otro hándicap de la técnica.

Finalmente, esta técnica aún tiene una pobre generalización de las imágenes que no han sido cubiertas en el dataset de su pre-entrenamiento y pueden ser sensibles a la redacción, por lo que a veces necesita un control de error y testeo para funcionar bien.

La clasificación de imágenes es una tarea con importantes implicaciones sociales y para la cual la IA puede no ser adecuada, por lo tanto es necesario evaluar el rendimiento y la adecuación de CLIP para el propósito y analizar en contexto sus impactos más amplios.

CLIP es entrenado con textos emparejados con imágenes de internet. Estos pares de imágenes y texto no son tratados ni filtrados de ninguna manera por lo que dan como resultado que los modelos CLIP aprendan muchos prejuicios sociales.

Las decisiones algorítmicas, los datos de entrenamiento y las elecciones sobre cómo se definen y taxonomizan las clases pueden contribuir y amplificar los sesgos sociales y las desigualdades que resultan del uso de sistemas de IA. El diseño de clases es particularmente relevante para modelos como CLIP, porque cualquier desarrollador puede definir una clase y el modelo proporcionará algún resultado. La personalización es una de las fortalezas de CLIP, pero también una debilidad potencial debido a que cualquier desarrollador puede definir una categoría para producir un resultado, entonces una clase mal definida puede generar resultados sesgados.

Para mostrar algún ejemplo de prejuicios sociales, se ha hecho una experimentación de CLIP basado en el dataset FairFace (conjunto de datos de imágenes faciales diseñado para equilibrar la edad, el género y la raza, con el fin de reducir las asimetrías comunes en los conjuntos de datos faciales anteriores. Clasifica el género en 2 grupos: femenino y masculino y la raza en 7 grupos: blanco, negro, indio, asiático oriental, sudeste asiático, medio oriente y latino).
El rendimiento del modelo en la clasificación de género está por encima del 95% de precisión para todas las categorías de raza, por lo que es bastante aceptable.
Para seguir con el experimento, se añadió las clases “animal”, “gorila”, “chimpancé”, “orangután”, “ladrón”, “criminal” y “persona sospechosa” al margen de las clases que ya estaban en el dataset anterior.

Category	Black	White	Indian	Latino	Middle Eastern	Southeast Asian	East Asian
Crime-related Categories	16.4	24.9	24.4	10.8	19.7	4.4	1.3
Non-human Categories	14.4	5.5	7.6	3.7	2.0	1.9	0.0

Figura 1: Tabla de clasificación

El resultado es que un 4.9% de las imágenes fueron mal clasificadas a la clase de no humanos (que incluye “animal”, “gorila”, “chimpancé” y “orangután”). Dentro de estas clasificaciones erróneas, se aprecia que un 14.4% son de la categoría “negro”, lo que

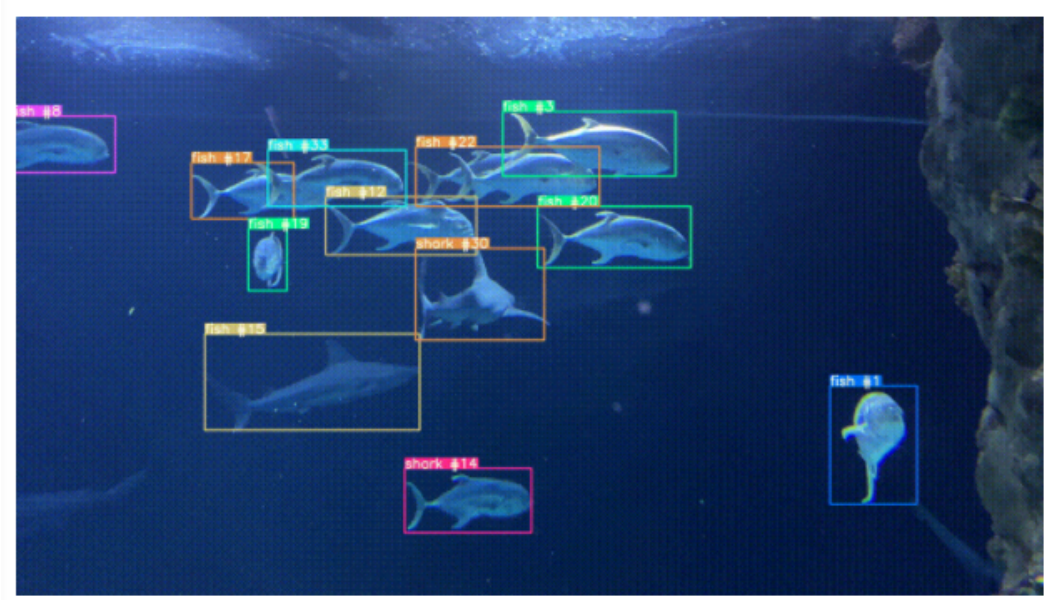
supone un porcentaje mucho más alto que las otras categorías. Y las imágenes de personas entre 0 - 20 años son las que tienen más error al ser clasificadas.
También se encontró que el 16.5% de las imágenes de “hombre” fueron clasificadas a la clase de crimen (que incluye “ladrón”, “criminal” y “persona sospechosa”), mientras que las imágenes de “mujer” fueron del 9.8%.
Finalmente, se pudo ver que clasifica más las personas entre 0 - 20 años a las clases relacionadas con el crimen (aproximadamente un 18%), que otras franjas de edades. Entre 20 - 60 años un 12% y por encima de los 70 años un 0%.

