NEURAL INFORMATION PROCESSING SYSTEMS

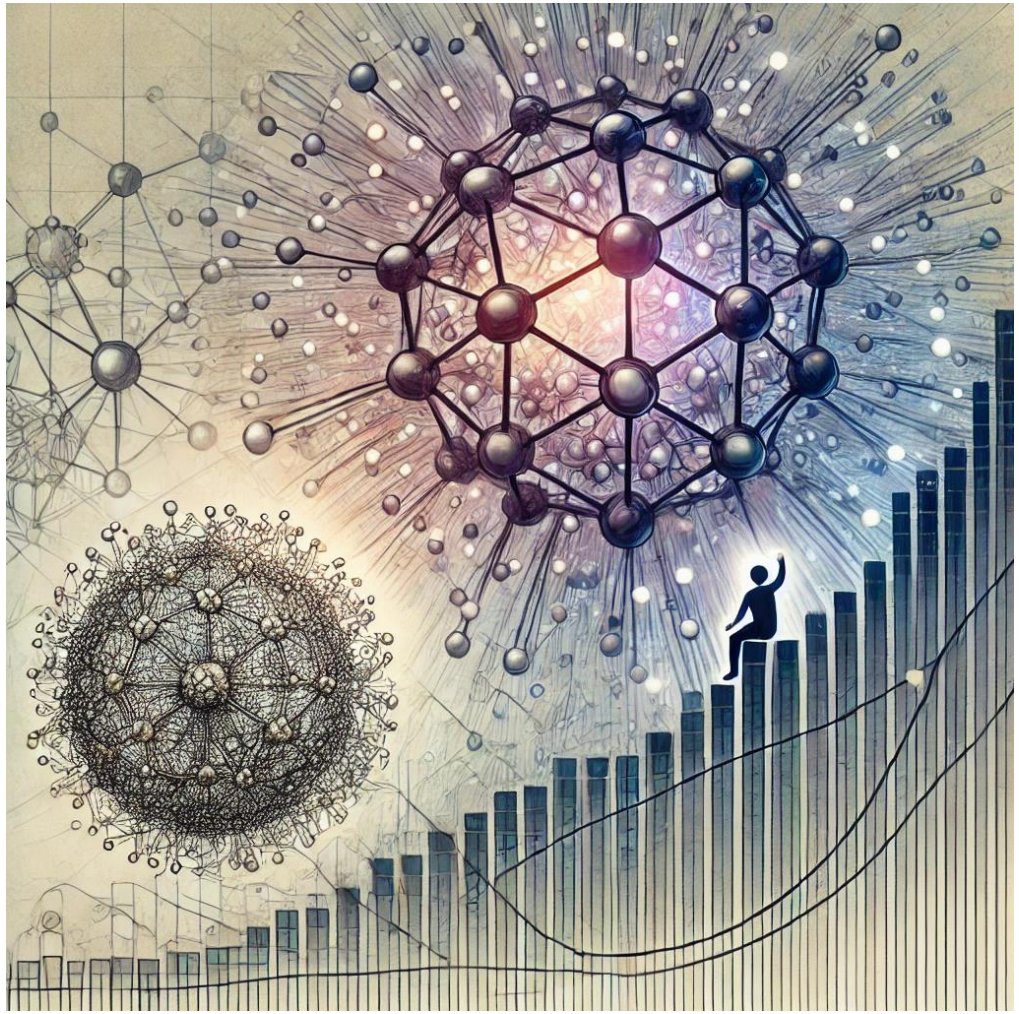
Presentation at NeurIPS 2024

Textual Training for the Hassle-Free Removal of Unwanted Visual Data

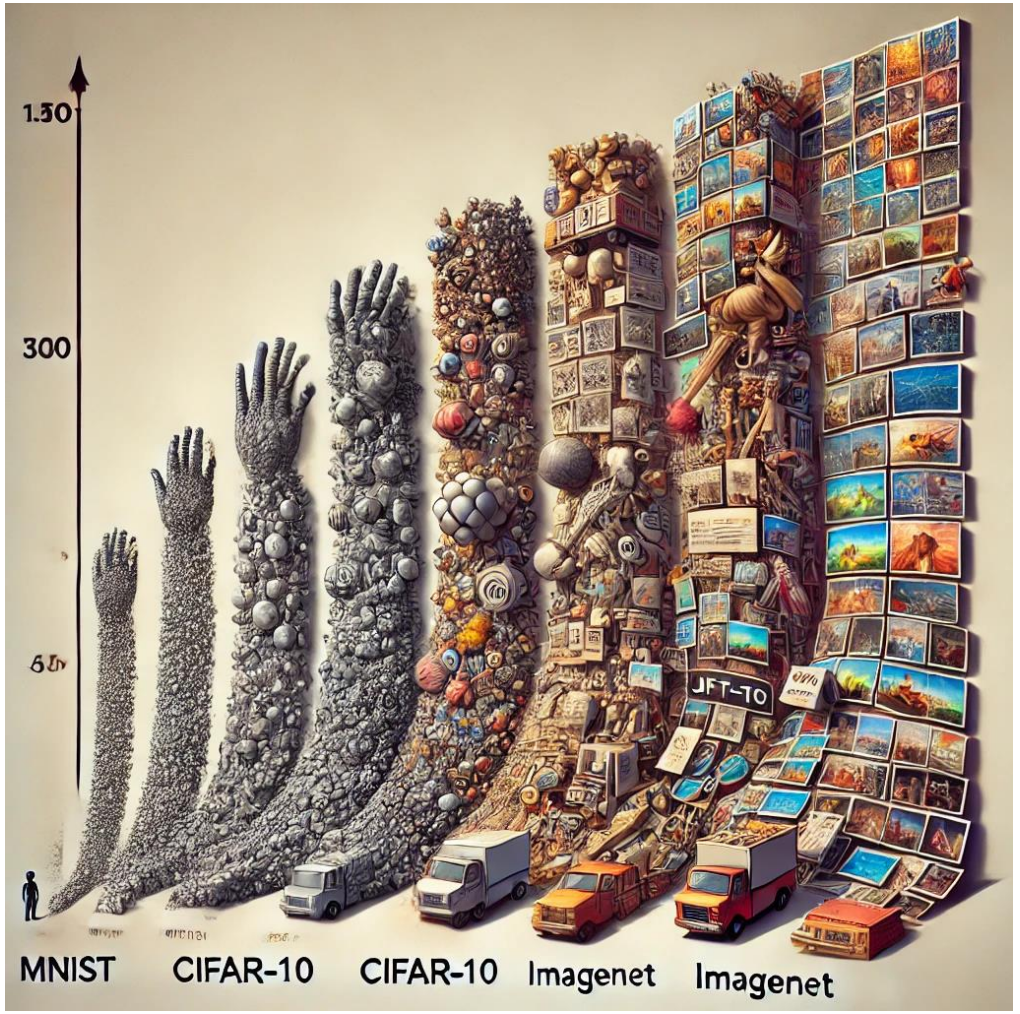
Saehyung Lee, Jisoo Mok, Sangha Park, Yongho Shin, Dahuin Jung, Sungroh Yoon

VERI TAS LUX MEA
DSAIL
Data Science & AI Laboratory

We are currently in the midst of what is known as the large-scale AI era.



Model size ↑



Dataset size ↑

generated by DALLE 3

What do the following two datasets have in common? Large-scale? Web crawling?

