Sound as input
for machine learning models

Image

vs

Mean math values

spectral_centroid
1379.421175

קול כקלט למודלים של בינה

ייצוג כתמונה

לעומת

ייצוג כערכי חציונים מתמטיים

spectral_centroid
1379.421175

מגיש: לב טוניק (tuniklev@Gmail.com, 307366450)

מנחה: פרופ' שאול מרקוביץ'

מבוא

1. רקע כללי ומוטיבציה ראשונית לפרויקט:

בסמסטרים הקודמים יצא לי להיחשף לתחום הבינה המלאכותית מזוויות שונות- בקורס מבוא נחשפתי לרקע תאורטי, ובשני פרויקטים מימשתי בפועל פתרונות לבעיות שונות (להלן לינקים לרקע על הפרויקטים הקודמים: "זיהוי אנשים במצוקה", "Automated Bokeh effect").
עם זאת, ככל שלמדתי יותר בקורסים הנ"ל, הרגשתי רצון עז לבדק דברים נוספים בתחום, ולכן ביקשתי מפרופ' שאול מרקוביץ' שינחה אותי בביצוע עבודה אשר תחקור מספר נושאים שעניינו אותי.

מטרת העל של עבודה זו הינה לבדק כיצד שינוי אופן ייצוגם של קבצי שמע, משפיע על דיוק הניבויים, כאשר ההשוואה תתבצע בין שני הייצוגים הבאים:

1. ייצוג של קבצי שמע בתור תמונה (שהינה בעצם גרף של תכונה מסוימת או מספר תכונות)
2. ייצוג של קבצי שמע בתור ערכי החציונים המתמטיים של וקטור התכונות.

- הבעיה תוסבר בהרחבה בעמוד הבא.

מטרות המשנה לעבודה זו (אשר נולדו לאחר העבודות הקודמות ככיוונים מעניינים לבדיקה):

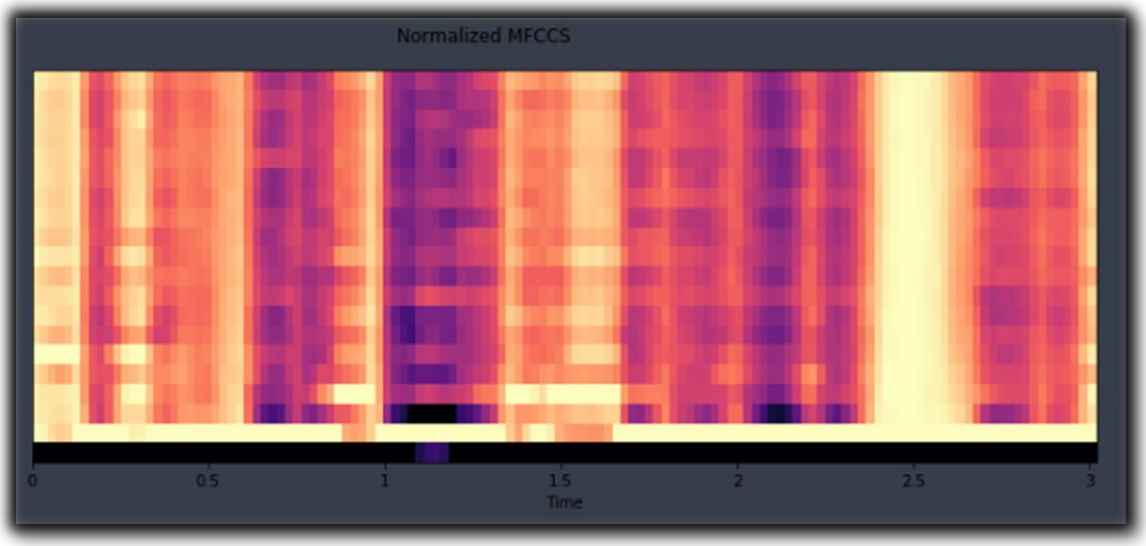
1. התנסות בביצוע Transfer Learning ולקיחת מודלים אשר הוכחו בעבר כמעולים לפיתרון בעיות סיווג של תמונות, ובדיקת ביצועיהם על הבעיה בעבודה זו.
2. שימוש בתמונות וב-Data Augmentation בתור פיתרון לכמות מועטה יחסית של דוגמאות ומניעת Over-fitting.
3. שימוש ב- Jupyter בתור כלי שניתן להציג ולהשמיע בו מידע כגון גרפים וקבצי שמע באופן רציף בין שורות הקוד עצמן.

