חלק

מאמר 1מודל CNN (Convolutional neural network)

הקדמה
FlowPic: Encrypted Internet Traffic Classification is as Easy as Image Recognition

המאמר הראשון מציג דרך חדשה לסיווג של תעבורה מוצפנת של מידע באמצעות רשת נוירונים מפותלת.

הרעיון המרכזי המתואר במאמר מציע לבצע העתקה של זרימת מידע ברשת אל תמונות בהתאם לגודל החבילה וזמן הגעתן, באמצעות שימוש בטכניקות

CNN-Based image Recognition, על מנת לסווג קטגוריות של תעבורה כגון:

(VoIP,video,browsing,וכו) ובכך ניתן לזהות אפליקציות ספציפיות.

תרומה עיקרית

אז התרומה העיקרית של מאמר זה הינה הגישה החדשה בה מבצעים העתקה של המידע המוצפן ברשת אל תמונות וסיווגן באמצעות CNN אשר במקור הינה גישה לזיהוי תמונות.

יתרונות:

- ממיר זרם מידע לתמונות, ובכך הופך את סיווג התעבורה לפשוט כמו בזיהוי תמונות.
- מגיע לרמת דיוק גבוהה בזיהוי אפליקציות(ליתר דיוק "99.7%").
- נדרש מינימום אחסון כוח חישוב, אשר הופך את הסיווג למעשי עבור ניתוח של זרם המידע בזמן אמת.
- לא משתמש בתוכן המידע אלא רק בגודל וזמני הגעה של החבילות מה ששומר על פרטיות המשתמש.

תכונות תעבורה שהמאמר משתמש בו

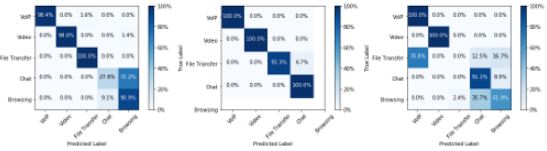
1. גודל חבילה: גודל של כל חבילה בזרם המידע.
2. זמן הגעת חבילה: משך הזמן של כל חבילה שהגיעה.
3. זרימה דו כיוונות לעומת זרימה חד כיוונית: שלא כמו בגישות ישנות, המאמר עוסק בזרימה חד כיוונית של מידע אשר מוריד באופן גבוה את הסיבוכיות.
4. הצגה של זרם-תמונה(Flow-Pic)- ממיר חבילות של מידע לתמונה על בסיס גודל וזמן הגעה של החבילה.
5. עמידות בהצפנה – אפקטיבי גם עבור VPN וגם Tor Traffic.
6. היסטוגרמת זמן-גודל הגעה – בו ציר ה-X מתאר את זמן ההגעה וציר ה-Y מתאר את גודל החבילה,ובכך יוצר את התמונה.
התכונות החדשות הינן 4,5,6.

תוצאות עיקריות והמסקנות מהן

1. דיוק גבוה סיווג התעבורה המוצפנת-

- תוצאה : זרם התמונה(Flow-Pic) שמשולב עם גישת ה- CNN מגיע לרמת דיוק למעלה מ-96% בסיווג זרם מידע מוצפן.
- טבלה 4 אשר מתארת פרוטוקול או מחלקה אל מול רמת הדיוק בפעולות הסיווג.

Class	Accuracy (%)			
VoIP	Training/Test	Non-VPN	VPN	Tor
	Non-VPN	**99.6**	99.4	48.2
	VPN	95.8	**99.9**	58.1
	Tor	52.1	35.8	**93.3**
Video	Training/Test	Non-VPN	VPN	Tor
	Non-VPN	**99.9**	98.8	83.8
	VPN	54.0	**99.9**	57.8
	Tor	55.3	86.1	**99.9**
File Transfer	Training/Test	Non-VPN	VPN	Tor
	Non-VPN	**98.8**	79.9	60.6
	VPN	65.1	**99.9**	54.5
	Tor	63.1	35.8	**55.8**
Chat	Training/Test	Non-VPN	VPN	Tor
	Non-VPN	**96.2**	78.9	70.3
	VPN	71.7	**99.2**	69.4
	Tor	85.8	93.1	**89.0**
Browsing	Training/Test	Non-VPN	VPN	Tor
	Non-VPN	**90.6**	-	57.2
	VPN	-	-	-
	Tor	76.1	-	**90.6**



(a) Non-VPN (b) VPN (c) Tor

- מסקנה : גם כאשר מאמנים את המודל בסביבה שהיא אינה מבוססת VPN, עדיין המודל מצליח לסווג תעבורה מסוג VPN בדיוק גבוה (99.4% - 78.9%).