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE

80 million tiny images: a large dataset for non-parametric object and scene recognition

Antonio Torralba, Rob Fergus and William T. Freeman

LAION-400-MILLION OPEN DATASET

by: Christoph Schuhmann, 20 Aug, 2021

We present LAION-400M: 400M English (image, text) pairs - see also our Data Centric AI NeurIPS Workshop 2021 paper

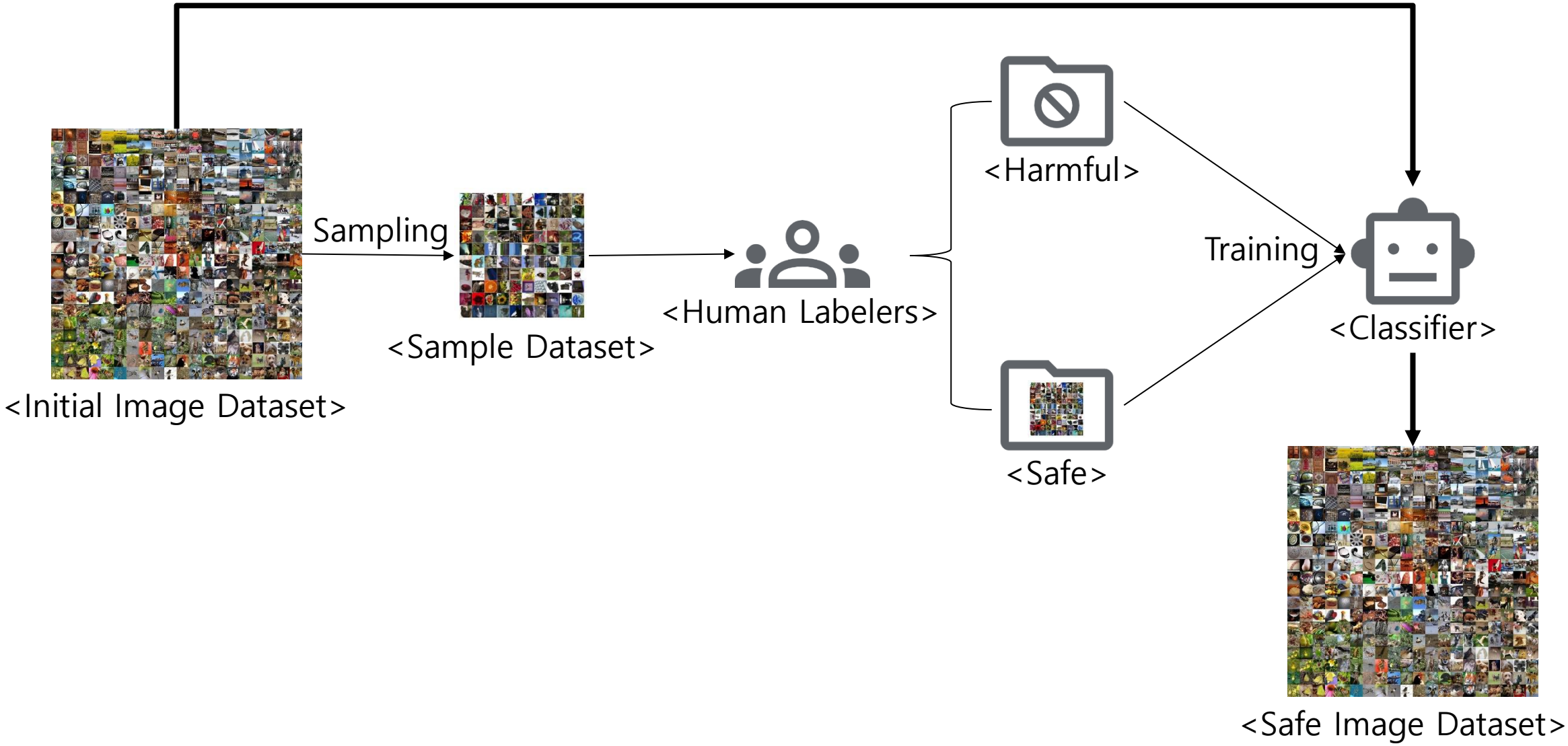
What do the following two datasets have in common?
Large-scale? Web crawling?