2. הצגת הבעיה:

השכבה הראשונה של כלל המודלים מקבלת כקלט מטריצות עם מספרים-
בין אם מדובר בתמונה שהפכו אותה למטריצה רב מימדית (מימד אחד עבור כל ערוץ RGBA), ובין אם מדובר בוקטור מתמטי שמייצג מאפיינים שונים (שהכנו כקלט עבור המודל).

עם זאת, כאשר מתמקדים בקבצי קול, שני הייצוגים מספקים למודל מידע **שונה**.
לדוגמא, אראה קובץ שמע בשני הייצוגים השונים:

1. ייצוג של קובץ השמע בתור תמונה (גרף) של תכונה ספציפית (תכונת ה-MFCC):



2. ייצוג של אותו הקובץ בתור ערכי החציונים המתמטיים של וקטור התכונות:

filename	spectral_centroid	zero_crossings	spectral_rolloff	chroma_stft	rms	mel_spec	mfcc_1	mfcc_2	...
vav	2902.358859	0.135362	5913.555908	0.408873	58.325471	65.212778	-116.627914	31.952889	...
man-	4081.163181	0.278320	6988.913653	0.316327	40.030305	10.216539	-111.718104	-38.222843	...

מבחינת הדגשת ההבדלים בין שני הייצוגים:

- מתמטית: כאשר אני מעביר למודל וקטור מאפיינים, אני מספק לו כמות קטנה יותר של מידע ביחס לכמות המידע הניתנת בתמונה (וקטור המאפיינים מכיל 26 תכונות [ערכים] סה"כ, בעוד שהתמונה שתועבר למודל תהיה בגודל של 150*150*3 (150*150 פיקסלים ב-RGB = 67500 ערכים). גודל התמונה נובע מההשערה שתמונות קטנות מדי יקשו על מודלים לחלץ מתוכן דברים חשובים.

- אינטואיטיבית:

  - מזווית המודל שיקבל ערכי חציונים: שוללים ממנו את היכולת להבין כיצד ההקלטה השתנתה לאורך זמן, אך בתמורה הוא מקבל ערכים מדויקים משלל של מאפיינים.
  - מזווית המודל שיקבל תמונות: שוללים ממנו את כלל המאפיינים פרט למאפיין ספציפי, כלומר שוללים ממנו מידע שעלול להיות קריטי, אך בתמורה הוא מקבל את היכולת להבין כיצד מאפיין מסוים השתנה לאורך כל ההקלטה.

לסיכום, קיימים הבדלים בין המידע שמועבר למודל בשני הייצוגים הנ"ל, לכן יהיה מעניין להבין איזה מהייצוגים יוביל לדיוקים גבוהים יותר בקבוצת המבחן.

3. על התהליך שעברתי:

בתחילת העבודה השקעתי זמן ניכר במחקר באינטרנט על האופן שבו ניגשים לבעיות של סיווג קבצי שמע, ומצאתי שכיום בעיה זו נחשבת כבעיה קשה יותר מבעיית סיווג תמונות, כאשר עיקר המאמצים של החברות המסחריות הגדולות מופנים לזיהוי ופירוש מילים, כחלק מאלגוריתמים שמופעלים במוצרים כגון Alexa או Google Home.

כיוון שלא רציתי להפוך "קבצי שמע למילים", אלא פשוט לזהות האם בקובץ שמע מסוים קיימת "צעקה של אדם במצוקה", הנחתי כי הדבר דומה יחסית לבעיה של סיווג סגנונות מוזיקה, או סיווג של צלילי חיות.

בבעיות מסוג זה הגישה העיקרית הינה חילוץ ערכי החציונים של תכונות מוזיקליות (פירטתי על התכונות העיקריות בפרק יצירת וקטור התכונות), ולצורך כך, השתמשתי בספריה Librosa.

במהלך העבודה נתקלתי באתגרים שאפרט עליהם בפרקים הבאים, כאשר עיקר הדילמות היו סביב האופן שבו יש להציג קובץ שמע בתור תמונה, וסביב תוצאות הניסויים, אשר הפתיעו אותי, ודרשו ממני למצוא דרכים "להקשות" על שני המודלים כדי שאוכל להשוות ביניהם.

4. האם הצלחתי?

מטרת העל הייתה לבדק את השפעתם של ייצוגים שונים על תוצאות הניבויים, לכן לא הייתה פה מטרה "שייצוג זה או אחר ינצח".

