AstraZeneca

Google DeepMind

ICML
International Conference
On Machine Learning

An Image is Worth Multiple Words: Discovering Object Level Concepts using Multi-Concepts Prompts Learning

Chen Jin, Ryutaro Tanno,
Amrutha Saseendran, Tom Diethe, Philip Teare

Paper: https://arxiv.org/abs/2310.12274
Github: https://github.com/AstraZeneca/MCPL
Website: https://astrazeneca.github.io/mcpl.github.io/

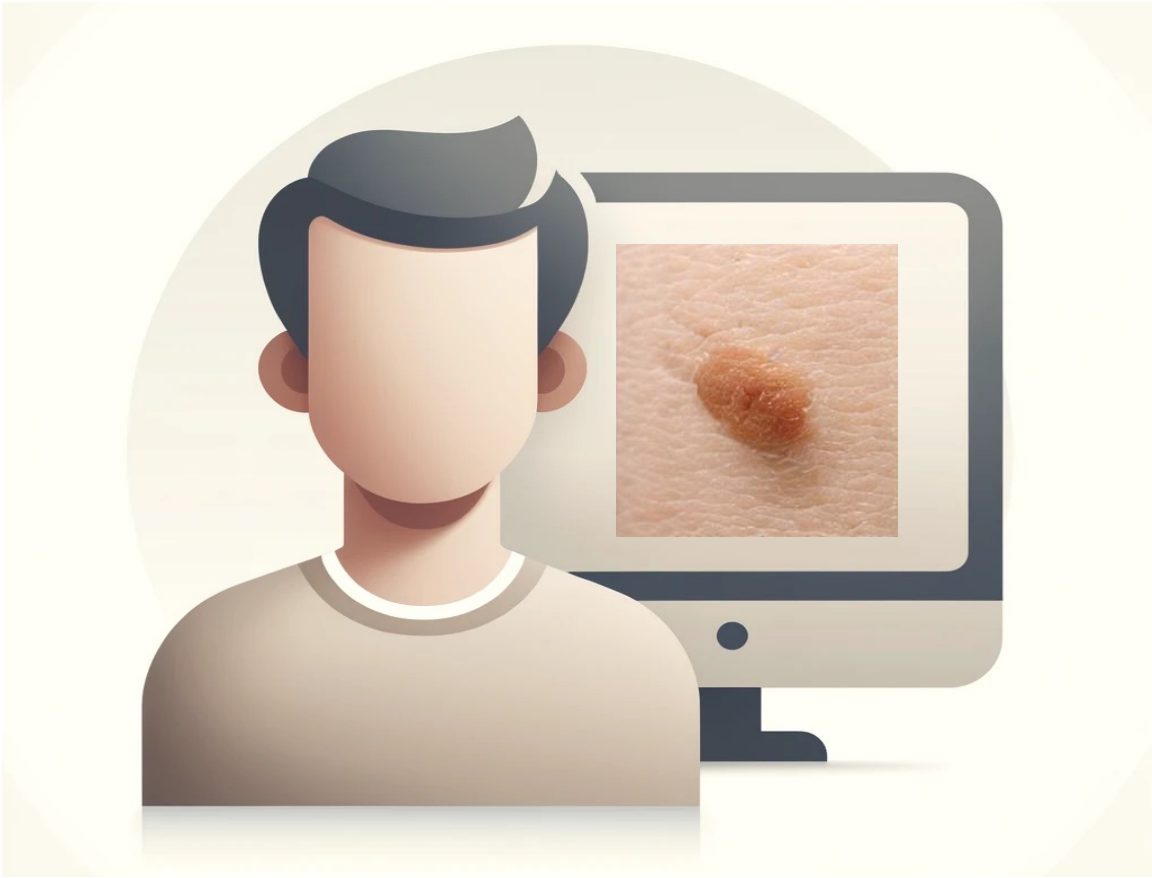
In nurseries, toddlers are shown pictures to learn new things.

Chen Jin

18/06/2024

Similarly, we explore teaching machines new concepts through natural language without requiring image annotations.

