הפקולטה להנדסה

המעבדה לעיבוד אותות

Deep audio-visual-based sound multi sources localization and tracking

אלין גולן

איתי אלמליח

פרויקט שנה ד' לקראת תואר ראשון בהנדסה

מנחה: גב' רננה שרה אופוצינסקי

מנחה אקדמי: פרופ' שרון גנות

אוקטובר 2023

תוכן עניינים

תודות 4

תקציר 5

I הצגת הבעיה 6

תיאור הבעיה 6

השיטות המוצעות: 7

תיאור הפתרון: 7

II רקע תיאורטי 8

למידת מכונה 8

Supervised Learning: 8

Unsupervised Learning: 8

למידה עמוקה 9

Deep Neural Networks (DNNs): 9

Transfer Learning: 9

Deep Fusion for Audio-Visual Integration: 10

Challenges and Research Directions: 10

רשתות נוירונים 10

Activation function: 10

Softmax function: 11

Fully connected NN: 11

CNN – Convolutional NN: 11

ResNet - Residual neural network: 12

RNN - Recurrent neural network: 12

מדד *Mean Average Precision (mAP)*: 13

III מבנה המערכת 14

***TALKNET* 14**

IV יישום ומימוש 18

הדאטה 18

היפר פרמטרים 19

Optimizer: 19
Learning rate: 19
Preprocessing **.....20**
תהליך האימון 20
V הצגת התוצאות 22
VI דיון ומסקנות: 26
רעיונות להמשך 27
סיכום: ..... **שגיאה! הסימניה אינה מוגדרת.**
ביבליוגרפיה ונספחים: 28

תודות

העבודה הזו בוצעה תחת ההנחיה של גב' רננה שרה אופוצינסקי ופרופ' שרון גנות, מהפקולטה להנדסה, אוניברסיטת בר-אילן. אנו מודים למנחה שלנו על כל התמיכה וההדרכה המתמדת במהלך הפרויקט, ועל הזמן שהקדישה לענות על שאלותינו.

תקציר

האם השמע שייך לקול אנושי? האם ישנה תנועת שפתיים? האם השמע מסונכרן עם תנועת השפתיים? ממצאים קוגניטיביים שאלו חלק מהפרמטרים לפיהם בני אדם מזהים האם אדם מדבר.

בעולם שבו התקשורת האודיו-ויזואלית היא חלק בלתי נפרד מחיינו, יש חשיבות לזיהוי הדובר. Active speaker detection (ASD) עוסק בזיהוי ואיכון דובר, אחד או יותר, בסצנה ויזואלית. משימה זו היא חזית חיונית עבור מגוון רחב של יישומים כגון יומן דובר, מיקוד מחדש של וידאו לפגישות, מערכות תמלול, שיפור דיבור ואינטראקציה בין אדם לרובוט.

במסגרת פרויקט הזה נחקור את משימת ה-ASD, נלמד באופן יסודי את תחום הלמידה עמוקה אליו אנו נחשפים לראשונה, לאחר מכן אנו נבחן מערכת מרובת מודלים לסיווג ASD המתבססת על המאמר "Is Someone Speaking? Exploring Long-term Temporal Features for Audio-visual Active Speaker Detection" נבין את כל שלבי המימוש ותתי הרשתות המרכיבות את המערכת על מנת לממש ולחלץ את תוצאות המערכת. לבסוף, ננתח את התוצאות באופן מפורט ונסיק את המסקנות הנדרשות.

הצגת הבעיה

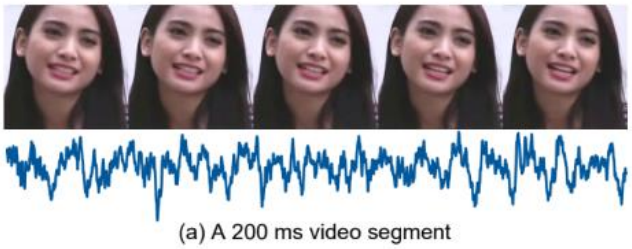
תיאור הבעיה

בעיית זיהוי דוברים פעילים (ASD) שואפת לזהות מי מדבר בסצנה ויזואלית של דובר אחד או יותר. ASD מוצלח תלוי בפרשנות מדויקת לטווח קצר ולטווח ארוך של האודיו והוידאו וגם ניתוח הקשר בין המידע הוויזואלי לקלט השמע.

נציין מספר אתגרים עיקריים בבניית מודל ASD מוצלח:

1. <u>Short-term ASD:</u> מערכות לפרקי זמן קצרים מחלצות פיצ'רים אודיו-חזותיים מקטע קצר באורך קבוע (למשל: 200ms, 300ms). המאפיינים הללו מייצגים את הרמזים הבולטים למשימת ה-ASD ולכן גם ניתן לראות המון מחקרים העוסקים בתחום זה. אם זאת, כפי שמומחש באיור העליון (a), קשה לשפוט האם התבצעה פעולת דיבור מקטע וידיאו כה קצר. עבור סרטון ארוך יותר בן 2 שניות, כפי שמוצג באיור התחתון (b) ההבחנה יותר ברורה. כך גם במציאות, כאשר בני אדם מזהים דובר, הם לוקחים בחשבון שנאמר משפט, משפט יכול להתפרש על פני מאות פריימים.

נוכל לטפל בבעיה ע"י הגדלת גודל הסגמנט, נוכל לקבל את המאפיינים הממוצעים במחיר של רזולוציית הזמן של פעילות הדיבור.



(a) A 200 ms video segment

(b) A 2-second video segment

2. <u>סנכרון בין מודלים:</u> על מנת לסווג דובר כפעיל צריך להתייחס לסנכרון בין המודלים כגון: דיבור-שפה, דיבור-פנים (לדוגמה: יכולות להיות תנועות שווא של שפתיים) וכן גם לסנכרון בין אותות האודיו לוויזואליים כאשר משתמשים במערכת לפרקי זמן קצרים קשה לסנכרן בין 2 אותות הקלט. עבור אותות הוידאו, היחידה המינימלית היא תמונה סטטית. ואילו עבור אותות השמע, היחידה המינימלית היא עשרות אלפיות שנייה.

השיטות המוצעות:

ישנם מחקרים רבים על ASD באמצעות אודיו, וידאו ושילוב של שניהם. זיהוי דוברים מתוך אודיו נקרא voice activity detector, תחום זה אנו לומדים לזהות נוכחות של דיבור בניגוד לרעשים אקוסטיים אחרים. החיסרון בהסתמכות על שיטה זו בלבד היא שבעולם האמיתי אותות אודיו נקלטים ע"י מיקרופונים מרוחקים וכתוצאה מכך הם מעורפלים מטבעם, דבר המציב אתגרים למשימת הVAD. מחקרים אחרים מציעים זיהוי דוברים מתוך קטע הוידאו, לצורך העניין ניתוח תנועות פנים ופלג גוף עליון עוזרות לזהות דובר ויזואלי. עם זאת, הביצועים נמוכים עקב קורלציה נמוכה בין תנועות הגוף ופעילות הדיבור (חיוך אכילה וכו' עלולים לפגוע בביצועי ה-ASD). מסקנות אלו הובילו להבחנה שקצב הדיבור והגיית המילים נמצאים בקורלציה הדוקה עם תנועת הפנים ולכן עיבוד אודיו-ויזואלי משיג שיפור משמעותי בשימוש בקשר זה.

מחקרים רבים מציגים גישות שונות לביצוע משימת ה-ASD המושפעת מהנחות היסוד שנקבעות תחילה ומתוך כך המודל שנבחר למימוש המשימה. נציג מספר גישות קיימות:

- <u>ASD כ-assignment task:</u>
  הנחת היסוד: דיבור שזוהה חייב להיות שייך לאחד הדוברים על המסך (לא תמיד נכון, יכול להיות דיבור מחוץ לפריים)
- <u>ASD כמשימת סיווג:</u>
  1. הערכת הפרצופים בפריים.
  2. שרשור תכונות אודיו-וידיאו שחולצו כקלט למסווג בינארי רב-שכבתי (MLP) לזיהוי הרמקול (/הדובר) הפעיל בכל קטע וידיאו קצר - ללא התחשבות בתלות בין פריימים זמניים.
  3. backend classifier עם מבנה זמני כמו רשת עצבית חוזרת (RNN) עם GRU, יחידה חוזרת משוערת ו-LSTM, זיכרון לטווח קצר-ארוך, שהשיג הצלחה ראשונית.

המאמר שאנו מסתמכים עליו בפרויקט מונע מגישת הסיווג באמצעות RNN לפי קלט אודיו-ויזואלי.

