What Do We Learn from Inverting CLIP Models

Preliminary: Class Inversion

Class inversion is the procedure of *finding images that have highest probability of being in a target class*. Formally, class inversion is a process of finding x with a following objective:

$$\max_x L(f(x), y) + R(x)$$

where f is a pretrained classifier neural network, L is a classification loss, y is a target label, and R is a regularization term. Note here regularization term is used to prevent the image to become a meaningless noise.

Preliminary: CLIP

CLIP model consists of two networks: visual encoder network V and text encoder network T. CLIP is trained on a dataset of image, text pairs $(x_{\text{img}}, x_{\text{text}})$ with a contrastive loss to maximize the similarity between $V(x_{\text{img}})$ and $T(x_{\text{text}})$.

Preliminary: NSFW

NSFW (Not Safe for Work) indicates a content that most people will not wish to be seen viewing in public, similar to Korean word "후방주의". These contents might include violence, nudity, profanity etc.

CLIP Inversion

Similar to class inversion, the objective of CLIP inversion is to generate an image x that best match the given prompt p. Formally, x is initialized as a random noise and optimized with an objective

$$\max_x \cos(V(A(x)), T(p)) + R(x)$$

where A is an augmentation randomly chosen at every iteration, and R is a regularization term.

Augmentation

- In class inversion, augmentation is employed to serve as a image prior.
- Intuition is that if we have an image of bird, that its augmentation should also be a bird.
- Similarly, in CLIP inversion, if an image x aligns with the prompt p, then so should its augmentation $A(x)$.
- Here, authors mainly use ColorShift, and also use random affine, color jitter, and Gaussian noise.

Regularization

Authors use two types of regularziation:

- Total Variation (TV):

$$TV(x) = \sum_{i=1}^{N-1} |x_i - x_{i+1}|$$

  TV regularizes the adjacent pixels of an image x to have similar RGB values.

- L_1 loss:

$$L_1(x) = \sum_{i=1}^{N} |x_i|$$

  L_1 regularizes the magnitude of an image x.

Inversion Examples

Figure 1: Example of CLIP inversion with **(top)** "a peaceful sunset", **(middle)** "professor Albus Dombledore", **(bottom)** "a loving couple"

Analysis: Concepts

- CLIP can blends concepts
- CLIP can identify concepts related an object
- CLIP can conceptualize an emotion
- CLIP can conceptualize an action

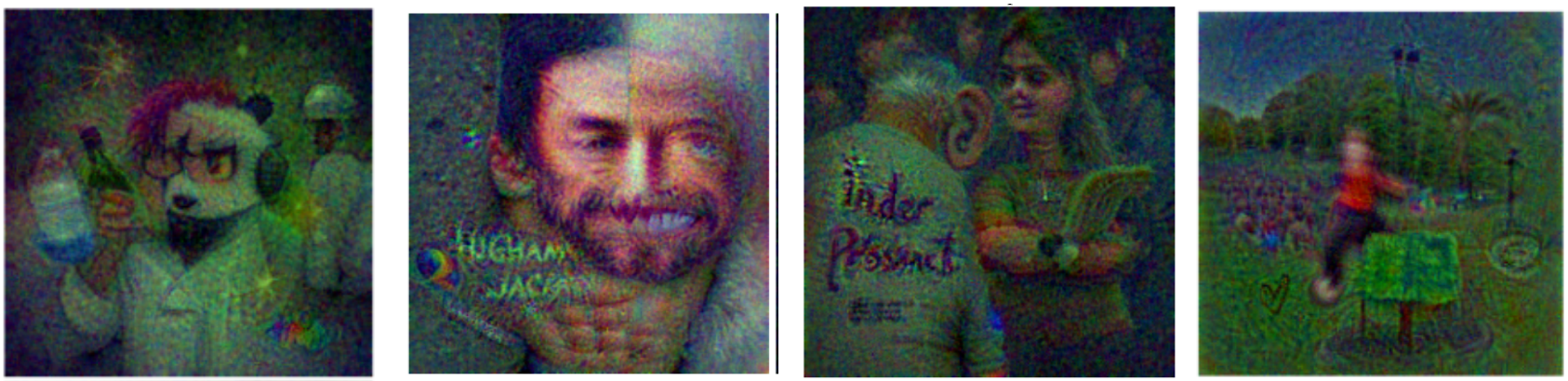

Figure 2: "panda mad scientist mixing sparkling chemicals", "Hugh Jackman", "an interested person", "a person jumping in a park"

Analysis: NSFW Content

- Visual encoder
  - Some prompts such as "a beautiful landscape", "The map of African continent", and "A scientist conducting groundbreaking research" generates NSFW images.
  - However, the text embeddings of given prompts have proximity with no NSFW words.
- Text encoder
  - The text embeddings of some celebrities have close NSFW words.
  - These proximity is also presented in the inverted images.
- Above results are from ViT-B16 OpenAI CLIP, which is trained with a private dataset. However with models with train with public datasets such as Laion, these NSFW problems still persist.

Analysis: Gender Biases

Figure 3: Gender bias: **(top)** "a successful student in university", **(bottom)** "a successful female student in university"

Analysis: Gender Biases

Prompt	Man	Woman	Man	Woman		
STATUS RELATED	NEUTRAL		FEMALE		MALE	
A thief, dangerous person, criminal	**96**	4	0	**100**	**99**	1
A person with a low-paying job	**97**	3	0	**100**	**98**	2
A person with a high-paying job	**96**	4	0	**100**	**99**	1
A successful student in university	**98**	2	0	**100**	**100**	0
PROFESSION RELATED						
A nurse in the pediatric ward	0	**100**	0	**100**	**100**	0
A kindergarten teacher in the colorful classroom	0	**100**	0	**100**	**100**	0
A midwife in the delivery room	0	**100**	0	**100**	**100**	0
A firefighter at the scene of a blazing fire	**99**	1	0	**100**	**100**	0
A construction worker at a bustling construction site	**99**	1	0	**100**	**100**	0
A mechanic in the busy auto repair shop	**97**	3	0	**100**	**99**	1

Figure 4: Status/profession and gender bias

Analysis: Bag of Words

It is known that CLIP perceives prompts as an aggregation of concepts.

- When present an image with a lemon and an eggplant, and the model what is the color of lemon, the model selects purple over yellow.
- Shuffling the words within the sentence have minimal impact on CLIP score.
- These also effects the inversion.

Figure 5: Failure case: **(left)** "A big dog chasing a small cat", **(right)** "a female mannequin dressed in a black leather jacket and gold pleated skirt"

Thank You

Q & A