A Few Papers about Security for CLIP Retrieval Unlearning

Paper Reading by Luohao Lin

2025.09.03

智能多媒体内容计算实验室
Intelligent Multimedia Content Computing Lab

- Safe-CLIP
- Hyperbolic-CLIP
- Hyperbolic vs Euclidean
- CLIPErase
- 总结

Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models

Samuele Poppi*[1,2], Tobia Poppi*[1,2], Federico Cocchi*[1,2], Marcella Cornia[1], Lorenzo Baraldi[1], and Rita Cucchiara[1,3]

[1] University of Modena and Reggio Emilia, Italy
name.surname@unimore.it
[2] University of Pisa, Italy
name.surname@phd.unipi.it
[3] IIT-CNR, Italy

Safe-CLIP

Tobia Poppi

University of Modena and Reggio Emilia, University of Pisa
在 unimore.it 的电子邮件经过验证 - 首页
Responsible AI　AI Safety　Vision-and-Language

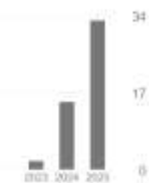
关注

标题	引用次数	年份
Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models S Poppi, T Poppi, F Cocchi, M Cornia, L Baraldi, R Cucchiara Proceedings of the European Conference on Computer Vision	47 *	2024
Hyperbolic Safety-Aware Vision-Language Models T Poppi, T Kasarla, P Mettes, L Baraldi, R Cucchiara IEEE/CVF Conference on Computer Vision and Pattern Recognition	2	2025
Uncovering the background-induced bias in RGB based 6-DOF object pose estimation E Govi, D Sapienza, C Scribano, T Poppi, G Franchini, P Ardón, ... arXiv preprint arXiv:2304.08230	1	2023
Improving LLM First-Token Predictions in Multiple-Choice Question Answering via Prefilling Attack S Cappelletti, T Poppi, S Poppi, ZX Yong, D Garcia-Olano, M Cornia, ... arXiv preprint arXiv:2505.15323		2025

创建我的个人资料

引用次数

	总计	2020 年至今
引用	50	50
h 指数	2	2
i10 指数	1	1



Rita Cucchiara

Università degli Studi di Modena e Reggio Emilia, Italia
在 unimore.it 的电子邮件经过验证 - 首页
Computer Vision　Pattern Recognition　Deep Learning　Multimedia　Artificial Intelligence

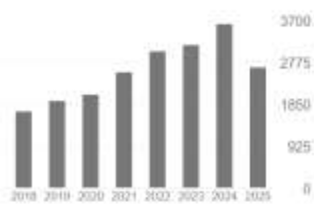
关注

标题	引用次数	年份
Performance measures and a data set for multi-target, multi-camera tracking E Ristani, F Solera, R Zou, R Cucchiara, C Tomasi European conference on computer vision, 17-35	3631	2016
Detecting moving objects, ghosts, and shadows in video streams R Cucchiara, C Grana, M Piccardi, A Prati IEEE transactions on pattern analysis and machine intelligence 25 (10), 1337 ...	2305	2003
Visual tracking: An experimental survey AWM Smeulders, DM Chu, R Cucchiara, S Calderara, A Dehghan, ... IEEE transactions on pattern analysis and machine intelligence 36 (7), 1442-1468	2090	2013
Meshed-memory transformer for image captioning M Cornia, M Stefanini, L Baraldi, R Cucchiara Proceedings of the IEEE/CVF conference on computer vision and pattern ...	1388	2020

