Machine Unlearning in Hyperbolic vs. Euclidean Multimodal Contrastive Learning: Adapting Alignment Calibration to MERU (CVPR workshop 2025)

presentor: Jihu Lee

IDEA lab
Department of Statistics
Seoul National University

September 7, 2025

Machine unlearning for concept removal

- concept c ("dog"), forget set $\mathcal{D}_f$ (positive pairs of images of dogs and captions).
- retain set $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_f$ (other pairs)
- GOAL) image encoder f_{img}, text encoder $f_{txt} \rightarrow f^*_{img}, f^*_{txt}$

  [1] $\mathrm{sim}(f^*_{img}(\text{image of a dog}), f^*_{txt}(\text{"this is a dog"})) << \mathrm{sim}(f^*_{img}(\text{image of a dog}), f^*_{txt}(\text{"this is an apple"}))$

  [2] $\mathrm{sim}(f^*_{img}(\text{image of a dog}), f^*_{txt}(\text{"this is a dog"})) << \mathrm{sim}(f^*_{img}(\text{image of an apple}), f^*_{txt}(\text{"this is a dog"}))$

  [3] $\mathrm{sim}(f^*_{img}(\text{image of an apple}), f^*_{txt}(\text{"this is an apple"})) \approx \mathrm{sim}(f_{img}(\text{image of an apple}), f_{txt}(\text{"this is an apple"}))$

Alignment Calibration (AC)

- batch - $\{(x_i^r, t_i^r)\}_{i=1}^N \subset \mathcal{D}_r$, $\{(x_i^f, t_i^f)\}_{i=1}^N \subset \mathcal{D}_f$

$$L_{\text{AC}} = L_{\text{retain}} + \varepsilon \cdot L_{\text{forget}} \tag{1}$$

[1] L_{retain}: same as CLIP loss calculated in the retain set

[2]

$$L_{\text{forget}} = \alpha \cdot L_{\text{neg}} + \beta \cdot L_{\text{pos}} + \gamma \cdot L_{\text{perf}} \tag{2}$$

- L_{neg}: maximizes similarity between negative pairs in the forget set
- L_{pos}: minimizes similarity between positive pairs in the forget set
- L_{perf}: ensures that general model capabilites remain intact

Alignment Calibration (AC) - Retain Loss

$$L_{\text{retain}} = -\frac{1}{2N}\sum_{i=1}^{N}\Bigg[\log\frac{\exp(\mathrm{sim}(x_i'^r, t_i'^r)/\tau)}{\sum_{j=1}^{2N}\exp(\mathrm{sim}(x_i'^r, t_j')/\tau)} + \log\frac{\exp(\mathrm{sim}(x_i'^r, t_i'^r)/\tau)}{\sum_{j=1}^{2N}\exp(\mathrm{sim}(x_j', t_i'^r)/\tau)}\Bigg] \quad (3)$$

Alignment Calibration (AC) - Forget Loss

- $$L_{\mathsf{neg}} = -\frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\frac{\mathrm{sim}(x_i'^f, t_j'^f) + \mathrm{sim}(x_j'^f, t_i'^f)}{\tau} \tag{4}$$

- $$L_{\mathsf{pos}} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{sim}(x_i'^f, t_i'^f)/\tau \tag{5}$$

- $$\begin{aligned} L_{\mathsf{perf}} =& \frac{1}{2N}\sum_{i=1}^{N}\Bigg[\log\left(\frac{1}{2N}\sum_{j=1}^{2N}\exp(\mathrm{sim}(x_i'^f, t_j')/\tau)\right) \\ &+ \log\left(\frac{1}{2N}\sum_{j=1}^{2N}\exp(\mathrm{sim}(x_j', t_i'^f)/\tau)\right)\Bigg] \end{aligned} \tag{6}$$

Hyperbolic Alignment Calibration (HAC) - Main contribution

- $$L_{\text{HAC}} = L_{\text{retain}} + \varepsilon \cdot L_{\text{forget}} + \omega_r \cdot L_{\text{r-ent}} + \omega_f \cdot L_{\text{f-ent}} \tag{7}$$
- cosine similarity $\leftarrow$ negative hyperbolic distance

  [1]

$$L_{\text{r-ent}} = \frac{1}{N} \sum_{i=1}^{N} \max(0, \mathrm{ext}(x_i'^r, t_i'^r) - \mathrm{aper}(t_i'^r)) \tag{8}$$

  [2]

$$L_{\text{f-ent}} = \frac{1}{N} \sum_{i=1}^{N} \max(0, \mathrm{aper}(t_i'^r) - \mathrm{ext}(x_i'^r, t_i'^r)) \tag{9}$$