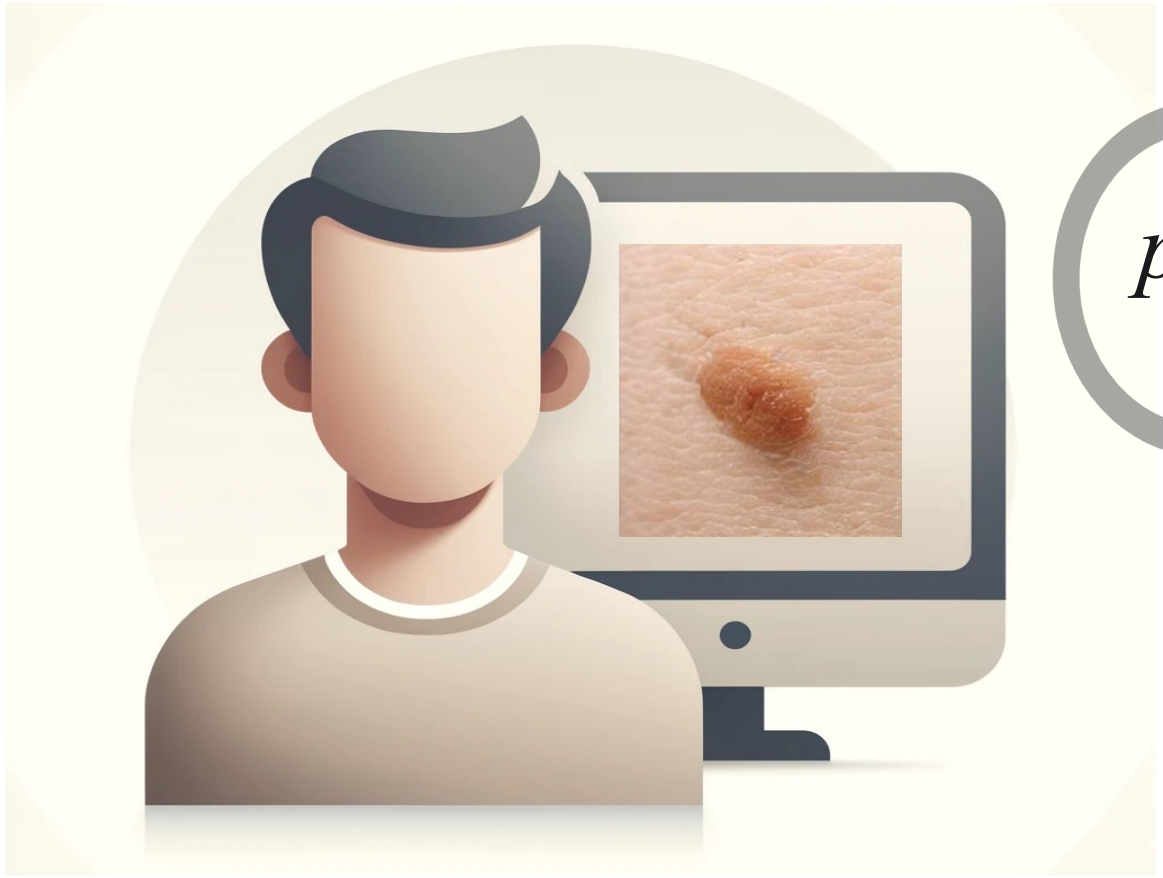
When humans see an image.



