Supplementary Materials - *DynaMO*: Protecting Mobile DL Models through Coupling Obfuscated DL Operators

ANONYMOUS AUTHOR(S)

1 PROOF

Lemma 1.1. ***(Coupled Weight Transformation on Linear Model)***: *A sub-network f consists of multiple linear layers $\{L_1, L_2, \cdots, L_n\}$ and the i-th layer is $L_i : X_i = W_{i-1}^\top X_{i-1} + b_{i-1}$ where $i \in [1, n]$. The output of the sub-network f w.r.t. to the input X_0 would be $f(X_0)$. If W_1 is transformed to aW_1, W_n is transformed to $\frac{1}{a}W_n$, and b_i is transformed to ab_i for $i \in [1, n-1]$, then the transformed network f^s w.r.t. to the input X_0 would be $f^s(X_0)$ and we have $f^s(X_0) = f(X_0)$.*

Proof. Let us denote the scaled layer as L^s. For L_k^s where $k \in [1, n-1]$, we have

$$\begin{aligned}
X_1^s &= aW_1^\top X_0 + ab_0 = aX_1 \\
X_2^s &= W_2^\top X_1^s + ab_1 = aW_2^\top X_1 + ab_1 = aX_2 \\
&\cdots \\
X_{n-1}^s &= W_{n-1}^\top X_{n-2}^s + ab_{n-1} = aW_{n-1}^\top X_{n-2} + +ab_{n-1} = aX_{n-1}.
\end{aligned} \tag{1}$$

Then for L_n^s, we have $X_n^s = \frac{1}{a} W_n^\top X_{n-1}^s + ab_n = \frac{1}{a} W_n^\top (aX_{n-1}) + b_n = X_n$. Therefore, this gives us $f^s(X_0) = f(X_0)$. □

Theorem 1.2. ***(Coupled Weights Obfuscation)***: *We consider a general non-linear layer $ReLU_\beta (\beta \geq 0)$*, i.e.,

$$ReLU_\beta(x) = \begin{cases} \beta, & \text{if} \quad x \geq \beta; \\ x, & \text{else if} \quad 0 < x < \beta; \\ 0, & \text{otherwise.} \end{cases}$$

If W_i and b_i in the sub-network f are scaled to aW_i and ab_i for $i \in [1, n]$ with $0 < a < 1$, respectively, then, $ReLU_\beta(\frac{1}{a} I_{n+1}^\top ReLU_\beta(f^s(X_0))) = I_{n+1}^\top ReLU_\beta(f(X_0))$.

Proof. From Equation (1), we can conclude that $f^s(X_0) = aX_n$. Then, for the original sub-network f we have

$$\mathrm{ReLU}_\beta(f(x_0)) = \mathrm{ReLU}_\beta(x_n) = \begin{cases} \beta, & \text{if} \quad x_n \geq \beta; \\ x_n, & \text{else if} \quad 0 < x_n < \beta; \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where x_0 and x_n are elements in X_0 and X_n, respectively. Thus,

$$\mathrm{ReLU}_\beta(f(x_0)) = \begin{cases} \beta, & \text{if} \quad x_n \geq \beta; \\ x_n, & \text{else if} \quad 0 < x_n < \beta; \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Then, for the scaled sub-network f^s we have

$$\mathrm{ReLU}_\beta(f^s(x_0)) = \mathrm{ReLU}_\beta(ax_n) = \begin{cases} \beta, & \text{if} \quad x_n \geq \frac{\beta}{a}; \\ ax_n, & \text{else if} \quad 0 < x_n < \frac{\beta}{a}; \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

$$\frac{1}{a}\mathrm{ReLU}_\beta(ax_n) = \begin{cases} \frac{\beta}{a}, & \text{if} \quad x_n \geq \frac{\beta}{a}; \\ x_n, & \text{else if} \quad 0 < x_n < \frac{\beta}{a}; \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Note that $\frac{\beta}{a} > \beta$, thus,

