Annotation d’images

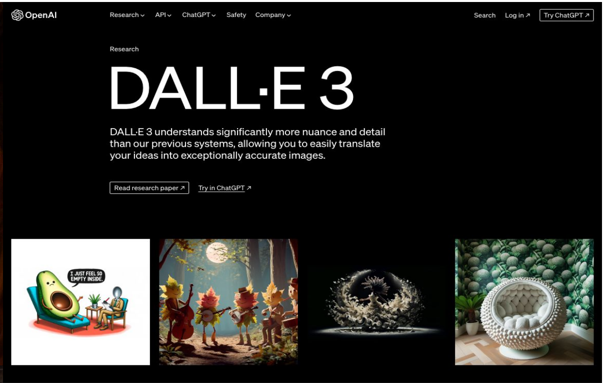
Mounir Bendali-Braham

Deep Learning Strasbourg - 20/01/2024

L’objectif de l’annotation d’images

Midjourney v6

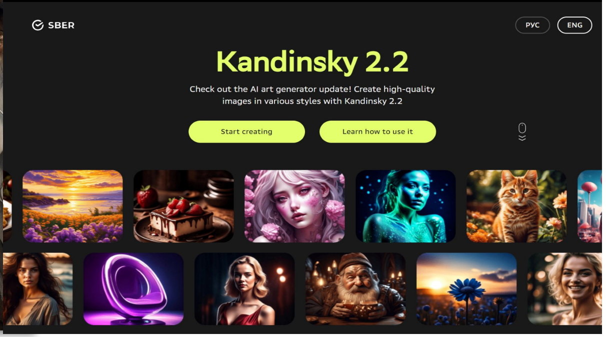
Stable Diffusion XL



Research

DALL·E 3

DALL·E 3 understands significantly more nuance and detail than our previous systems, allowing you to easily translate your ideas into exceptionally accurate images.