Chen Jin

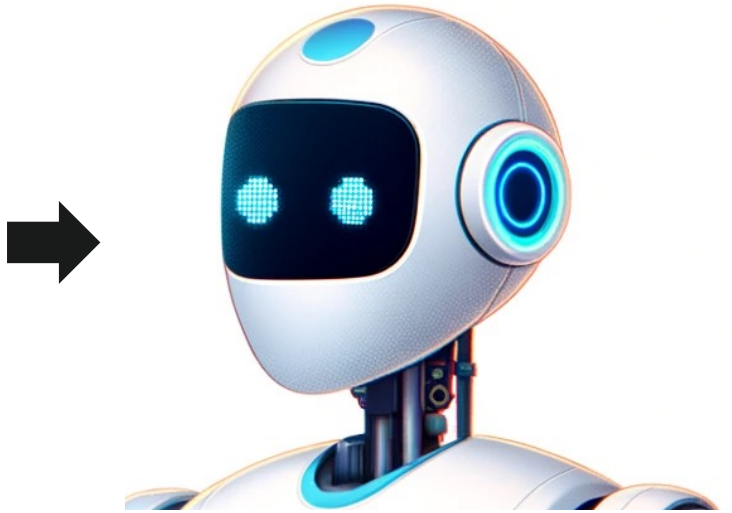
18/06/2024

The human may describes an image, leaving out multiple unfamiliar concepts.



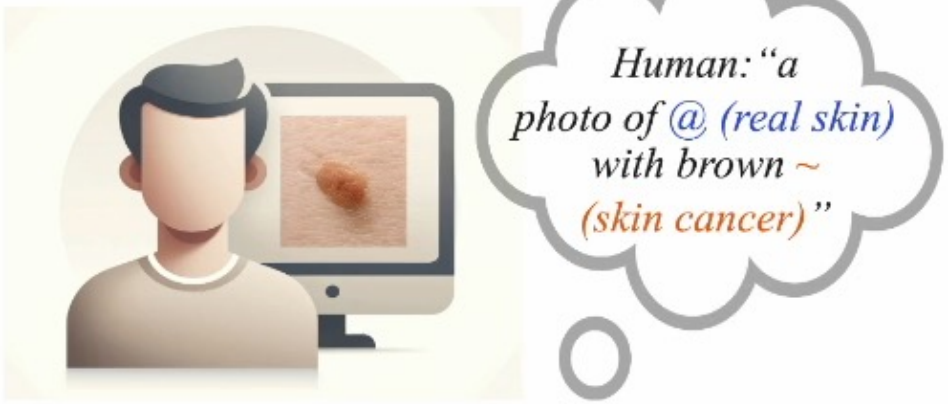
Human: "a photo of @ (real skin) with brown ~ (skin cancer)"

The machine then learns to
link each new concept with a learnable prompt (pseudo words)
from single sentence-image pair.



"a photo of @ (real skin) with brown ~ (skin cancer)"

Once learned we can explore hypothesis generation by only changing the weight of the new concept

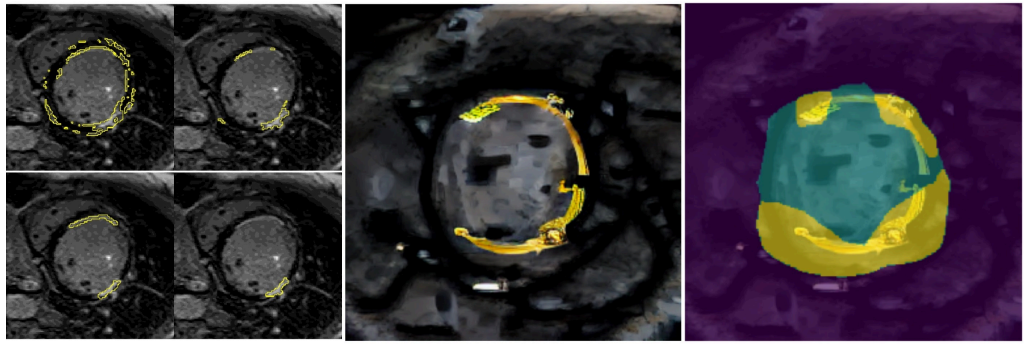


Human: "how skin cancer may develop?"

