My Research Topic

related to natural language processing

2023-11-30

Hwang Seunghyeon | Augmented Intelligence Lab

Contents

- What is image capturing?
  - LG Captioning AI
  - Video captioning
- Real-time video captioning with MR devices
  - Zero-Shot Image Captioning
  - Research and implementation goals

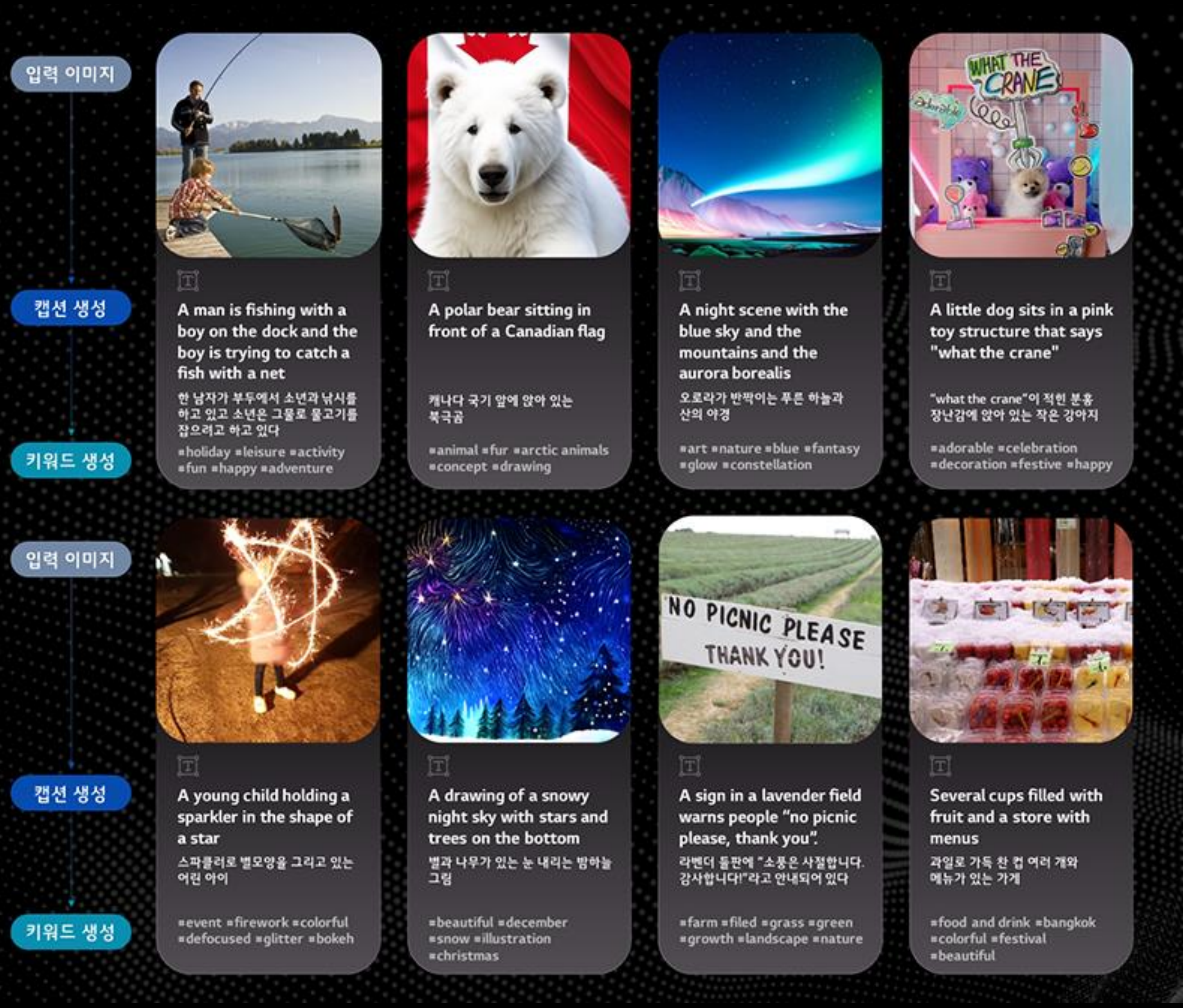
Image Captioning

Image to Text

Input: photos or drawings

Output: sentences or keywords

CVPR 2023 - LG Captioning AI



A man is fishing with a boy on the dock and the boy is trying to catch a fish with a net