创建我的个人资料

引用次数　查看全部

	总计	2020 年至今
引用	32996	17152
h 指数	75	53
i10 指数	357	207

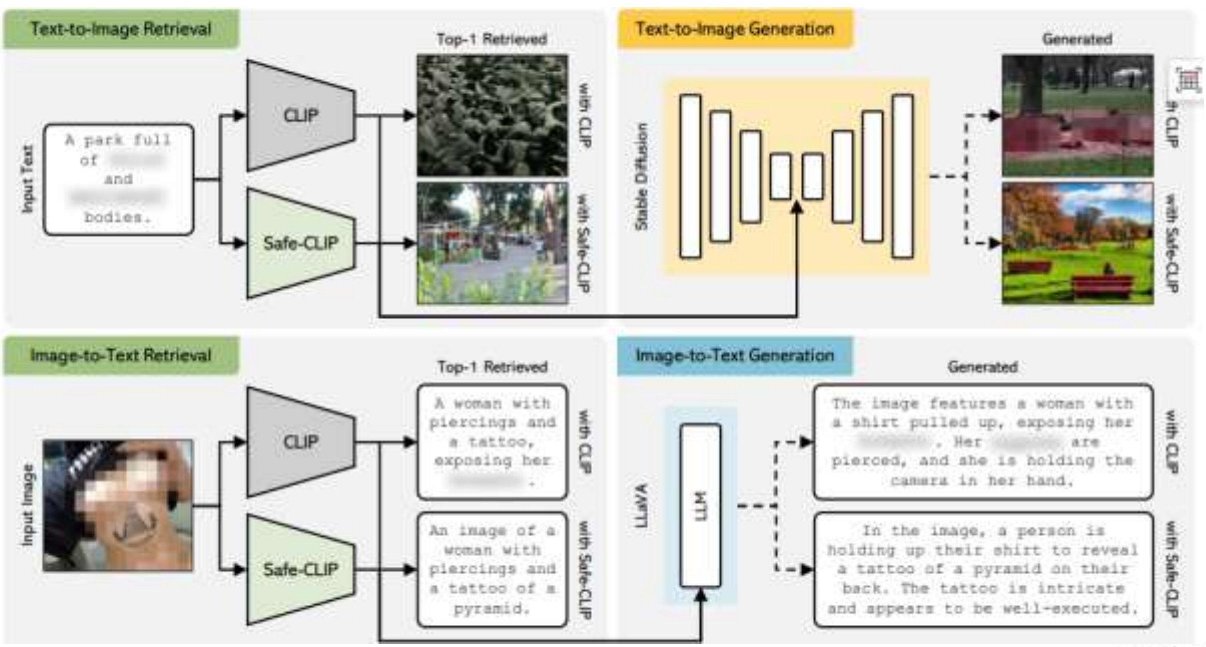


体内容计算实验室
imedia Content Computing Lab

Safe-CLIP

模型训练过程中混入NSFW数据，导致：

- ➢ 跨模态检索时，NSFW文本检索到有害图片
- ➢ T2I时，生成包含NSFW概念的图片
- ➢ I2T时，触发有毒的描述性文本

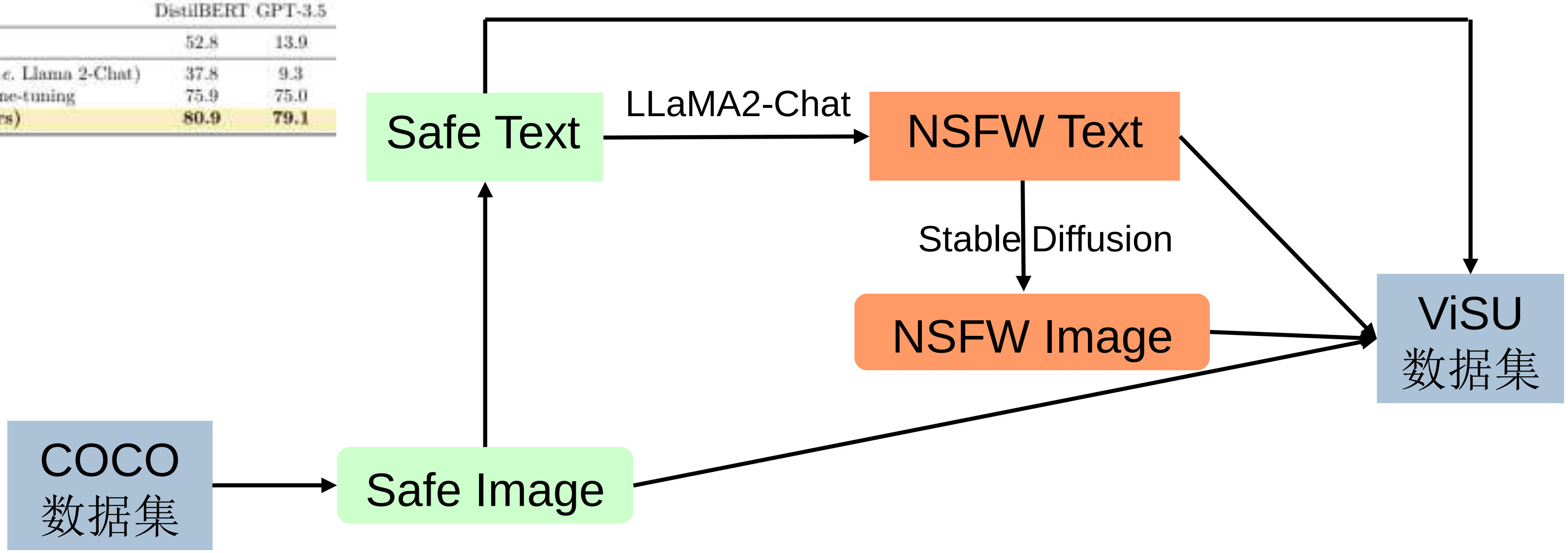


Safe-CLIP

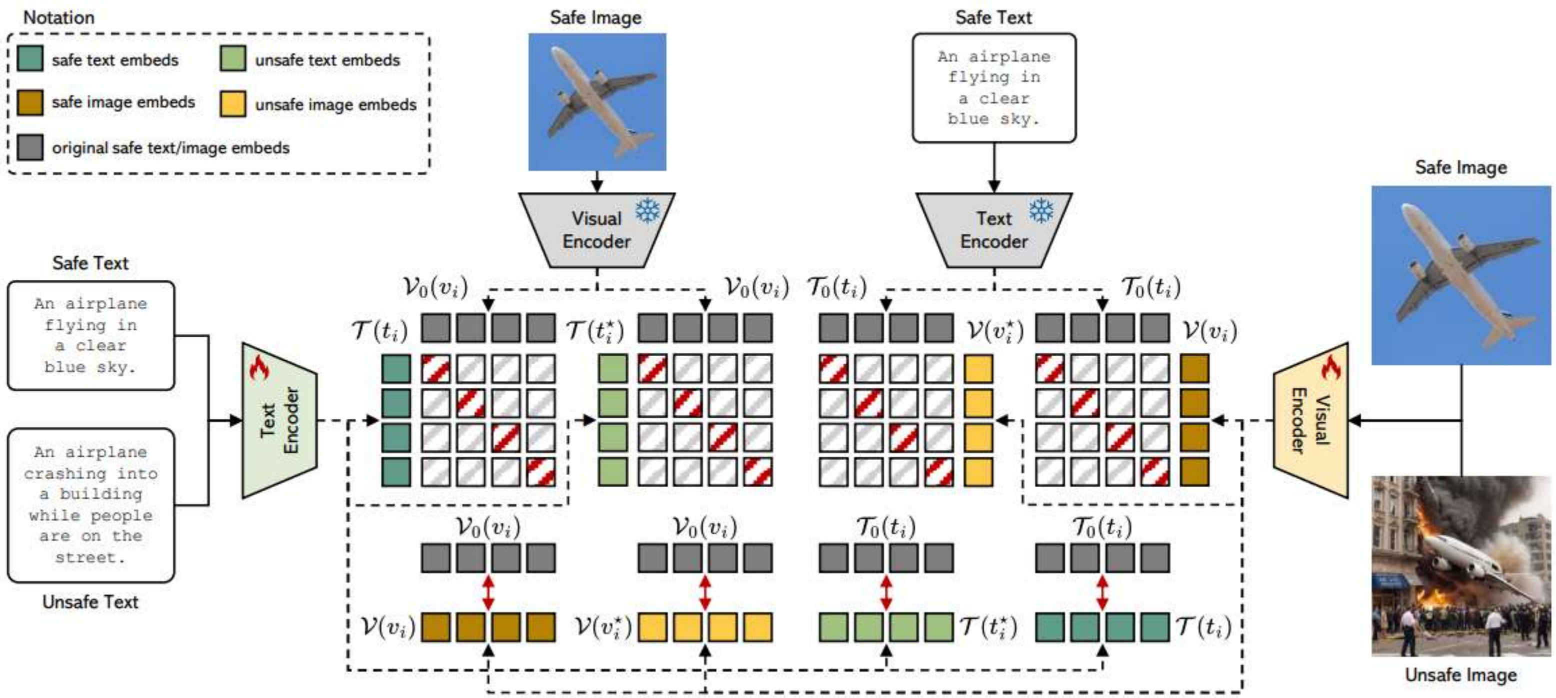

```
{
    "incremental_id": 0,
    "safe": "A young man holding an umbrella next to a herd of cattle.",
    "nsfw": "A young man holding a gun next to a herd of cattle, with a grin on his face and a wicked look in his eyes. ",
    "coco_id": "000000184613",
    "tag": "weapons",
    "prompt_id": 16
},
```


$$\text{rank}(t_i^\star, t_i) = \text{CLIP-Sim}(t_i^\star, t_i) + \text{NSFWRate}(t_i^\star),$$

Dataset	% NSFW DistilBERT	% NSFW GPT-3.5
I2P [44]	52.8	13.9
w/o SFT (*i.e.* Llama 2-Chat)	37.8	9.3
w/o DPO fine-tuning	75.9	75.0
ViSU (Ours)	**80.9**	**79.1**



Safe-CLIP



$$L_{\text{redir},1} = -\frac{1}{N}\left(\sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{T}(t_i^\star),\mathcal{V}_0(v_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{T}(t_j^\star),\mathcal{V}_0(v_i))/\tau)} + \sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{T}(t_i^\star),\mathcal{V}_0(v_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{T}(t_i^\star),\mathcal{V}_0(v_j))/\tau)}\right. \tag{4}$$