80 million tiny images: a large dataset for
non-parametric object and scene recognition

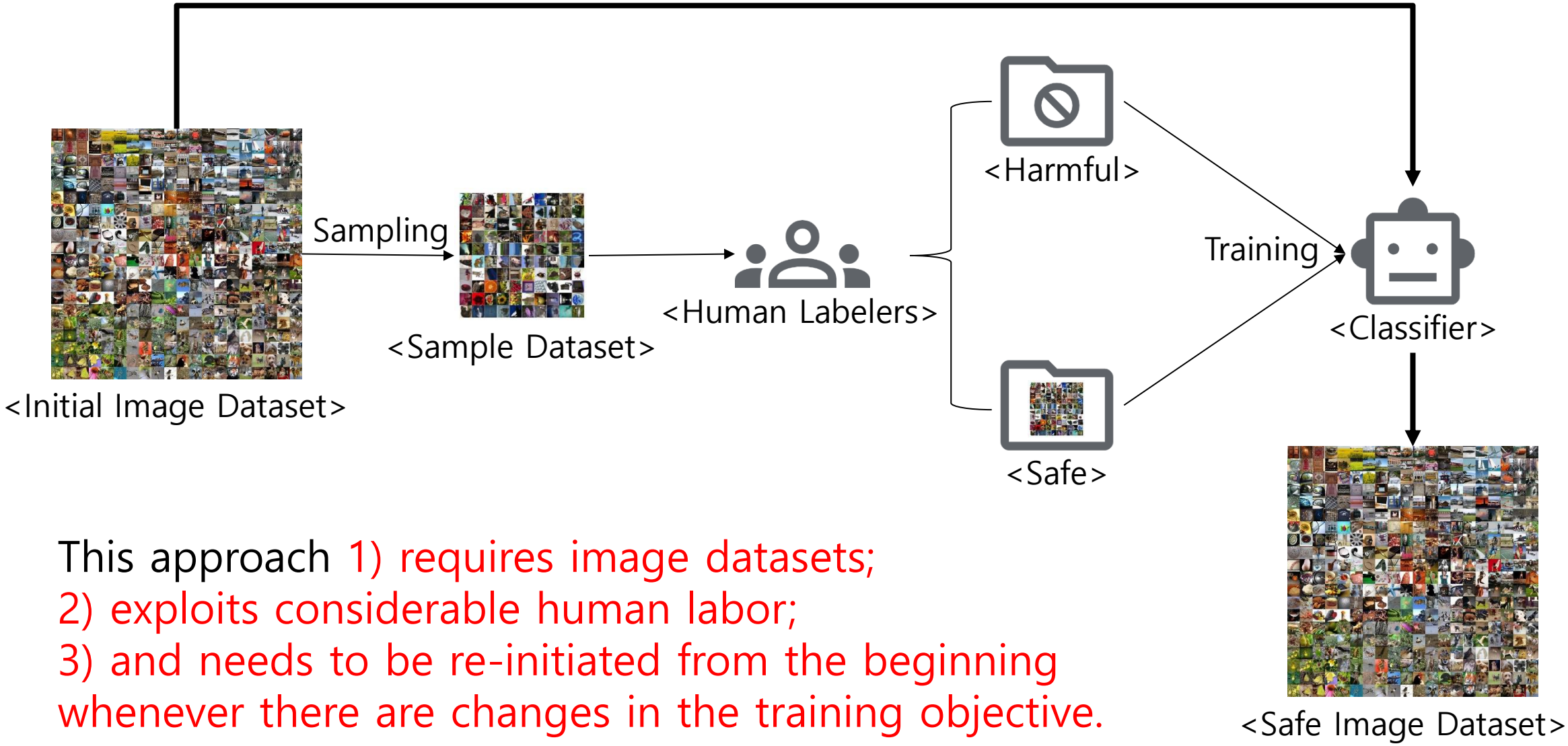
They contain harmful content.