Check interactive demos on our website: https://astrazeneca.github.io/mcpl.github.io/

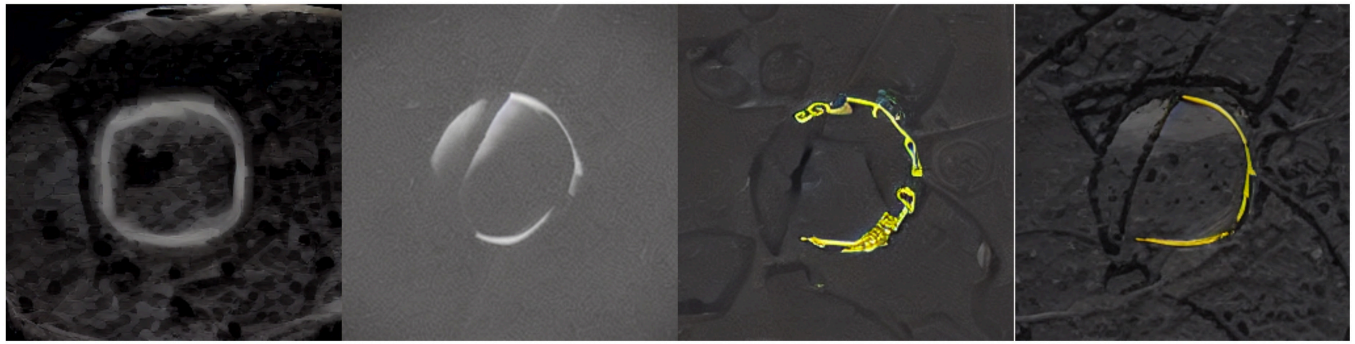
Discovering OOD Concepts from Medical Image and Disentangling

Reference image(s) Generated image Cross-attention mask

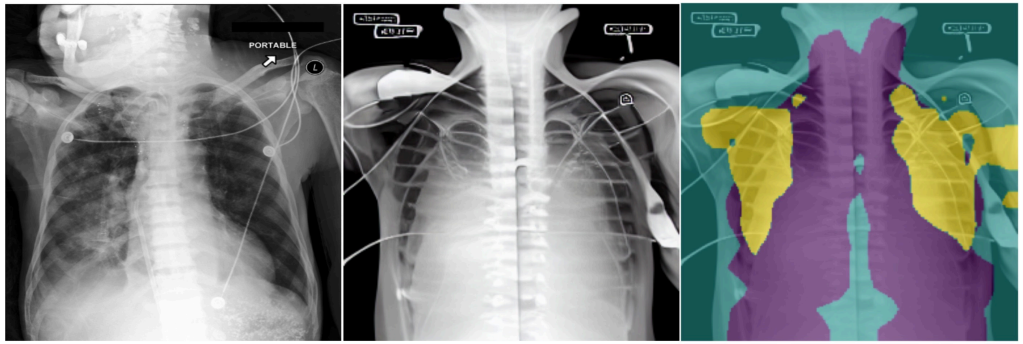


*"a photo of ! with round * and thin & on the side circled by yellow lines"*

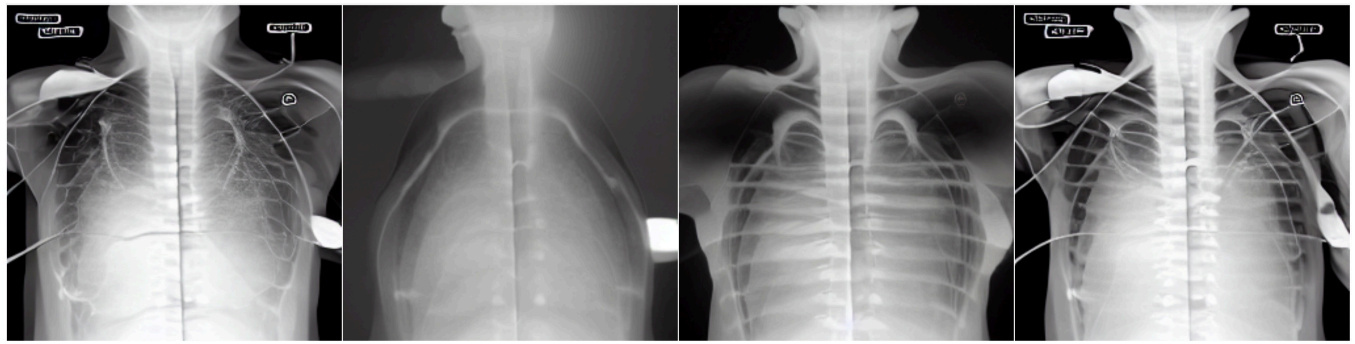
Generating or editing each disentangled concept



! round * thin & yellow lines



*"a photo of white ! (chest X-ray) and black @ (lung) which have smoky * (consolidation)"*



"white !" "smoky *" remove "white !" remove "smoky *"

Check interactive demos on our website: https://astrazeneca.github.io/mcpl.github.io/

Chen Jin

18/06/2024

How?