$$\left.+\sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{V}(v_i^\star),\mathcal{T}_0(t_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{V}(v_j^\star),\mathcal{T}_0(t_i))/\tau)} + \sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{V}(v_i^\star),\mathcal{T}_0(t_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{V}(v_i^\star),\mathcal{T}_0(t_j))/\tau)}\right)$$

$$L_{\text{redir},2} = -\frac{1}{N}\left(\sum_{i=1}^{N}\cos(\mathcal{T}(t_i^\star),\mathcal{T}_0(t_i)) + \sum_{i=1}^{N}\cos(\mathcal{V}(v_i^\star),\mathcal{V}_0(v_i))\right)$$

$$L_{\text{pres},1} = -\frac{1}{N}\left(\sum_{i=1}^{N}\cos(\mathcal{T}(t_i),\mathcal{T}_0(t_i)) + \sum_{i=1}^{N}\cos(\mathcal{V}(v_i),\mathcal{V}_0(v_i))\right).$$

$$L_{\text{pres},2} = -\frac{1}{N}\left(\sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{V}_0(v_i),\mathcal{T}(t_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{V}_0(v_i),\mathcal{T}(t_j))/\tau)} + \sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{V}_0(v_i),\mathcal{T}(t_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{V}_0(v_j),\mathcal{T}(t_i))/\tau)}\right. \tag{7}$$

$$\left.+\sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{T}_0(t_i),\mathcal{V}(v_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{T}_0(t_i),\mathcal{V}(v_j))/\tau)} + \sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{T}_0(t_i),\mathcal{V}(v_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{T}_0(t_j),\mathcal{V}(v_i))/\tau)}\right).$$