$$\mathrm{ReLU}_\beta(\frac{1}{a}\mathrm{ReLU}_\beta\big(f^s(x_0)\big)) = \begin{cases} \beta, & \text{if} \quad x_n \geq \beta; \\ x_n, & \text{if} \quad 0 < x_n < \beta; \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Therefore, we can conclude that $\forall x_0 \sim X_0$, the following holds: $\mathrm{ReLU}_\beta(\frac{1}{a}\mathrm{ReLU}_\beta\big(f^s(x_0)\big)) = \mathrm{ReLU}_\beta\big(f(x_0)\big)$. Thus, it follows that: $\mathrm{ReLU}_\beta(\frac{1}{a}\mathrm{ReLU}_\beta\big(f^s(X_0)\big)) = \mathrm{ReLU}_\beta\big(f(X_0)\big)$. Note that since the weights I_{n+1} forms an Identity matrix, the equation $\mathrm{ReLU}_\beta(\frac{1}{a}I_{n+1}^\top\mathrm{ReLU}_\beta\big(f^s(X_0)\big)) = I_{n+1}^\top\mathrm{ReLU}_\beta\big(f(X_0)\big)$ still holds. □

2 PERFORMANCE OF *DYNAMO* USING DIFFERENT NUMBERS OF EXTRA OBFUSCATING OPERATORS

Table 1. Deobfuscation performance using *DLModelExplorer* on the models obfuscated by our proposed *DynaMO*. **The number of extra obfuscating operators is 20.** WER, WEA, OCA, and NIR are the metrics used to measure deobfuscation performance. 'TN': True negative (*i.e.*, correct identification for obfuscating operators) rate of operator classification.

Metric	Fruit	Skin	MobileNet	MNASNet	SqueezeNet	EfficientNet	MiDaS	LeNet	PoseNet	SSD	**Average value**	**Difference**
WER	58.62%	48.28%	57.14%	72.22%	88.89%	70.59%	77.45%	75.00%	65.62%	73.61%	**68.74%**	**30.02%** ↓
WEE	2.74	1.95	0.82	0.52	0.07	0.71	0.39	0.001	0.20	0.05	**0.75**	**0.75** ↑
NIR	60.78%	62.00%	58.82%	81.58%	79.17%	80.52%	82.67%	60.00%	35.87%	86.21%	**68.76%**	**29.63%** ↓
OCA	60.78%	62.00%	60.78%	81.82%	82.00%	80.52%	90.67%	66.67%	67.35%	91.46%	**74.41%**	**25.48%** ↓
TN (OCA)	0%	0%	0%	0%	10.00%	0%	0%	27.27%	0%	0%	**3.73%**	N/A

Table 2. Deobfuscation performance using *DLModelExplorer* on the models obfuscated by our proposed *DynaMO*. **The number of extra obfuscating operators is 10.** WER, WEA, OCA, and NIR are the metrics used to measure deobfuscation performance. 'TN': True negative (*i.e.*, correct identification for obfuscating operators) rate of operator classification.

Metric	Fruit	Skin	MobileNet	MNASNet	SqueezeNet	EfficientNet	MiDaS	LeNet	PoseNet	SSD	**Average value**	**Difference**
WER	75.86%	68.97%	78.57%	74.07%	88.89%	80.39%	86.27%	50.00%	84.38%	73.61%	**76.10%**	**22.66%** ↓
WEE	0.47	0.43	0.11	0.54	0.01	0.56	0.09	0.001	0.25	0.06	**0.25**	**0.25** ↑
NIR	75.61%	75.61%	76.92%	86.11%	88.89%	88.57%	87.94%	60.33%	50.77%	84.27%	**77.50%**	**20.89%** ↓
OCA	75.61%	75.61%	79.49%	86.30%	90.91%	88.57%	96.45%	57.14%	80.49%	91.46%	**82.21%**	**17.68%** ↓
TN (OCA)	0%	0%	0%	0%	0%	0%	0%	28.57%	0%	0%	**2.86%**	N/A