מבחינתי ההצלחה של עבודה זו טמונה בכך שלמדתי ששני הייצוגים חשובים, וכל אחד מהם תורם בדרכו לניבוי, כאשר ניבוי באמצעות תמונה צריך להיעשות בדרך ספציפית מאד, אחרת הוא מוביל לתוצאות גרועות משמעותית מהניבוי שמתבצע ע"י ערכי חציונים מתמטיים של וקטור התכונות.

בסופו של דבר הצלחתי להגיע למצב שבו המודל שקיבל ייצוג של תמונה בתור קלט הגיע לערכי דיוק גבוהים יותר.

פירוט נוסף על כך קיים בפרק הניסויים ובפרק הסיכום.

תיאור פתרון הבעיה

כיוון שלא הצגתי בעיה, אלא הצגתי "שאלה":

כיצד שינוי אופן ייצוגם של קבצי שמע, משפיע על דיוק הניבויים, כאשר ההשוואה תתבצע בין שני הייצוגים הבאים:

1. ייצוג של קבצי שמע בתור תמונה (שהינה בעצם גרף של תכונה מסוימת או מספר תכונות)
2. ייצוג של קבצי שמע בתור ערכי החציונים המתמטיים של וקטור התכונות.

התשובה בעצם מוצגת בסוף כל ניסוי בפרק הניסויים, והדיון בתוצאות נכתב בפרק הסיכום.

עם זאת אציין כבר כעת כי חד משמעית אופן הייצוג של קבצי השמע משפיע על תוצאות דיוק הניבויים.

תיאור המערכת

נתונים גולמיים על מאגר המידע

מאגר המידע שהתבססתי עליו בפרויקט זה בנוי משני חלקים:

1. מאגר שבניתי לצורך פרויקט זה לטובת קבוצת המבחן:

הדגימות נלקחו מסירטוני YouTube של **תרחישים אמיתיים:**

- שיחת הטלפון למוקד 100 מאירוע חטיפת הנערים (לערך בויקיפדיה לחץ כאן)
- Shocking 911 Calls Detail Home Invasion Attack
- Top 15 Disturbing 911 Calls
- Woman secretly called 911
- Woman whispers for help in a chilling 911 call

ושל **כמה סצנות מסרט** שבו מישהי נחטפת ומתקשרת למשטרה לדווח על כך:

- (The Call, 2013)

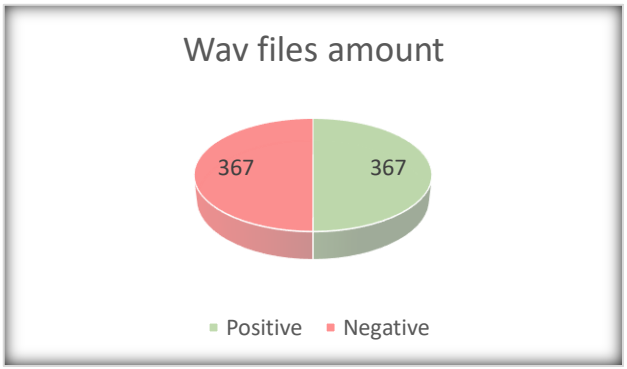
מאגר זה נבנה לאחר קבלת התוצאות של הניסוי הראשון, אשר הראו את הצורך ב"העלאת רמת הקושי" על המסווגים השונים. הוא בנוי מ-50 קבצי שמע סה"כ, ביחס של 1:1 לקבצים שמייצגים "צעקת מצוקה **אמיתית**", וקבצים שמייצגים "שיח טלפוני" שאינו צעקה. כל קובץ באורך של כ-3 שניות.

2. חלק ספציפי ממאגר שבניתי והשתמשתי בו בקורס קודם (קורס "פרויקט בבינה מלאכותית" 236502). מאגר זה מהווה את קבוצת האימון והולידציה, והוא כולל:

367 קבצי קול של "צעקה" שאורכם כ-3 עד 5 שניות, קבצים אלה הינם ה"Positives".