한 남자가 부두에서 소년과 낚시를 하고 있고 소년은 그물로 물고기를 잡으려고 하고 있다

#holiday #leisure #activity #fun #happy #adventure

A polar bear sitting in front of a Canadian flag

캐나다 국기 앞에 앉아 있는 북극곰

#animal #fur #arctic animals #concept #drawing

A night scene with the blue sky and the mountains and the aurora borealis

오로라가 반짝이는 푸른 하늘과 산의 야경

#art #nature #blue #fantasy #glow #constellation

A little dog sits in a pink toy structure that says "what the crane"

"what the crane"이 적힌 분홍 장난감에 앉아 있는 작은 강아지

#adorable #celebration #decoration #festive #happy

A young child holding a sparkler in the shape of a star

스파클러로 별모양을 그리고 있는 어린 아이

#event #firework #colorful #defocused #glitter #bokeh

A drawing of a snowy night sky with stars and trees on the bottom

별과 나무가 있는 눈 내리는 밤하늘 그림

#beautiful #december #snow #illustration #christmas

A sign in a lavender field warns people "no picnic please, thank you".

라벤더 들판에 "소풍은 사절합니다. 감사합니다!"라고 안내되어 있다

#farm #filed #grass #green #growth #landscape #nature

Several cups filled with fruit and a store with menus

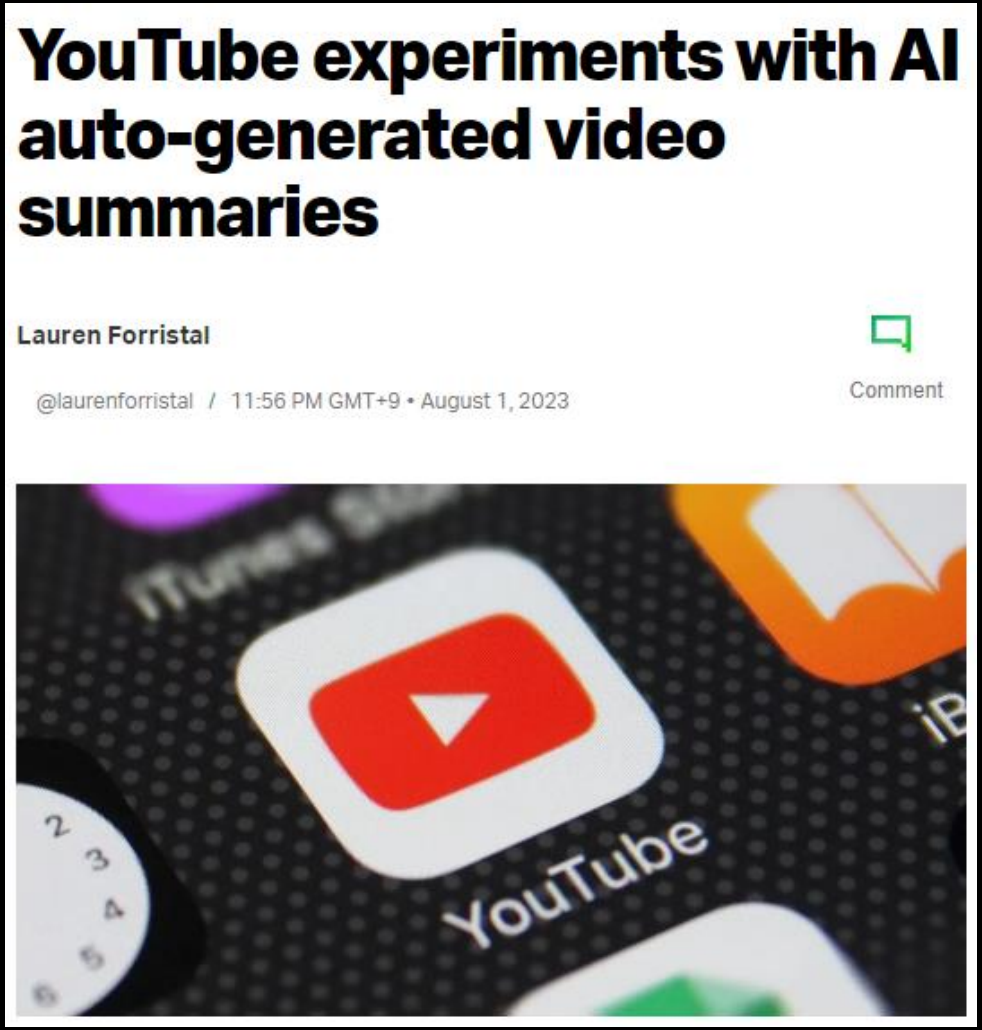
과일로 가득 찬 컵 여러 개와 메뉴가 있는 가게

#food and drink #bangkok #colorful #festival #beautiful

Video Captioning

Generate captions that match the words and actions spoken on the screen.

Designed to be used by people with hearing impairments.



YouTube experiments with AI auto-generated video summaries

Lauren Forristal

@laurenforristal / 11:56 PM GMT+9 • August 1, 2023

Comment

Zero-shot Image Captioning

ZeroCap

CVPR 2022

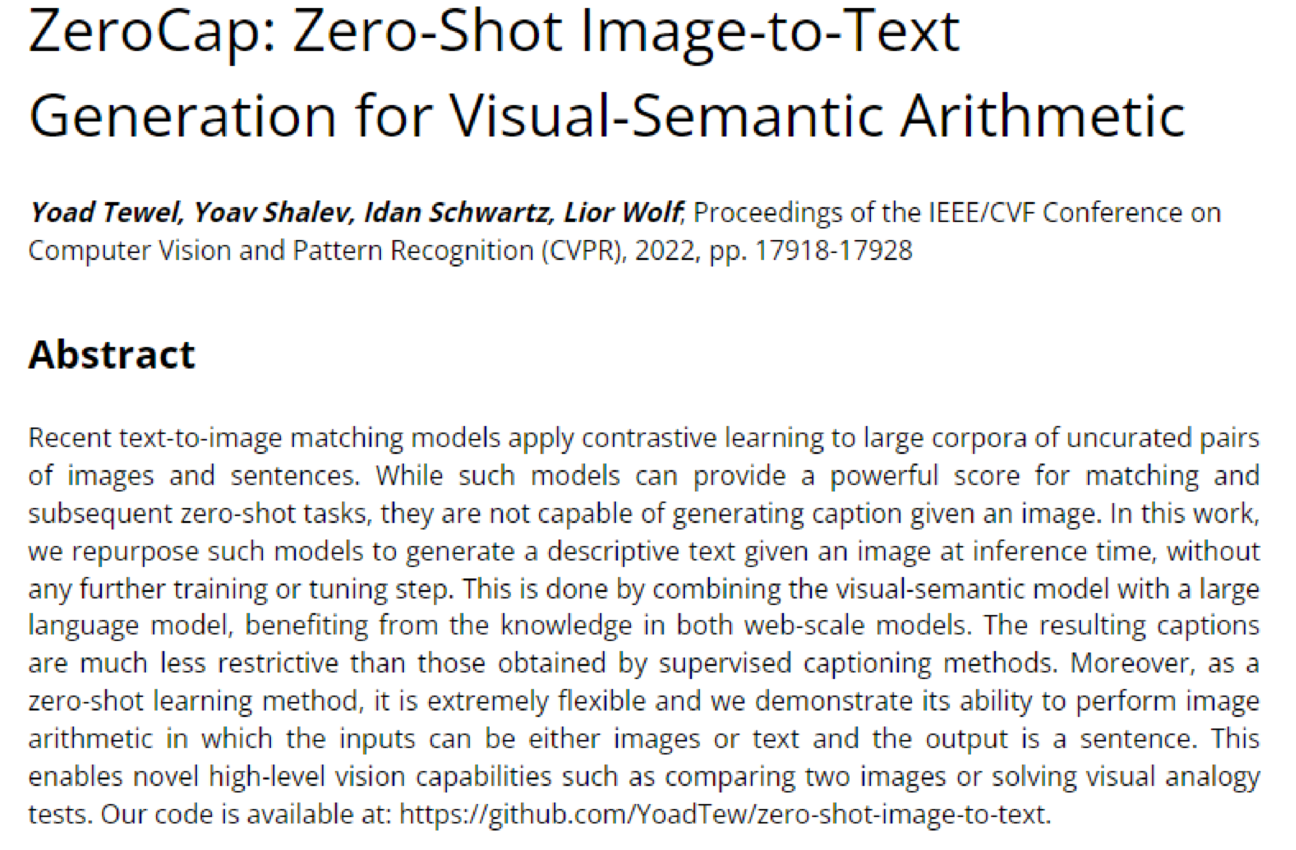
Yoad Tewel, Yoav Shalev,
Idan Schwartz, Lior Wolf

