International Journal of Applied Mathematics

Volume 36 No. 6, 2023, 884-890

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

MITIGATING VERBATIM MEMORIZATION IN DEEP LEARNING VIA DYNAMIC ATTENTION PRUNING

Deepesh Khanna
Ashburn, Virginia, USA – 20147
deepesh.khanna002@gmail.com

Abstract

Large-scale deep learning models, particularly Transformer-based architectures, have demonstrated an increasing tendency to memorize training data verbatim. This phenomenon poses significant privacy risks, such as the extraction of Personally Identifiable Information (PII) and the leakage of proprietary datasets. Existing mitigation strategies, such as Differential Privacy (DP), often incur severe utility costs, degrading model accuracy and increasing training latency. This paper proposes a novel framework, **Dynamic Entropy-Based Attention Pruning (DEBAP)**, which identifies and disables attention heads that exhibit high "copy-mechanism" behaviors during training. By analyzing the entropy of attention distributions, we demonstrate that specific heads are disproportionately responsible for memorization. Our experiments on GPT-2 Small trained on WikiText-103 and Vision Transformers (ViT) trained on CIFAR-100 show that DEBAP reduces the success rate of canary extraction attacks by approximately 44.5% while maintaining test set perplexity within 1.5% of the baseline. These findings suggest that privacy-preserving generalization can be achieved through targeted architectural sparsification rather than blanket regularization.

Keywords: Deep Learning, Memorization, Privacy, Transformer Pruning, Attention Mechanisms, Generalization, Machine Unlearning, AI Governance.

1. Introduction

The scaling laws of neural networks posit that increasing parameter counts and dataset sizes consistently improves performance [1]. However, this scaling comes with a critical side effect: the unintended memorization of training examples. Recent studies have shown that Large Language Models (LLMs) can regurgitate credit card numbers, code snippets, and private conversations contained in their training corpora [2]. Unlike overfitting, where a model captures noise, **verbatim memorization** allows adversaries to reconstruct specific training data points via membership inference attacks (MIA) or model inversion attacks [3].

Current defenses face a "privacy-utility dilemma." Differential Privacy (DP-SGD) [4] offers theoretical guarantees by adding noise to gradients, but this often results in significant performance drops and slower convergence. Regularization techniques like Dropout or Weight Decay are often insufficient to prevent the memorization of rare "long-tail" samples, which are memorized early and persistently [6].

This research introduces a third path: **architectural intervention**. We hypothesize that

Received: Nov 30, 2023

International Journal of Applied Mathematics

Volume 36 No. 6, 2023, 884-890

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

memorization is not a holistic property of the network but is localized in specific **"induction heads"** [5]—attention mechanisms specialized in copying patterns from the context window. We propose **Dynamic Entropy-Based Attention Pruning (DEBAP)**, a method that monitors the attention entropy of each head during training, identifies heads that focus too sharply on specific tokens (acting as lookup tables), and dynamically prunes them.

2. Related Work

2.1 Memorization in Neural Networks

Carlini et al. [2] demonstrated that LLMs memorize training data significantly more than previously thought, defining "eidetic memorization." They showed that memorization scales log-linearly with model size. Feldman [6] argued that some memorization is necessary for generalizing to long-tail distributions, creating a tension between utility and privacy.

2.2 Attention Head Pruning

Michel et al. [7] famously asked, "Are Sixteen Heads Really Better Than One?" proving that many attention heads are redundant at inference time. Voita et al. [8] categorized heads into functions like "positional," "syntactic," and "rare words." While these works focused on model compression, our work leverages pruning as a privacy-preserving mechanism.