Birhane et al. "Into the laion's den: Investigating hate in multimodal datasets." NeurIPS 2023.
Birhane, Abeba, and Vinay Uday Prabhu. "Large image datasets: A pyrrhic win for computer vision?." WACV 2021.

How can we filter out harmful content lurking in our large image dataset?

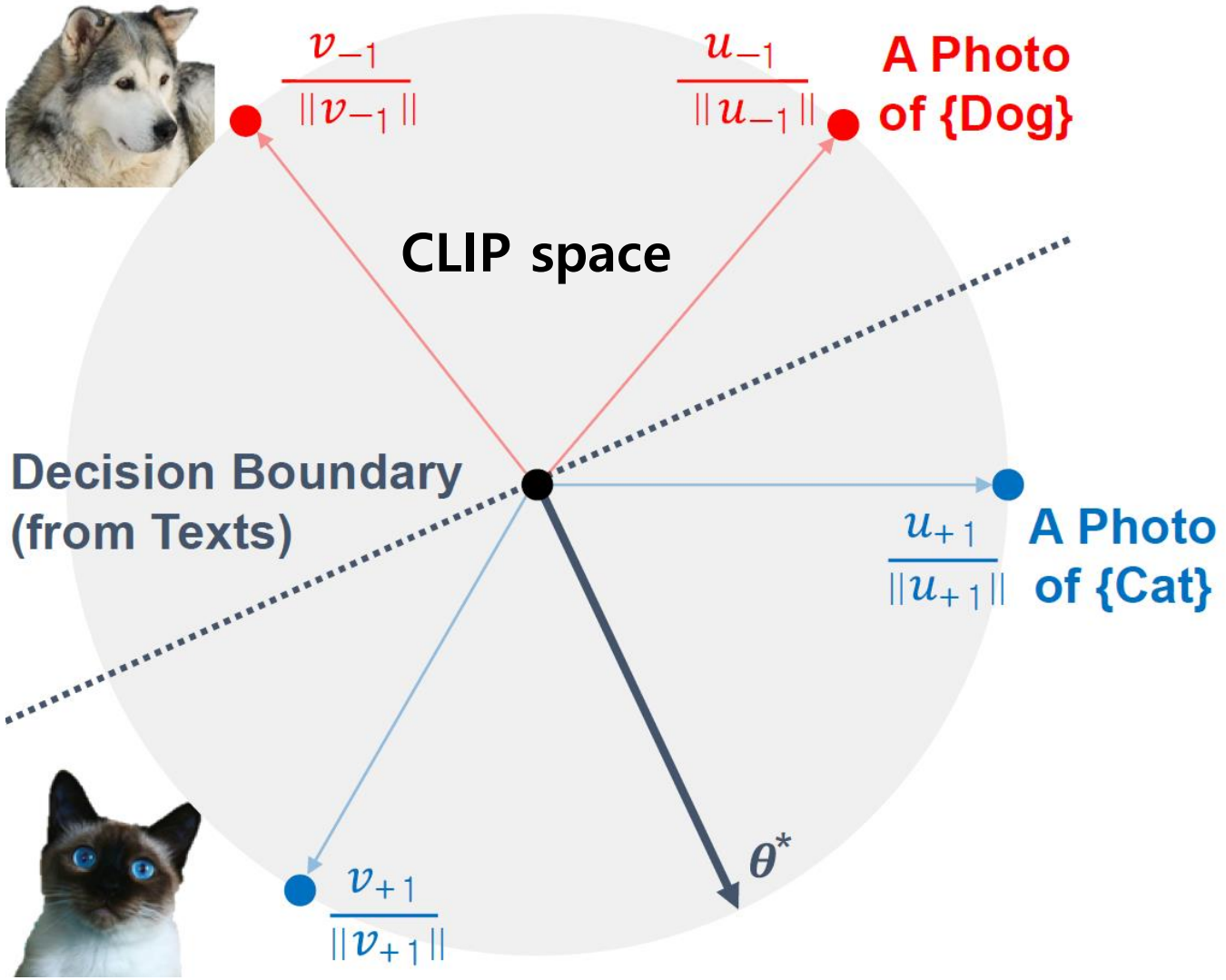


How can we filter out harmful content lurking in our large image dataset?

