9.3.1 Face Recognition

אחד מהיישומים החשובים בעיבוד תמונה הינו זיהוי פנים, כאשר ניתן לחלק משימה זו לשלושה שלבים:

1. Detection – מציאת הפרצופים בתמונה.
2. Embedding – מיפוי כל פרצוף למרחב חדש, בו הפיצ'רים שאינם קשורים לתיאור הפנים (למשל: זווית, מיקום, תאורה וכדו') אינם משפיעים על הייצוג.
3. Searching – חיפוש במאגר של תמונות למציאת תמונת פנים הקרובה לתמונת הפנים שחולצה מהתמונה המקורית.

גישה פשטנית, כמו למשל בניית מסווג המכיל מספר יציאות כמספר הפנים אותם רוצים לזהות, הינה בעייתית משתי סיבות עיקריות: ראשית יש צורך באלפי דוגמאות לכל אדם (שלא ניתן בהכרח להשיג). כמו כן, נצטרך ללמד את המערכת מחדש בכל פעם שרוצים להוסיף מישהו חדש. כדי להתגבר על בעיות אלו מבצעים "למידת מטריקה" (metric learning) בה מזקקים פיצ'רים של פנים ויוצרים וקטור יחסית קצר, למשל באורך 128, המכיל את האלמנטים המרכזיים בתמונת הפנים. כעת נפרט את שלושת השלבים:

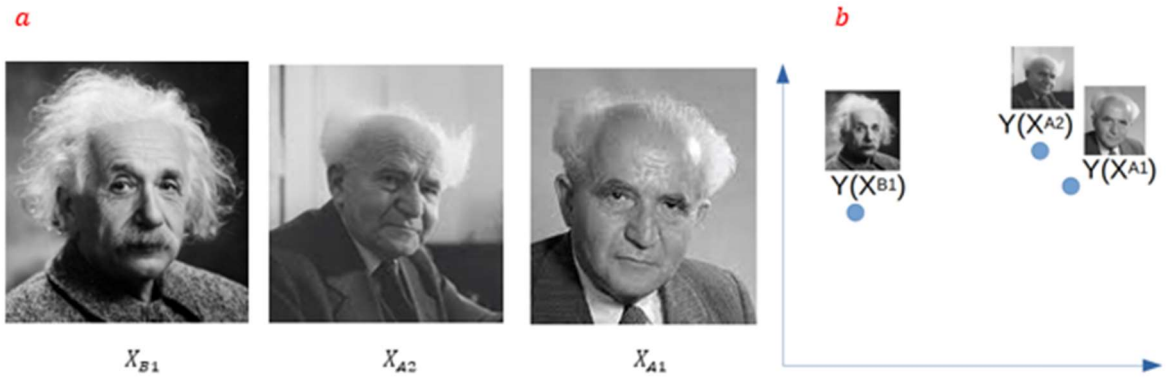
1. מציאת פנים.

כדי למצוא פרצופים בתמונה ניתן להשתמש ברשתות המבצעות detection, כפי שתואר בפרק 9.1. שיטה מקובלת למשימה זו הינה Yolo, המבוססת על חלוקת התמונה למשבצות, כאשר עבור כל משבצת בוחנים האם יש בה אובייקט מסוים, מהו אותו אובייקט, ומה ה-bounding box שלו.

2. תיאור פנים.

כאמור, המשימה בתיאור פנים נעשית בעזרת metric learning, כאשר הרעיון הוא לזקק פנים לוקטור שאינו מושפע מפיצ'רים שלא שייכים באופן מהותי לפנים הספציפיות האלה, כגון זווית צילום, רמת תאורה וכדו'. בכדי לעשות זאת יש לבנות רשת המקבלת פנים של בנאדם ומחזירה וקטור, כאשר הדרישה היא שעבור שתי תמונות של אותו אדם יתקבלו וקטורים מאוד דומים, ועבור פרצופים של אנשים שונים יתקבלו וקטורים שונים. למעשה, פונקציית ה-loss תקבל בכל פעם minibatch, ותעניש בהתאם לקרבה בין וקטורים של אנשים שונים וריחוק בין וקטורים של אותו אדם.

כעת נניח שיש לנו קלט X, המכיל אוסף פרצופים. כל איש יסומן באות אחרת – A,B,C, ותמונות שונות של אותו אדם יסומנו על ידי אות ומספר, כך שלמשל X_{A1} זוהי התמונה הראשונה של אדם A בסט הקלט X, וכמובן ש-X_{A1} ו-X_{A2} הן שתי תמונות של אותו אדם. באופן גרפי, בדו-ממד ניתן לתאר זאת כך (בפועל הווקטורים המייצגים פנים יהיו בממד גבוה יותר):



איור 9.1 a) דוגמאות מסט הפרצופים X. b) איך נרצה שהדאטה ימופה לממד חדש Y.