תיאור הפתרון:

הפרויקט שלנו מתבסס על השיטה המתוארת במאמר " Is Someone Speaking? Exploring Long-term Temporal Features for Audio-visual Active Speaker Detection". מחקר זה, מציג גישה חדשנית (לכשיצא ב2020) המבוססת על תכונות קצרות טווח ותכונות ארוכות טווח בניגוד למחקרים קודמים שהשתמשו בתכונות קצרות טווח בלבד. מערכת TalkNet המוצעת במאמר מורכבת ממקודדי אודיו ווידיאו זמניים, ממנגנון cross-attention וממנגנון self-attention שעליהם נרחיב בהמשך.

II רקע תיאורטי

בפרק זה יעסוק בבסיס התיאורטי של הפרויקט ויכלול דיון מקיף על התיאוריות, המודלים והעקרונות המדעיים המרכזיים הרלוונטיים למחקר שלנו.

למידת מכונה

למידת מכונה (ML) היא אבן הפינה של הבינה המלאכותית המודרנית (AI), הממלאת תפקיד מרכזי בהפיכת נתונים גולמיים לתובנות ניתנות לפעולה. בתחום של זיהוי רמקולים פעילים (ASD), למידת מכונה היא מרכיב בסיסי המאפשר למחשבים ללמוד דפוסים ולקבל החלטות מושכלות, ובכך לאפשר זיהוי ומעקב אחר רמקולים בתוכן אודיו-ויזואלי. חלק זה בוחן את ההיבטים השונים של למידת מכונה כשהם קשורים לפרויקט ASD.

Supervised Learning:

Supervised Learning משמשת אבן דרך לפיתוח מודל ASD. בפרדיגמה זו, מודלים מאומנים באמצעות מערכי נתונים מסומנים, כאשר דגימות קלט משולבות עם תוויות פלט מתאימות. גישה מפוקחת זו מאפשרת למודל ללמוד את הקשר בין תכונות הקלט לבין לוקליזציית הדובר או תוצאות המעקב הרצויות. אלגוריתמי למידה מפוקחים נפוצים המיושמים ב-ASD כוללים:

- **רגרסיה לינארית:** טכניקת יסוד למשימות רגרסיה, רגרסיה ליניארית מבקשת לבסס קשר ליניארי בין תכונות קלט ותוויות פלט. בעוד רגרסיה ליניארית היא פשוטה, ייתכן שהיא לא תלכוד דפוסים לא ליניאריים מורכבים בנתונים אודיו-ויזואליים.
- ***Support Vector Machines (SVMs):*** SVMs הם מסווגים מגוונים המשמשים ב-ASD עבור משימות סיווג בינאריות ורב-מחלקתיות. הם שואפים למצוא את המישור המפריד בצורה הטובה ביותר את מרחב תכונת הקלט למחלקות נפרדות, מה שהופך אותם מתאימים להבחנה בין דוברים ללא דוברים.
- **עצי החלטה ויערות אקראיים:** עצי החלטה ועמיתיהם במכלול, יערות אקראיים, משמשים ליצירת מבנים היררכיים המחלקים את מרחב נתוני הקלט. שיטות אלה יכולות להתמודד הן עם משימות סיווג והן במשימות רגרסיה והן משמשות ללכידת גבולות החלטה מורכבים ב-ASD.

Unsupervised Learning:

ב-ASD, שבו נתונים מסומנים יכולים להיות מוגבלים, טכניקות למידה ללא פיקוח נכנסות לפעולה. למידה ללא פיקוח מתמקדת בחילוץ דפוסים או מבנים משמעותיים מנתונים לא מסומנים. אמנם גישה זו אינה מסתמכת על תוויות פלט מוגדרות מראש, אך היא בעלת ערך עבור משימות שונות, כולל:

- **Clustering Algorithms:** אלגוריתמי אשכולות, כגון ממוצעי k ואשכולות היררכית, מקבצים נקודות נתונים דומות על סמך התכונות שלהן. זה יכול להיות שימושי ב-ASD לזיהוי אשכולות של רמקולים או קטעים של נתונים אודיו-ויזואליים עם מאפיינים דומים.
- ***Dimensionality Reduction Techniques:*** שיטות הפחתת מימדיות, כמו Principal Component Analysis (PCA) ו-t-Distributed Stochastic Neighbor Embedding (t-

(SNE, שואפות להפחית את המורכבות של נתונים בעלי ממד גבוה. על ידי הפיכת נתונים לייצוגים בממד נמוך יותר, טכניקות אלו מפשטות את החילוץ וההדמיה של תכונות ASD.

למידה עמוקה

למידה עמוקה היא תת-תחום של למידת מכונה שחולל מהפכה באופן שבו אנו ניגשים למשימות מורכבות בבינה מלאכותית (AI). היכולת יוצאת הדופן שלו ללמוד ייצוגים היררכיים באופן אוטומטי מנתונים הפכה אותו לאבן יסוד ביישומי AI מודרניים, כולל זיהוי רמקולים פעיל (ASD). בחלק זה, אנו מתעמקים במורכבות של למידה עמוקה, בארכיטקטורות הרשת העצבית שלה, והשפעתה העמוקה על תחום ASD.

Deep Neural Networks (DNNs):

בליבת הלמידה העמוקה נמצאות רשתות עצביות עמוקות (DNNs). רשתות אלו מורכבות משכבות מרובות של צמתים או נוירונים מחוברים, והן אחראיות למונח "עמוק" בלמידה עמוקה. DNNs מצטיינים בלכידת דפוסים ותכונות מורכבות מנתונים. בהקשר של ASD רשתות ה-DNNs חשובים במיוחד לעיבוד המידע העשיר הזמין הן באודיו אודיו והן בוויזואליות.

Convolutional Neural Networks (CNNs):

CNNs הם מחלקה של DNNs המותאמים בעיקר לנתונים דמויי רשת, כגון תמונות ומסגרות וידאו. הם הוכיחו שהם חלק בלתי נפרד במשימות ראייה ממוחשבת. CNNs מעסיקים שכבות קונבולוציוניות כדי לחלץ אוטומטית תכונות מרחביות מנתוני הקלט. במשימת ה-ASD רשתות CNNs מיומנים בניתוח רמזים חזותיים, כולל הבעות פנים ומחוות, שהן חיוניות ללוקליזציה של הדוברים.

Recurrent Neural Networks (RNNs):

RNNs מתוכננים לטיפול בנתונים רציפים על ידי הכנסת קטעים מחזוריים. קטעים אלה מאפשרים לרשת לשמור על זיכרון של כניסות קודמות, מה שהופך אותם מתאימים למשימות עם תלות זמנית. בהקשר של ASD, שבו זרמי אודיו מכילים לעתים קרובות מידע רציף כמו דפוסי דיבור, RNNs ממלאים תפקיד מרכזי בעיצוב ההיבטים הזמניים של נתוני אודיו.

Transfer Learning:

סימן ההיכר של למידה עמוקה הוא היכולת שלה למנף מודלים מאומנים מראש על מערכי נתונים גדולים. ב-ASD, שבו נתונים מתויגים יכולים להיות נדירים ויקרים להשגה, Transfer Learning חשובה ביותר. על ידי כוונון עדין של מודלים שהוכשרו מראש, החוקרים יכולים להתאים אותם למשימה הספציפית של זיהוי רמקולים. Transfer Learning מפחיתה משמעותית את זמן האימון ודרישות הנתונים תוך שיפור ביצועי המודל.

Deep Fusion for Audio-Visual Integration:

הכוח האמיתי של למידה עמוקה מגיע כאשר משלבים מידע ממגוון דרכים, כגון אודיו וויזואלי. שיטות היתוך עמוק מנצלות רשתות עצביות עמוקות כדי לחלץ תכונות ברמה גבוהה באופן עצמאי מכל אופנה ולאחר מכן למזג אותן בשכבה עמוקה יותר, מה שמשפר את היכולת של המערכת לבצע לוקליזציה ולעקוב אחר רמקולים במדויק. האיחוד של רמזים אודיו וחזותיים ב-ASD יכול לשפר משמעותית את החוסן של זיהוי הרמקולים בסביבות שונות.

Challenges and Research Directions:

למרות הצלחתה המדהימה, למידה עמוקה ב-ASD מתמודדת גם עם אתגרים. אתגר משמעותי אחד הוא הצורך במערכי נתונים גדולים ומגוונים להכשרת מודלים חזקים. בנוסף, פרשנות והסבר מודלים הם קריטיים ביישומים כמו ASD כדי להבטיח שקיפות ואמינות.