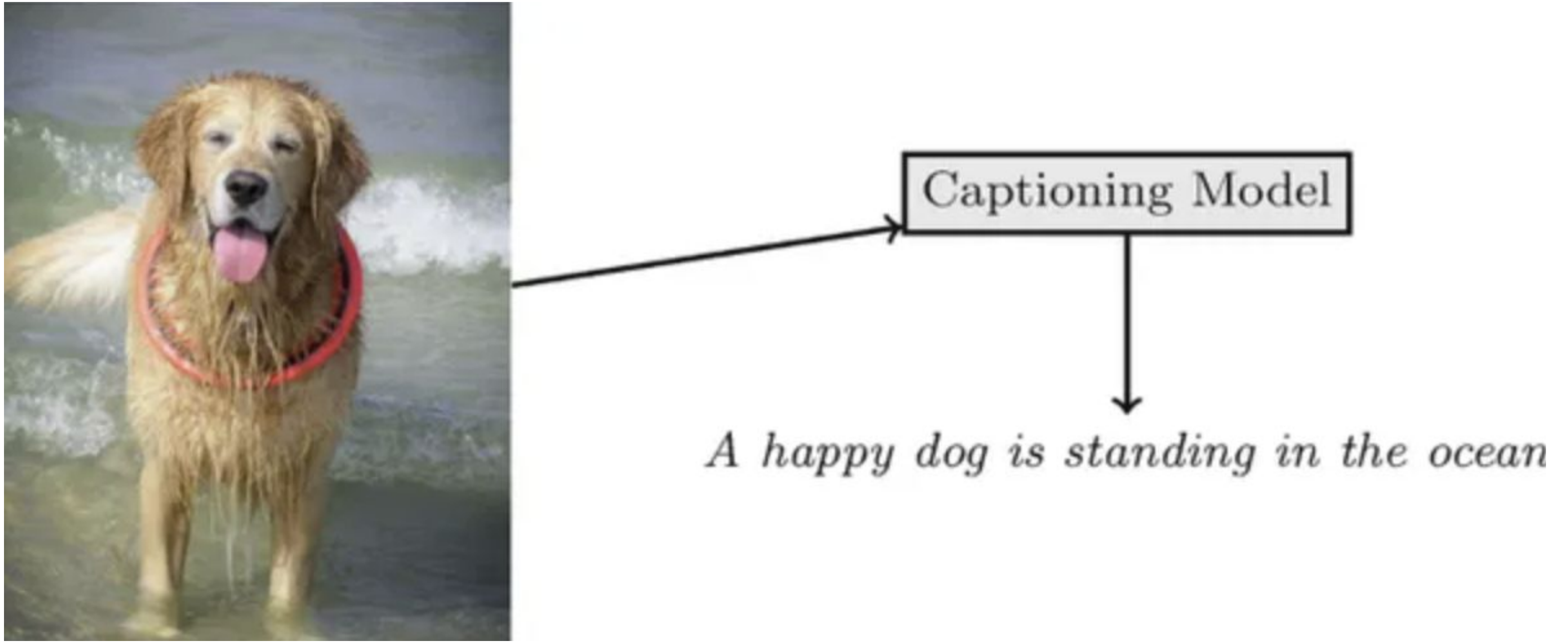


Kandinsky 2.2

Check out the AI art generator update! Create high-quality images in various styles with Kandinsky 2.2

c.f. slide “Références” pour les sources de ces images

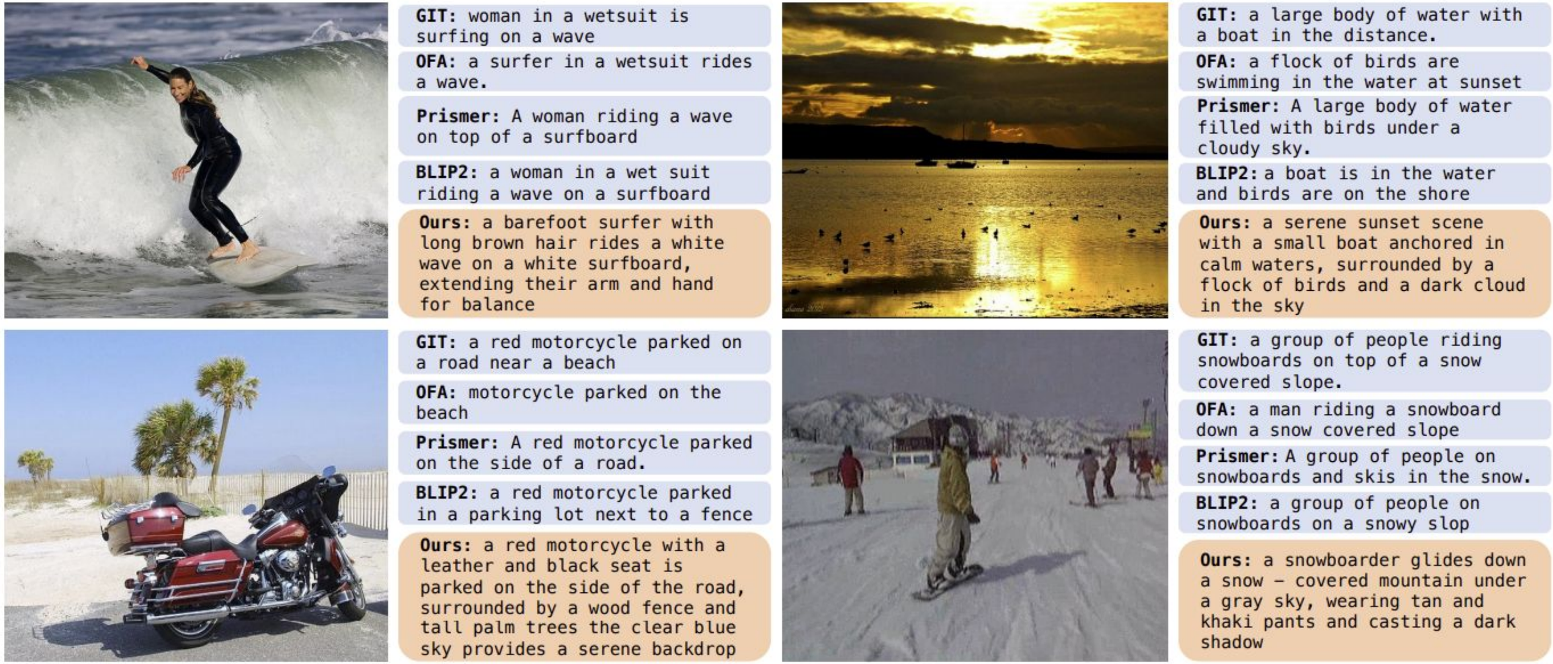
Qu’est-ce que l’annotation d’images ?



