סיכום פרויקט נושאים בחזית המחקר

מנחה: ד"ר אורן פרייפלד

נושא הפרויקט למידה חצי מונחית

רקע-

למידה חצי מוחית היא גישה בלמידת מכונה שמטרתה סיווג מידע. גישה זו משלבת מאגר קטן של מידע מתויג יחד עם מאגר גדול של מידע לא מתויג. ענף זה של למידת מכונה ממוקם בין למידה לא מונחית (training data לא מתויג בכלל) לבין למידה מונחית (training data מתויג לגמרי).

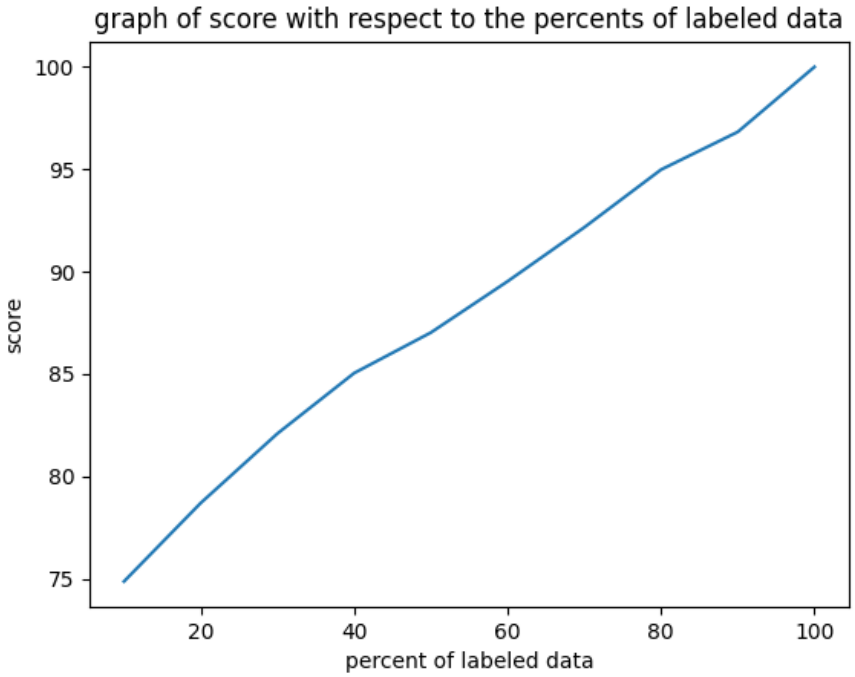
ההנחה הכללית הינה שניתן לשפר את המודל בעזרת שימוש במישע מקדים.

אנחנו בחנו האם שימוש במידע לא מתויג משפרת את ההצלחה של מודל המזהה ספרות מ-0 עד 9.

שלב מקדים-

ראשית, רצינו להבין מהי למידה חצי מונחית ולהרגיש את מגמת התוצאות. לכן הורדנו data-set מתויג של ספרות מ-0 עד 9 והשתמשנו באלגוריתם label propagation של sklearn. ריצת האלגוריתם בוצעה עם מידע מתויג באחוזים שונים ואת התוצאות שקיבלנו השוונו למידע האמיתי שקיבלנו.

מצאנו בקירוב קורלציה לינארית בין כמות המידע המתויג לבין ההצלחה של המודל כפי שניתן לראות בתמונה הבאה:



כאשר 10% מהמידע מתויג האלגוריתם צדק בכ75% מהתיוגים, וכמובן שעבור 100% האלגוריתם צדק לכל המידע.

השערת הפרויקט

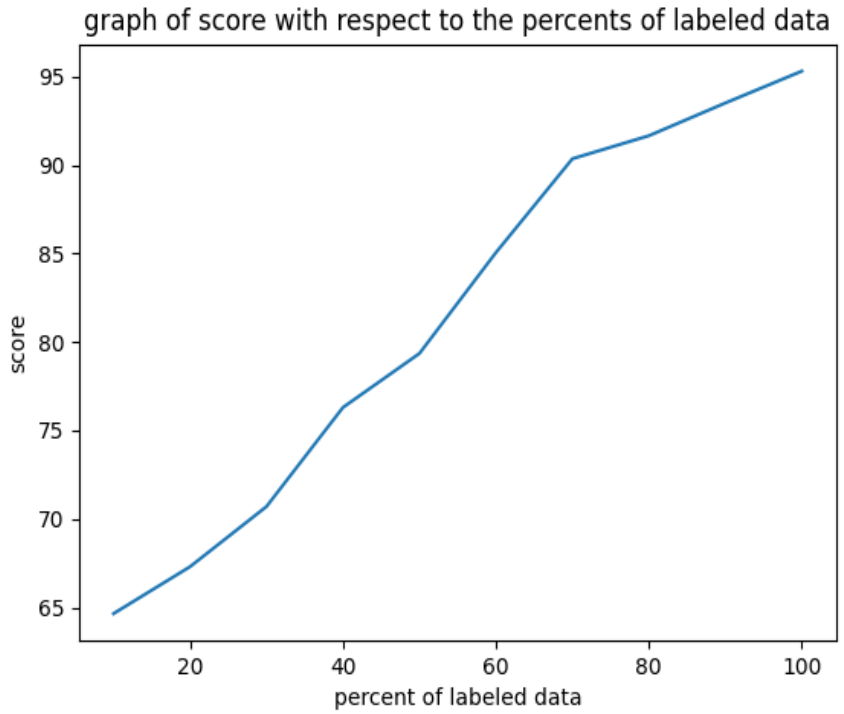
נרצה לבחון האם שימוש בטכניקות של למידה לא מונחית יכול להגדיל את ה-training set ולהוביל למגמת שיפור בביצועים של המודל.