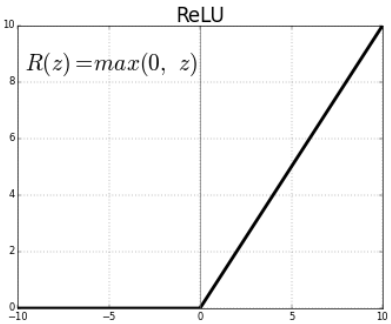
לסיכום, למידה עמוקה חוללה מהפכה בזיהוי רמקולים פעילים על ידי מתן כלים רבי עוצמה לאינטגרציה אודיו-ויזואלית, מודלים זמניים וחילוץ תכונות. CNNs ו-RNNs, יחד עם Transfer Learning, שיפרו משמעותית את הדיוק והחוסן של מערכות ASD. ככל שהתחום ממשיך להתפתח, טיפול באתגרים וחקירת כיווני מחקר חדשים יובילו ללא ספק לפתרונות זיהוי רמקולים מתקדמים עוד יותר.

רשתות נוירונים

עד כה, דנו בכלליות על למידה עמוקה ועל הרשתות השונות. כעת, נעמיק במושג של רשת נוירונים ובמושגים הנלווים אליו. רשת נוירונים, הוא מבנה מתמטי חישובי אשר פותח בהשראת תהליכים קוגניטיביים המתרחשים ברשת עצבית טבעית. רשת נוירונים הינה חיבור של יחידות עיבוד בסיסיות (נוירונים מלאכותיים) על ידי משקלים ופונקציות לא ליניאריות. רשת נוירונים נקראת עמוקה אם היא מכילה יותר משכבה חבויה אחת.

Activation function:

רשתות נוירונים מורכבות משכבות ליניאריות ולא ליניאריות לסירוגין. הפונקציה הלא ליניארית המפרידה בין השכבות הליניאריות נקראת פונקציית אקטיבציה. פונקציית האקטיבציה הנפוצה ביותר בשנים האחרונות, וגם זו שנשתמש בפרויקט הינה ReLU.



היתרונות של ReLU שהיא זולה חישובית ושה-gradient decent הסטוכסטי מתכנס מהר.

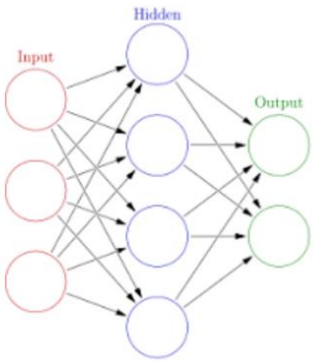
Softmax function:

Softmax הוא הכללה של binary logistic regression לסיווג למס' מחלקות. בהינתן וקטור של k מספרים ממשיים מחזירה וקטור להתפלגות הסתברות של K תוצאות אפשריות. לעתים קרובות, נעשה שימוש בפונקצית softmax כפונקציית ההפעלה האחרונה של רשת עצבית כדי לנרמל את הפלט של הרשת להתפלגות הסתברות על מחלקות פלט חזויות, בהתבסס על אקסיומת הבחירה של Luce.

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad \text{for } i = 1, \ldots, K \text{ and } \mathbf{z} = (z_1, \ldots, z_K) \in \mathbb{R}^K$$

Fully connected NN:

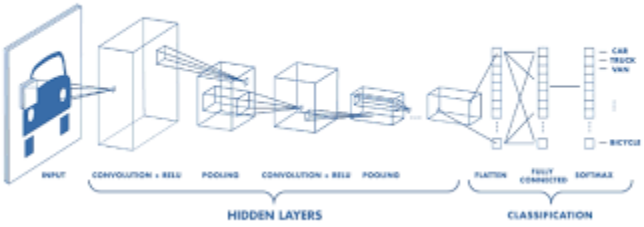
מודלי רשתות נוירונים מאורגנים לרוב בשכבות נפרדות של נוירונים. קיימים סוגים שונים של רשתות נוירונים, כאשר ניתן לשלב בין מספר ארכיטקטורות באותו מודל. הסוג הבסיסי ביותר נקרא Fully NN connected – כל נוירון מחובר לכל הנוירונים בשכבה הסמוכה ואינו מחובר לאף נוירון משכבתו.



מבנה סכמתי של רשת נוירונים בעלת שכבה חבויה אחת. ערכו של כל נוירון נקבע לפי סכמה מסוימת של כל הנוירונים מהשכבה הקודמת המחוברים אליו.

CNN – Convolutional NN:

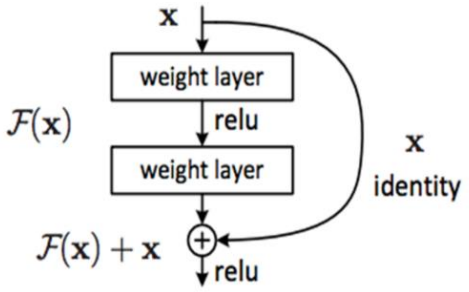
רשת CNN נפוצה כאשר עוסקים במידע חזותי, שכן מבנה הרשת משמר את היחסים המרחביים בין המאפיינים. בארכיטקטורה זו, בכל שכבה אנו מבצעים קונבולוציה דו מימדית של המאפיינים עם סדרת פילטרים, שאת הפרמטרים שלהם המודל לומד ומעדכן על מנת למקסם את הביצועים. כך ממדי התמונה מצטמצמים, והמודל בעצם יכול להתייחס למאפיינים מרחביים הולכים וגדלים של הקלט. כיוון שהמודל שלנו עוסק בסרטונים, אנו משתמשים בקונבולוציות בשלושה ממדים, בכדי לשמר מאפיינים בממד הזמן, ולא רק בממדים המרחביים.



מבנה סכמתי של CNN. בכל שכבה, ממדי התמונה מצומצמים ע"י סדרת פילטרים
במטרה למצוא מאפיינים מרחביים אשר מסייעים להחלטת המודל בתהליך הלמידה.

ResNet - Residual neural network:

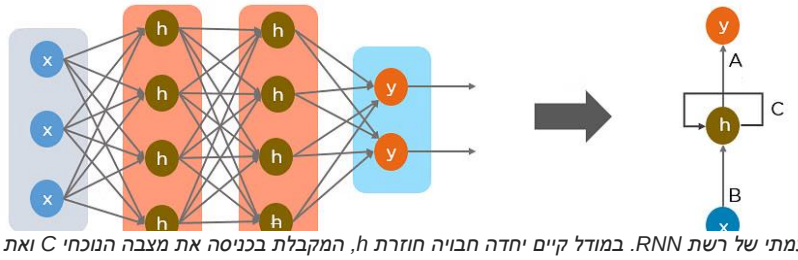
טכניקה זו עוזרת לנו להימנע מבעיית הגרדיאנטים המתאפסים שיש לה נטייה לקרות במודלים עם שכבות עמוקות ע"י זיכרון למרחק גדול יותר לאחור בעומק השכבות ברשת. בטכניקה זו אנו יוצרים מעקף בין שכבה קדומה יותר לשכבה מאוחרת יותר וכך מתקיים הזיכרון.



מבנה סכמתי של ResNet. במודל קיים מעקף כך שחלק מהמידע מדלג על
כמה שכבות ללא שום למידה כלשהי ולאחר מכן עושים חיבור כלשהו עם המידע שנלמד.

RNN - Recurrent neural network:

רשת עצבית חוזרת (RNN) היא סוג של רשת עצבית מלאכותית שמשתמשת במידע רציף, כלומר, היא מזינה את הפלט של השלב הקודם יחד עם הקלט של השלב הנוכחי. זה מאפשר ל-RNN לזכור מה היא למדה באיטרציות השונות, ולכן היא מתאימה לעיבוד נתונים רציפים או סדרתיים, כמו טקסט, שמע או סדרות זמן. תכונה זו מאפשרת לרשת לזהות דפוסים וקורלציות בנתונים. בפועל RNN מצליחה במיוחד בתחום של עיבוד שפה טבעי (NLP) ומשמשת למשימות כמו זיהוי דובר, דגמי שפה יצירת טקסטים ועוד.



מבנה סכמתי של רשת RNN. במודל קיים יחדה חבויה חוזרת h, המקבלת בכניסה את מצבה הנוכחי C ואת המצב הקודם B

רשת עצבית חוזרת (RNN) מורכבת מכמה יחידות, כאשר כל יחידה מקבלת קלט מהיחידה הקודמת ומעבירה את הפלט שלה ליחידה הבאה. זה מאפשר לרשת לשמור על מצב נוכחי שמשפיע על הקלטים הבאים. במבנה הסטנדרטי שלRNN, ישנן שלושה משקולות: אחד מהקלט למצב הנוכחי (W_x), אחד מהמצב הקודם למצב הנוכחי (W_h), ואחד מהמצב הנוכחי לפלט (W_y).