Knowledge mining with text-guided image generation model

Textural Inversion Gal et al. (2022)

Preprint: https://arxiv.org/abs/2310.12274

Chen Jin

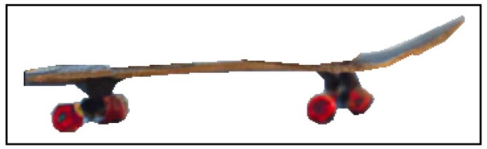
18/06/2024

Finer granular (word-object) learning and challenges

Textural Inversion

learning multiple concepts separately

*"a photo of * (teddybear) "*



"a photo of & (skateboard)"

composing multiple concepts in a single scene

*"a photo of brown * (teddybear) on a rolling & (skateboard) at times square"*

Ours versus other approaches

Crop-based

Cones / Custom Diffusion

*"a photo of * (teddybear) "*

"a photo of & (skateboard)"

Mask-based

Break-A-Scene

*"a photo of brown * (teddybear) on a rolling & (skateboard) at times square"*

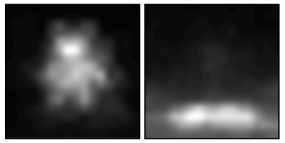
Mask-Free learning and editing multiple new concepts with MCPL (ours)

Discovering multiple concepts from single image

*"a photo of brown * (teddybear) on a rolling & (skateboard) at times square"*

Recomposing concepts with aligned cross-attention masks

generated image



** cross-attention &*

Editing each disentangled concept

~~*(teddybear)~~ *panda*

~~*(teddybear)~~ *cat*

~~rolling & (skateboard)~~ *surfing board*

~~rolling & (skateboard)~~ *flying blanket*

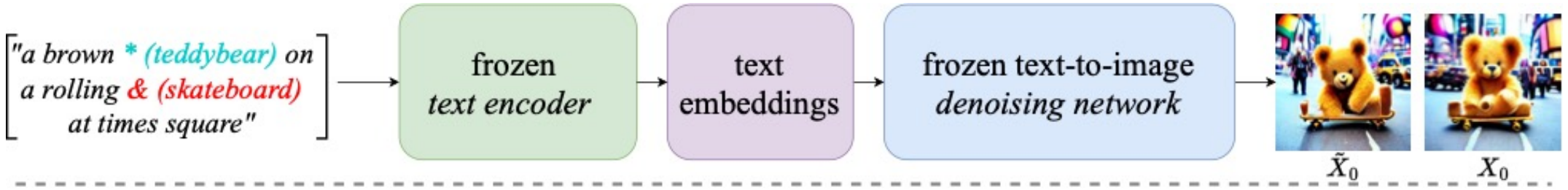
learnable prompts are represented as coloured pseudo words

Website: https://astrazeneca.github.io/mcpl.github.io/ Github: https://github.com/AstraZeneca/MCPL Preprint: https://arxiv.org/abs/2310.12274

