Differentially Private Representation Learning via Image Captioning

AUTHORS

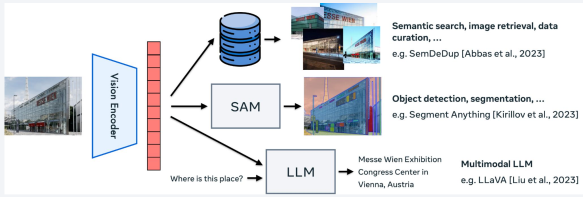
Tom Sander[1,2,*], Yaodong Yu[1,3,*], Maziar Sanjabi[1], Alain Durmus[2], Yi Ma[3], Kamalika Chaudhuri[1], Chuan Guo[1]

AFFILIATIONS

(1) Meta
(2) Ecole polytechnique, Paris
(3) BAIR, UC Berkeley
* Equal Contribution

Context: Foundation models can heavily memorize their pre-training data!

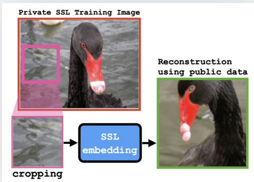
Foundation models learn transferable representations that are useful for various modern AI tasks



However, training foundation models on web-scrawled data comes with risks:

- Unintended memorization
- Copyright issues
- Privacy issues (e.g. PIIs leaks)

Somepalli et al, 2023 [1]



Meehan et al, 2023 [2]

Mitigation: Differentially private representation learning

Differential Privacy (DP) provably guarantees that the model memorizes at most ε nats of information from each training sample

Traditionally DP representation learning is very hard

– ViP [3] made strides and trained a masked autoencoder with DP-SGD on a 233M subset of LAION-2B, obtaining image representations under ε = 8 that are comparable to AlexNet, but is still far from SOTA non-private performance

Can we achieve better DP representation through multimodal supervision?

TL;DR A Differentially Private Captioner with better image representations than regular MAE!

Method: Captioning, Predicting captions from images

- Hypothesis: More efficient information extraction due to concise text captions providing better supervision than image-only SSL
- DP-Cap trained on the same dataset significantly outperforms previous SOTA, demonstrating the effectiveness of the captioning approach under DP constraints

1) MAE training of the image encoder on a synthetic dataset similar to ViP

2) Joint optimization of the image encoder and text decoder with transformer architectures, on a 233M filtered and deduplicated subset of LAION-2B

$$L_{\mathrm{Cap}}([\mathbf{x}^{\mathrm{img}}, \mathbf{z}^{\mathrm{text}}]; \boldsymbol{\theta}) := \frac{1}{T} \sum_{t=0}^{T-1} \ell_{\mathrm{CE}}\Big(z_{t+1}^{\mathrm{text}}, \varphi\big(\underbrace{\psi(\mathbf{x}^{\mathrm{img}}; \boldsymbol{\theta}_{\mathbf{enc}})}_{\text{image embedding}}, \underbrace{z_1^{\mathrm{text}}, \ldots, z_t^{\mathrm{text}}}_{\text{first } t \text{ tokens}}; \boldsymbol{\theta}_{\mathbf{dec}}\big)\Big)$$

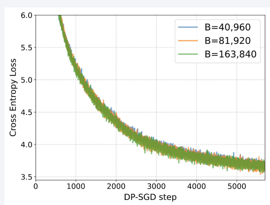
We apply DP-SGD to this loss and obtain privacy guarantees (i.e. ε) w.r.t. the 233M LAION-2B subset

Noisy DP-SGD update: $\widetilde{\mathbf{g}}_k := \frac{1}{B}\left[\sum_{i \in \mathcal{B}_k} \mathrm{clip}_C\left(\nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}_k)\right) + \mathcal{N}\left(0, C^2\sigma^2\mathbf{I}\right)\right]$

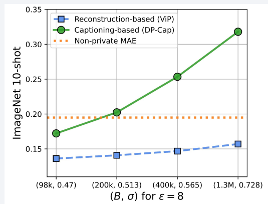
Under the hood: Captioning handles 1.3M batch sizes!

Small ε ↔ low memorization can only be achieved for σ>0.5

But good performance necessitates a low effective noise σ/B. Huge B is thus the only solution. Good news: DP-Cap scales gracefully with batch size!

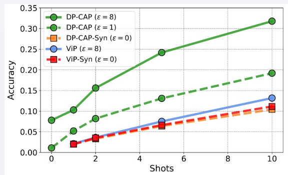


(a) Constant performance at fixed effective noise σ/B and S.

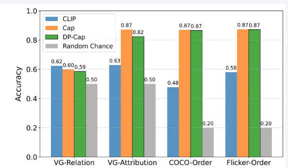


(b) Thanks to (a), we can reach B=1.3M. Achieving low effective noise so great performance and strong guarantees!

Results



(a). Few-shot learning on ImageNet



(b). ARO Evaluation

(c). Captions generated by DP-Cap

Results overview: Linear probing, ARO benchmark, Captioning

Model	pretraining data	DP?	ImageNet-1K	Places-365	Places-205	iNat-2021
DP-NFNet	ImageNet-1K	✓	45.3%	40.1%	39.2%	28.2%
TAN	ImageNet-1K	✓	49.0%	40.5%	38.2%	31.7%
AlexNet	ImageNet-1K	✗	56.5%	39.8%	35.1%	23.7%
SimCLR	ImageNet-1K	✗	67.5%	46.8%	49.3%	34.8%
Cap	Dedup-LAION-233M	✗	77.5%	56.3%	63.9%	63.9%
MAE	Dedup-LAION-233M	✗	62.5%	51.0%	54.7%	42.3%
ViP	Dedup-LAION-233M	✓	56.5%	47.7%	49.6%	38.2%
DP-Cap	Dedup-LAION-233M	✓	63.4%	51.9%	54.3%	44.5%

Under ε=8, DP-Cap achieves 63.4% linear probing accuracy on ImageNet, better than non-private MAE trained on the same dataset!

Model	Config	# parameters	ImageNet-1K (Vision)					ARO (Vision-Language)			
			1-shot	2-shot	5-shot	10-shot	LP	VGR	VGA	COCO	Flickr
ViP	Base	86.6M	2.5%	4.2%	8.5%	14.3%	56.5%	/	/	/	/
DP-Cap	Tiny	22.0M	7.9%	12.1%	18.7%	25.2%	57.5%	58.6%	79.1%	85.7%	87.1%
DP-Cap	Small	49.0M	9.0%	14.0%	21.6%	28.9%	61.1%	59.1%	80.5%	86.0%	86.6%
DP-Cap	Base	86.6M	10.3%	15.6%	24.2%	31.8%	63.4%	58.6%	82.4%	86.6%	87.2%
DP-Cap	Large	407.3M	11.8%	17.5%	26.2%	34.0%	65.8%	59.5%	80.1%	86.6%	86.5%

DP-Cap scales well with model size! DP-Cap-Large reaches 65.8%

References

(1) Gowthami Somepalli and Vasu Singla and Micah Goldblum and Jonas Geiping and Tom Goldstein. "Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models", https://arxiv.org/abs/2212.03860
(2) Casey Meehan, Florian Bordes, Pascal Vincent, Kamalika Chaudhuri and Chuan Guo. "Do SSL Models Have Déjà Vu? A Case of Unintended Memorization in Self-supervised Learning", https://arxiv.org/abs/2304.13850
(3) Yaodong Yu and Maziar Sanjabi and Yi Ma and Kamalika Chaudhuri and Chuan Guo. "ViP: A Differentially Private Foundation Model for Computer Vision", https://arxiv.org/abs/2306.08842