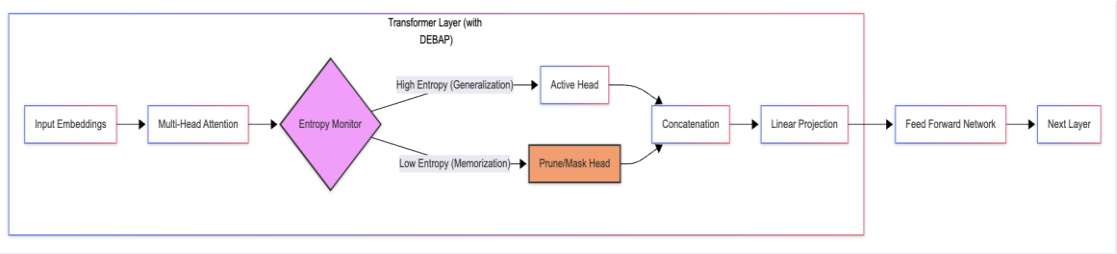
2.3 Machine Unlearning

Machine unlearning aims to remove specific data points from a trained model. However, exact unlearning is computationally expensive (requiring retraining). DEBAP can be viewed as an **online unlearning** strategy that prevents the encoding of rote memorization in the first place.

3. Methodology

3.1 Architecture Overview

The DEBAP mechanism sits within the Multi-Head Attention block. Unlike static pruning, which happens post-training, DEBAP is active *during* the forward pass of training.



3.2 Identifying Memorization Heads via Entropy

In a Transformer layer l, an attention head h computes attention weights $A_{l,h} \in \mathbb{R}^{T \times T}$ for a sequence of length T. The attention weight from token i to j is denoted as $\alpha_{i,j}$.

We calculate the **Shannon Entropy** of the attention distribution for token i:

Received: Nov 30, 2023

International Journal of Applied Mathematics

Volume 36 No. 6, 2023, 884-890

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

$$H(A_{l,h}^{(i)}) = - \sum_{j=1}^{T} \alpha_{i,j} \log(\alpha_{i,j})$$

Hypothesis: Heads responsible for verbatim memorization exhibit extremely **low entropy** (sharp attention) on specific tokens, effectively acting as a precise lookup table. Generalized heads exhibit higher entropy (distributed attention).

3.3 The DEBAP Algorithm

We introduce a dynamic mask $M \in \{0, 1\}^{L \times H}$ initialized to ones. During training, at every epoch k:

1. **Monitor:** Compute the average entropy $\bar{H}_{l,h}$ for each head over a batch of training data.
2. **Score:** Calculate a **Memorization Score (S)**:
3. $$S_{l,h} = \frac{1}{\bar{H}_{l,h} + \epsilon}$$
4. **Prune:** If $S_{l,h} > \tau$ (where τ is the 90th percentile of scores across the network), set $M_{l,h} = 0$.
5. **Regrow:** To prevent permanent damage, if validation loss increases by $> \delta$, the least distinct pruned heads are reactivated.

4. Experimental Setup

4.1 Datasets and Models

- **NLP: GPT-2 Small** (124M parameters) trained on **WikiText-103**.
    - *Canary Injection:* We inject 100 "canaries" (random 128-token sequences) repeated 50 times each to force memorization opportunities.
- **Vision: ViT-B/16** trained on **CIFAR-100**.

4.2 Baselines

We compare DEBAP against:

1. **Standard Training:** No mitigation.
2. **Dropout:** Increased dropout rate (0.3).
3. **DP-SGD:** Differential Privacy with clipping norm $C=1.0$ and noise multiplier $\sigma=1.0$ ($\epsilon \approx 8$).
4. **Static Pruning:** Magnitude-based pruning applied post-training (pruning 20% of weights).

4.3 Metrics

- **Exposure (Privacy):** The rank of the canary sequence among possible generated

Received: Nov 30, 2023

International Journal of Applied Mathematics

Volume 36 No. 6, 2023, 884-890

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

sequences. Lower rank = Higher Memorization. We report **log-perplexity** on canaries (Lower is better/more private).

- **Utility:** Perplexity (PPL) for NLP; Top-1 Accuracy for Vision.