חישוב המצב הנוכחי מתבצע ע"י הנוסחה הבאה:

$$curr = f(input * W_x + prev * W_h)$$

כאשר f היא פונקציית ההפעלה, כמו tanh או ReLU.

חישוב המצב הנוכחי מתבצע ע"י הנוסחה הבאה:

$$output = g(curr * W_y)$$

כאשר f היא פונקציית ההפעלה, כמו softmax או sigmoid, תלוי במשימה.

מדד Mean Average Precision (mAP):

ממוצע של ערכי הדיוק הממוצע (AP) על פני כל הקטגוריות. הדיוק הממוצע (AP) מחשב את הממוצע של ערכי הדיוק לכל ערך של ה .Recall-במילים אחרות, הנוסחה לחישוב ה-mAP היא:

$$mAP = \frac{1}{|C|}\sum_{c=1}^{|C|} AP(c)$$

כאשר:

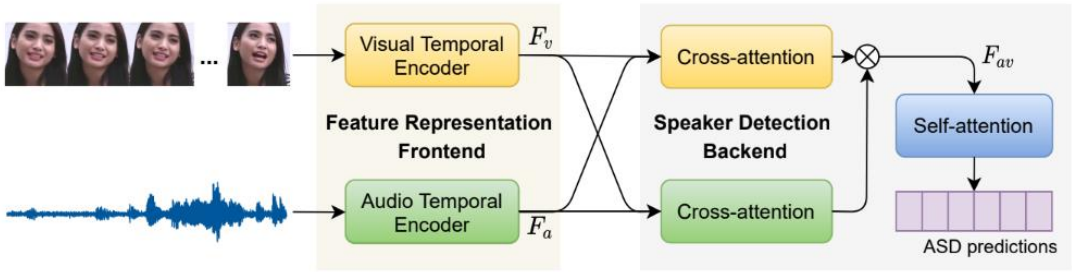
- |C| הוא מספר הקטגוריות.
- $AP(c)$ הוא הדיוק הממוצע של הקטגוריה c.

הדיוק הממוצע (AP) מחושב כך: $AP = \int_0^1 p(r)dr$

כאשר $p(r)$ הוא הדיוק ברמת ה-recall r. במילים אחרות, השטח מתחת לעקומת ה- precision recall.

III מבנה המערכת

בחלק זה נעמיק אודות מבנה המערכת, תתי המערכות המרכיבות אותה וזרימת הנתונים דרך המערכת כולה. נתחיל בתיאור כללי, לאחריו נפרט על כל אחת מתתי המערכות בצורה נרחבת, נסביר אודות ארכיטקטורת רשתות הנוירונים, נבין את קלטי המערכת ואת תוצריה.



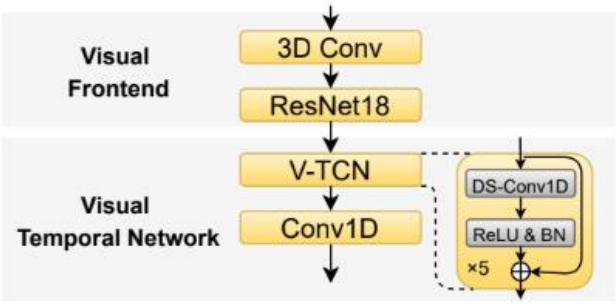
תיאור כל שלבי המערכת והזרימה בין השלבים השונים

TALKNET

ארכיטקטורת TalkNet נועדה להתמודד עם האתגרים בזיהוי רמקולים פעילים (ASD) על ידי שילוב של תכונות זמניות לטווח קצר ולטווח ארוך. המערכת כוללת שלושה מרכיבים עיקריים: מקודדים זמניים אודיו וויזואליים, מנגנון אודיו-ויזואלי של Cross-Attention ומנגנון Self-Attention. TalkNet הוא מערכת פיפליין מקצה-לקצה הלוקחת קטע אודיו-ויזואלי ← וקובעת האם האדם מדבר בכל פריים בוידאו. המערכת מורכבת משני שלבים, השלב הראשון אחראי על מיצוי התכונות של קלט ובשלב השני מתבצע זיהוי הדוברים.

בהרחבה, השלב הראשון מכיל מקודד שמע ומקודד חזותי שלומדים את הייצוג הזמני של הקלט. השלב השני מורכב ממנגנון **cross-attention** כדי ללכוד ראיות בין-מודליות (כלומר מתרחשות גם באותות האודיו וגם בוידאו). וממנגנון **self-attention** ללכידת ראיות דיבור ארוכות טווח תוך-מודליות ברמת האמירה. ולבסוף, מתבצע סיווג תוך מתן דגש להקשר הזמני לטווח ארוך ולאינטראקציות בין-מודאליות. TalkNet מדגימה שיפורים משמעותיים ביחס למערכות קיימות בתרחישים מאתגרים בעולם האמיתי.

מקודד זמני-חזותי:



מקודד זמני חזותי שואף ללמוד את הייצוג לטווח ארוך של דינמיקת הבעות הפנים. כפי שמוצג בתרשים, המקודד החזותי מורכב משני חלקים:

visual frontend - חילוץ מידע מרחבי מכל וידאו-פריים. כולל שכבת קונבולוציה תלת ממדית ואחריה בלוק ResNet18 שמטרתו להפריד בין קולו של הדובר באזורי שפתיים בווידאו המתאים באודיו מרובה דוברים. חלק זה מקודד את זרם הוידאו-פריים לרצף של הטבעות מבוססות פריימים.

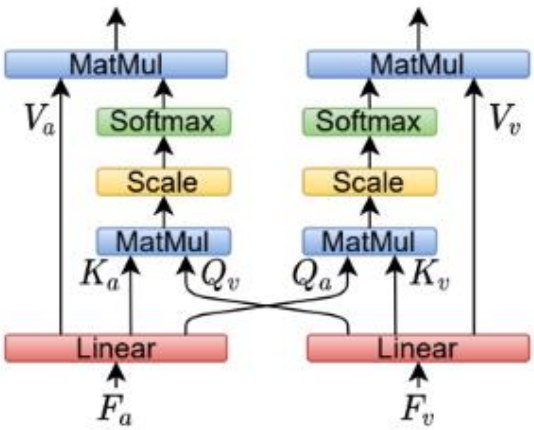
visual temporal network - רשת זו לוכדת מבנים מרחביים-זמניים חזותיים לטווח ארוך. מיוצגת על ידי בלוק פיתול זמני של וידאו (V-TCN), כוללת 5 יחידות לינאריות מקושרות שיורית (ReLU), נורמליזציה של אצווה (BN) ושכבות פיתול הניתנות להפרדה מבחינה עומק (DS Conv1D).

Audio Temporal Encoder:

מקודד שמע זמני שואף ללמוד את ייצוג תוכן האודיו מהדינמיקה הזמנית.
המקודד מורכב מרשת 2D ResNet34 עם מודול סחיטה-ועירור (squeeze-and-excitation) ללימוד ייצוג תוכן אודיו מדינמיקה זמנית. סחיטה-ועירור היא טכניקה לשיפור ייצוג תכונות השמע באמצעות כיול מחדש של חשיבות ערוץ התכונות הקונבולוציוניות. הרשת מקבלת כקלט רצף מסגרות שמע, אשר כל פריים שמע מיוצג על ידי וקטור מקדמי MFC (Mel-frequency cepstral) באורך 13 שהם מאפיינים הלוכדים את המאפיינים הספקטרליים של הצליל. והמקודד מייצר רצף של הטבעות אודיו (Fa) כפלט, כלומר רצף של וקטורים המקודדים את ההקשר הזמני של הצליל חלון הזמן המכיל מסגרות אודיו מרובות. המקודד מבטיח שרזולוציית הזמן של הטבעות האודיו תואמת את זו של ההטבעות החזותיות כדי להקל על מנגנוני הקשב הבאים.