נעבוד לפי ההנחה שנתונים זהים בעלי נטייה להשתכן קרוב זה לזה במרחב. לכן אלגוריתם ה-clustering יצליח לתייג אותם באותו אשכול, ובעזרת המידע המתויג שבידנו נצליח לתייג את המידע באחוזים גבוהים. כמו כן, נניח שיש יחס ישיר בין גודל ה-training set לאיכות המודל. כלומר, ככל שה-training set גדול יותר, כך אחוזי ההצלחה של המודל יגדלו.

בדיקת ההשארה-

השתמשנו במידע הלא מתויג על מנת לאמן את המודל. מתוך 10,000 הדוגמאות של הdata set המתויג, לקחנו 8,000 תמונות שיהוו training-set למודל, ואת ה2,000 הנוספים השארנו בתור test set. לצורך זאת ביצענו את השלבים הבאים:

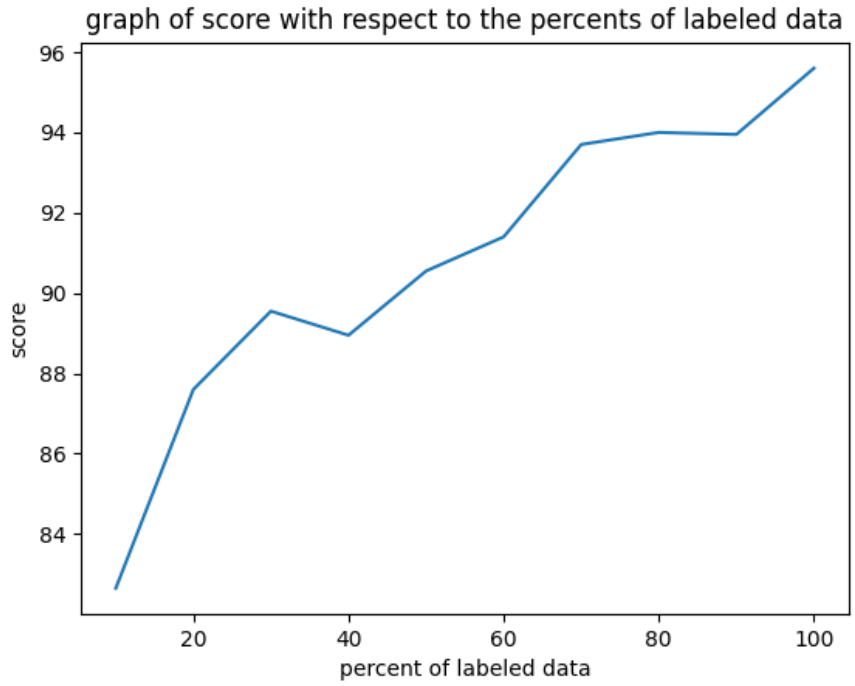
1) ביצוע PCA על מנת להוריד את הממד של המידע מ28*28 ל128.
2) הרצת אלגוריתם למידה לא מונחית וקבלת חלוקה לאשכולות של המידע.
3) בחירה שרירותית ,עבור אחוזים שונים, של מידע שיהיה מתויג. לאחר מכן נתנו תוויות חדשות לכל אשכול שקיבלנו. תיוג המידע הלא מתויג בוצע על פי שיטת "הרוב קובע", כלומר התווית המתאימה לרוב הנקודות המתויגות באשכול.
4) בעזרת ה-training set והמידע שתייגנו, אימנו מודל שיזהה ספרות.
5) בדקנו עבור הtest-set את ההצלחה בזיהוי, ושמרנו את ההצלחה עבור כל כמות שונה של אחוזי תיוג