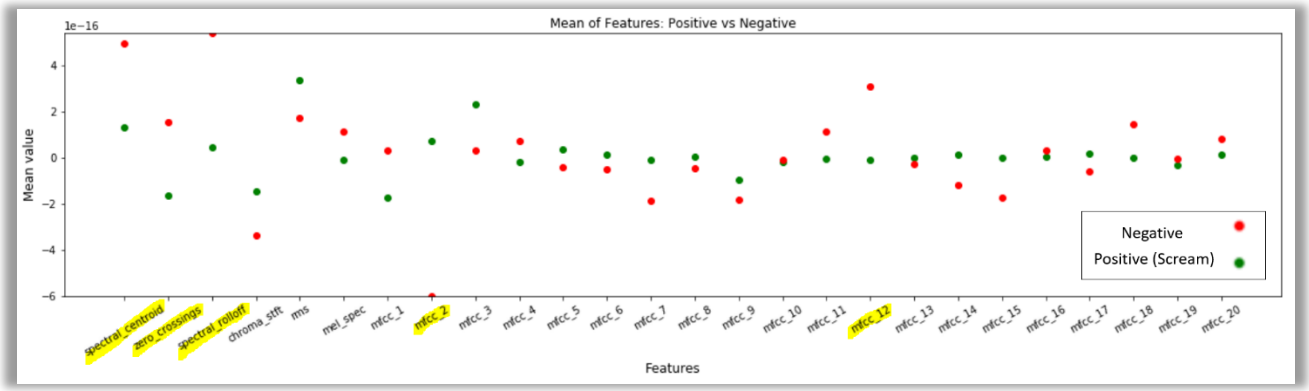
בנוסף המאגר כולל כמות זהה (367) של קבצי קול משיחות טלפון בהן לא היו צעקות, והם מהווים את ה-"Negatives" כאשר כמות הקבצים, ואורכו של כל קטע זהה לנתוני קבצי ה"צעקה". (3-5 שניות).

להלן פילוח הקבצים של קבוצת האימון והולידציה:



היבט ה-Math במידע הגולמי

עבור כל קבוצה מהפילוח הנ"ל (Positives, Negatives), חישבתי את ערך החציון שלה, לכל אחד מה-Features שאשתמש בהם. להלן הגרף שהתקבל לאחר נירמול (למידע נוסף על ההבדל בין הנירמול בפרק זה לנירמול שבוצע בתהליך האימון לחץ פה):



מסקנות:

מסומנים בצהוב חמשת המאפיינים שנתנו את ההבדלים הכי גדולים בין ערכי החציון של שתי הקבוצות (Pos', Neg'). אני משער שאלה המאפיינים שיעזרו למודל המתמטי לנבא טוב יותר לאיזה קבוצה קובץ שמע ישתייך.

מעניין לציין כבר כעת כי המאפיין mel_spec (המאפיין השישי משמאל), אשר אינו אחד מחמשת המאפיינים המובילים כאן, היה המאפיין שכאשר ייצגתי את קבצי השמע בתור תמונה- נתן דיוקים גבוהים יותר מהדיוקים של המודל שהתבסס על כל ערכי החציונים של וקטור התכונות הנ"ל. פירוט נוסף על כך נכתב בפרק הניסויים והסיכום.

היבט ה-Image במידע הגולמי

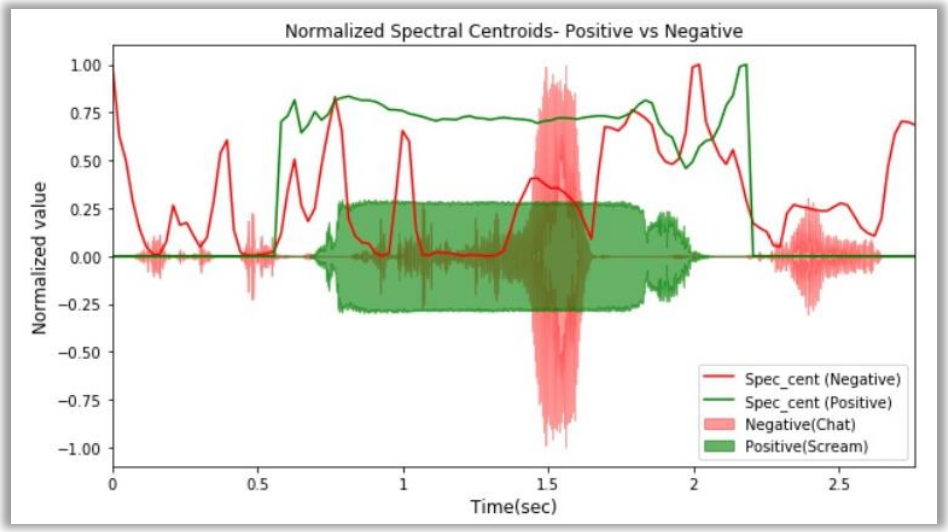
בחלק זה לא ניתן להציג "חציון" של תמונות כפי שמוצג למעלה, לכן אבחר באופן רנדומלי קובץ שמע בודד מכל קבוצה (Positives & Negatives) , ואציגם באופן חזותי עבור כל מאפיין עיקרי בוקטור המאפיינים שלי (להסבר על כל מאפיין יש לקרוא בפרק "יצירת וקטור התכונות).