2. זיהוי אפליקציות

- תוצאה : ה-CNN יכול לסווג אפליקציות ספציפיות(סקייפ,יוטיוב,סרטון פייסבוק,וכו) בדיוק של 99.7%.
- טבלה 5 : המהווה מדד לביצוע של מודל ה-CNN עבור VoIP ו אפליקציות וידיאו.

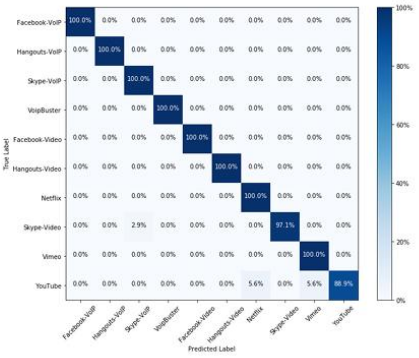


Figure 5: A confusion matrix of the VoIP and video applications identification problem.

- מסקנה : המודל מכליל באופן תקין את הסיווג והזיהוי של אפליקציות ואף מצליח לזהות אפליקציות אשר לא התאמן עליהן.

3. ביצועי סיווג חזקים על פני טכניקות הצפנה

- תוצאה : מודל הCNN מצליח באופן נכון לסווג זרמי מידע מוצפנים מסוגים שונים(Non-VPN,VPN,Tor).
- טבלה : אשר מהווה סיכום של תוצאות הצלחה של סיווג זרמי מידע.

Problem	FlowPic Acc. (%)	Best Previous Result	Remark
Non-VPN Traffic Categorization	85.0	84.0 % Pr., Gil *et al.* [15]	Different categories. [15] used unbalanced dataset
VPN Traffic Categorization	98.4	98.6 % Acc., Wang *et al.* [7]	[7] Classify raw packets data. Not including browsing category
Tor Traffic Categorization	67.8	84.3 % Pr., Gil *et al.* [15]	Different categories. [15] used unbalanced dataset
Non-VPN Class vs. All	97.0 (Average)	No previous results	
VPN Class vs. All	99.7 (Average)	No previous results	
Tor Class vs. All	85.7 (Average)	No previous results	
Encryption Techniques	88.4	99. % Acc., Wang *et al.* [7]	[7] Classify raw packets data, not including Tor category
Applications Identification	99.7	93.9 % Acc., Yamansavascilar *et al.* [10]	Different classes

- מסקנה : מודל ה CNN הינו עמיד להצפנות ואינו דורש בדיקה של המטען(Payload)בכך נמנעת חדירה לפרטיות.

מאמר ניתוח תעבורה מוצפנת מסוג HTTPS

הקדמה :

מאמר זה מעמיק בשאלה כיצד עדיין ניתן לחשוף מידע על משתמש כלשהו דרך זרם מידע מוצפן תחת פרוטוקול HTTPS,כגון :מערכת הפעלה,דפדפנים,ואפליקציות.

על אף ש -HTTP מתוכנן להגן על פרטיות המשתמש,המאמר מראה שהתוקף(האקר) יכול להשתמש בדפוסי זרם המידע,גודל החבילה,וכדומה על מנת לסווג את פעילותו של המשתמש בדיוק גבוה.

באמצעות שימוש בלמידת מכונות,כותבי המאמר הגיעו לרמת דיוק של 96.06% בזיהוי מידע על מערכת המשתמש מבלי לפענח את זרם המידע המוצפן.

תרומה עיקרית

כפי שתואר בהקדמה התרומה העיקרית של מאמר זה הינה ההדגשה על הסכנה של חשיפת מידע אישי של המשתמש על אף השימוש בפרוטוקול בטוח כמו HTTPS.

נקודות נוספות :

- שימוש בתכונה חדשה שמבוססת TLS/SSL ועל דפוסי גלישה בו הדפדפן שולח צרורות של מידע,כלומר ישנם תקופות קצרות טווח של רמת פעילות גבוהה אשר אחריה מגיע ירידה חדה בפעילות.

- משתמש במכונת תמיכה ווקטורית (Support Vector Machine)- אלגוריתם ללמידה של מכונות אשר משתמשים בו עבור סיווג.
אלגוריתם זה מוצר את בחירת החסם האופטימלי אשר מפריד מחלקות במבני נתונים.

- מהווה הוכחה שהצפנה בלבד אינה מבטיחה בטיחות מבחינת הפרטיות.

