# Review 8: Adversarial Concurrent Training: Optimizing Robustness and Accuracy Trade-off of Deep Neural Networks

**Paper: https://arxiv.org/abs/2104.00322v4**

תחום: adversary training (אימון עם יריב או אימון לעומתני בעברית)

תקציר מאמר בשתי שורות : המאמר מציע שיטה הנקראת (adversarial concurrent training(ACT שבליבה אימון בו זמני של הרשת לדוגמאות הרגילות והשניה לדוגמאות האדברסריות (שזה הרשת הרובאסטית נגד התקפות)

הסבר על תחום: המטרה האולטימטיבית של אימון עם יריב הינה להעניק לרשתות נוירונים רובסטיות נגד דוגמאות אדוורסריות(לא מצאתי תרגום יותר טוב - אשמח להצעות). הדוגמא אדוורסרית נוצרת (בהינתן רשת מאומנת) עי" הוספת רעש קטן לתמונה רגילה כדי לעוות את הסיווגה ע'" הרשת. אתם זוכרים את התמונה של הפנדה החמודה שמוסיפים לה קצת רעש (העין האנושית עדיין מזהה של פנדה בקלות) אבל הרשת מסווגת אותה כפרפר בהסתברות 0.95 בערך. אז בגישה של האימון עם היריב מנסים לחשל את הרשת נגד הדוגמאות האלו.

גישות קלאסית: הגישה הקלאסית לאימון עם יריב הינה לאמן רשת שתיתן פרדיקציות קרובות לדוגמאות אדוורסריות ולדוגמאות הרגילות שממנה הדוגמא אדוורסרית נוצרה בדרך כלל). הגישה הזו פוגעות ביכולת ההכללה של המודל כי פונקציית לוס לא ממזערת את הלוס על הדוגמאות הנקיות באופן ישיר.הגישה היותר מתקדמת זה לשלב את הלוס על הדוגמאות הנקיות עם adversarial loss שזה שני לוסים ניגודיים ולרוב זה עובד לא כל כך טוב.

תקציר המאמר: אז בשביל לנסות לשפר את ההיבטים הנ"ל המאמר מציע לאמן שתי רשתות : אחד לדוגמאות הרגילות והשנייה G רובסטית על הדוגמאות האדברסריות. כל אחד מהלוסים מורכב משני מחוברים: הראשון זה הלוס על דוגמא אדוורסרית (או נקייה בהתאם) והשני הוא "מרחק" KL בין הסיווגים של שתי הרשתות (KL הפוך -inverse KL).

בנוסף בלוס של הרשת עם יריב מבוצע מיקסום על פני התוספת לדוגמא רגילה כדי ליצור דוגמא אדוורסרית "טובה" שהופך את הבעיה ללוס הראשון לבעיית minimax.

בעצם הטענה במאמר שהרעיון הזה מאפשר אימון של רשת רובסטית נגד דוגמאות אדוורסריות עם פגיעה יותר קטנה ביכולת הכללה שלה.

שיפורים שהושגו: הם בודקים את הביצועים של הרשת המאומנת נגד דוגמאות אדוורסריות:

הם משווים את הסטייה המינימלית הנדרשת בשביל לגרום לרשת לא לסווג נכון את הדוגמא

ביצועים בהתקפות ידועות.