אציין כי קיים הבדל בין התמונות שאעביר למודל הלמידה לבין התמונות בהמשך פרק זה. ההבדל נובע מכך שבפרק זה חשוב לי לתת לקורא "לראות בתמונה אחת" את קובץ הקול החיובי והשלילי. למודל עצמו אעביר תמונות שמייצגות את הדגימות החיוביות והשליליות בניפרד (לא יתקיים מצב שאעביר לו תמונה בודדת עם דגימה חיובית ושלילית יחדיו שכן זה יסתור את האופן שבו מלמדים מודל). יש לפנות לנספחים לקבלת דוגמא מדויקת באשר לתמונות שיועברו למודל.

הסבר כללי על התמונות:

- בתמונות הבאות כל מקטע צפוף ייצג הקלטה קולית, בעוד שכל קו דק ייצג את המאפיין.
- בזמנים שבהם המיקרופון פעל והקליט, אך לא היה דיבור, נצפה לראות מדי פעם קפיצות גבוהות בערכי המאפיינים כיוון שהשינוי בין שקט מוחלט לבין צליל קצר בא לידי ביטוי בקפיצה בערכים הנמדדים. לכן כדאי להתייחס לערכי הגרף רק בזמנים שבהם היה שיח נראה לעיין.

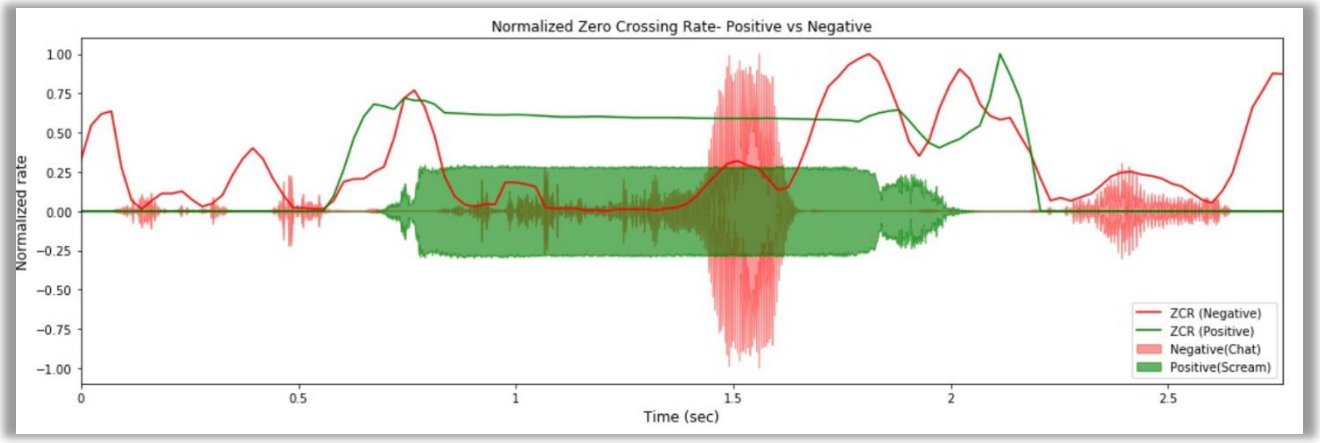
Spectral centroids



מסקנות:

ניתן לראות כי ה"צעקה" בהקלטה זו שומרת על מרכז כובד אחיד של תדר מסוים, בעוד ששיחה רגילה משנה באופן מהיר את מרכז הכובד של התדר. אשער כי אם מודל יוכל לשים לב לכך, זה יעזור לו להבחין בין צעקות מצוקה לבין שיחה רגילה.

