АНАЛИЗ УСТОЙЧИВОСТИ ФИЗИЧЕСКИ НЕКЛОНИРУЕМОЙ ФУНКЦИИ ТИПА АРБИТР К КРИПТОГРАФИЧЕСКИМ АТАКАМ С ИСПОЛЬЗОВАНИЕМ НЕЙРОННЫХ СЕТЕЙ ДОЛГОЙ КРАТКОСРОЧНОЙ ПАМЯТИ

Шинкевич Н.Н., Шамына А.Ю.
Кафедра программного обеспечения информационных технологий,
Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: nn5h@yahoo.com, shamyna@bsuir.by

Рассматриваются способы криптографических атак на классические реализации физически неклонируемых функций типа арбитр (АФНФ) с использованием ИНС выбранной конфигурации, экспериментально подтверждается эффективность выбранного подхода и необходимость использования дополнительных решений для повышения устойчивости к подобного рода атакам. Предложен способ атаки на основе сетей долгой краткосрочной памяти (NNLSTM) на АФНФ, реализованных на FPGA. Экспериментальные результаты предложенного метода NNLSTM показывают его высокую точность прогноза, до 99,96%.

Введение

В настоящее время существует проблема обеспечения цифровой безопасности, а именно: защиты устройств и их проектных описаний от несанкционированного копирования, идентификации устройств и т.д. Использование физически неклонируемых функций (ФНФ) [1] позволяет решить перечисленные выше проблемы с высокими показателями случайности, уникальности и стабильности в зависимости от решаемой задачи, и в то же время с низкими аппаратными и финансовыми затратами. Широкое распространение получили ФНФ типа арбитр (АФНФ) благодаря относительной простоте реализации и сравнительно низкими аппаратными затратами. Однако распространение сигналов по блоку симметричных путей (БСП) АФНФ и, следовательно, значения задержек имеют линейную природу [2], что является потенциальной уязвимостью для криптографических систем, в которых они применяются [3, 4]. В работах [5, 6] описываются способы атак, основанные на применении методов машинного обучения. Самыми эффективными при моделировании классических АФНФ являются: линейная регрессия, классификация при помощи SVM, нейронные сети. В работах [5, 6] рассматриваются атаки с использованием ИНС, которые представляют собой полносвязную структуру из плотных (Dense) слоев. Каждый такой слой обрабатывает данные без учета порядка, применяя к обучающей выборке операцию взвешенной суммы (формула 1). Такой подход является наиболее распространенным [5, 6], предсталенные ИНС отличаются числом плотных слоев, функциями активации, наличием либо отсутствием слоев предобработки данных, но глобально практически идентичны по своей структуре, и представляют собой ничто иное, как модификации полносвязных нейронных сетей. В настоящей статье авторами предлагается другая структура ИНС, учитывающая факт наличия временных задержек. Экспериментально подтверждается эффективность использования предложенной конфигурации ИНС.

I. Описание конфигурации ИНС

Выбранная авторами конфигурация ИНС состоит из слоев следующих типов:

- Embedding, преобразовывает положительные целые числа (индексы) в плотные (dense) векторы фиксированного размера;
- LSTM, особый тип РНС, способный обучаться долговременным зависимостям;
- Bidirectional, или двунаправленный слой, необходим для определения выходной последовательности с учетом всего контекста, а не только его линейной интерпретации;
- Dense, вычисляет f_a от взвешенной суммы (формула 1);
- Dropout, предотвращает переобучение сети путем удаления некоторой доли ответов случайным образом.

$$Y = f_a(X * W + B), x_i \in X, y_i \in Y, \forall i \in N \quad (1)$$

где W, B - матрицы весов и смещений соответственно, f_a - функция активации. Чтобы воспользоваться преимуществами Embedding слоев, а именно возможностью формирования начального представления для каждого i-го элемента j-й последовательности входных данных, необходимо представить их в символьном формате. Далее с данными необходимо произвести следующие манипуляции:

1. балансировка выборки;
2. преобразование бинарной строки challenge по формулам 2, 3.

$$B_i = A, b_i == 0 \quad (2)$$

$$B_i = B, b_i == 1 \quad (3)$$

Схема построенной нейронной сети представлена на рисунке 1.

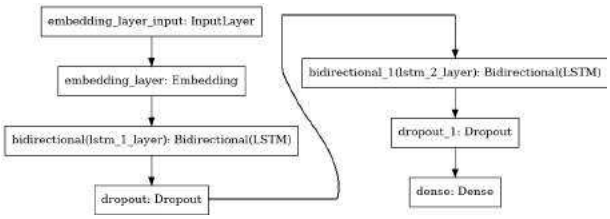

Рис. 1 – Схема модели NN LSTM

II. Проведение эксперимента и сбор данных

Для оценки устойчивости АФНФ различных реализаций к рассматриваемым способам атак было создано несколько наборов экспериментальных данных: собранных как с аппаратной реализации АФНФ на FPGA Artix 7, так и с параметрической модели «Post place & route», созданной с использованием САПР Xilinx 14.7 и HDL языка Verilog, а также средства моделирования ISim. Время обучения SVM и NNLSTM было ограничено 20 эпохами для каждого тестируемого подхода.