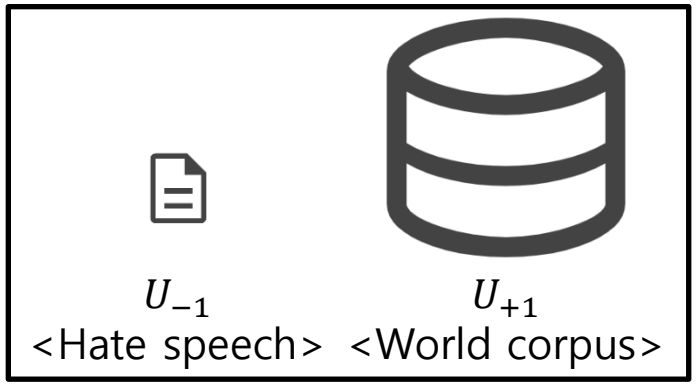


This approach 1) requires image datasets;
2) exploits considerable human labor;
3) and needs to be re-initiated from the beginning whenever there are changes in the training objective.

IDEA 1: classifiers obtained solely using textual data can operate on visual data as well!



IDEA 2: No need to divide the world corpus!



- The original loss function

$$\sum_{u \in U_{-1}} \lambda L\left(u, -1\right) + \sum_{u \in U_{+1}} \left(1 - \lambda\right) L\left(u, +1\right), \lambda \in [0, 1]$$

- We reformulate the original loss function as

$$= \sum_{u \in U_{-1}} L\left(u, -1\right) - \sum_{u \in U_{-1}} \left(1 - \lambda\right) L\left(u, -1\right) + \sum_{u \in U_{+1}} \left(1 - \lambda\right) L\left(u, +1\right).$$

- We propose changing it to

$$\sum_{u \in U_{-1}} L\left(u, -1\right) + \sum_{u \in U_{-1}} \left(1 - \lambda\right) L\left(u, +1\right) + \sum_{u \in U_{+1}} \left(1 - \lambda\right) L\left(u, +1\right)$$

$$= \sum_{u \in U_{-1}} L\left(u, -1\right) + \sum_{u \in U_{-1} \cup U_{+1}} \left(1 - \lambda\right) L\left(u, +1\right).$$

The union of the two sets $U_{+1} \cup U_{-1}$ allows us to train unwanted data detectors without needing to strictly differentiate between desired and unwanted data within the world corpus, providing a solution to the complexities involved in dataset construction

Our proposed loss function

- We incorporate the concept of focal loss into our loss function to further encourage the formation of the decision boundary near the in-distribution.

- Our proposed loss can thus be defined as:

Definition 1. Let $B_{-1} = \{x_i\}_{i=1}^N$ and $B = \{\tilde{x}_i\}_{i=1}^N$ denote mini-batches that are respectively drawn from the specified in-distribution and the overall data distribution. Let L be the cross-entropy loss. Then our proposed loss function is

$$\sum_{x_i \in B_{-1}} L\left(x_i, -1\right) + (1-\lambda) \sum_{x_j \in B} \beta_j L\left(x_j, +1\right); \;\; \beta_j = \frac{N\alpha_j}{\sum_{x_k \in B} \alpha_k} \text{ and } \alpha_j = \left(1 - p\left(x_j\right)\right)^{\gamma}.$$

- $\gamma \geq 0$ is a hyper-parameter of the focal loss.

