OpenAI's CLIP model

Joan Allés
Pau Bosch
Jiabo Wang

Índice

Descripción

C → Constrastive

L → Language

I → Image

P → Pre-Training

Preentrenamiento constrativo de lenguaje e imágenes

Desarrollo

- Organización dedicada al desarrollo e investigación de proyectos de inteligencia artificial para crear soluciones a problemas actuales

- CLIP ha sido creado como una herramienta que permita solucionar los principales problemas actuales de visión por computadora

Características básicas

Aprendizaje Zero-shot → Testeo del sistema utilizando muestras no utilizadas durante el aprendizaje

Supervisión del lenguaje natural → Utilización del lenguaje natural para la clasificación de imágenes

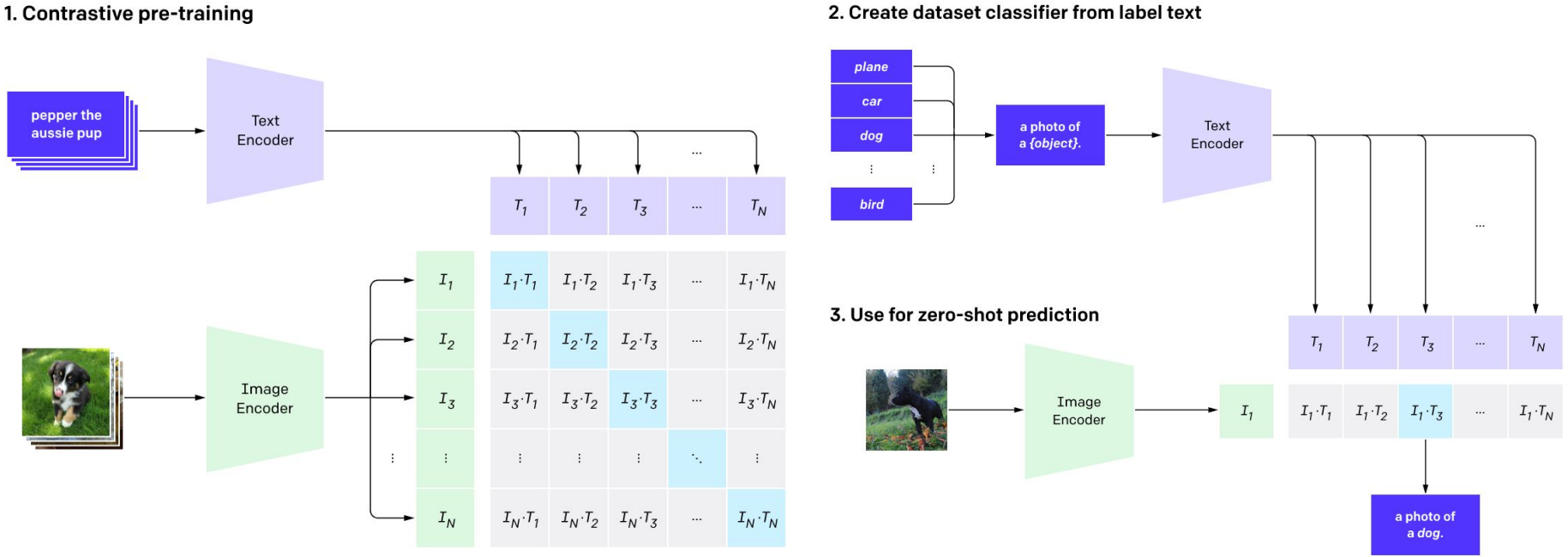
Aprendizaje multimodal → Uso de diversos tipos de aprendizaje

Principales antecedentes

Trabajo de Ang Li y sus coautores en FAIR que demostraron en 2016 el uso del lenguaje natural para permitir aprendizaje a varios conjuntos de datos de clasificación de imágenes.

Arquitecturas punteras y modernas como Transformer e incluye VirTex, herramientas innovadoras en el procesamiento del lenguaje natural.

Proceso del sistema CLIP



Innovación

Problemas de los modelos clásicos de deep learning de visión por ordenador:

- La base de datos necesaria para entrenarlos requiere mucho trabajo y son costosas de crear.

- Solo son buenos en una tarea concreta.

- Es muy costoso adaptar un modelo para que realicen una nueva tarea.

Innovación

- Aprendizaje a partir de descripciones en lenguaje natural.

- Parejas formadas por descripciones e imágenes fácilmente accesible en grandes cantidades en internet.

Ventajas del lenguaje natural

- ImageNet : 14 millones de imágenes etiquetadas para coincidir con alguna de las 22000 categorías de objetos que contiene la base de datos. Fueron necesarios 25000 trabajadores.

- CLIP 400 millones de parejas descripción-imagen, sacadas directamente de internet.

Ventajas del lenguaje natural

- Puede recibir instrucciones de que clasificar en lenguaje natural, dando lugar a una gran variedad de tareas de clasificación que no viene limitada por las categorías etiquetadas como en los diseños tradicionales.

- Consigue igualar el rendimiento de ResNet-50 en ImageNet zero-shot, sin usar ninguna de las 1.28M de ejemplos etiquetados.

Ventajas del lenguaje natural

- En los modelos clásicos si se quiere utilizar el mismo modelo para clasificar una nueva categoría será necesario modificar y refinar el modelo para que vuelva a funcionar

- En CLIP solo es necesario indicar al text-encoder los conceptos visuales de las nuevas tareas.

Otras diferencias

- Los modelos más estándar hasta el momento 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

Limitaciones

- Aprendizaje de tareas y capacidad de transferencia de datos.

- Fácil reconocimiento de objetos comunes, pero no de tareas complejas o objetos abstractos, como por ejemplo: contar el número de objetos de una imagen, estimar distancias relativas entre dos objetos, identificación de la diferencia entre dos modelos de coches, …

- Tareas o conceptos visuales que son difíciles de ser descritos de manera clara, también supone una dificultad para la técnica.

- Pobre generalización de imágenes no cubiertas en el pre-entrenamiento.

Implicaciones sociales

- CLIP es entrenado con textos emparejados con imágenes de internet, estos no son tratados ni filtrados de ninguna manera por lo que los modelos CLIP pueden aprender 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.

Experimento

- Dataset Fairface (distingue dos grupos de género: hombre y mujer; y 7 de raza: blanco, negro, indio, asiático oriental, sudeste asiático, medio oriente y latino).
- Se añaden las clases siguientes: “animal”, “gorila”, “chimpancé”, “orangután”, “ladrón”, “criminal” y “persona sospechosa”

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

4.9 % de imágenes mal clasificadas a la categoría de “no humano”
16.5 % de hombres mal clasificadas a la categoría de “crimen” por 9.8 % de mujeres, la mayoría de los cuales son jóvenes (entre 0 - 20 años).

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).
- Búsqueda de imágenes.
- Seguimiento de objetos dentro de videos identificando diferentes objetos que aparecen en este.
- Poner texto como pie de foto.
- Moderador de imágenes.
- Similaridad de imágenes.

Impacto

CLIP se utilizará de muchas formas más creativas en el futuro. Posibles usos que puede tener en un futuro no muy lejano:

- Indexación de vídeos
- Detección de objetos.
- etc.

Impacto más significativo que ha tenido sobre la empresa ha sido un aumento significativo de su prestigio.

Conclusiones

- Herramienta innovadora, que permite dar un salto de calidad en la clasificación de imágenes.

- Presenta ciertas limitaciones y margen de mejora.

- Tiene un gran variedad de casos de uso y un gran potencial de cara al futuro.