# Review 7: PeerNets: Exploiting Peer Wisdom Against Adversarial Attacks

**Paper: https://arxiv.org/abs/1806.00088v1**

תחום מאמר: adversary training, graph attention network

תקציר מאמר בשתי שורות: המאמר מציע שיטה להתמודד עם תקיפות אדוורסריות בעזרת הוספת שכבות מנצלות את הדמיון בין הדוגמא הנתונה לבין התמונות האחרות בדאטה סט. ויצא בערך לפני שנתיים.

אינטואיציה: קודם כל בואו ניזכר מה זה התקפה אדוורסרית על רשת נוירונים מאומנת. המשמעות של תקיפה אדוורסרית היא יצירה של דוגמאות אדוורסריות. דוגמא אדוורסרית מתחילה מדוגמא רגילה x עם לייבל y. המטרה היא להוסיף פרטורבציה (רעש קטן) לתמונה וליצור תמונה כך הרשת תזהו את אותה עם לייבל שונה מ y בזמן שהעין האנושית בקלות מסווגת את הדוגמא הזו עם הלייבל המקורי. כמו שאמרתי המאמר מציע לנצל את הדמיון בין הדוגמא המסווגת לתמונות אחרות בדאטה סט שזה בעצם מקשה על יצירה דוגמא אדוורסרית כי עכשיו המסווג יידע שאפילה עם הרעש הקטן המוסף התמונה עדיין ״דומה״ לתמונות אחרות בעלות אותו לייבל.

הרעיון: במאמר התמונה מוגדרת כמטריצה n x d, כאשר n זה מספר הפיקסלים בתמונה ו d זה כמות הפיצ'רים פר פיקסל. אז לכל פיקסל בתמונה מחפשים את K (הם בחרו K =10) פיקסלים הקרובים ביותר במונחי מרחק קוסיין (cosine distance). החיפוש מתבצע בין כל (!!) הפיקסלים של כל התמונות (הנקראות תמונות peer במאמר). הבנייה הזו היא בעצם יצירה של הגרף K-שכנים ברמת הפיקסל עבור הדאטה סט. לאחר מכן מעבירים את הגרף הזה דרך (graph attention network(GAT כדי לקבל את ייצוג אגרגטיבי התמונה עי״ כל ה peers. בעצם לכל פיקסל בתמונה ולכל פיקסל בתמונת peer בגרף K-שכנים מחשבים attention score בין 0 ל 1 בדומה למה שעושים בשכבות attention קלאסיות המשמשים למשל למשימות סגמנטציה וגם לתרגום אוטומטי.

בשלב האחרון לכל פיקסל בתמונה סוכמים את כל ה scores עבור כל השכנים שלו. הם קוראים לשכבה הזו.Peer Regularization (PR) layer.

אימון: מכיוון שבניית גרף כזה לדאטה סט גדול זו משימה מאוד כבדה מבחינת כמות החישובים באימון ה peers נבחרים מהתמונות במיניבאטץיֿ

אינפרנס: peers נבחר כסט קבוע של תמונות ולכל דוגמת טסט הגרף נבנה רק בן הדוגמא הזו לסט הזה.

תוצאות: המאמר מוכיח כי עבור סוגי תקיפות מגוונים:

Universal adversarial perturbations ,Gradient descent attack, Fast Gradient Sign attack.

הוספת שכבת PR לרשת מגבירה את הרובסטיות של הרשת (מקטינה fooling rate) לעוצמת פרטורבציה נתונה מול שיטות הגנה אחרות.

דאטה סטים שנבחנו: MNIST, CIFAR10

מסקנה: השיטה נראית די מבטיחה אך לא ברור עד כמה היא מאריכה את זמן האימון. גם בחירה של גודל הבאטץ' לא נראה טריוויאלי. אבל שווה לעקוב אחרי התפתחויות :)