טכנולוגיות דיפ־פייק

המונח דיפ־פייק (deepfake) מתאר תמונה, קטע קול או סרטון שעברו מניפולציה מלאכותית אך הם בעלי מראה מציאותי ואמין (Sayler & Harris, 2021). המונח דיפ־פייק הוא למעשה הֶלחם של למידה עמוקה (deep learning) וזיוף (fake), שכן שורשי מהפכת הזיוף העמוק נעוצים במהפכה טכנולוגית נוספת שהתחוללה לפני פחות מעשור: למידה עמוקה (Sejnowski, 2018).

למידה עמוקה (deep learning) היא תת־תחום של למידת מכונה, אשר עושה שימוש ברשתות נוירונים מלאכותיות (artificial neural networks). רשתות נוירונים אלה הן אלגוריתמים השואבים השראה מאופן הפעולה של רשת העצבים במוח האדם. הרשת העצבית לומדת על ידי תיקון הקשרים הרבים בתוכה; היא עורכת תיקונים קטנים באמצעות בחינת מידע רב על מנת להיטיב את דיוקה, וכך הפלט של נוירון אחד הוא הקלט של נוירון אחר. בזכות הצלחות ראויות לציון, רשתות הנוירונים הפכו עם השנים לנפוצות ביותר מבין הגישות של למידת המכונה, והן אחראיות למגוון הישגים בתחום הבינה המלאכותית ובתוכם: זיהוי פנים, זיהוי עצמים בתמונות, תמלול ותרגום, שליטה בכלי רכב אוטונומיים וברחפנים ועוד (ענתבי, 2020).

קיימות מספר שיטות ליצירת דיפ־פייק, הנפוצה ביותר היא בתחום הווידאו ומסתמכת על שימוש ברשתות נוירונים עמוקות (deep neural networks) הכוללות אלגוריתמים המשתמשים בטכניקה של החלפת פנים של אדם אחד בפניו של אדם אחר (face-swapping). טכנולוגיה זו פותחה ושוכללה על ידי מפתחים וקהילות אינטרנטיות של קוד פתוח במטרה ליצור אפליקציות נגישות וידידותיות למשתמש להחלפת פנים, כמו FakeApp ו־FaceSwap (Khalil & Maged, 2021).

האפליקציות המדוברות ודומותיהן עובדות לרוב באופן דומה. הן מתבססות על מספר תמונות או סרטונים של אדם א׳ — האדם שאנחנו רוצים שיופיע בסרטון — ומתוך מידע זה האלגוריתם רוכש ולומד את המראה, ההגייה וצורת הדיבור של אדם א׳. נוסף על כך יש צורך בסרטון בסיס שבו אדם ב׳ אומר או עושה את מה שנרצה שהוא יציג כאילו אדם א׳ עושה או אומר (אף שלא עשה או אמר). האלגוריתם יודע לשנות את פניו וקולו של אדם ב׳ כך שייראה כאילו אדם א׳ הופיע בסרטון. האפליקציה חותכת למעשה את הפנים מתוך כל תמונה נתונה ומאמנת שתי רשתות נוירונים מסוג אוטואנקודר (autoencoder) — אחת על פניו של המוחלף והשנייה על פניו של המחליף. לאחר האימון, האפליקציה מסוגלת לקבל כל תמונה של המוחלף, להחליף את פניו בפנים האחרות ולחברן לגוף באופן שנראה אמין.

רשת אוטואנקודר, שהיא הבסיס לתהליך זה, היא רשת נוירונית שמשחזרת את הקלט שהיא מקבלת בשכבת הפלט. היא מורכבת משני חלקים: מקודד (encoder) ומפענח (decoder). המקודד מקבל קלט — תמונת פנים, מעבד אותו ומחזיר פיסת נתונים קטנה יותר. אם למשל תמונת הפנים מיוצגת על ידי עשרות אלפי ערכים מספריים, אזי ״הנתונים הנסתרים״ (latent variables) — הפלט במוצא המקודד — יכולים להיות בסך הכול עשרות מספרים. כלומר, המקודד מעבד ומאבד מידע, אך השאיפה היא לאמן אותו כך שהמידע שיוציא ייצג את הקלט בדרך שתאפשר לשחזר את הקלט המקורי. המפענח מקבל את הקוד המקוצר ומטרתו היא לעבד אותו ולהרחיבו לגודלו מקורי, דהיינו שוב לתמונת פנים. האימון של שני המרכיבים —