Safe-CLIP

Model	Text-to-Image ($\mathbf{T}$-to-$\mathbf{V}$)			Image-to-Text ($\mathbf{V}$-to-$\mathbf{T}$)			Text-to-Image ($\mathbf{T}^\star$-to-$\mathbf{V} \cup \mathbf{V}^\star$)			Image-to-Text ($\mathbf{V}^\star$-to-$\mathbf{T} \cup \mathbf{T}^\star$)		
	R@1	R@10	R@20	R@1	R@10	R@20	R@1	R@10	R@20	R@1	R@10	R@20
CLIP (ViT-L) [39]	36.8	71.6	81.5	39.8	74.2	83.5	2.0	24.8	33.2	4.5	32.9	40.6
DataComp-1B (ViT-L) [13]	46.7	79.7	87.4	47.0	81.3	88.9	1.6	28.1	35.6	5.5	37.5	44.9
w/o inap. content redirection	49.9	83.7	90.3	48.1	83.6	90.5	1.6	30.4	40.1	6.1	35.2	42.6
w/o negative cosine similarities	41.9	78.5	87.3	41.5	77.8	86.9	**8.2**	46.0	56.6	13.7	60.4	68.2
Safe-CLIP	45.9	81.8	89.7	45.3	82.3	89.7	8.0	**46.9**	**58.0**	**19.1**	**62.9**	**71.1**

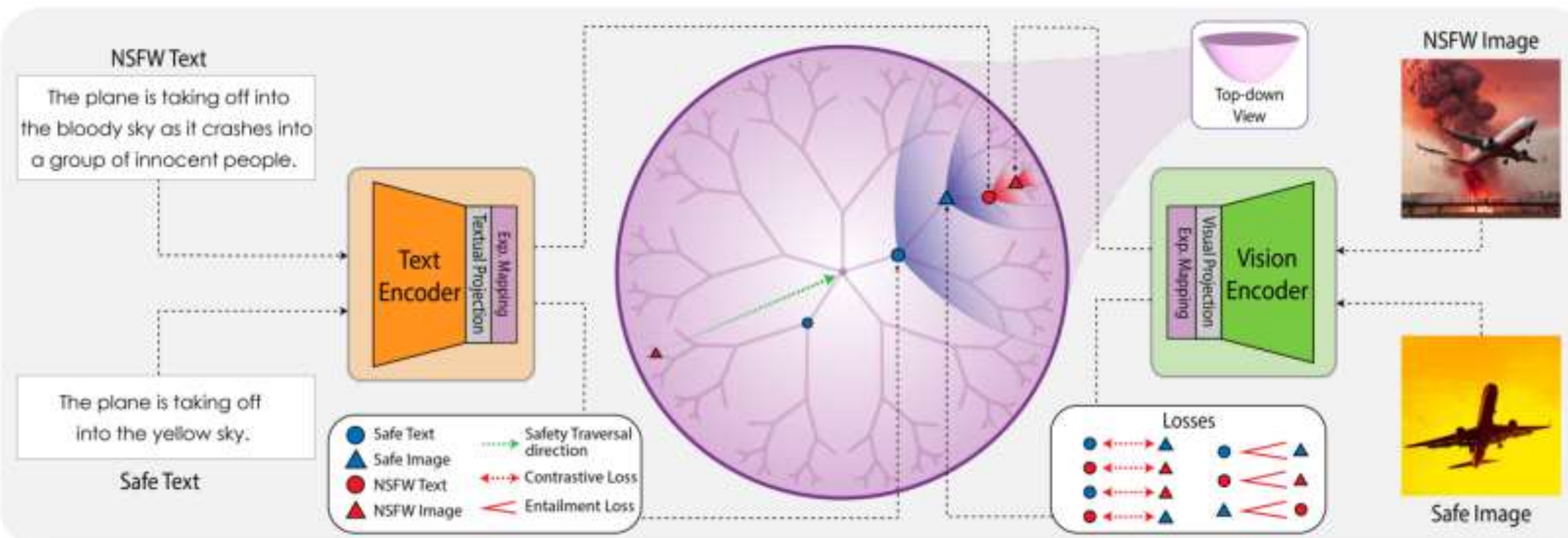
- Safe-CLIP
- Hyperbolic-CLIP
- Hyperbolic vs Euclidean
- CLIPErase
- 总结

Hyperbolic Safety-Aware Vision-Language Models

Tobia Poppi*[1,2] Tejaswi Kasarla*[3] Pascal Mettes[3] Lorenzo Baraldi[1] Rita Cucchiara[1,4]
[1]University of Modena and Reggio Emilia, Italy [2] University of Pisa, Italy
[3] University of Amsterdam, Netherlands [4]IIT-CNR, Italy
[1]{name.surname}@unimore.it [2] {name.surname}@phd.unipi.it [3] {initial.surname}@uva.nl