Audio-visual Cross-Attention:

המוטיבציה לשימוש במנגנון Cross-Attention נובעת מהעובדה שסנכרון אודיו-ויזואלי הוא רמז אינפורמטיבי לפעילות הדיבור. מנגנון Cross-Attention משפר את היכולת ללמוד אינטראקציות בין-מודאליות. המנגנון מורכב משכבת קשב, שכבת הזנה קדימה, חיבור שיורי ושכבת נרמול. הפלטים של רשת תשומת לב צולבת משורשרים לאורך הכיוון הזמני ליצירת תכונה אודיו-ויזואלית משותפת.
*כיוון שלזרימת אודיו ולזרימת וידיאו יש דינמיקה ייחודית, הם לא בדיוק מיושרים בזמן. היישור האודיו-ויזואלי עשוי להיות תלוי בגורמים חיצוניים כמו התנהגות הרמקולים. לכן אנו מציעים 2 רשתות cross-attention לאורך הממד הזמני כדי לתאר באופן דינמי אינטראקציה אודיו-ויזואלית כזו.



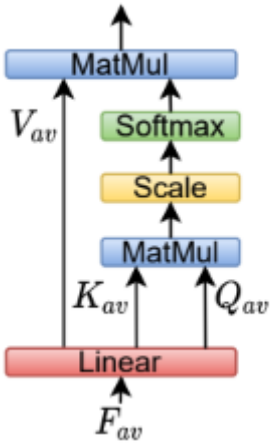
קלט המנגנון מתקבל מההטמעות אודיו וויזואלית בשלב הקודם והוא: וקטורי השאילתה (Qa, Qv), מפתח (Ka, Kv) וערך (Va, Vv). באמצעות פעולת softmax מחושב הדמיון בין וקטורי השאילתה, המפתח והערך מהטבעות האודיו והוויזואליות. יצירת תכונת הקשב-שמע F_{av} נעשית ע"י הגדרת תכונת האודיו F_a כמקור ואת התכונה החזותית F_v כיעד (חישוב F_{va} להפך). וכן מוצאי הרשת הן תכונת הקשב-שמע F_{av} ותכונת הקשב-חזותי F_{va} כדלהלן:

$$F_{a\to v} = softmax(\frac{Q_v K_a^T}{\sqrt{d}})V_a$$

$$F_{v\to a} = softmax(\frac{Q_a K_v^T}{\sqrt{d}})V_v$$

<u>Self-Attention and Classifier:</u>

רשת Self-Attention מיושמת לאחר ה-Cross-Attention. , תוך התמקדות במודלים של מידע זמני ברמת הבעה אודיו-ויזואלית. מבנה הרשת דומה ל-Cross-Attention, אך בשונה, קלט הרשת הינו התכונה האודיו-ויזואלית המשותפת, Fav. הוא משתמש בתכונות אודיו-ויזואליות משותפות (Fav) עבור שאילתות, מפתחות וערכים, תוך הבחנה בין מסגרות מדברות ובלתי מדברות.



Loss Function:

השלב האחרון במערכת הוא ביצוע fully connected layer ולאחריה פעולת softmax, המקרינה את הפלט של רשת הקשב העצמי לרצף תוויות ASD. אנו רואים ב-ASD משימת סיווג ברמת המסגרת. לבסוף, נעשה שימוש ב-cross-entropy loss כדי למדוד את השונות בין תוויות ASD חזויות ותוויות אמת מבוססות עבור כל פריים וידאו. הנוסחה כוללת את הלוגריתם השלילי של ההסתברות החזויה עבור המחלקה הנכונה, והניתנת ע"י:

$$Loss = -\frac{1}{T}\sum_{i=1}^{T}(y_i \cdot \log s_i + (1 - y_i) \cdot \log (1 - s_i))$$

כאשר si, yi הם תוויות משוערות ואמיתיות בהתאמה, T מספר הוידיאו-פריימים.

הגדלת שמע עם דגימה שלילית:

כדי לשפר את עוצמת הרעש, TalkNet מציגה טכניקת הגדלת אודיו באמצעות דגימה שלילית.

במהלך האימון, רצועת אודיו של סרטון אחד נבחרה באופן אקראי כרעש, ומציעה רמקולים של רעש והפרעות בתוך התחום ממערך האימון עצמו.

IV יישום ומימוש

בחלק זה נפרט אודות יישום ומימוש הפרויקט. נדון במערך הנתונים שבו השתמשנו, בתהליך העיבוד המקדים של הנתונים, באופן מימוש הקוד, בספריות המרכזיות בהן השתמשנו, ועוד.

הדאטה

בפרויקט זה למידת המודל מתבצעת ע"י מערך נתונים אודיו-ויזואלי בשם AVA-ActiveSpeaker. מערך הנתונים נגזר מסרטים הוליוודיים וגודלו 29,723 קטעי וידיאו לסט האימון, 8,015 לסט הוולידציה ו-21,316 לסט הבדיקה. כאשר אורך ה- video utterances נעות בין 1 ל-10 שניות. הדאטה בנוי מרצועות פנים מסווגות באופן זמני בסרטונים, כאשר כל מופע פנים מסומן כמדבר או לא, והאם הדיבור נשמע. מערך הנתונים מכיל כ-3.65 מיליון פריימים מסומנים, כ-38 שעות של רצועות פנים והשמע המתאים.

<u>תיוג מערך הנתונים:</u>

ישנן 3 מחלקות לסיווג כל פריים בסרטון - לא מדבר (NS), מדבר ונשמע (S&A), מדבר אך לא נשמע (S&NA). מדבר אבל לא נשמע מכסה מקרים שבהם אדם עשוי להיראות מדבר למרות שהדיבור שלו אינו נשמע בפס הקול. למחלקה זו יש השפעה עבור גישות חזותיות בלבד.

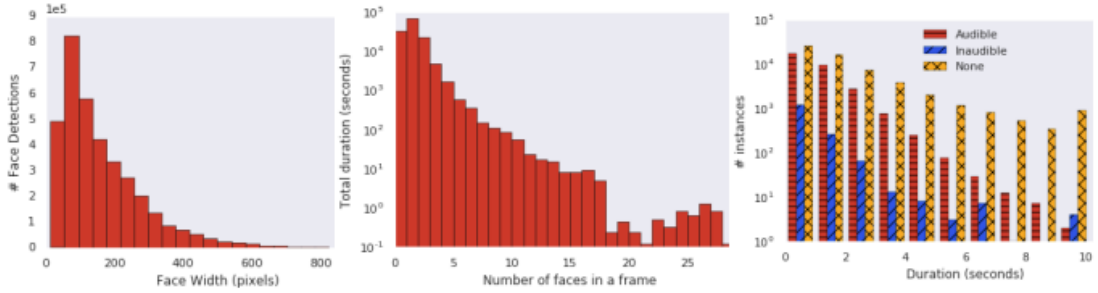
Label	Time	# Segments	Mean Duration
NS	28.10 hours	58, 171	1.74 seconds
S&A	9.46 hours	30, 623	1.11 seconds
S&NA	0.35 hours	1, 547	0.83 seconds

טבלה 1: מתארת את הסטטיסטיקה המצטברת על מערך הנתונים של AVA-ActiveSpeaker עבור שלוש התוויות: לא מדבר, מדבר ושמע, מדבר אך לא נשמע.

הגרפים הבאים מתארים את מאפייני הנתונים, כוללים התפלגויות של גדלי הפנים שזוהו בסרטונים, מספר הפנים שזוהו במקביל, התפלגות ואופי תנאי הרעש בסרטון. מידע זה מסייע להבנה עמוקה יותר של בסיס הנתונים, מה הוא מכיל ואולי גם אילו אתגרים עשויים להיות בזיהוי הדובר.

Label	Visible	Not Visible
Clean Speech	35%	23%
Speech with Music	23%	28%
Speech with Noise	42%	50%

טבלה 2: תנאי רעש התואמים לרמקולים גלויים.



חלוקת רוחב פנים מסומנת: פנים קטנות יותר מאתגרות יותר לזיהוי. (ב) (log) משך הזמן הכולל לעומת #faces במסגרת: ככל שמספר הפנים הגלויות גדול יותר, כך המשימה קשה יותר. (ג) התפלגות אורכי הסגמנטים לפי תווית: קטעים קצרים בסרט הופכים את המעקב אחר הדוברים למאתגר.

מערך הנתונים AVA-ActiveSpeaker מציג מספר אתגרים:

- מכיל סרטים במגוון שפות.
- קצב הפריים לשנייה (fps) של הסרטים משתנה.
- מספר לא מבוטל של סרטונים בעלי תמונות מטושטשות ואודיו רועש.
- מכיל סרטים ישנים עם דיאלוגים מדובבים.

כל אחד מהגורמים האלה מקשה על היכולת לבצע סנכרון מדויק של האותות האודיו-ויזואליים.
← ולכן עיבוד מקדים של הנתונים עשוי לסייע בקבלת תוצאות טובות יותר.

