תרגיל 1 | רשתות נוירונים לתמונות

יואב שפירא 312492838

30 במרץ 2023

חלק פרקטי

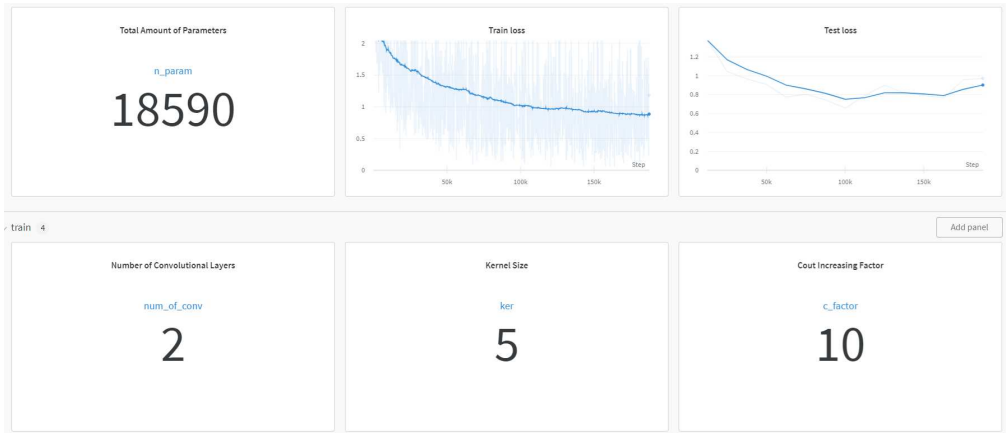
שאלה 1

בכל הגרפים בשאלה הזו ביצעתי החלקה של גרפים של ה$Train\ Loss$ כדי שיהיה יותר נוח לקריאה, ולעתים גם ל$Test\ Loss$. במקרה כזה אפשר לראות ברקע את הגרף המקורי מעט שקוף יותר.