תכונות תעבורה שהמאמר משתמש בו

1.מונה חבילות

2.זמן הגעה

.כמות ביטים כוללת

4.דפוס התנהגות של הדפדפן

5.מנגנון Keep-Alive Packets Count – מאתר חיבורי HTTPS persistent.

6. מנגנון התנהגות מסוג TLS/SSL.

7.מטריצת מדידה ל-Throughput.

התכונות החדשות הינן 4,5,6,7.(טבלה 1)

תוצאות עיקריות והמסקנות מהן

1. דיוק גבוה בסיווג של מערכות הפעלה,דפדפן ואפליקציות

- תוצאה : המודל (SVM) מגיע לרמת דיוק של 96.06% בסיווג של מערכות הפעלה של הדפדפנים ואפליקציות מדפדפני HTTPS מוצפנים.

- איור 2 : המתאר את רמת הדיוק של המודל עם קבוצות תכונות שונות :

(a) Tuple Accuracy Results (b) OS Accuracy Results

(c) Browser Accuracy Results (d) Application Accuracy Results

-מסקנה : תכונות ה SSL/TLS וצרורות התעבורה משפרות דרסטית את דיוק הסיווג בהשוואה לשימוש בתכונות הבסיסיות.

2.SSL/TLS & Bursty Features Improve classifications:

-תוצאה : הוספת תכונה זו מגדילה את רמת הדיוק של הסיווג מ-9.52% ל- 96.06%

- טבלה 1 : המתארת את התכונות הבסיסיות והחדשות-

Forward packets
Forward total Bytes
Min forward inter arrival time difference
Max forward inter arrival time difference
Mean forward inter arrival time difference
STD forward inter arrival time difference
Mean forward packets
STD forward packets
Backward packets
Backward total Bytes
Min backward inter arrival time difference
Max backward inter arrival time difference
Mean backward inter arrival time difference
STD backward inter arrival time difference
Mean backward packets
STD backward packets
Mean forward TTL value
Minimum forward packet
Minimum backward packet
Maximum forward packet
Maximum backward packet
Total packets
Minimum packet size
Maximum packet size
Mean packet size
Packet size variance

(a) base features

TCP initial window size
TCP window scaling factor
SSL compression methods
SSL extension count
SSL chiper methods
SSL session ID len
Forward peak MAX throughput
Mean throughput of backward peaks
Max throughput of backward peaks
Backward min peak throughput
Backward STD peak throughput
Forward number of bursts
Backward number of bursts
Forward min peak throughput
Mean throughput of forward peaks
Forward STD peak throughput
Mean backward peak inter arrival time diff
Minimum backward peak inter arrival time diff
Maximum backward peak inter arrival time diff
STD backward peak inter arrival time diff
Mean forward peak inter arrival time diff
Minimum forward peak inter arrival time diff
Maximum forward peak inter arrival time diff
STD forward peak inter arrival time diff
Keep alive packets
TCP Maxiumu Segment Size
Forward SSL Version

(b) new features

-מסקנה: גישת הסיווג הנוכחית מפספסת תכונות ספציפיות של דפדפנים אשר התכונות (features) החדשות אכן מיישמות.

מאמר 3 Early Encrypted Traffic Classification

הקדמה:

מאמר זה עוסק בסטנדרט ה - Encrypted ClientHello, תכונת אבטחה של -TLS Transport Layer Security גרסא 1., אשר מסתיר מידע קריטי שמנוצל לסיווג זרם המידע.

הצפנת מידע מקשה על סיווג התעבורה באמצעות הגישות הרגילות:

1.Deep Packet Inspection

2.Server Name Indication Inspection

.Flow-Based Fingerprinting

מכיוון ECH מסתיר את המידע הקריטי, דרושה טכניקה חדשה על מנת לזהות את סוגי השירותים בחיבור מבלי להרוס את ההצפנה.

תרומה עיקרית של המאמר

התרומה העיקרית של המאמר הינה ההדגמה של היכולת של ECH ב-TLS. 1
מפריע לגישות ההצפנה הרגילות ומציע פתרון יעיל שמבוסס על למידת מכונות

Hybrid Random Forest Traffic Classifier - (hRFTC) על מנת לסווג זרם מידע מוצפן מבלי להסתמך על המידע הקריטי של TLS.

- hRFTC : מודל חדש בעל רמת דיוק גבוהה מבחינת סיווג התעבורה.
- מראה את קצה גבול היכולת של גישות סיווג הסטנדרטיות.
- מציע אלטרנטיבה לתכונות הסיווג הסטנדרטיות.