School of Computer Science,
Tel Aviv University

Using a visual-semantic model with a large language model to generate caption

visual-semantic model : CLIP

large language model : GPT-2



ZeroCap: Zero-Shot Image-to-Text Generation for Visual-Semantic Arithmetic

Yoad Tewel, Yoav Shalev, Idan Schwartz, Lior Wolf; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17918-17928

Abstract

Recent text-to-image matching models apply contrastive learning to large corpora of uncurated pairs of images and sentences. While such models can provide a powerful score for matching and subsequent zero-shot tasks, they are not capable of generating caption given an image. In this work, we repurpose such models to generate a descriptive text given an image at inference time, without any further training or tuning step. This is done by combining the visual-semantic model with a large language model, benefiting from the knowledge in both web-scale models. The resulting captions are much less restrictive than those obtained by supervised captioning methods. Moreover, as a zero-shot learning method, it is extremely flexible and we demonstrate its ability to perform image arithmetic in which the inputs can be either images or text and the output is a sentence. This enables novel high-level vision capabilities such as comparing two images or solving visual analogy tests. Our code is available at: https://github.com/YoadTew/zero-shot-image-to-text.

Zero-shot Image Captioning

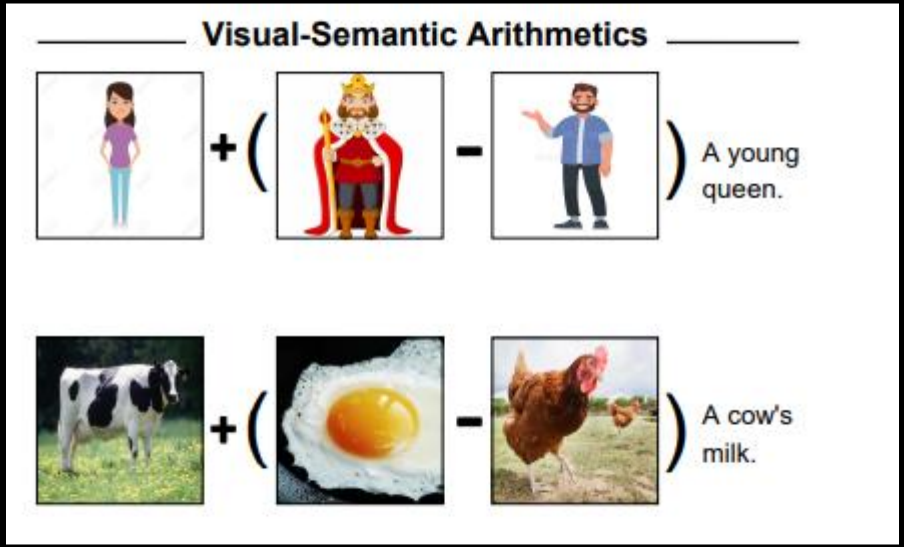
- AI makes inferences about objects it sees for the first time.
  - Based on pre-trained data.
  - Just like a human.
- Combining
  - visual-semantic model.
  - Large language model. (LLM)

LLM

- Inferring the word that follows the caption
  - Image of a ……
- Inference Methods
  - Generate words in the correct orientation for a given image
    - CLIP loss L_{CLIP}
  - Preserves language attributes
    - loss term L_{CE}

Visual-Semantic Arithmetic

- Word Embedding arithmetic
  - queen ='king'-'man'+ 'woman'
  - Embedding words as vectors
- Image Embedding arithmetic
  - Embedding Vectors in Image Space