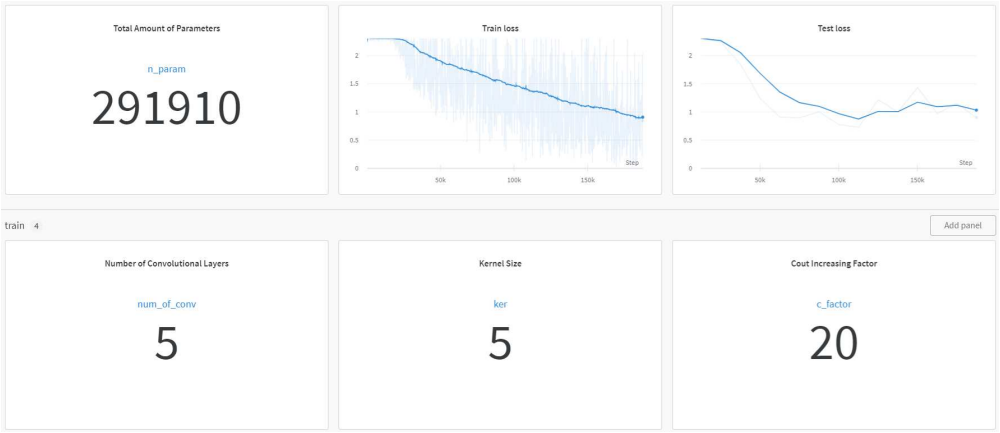
ארכיטקטורה

במהלך המימוש שיניתי את מספר השכבות קונובלוציה, גודל הפילטר, ומספר ערוצי הפלט מכל שכבה. לאחר ניסויים ראיתי ששינוי גודל הקרנל לא שינה השפיע רבות על התוצאה ולכן קיבעתי אותו על 5 לכל השכבות קונבולוציה. בנוגע למספר ערוצי הפלט, האסטרטגיה הייתה להחליט על מספר ערוצים מהשכבה הראשונה, ועל פקטור $cout\ factor$ כאשר בכל שכבה מספר ערוצי הפלט גדל ב$cout\ factor$. אחרי הרבה ניסויים ראיתי ש$cout\ factor$ הוא הפקטור החשוב וקיבעתי את מספר ערוצי הפלט מהשכבה הראשונה ל10. שינוי $cout\ factor$ היה הגורם המרכזי שהשפיע על התוצאות, אם כי גם מספר השכבות. עבור מספר קטן מדי של ערוצים, הרשת הגיעה לתוצאות נמוכות יחסית גם עם 5 שכבות, אך עבור מספר גדול של ערוצים הרשת הגיעה ל$Overfit$ עם 5 שכבות קונבולוציה. עבור רשת מאוד קטנה - 2 שכבות קונבולוציה ו$cout\ factor = 10$ - הרשת הייתה ב$Underfit$.

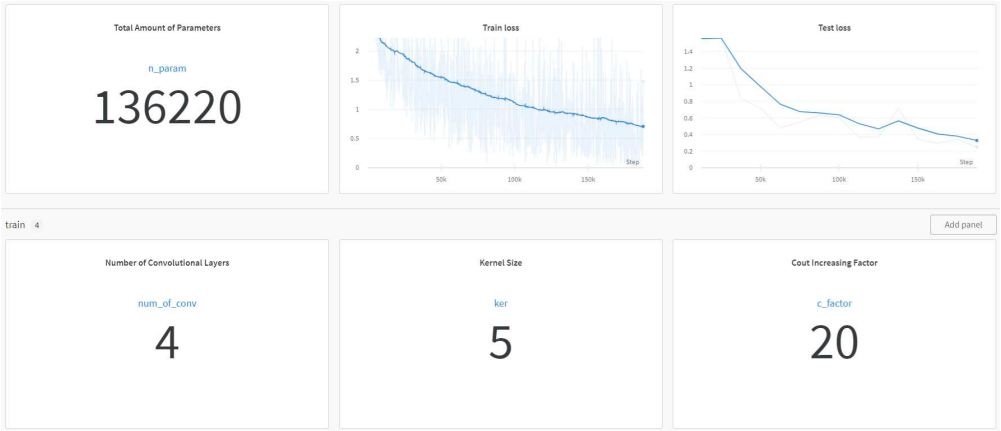
1. $Underfit$: עבור רשת עם 2 שכבות קונבולוציה, כשהראשונה מוציאה 10 ערוצי פלט והשנייה 20 ערוצי פלט, כך שבסה"כ יש 18590 פרמטרים נלמדים. אפשר לראות שהגרף של ה$Train\ loss$ שהשיפוע הולך ומתמתן מאוד מהר, כלומר תהליך הלמידה עומד למעשה להתכנס, בזמן שאחוזי הדיוק ברשת הזו עמדו על בערך 0.6 .



2. $Overfit$: עבור רשת עם 5 שכבות קונבולוציה, כשהראשונה מביניהן מוציאה 10 ערוצי פלט וכל שכבה שאחריה מגדילה ב20 ערוצי פלט, ובסה"כ עם 291910 פרמטרים נלמדים. אפשר לראות שגרף ה$Train\ loss$ ממשיך לרדת בשיפוע די קבוע , כלומר אם היינו ממשיכים עוד הוא היה משתפר אפילו יותר, אבל ה$Test\ loss$ כבר מהאמצע מתחיל לעלות, התנהגות של $overfit$.