HAC-reg

- $$L_{\text{HAC-reg}} = L_{\text{HAC}} + \lambda \cdot L_{\text{norm-reg}} \tag{10}$$

- $$L_{\text{norm-reg}} = \frac{1}{N} \sum_{i=1}^{N} (\|x_i'^f\|_{\mathcal{L}} + \|t_i'^f\|_{\mathcal{L}}) \tag{11}$$

Experiments

- Task: Zero-shot image classification
  - "a picture of a [CLASS]"
  - R-acc: accuracy on retained classes ($\uparrow$)
  - F-acc: accuracy on forgotten classes ($\downarrow$)

- Train: RedCaps2

Concept-class	Subreddits	Image-text samples	% on Redcaps	% on Redcaps2
dogs	dogpictures, bordercollie, bostonterrier, lookatmydog, doggos, bulldogs, australiancattledog, frenchbulldogs, bernesemountaindogs, australianshepherd, beagle, chihuahua, corgi, dobermanpinscher, husky, labrador, pitbulls, pomeranians, pug, pugs, rarepuppers, rottweiler	511585	4.26%	7.33%
cats	cats, blackcats, supermodelcats, catpictures, siamesecats, bengalcats, siberiancats	532640	4.43%	7.63%
food	food, foodporn, veganfoodporn, healthyfood, breakfastfood, chinesefood, tastyfood, budgetfood, baking, bento, breadit, breakfastfood, breakfast, burgers, chefit, pizza, sushi, tacos, veganrecipes, vegetarian	630971	5.25%	9.04%
plants	houseplants, plants, plantedtank, airplants, plantbaseddiet, plantsandpots, carnivorousplants, flowers, bonsai, botanicalporn, cactus, microgreens, monstera, orchids, permaculture, roses, succulents, vegetablegardening, gardening	587798	4.89%	8.42%
Total	68 subreddits	2262994	18.85%	32.43%

Table 6. Grouping of subreddits to higher-order concepts.

Results: performances

Table 1. Zero-shot classification accuracy in retain and forget sets, varying positive alignment calibration. Largest difference in retain and forget performance in green. Best value for each column and geometry in **bold**.

Method	Weights α, γ	Weights β	CIFAR-10 R-acc↑	CIFAR-10 F-acc↓	O-IIIT Pets R-acc↑	O-IIIT Pets F-acc↓
AC	0.75	0	**60.5**	45.6	73.6	66.2
		0.25	60.3	31.7	73.5	48.7
		0.5	58.7	**21.2**	**74.9**	31.5
		0.75	58.4	24.9	73.9	**24.6**
f-C	0	0	58.9	48.1	74.6	69.9
f-C-R			60.6	63.6	72.3	72.2
O-C			59.4	66.4	74.6	73.2
HAC	0.5	0	55.2	73.6	74.8	63.9
		0.25	34.7	**0.0**	62.1	**15.8**
		0.5	49.9	**0.0**	69.6	17.9
		0.75	39.6	0.03	63.9	16.1
f-M	0	0	**56.4**	73.0	**75.2**	65.9
f-M-R			41.7	95.7	71.8	65.8
O-M			38.1	94.6	72.0	70.8

Results: performances

Table 2. Zero-shot classification accuracy in retain forget sets, varying negative alignment calibration and performance preserving. Largest difference in retain and forget performance in green. Best value for each column and geometry in **bold**.

Method	Weights		CIFAR-10		O-IIIT Pets	
	α, γ	β	R-acc↑	F-acc↓	R-acc↑	F-acc↓
AC	0.5	0.5	**58.8**	24.0	74.6	32.9
	0.75		58.7	**21.2**	**74.9**	**31.5**
	1		57.2	27.4	73.6	41.5
HAC	0.5	0.5	**49.9**	**0.0**	**69.6**	**17.9**
	0.75		40.7	0.02	67.5	18.0
	1		42.7	0.04	68.6	19.8

Results: performances

Table 3. Zero-shot classification accuracy in retain and forget sets, varying the weight entailment losses. Largest difference in retain and forget performance in green. Best value for each column and geometry in **bold**.

Weights			CIFAR-10		O-IIIT Pets	
ϵ	ω_r	ω_f	R-acc↑	F-acc↓	R-acc↑	F-acc↓
0	0.2	1.0	**56.3**	73	**74.9**	66.6
	1.0	0.2	47.2	85.9	68.5	64.6
0.05	0.2	1.0	39.9	**0.0**	60.7	**0.08**
	1.0	0.2	49.6	37.0	40.1	0.10
0.1	0.2	1.0	52.7	**0.0**	67.9	15.1
	1.0	0.2	44.0	48.7	56.8	18.6

Results: performances

Table 4. Zero-shot classification accuracy in retain and forget sets, varying the hyperbolic norm regularization. Largest difference in retain and forget performance in green. Best value for each column and geometry in **bold**.

