From Ensemble and DistilBERT Baselines to Deep Metric Learning: An Extended Study for Imbalanced Windows PE Malware Detection

Ly Ngoc Vu
Industrial University of Ho Chi Minh City
Ho Chi Minh City, Vietnam
dezzhuge@gmail.com

Dang Thi Phuc
Industrial University of Ho Chi Minh City
Ho Chi Minh City, Vietnam
phucdt@iuh.edu.vn

***Abstract*—Windows PE malware detection requires high recall under strict false-positive constraints, especially for imbalanced real-world data. This paper extends a prior ATC 2024 study that evaluated classical machine learning and text-based deep learning baselines. The extension introduces a deep metric learning (DML) branch on a shared Transformer backbone and compares five variants: baseline classification, ArcFace, Contrastive, Triplet, and Multi-Similarity. Evaluation uses multi-seed reporting and deployment-oriented metrics, including TPR@FPR operating points. Experimental results show that DML is beneficial but objective-dependent: Multi-Similarity provides the strongest overall operating profile, while ArcFace is unstable in strict low-FPR conditions despite high ranking metrics. The study provides practical guidance for objective selection in operational malware detection pipelines.**

***Index Terms*—malware detection, PE files, deep metric learning, Transformer, low-FPR evaluation**

I. Introduction

Malware detection on Windows PE files remains a high-impact cybersecurity task under evolving adversarial behavior and class-imbalanced data distributions [6], [7]. Although modern classifiers often report high aggregate accuracy, real deployments are constrained by strict false-positive-rate (FPR) budgets, where threshold behavior is more informative than global metrics [9], [10]. Prior results in the same project line showed that both optimized machine learning and text-based deep learning are strong baselines for PE classification [1].

This paper extends that baseline study with deep metric learning (DML), motivated by the need for better class separation in highly imbalanced settings [2]–[5]. The central question is not whether DML can improve ranking metrics alone, but whether it improves low-FPR operational behavior.

The main contributions are:

1) A controlled objective-level comparison of baseline, ArcFace, Contrastive, Triplet, and Multi-Similarity on a shared Transformer backbone.
2) Multi-seed evaluation with deployment-oriented metrics, including TPR@FPR targets.
3) An operational analysis showing that objective choice materially affects low-FPR reliability.

The remainder of this paper is organized as follows. Section II positions the extension relative to prior work. Section III summarizes the methodology and system design. Section IV presents experiments and results. Section V discusses implications and validity threats. Section VI concludes and outlines future work.

II. Related Work and Extension Scope

A. Prior ATC 2024 Results

The prior paper reported:

1) Classical ML models: Logistic Regression, Random Forest, SVC, XGBoost, plus ensemble voting/stacking.
2) Text-based deep learning models: LSTM, BiLSTM, DistilBERT.
3) Large PE dataset setting (˜34k samples) with high performance.

B. Extension in This Paper

Recent metric-learning objectives, including ArcFace, Triplet, Contrastive, and Multi-Similarity, are widely adopted for discriminative embedding learning [2]–[5], [8]. However, comparative evidence in PE malware settings under strict operating points is limited.

This work extends prior results by adding:

1) Objective-level study of DML variants.
2) Multi-seed statistical evaluation.
3) Deployment-oriented low-FPR analysis.

III. Methodology

A shared Transformer backbone is used, while objectives/heads vary:

1) baseline: cross-entropy/focal classification.
2) arcface: angular margin objective.
3) contrastive: pairwise distance objective + classification.
4) triplet: relative distance objective with hard mining.
5) multi_similarity: weighted hard-pair similarity objective.

A. Experimental Setup

The dataset setting follows the prior project pipeline and preserves the same PE text-feature processing strategy [1]. The extended pipeline evaluates five variants through a single runner and stores per-variant outputs for reproducibility. Model quality is assessed by Accuracy, Precision, Recall, F1, ROC-AUC, and PR-AUC. Operational reliability is measured with TPR@FPR points to reflect defender-oriented deployment constraints.