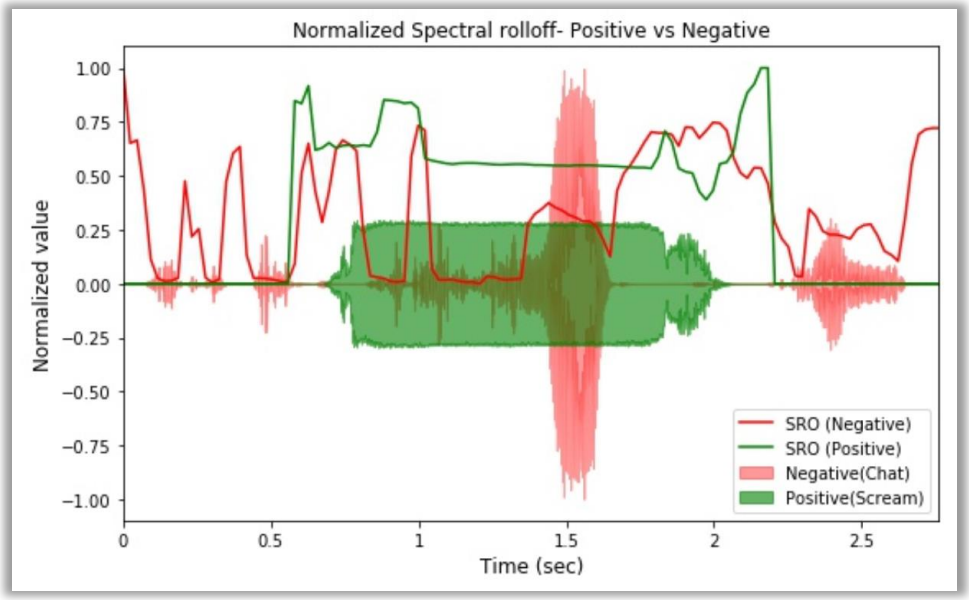
Zero crossings rate



מסקנות:

ניתן לראות כי ערכו של מאפיין זה גובר כאשר אין דיבור והמיקרופון פתוח (הסבר על כך בפרק יצירת וקטור התכונות) בין אם מדובר בהקלטה החיובית או השלילית. עם זאת, ניתן להבחין כי בהקלטה החיובית קצב מעברי האפס נשאר אחיד יחסית בעוד שבעת דיבור קצב מעברי האפס משתנה באופן תכוף יותר. אשער כי גם מאפיין זה יוכל להועיל למודל אשר ידע לזהות כי בצעקות קצב מעברי האפס נשאר אחיד יחסית למשך זמן ארוך ביחס לקבצי הקלטה שאינם צעקות.