תכונות תעבורה שהמאמר משתמש בן

תכונות תעבורה בסיסיות :

1.אורך חבילה סטטיסטי- מינימום,מקסימום,שונות של גודל החבילה.

2.זמן בין הגעה- הפרש הזמנים בין חבילות רצופות.

.כמות חבילות ומשך זרימה- המספר הכולל של החבילות ואורך כל המעבר.

תכונות תעבורה חדשות(Novel) :

1.TLS Encrypted ClientHello Length- מודד את אורכה של ההודעה המוצפנת של ECH.

2.TLS Record Layer Statistic- מחלץ דפוסי תזמון וגודל בשכבת הרשומות של TLS.

3.TLS Handshake Timing Features- משתמש במשך "לחיצת הידיים" כטביעת אצבע מכיוון שלשירותים שונים יש לחיצת ידיים שונה.

תוצאות עיקריות והמסקנות מהן

1. hRFTC מגיע לדיוק סיווג גבוה למרות מכשול ההצפנה

- תוצאה : דיוק המודל הגיע ל94.6%, מבחינת דיוק ההצפנה בהשוואה למודלים אחרים.

-טבלה 11 :

TABLE 11. Full dataset per class F-score for different classifiers.

Class	F-score [%]						
	Hybrid Classifiers			Flow-based Classifier	Packet-based Classifiers		
	hRFTC [proposed]	**UW** [35]	**hC4.5** [34]	**CESNET** [63]	**RB-RF** [24]	**MATEC** [33]	**BGRUA** [32]
BA-AppleMusic	**92.1**	89.5	80.2	89.2	25.5	13.1	14.5
BA-SoundCloud	**99.6**	98.9	97.8	98.7	84.4	81.8	82.0
BA-Spotify	**93.6**	90.8	89.0	88.5	16.3	0.0	3.6
BA-VkMusic	**95.7**	89.7	88.5	91.8	2.6	2.1	3.2
BA-YandexMusic	**98.5**	93.2	93.7	92.5	1.8	0.2	0.1
LV-Facebook	**100.0**	99.7	99.8	99.8	**100.0**	**100.0**	**100.0**
LV-YouTube	**100.0**	**100.0**	99.9	**100.0**	**100.0**	99.0	98.4
SBV-Instagram	**89.7**	74.7	76.5	78.8	10.0	6.3	6.4
SBV-TikTok	**93.3**	81.8	81.8	76.3	38.3	34.3	34.5
SBV-VkClips	**95.7**	94.0	91.3	92.4	53.2	37.7	46.0
SBV-YouTube	**98.2**	96.6	94.7	96.4	1.1	0.2	0.2
BV-Facebook	**87.7**	78.2	79.7	77.6	5.6	3.2	3.8
BV-Kinopoisk	**94.1**	84.1	85.8	89.8	5.4	4.0	4.1
BV-Netflix	**98.5**	97.2	95.2	93.7	50.7	52.3	56.1
BV-PrimeVideo	**91.3**	86.7	84.1	84.7	32.5	24.7	26.8
BV-Vimeo	**94.8**	90.5	90.2	81.4	72.0	19.5	68.6
BV-VkVideo	**88.6**	80.5	80.4	79.7	10.5	0.0	0.1
BV-YouTube	**85.9**	84.3	77.0	78.5	22.3	19.6	20.2
Web (known)	**99.7**	99.5	99.4	99.4	98.0	98.0	98.0
Macro-F-score (average)	**94.6**	89.9	88.7	88.9	38.4	31.4	35.1

LV is Live Video, (S)BV is (Short) Buffered Video, and BA is Buffered Audio.

-מסקנה : על אף הצפנת המידע על ידי מנגנון ה-ECH,תכונות המבוססות על זרימה עדיין מאפשרות סיווג בלע דיוק גבוה.

2.מנגנון ה- ECH מקטינה את האפקטיביות של גישות סטנדרטיות :

- תוצאה : שיטות הסיווג אשר נשענות על מידע של ה -TLS בצורה של טקסט הינן בעלות ביצועים גרועים כאשר ECH מופעל.