Source: https://www.analyticsvidhya.com/blog/2023/06/vision-transformers/

Comparaison de quelques modèles d’annotation

Picture taken from FuseCap’s article



GIT: woman in a wetsuit is surfing on a wave

OFA: a surfer in a wetsuit rides a wave.

Prismer: A woman riding a wave on top of a surfboard

BLIP2: a woman in a wet suit riding a wave on a surfboard

Ours: a barefoot surfer with long brown hair rides a white wave on a white surfboard, extending their arm and hand for balance

GIT: a large body of water with a boat in the distance.

OFA: a flock of birds are swimming in the water at sunset

Prismer: A large body of water filled with birds under a cloudy sky.

BLIP2: a boat is in the water and birds are on the shore

Ours: a serene sunset scene with a small boat anchored in calm waters, surrounded by a flock of birds and a dark cloud in the sky

GIT: a red motorcycle parked on a road near a beach

OFA: motorcycle parked on the beach

Prismer: A red motorcycle parked on the side of a road.

BLIP2: a red motorcycle parked in a parking lot next to a fence

Ours: a red motorcycle with a leather and black seat is parked on the side of the road, surrounded by a wood fence and tall palm trees the clear blue sky provides a serene backdrop

GIT: a group of people riding snowboards on top of a snow covered slope.

OFA: a man riding a snowboard down a snow covered slope

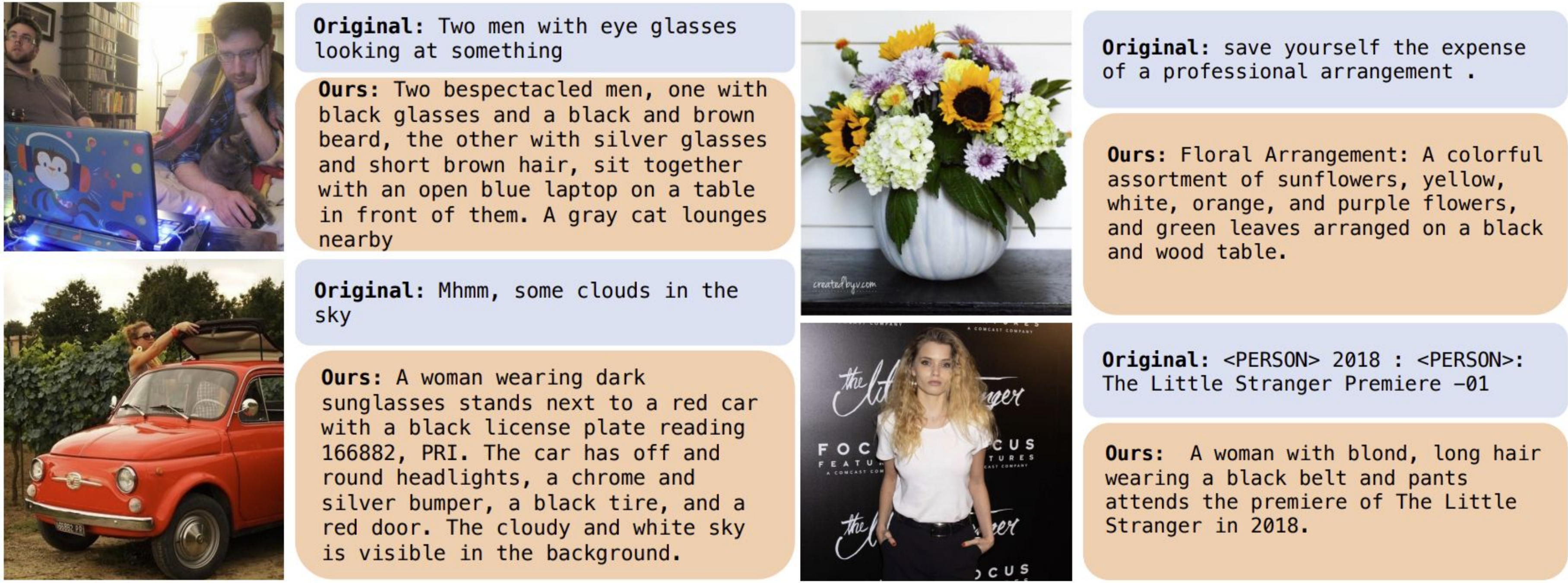
Prismer: A group of people on snowboards and skis in the snow.