תוצאות-

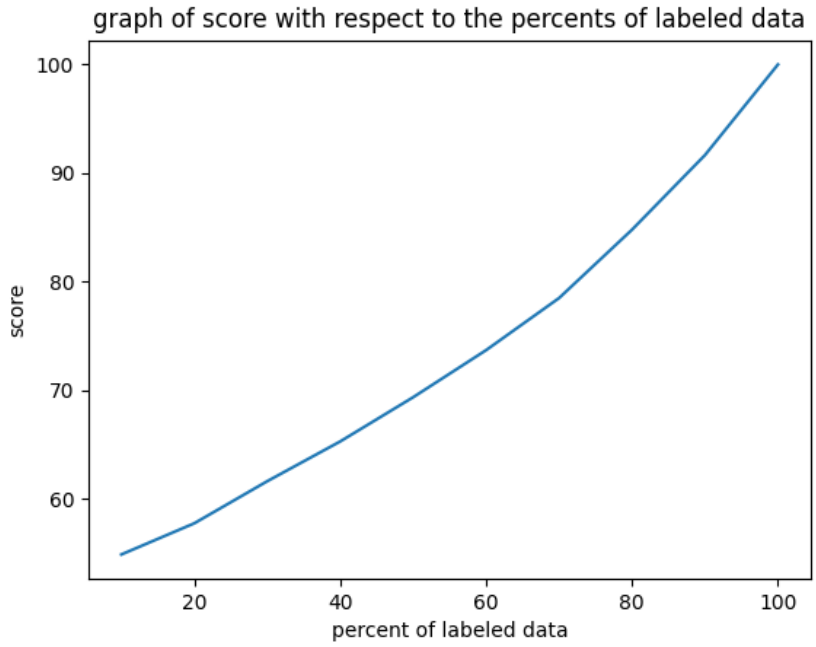
אכן, כפי שציפינו, ניתן לראות שככל שהמידע מתויג יותר כך אחוזי ההצלחה של המודל

אך על מנת לבדוק אם השימוש בטכניקה משפר את התוצאות בדקנו מה ההצלחה באימון המודל רק בעזרת המידע המתויג. דגמנו כמות זהה של מידע ואימנו מודל רק באמצעות המידע המתויג. לאחר מכן בדקנו על אותו הtest set את ההצלחה של המודל. וקיבלנו את הגרף הבא:



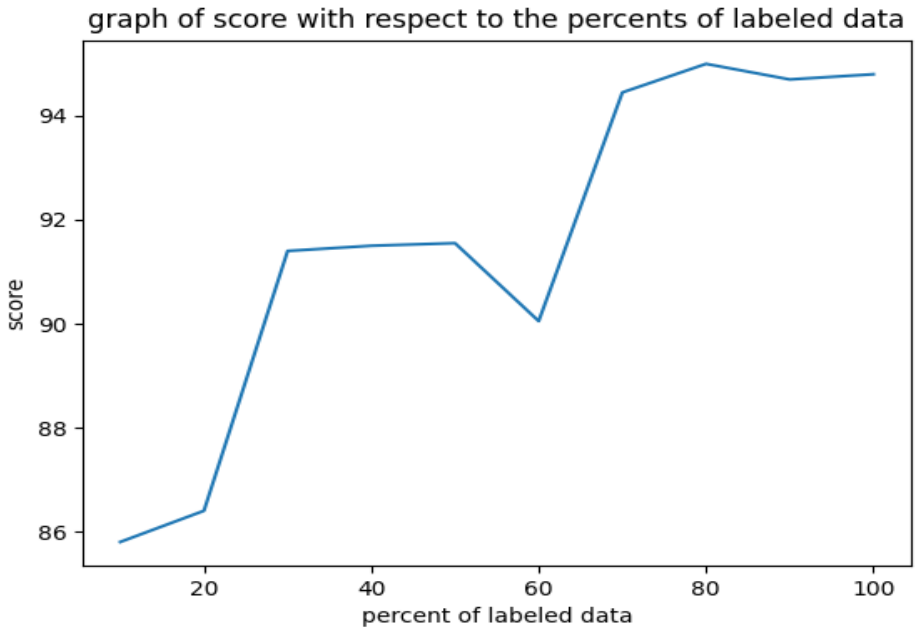
כלומר, התוצאה של אימון המודל בעזרת המידע המתויג בלבד וללא תוספת המידע שהושג בעזרת הלמידה החצי מונחית טובה יותר. הסיבה לכך היא שהתיוג של המידע החדש בעזרת ה majority rule אינה טובה.

עקב התוצאות הנ"ל, בדקנו את איכות התוצאות של סיווג המידע שלנו על פי שיטת majority rule ביחס למידע האמיתי. ניתן לראות את תוצאות הבדיקה בגרף הבא:



קיבלנו כי השיטה של "הרוב קובע" צדקה בפחות מ-55% מהתיוגים בלבד. עובדה זו גרמה למודל להתאמן עם למעלה מ-3,000 דוגמאות לא נכונות שמהוות כמעט חצי מהמודל.

עקב איכות התוצאות לפי שיטת majority rule, ביצענו שימוש באלגוריתם העובד על פי label propagation. זאת על מנת לבדוק האם שימוש בשיטה זו ישפר את התוצאות. התקבל הגרף הבא:



ניתן לראות שהשימוש בלמידה חצי מונחית משפר את הביצועים של הרשת ברוב גדול מהמקרים ובפרט כאשר מתבצע שימוש בכמות גדולה של מידע לא מתויג.

מסקנות-

שימוש בטכניקות של למידה חצי מונחית עשוי לשפר את המודל למרות תוספת של דוגמאות לא נכונות, צריך מחקר נוסף כיצד ניתן לשפר את התיוג בעזרת המידע הלא מתויג.