Variance-Based Membership Inference Attacks Against Large-Scale Image Captioning Models

Supplementary Material

A. Evaluation of public and fine-tuned models

Table 1 presents the evaluation results of attack performance on the public BLIP model fine-tuned by us on the Textcaps and Flickr30K datasets after 5 and 10 fine-tuning epochs.

Overall, our proposed methods, MVTA and C-WSA, consistently outperform the baseline CSA and WSA attacks across key metrics, highlighting their robustness and precision in distinguishing member from non-member data.

Furthermore, the results reveal a clear trend: increasing the number of fine-tuning epochs enhances the effectiveness of all attack methods. Specifically, as the target model is trained for more epochs, it memorizes more of the training dataset, making it more susceptible to our attacks.

Table 1. Evaluation of attack performance on the publicly available BLIP model fine-tuned by us on the Textcaps and Flickr30K datasets after 5 and 10 epochs.

Dataset [Model]	Textcaps [BLIP]					
Epochs	5 Epochs			10 Epochs		
Method	AUC	ACC	TPR@FPR=1%	AUC	ACC	TPR@FPR=1%
CSA	0.5474	0.5373	**0.0401**	0.5565	0.5424	**0.0417**
MVTA	**0.6120**	**0.5826**	0.0250	**0.6482**	**0.6080**	0.0346
WSA	0.8483	0.8255	0.4721	0.8441	0.8182	0.4478
C-WSA	**0.9363**	**0.8820**	**0.5269**	**0.9419**	**0.8953**	**0.5560**
Dataset [Model]	Flickr30K [BLIP]					
Epochs	5 Epochs			10 Epochs		
Method	AUC	ACC	TPR@FPR=1%	AUC	ACC	TPR@FPR=1%
CSA	0.6200	0.5872	0.0169	0.6403	0.5996	0.0190
MVTA	**0.6245**	**0.5899**	**0.0196**	**0.6612**	**0.6185**	**0.0242**
WSA	0.9074	0.8513	**0.4893**	0.9173	0.8615	0.4333
C-WSA	**0.9456**	**0.8869**	0.4450	**0.9610**	**0.9050**	**0.5436**

B. Confidence-based training selection

The warm-epoch and confidence threshold were selected based on empirical analysis, with a warm-up epoch of 0.25 (a quarter epoch) and a confidence threshold of 0.4 yielding the best AUC performance, as shown in Figure 1.

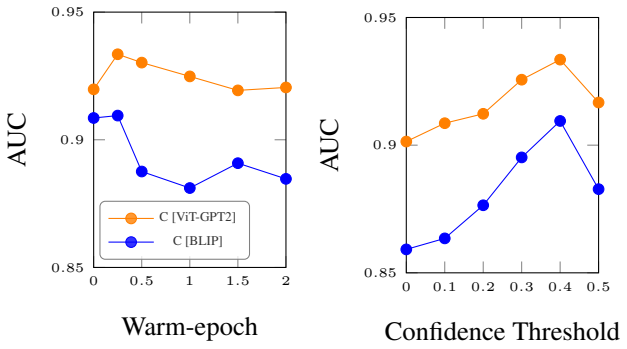

Figure 1. Comparison of AUC metrics for C-WSA configurations across different warm-epochs and confidence thresholds. C refers to MSCOCO.