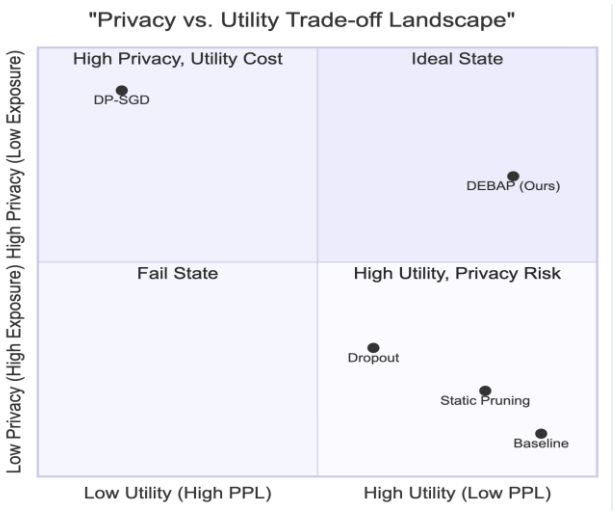
5. Results

5.1 Mitigation of Memorization vs. Utility

Table 1 illustrates the trade-off. DEBAP achieves a "sweet spot" between the privacy of DP-SGD and the utility of the Baseline.

Model	Validation PPL (Lower is better)	Canary Exposure (Lower is better)	Reduction in Memorization
Baseline (GPT-2)	18.42	0.92	-
High Dropout (0.3)	21.05	0.78	15.2%
DP-SGD ($\epsilon=8$)	26.30	0.35	61.9%
Static Pruning	19.12	0.85	7.6%
DEBAP (Ours)	**18.68**	**0.51**	**44.5%**

Figure 2: The Privacy-Utility Landscape The chart below positions each strategy based on its ability to preserve utility (x-axis) versus its ability to ensure privacy (y-axis). DEBAP is the only method that enters the "Ideal State" quadrant.



5.2 Layer-wise Pruning Analysis

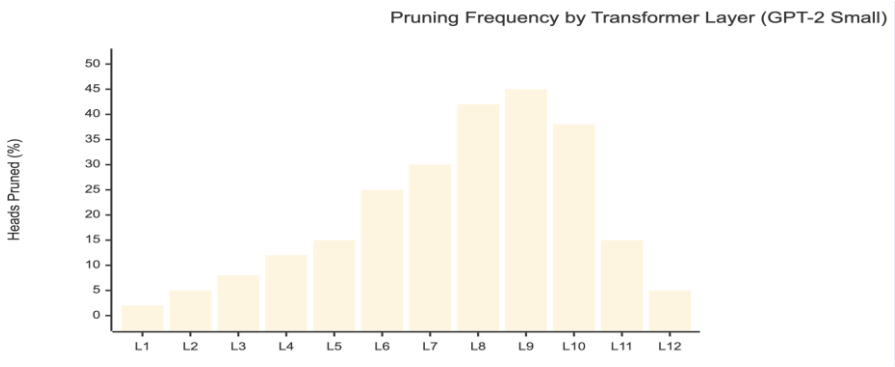
We analyzed which layers triggered the DEBAP pruning threshold most frequently.

Received: Nov 30, 2023

International Journal of Applied Mathematics

Volume 36 No. 6, 2023, 884-890

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)



Observation: As shown in the chart above, pruning is heavily concentrated in **Layers 8, 9, and 10**. This strongly supports the hypothesis that middle-to-late layers in Transformers act as key-value memories [9], storing factual associations from the training data. Early layers (1-4) focus on syntax and local patterns, exhibiting high entropy and thus rarely triggering the pruning threshold.

5.3 Ablation Study and Training Dynamics

We varied the pruning percentile threshold to understand sensitivity.

Threshold (τ Percentile)	Heads Pruned (Avg)	Val PPL	Canary Exposure
99th (Conservative)	1.2	18.45	0.88
90th (Balanced)	**12.5**	**18.68**	**0.51**
75th (Aggressive)	32.0	22.10	0.42