Hyperbolic-CLIP

单纯的遗忘 -> 对安全和NSFW内容的有效区分



$$g_T(T_k) \ll g_I(I_k) \ll g_T(T_k^\star) \ll g_I(I_k^\star).$$

Hyperbolic-CLIP

$$L(I,T,I^\star,T^\star) = L_{\text{hSC}}(I,T,I^\star,T^\star) + L_{\text{hSE}}(I,T,I^\star,T^\star).$$

$$L_{\text{hSC}}(I,T,I^\star,T^\star) = L^\star_{cont}(I,T) + L^\star_{cont}(I^\star,T^\star) + L^\star_{cont}(I,T^\star) + L^\star_{cont}(I^\star,T).$$

$$L_{\text{hSE}}(I,T,I^\star,T^\star) = L^\star_{ent}(I,T) + L^\star_{ent}(T^\star,I) + L^\star_{ent}(I^\star,T^\star).$$

$$L^*_{cont}(I,T) = -\sum_{i\in B} \log \frac{\exp(d_{\mathcal{L}}(g_I(I_i),g_T(T_i))/\tau)}{\sum_{k=1,k\neq i}^{B} \exp(d_{\mathcal{L}}(g_I(I_i),g_T(T_k))/\tau)},$$

$$L^\star_{ent}(I,T) = \max(0, \phi(I_k,T_k) - \eta\omega(T_k)) \text{ and}$$
$$L^\star_{ent}(I^\star,T^\star) = \max(0, \phi(I^\star_k,T^\star_k) - \eta\omega(T^\star_k)),$$

Hyperbolic-CLIP

Model	Text-to-Image (T-to-I) R@1	R@10	R@20	Image-to-Text (I-to-T) R@1	R@10	R@20	Text-to-Image ($T^\star$-to-$I \cup I^\star$) R@1	R@10	R@20	Image-to-Text ($I^\star$-to-$T \cup T^\star$) R@1	R@10	R@20
CLIP [69]	36.8	71.6	81.5	39.8	74.2	83.5	2.0	24.8	33.2	4.6	32.9	40.6
MERU [20]	14.9	43.0	54.2	14.7	42.3	53.8	2.2	15.2	21.5	4.4	22.6	29.4
HyCoCLIP [63]	34.3	71.2	80.6	34.4	71.3	82.2	2.8	25.3	33.2	8.2	37.8	45.7
Safe-CLIP [66]	45.9	81.8	89.7	45.3	82.3	89.8	8.0	46.9	58.0	19.1	62.9	71.1
MERU*	50.0	84.1	91.1	51.2	85.3	92.3	2.3	39.9	49.4	5.7	47.9	54.7
HyCoCLIP*	47.7	81.9	89.1	46.7	82.7	90.4	1.5	32.7	42.3	6.9	45.2	53.6
HySAC	**49.8**	**84.1**	**90.7**	**48.2**	**84.2**	**91.2**	**30.5**	**62.8**	**71.8**	**42.1**	**73.3**	**79.8**

Table 1. **Safe content retrieval performance on ViSU test set.** Across all tasks and recall rates, HySAC improves over existing safety unlearning CLIP and hyperbolic CLIP models, highlighting that our approach is able to navigate unsafe image or text inputs towards relevant but safe retrieval outputs. * CLIP fine-tuned in hyperbolic space on ViSU training set with MERU/HyCoCLIP losses.

Model	Text-to-Image ($T^\star$-to-$I^\star$) R@1	R@10	R@20	Image-to-Text ($I^\star$-to-$T^\star$) R@1	R@10	R@20	Text-to-Image ($T^\star$-to-$I^\star \cup I$) R@1	R@10	R@20	Image-to-Text ($I^\star$-to-$T^\star \cup T$) R@1	R@10	R@20
CLIP [69]	73.1	94.9	97.6	72.8	95.2	97.7	68.4	92.3	95.9	67.1	93.3	96.7
MERU [20]	29.4	62.4	72.2	25.8	57.7	67.8	23.5	54.0	64.3	19.5	51.1	61.2
HyCoCLIP [63]	69.5	93.1	95.8	65.0	91.1	95.0	63.7	89.7	93.7	55.2	88.0	92.7
Safe-CLIP [66]	58.0	86.2	91.4	56.0	85.1	91.0	47.7	80.0	85.8	32.1	77.1	84.6
HySAC	**81.4**	**98.4**	**99.4**	**82.2**	**97.8**	**99.2**	**81.1**	**98.4**	**99.4**	**80.5**	**97.2**	**98.9**

Table 2. **Unsafe content retrieval performance on ViSU test set.** Akin to safe content retrieval, our approach performs best. This is a result of our objective, as we assign different content to different regions, enabling us to maintain valuable safety information.