Spectral rolloff

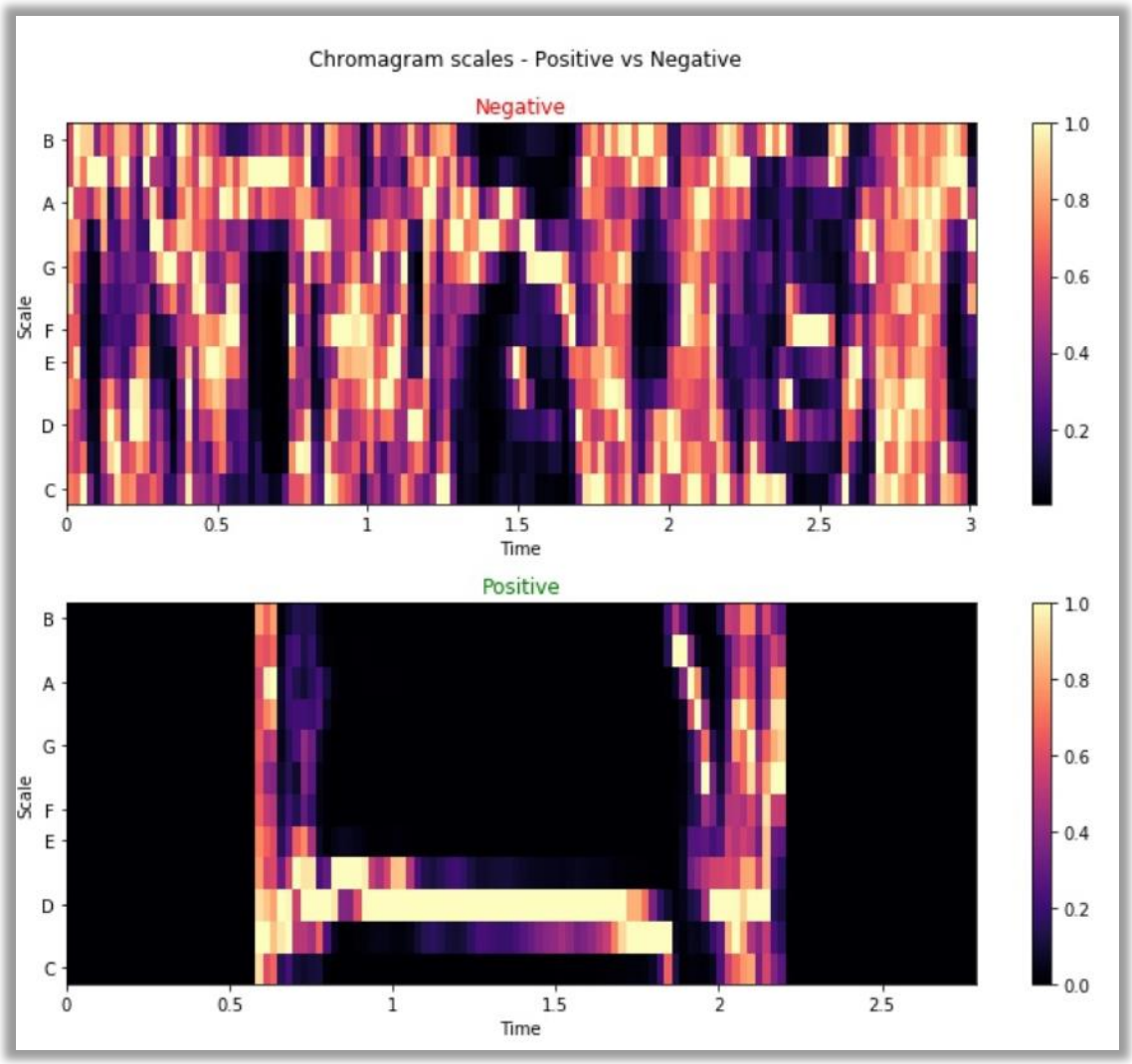


מסקנות:

גם במאפיין זה קיימת אחידות מסוימת בערך המתקבל לאורך הדגימה החיובית. מעבר לכך קשה לי לזהות משהו שמבדיל באופן חד בין החיובי לשלילי, פרט לעובדה שהדגימה החיובית נמצאת בערך מתמטי גבוה יותר מהדגימה השלילית ברוב ההקלטה. איני יודע אם מודל יוכל להבחין בכך שגובה קו מסוים בתמונה מרמז על שיוך הדגימה לקבוצה החיובית. יהיה מעניין לראות זאת בפרק הניסויים.

(בניסויים שביצעתי מאפיין זה היה אחד משלושת המאפיינים היעילים עבור מסווגים שקיבלו כקלט תמונה, אך הדיוק שלו היה נמוך משמעותית מדיוקים שהתקבלו ממסווג שהתבסס על ערכי החציונים של וקטור התכונות).

Chromagram (chroma feature)

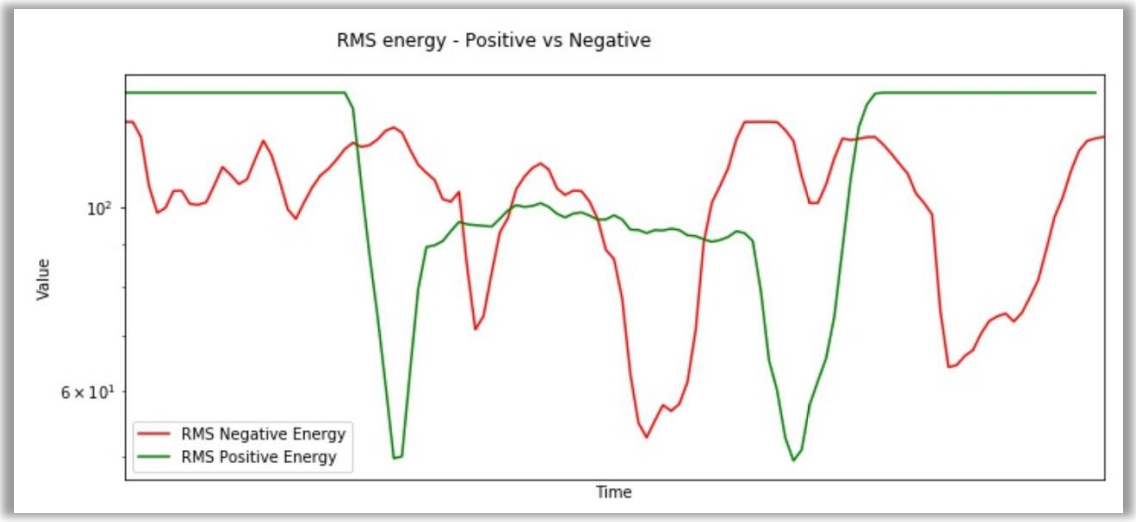


מסקנות :

במאפיין זה ניתן לזהות בקלות כי בצעקה (Positive) קיימים סולמות מעטים, בעוד שבשיח רגיל (Negative) יש שימוש במנעד רחב יותר של סולמות ווקליים. אני משער שסוג זה של תמונות יקל על מודל הלמידה בהנחה וידע לחפש רצף של סולם כלשהו, או דפוסים שבהם "רב השחור על הצבעוני" (שחור מסמל היעדר צליל בסולם מסוים).