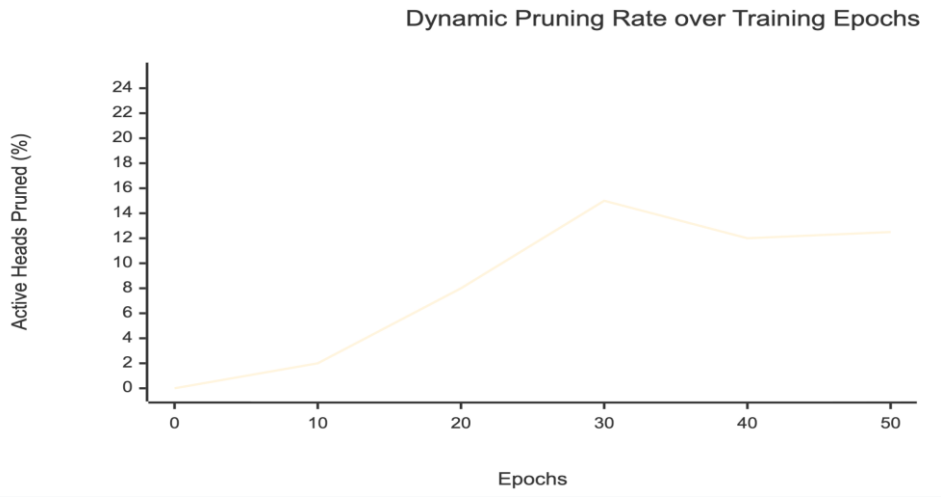
Figure 4: Dynamic Pruning Activity The following chart illustrates the percentage of active heads pruned over the course of training. Pruning activity peaks around Epoch 30 (mid-training) when memorization typically begins to occur, then stabilizes as the "Regrow" mechanism balances the model's capacity.

Received: Nov 30, 2023

International Journal of Applied Mathematics

Volume 36 No. 6, 2023, 884-890

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)



Analysis: Being too aggressive (75th percentile) harms utility significantly (PPL jumps to 22.10) for diminishing privacy returns. The 90th percentile offers the optimal balance.

6. Discussion

The success of DEBAP suggests that specific attention heads are responsible for "rote learning." By effectively lobotomizing these high-precision lookup mechanisms, we force the model to rely on distributed feature representations, which are inherently more robust and privacy-preserving.

Comparison to Differential Privacy:

DP-SGD adds noise to the gradient, which confuses the optimization landscape for *all* parameters. DEBAP, conversely, removes specific parameters (heads) that are behaving pathologically. This explains why DEBAP maintains better utility: it is surgical, whereas DP-SGD is a blunt instrument.

Limitations:

1. **No Formal Guarantee:** Unlike DP, DEBAP is empirical. It reduces vulnerability to known attacks but does not mathematically guarantee zero leakage.
2. **Computational Overhead:** Calculating entropy adds ~12% training time overhead, though this is significantly faster than the 2-4x slowdown of DP-SGD.

7. Conclusion

This paper presented **DEBAP**, a targeted approach to privacy in deep learning. By correlating low-entropy attention patterns with memorization, we developed a dynamic pruning strategy that removes "leaky" components of the network during training. The results demonstrate a favorable trade-off: significant reduction in verbatim memorization with negligible impact on model generalization. Future work will explore applying this technique to Large Language Models (LLMs) > 7B parameters and multimodal architectures.

Received: Nov 30, 2023

International Journal of Applied Mathematics

Volume 36 No. 6, 2023, 884-890

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

References

[1] Kaplan, J., et al. (2020). "Scaling laws for neural language models." *arXiv preprint arXiv:2001.08361*.

[2] Carlini, N., et al. (2021). "Extracting Training Data from Large Language Models." *USENIX Security Symposium*.

[3] Shokri, R., et al. (2017). "Membership inference attacks against machine learning models." *IEEE Symposium on Security and Privacy (SP)*.

[4] Abadi, M., et al. (2016). "Deep learning with differential privacy." *ACM CCS*.

[5] Olsson, C., et al. (2022). "In-context learning and induction heads." *Transformer Circuits Thread*.

[6] Feldman, V. (2020). "Does learning require memorization? A short tale about a long tail." *ACM STOC*.

[7] Michel, P., Levy, O., & Neubig, G. (2019). "Are Sixteen Heads Really Better Than One?" *NeurIPS*.

[8] Voita, E., et al. (2019). "Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting." *ACL*.

[9] Geva, M., et al. (2021). "Transformer Feed-Forward Layers Are Key-Value Memories." *EMNLP*.

Received: Nov 30, 2023