Membership Inference Attacks against Large Vision-Language Models

Zhan Li[*,1,3], Yongtao Wu[*,1], Yihang Chen[*,1,2], Francesco Tonin[1], Elias Abad Rocamora[1], and Volkan Cevher[1]

[1]EPFL [2]UCLA [3]Ant Group [*]Equal contribution

Paper Codes

TL;DR

Membership **I**nference **A**ttack (**MIA**) [1]: a type of attack on ML models that attempts to whether a particular data record is part of the training dataset. Our contributions can be summarized as follows.

- We release the first benchmark tailored for the detection of training data in VLLMs, called **V**ision **L**anguage **MIA** (**VL-MIA**). By leveraging Flickr and GPT-4, we construct VL-MIA that contains two images MIA tasks and one text MIA task for various VLLMs, including MiniGPT-4 [2], LLaVA 1.5 [3] and LLaMA-Adapter V2 [4].
- We perform the first individual image or description MIAs on VLLMs in a cross-modal manner. Specifically, we demonstrate that we can perform image MIAs by computing statistics from the image or text slices of the VLLM's output logits.
- We propose a target-free MIA metric, `MaxRényi-K%`, and its modified target-based `ModRényi`. We demonstrate their effectiveness on open-source VLLMs and closed-source GPT-4.

MaxRényi MIA

Rényi entropy of order α Given a probability distribution p, the Rényi entropy [5] of order α is defined as $H_\alpha(p) = \frac{1}{1-\alpha} \log \left(\sum_j (p_j)^\alpha\right), \quad 0 < \alpha < \infty, \alpha \neq 1$. At $\alpha = 1$ and $\alpha = \infty$, the entropy is defined as:

- $H_1(p) = -\sum_j p_j \log p_j$, $H_\infty(p) = -\log \max p_j$.

`MaxRényi-K%` For a token sequence $X := (x_1, x_2, \ldots, x_L)$, let the next-token probability distribution at the i-th token be: $p^{(i)}(\cdot) = \mathbb{P}(\cdot | x_1, \ldots, x_i)$. Define Max-K%$(X)$ as the subset of X containing the top K% tokens with the largest Rényi entropies. The `MaxRényi-K%` score of X is given by:

$$\texttt{MaxRényi-K\%}(X) = \frac{1}{|\text{Max-K\%}(X)|} \sum_{i \in \text{Max-K\%}(X)} H_\alpha(p^{(i)}).$$

Special cases:

- $\alpha = 1, K = 100$: standard entropy-based MIA.
- $\alpha = \infty$: the most likely next token probability. In comparison, Min-K [6] deals with the target next token probability.

Extension to Target-Based Scenarios: `ModRényi` We propose `ModRényi` for scenarios where the target token ID is known. Using a linearized Rényi entropy, $\overline{H}_\alpha(p)$, we define: $\overline{H}_\alpha(p) = \frac{1}{1-\alpha}\left(\sum_j (p_j)^\alpha - 1\right), \quad 0 < \alpha < \infty, \alpha \neq 1$. Given next token ID y, we define the modified Rényi entropy as:

$$\overline{H}_\alpha(p, y) = -\frac{1}{|\alpha - 1|}\left((1 - p_y)p_y^{|\alpha-1|} - (1 - p_y) + \sum_{j \neq y} p_j(1 - p_j)^{|\alpha-1|} - p_j\right).$$