היפר פרמטרים

Optimizer:

אופטימייזר הוא אלגוריתם שמשמש למציאת הפרמטרים האופטימליים של המודל. האופטימייזר מנסה למצוא את הפרמטרים שממזערים את פונקציית ה-loss בכל epoch, כלומר את השגיאה של המודל על נתוני האימון. קיים מגוון רחב של אופטימייזרים, אנו משתמשים ב-Adam optimizer.

אלגוריתם Adam במושג שנקרא Exponentially Moving Average שמאפשר לו לקחת בחשבון את הערכים של ה-epochs הקודמים. זה מאפשר לו להתאים את גודל הצעדים שהוא עושה בהתאם לשיפוע, כלומר, להקטין את הצעדים כשהמדרון תלול ולהגדיל אותם כשהמדרון רדוד.

Learning rate:

Learning rate הוא משתנה שמשפיע על המהירות שבה מודל למידת מכונה "לומד" או מתאים את עצמו לנתונים. במילים אחרות, הוא משפיע על כמה מהמידע החדש שהמודל מקבל משפיע על המידע הישן. במהלך האימון, ה-Learning rate משפיע על גודל הצעד שהמודל עושה בכל איטרציה. ישנה חשיבות לקביעת ערך טוב לפרמטר זה, שכן אם ה-Learning rate גבוה מידי, האימון עשוי לדלג מעל המינימום שהוא מנסה למזער. מצד שני, אם ה-Learning rate נמוך מידי האימון עשוי לקחת זמן רב מידי או להסתיים במינימום מקומי שאינו אופטימלי.

במקרה שלנו, ה-Learning rate מוגדר להיות 4-10, והוא מופחת ב-5% בכל איטרציה. זה מאפשר למודל להתאים את עצמו לנתונים בצורה מדודה וממוקדת.

Preprocessing

העיבוד המקדים הוא שלב חיוני שבו מתבצעת הכנה וניקוי של הנתונים לפני שהם מוזנים למודל. המטרה היא להפוך את הנתונים לצורה שבה המודל יכול לעבוד באופן מיטבי ולהגיע לתוצאות מדויקות יותר. במקרה שלנו, לא ביצענו עיבוד מקדים מורכב והוא מתבצע בשלב האימון. למרות שהנתונים נשארים במצבם הגולמי, אנו משתמשים בטכניקת האוגמנטציה הוויזואלית כדי להרחיב את מאגר הנתונים שלנו ולהגדיל את הגיוון של התמונות. זה נעשה על ידי היפוך אקראי, סיבוב, וחיתוך של התמונות המקוריות. פעולה זו משפרת את היכולת של המודל להתמודד עם תמונות חדשות ולא ידועות מראש.

בנוסף, כדי לעזור למודל להתמודד עם סיטואציות שונות של רעשים ותנאים אקוסטיים מתבצעת גם אוגמנטציית שמע על ידי שימוש במקורות נוספים ממאגרי נתונים של RIRs ו-MUSAN. הם מכילים קטעי שמע שונים כמו רעשי רקע, שיחות, מוזיקה ועוד.

ולסיום, מתבצעת דיגמה שלילית המטרה להכניס למודל דוגמאות שאינן מכילות את התכונה או המאפיין שאנו מעוניינים לזהות. זה מאפשר למודל ללמוד להבחין בין דוגמאות חיוביות לשליליות ולשפר את היכולת שלו לזהות את התכונה או המאפיין בנתונים חדשים. במערכת TalkNet, הדיגום השלילי נדגם ממערך האימון של AVA-ActiveSpeaker מה שמאפשר להוסיף רעשים שמקורם באותו המאגר שממנו הוא מתאמן. במילים אחרות, המודל מתאמן לא רק על הנתונים ה"נקיים", אלא גם על נתונים שמכילים רעשים שונים.

תהליך האימון

תחילה, נציין את הספריות המרכזיות שעליהן אנו מתבססים למימוש המערכת. PyTorch, ספריית הלמידה עמוקה שמשמשת ליישום TalkNet עם כלי האופטימיזציה של Adam. בנוסף, Official Tool, ספריה המשמשת להערכת הביצועים של TalkNet במערך הנתונים של AVA-ActiveSpeaker.

נקודה חשובה לציון, אנחנו יכולים להציג רק תוצאות איכותיות (ויזואליות) מהנתונים של ה-test, ולא תוצאות כמותיות, כלומר, להעריך כמותית את הביצועים, מאחר ואין לנו גישה לתיוגים של ה-test ממערך הנתונים. AVA-ActiveSpeaker הוא מאגר נתונים שמשמש כאתגר שנתי בו סטודנטים מתמודדים להשגת דיוק מרבי כדי לנצח בתחרות. מסיבה זו תיוג ה-test אינו גלוי לציבור, וכדי להשתמש בו נדרשת הרשאה שאין ברשותנו.

כעת, נדון בתהליך האימון. המודל שלנו אומן באופן מקבילי על פני 4 מעבדים גרפים (GPU) שונים, כל אחד בגודל של כ-11 ג'יגה. במהלך תהליך האימון, ביצענו 25 אפוקים, כאשר בכל אפוק השתמשנו בקבוצות נתונים (batches) בגודל של 1500 ושיעור למידה (learning rate) של 0.0001. משך הריצה של כל אפוק היה כ-4 שעות, כלומר, סה"כ תהליך האימון נמשך כ-100 שעות. את הדיוק אנו מערכים לפי מדד mAP (עליו דיברנו בהרחבה ברקע תיאורטי) , ככל שהמדד גבוה יותר כך המודל מדויק יותר.

החוקרים במאמר מדווחים על דיוק מירבי (mAP) של 92.3% בסט הוולידציה ו-90.8% בסט נתוני הבדיקה. במימוש שלנו, השגנו דיוק של 91.92%, ההבדל נובע מהקטנת גודל ה- batch ל-1500 במקום 2500 כפי שאימנו החוקרים במקור. הסיבה לכך היא בשל מחסור במשאבי זיכרון שהיו ברשותנו (זהו הבדל מינורי, שנוכל לראות את ההשפעות שלו בוויזואליזציה בפרק הצגת התוצאות). התגובה הזו לשינוי גודל ה-batch הינה הגיונית כיוון שאנו יודעים שככל שה-batch size גבוה, כלומר, מספר הדוגמאות שנלקחות בכל איטרציה של האלגוריתם האימון גבוה יותר, כך איכות המודל גבוה יותר כיוון שהמודל מתאים את עצמו למגוון רחב יותר של דוגמאות.

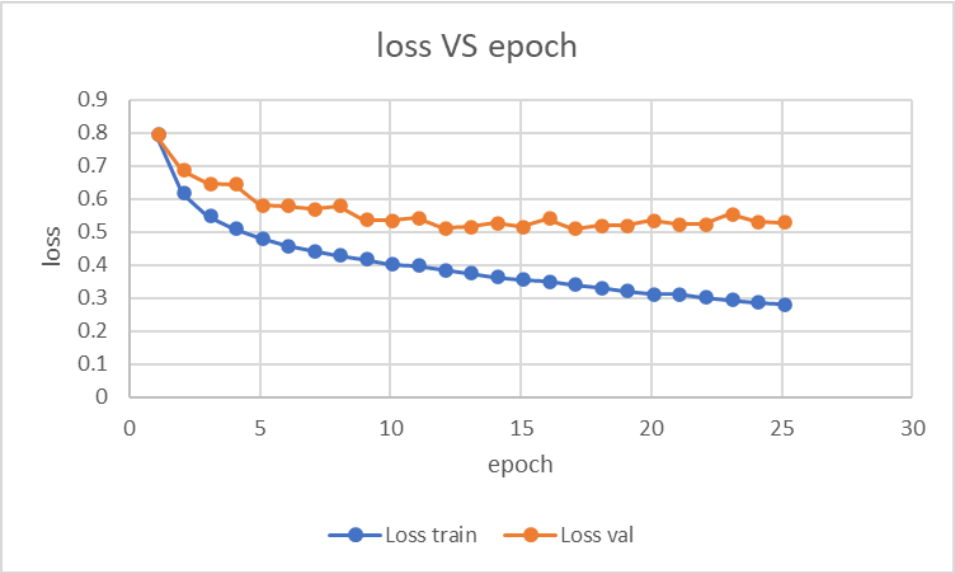
בנוסף, כתבנו סקריפט שמחשב את מדדי ה-ACC וה-loos מייצר מהן גרפים שמתארים את התנהגות המודל במהלך האימון. את הגרפים הללו נראה וננתח בפרק הבא. זאת על מנת שנוכל לבחון את איכות הלמידה של המודל והתוצאות שהוא משיג.