- Safe-CLIP
- Hyperbolic-CLIP
- Hyperbolic vs Euclidean
- CLIPErase
- 总结

Machine Unlearning in Hyperbolic vs. Euclidean Multimodal Contrastive Learning: Adapting Alignment Calibration to MERU

Àlex Pujol Vidal[1,3]
alexpv@create.aau.dk

Kamal Nasrollahi[1,3,4]
kna@milestone.dk

Thomas B. Moeslund[1,4]
tbm@create.aau.dk

Sergio Escalera[1,2,4]
sescalera@ub.edu

Hyperbolic vs Euclidean

Àlex Pujol Vidal

关注

PhD student, Aalborg University
在 create.aau.dk 的电子邮件经过验证 - 首页
Machine Learning Artificial Intelligence

标题	引用次数	年份
Verifying machine unlearning with explainable ai ÀP Vidal, AS Johansen, MNS Jahromi, S Escalera, K Nasrollahi, ... International Conference on Pattern Recognition, 458-473	2	2024
Machine Unlearning in Hyperbolic vs. Euclidean Multimodal Contrastive Learning: Adapting Alignment Calibration to MERU ÀP Vidal, K Nasrollahi, TB Moeslund, S Escalera Proceedings of the Computer Vision and Pattern Recognition Conference, 1644-1653	1	2025
Machine Unlearning in Hyperbolic vs. Euclidean Multimodal Contrastive Learning: Adapting Alignment Calibration to MERU À Pujol Vidal, S Escalera, K Nasrollahi, TB Moeslund arXiv e-prints, arXiv: 2503.15166		2025

创建我的个人资料

引用次数

	总计	2020 年至今
引用	3	3
h 指数	1	1
i10 指数	0	0

开放获取的出版物数量 查看全部

0 篇文章 1 篇文章

无法查看的文章 可查看的文章

根据资助方的强制性开放获取政策

Sergio Escalera

关注

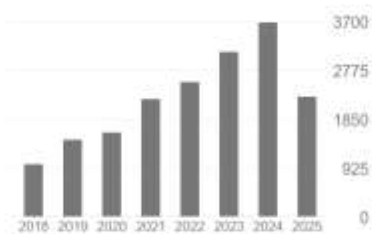
Prof., ICREA Academy, University of Barcelona, Computer Vision Center, ELLIS & IAPR & AAIA Fellow
在 ub.edu 的电子邮件经过验证 - 首页
Human Behavior Analysis Machine Learning Computer Vision Affective Computing Social Signal Processing

标题	引用次数	年份
Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications CA Corneanu, MO Simón, JF Cohn, SE Guerrero IEEE transactions on pattern analysis and machine intelligence 38 (8), 1548-1568	673	2016
Survey on emotional body gesture recognition F Noroozi, CA Corneanu, D Kamińska, T Sapiński, S Escalera, ... IEEE transactions on affective computing 12 (2), 505-523	604	2018
Bi-directional ConvLSTM U-Net with densley connected convolutions R Azad, M Asadi-Aghbolaghi, M Fathy, S Escalera Proceedings of the IEEE/CVF international conference on computer vision ...	601	2019

创建我的个人资料

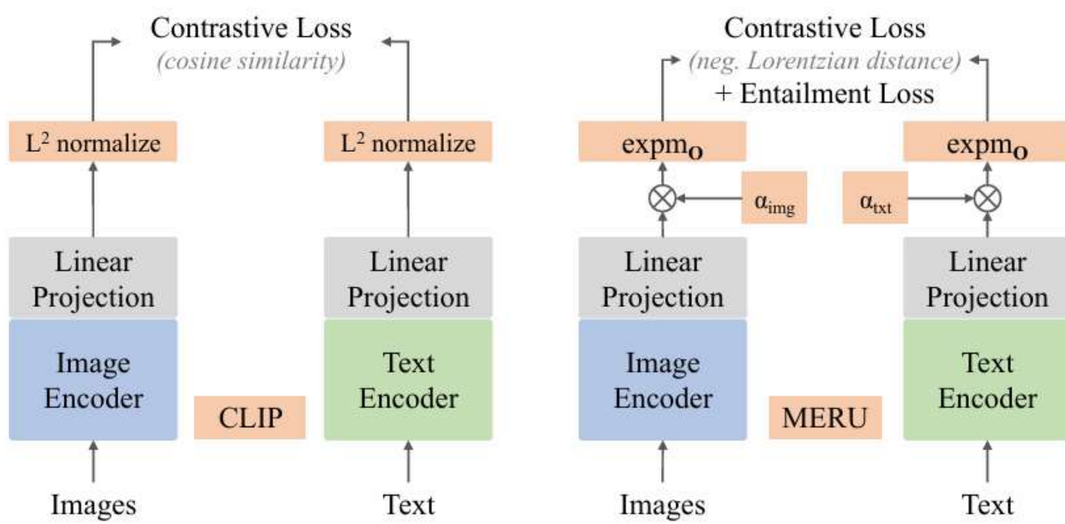
引用次数 查看全部

	总计	2020 年至今
引用	21383	15514
h 指数	77	67
i10 指数	269	207



内容计算实验室
...dia Content Computing Lab

Hyperbolic vs Euclidean



Hyperbolic vs Euclidean

AC（Alignment Calibration）：

Published in Transactions on Machine Learning Research (08/2025)

MUC: Machine Unlearning for Contrastive Learning with Black-box Evaluation

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

$$L_{\text{pos}} = \frac{1}{N}\sum_{i=1}^{N}\text{sim}(x_i'^f, t_i'^f)/\tau.$$

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

$$L_{\text{perf}} = \frac{1}{2N}\sum_{i=1}^{N}\left[\log\left(\frac{1}{2N}\sum_{j=1}^{2N}\exp(\text{sim}(x_i'^f, t_j')/\tau)\right)\right. \tag{9}$$

$$\left.+\log\left(\frac{1}{2N}\sum_{j=1}^{2N}\exp(\text{sim}(x_j', t_i'^f)/\tau)\right)\right] \tag{10}$$