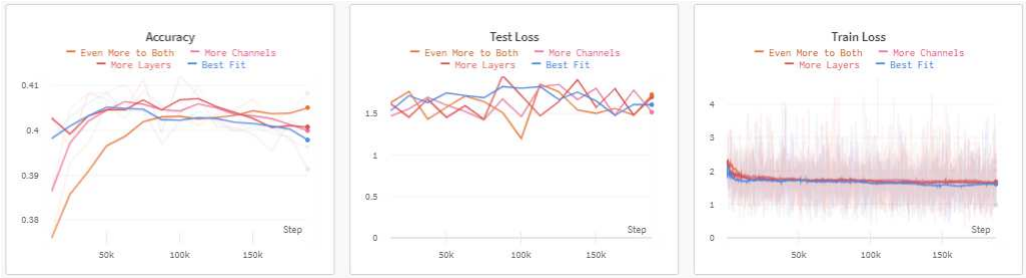
3. $Best-fit$: כשהגעתי לאוברפיט, הורדתי שכבה אחת של קונבולוציה כדי להוריד את מספר ערוצי הפלט ואת מספר הפרמטרים הכללי (הרבה מהFC) בתקווה שיביא לתוצאות טובות, ואכן הרשת הגיעה לאחוז דיוק של כ0.7 (הכי טוב שהגעתי), עם רשת של 4 שכבות קונבולצה כשהראשונה מוציאה 10 ערוצים ואחריה כל שכבה מגדילה ב20 ערוצים, ובסה"כ 136220 פרמטרים נלמדים. אפשר לראות בגרף שהשיפוע של שני ה$loss$ ממשיך לרדת, וביחד עם ממצא של אחוז דיוק גבוה יחסית מסיקים שזו למידה טובה של הרשת.



שאלה 2

חשיבות של אי־ליניאריות

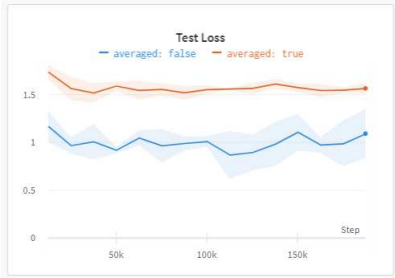
האופרטורים הלא ליניאריים במודל הם פעולות ה$MaxPooling$ והאקטיבציה $ReLU$. כשמורידים את השכבות האלה הביצועים של המודל יורדים משמעותית, כי הוא יכול לבטא הרבה פחות פונקציות, כשבמציאות ההחלטה האם עצם הוא מטוס או כלב היא פונקצייה די מסובכת... למעשה, כל פעולות הרשת יכולות להסתכם במטריצה אחת והיא מבצעת פעולה ליניארית אחת. גם לאחר הגדלה משמעותית של הרשת, הביצועים נשארים כמעט אותו דבר. בגרפים למטה ניתן לראות את הרשת המקורית, רשת שהוספתי בה ערוצים, רשת שהוספתי בה שכבות, ואחת שהוספתי בה גם וגם. כל הגרפים של ה$Train\ Loss$ מאוד סטטיים באופן כללי, עם ירידה קטנה בהתחלה. הוספתי כאן $Acuuracy$ כי מעניין לראות שהירידה הקטנה בהתחלה היא בכל זאת למידה כלשהי - אפשר לפרש זאת על ידי כך שיש גורמים בתמונות שהינם כן מאוד פשוטים שאפשר לבטא בפעולת ליניארית אחת, ואפשר בעזרתם להגיע לדיוק **כלשהו** כלומר יותר טוב מצ'אנס (תזכורת, צ'אנס כאן הוא דיוק של 0.1 כי יש 10 קלאסים).
הערה: במימוש המודל השארתי רק את הקונבולוציות, ללא $Padding$ כך שלמעשה עדיין יש צמצום של מימד המטריצות (מבנה טלסקופי). הייתה אפשרות לבצע $Average\ Pool$ (על כל ערוץ בנפרד) אבל כמו שלמדנו זו פעולה די מיותרת כי הפרמטרים בפילטר הקונבולוציה יכולים פשוט ללמוד את המשקול.