B. LVModel Architecture Details

The base LVModel is a Transformer encoder classifier implemented in Flax/JAX. Let input token IDs be $\mathbf{X} \in \mathbb{N}^{B\times T}$, where B is batch size and T is sequence length (capped by $T \leq 380$ in the current setup).

Embedding and positional encoding:

$$\mathbf{H}_0 = E_{\text{tok}}(\mathbf{X}) + E_{\text{pos}}(1{:}T), \tag{1}$$

where both embedding tables use normal initialization ($\sigma = 0.02$).

Shared MHA block implementation: for each encoder layer, a single dense projection computes concatenated QKV:

$$\text{QKV} = W_{qkv}\mathbf{H} \in \mathbb{R}^{B\times T\times 3d}, \tag{2}$$

then reshaped to $(3, B, h, T, d_h)$ where $d_h = d/h$ and h is head count. Attention logits use scaled dot-product:

$$\mathbf{S} = \frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}}, \quad \mathbf{A} = \text{softmax}(\mathbf{S}), \tag{3}$$

followed by attention dropout and output projection back to d dimensions.

Encoder block design (pre-norm): each layer applies LayerNorm before attention and feed-forward modules:

$$\tilde{\mathbf{H}} = \text{LN}(\mathbf{H}_{l-1}), \tag{4}$$

$$\mathbf{A} = \text{MHA}(\tilde{\mathbf{H}}), \quad \mathbf{H}' = \mathbf{H}_{l-1} + \mathbf{A}, \tag{5}$$

$$\mathbf{F} = W_2\,\text{Dropout}(\text{GELU}(W_1\,\text{LN}(\mathbf{H}'))), \tag{6}$$

$$\mathbf{H}_l = \mathbf{H}' + \text{Dropout}(\mathbf{F}). \tag{7}$$

This is a **pre-norm Transformer** with two residual paths (attention and FFN), improving optimization stability.

Pooling and classification: the sequence representation is mean pooled and passed through a tanh pooler and LayerNorm classifier:

$$\mathbf{z} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{H}_{L,t}, \tag{8}$$

$$\mathbf{p} = \tanh(W_p\mathbf{z}), \tag{9}$$

$$\mathbf{y} = W_c\,\text{LN}(\text{Dropout}(\mathbf{p})). \tag{10}$$

In experiments, the instantiated architecture uses $d_{\text{model}} = 256$, $h = 8$, $d_{\text{ff}} = 512$, and $L = 2$.

Metric-learning variants: ArcFace, Contrastive, Triplet, and Multi-Similarity share the same encoder trunk, then add an embedding projection ($d_{\text{emb}} = 256$), LayerNorm, and ℓ_2 normalization. The projected embedding is consumed by a task-specific head: ArcFace angular-margin classifier, or linear logits with auxiliary metric losses (contrastive/triplet/MS).

C. Pipeline Overview

Figure 1 summarizes the end-to-end training and evaluation flow used for baseline and DML variants.

IV. Results

A. Prior ATC 2024 Baselines

TABLE I
ML BASELINES FROM PRIOR PAPER

Model	Precision	Recall	F1	Accuracy
Logistic Regression	0.990086	0.990232	0.990111	0.990112
Random Forest	0.990000	0.990382	0.990402	0.990403
SVC	0.990000	0.988060	0.988075	0.988076
XGBoost	0.990000	0.991542	0.991566	0.991566

TABLE II
DEEP LEARNING BASELINES FROM PRIOR PAPER

Model	Precision	Recall	F1	Accuracy
DistilBERT	0.9864895	0.9865220	0.9864705	0.986471
LSTM	0.9674135	0.9674490	0.9674130	0.967413
BiLSTM	0.9507005	0.9499545	0.9500705	0.950102

B. New DML Results

TABLE III
AGGREGATED 5-SEED METRICS (MEAN)