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

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

智能多媒体内容计算实验室

Intelligent Multimedia Content Computing Lab

Hyperbolic vs Euclidean

HAC（Hyperbolic Alignment Calibration）：

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

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

$$L_{\text{HAC}} = L_{\text{retain}} + \varepsilon \cdot L_{\text{forget}} + \omega_r \cdot L_{\text{r-ent}} + \omega_f \cdot L_{\text{f-ent}}.$$

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

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

Hyperbolic vs Euclidean

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

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	0.5	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

Task	Method	Unlearn Set	CIFAR-10[18] R-acc↑	CIFAR-10[18] F-acc↓	CIFAR-100[19] R-acc↑	CIFAR-100[19] F-acc↓	STL-10[1] R-acc↑	STL-10[1] F-acc↓	O-IIIT Pets[26] R-acc↑	O-IIIT Pets[26] F-acc↓	Food101[3] R-acc↑	Food101[3] F-acc↓	Flowers102[25] R-acc↑	Flowers102[25] 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**

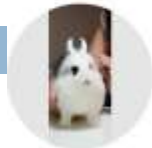
- Safe-CLIP
- Hyperbolic-CLIP
- Hyperbolic vs Euclidean
- CLIPErase
- 总结

CLIPErase: Efficient Unlearning of Visual-Textual Associations in CLIP

Tianyu Yang[1], Lisen Dai[2], Xiangqi Wang[1], Minhao Cheng[3],
Yapeng Tian[4], Xiangliang Zhang[1*]

[1]University of Notre Dame
[2]Columbia University
[3]Pennsylvania State University
[4]University of Texas at Dallas

CLIPErase



Tianyu Yang

关注

University of Notre Dame

在 nd.edu 的电子邮件经过验证 - 首页

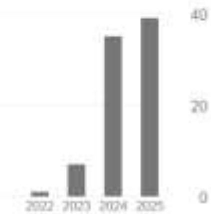
Deep Learning Computer Vision Multi-Modal Learning Natural Language Processing

标题	引用次数	年份
Unimath: A foundational and multimodal mathematical reasoner Z Liang, T Yang, J Zhang, X Zhang Proceedings of the 2023 Conference on Empirical Methods in Natural Language ...	32	2023
Scemqa: A scientific college entrance level multimodal question answering benchmark Z Liang, K Guo, G Liu, T Guo, Y Zhou, T Yang, J Jiao, R Pi, J Zhang, ... Proceedings of the 62nd Annual Meeting of the Association for Computational ...	28 *	2024
CLIPErase: Efficient Unlearning of Visual-Textual Associations in CLIP T Yang, L Dai, X Wang, C Minghao, Y Tian, X Zhang Proceedings of the 63nd Annual Meeting of the Association for Computational ...	10	2025
Ensemble Learning for Interpretable Concept Drift and Its Application to Drug Recommendation	4	2023

创建我的个人资料

引用次数

	总计	2020 年至今
引用	82	82
h 指数	4	4
i10 指数	3	3



Xiangliang Zhang

关注

Leonard C. Bettex Collegiate Professor, Computer Science and Engineering, University of Notre Dame

在 nd.edu 的电子邮件经过验证 - 首页

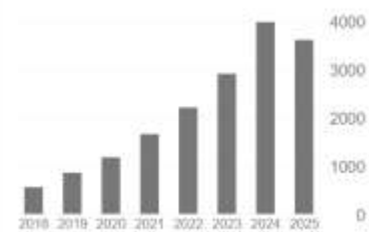
Machine Learning AI for Science

标题	引用次数	年份
The soundscape of the Anthropocene ocean CM Duarte, L Chapuis, SP Collin, DP Costa, RP Devassy, VM Eguiluz, ... Science 371 (6529)	881	2021
Large Language Model based Multi-Agents: A Survey of Progress and Challenges T Guo, X Chen, Y Wang, R Chang, S Pei, NV Chawla, O Wiest, X Zhang IJCAI 2024	803	2024
Self-supervised hypergraph convolutional networks for session-based recommendation X Xia, H Yin, J Yu, Q Wang, L Cui, X Zhang Proceedings of the AAAI Conference on Artificial Intelligence 35 (5), 4503-4511	668	2021
CreditCoin: A Privacy-Preserving Blockchain-Based Incentive Announcement Network for Communications of Smart Vehicles	650	2018