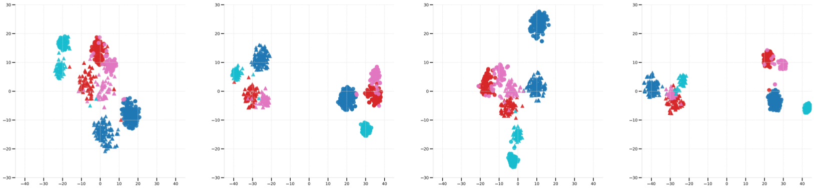
Method	Weight	CIFAR-10		O-IIIT Pets	
	λ	R-acc↑	F-acc↓	R-acc↑	F-acc↓
HAC	0	52.0	13.0	46.3	**0.04**
HAC-reg	0.1	52.7	**0.0**	67.9	15.1
	0.5	54.0	**0.0**	66.3	10.8
	2.0	**56.8**	49.2	**74.8**	35.9

Results: performances

Table 5. Zero-shot classification accuracy in retain set (R-acc) and forget set (F-acc), across different tasks, after unlearning: (A) "dog"; (B) "dog" and "cat"; (C) "dog", "cat", "food" and "plant". We report results for both CLIP and MERU after alignment calibration using the optimal configuration from Section 4.3. Values in **bold** indicate better at retaining or unlearning across A, B and C. A blank space - indicate that for that experiment and dataset there is no forget or retain set.

Task	Method	Unlearn Set	CIFAR-10[18]		CIFAR-100[19]		STL-10[1]		O-IIIT Pets[26]		Food101[3]		Flowers102[25]	
			R-acc↑	F-acc↓	R-acc↑	F-acc↓	R-acc↑	F-acc↓	R-acc↑	F-acc↓	R-acc↑	F-acc↓	R-acc↑	F-acc↓
Zero-shot Classification	AC	A	**58.7**	21.2	**27.9**	-	**88.1**	83.1	**74.9**	31.5	**72.4**	-	**44.7**	-
		B	**90.3**	71.4	**26.6**	-	**90.3**	71.4	-	53.4	**72.5**	-	**45.0**	-
		C	**90.0**	77.0	**23.4**	57.2	**90.0**	77.0	-	64.0	-	0.16	-	19.2
	HAC-reg	A	54.0	**0.0**	20.6	-	84.3	**38.0**	66.3	**10.8**	67.6	-	40.1	-
		B	83.5	**2.1**	21.8	-	83.5	**2.1**	-	**25.7**	59.6	-	36.4	-
		C	82.7	**22.1**	18.8	**21.6**	82.7	**22.1**	-	**28.7**	-	**0.08**	-	**0.04**

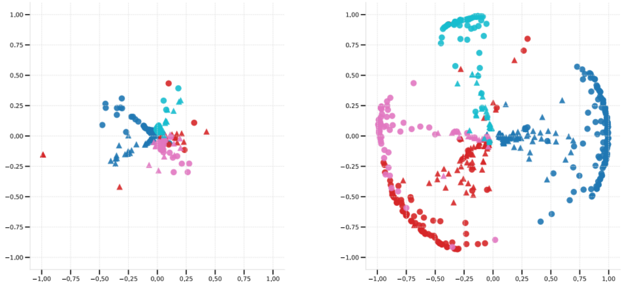
Results: visualizations



(a) Original CLIP (b) Unlearned CLIP (c) Original MERU (d) Unlearned MERU

Figure 1. Latent space visualizations with T-SNE of CLIP and MERU before and after removing the concept "dog". △ refer to text embeddings, ○ to image embeddings, and colors to **dogs**, **cats**, **pizzas**, and **buses**.

Results: visualizations



(a) Original MERU

(b) Unlearned MERU

Figure 2. Latent space visualizations with hyperbolic T-SNE of MERU before and after removing the concept ”dog”. △ refer to text embeddings, ○ to image embeddings, and colors to **dogs**, **cats**, **pizzas**, and **buses**.

[1] Vidal, À. P., Nasrollahi, K., Moeslund, T. B., & Escalera, S. (2025). Machine Unlearning in Hyperbolic vs. Euclidean Multimodal Contrastive Learning: Adapting Alignment Calibration to MERU. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 1644-1653).


THANK
you!