BLIP2: a group of people on snowboards on a snowy slop

Ours: a snowboarder glides down a snow – covered mountain under a gray sky, wearing tan and khaki pants and casting a dark shadow

Vue d’ensemble des modèles d’annotation: FuseCap

From FuseCap article “how a good a bad captionner describe an image”



Original: Two men with eye glasses looking at something

Ours: Two bespectacled men, one with black glasses and a black and brown beard, the other with silver glasses and short brown hair, sit together with an open blue laptop on a table in front of them. A gray cat lounges nearby

Original: Mhmm, some clouds in the sky

Ours: A woman wearing dark sunglasses stands next to a red car with a black license plate reading 166882, PRI. The car has off and round headlights, a chrome and silver bumper, a black tire, and a red door. The cloudy and white sky is visible in the background.

Original: save yourself the expense of a professional arrangement .

Ours: Floral Arrangement: A colorful assortment of sunflowers, yellow, white, orange, and purple flowers, and green leaves arranged on a black and wood table.

Original: <PERSON> 2018 : <PERSON>: The Little Stranger Premiere -01

Ours: A woman with blond, long hair wearing a black belt and pants attends the premiere of The Little Stranger in 2018.

Vue d’ensemble des modèles d’annotation: MiniGPT-4

Who are the people in the image?

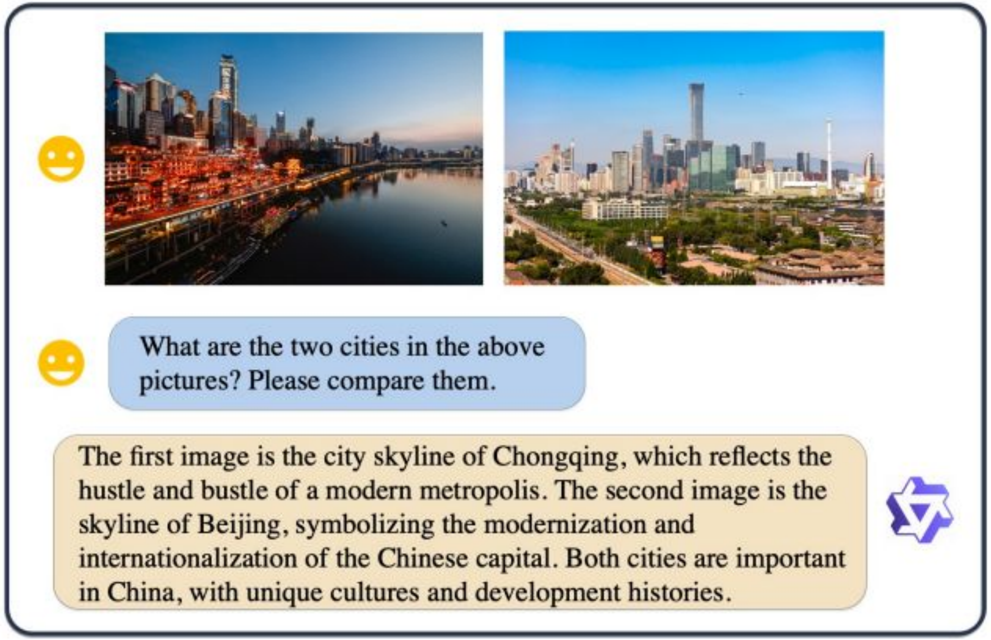
The people in the image are Goku and Vegeta, two characters from the Dragon Ball series. Goku is a saiyan warrior and the main protagonist of the series, while Vegeta is a saiyan prince and the main antagonist. They are both known for their powerful abilities and their rivalry.

Can you give me a short introduction to this movie?

The Godfather is a classic American crime drama film directed by Francis Ford Coppola and released in 1972. It is based on the novel of the same name by Mario Puzo and follows the story of the Corleone family, a powerful mafia clan in New York City. The film stars Marlon Brando as the patriarch of the family, Vito Corleone, and Al Pacino as his son, Michael Corleone. The Godfather is known for its iconic performances, memorable dialogue, and its portrayal of the American mafia lifestyle. It is considered one of the greatest films of all time and has had a lasting impact on popular culture.