Chen Jin

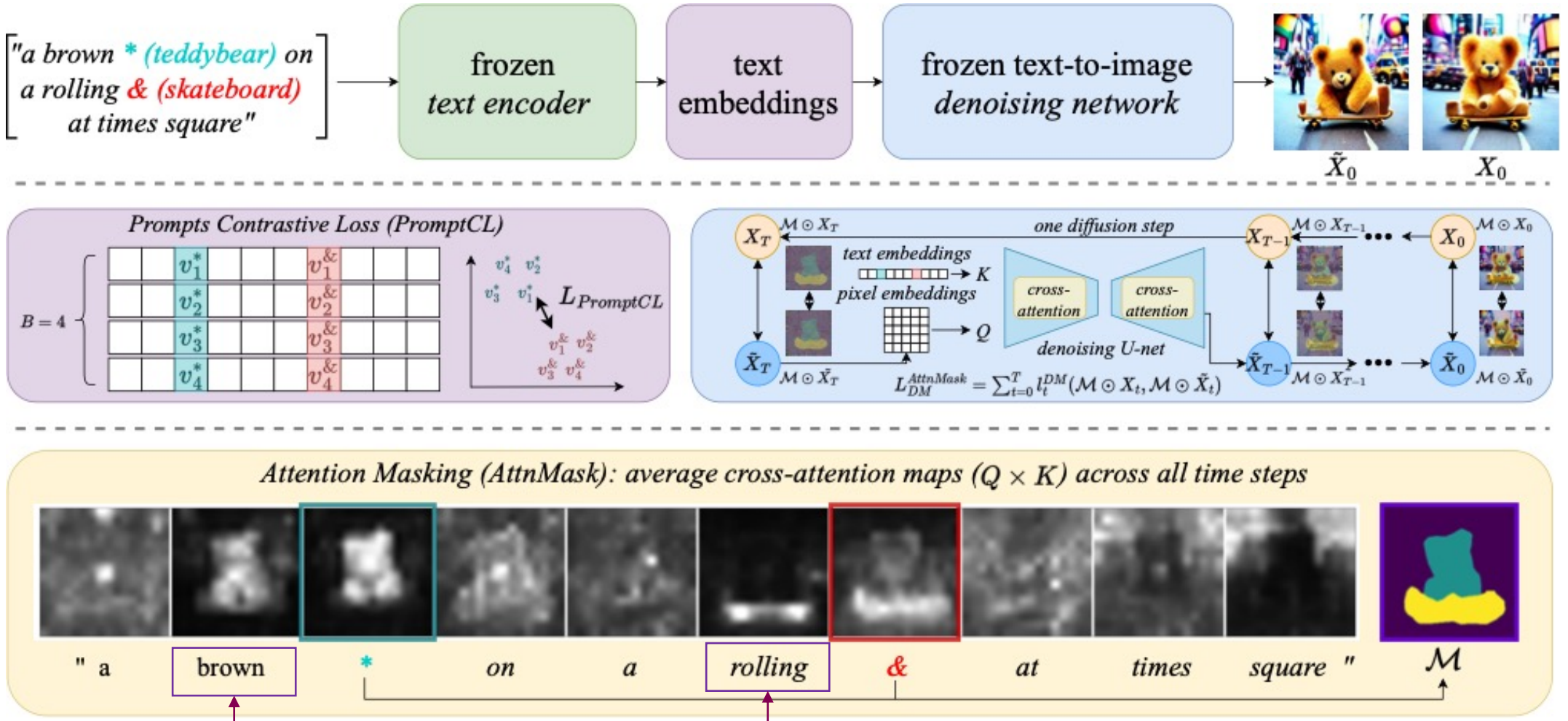
18/06/2024

Our solution: Multi-Concepts Prompts Learning



Chen Jin

Our solution: Multi-Concepts Prompts Learning



Bind adjective (Bind adj.) to associate new "words" with known words;