V הצגת התוצאות

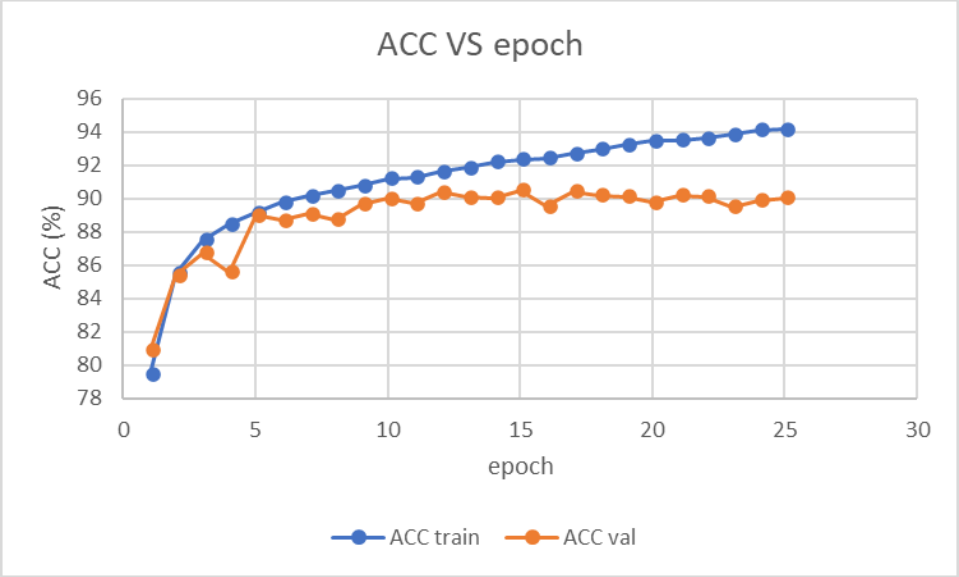
בחלק זה נציג את תוצאות המודל. נעשה זאת על ידי הצגת גרפים של פונקציית ה- Loss וה-Accuracy עבור הרשת אשר מעידים על אימון נכון ולמידה של המודל. כמו כן, נראה את פלט רשת ה- TalkNet, על ידי ויזואליזציה של סרטונים מה-test dataset ונראה כי אכן מזהה דוברים בסרטון.

Loss And Accuracy

האיור מטה מתאר את פונקציית ה- Loss וה- Accuracy של רשת ה- TalkNet כתלות במספר ה- epoch. הקו הכחול מתאר את ערכי ה- loss וה- Accuracy במהלך האימון והקו הכתום במהלך הוולידציה.



הגרף למעלה מציג את התנהגות ה-loss כתלות במספר ה-epoch במהלך אימון הרשת. כפי שאנו מצפים, ניתן לראות ירידה בערך ה- loss ככל שמספר ה-epoch עולה. דבר זה מראה אימון בריא של המודל, כלומר, ככל שהמודל מתאמן יותר על סט האימון כך משפר את למידת השגיאה שלו לנתוני האימון וכן גם על נתוני הוולידציה. במילים אחרות זה מצביע על כך שהמודל מצליח להכליל היטב. בתחום בו 15<epoch<25 ישנה התייצבות של ערך ה-loss בוולידציה מה שיכול לסמן שהמודל מתחיל להתאים את עצמו יותר מידי נתוני האימון (נכנס ל-overfitting).



מדד ה- ACC מתאר את ביצועי המודל על סט האימון וסט הוולידציה (בהתאם למקרא) כתלות במספר הepoch. מהתבוננות בגרף, נוכל להבחין בעלייה אקספוננציאלית מגמתית ועקבית של המדד עבור סט האימון ולבסוף מגיע לרוויה, מה שמעיד על אימון תקני של הרשת. הערך המירבי אליו מגיע ה-ACC בנתוני האימון הוא 94.03% . עליית המדד עשויה להצביע על התקדמות ביכולת המודל לסיים את האימון וללמוד מספר רב של אפוקים בצורה יעילה.

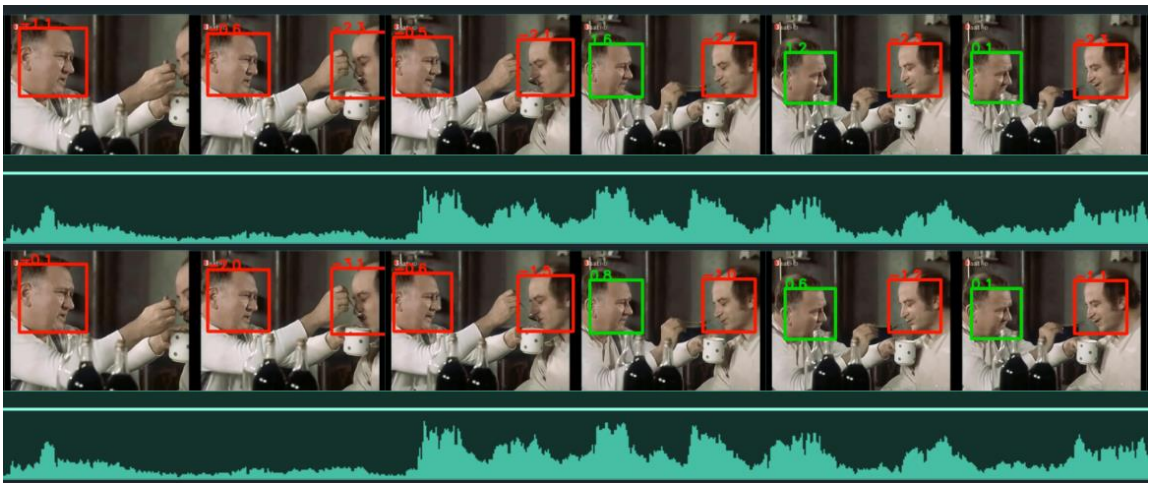
עבור סט הוולידציה, בשונה, נוכל להבחין בעלייה לא יציבה של מדד ה-ACC והתכנסות של האימון על נתוני הוולידציה. הערך המירבי אליו מגיע מדד ה- ACC בנתוני הוולידציה הוא 90.08%. חוסר היציבות של המדד על פני האפוקים השונים עשוי להצביע על אתגרים נוספים בזיהוי האובייקטים בסט הולידציה.

עבור שני הסטים, סט הוולידציה וסט האימון ניתן להבחין בצורה ברורה בתהליך למידה. אנו רואים קורלציה בין עלייה בדיוק בסט האימון לעלייה בדיוק של סט הוולידציה, כפי שציפינו. במילים אחרות, העלייה בדיוק על הוולידציה מצביעה על כך שהמודל מצליח להכליל היטב ואינו מאומן מידי (overfitting) על סט האימון.

כעת, נציג תוצאות ויזואליות של סרטוני וידאו חדשים (שהמודלים לא אומנו עליהם) מתוך סט הנתונים של ה-test, ונערוך השוואה בין המודל שהוצג במאמר למודל שאימנו. התוצאות שנציג הן רק חלק מהתוצאות שבחנו. כפי שציינו בעבר, המודל שהוצג במאמר משיג דיוק של 92.3%, בעוד שמודל שלנו משיג 91.92%. נציג מקרים שבהם המודל שהוצג במאמר מגיע לתוצאות טובות יותר מהמודל שלנו, מקרים שבהם שני המודלים מגיעים לאותה תוצאה, ומקרים שבהם המודל שלנו מגיע לתוצאות טובות יותר.

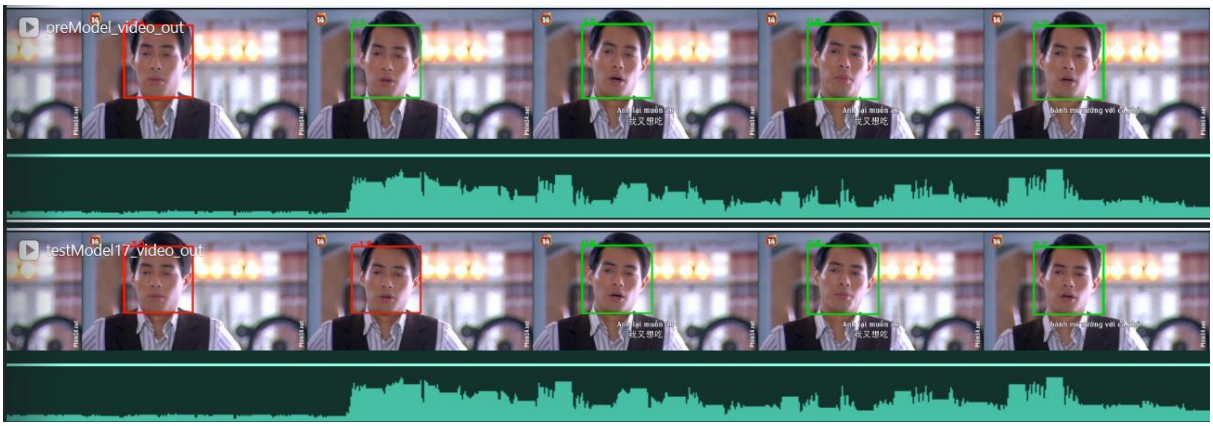
שימו לב, מבנה הוויזואליזציה לכל התוצאות שנציג מכיל סגמנט אודיו-וידאו עליון שנוצר מהמודל שהוצג במאמר וסגמנט אודיו-וידאו תחתון שנוצר מהמודל שאימנו.