Comparison

Traditional Captioning

- Pre-trained models
- Doesn't work well with new data

Zero-shot Captioning

- LLM
- No need to learn specific visuals

Real-time video captioning with MR devices

Apple Vision Pro applications for the deaf and hard of hearing

Research goals

- Enabling real-time video captioning with zero-shot captioning technology
- Combining **multiple** state-of-the-art
  - Visual-semantic model using Diffusion Models
  - GPT-4
- Analyze usability differences with existing products

Implementation goals

- Purpose
  - Real-time captioning in movie theaters, performance venues, etc.
  - Filling the gap that OTT cannot fill for the hearing impaired.
- Environment
  - Apple Vision Pro
  - Microsoft Holo Lens 2

Existing

TranscribeGlass™

Computer Assisted Real-time Translation (CART)

Existing

Imagen

Diffusion model + LLM(T5)

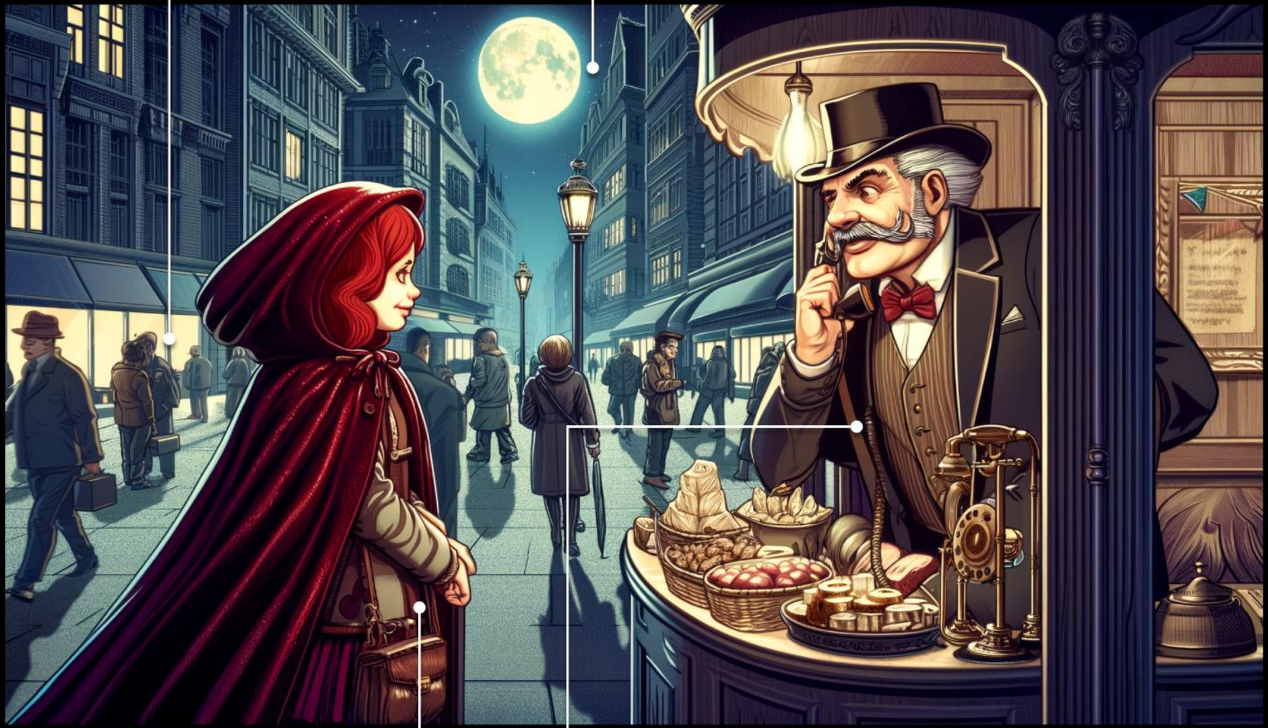
Text to image

Existing

DALL-E 3

Diffusion model + LLM(GPT-4)

Text to image



The sidewalks bustling with **pedestrians enjoying the nightlife.**

A bustling city street under the shine of a **full moon.**

At the corner stall, a **young woman** with fiery red hair, dressed in a signature velvet cloak, is **haggling with the grumpy old vendor.**

The grumpy vendor, a **tall, sophisticated man,** is wearing a sharp suit, sports a **noteworthy moustache** and is animatedly conversing on his **steampunk telephone.**