MIAs against VLLMs: pipeline & benchmark

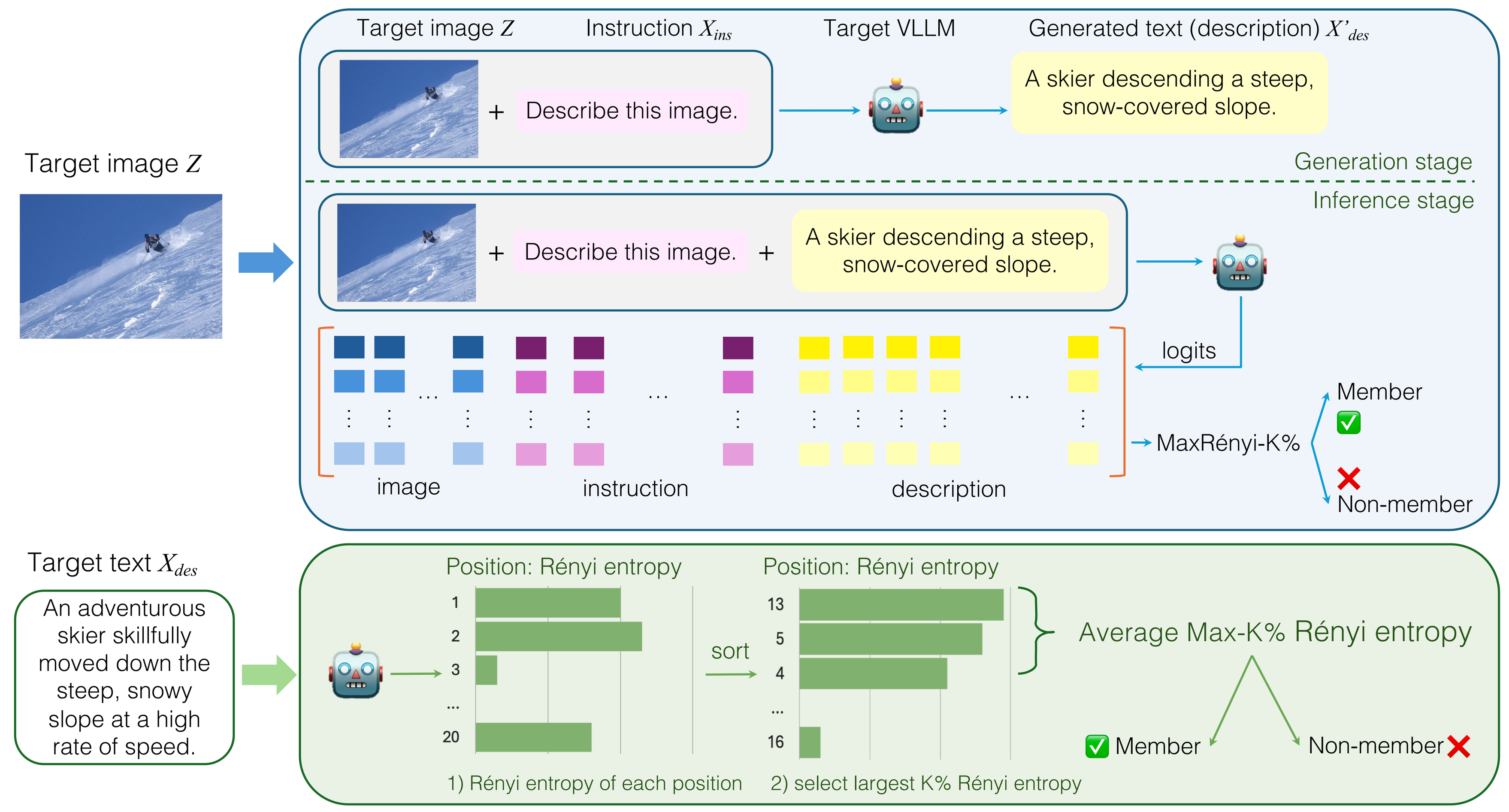

Table 1: **Overview of VL-MIA dataset**: VL-MIA covers image and text modalities and can be applied for dominant open-sourced VLLMs.

Dataset	Modality	Member data	Non-member data	Application
VL-MIA/DALL-E	image	LAION_CCS	DALL-E-generated images	LLaVA 1.5 MiniGPT-4 LLaMA_adapter v2
VL-MIA/Flickr	image	MS COCO (from Flickr)	Latest images on Flickr	LLaVA 1.5 MiniGPT-4 LLaMA_adapter v2
VL-MIA/Text	text	LLaVA v1.5 instruction-tuning text	GPT-generated answers	LLaVA 1.5 LLaMA_adapter v2
		MiniGPT-4 instruction-tuning text	GPT-generated answers	MiniGPT-4

Dataset examples

Table 2: Examples in VL-MIA/image non-member data are generated by DALL-E or collected from recent Flickr websites.

Dataset	Member	Non-member	Member	Non-member
VL-MIA/DALL-E				
VL-MIA/Flickr				

Experiments

We conduct MIAs on open-source LLaVA and closed-source model GPT-4.

Table 3: **Image MIA on VL-MIA/Flickr on LLaVA** with a size of 2000.

Metric		img	inst	desp	inst+desp
Perplexity*		N/A	0.365	0.665	0.561
Min_10% Prob*		N/A	0.353	0.606	0.336
Min_20% Prob*		N/A	0.335	0.619	0.345
Aug_KL		0.586	0.535	0.483	0.504
Max_Prob_Gap		0.602	0.516	0.639	0.637
ModRényi*	$\alpha = 0.5$	N/A	0.528	0.658	<u>0.681</u>
	$\alpha = 1$	N/A	0.379	0.656	0.513
	$\alpha = 2$	N/A	0.528	0.659	0.680
Rényi ($\alpha = 0.5$)	Max_0%	0.559	<u>0.647</u>	0.656	0.648
	Max_10%	0.561	<u>0.647</u>	0.659	0.675
	Max_100%	**0.711**	**0.685**	**0.687**	**0.695**

Table 4: **Image MIA on GPT-4**.

Metric		VL-MIA/ DALL-E	VL-MIA/ Flickr
Perplexity/zlib*		<u>0.807</u>	0.520
Max_Prob_Gap		0.516	0.486
Rényi ($\alpha = 0.5$)	Max_0%	0.697	0.571
	Max_10%	0.749	0.604
	Max_100%	**0.815**	<u>0.605</u>
Rényi ($\alpha = 1$)	Max_0%	0.688	0.572
	Max_10%	0.747	0.591
	Max_100%	0.790	**0.630**
Rényi ($\alpha = \infty$)	Max_0%	0.685	0.561
	Max_10%	0.708	0.549
	Max_100%	0.781	0.583

See more experiments in the paper.

Future work

- We would like to extend the method to a broader class of multimodal models that incorporate speech or video modalities.
- Our proposed method is semi-black-box, and requires the full probability distribution of the next token prediction. We would like to tackle the case where more or fewer internal workings of VLLMs are available.

References

[1] Reza Shokri et al. "Membership inference attacks against machine learning models". In: *2017 IEEE symposium on security and privacy (SP)*. IEEE. 2017, pp. 3–18.

[2] Deyao Zhu et al. "MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models". In: *The Twelfth International Conference on Learning Representations*. 2023.

[3] Haotian Liu et al. "Visual instruction tuning". In: *Advances in neural information processing systems* 36 (2023).

[4] Peng Gao et al. "LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model". In: *arXiv preprint arXiv:2304.15010* (2023).

[5] Alfréd Rényi. "On measures of entropy and information". In: *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*. Vol. 4. University of California Press. 1961, pp. 547–562.

[6] Weijia Shi et al. "Detecting Pretraining Data from Large Language Models". In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=zWqr3MQuNs.

Acknowledgements