Hassle-Free Textual Training (HFTT)

(a) Training Phase

$\mathcal{L}_{CE}$

$$P(Y=1|X) = \frac{\sum \exp(CosSim(\leftrightarrow))}{\sum \exp(CosSim(\leftrightarrow)) + \sum \exp(CosSim(\leftrightarrow))}$$

Task Embeddings

Dog ... Cat Bus

Trainable Embeddings

Measure Cosine-similarity

Text Encoder

$X_T \sim Text\ Corpus$

(b) Test Phase

Thresholding

$$P(Y=1|X) = \frac{\sum \exp(CosSim(\leftrightarrow))}{\sum \exp(CosSim(\leftrightarrow)) + \sum \exp(CosSim(\leftrightarrow))}$$

Task Embeddings

Dog ... Cat Bus

Trainable Embeddings

Measure Cosine-similarity

Image Encoder

$X_I\ (Images)$

Application 1: Out-of-Distribution Detection (in-distribution=ImageNet)

Method	OOD Dataset								Average	
	iNaturalist		SUN		Places		Texture			
	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC
In-distribution images required										
Mahalanobis	99.33	55.89	99.41	59.94	98.54	65.96	98.46	64.23	98.94	61.51
MSP	40.17	89.76	63.99	79.40	63.50	80.19	67.01	79.33	58.67	82.17
KNN	29.17	94.52	35.62	92.67	**39.61**	**91.02**	64.35	85.67	42.19	90.97
NPOS	**16.58**	**96.19**	43.77	90.44	45.27	89.44	46.12	**88.80**	37.94	91.22
In-distribution images not required										
Energy	34.70	90.55	32.33	90.58	40.29	89.32	51.24	72.36	39.64	85.70
MaxLogit	35.03	89.46	32.86	90.33	41.15	89.60	68.17	75.63	44.30	86.26
ZOC	87.30	86.09	81.51	81.20	73.06	83.39	98.90	76.46	85.19	81.79
MCM	34.33	91.36	32.27	91.86	47.48	88.68	50.90	87.52	41.25	89.86
HFTT (ours)	27.44	93.27	**19.24**	**95.28**	43.54	90.26	**43.08**	88.23	**33.33**	**91.76**

Application 2: Hateful Image Detection (in-distribution=Antisemitic/Islamophobic)

Method	Innocuous Dataset										Average	
	iNaturalist		SUN		Places		Texture		NINCO			
	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC
Energy	12.30	97.53	3.88	98.51	13.87	97.40	39.70	94.98	26.89	96.12	17.43	97.10
MaxLogit	23.65	96.89	18.49	97.49	27.84	96.48	33.33	96.00	33.03	95.99	25.82	96.71
ZOC	87.76	71.05	66.51	85.23	69.96	82.57	65.48	83.22	81.06	78.36	74.15	84.09
MCM	80.53	76.70	87.54	69.38	81.37	74.12	60.39	84.97	81.95	78.00	77.45	76.29
CLIPN	47.71	92.78	36.36	95.16	40.62	94.52	53.36	92.36	68.40	89.58	49.29	92.88
NegLabel	**0.03**	**99.84**	1.09	99.10	5.16	98.50	3.56	98.82	12.62	97.86	4.49	98.82
HFTT (ours)	0.17	99.44	**1.05**	**99.13**	**4.38**	**98.60**	**1.73**	**99.08**	**4.18**	**98.52**	**1.83**	**99.06**

[1] Felipe González-Pizarro and Savvas Zannettou. Understanding and detecting hateful content using contrastive learning. AAAI 2023.

Wrap-Up

- We eliminate the need to build an image dataset for training unwanted visual data detectors by using pre-trained vision-language models.
- We introduce a new loss function. Its use eliminates the need for labor in annotating out-distribution data.
- We empirically analyze HFTT, a method composed of the above proposals. Our experiments show that HFTT is effective in a range of scenarios, from traditional OOD detection to situations involving abstract concepts, like the identification of hateful images.

Code

Paper