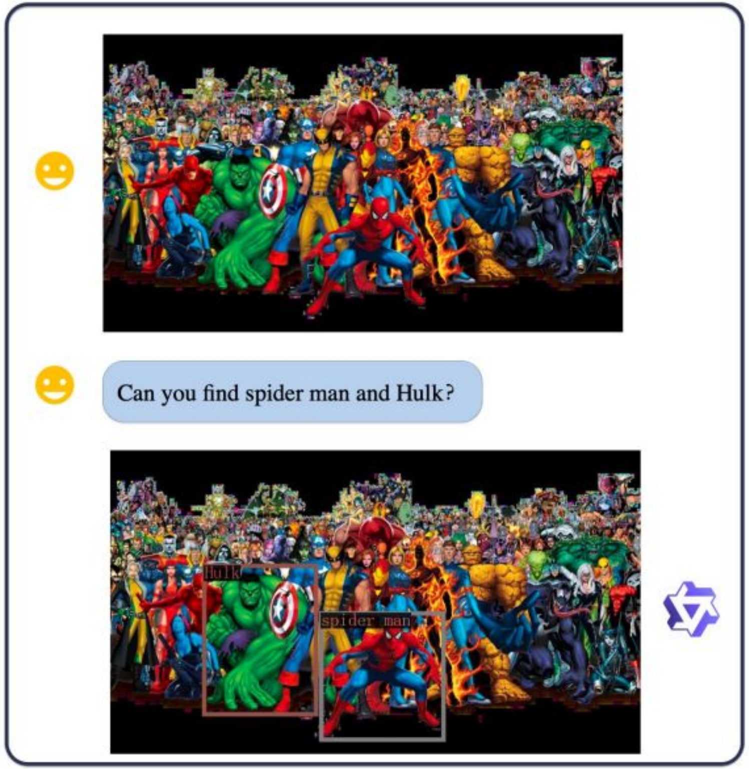
https://minigpt-4.github.io/

Vue d’ensemble des modèles d’annotation: Qwen-VL (1)



What are the two cities in the above pictures? Please compare them.

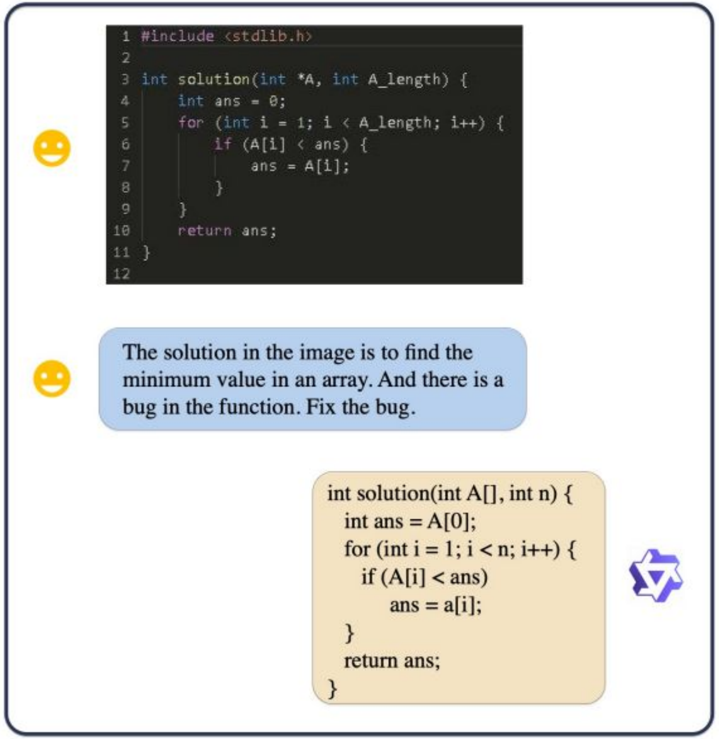
The first image is the city skyline of Chongqing, which reflects the hustle and bustle of a modern metropolis. The second image is the skyline of Beijing, symbolizing the modernization and internationalization of the Chinese capital. Both cities are important in China, with unique cultures and development histories.