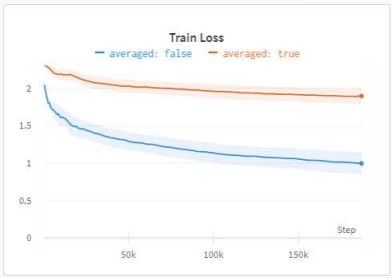


שאלה 3

חשיבות של הגדלת רצפטיב־פילד

האפשרות הטובה יותר היא **לא** לבצע $Global\ Average\ Pooling$, כי אין בה יתרון לעומת פעולה ליניארית של FC - השכבת FC יכולה ללמוד את פעולת המשקול של הממוצע הגלובלי. בנוסף, היא רק מצמצת את המימד של הדאטא ובכך מוותרת על חלק מהמידע, בעוד ששכבת הFC הייתה יכולה להשתמש במידע הזה ללמוד את המשקולות שלה יותר במדויק. ביצעתי ניסוי על רשתות קטנה (10 ערוצי אאוטפוט מהקונבולוציה) וגדולה יותר (20 ערוצי אאוטפוט), עם $Global\ Average\ Pooling$ ובלי, ואכן התוצאות מראות את זה באופן ברור. בגרף ניתן לראות את הממוצע עם הסטיית תקן של הגרפים, מקובצות לפי האם בוצע $Global\ Average\ Pooling$ (באדום) או לא בוצע (בכחול). אפשר לראות שבשני הגרפים של ה$Loss$ הרשתות שלא עברו $Global\ Average\ Pooling$ הגיעו לתוצאות טובות יותר. בכל מקרה, הביצועים של הרשתות האלה לא היו טובים יותר מהביצועים של הרשתות בסעיף 1, וזאת בגלל שאין כאן הרחבה של הרצפטיב־פילד של כל נוירון לאורך הדרך. ברשתות האלה, כל נוירון בשכבה הFC 'מסתכל' על נקודה אחת בתמונת אינפוט ולא משקלל את המידע הכולל מכל שאר התמונה. למשל אם המשימה היא להפריד בין חתול לכלב, הנוירונים שנמצאים על האישונים בעיניים לא יהיו רלוונטיים בכלל, כי אין להם את ההקשר הכולל של התמונה והרי גם לחתול וגם לכלב יש אישונים שחורים. כשמגדילים את הרצפטיב־פילד, הנוירונים יפתחו רגישות גם למה שמסביב לאישון ויהיו רלוונטיים להחלטה הסופית.





חלק תאורטי

שאלה 1

יהי L אופרטור ליניארי, כך שמתקיים התנאי

$$(\star): \quad L\left[x(i+k)\right](j) = L\left[x(i)\right](j+k)$$

ונראה כעת כי L טומן בחובו את פעולת הקונבולוציה. נגדיר לכל x את הפונקצייה $g(x)$ כך:

$$g(x) = L\left[\delta(x)\right]$$

כשδ היא פונקציית דלתא:

$$\delta(x) = \begin{cases} 1 & x = 0 \\ 0 & else \end{cases}$$

למעשה הגדרנו כאן קרנל של קונבולוציה (g), כמו שיתבהר למטה.
נתייחס לx כסיגנל בדיד ונציג אותו כסכום ממושקל של פונקציות דלתא, הנתון על ידי:

$$x(i) = \sum_{n=-\infty}^{\infty} x(n)\delta(i-n)$$

כש$x(k)$ הם קבועים שנתונים מהסיגנל (במקרה זה מתייחסים לסיגנל כאל סיגנל אינסופי לצורכי נוחות החישוב. בפועל לסיגנל יש התחלה וסוף, אך לפני ואחרי הסיגנל אפשר פשוט לרפד באינסוף אפסים).
נתבונן כעת על הביטוי:

$$\begin{aligned}
L\left[x(i+k)\right] &= L\left[\sum_{n=-\infty}^{\infty} x(n)\delta(i+k-n)\right] \\
(1) &= \sum_{n=-\infty}^{\infty} x(n)L\left[\delta(i+k-n)\right] \\
(2) &= \sum_{n=-\infty}^{\infty} x(n)L\left[\delta(i-n)\right] \\
(3) &= \sum_{n=-\infty}^{\infty} x(n)g\left(i-n\right) \\
(4) &= x * g
\end{aligned}$$