Variant	Accuracy	F1	ROC-AUC	PR-AUC	TPR@FPR=1e-2
baseline	0.9924	0.9960	0.9983	0.9999	0.9754
arcface	0.7979	0.7934	0.9704	0.9984	0.0000
contrastive	0.9931	0.9964	0.9971	0.9997	0.9533
triplet	0.9932	0.9964	0.9969	0.9998	0.9351
multi_similarity	**0.9946**	**0.9972**	0.9978	0.9999	**0.9851**

V. Discussion

The results indicate that DML is helpful but not uniformly so. Multi-Similarity yields the strongest overall profile, improving low-FPR sensitivity while maintaining top-tier F1 and accuracy. Baseline remains highly competitive, indicating a strong representation backbone and training recipe. ArcFace shows a notable mismatch between ranking quality and low-FPR behavior, suggesting sensitivity to margin-scale calibration under class imbalance.

Main findings:

1) DML helps, but improvements are objective-dependent.
2) Multi-Similarity is strongest overall in this setup.
3) Baseline remains highly competitive.
4) ArcFace requires further calibration/tuning for strict low-FPR deployment.

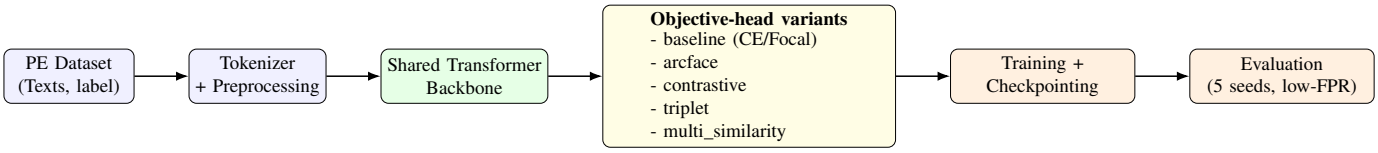

Fig. 1. Unified pipeline for baseline and DML variants.

A. Threats to Validity

1) **Single-domain evaluation**: results are derived from one PE data pipeline; broader cross-dataset validation is needed.
2) **Seed count**: five-seed reporting improves stability estimates but does not eliminate uncertainty.
3) **Calibration sensitivity**: thresholded performance may shift under temporal drift and distribution changes.
4) **Static-feature bias**: packed or heavily obfuscated binaries may require additional dynamic/contextual signals.

VI. Conclusion

This extended study confirms that text-based PE malware detection can achieve high performance and that DML can further improve deployment-oriented behavior when the objective is carefully selected. Multi-Similarity is the recommended candidate in the current experiments.

Future Work

Planned next steps include explicit calibration studies, statistical significance testing across objectives, cross-dataset generalization experiments, and robustness analysis under packing/obfuscation and temporal drift.

Acknowledgment

This work extends the authors' prior ATC 2024 study and integrates new metric-learning experiments and analysis.

References

[1] L. N. Vu and D. T. Phuc, "Windows Malware Detection: Exploring from Machine Learning to Text-Based Deep Learning Approaches," in Proc. ATC, 2024.
[2] J. Deng et al., "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," CVPR, 2019.
[3] X. Wang et al., "Multi-Similarity Loss with General Pair Weighting for Deep Metric Learning," CVPR, 2019.
[4] F. Schroff et al., "FaceNet: A Unified Embedding for Face Recognition and Clustering," CVPR, 2015.
[5] R. Hadsell et al., "Dimensionality Reduction by Learning an Invariant Mapping," CVPR, 2006.
[6] A. Harang and E. M. Rudd, "SOREL-20M: A Large Scale Benchmark Dataset for Malicious PE Detection," arXiv:2012.07634, 2020.
[7] BODMAS Dataset, "Benchmark for malware analysis at scale," 2021.
[8] W. Zheng, J.-H. Lai, and P. C. Yuen, "A Survey on Deep Metric Learning: Approaches and Applications," arXiv:2303.15032, 2023.
[9] C. Guo et al., "On Calibration of Modern Neural Networks," ICML, 2017.
[10] S. Wang et al., "A Comprehensive Survey of Out-of-Distribution Detection Methods for AI Security," ACM Comput. Surv., 2024.