Evaluation – Multi-Concept-Dataset

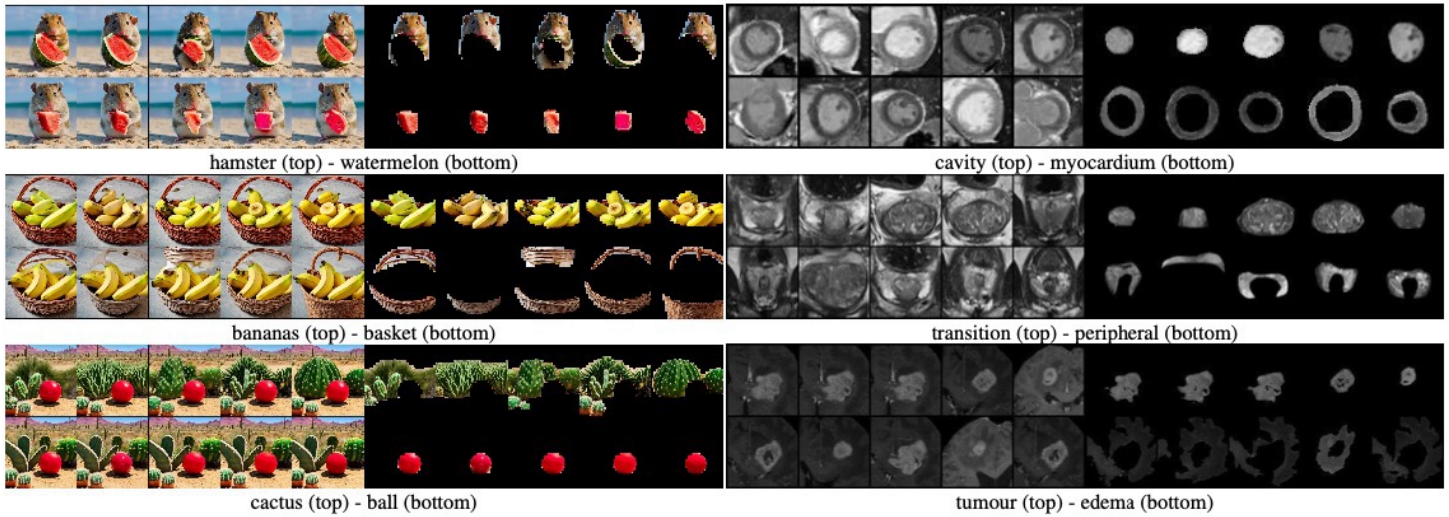


Figure 43. **Evaluation dataset (two concepts).** We prepared five sets of in-distribution natural images and three sets of out-of-distribution biomedical images, each containing two concepts resulting in a total of 16 concepts.

Figure 15. Visual examples of real images containing various number of objects.

Figure 16. Challenge segmentation examples. To streamline segmentation tasks, we utilize the MaskFormer model (Cheng et al., 2021) trained on the COCO (Lin et al., 2014) and ADE20K (Zhou et al., 2017) datasets, along with the Segment Anything (SAM) model (Ma & Wang, 2023). We observe that MaskFormer occasionally under-segments, whereas SAM is prone to over-segmentation. Consequently, manual adjustments are necessary to ensure the datasets are accurately prepared for evaluation.

Figure 44. **Evaluation dataset (three to five concepts).** We generate nine sets containing 9 more object-level concepts.

Dataset: https://github.com/AstraZeneca/MCPL/tree/master/dataset

Prompt and image fidelity: embedding similarity comparing to the estimated "ground truth"