כש: 1 נובע מליניאריות של L, 2 נובע מהתנאי ($\star$) 3 נובע מהגדרת g, ו4 נובע מהגדרת קונבולוציה. וזה בדיוק מה שהיינו צריכים להראות.
על מנת לחשוף את הגרעין של הקונבולוציה, הסיגנל צריך להיות $x(i) = \delta(i)$, כך שהוא ישמש בעצמו כגרעין זהות בפעולה $x * g$ (מההגדרה של פונקציית דלתא, הגרעין יראה $[...0,1,0...]$- גרעין זהות בקונבולוציה.

שאלה 2

הסדר לא חשוב מלכתחילה, אך חשוב שהדאטא סט יהיה מנורמל כדי לקבל תוצאות משמעותיות. **הסבר**: השכבת $fully-connected$ מכילה פרמטרים נלמדים, שאינם משותפים (לא כמו בשכבת קונבולוציה), מכך שהפעולה היא לא LTI. לכן, היא תלמד כבר לבד איזה חלקים חשובים יותר ופחות בתוך הוקטור, באופן עצמאי. אם הסדר של הערכים בוקטור היה משתנה - המשקולות בשכבה פשוט היו משנות את הסדר שלהם בהתאם. אבל, חשוב להדגיש שהדאטא צריך להיות מנורמל בצורה מסויימת - במובן של $make\ sense$ - כך שהמשקולות הנלמדות יהיו כמה שיותר מכלילות ויהיו ביצועים מוצלחים.

שאלה 3

1. **כן** LTI. **הסבר:** הפעולה $ReLU$ היא פעולה מקומית שפועלת על כל נוירון (פיקסל) בצורה בלתי תלויה ($element-wise$). כלומר היא לא מתחשבת בכלל באיזה פיקסלים נמצאים מסביב, ובפרט לא משנה לה מה המיקום של הפיקסל.

2. **לא** LTI. **הסבר:** הפעולה שלוקחת את הפיקסל הימני העליון תמיד, היא רגישה להזזה. לדוגמא נסתכל על החלון (שלקוח נניח מתוך תמונה)::

$$M = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 1 \end{bmatrix}$$

נרצה להפעיל את הפעולה עם גרעין של $2 * 2$, ונתחיל משמאל. כלומר על התת־חלון הבא:

$$m = \begin{bmatrix} 1 & 2 \\ 4 & 5 \end{bmatrix}$$

התוצאה תהיה הפיקסל שמכיל 2. אבל, אם נזיז שמאלה את החלון M, נקבל פעולה על התת חלון הבא:

$$m = \begin{bmatrix} 2 & 3 \\ 5 & 1 \end{bmatrix}$$

התוצאה תהיה הפיקסל שמכיל 3.
הפעולה $MaxPooling$ לעומת זאת היא אכן קירוב לפעולה שהיא LTI, בכך שהיא מורידה את הסיכויים לכך שהזזה תגרום לתוצאה שונה, ואכן בשני המקרים הייתה מתקבלת אותה תוצאה ־ הפיקסל שמכיל 5.

3. **כן** LTI. **הסבר:** ה$bias$ הוא משותף לכל הפיקסלים בערוץ (אפשר להסתכל עליו כנרמול של הערכים במטריצה, ועל כן כהזזה). מאחר שהוא משותף לכל הפיקסלים, בפרט אם נזיז את החלון M נקבל אחרי הוספת ה$bias$ את אותה התוצאה.

4. **לא** LTI. **הסבר:** השכבה FC נועדה לאפשר רגישויות שונות באיזורים שונים לאחר ששכבות הקונבולוציה חיפשו את התבניות הרצויות. בפרט היא לא LTI מכך שהיא לומדת משקולות שאינם משותפות לפיקסלים.