创建我的个人资料

引用次数 查看全部

	总计	2020 年至今
引用	18321	15622
h 指数	68	59
i10 指数	259	237



计算实验室

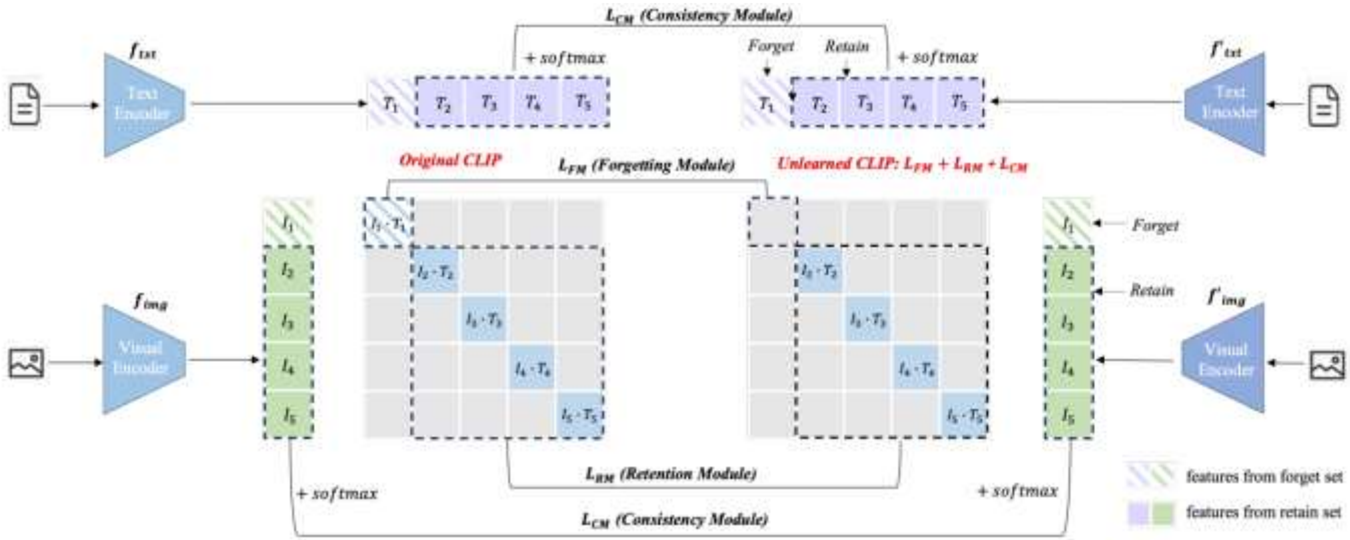
tent Computing Lab

CLIPErase

两个挑战：

- 跨模态关联破坏风险
- 概念区分精度不足

CLIPErase



$$\mathcal{L}_{\mathrm{FM}} = \frac{1}{N_f} \sum_{n=1}^{N_f} \left(f_{\mathrm{img}}(x_i^n) \cdot f_{\mathrm{txt}}(x_t^n) \right)$$

$$\mathcal{L}_{\mathrm{RM}} = -\frac{1}{N_r} \sum_{n=1}^{N_r} \log \operatorname{softmax}(f_{\mathrm{img}}(x_i^n) \cdot f_{\mathrm{txt}}(x_t^n)/\tau) \quad (3)$$

$$\mathcal{L}_{\mathrm{CM}} = \frac{1}{N_r} \sum_{n=1}^{N_r} \left[\mathrm{KL}\left(\mathbf{p}_o^{\mathrm{img}} \parallel \mathbf{p}_u^{\mathrm{img}} \right) + \mathrm{KL}\left(\mathbf{p}_o^{\mathrm{txt}} \parallel \mathbf{p}_u^{\mathrm{txt}} \right) \right]$$

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\mathrm{RM}} + \lambda_2 \mathcal{L}_{\mathrm{FM}} + \lambda_3 \mathcal{L}_{\mathrm{CM}}$$

CLIPErase

Dataset	Method	ZS Prediction (%)		ZS Retrieval (%)	
		Acc.$D_f \downarrow$	Acc.$D_r \uparrow$	Acc.$D_f \downarrow$	Acc.$D_r \uparrow$
CIFAR-100	CLIP	86.08	72.85	88.61	73.43
	CLIP+GA	4.43	5.22	0.63	5.39
	CLIP+GradDiff	0.00	89.96	0.00	90.64
	CLIP+KL	91.88	80.88	91.77	81.51
	CLIP+ENMN	0.00	12.46	0.00	17.94
	CLIPErase (ours)	0.00	90.99	0.00	91.85
Conceptual 12M	CLIP	96.20	93.60	94.48	92.77
	CLIP+GA	38.22	4.15	1.17	5.38
	CLIP+GradDiff	4.96	97.01	5.64	97.46
	CLIP+KL	99.04	98.41	98.83	98.02
	CLIPErase (ours)	0.74	97.10	0.74	97.62

FM	RM	CM	Accuracy (%)		Improvement (%)	
			$\downarrow D_f$	$\uparrow D_r$	$\downarrow D_f$	$\uparrow D_r$
✗	✗	✗	86.08	72.85	-	-
✓	✗	✗	18.57	64.12	$\downarrow$ 67.5	$\downarrow$ 8.73
✓	✓	✗	9.40	73.14	$\downarrow$ 76.68	$\uparrow$ 0.56
✓	✓	✓	0	90.80	$\downarrow$ 86.08	$\uparrow$ 17.95

- Safe-CLIP
- Hyperbolic-CLIP
- Hyperbolic vs Euclidean
- CLIPErase
- 总结

总结与思考

- 遗忘不彻底，往往有残留
- 遗忘特定有害数据而不影响相似但无害的内容
- 单文本多概念的选择性遗忘

谢谢大家！