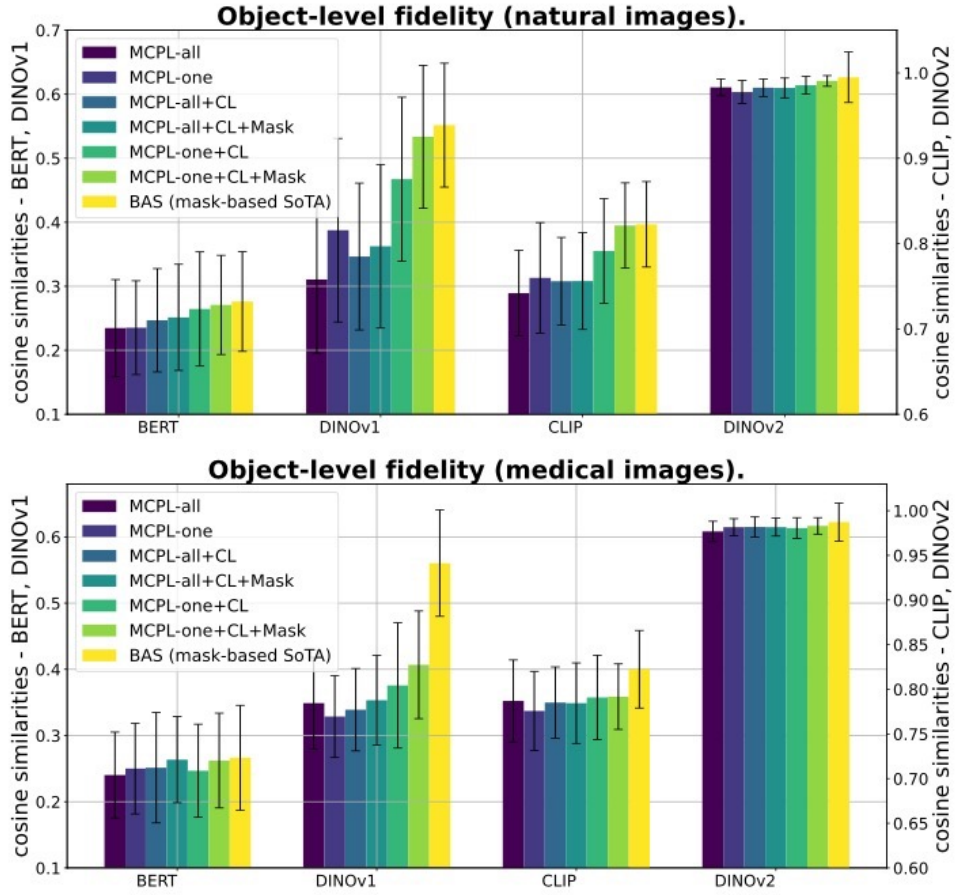


Figure 6. **Embedding similarity in learned object-level concepts compared to masked "ground truth" (two concepts per image).**

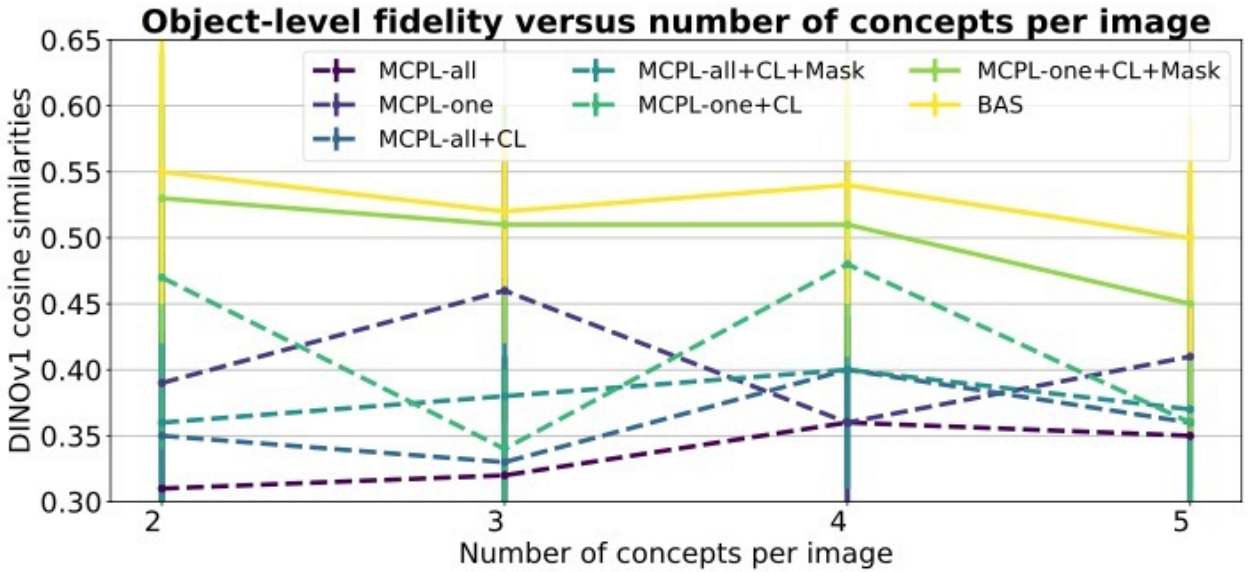


Figure 8. **Evaluate the learning as the number of concepts per image increases.** Here each data point represents an *average of 20∼40k* pairwise cosine similarities measured by DINOv1.

Preprint: https://arxiv.org/abs/2310.12274

More results can be found on our project page

https://astrazeneca.github.io/mcpl.github.io/