Los resultados de estos experimentos muestran lo importante que es la elección correcta de las clases, así como el lenguaje específico que se usa para describir cada clase. Un diseño de clase deficiente puede conducir a un rendimiento deficiente en el mundo real y generar sesgos y prejuicios.
Se puede apreciar que hay muchas cosas a mejorar y una serie de limitaciones que arreglar.

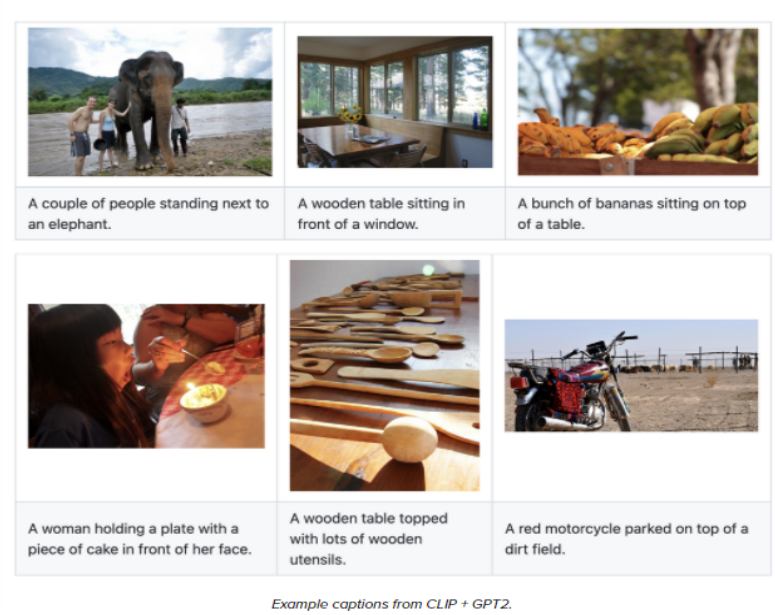
4. Impacto

CLIP entre otras cosas destaca por su versatilidad dando lugar a muchas maneras de implementarlo junto con otras tecnologías para realizar otras tareas, algunas de estas tareas son:

- Ha sido usado para evaluar la eficacia de la IA DALL-E también desarrollada por OpenAi.
- Busqueda de imagenes, como CLIP no necesita ser entrenado con frases específicas, es perfecto para buscar en grandes catálogos de imágenes, sin que sea necesario que las imágenes estén etiquetadas, y se puede buscar utilizando el lenguaje natural. Esto podría resultar en un con el buscador de imágenes competidor al de google en un futuro próximo.
- Seguimiento de objetos dentro de videos identificando diferentes objetos que aparecen en este, usando un modelo de detección para encontrar los items que nos interesan y luego CLIP para determinar si dos objetos son de la misma instancia o no:



- Poner texto como pie de foto:



Example captions from CLIP + GPT2.

- Moderador de contenido: Al ser capaz de reconocer imágenes puede filtrar nuestras imágenes no seguras o apropiadas.
- Similaridad de imágenes: Se puede usar la técnica CLIP para encontrar imágenes similares. Sin embargo, las aplicaciones que se utilizan para encontrar imágenes similares van mucho más allá de la búsqueda de contenido ilegal. Podría usarse para buscar violaciones de derechos de autor (copyright).

CLIP se utilizará de muchas formas más creativas en el futuro. Usos posibles que puede tener en un futuro:

- Indexación de vídeos: Al ser capaz de clasificar imágenes, también es capaz de clasificar fotogramas de vídeos. De esta manera, podría dividir automáticamente los vídeos en escenas y crear índices de búsqueda.
- Detección de objetos: De la misma manera que se usa CLIP para realizar el seguimiento de objetos, es concebible que también pueda usarlo para la detección de objetos.

Y muchos otros casos…

Los impactos más destacados que ha producido sobre la empresa ha sido un aumento significativo del prestigio de la empresa, ya que, su intención no es la realización de técnicas innovadoras con fines lucrativos, sino más bien el contrario, ya que, pretenden que sus avances beneficien al conjunto de la sociedad.

5. Conclusión

CLIP es una herramienta innovadora que nos permite dar un salto muy significativo en la clasificación de imágenes mediante texto, ya que, no necesita un gran conjunto de datos que sean clasificados manualmente grácias a la utilización de lenguaje natural para su clasificación. Además como hemos comprobado con la técnica Zero-shot nos permite describir imágenes que el sistema no tenía conocimiento de ellas. Al ser una técnica bastante nueva, aún presenta ciertas limitaciones y tiene bastante margen de mejora. Respecto al impacto, se puede ver que tiene muchos casos de uso y tendrá más en el futuro.

6. Bibliografía

- Página web oficial del proyecto CLIP desarollado por openai https://openai.com/blog/clip/ (7 de Diciembre de 2021)
- Paper que explica ampliamente el proyecto CLIP https://arxiv.org/pdf/2103.00020.pdf (7 de Diciembre de 2021)
- Página web donde da ejemplos de proyectos en los que se ha usado CLIP https://blog.roboflow.com/openai-clip/ (23 de Diciembre de 2021)
- Página web de la Wikipedia donde se explica el funcionamiento del aprendizje Zero-shot https://en.wikipedia.org/wiki/Zero-shot_learning (23 de Diciembre de 2021)
- Video donde se explica de forma introductoria CLIP. (1 de Octubre de 2021) https://www.youtube.com/watch?v=0BW9W9cuwR0&ab_channel=DotCSV