# Review 1: Curriculum by Smoothing

**Paper: https://arxiv.org/abs/2003.01367v5**

תחום מאמר: ארכיטקטורות CNN, למידת curriculum

מושגים וכלים מתמטיים במאמר: קרנלים גאוסיאניים

מאמר הוצג בכנס: NeurIPS

תמצית מאמר: המאמר מציע להוסיף קרנלים(מסננים) גאוסיאניים מחליקים עם עוצמת החלקה יורדת אחרי כל שכבת קונבולוציה (לפני אקטיבציה) של CNN. קרנלים אלו נועדים להתמודד עם הארטיפקטים בעלי תדרים גבוהים הנוצרים בפיצ'רים של שכבות CNN (לטענת מחברי המאמר) בשלבים המוקדמים של אימון הרשת. בעצם הגישה שלהם (הנקראת CBS) מזכירה את העיקרון של למידת עם curriculum: אימון רשתות נוירונים על משימות בעלות רמת קושי הולכת וגדלה. במקרה זה הפיצ'רים בעלי תדרים נמוכים הינם ״משימות קלות״ ואלו בעלי תדרים גבוהים מיוחסים ל״משימות קשות יותר״. בעצם CBS מנסה לווסת (לשלוט) בכמות האינפורמציה בעלת תדרים גבוהים המגיע לשכבות שונות של CNN.

רעיון בסיסי: אז מה בעצם קורה כאן? בשכבות המוקדמות מוסיפים קרנלים המעבירים רק תדרים נמוכים (ערך גבוה של פרמטר סיגמא של גאוסיאן). זה אומר בתכלס CBS ״מכריחה״ את השכבות המוקדמות ללמוד רק מידע בעל תדרים נמוכים (״דרגת קושי נמוכה״ בשפה של למידת curriculum). ככל שמתקדמים לשכבות עמוקות יותר אנחנו לאט לאט משחררים ו״מרשים״ לרשת ללמוד אינפורמציה בעלת תדרים הולכים וגדלים ("רמת קושי עולה״). בנוסף בדרך זו הם חוסמים העברת מידע רועש מהשכבת המוקדמות לשכבות יותר עמוקות בתחילת תהליך אימון של רשת.

תקציר מאמר: כמו שכבר נאמר למעלה המאמר מציע להוסיף קרנלים גאוסיאניים דו-מימדיים מחליקים ליציאה של קונבולוציה ולפני האקטיבציה בכל שכבה. אז קודם כל בואו ניזכר מהם בעצם הקרנלים האלו.

קרנל גאוסיאני: קרנל גאוסיאני דו-מימדי זה בעצם טרנספורמציה המוגדרת ע"י פונקצית צפיפות גאוסית דו-מימדית בעלת מטריצה קווריאנס I\*sigma כאשר וקטור התוחלת שלה ההמצא בנקודה (x, y) שבה קרנל זה מופעל. לקרנל זה יש תכונות של מסנן מעביר נמוכים וככל שסיגמא הולכת וקטנה הוא מעביר תדרים יותר ויותר גבוהים (קרי יכולת החלקה יורדת).

בחירת סיגמאות של קרנלים: הם מציינים שבחירה של עוצמות ההחלקה (פרמטרים סיגמא) תלוי רק בגודל המודל ובמשך זמן אימון של רשת (שיש ביניהם קורלציה חיובית). ההסבר שלהם (ללא הוכחה) הוא כזה: ככל שמשך זמן אימון של רשת יותר ארוך, המשקלים של הרשת רחוקים מערכם האופטימליים ואז צריך "להחליק אותם יותר" כדי למנוע יצירת ארטיפקטים. מעניין שקצב הדעיכה של סיגמאות נשאר קבוע לכל הארכיטקטורות שהם בדקו (שווה ל 0.9).

הישגי מאמר:

שיטות השוואה עם SOTA: הרעיון של המאמר הינו מאד פשוט ומובן אך נדרשת הוכחה שגישה זה משפרת ביצועים למגוון ארכיטקטורות ומשימות. אז הם הראו CBS טובה מ SOTA עבור 3 סוגים של משימות:

אימון רשת לטאסק ספציפי (סיווג תמונות)

הם השוו אימון רשת בצורה רגילה מול CBS למשימת סיווג הוכיחו את עליונותה של CBS למגוון ארכיטקטורות רשת: VGG-18, ResNet18, Wide ResNet-50, ResNext-50

דאטה סטים: SVHN, CIFAR10, CIFAR100, ImageNet

הפקת פיצ'רים חזקים לאותו סוג של משימה בדומיין שונה :הם עשו פריטריין של רשת VGG16 על Imagenet רגיל ועם CBS והקפיאו את משקלי הרשת. אחר כך הם הוסיפו 3 שכבות FC ואימנו אותם על דאטה סטים אחרים ( CIFAR10, CIFAR100, SVHN). התוצאה - ניצחון ל CBS בכולם

הפקת פיצ'רים חזקים לסוג אחר של משימה: דומה לסעיף הקודם רק המשימה Downstream היתה סגמנטציה סמנטית (Faster-RCNN, Pascal-Voc). התוצאה: ניצחון ל CBS

וזה הכי מעניין לדעתי: הם השתמשו ב CBS בשביל לאמן שני סוגים של VAE וניצחו את הארכיטקטורה הרגילה מבחינת NLL (שערוך של לוס) וגם מבחינת המידע ההדדי בין הייצוג הלטנטי ובין התמונה המגונרטת עבור דאטה סטים: MNIST ו-CelebA.

נ.ב. מאוד אהבתי את הרעיון של המאמר, הוא פשוט, קל למימוש ולא דורש כיול הייפר פרמטרים. הם גם הצליחו לשכנע אותי ש CBS משפרת את ביצועי הרשת למשימות מגוונות. מה שטיפה חסר לי במאמר זו הצדקה תיאורטית כלשהי של ״נפנופי ידיים״ רבים בו - התקווה שזה יבוא בהמשך.