לילה טוב חברים, היום אנחנו שוב בפינתנו DeepNightLearners עם סקירה של מאמר בתחום הלמידה העמוקה (לא הספקתי לסיים בלילה). היום בחרתי לסקירה את המאמר שנקרא " Neuron Shapley: Discovering the Responsible Neurons" שפורסם לראשונה לפני כ- 9 חודשים. נכון שבדרך כלל אני סוקר מאמרים טריים יותר אבל זה משך את תשומת ליבי כי הוא מציג שיטה מגניבה לחקר התנהגות של רשתות נוירונים(explainable NN) ומשתמשת באופן מפתיע בכלים מעולם בעיות שודדי מרובי ידיים(MAB- multi-arm bandits).

תחומי מאמר: חקר התנהגות של רשתות, נוירונים מאומנות, תורת המשחקים

מאמר יוצג בכנס: NeurIPS 2020

כלים מתמטיים, מושגים וסימונים: ערכי SHAP, שיטת מונטה קרלו לדגימה, בעיות שודדי מרובי ידיים, רווח סמך (confidence interval), חשיבות של פיצ'רים (feature importance)

תמצית מאמר: המאמר מציע שיטת למדידת תרומה של נוירון על הביצועים של רשת נוירונים מאומנת. במילים אחרות עם איפוס של נוירון גורם לירידה משמעותית בביצועים של רשת נוירונים, החשיבות(תרומה) שלו גבוהה, אחרת היא נמוכה. במידה מסוימת זה מזכיר "חשיבות של פיצ'ר" (feature importance) רק שכאן אנו לא בוחנים את הפיצ'רים של מודל עצמו ולא של הקלט שלו. המחברים בחרו בגישה דומה לערכי SHAP הקלאסיים, שהפכו לאחרונה לאחד הכלים הפופולריים בשערוך חשיבות הפיצירים, ככלי למדידת לחשיבות של נוירונים. באופן לא מפתיע "חשיבות של נוירון" נקראת במאמר ערך שאפלי של נוירון (Neuron Shapley - נקרא לזה N-Shap בהמשך).

אז מה זה בעצם N-Shap? למעשה N-Shap של נוירון N_i מודד את התרומה הממוצעת לביצועים, מושגת עי״ הוספת נוירון N_i לכל תת-הרשתות של רשת N שלא מכילות את N_i. כלומר לוקחים כל תת-הרשת של N ,מודדים את הביצועים שלה ואז מוסיפים לכל אחת מהם את N_i ושוב מודדים את הביצועים ומחשבים את ההפרש שלהם(נדגיש שאנו <u>לא מאמנים את תת-הרשתות אלא רק מודדים את הביצועים שלה על דאטה סט נתון</u>).
אז N-Shap של נוירון זה מחושב עי״ מיצוע של כל הפרשי הביצועים עבור כל תת רשתות של N. שימו לב שבנוסחה (1) במאמר שמגדירה את N-Shap מופיעות מקדמים בינומיאלי (מספר תת-רשתות בגודל S) המיועד בשביל מיצוע את התרומות של כל תתי רשתות בגודל S).

כידוע המספר הכולל של תת-רשתות של רשת נוירונים הינו אקספוננציאלי בכמות הנוירונים ברשת. לכן גישה זו אינה ישימה אפילו עבור רשתות לא גדולות במיוחד (מאות אלפי נוירונים). כדי להתגבר על בעיה זו מחברי המאמר מציע שתי גישות:

- <u>גישת מונטה-קרלו:</u> עבור כל נוירון N_i, דוגמים מספר תת-רשתות M(את הנוירונים המהווים אותן) באופן רנדומלית, כלומר כל תת רשת מקבלת הסתברות שווה להיבחר. אז N-SHAP של כל נוירון זה בעצם ממוצע של כל התרומות של על כל תתי-הרשתות שנדגמו עבורו. מכיוון שמספר תתי רשתות הינו אקספוננציאלי במספר המשקלים ברשת הגישה הזו לא יעילה עקב השונות הגבוהה של האומדנים של N-Shap המחושבים באמצעותה (כאשר מספר הדגימות M הינו הרבה יותר קטן ממספר הנוירונים הכולל N_num).
- <u>גישת דגימה אדפטיבית המבוססת על הכלים מעולם MAB:</u> הם מציינים במאמר שבעצם אנו רוצים לבחור K נוירונים בעלי N-Shap הגבוהים ביותר. עם ניסוח כזה הבעייה הופכת דומה לבעיה הקלאסית בתחום של MAB קרי מציאת "מכונת הימורים בעלת הסתברות זכייה מקסימלית". במילים אחרות המחברים שמו לב שבעיה זו דומה לבעית מציאה של K משתנים מקריים בעלי תוחלת הגבוהה ביותר עבור סט גדול של משתנים כאשר נדונה באופן נרחב בתחום של MAB (יפורט בהמשך). בהתבסס על הבחנה זו המאמר מציע אלגוריתם הנקרא (truncated MAB Shapley T-MAB-S) שעבור K נתון מזהה K נוירונים עם התרומה הגבוהה ביותר. בגדול בכל איטרציה, עבור נוירון N_i האלגוריתם דוגם תת רשת אחת ומחשב את תרומתו. לאחר מכן מצמצמים את סט הנוירונים הנדגמים עי״ הוצאת נוירונים שרווח סמך שלהם (על

פני האיטרציות) לא מכיל את ערך התרומה ה K המקסימלי (k-th largest) עבור אותה איטרציה. האלגוריתם עוצר כאשר לא נותרו נוירונים בסט הנדגם (התהליך והאינטואיציה יפורטו בפרק הבא)

הסבר של רעיונות בסיסיים :

תכונות של N- Shap: קודם כל נדון בשלוש תכונות הבסיסיות של מטריקת N-Shap:

- תרומה 0 לנוירון N_i יכולה לקרות רק אם הוספתו לכל תת רשת לא משפיע בכלל על הביצועים
- אם השינויים בביצועים לאחר הוספה של שני נוירונים נתונים לכל תת רשת אפשרית (שלא מכילות את שני הנוירונים) הינם שווים, אז N-Shap של נוירונים אלו גם יהיו שווים
- אדיטיביות: נניח שיש לנו שני דאטה סטים שחישבנו עבורם N_Shaps של נוירון כלשהו. ניתן לראות ש -N Shap עבור נוירון זה המחושב על איחוד דאטה סטים אלו יהיה שווה לסכום של ערכיו

בזכות תכונות אלו (שהם מוכיחים בצורה ריגורוזית) המאמר טוען N-Shap מהווה מטריקה "טובה והגיונית" למדידה של תרומת נוירון לביצועי רשת (אני חושב ש- N-Shap הינה מטריקה טובה בהקשר המדובר כי היא מהווה הרחבה טבעית של ערכי שאפלי קלאסיים לרשתות נוירונים)

פריטים ואינטואציה של האלגוריתם :T-MAB-S

- מגדירים את סט הנוירונים הנדגמים U כסט המכיל את כל הנוירונים של רשת N
- עבור כל נוירון N_i האלגוריתם דוגם תת-רשת אחת ומודד את התרומה של N_i עבור תת רשת זו. כאשר התרומה של נוירון קטנה מאיזשהו סף, לוקחים את הערך שהמתקבל באיטרציה הקודמת. מחשבים את ממוצע, שונות רווח סמך של ערכי N-Shap עבור כל הנוירונים שנותרו ב U. נזכיר כי רווח סמך נבנה סביב הממוצע ורוחבו נמדד במספר שונויות סביב התוחלת (ראה הסבר על בניית רווח סמך ליותר פרטים)
- מחשבים את הערך ה K-th המקסימלי Max_K עבור ערכי N_Shap שהתקבלו באיטרציה זו
- מוציאים מ- U את כל הנוירונים Max_K לא שייך לרווח סמך שלהם (עם איזשהו מרג'ין קטן משני הצדדים). בואו נבין את האינטואיציה של מה שקורה בשלב הזה. למעשה Max_K הינו אומדן של מקסימום ה- K של כל ערכי N-Shap שנדגמו באיטרציה זו. כאשר אנו מוציאים את הנוירונים שעבורם הערך הזה לא שייך לרווח סמך שלהם (שזה האינטרוול שבו ערך N_Shap של נוירון זה אמור להיות בהסתברות גבוהה), אנו מוציאים את הנוירונים שהסבירות שערך N-Shap שלהם יהיה בין טופ-K הינה נמוכה. כך מצמצמים את מספר הנוירונים הנדגמים עי" הוצאתם של "מועמדים לא טובים להיות בין טופ-K
- עוצרים כאשר סט הנוירונים הנדגמים נהיה ריק
- בוחרים K הנוירונים עם ערכי N-Shap המקסימליים

