Recenzie Articol
Proiectarea Rețelelor

Grigore Lucian-Florin 343C4
Facultatea de Automatică și Calculatoare
Universitatea Politehnica, București
Ianuarie 2022

Titlu articol: SIRNN: A Math Library for Secure RNN Inference
Link articol: https://arxiv.org/pdf/2105.04236.pdf

Acest articol a fost publicat în cadrul conferinței IEEE Symposium on Security and Privacy, care a avut loc în perioada 24-27 mai 2021, fiind realizat de o echipa de cercetători din cadrul Microsoft Research India. Într-o industrie care orbiteaza din ce în ce mai mult în jurul inteligenței artificiale și a mecanismelor și modelelor consacrate care fac parte din aceasta, munca descrisă în articol este, pe de o parte, simpla si, pe de alta parte, incontestabil de necesara.

Pentru a putea explica cu eficienta conținutul articolului, trebuie oferite descrieri anumitor termeni prezenți în acesta:

- **inferenta**: operația prin care se ajunge de la un set de date la o anumită concluzie
- **calcul/computatie bipartit/a securizat**: protocol de comunicare între doua entități prin care acestea sa calculeze un rezultat comun, nestiind valorile interne ale celeilalte entități; valoarea comuna calculata poate include valori atat comune si deci vizibile, dar si interne, situația în care este necesara pastrarea confidentialitatii acestor variabile
- **entitate semi-onesta**: categorie de entități despre care se presupune ca nu vor încerca sa exploateze slabiciunile unui protocol în scopul obținerii de date ilegal; nivelul de securitate semi-onest este unul naiv, care în situații reale nu este recomandat; în schimb, acesta previne scurgerea accidentală de informații, fiind în același timp, considerabil de eficient
- **DNN**: rețea neurala adanca
- **CNN**: rețea neurala convolutionala
- **RNN**: rețea neurala recurenta
- **LSTM/GRU**: tipuri de celule folosite în RNN
- **ML**: învățare automată

Echipa de cercetători din cadrul Microsoft a observat un deficit de performanță pentru arhitecturile RNN, pentru care nu s-au dezvoltat protocoale de calcul care sa asigure inferenta securizata în același timp cu performanta. În schimb, la polul opus se afla rețele de tipul CNN sau DNN. Acestea din urma au beneficiat în ultima vreme de librării specializate pentru tipurile de calcul specifice fiecăreia. Folosirea în cadrul rețelelor recurente a librăriilor deja existente aduce un cost de performanta considerabil. Astfel s-a ajuns la crearea SIRNN (citită "siren"), care înseamnă "Secure Inference for RNNs".

Inferenta securizata presupune existenta a doua entități semi-oneste: un client si un server. Serverul deține un model ML (în cazul de fata, un model RNN) despre care se dorește ascunderea implementării interne fata de client. Clientul deține un set de date privat. Scopul este ca în urma învățării și a predicțiilor generate de rețea, aceasta sa nu acceseze datele clientului, iar clientul sa nu poată exploata implementarea internă a rețelei mai mult decat se poate deduce din predicțiile oferite. Acest canal de comunicare dintre client și server este realizat printr-un protocol bine structurat, care sa împiedice orice tentativa de hacking din partea unei parti. În cadrul comunicarii, datorită fluxului mare de date partajat, librăriile clasice, chiar dacă respecta normele de securitate, aduc un overhead considerabil procesului de antrenare și predicție a modelelor ML.

SIRNN este prima librărie care oferă inferenta securizata modelelor de rețele recurente fără compromis de performanta. Pentru aceasta, echipa de cercetători a avut în vedere mai multe aspecte.

În primul rand, au folosit noi aproximări pentru funcțiile matematice, fata de cele pre-existente. Rețele recurente au nevoie de operații precum exponențială, sigmoid, tangenta hiperbolica si reciproca radacinii patrate. Folosirea unor tabele de lookup pentru a obține valoarea inițială a funcției, urmată de un algoritm iterativ precum Goldschmidt oferă o aproximare foarte buna, reducând în același timp overhead-ul criptografic. Toate aceste aproximări sunt demonstrabil corecte.

În al doilea rand, s-au integrat protocoale pentru eficientizarea latimii de banda, și anume truncare, extensie, multiplicare și descompunerea cifrelor. Astfel, traficul de date dintre client și server se asigura ca folosește latime de banda exact cat are nevoie. În urma testelor efectuate, s-a observat o utilizare de 423 de ori mai mica a canalelor de comunicație dintre client și server, unde orice mesaj sau informație transmisă reduce performanța.

Implementarea acestor mecanisme în SIRNN o transforma dintr-o librărie eficienta si exacta. Fata de alte modele consacrate care folosesc librării pre-existente, îmbunătățirile consumului de timp și spațiu sunt impresionante. De exemplu, in figura de mai jos (extrasa din articol) se observa ca, folosind arhitectura Google-30, în timpul utilizarii batch-urilor in antrenare și testare, speedup-ul de timp este de peste 3000, iar ca memorie se consuma de aproximativ 1300 ori mai puțină.

Benchmark	Batch	Runtime (sec)		Comm.	
		[41]	SIRNN	[41]	SIRNN
Industrial-72	1	68.33 (18x)	3.7	11.84 GB (510x)	23.8 MB
	128	8746* (661x)	13.2	1.47 TB* (1451x)	1.04 GB
Google-30	1	3337 (67x)	49.6	259 GB (574x)	0.45 GB
	128	$4.3x10^5$* (3050x)	140	32.38 TB* (1316x)	25.2 GB
Heads	1	NA	409.7	NA	85.5 GB

În articol sunt oferite formule explicite și pseudocod pentru cele 4 funcții matematice menționate mai sus, precum și demonstrații ale corectitudinii lor. Mai mult decat atat, este efectuata o analiza de specialitate pentru acestea, comparand rezultatele obtinute cu alte librării deja consacrate în industrie, fata de care îmbunătățirile sunt cel putin solide.

În concluzie, SIRNN este o librărie crucială pentru rețelele recurente in contextul utilizarii bipartite, care facilitează comunicarea într-o maniera evolutionara, fără a sacrifica din normele de securitate necesare.