- איור 4 :

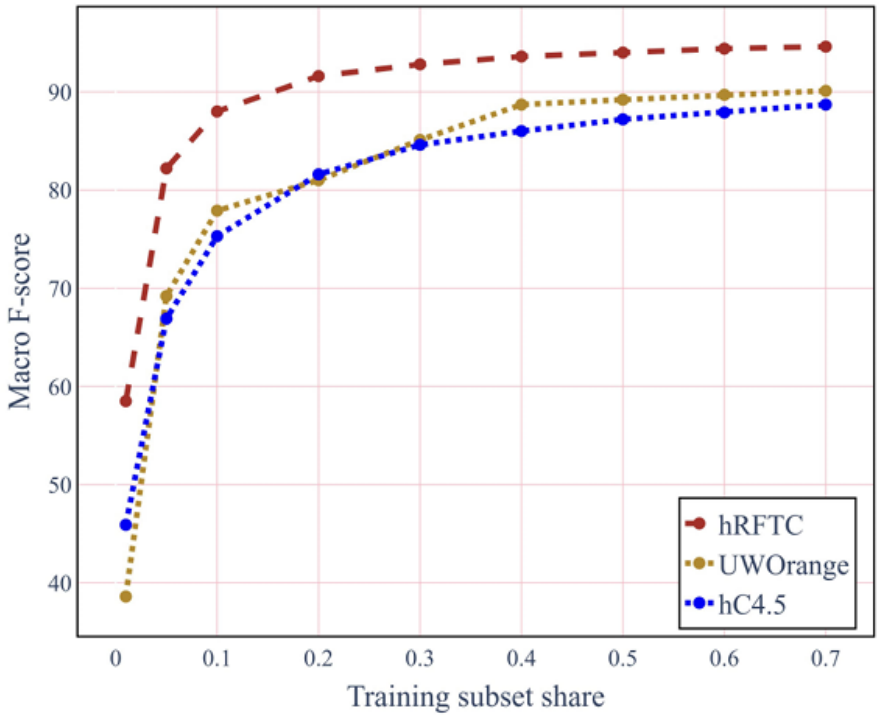


FIGURE 4. **F-score depending on the training subset share.**

-מסקנה : הצפנה של מידע בלבד אינה מספיקה למנוע סיווג, אך זה מאלץ מעבר לטכניקות מבוססות זרימה וסטטיסטיות במקום בדיקת הנתונים(.Meta-Data)

.החשיבות של תכונות שכבת הרשומות של TLS בסיווג :

-תוצאה : תכונות אשר מתקבלות ממודל הרשומות של TLS (כגון דפוסי זמן וגודל) מגביר באופן גבוה את הדיוק גם כאשר נתוני ClientHello.

-טבלה 2 :

TABLE 2. **Summary of most notable early traffic classification studies.**

Feature Type	Ref.	NN/ML	Study Year	Classification Problem	Traffic	ECH	Dataset Size, Number of flows	Dataset Year
Packet-based	[38]	NN	2017	Traffic Type	Multi-protocol	No	160k	2016
	[39]	NN	2018	Traffic Type	Multi-protocol	No	260k	2016
	[40]	NN	2019	Traffic Type	Multi-protocol	No	260k	2016
	[41]	NN	2019	Traffic Type	Multi-protocol	No	260k	2016
	BGRUA, [32]	NN	2020	Service	TLS (hidden SNI)	No	590k	2016
	MATEC, [33]	NN	2021	Service	TLS (hidden SNI)	No	590k	2016
	[42]	NN	2022	Traffic Type	Multi-protocol	No	260k	2016
	RB-RF, [24]	ML	2022	Service & Traffic Type	TLS+ECH	Yes	3.5k	2021
	[25]	NN	2023	Service & Traffic Type	TLS (hidden SNI)	No	380k	2021
Flow-based	[55]	NN	2017	Protocol & Service	Multi-protocol	No	22k	2017
	[56]	NN	2017	Protocol	Multi-protocol	No	260k	2017
	[46]	ML	2020	Traffic Type	Multi-protocol	No	260k	2016
	[57]	NN	2022	Service	TLS	No	65k	2022
	CESNET, [63]	NN	2023	Service & Traffic Type	TLS	No	140M	2022
Hybrid	**hC4.5**, [34]	ML	2020	Service	TLS (hidden SNI)	No	590k	2016
	[64]	NN	2022	Service	TLS (hidden SNI)	No	240k	2018
	UW, [35]	NN	2023	Service & Traffic Type	TLS (hidden SNI)	No	450k	2021

Note: We emphasize with **a bold font** the algorithms considered in this paper as baselines.

-מסקנה : שכבת הרשומות של מנגנון ה TLS מספקת מאפיינים חשובים של טביעת אצבע, המאפשרת למודלים לסיווג להישאר יעילים למרות ההצפנה.