Can you find spider man and Hulk?

https://arxiv.org/pdf/2308.12966.pdf

Vue d’ensemble des modèles d’annotation: Qwen-VL (2)

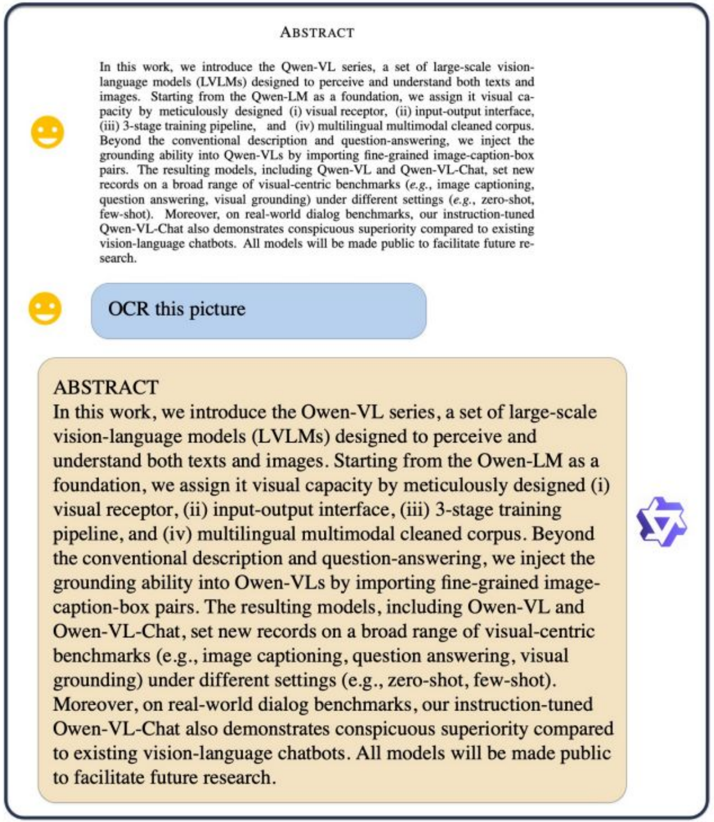


```
#include <stdlib.h>

int solution(int *A, int A_length) {
    int ans = 0;
    for (int i = 1; i < A_length; i++) {
        if (A[i] < ans) {
            ans = A[i];
        }
    }
    return ans;
}

```

The solution in the image is to find the minimum value in an array. And there is a bug in the function. Fix the bug.

```
int solution(int A[], int n) {
  int ans = A[0];
  for (int i = 1; i < n; i++) {
    if (A[i] < ans)
       ans = a[i];
  }
  return ans;
}
```

https://arxiv.org/pdf/2308.12966.pdf



ABSTRACT

In this work, we introduce the Qwen-VL series, a set of large-scale vision-language models (LVLMs) designed to perceive and understand both texts and images. Starting from the Qwen-LM as a foundation, we assign it visual capacity by meticulously designed (i) visual receptor, (ii) input-output interface, (iii) 3-stage training pipeline, and (iv) multilingual multimodal cleaned corpus. Beyond the conventional description and question-answering, we inject the grounding ability into Qwen-VLs by importing fine-grained image-caption-box pairs. The resulting models, including Qwen-VL and Qwen-VL-Chat, set new records on a broad range of visual-centric benchmarks (*e.g.*, image captioning, question answering, visual grounding) under different settings (*e.g.*, zero-shot, few-shot). Moreover, on real-world dialog benchmarks, our instruction-tuned Qwen-VL-Chat also demonstrates conspicuous superiority compared to existing vision-language chatbots. All models will be made public to facilitate future research.

OCR this picture

ABSTRACT

In this work, we introduce the Owen-VL series, a set of large-scale vision-language models (LVLMs) designed to perceive and understand both texts and images. Starting from the Owen-LM as a foundation, we assign it visual capacity by meticulously designed (i) visual receptor, (ii) input-output interface, (iii) 3-stage training pipeline, and (iv) multilingual multimodal cleaned corpus. Beyond the conventional description and question-answering, we inject the grounding ability into Owen-VLs by importing fine-grained image-caption-box pairs. The resulting models, including Owen-VL and Owen-VL-Chat, set new records on a broad range of visual-centric benchmarks (e.g., image captioning, question answering, visual grounding) under different settings (e.g., zero-shot, few-shot). Moreover, on real-world dialog benchmarks, our instruction-tuned Owen-VL-Chat also demonstrates conspicuous superiority compared to existing vision-language chatbots. All models will be made public to facilitate future research.

Jeux de données d'annotation d'images

A computer screen with a Windows message about Microsoft license terms.

A can of green beans is sitting on a counter in a kitchen.

A photo taken from a residential street in front of some homes with a stormy sky above.