המקודד והמפענח — נעשה במשותף, והגמול לכל אחד על הצלחתו הוא כאשר התמונה במוצא המפענח זהה לתמונה בכניסה למקודד. אימון מוצלח של אוטואנקודר הוא למעשה דחיסת נתונים או קידוד של הרבה מידע במעט מידע. ליכולת הזו שימושים רבים, ובמקרה של דיפ־פייק זהו ייצוג תמציתי של תמונת פנים של אדם א׳ שנועד לשחזר — באמצעות אוטואנקודר שאומן על מקבץ אחר — תמונת פנים של אדם ב׳. מדובר בשיטה פופולרית המאפשרת ליצור תוצרי דיפ־פייק בתוכנות חינמיות נגישות בקלות ובמחיר נמוך, כאשר התוצאות אמינות ומציאותיות יחסית (Sayler & Harris, 2021).

דרך נוספת ליצור דיפ־פייק עושה שימוש ברעיון שהגו איאן גודפלו(Ian Goodfellow) ועמיתיו ופורסם ב־2014 במאמר שכתבו, אשר הפך לאחד המצוטטים ביותר בשנים האחרונות בתחום מדעי המחשב (כ־35 אלף ציטוטים על פי מנוע החיפוש Google Scholar). הטכנולוגיה שגודפלו ועמיתיו יצרו מכונה Generative Adversarial Nets (להלן: GANs). דרך מערכת GAN יכולות שתי רשתות מחשבים להשוות נתונים זו עם זו בו־זמנית. מערכת GAN מורכבת משתי רשתות נוירונים מלאכותיים שלכל אחת מהן תפקיד מוגדר: האחת היא הרשת הגנרטיבית — הרשת שמייצרת מידע חדש. היא אמונה על יצירת נתונים מזויפים כגון תמונות, קטעי שמע או קטעי וידאו המשכפלים את המאפיינים של מערך הנתונים המקורי; הרשת השנייה היא הרשת הדיסקרימינטורית — המבחינה (מלשון to discriminate), שאמונה על זיהוי זיוף הנתונים (Goodfellow et al,. 2014).

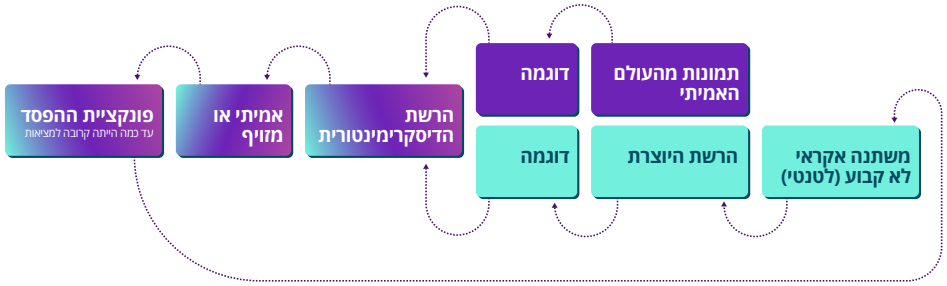
בשלב הראשון מאמנים את הרשת הגנרטיבית באימון בסיסי בנפרד מהרשת השנייה, כדי שתלמד כיצד נראות פנים אנושיות. לצורך זה המערכת מוזנת באלפי תמונות של דיוקנאות שונים, ומתוכם הרשת תסיק את המאפיינים הבסיסיים שמגדירים פנים ואת הקשרים החבויים בין המאפיינים הללו: מיקום העיניים, האף או הסנטר, מה מאפיין פנים נשיות ומה מאפיין פנים גבריות, וכדומה. המטרה אינה לימוד בעל־פה על ידי המערכת היכן נמצא כל איבר בפנים, אלא שהיא תגלה ותגדיר דפוסים כלליים ועקרוניים במידע שהיא מקבלת, כדי שבהמשך תהיה מסוגלת ליצור דיוקנאות חדשים ושונים זה מזה. שלב האימון הראשוני הזה חשוב כדי להביא את הרשת הגנרטיבית, היוצרת, לרמה של יכולת בסיסית. לאחר האימון הבסיסי של הרשת הגנרטיבית, גם הרשת הדיסקרימינטורית נדרשת לעבור אימון בסיסי בנפרד, כדי שתדע גם היא לזהות פנים אנושיות, וכך הן יהיו יריבות ראויות זו לזו (Goodfellow et al., 2014).