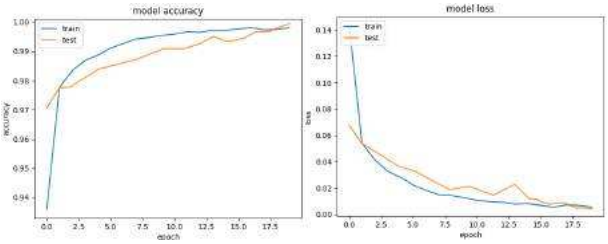

Рис. 2 – Изменение а – точности; б – ошибки модели в процессе обучения

Для достижения точности 99.96% потребовалось 20 эпох. При дообучении модели за 20 дополнительных эпох точность колеблется в пределах от 98.07 до 99.972% на тренировочной выборке, откуда следует, что незначительный прирост точности в 0.012% не оправдывает временные затраты. Процесс обучения, в частности графики изменения точности и ошибки на тренировочной и тестовой выборках представлен на рис. 2а, 2б.

Таблица 1 – Точность моделирования АФНФ на первом наборе данных

	SVM	SVM	NN LSTM	NN LSTM
	64bit	128bit	64bit	128bit
5000	54.67%	49.93%	52.11%	53.16%
10000	56.67%	51.19%	95.01%	98.28%
100000	57.77%	55.52%	99.15%	99.96%

Таблица 2 – Точность моделирования АФНФ на втором наборе данных

	SVM	SVM	NN LSTM	NN LSTM
	64bit	128bit	64bit	128bit
5000	52.92%	31.89%	61.81%	45.93%
10000	53.99%	33.16%	97.89%	97.20%
100000	58.11%	52.08%	99.94%	99.93%

III. Заключение

Исходя из результатов (табл. 1, 2), можно сделать вывод: NNLSTM успешно смоделировала несколько видов АФНФ, более того, сеть смогла найти зависимости в данных, используя выборку из 10^4 примеров, а при наличии 10^5 примеров сеть достигла почти 100%-й точности, чего нельзя сказать о методе SVM [7]; как показала практика, ему недостаточно 20 эпох обучения для достижения приемлемого уровня точности (предел был достигнут за 94 эпохи, точность 97.16% для набора данных из 10^5 примеров). Более того, при моделировании более сложной АФНФ точность метода SVM падала, тогда как на NNLSTM это почти не повлияло (наблюдается лишь малое снижение точности). Стоит отметить, что все выборки являются относительно небольшими по размеру; обычно ИНС требуется 10^6 примеров. Полученные результаты демонстрируют необходимость применения дополнительных решений при использовании АФНФ в качестве криптографического примитива. Например, в работе [8] предлагается использование нелинейной обфускации запросов к АФНФ при помощи MISR.

IV. Список литературы

1. Security with Noisy Data / P. Tuyls, B. Skoric, T. Kevenaar. – Switzerland : Springer, 2007. – 344 p.
2. Physical unclonable functions / Y. Gao, S. F. Al-Sarawi, and D. Abbott, Nature Electron., vol. 3, no. 2, pp. 81–91, Feb. 2020.
3. Modeling attacks on physical unclonable functions / U. Ruhrmair [et al.] // Proc. 17th ACM conf. on Comp. and comm. secur. (CCS'10). – Chicago, USA, 2010. – P. 237–249.
4. PUF modeling attacks on simulated and silicon data / U. Ruhrmair [et al.] // IEEE Trans. on Inform. Forens. and Secur. – 2013. – № 8 (11). – P. 1876–1891.
5. Going Deep: Using deep learning techniques with simplified mathematical models against XOR BR and TBR PUFs / M. Khalafalla, M. A. Elmohr and C. Gebotys, 2020 IEEE International Symposium on Hardware Oriented Security and Trust (HOST), 2020, pp. 80-90.
6. Deep Learning based Model Building Attacks on Arbiter PUF Compositions / Santikellur, P., Bhattacharyay, A., & Chakraborty, R. (2019). IACR Cryptol. ePrint Arch., 2019, 566.
7. Security evaluation and enhancement of bistable ring PUFs / X. Xu [et al.] // Proc. Int. Worksh. on Rad. Freq. Ident.: Secur. and Priv. Iss.(RFIDSec'2015). — New York, USA, 2015. — P. 3–16.
8. S. S. Zalivaka, A. A. Ivaniuk and C. Chang, "Low-cost fortification of arbiter PUF against modeling attack"2017 IEEE International Symposium on Circuits and Systems (ISCAS), 2017, pp. 1-4.