כאמור, נרצה לבנות פונקציית loss שמעודדת קירבה בין X_{A1} ו-X_{A2}, וריחוק בין X_{A1} ו-X_{B1}. פונקציית ה-loss מורכבת משני איברים, המודדים מרחק אוקלידי בין וקטורים שונים:

$$L = \sum_{X} \left\| Y(X^{Ai}) - Y(X^{Aj}) \right\| - \left\| Y(X^{Ai}) - Y(X^{Bj}) \right\|$$

כאשר האיבר הראשון ינסה להביא למינימום וקטורים של אותו אדם, והאיבר השני ינסה להביא למקסימום וקטורים של פרצופים שאינם שייכים לאותו אדם. כיוון שנרצה להימנע מקבלת ערכים שליליים, נוסיף פונקציית מקסימום.

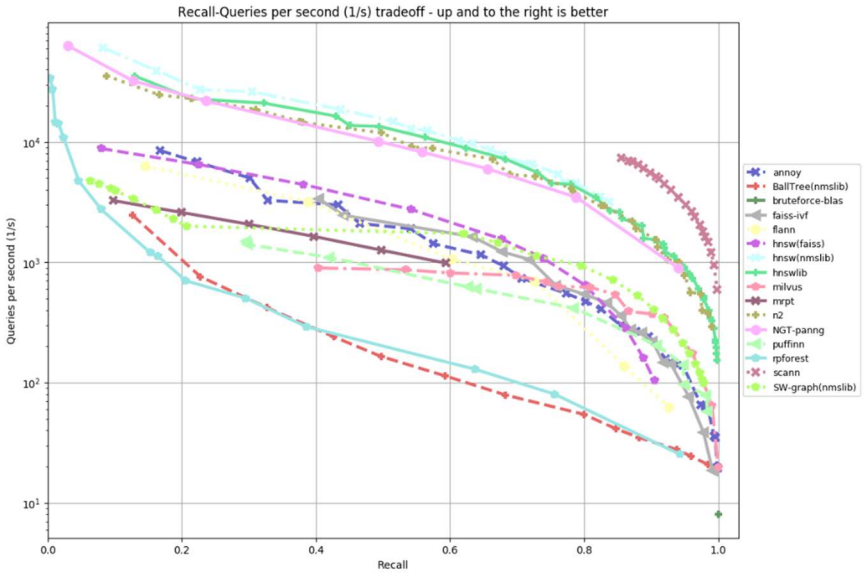
בנוסף, ניתן 'להרחיק' תוצאות של פרצופים שונים על ידי הוספת קבוע k, כך שהפרש בין המרחק של פרצופים של אנשים שונים לבין המרחק של פרצופים של אותו איש יהיה לפחות k:

$$L = \sum_{X} max(\|Y(X^{Ai}) - Y(X^{Aj})\| - \|Y(X^{Ai}) - Y(X^{Bj})\| + k, 0)$$

loss כזה נקרא triplet loss, כיוון שיש לו שלושה איברי קלט – שתי תמונות של אותו אדם ואחת של מישהו אחר. כאמור, הפלט של הרשת הנלמדת צריך להיות וקטור המאפיין פנים של אדם, ומטרת הרשת היא למפות פרצופים שונים של אותו אדם לווקטורים דומים עד כמה שניתן, ואילו פרצופים של אנשים שונים יקבלו וקטורים רחוקים זה מזה.

3. מציאת האדם

בשלב הקודם, בו התבצע האימון, יצרנו למעשה מאגר של פרצופים במרחב חדש. כעת כשיגיע פרצוף חדש, כל שנותר זה למפות אותו למרחב החדש, ולחפש במרחב זה את הוקטור הקרוב ביותר ולקטור המייצג את הפנים החדשות. בכדי לעשות זאת ניתן להשתמש בשיטות קלאסיות של machine learning, כמו למשל חיפוש שכן קרוב (כפי שהוסבר בחלק 2.1.3). שיטות אלו יכולות להיות איטיות עבור מאגרים המכילים מיליוני וקטורים, וישנן שיטות חיפוש מהירות יותר (ובדרך כלל המהירות באה על חשבון הדיוק). בעזרת השיטה המובילה כרגע (SCANN) ניתן להגיע לכמה מאות חיפושים שלמים בשנייה (החיפוש ב-100 מימדים מתוך מאגר של 10000 דוגמות).



איור 9.2 עבור פרצוף נתון, מחפשים עבורו וקטור תואם במימד החדש המכיל ייצוג וקטורי של הפרצופים הידועים. בכל שיטה יש טרייד-אוף בין מהירות החיפוש לבין הדיוק, ובגרף זה מוצגות שיטות שונות.

מלבד זיהוי וסיווג פנים, יש גם שיטות של מציאת אלמנטים של פנים הכוללות אף, עיניים וכו'. אחת השיטות המקובלות משתמשת בשערוך הצורה של פנים אנושיות, וניסיון למצוא את איברי הפנים לפי הצורה הסטנדרטית. בשיטה זו ראשית מבצעים יישור של הפנים והתאמה לסקאלה אנושית (על פי מרחק בין האיברים השונים בפנים), ולאחר מכן מטילים 68 נקודות מרכזיות על התמונה המיושרת, מתוך ניסיון להתאים בין הצורה הידועה לבין התמונה המבוקשת.

איור 9.3 זיהוי אזורים בפנים של אדם על ידי התאמת פנים לסקאלה אנושית והשוואה למבנה של פנים המכיל 68 נקודות מרכזיות.