הסבר על מושגים חשובים במאמר:

ערכי שאפל: ערכי שאפלי הינו כלי קלאסי לשערוך של חשיבות של פיצ'רים בהינתן מודל מאומן. למעשה עושים משהו מאוד דומה לנעשה במאמר הנסקר - מודדים את השינוי בביצועים המתקבל עי" הוספת של פיצ'ר f לכל תת-קבוצה של פיצ'רים (כאשר יש מספר רב של פיצ'ר משתמשים בקירובים בצורה דומה למה שנעשה במאמר)

בעיית "שודד מרובה ידיים" ((MAB: נניח שיש לנו N מכונות מזל שלכל אחת יש הסתברות שונה לזכייה שלא ידועה למהמר. המטרה העיקרית בבעיות MAB הינה(בגדול מאוד) למקסם את הרווח של המהמר (הסבר על בעיות MAB).

הישגי מאמר: המאמר מראה כמה תוצאות מעניינות ודי לא צפויות לגבי ההשפעה של נוירונים טופ-K על ביצועים המודל(עבור 3InceptionV שאומן על Imagenet). למשל הם מראים שהוצאה של 10 נוירונים (למעשה זה איפוס של 10 קרנלים שמחשבים אותם) לירידה של 50% בדיוק כאשר האיפוס של 20 נוירונים כאלו מרסק את הביצועים ל 8%(!!) דיוק. עוד דבר מעניין שהם מראים שאם מוציאים את הנוירונים החשובים לזיהוי של קטגוריה ספציפית,

הביצועים עלייה מתרסקים אך הפגיעה בדיוק בקטגוריות אחרות היא די קטנה. צריך לציין שהמסקנות האלו הן לא אינטואיטיביות כלל (לפחות מבחינתי) כי כאשר מאמנים רשת עם דרופאאוט חשיבות של כל נוירון בודד נוטה להיות לא גבוהה במיוחד. לא הייתי משער שההורדה של 20 נוירונים תוביל לקריסה מוחלטת של ביצועים....

הם גם בדקו נוירונים הכי רגישים (מבחינת תרומתם להתקפה) להתקפות אדוורסריות כאשר המטריקה שהם השתמשו בה היא הפרש בין ה"הצלחה של ההתקפה(עד כמה ההתקפה מצליחה לגרום לרשת לחזות לייבלים אקראיים לדוגמאות שעברו שינוי קל) לבין ביצועי מודל על התמונות הנקיות. הם מצאו שאיפוס נוירונים עם התרומה הכי גבוהה בהקשר הזה די מנטרלת כמעט את ההתקפה(אחוז הצלחת קרוב לאפס) בזמן של הביצועים על הדוגמאות הרגילות יורדות בצורה לא משמעותית. שימו לב שגישה זו אינה דרך טובה להתגונן נגד התקפות אדוורסריות. הסיבה היא שאיפוס נוירונים הכי חשובים (בהקשר זה) מעניק הגנה נגד ההתקפה הספציפית וניתן די בקלות לבנות התקפות אחרות מהסוג הזה לרשת עם הנוירונים המאופסים שתבחר נוירונים אחרים בשביל "להתמקד עליהם". מעניין שהנוירונים בעלי התרומה הכי גבוהה בהקשר האדוורסרי יצאו להם די שונים (קורלציה נמוכה) לנוירונים החשובים עבור משימת הסיווג המקורית.

לינק למאמר:https://arxiv.org/pdf/2002.09815.pdf

לינק לקוד: https://github.com/amiratag/neuronshapley

נ.ב. מאמר מעניין המשלב שיטות מתחום MAB וערכי שאפלי לאנליזה של מה שקורה בתוך רשתות נוירונים מאומנות. התוצאות לא כל כך אינטואיטיביות והייתי שמח לראות עוד מאמרים בודקים את הסוגייה הזו על משימות וארכיטקטורות רשת אחרות

#deepnightlearners