תוצאות זהות בשני המודלים:

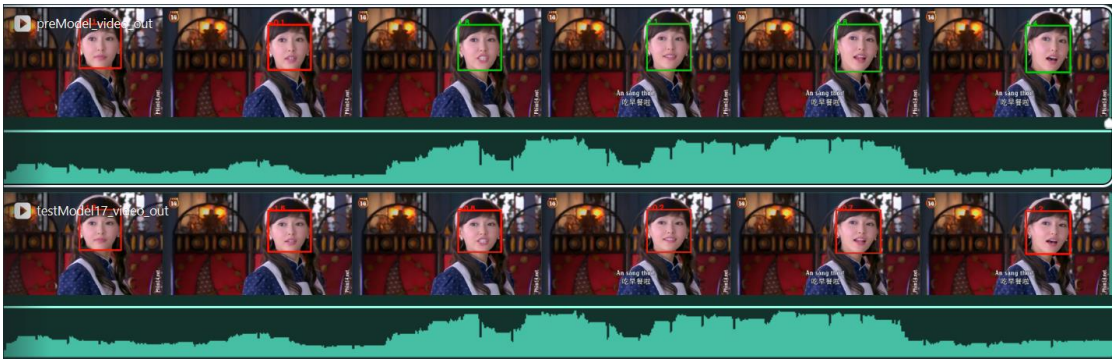


בקטע הנ"ל מופיע 2 דמויות מצולמות מהפרופיל שזוהו ע"י הרשת, אחד הדמויות הוא דובר שמתחיל לדבר באמצע הסגמנט, והוא זוהה נכונה ע"י 2 המודלים. נוכל להבחין שישנם הבדלים קלים במשקלי הסיווג של שני המודלים, שנגרם מהשוני המינורי בין 2 המודלים כתוצאה משינוי פרמטר ה-batch size.

תוצאות בהם המודל שהוצג במאמר מגיע לתוצאות טובות יותר:



בסרטון מופיע דובר יחיד שמתחיל לדבר במהלך הסגמנט, ניתן להבחין בבירור ע"י התבוננות בסגמנט האודיו. גם במודל המוצג במאמר וגם במודל שלנו ישנו זיהוי דובר נכונה, אפשר לראות שהמודל שאנחנו אימנו מזהה את הדובר עם דיליי של פריים אחד לעומת המודל שהוצג במאמר



בקטע הבא מופיעה דמות יחידה המתחילה לדבר באמצע הסגמנט, נוכל לראות זאת לפי סגמנט השמע וגם סגמנט הוידיאו.
נשים לב, כי במקרה זה המודל שאימנו אינו מצליח לזהות כלל את הדמות כדוברת, בעוד שהמודל שהוצג במאמר כן.

תוצאות בהם המודל שאימנו מגיע לתוצאות טובות יותר:

בקטע הנ"ל מוצג דמות אחת שאינו מדבר, לרוב מצולמת מהפרופיל. נוכל לראות את שהמודל המוצג במאמר שוגה בזיהוי דובר בעוד שהמודל שאימנו איננו מסווג את הדמות כדובר לאורך כל הסגמנט.

בקטע המוצג מופיעים 2 דמויות המצולמות מהפרופיל, כאשר הדמות השמאלית מדבר ב-2 אמירות במהלך סגמנט הנתונים, נוכל לראות את השפעת הדיבור בסגמנט השמע. הסגמנט העליון שמייצג את המודל המוצג במאמר לא מצליח לזהות את הדובר לאורך כל הסגמנט ואילו בסגמנט התחתון, הרשת שאימנו מזהה נכונה את הדובר ב-2 מקרי הדיבור.

VI דיון וסיכום:

במהלך הפרויקט, חקרנו באופן מעמיק את משימת ה-Active speaker detection , השקענו בלמידת היסודות של למידה עמוקה, כולל מושגים בסיסיים וסוגים שונים של רשתות נוירונים, באופן עצמאי באמצעות קורס אינטרנטי של סטנפורד. לאחר חיפוש ממושך, בחרנו לעבוד עם רשת ASD מורכבת ומרובת מודלים, שדרשה מאיתנו לבצע התאמות זיכרון ולכתוב קטעי קוד לחילוץ תוצאות המודל. לשם כך, נדרשנו להבין את כל שלבי המימוש ואת כל תתי הרשתות המרכיבות את המערכת הכוללת. בסוף, הצלחנו להגיע לתוצאות בעלות דיוק גבוה ולנתח אותן באופן מעמיק. במהלך הפרויקט, נחשפנו לתחומים חדשים, מערכות וטכניקות שלא הכרנו בעבר, נדרשנו להבין אותם ולהתאים אותם לצורכי המערכת שמימשנו.

בגרפים שהצגנו ראינו כי תהליך האימון של הרשת מתנהל באופן תקני, התנהגות המדדים כתלות במספר ה-epochs היא טובה ומראה שהרשת בתהליך של למידה, שיפור והתכנסות.

רעיונות להמשך

פיצוח משימת ה-ASD מורכבת דו שלבי הכולל מיצוי תכונות הקלט וצבירת הקשר זמני-מרחבי.
בנוגע להמשך פיתוח המערכת לקבלת תוצאות טובות יותר וייעול המודל, נוכל לבחון 2 גישות.

<u>גישה ראשונה:</u> הוספת פרדיקציה נוספת לאימון.
השערה זו סובבת סביב הרעיון שניתוח נוסף של רכיב האודיו או הרכיב החזותי יכול להוביל לשיפורים משמעותיים. דבר זה כרוך בחילוץ וייצוג של תכונות אודיו באופן שמשפר את כוח ההבחנה של המודל. תכונות אלו עשויות להכיל הן מאפיינים זמניים לטווח קצר והן לטווח ארוך של אותות אודיו, מה שיאפשר לוקליזציה טובה יותר של רמקולים בסביבות אקוסטיות מורכבות.

<u>גישה שניה:</u> עיבוד מקדים.
כאשר בחנו את מערך הנתונים AVA-ActiveSpeaker הן בקריאת מאמרים עליו והן באופן ידני, נוחכנו לדעת כי הוא מכיל דאטה יחסית גולמי ולא מעובד. ולכן יתכן שעיבוד מקדים פשוט למערך הנתונים יניב שיפור משמעותי למודל.

לסיכום, הפתרון המוצע מבקש להתבסס על TalkNet ועל ידי חידוד השיטה הקיימת באמצעות שילוב טכניקות חדשניות בניתוח אודיו וביצוע ניסויים יסודיים של עיבוד מקדים נוכל להשיג לשפר את תוצאות המודל לזיהוי רמקולים פעילים. מאמץ זה עולה בקנה אחד עם המטרה הרחבה יותר של קידום יישומים רב-מודאליים כגון זיהוי דיבור אודיו-ויזואלי, יומן דובר ומעקב אחר רמקולים.

ביבליוגרפיה ונספחים:

[1] Ruijie Tao, Zexu Pan, Rohan K. Das, Xinyuan Qian, Mike Z. Shou, and Haizhou Li. "Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection." (2021)

[2] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gal-lagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, et al. 2020. "AVA active speaker: An audio-visual dataset for active speaker detection". In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2020. IEEE, 4492–4496.

[3] Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, and Icksang Han. 2020. In defence of metric learning for speaker recognition. In Interspeech 2020.

[4] https://cs231n.github.io/

[5] https://github.com/AlinGolan/Deep-Learning-in-Hebrew

[6] קישור לדרייב המכיל את הסרטונים עם תוצאות הרשת
https://drive.google.com/drive/folders/1YhkoeLXBpdkpRaxrxq_qxSfsC3w4LeFI?usp=sharing