בשלב הבא מחברים בין הרשתות: מחברים את המוצא של הרשת הגנרטיבית לכניסה של הרשת הדיסקרימינטורית. מחברים לרשת הדיסקרימינטורית כניסה נוספת שלתוכה יוזנו תמונות של אנשים אמיתיים, לא מזויפים. הרשת הדיסקרימינטורית תידרש להבחין בין התמונות האמיתיות לבין התמונות המזויפות שהפיקה הרשת הגנרטיבית. כעת האימון מתחיל. כדי שהרשת הגנרטיבית תפיק תמונה מזויפת, ראשית יש להזין לתוכה מספר אקראי כלשהו. המספר האקראי הזה ישמש גרעין, שעליו יכולה הרשת הגנרטיבית לבנות את הפנים הספציפיות שהיא מציירת, בעזרת הידע שצברה על אודות הקשרים החבויים שבין חלקי הפנים. האקראיות הזו תבטיח שהרשת הגנרטיבית תפיק בכל פעם תמונה חדשה וייחודית. התמונה החדשה הזו תועבר לרשת הדיסקרימינטורית ולצידה תמונה אמיתית, לא מזויפת. הרשת הדיסקרימינטורית תבחן את התמונות ותקבע מי מהן אמיתית ומי מזויפת. אם הרשת הדיסקרימינטורית צדקה, סימן שהרשת הגנרטיבית לא עשתה עבודה טובה מספיק. המערכת תשנה את עוצמת הקשרים בין הנוירונים ברשת הגנרטיבית כדי

לנסות לשפר אותה ותנסה שוב ליצור תמונה. היא תיצור תמונה חדשה שונה מעט מקודמתה: פנים שונות, שיער אחר וכדומה. אם הרשת הדיסקרימינטורית טעתה וחשבה שהתמונה המזויפת היא בעצם תמונה אמיתית, סימן שהתמונה לא טובה מספיק, ולכן המערכת תשחק עם הקשרים בין הנוירונים של הרשת הדיסקרימינטורית ותנסה לשפר ולחדד את יכולת ההבחנה שלה.

תהליך זה נמשך בלולאה: הרשתות ממשיכות להתחרות זו בזו, לעיתים קרובות במשך אלפי או מיליוני חִזָרורים (איטרציות). בכל סיבוב שתי הרשתות ישתפרו בהדרגה עד התוצאה הרצויה, והיא שהרשת הדיסקרימינטורית תטעה בזיהוי התמונה המזויפת במחצית מהמקרים. בסיום האימון אפשר לפרק את מערכת ה־GAN ולשלוף ממנה את הרשת הגנרטיבית — שכן מטרת האימון היא ליצור רשת גנרטיבית טובה שאפשר יהיה להשתמש בה לתכליות נוספות (Citron & Chesney, 2019 ;Goodfellow et al., 2014).

רשתות יריבות יוצרות



לשיטת GAN יש יתרון משמעותי — אימון רשתות הנוירונים ללא מעורבות אנושית — מה שמכונה "לימוד לא מפוקח" (unsupervised training). כל ההחלטות של המערכת וכל התיקונים והשינויים ברשתות הנוירונים עצמן נעשים באופן עצמאי לחלוטין, ואין צורך באדם שיתערב בתהליך ויתייג כל תמונה כמזויפת או אותנטית (Salimans et al., 2016).

בשימוש ב־GANs כמעט בלתי אפשרי להבחין בעין אנושית לא מאומנת בין אמת לשקר, אולם השימוש ב־GANs מצריך כמות עצומה של נתוני אימון, והוא יעיל ליצירת תמונות סינתטיות ולא סרטונים. עם זאת, ההערכה הרווחת בקרב מומחים היא כי GANs יהיו המנוע העיקרי לפיתוח דיפ־פייק מתוחכם בעתיד (Adee, 2020).