A blue sky with fluffy clouds, taken from a car while driving on the highway.

A hand holds up a can of Coors Light in front of an outdoor scene with a dog on a porch.

A digital thermometer resting on a wooden table, showing 38.5 degrees Celsius.

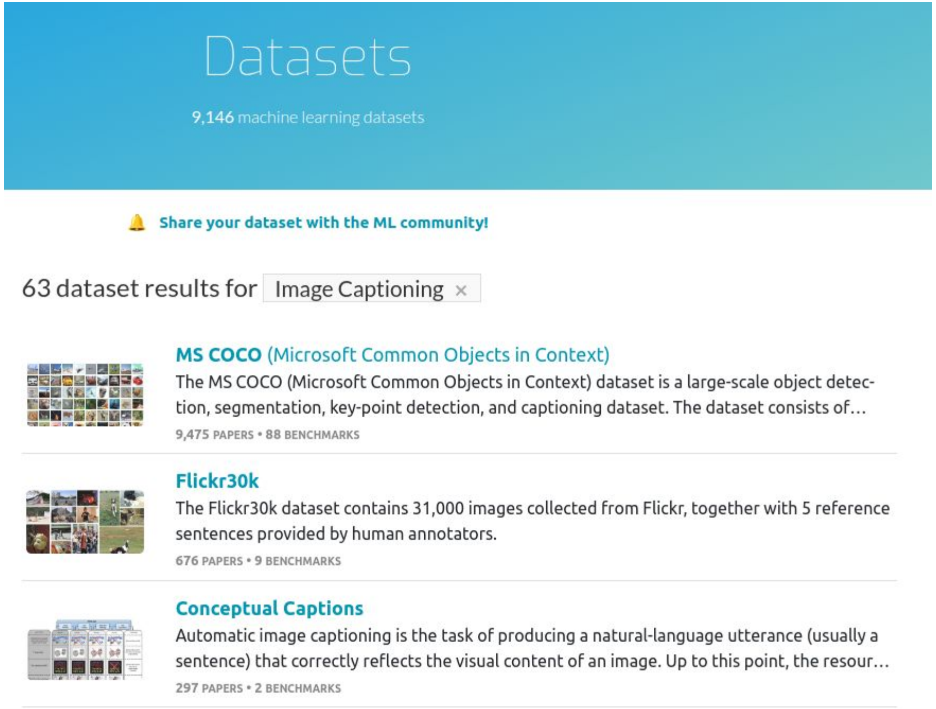
A Winnie The Pooh character high chair with a can of Yoohoo sitting on it in front of a white wall.

A cup holder in a car holding loose change from Canada.

https://vizwiz.org/tasks-and-datasets/image-captioning/

D’autres jeux de données en annotation d’images

https://paperswithcode.com/datasets?task=image-captioning



Datasets

9,146 machine learning datasets

Share your dataset with the ML community!

63 dataset results for Image Captioning ×

MS COCO (Microsoft Common Objects in Context)
The MS COCO (Microsoft Common Objects in Context) dataset is a large-scale object detection, segmentation, key-point detection, and captioning dataset. The dataset consists of…
9,475 PAPERS • 88 BENCHMARKS

Flickr30k
The Flickr30k dataset contains 31,000 images collected from Flickr, together with 5 reference sentences provided by human annotators.
676 PAPERS • 9 BENCHMARKS

Conceptual Captions
Automatic image captioning is the task of producing a natural-language utterance (usually a sentence) that correctly reflects the visual content of an image. Up to this point, the resour…
297 PAPERS • 2 BENCHMARKS

Captioning with ClipInterrogator

https://colab.research.google.com/github/pharmapsychotic/clip-interrogator/blob/main/clip_interrogator.ipynb

Captioning with MiniGPT-4

https://github.com/camenduru/MiniGPT-4-colab

Captioning with Qwen-VL

https://github.com/camenduru/Qwen-VL-Chat-colab/tree/main

Références

- Image midjourney https://mid-journey.ai/midjourney-v6-release/
- Image Stable Diffusion XL https://generativeai.pub/stable-diffusion-xl-is-here-whats-new-4e6ed27df70c
- Image Dall-E 3 (capture du site web https://openai.com/dall-e-3)
- Image Kandinsky v2.2 https://www.gadgetvoize.com/2023/07/12/sber-presents-its-neural-networks-new-version-kandinsky-2-2/