(בפרק הניסויים מאפיין זה נתן תוצאות דיוק נמוכות מאד ביחס למאפיינים אחרים שיוצגו בתור תמונה).

RMS (root mean square)

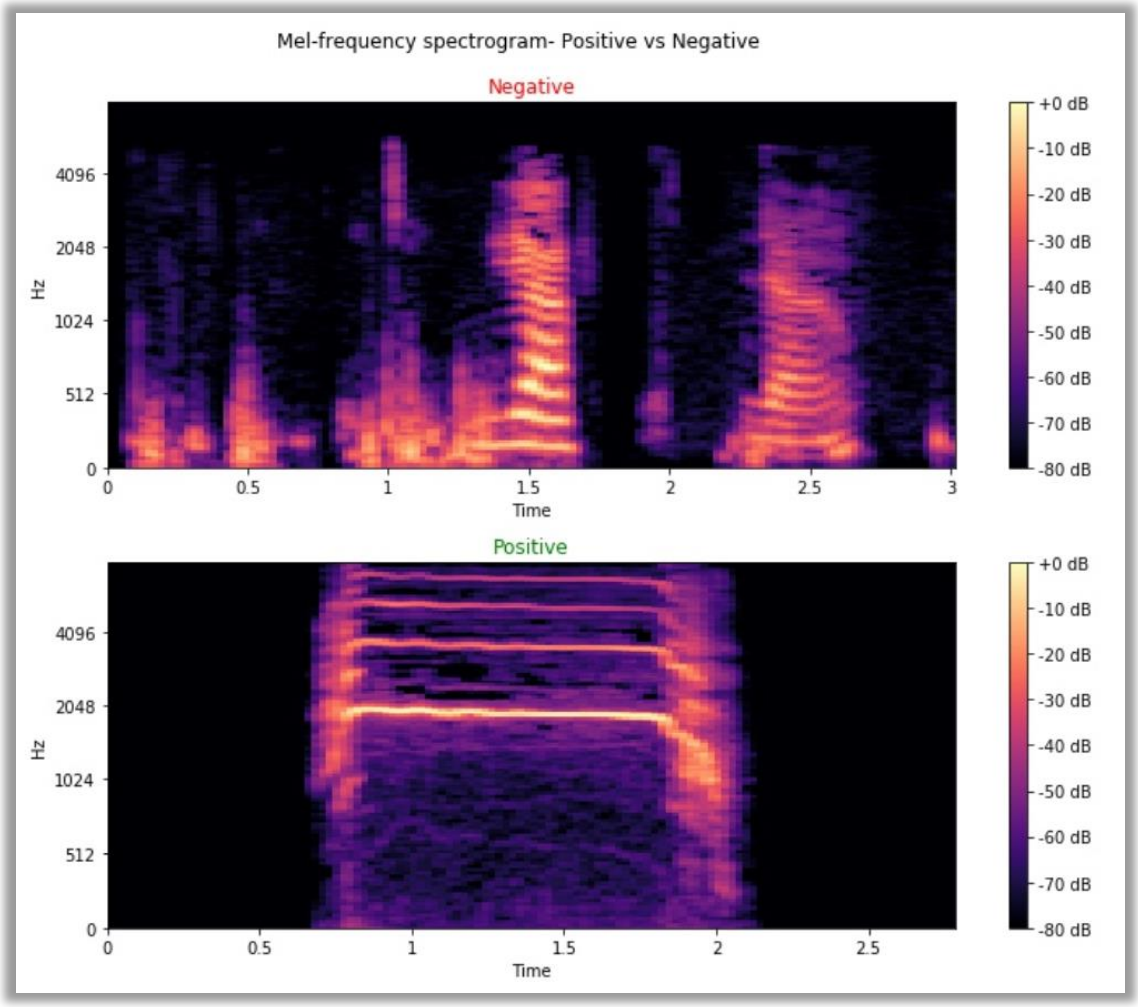


מסקנות :

מאפיין זה היה ניסיוני מבחינתי כיוון שראיתי איזושהי המלצה להשתמש בו במודלי בינה שקשורים לסיווג צלילים. עם זאת מהתמונה איני רואה דפוס מסוים שעוזר לי להבדיל בין הדגימה החיובית לשלילית, לכן לא אתעדף את הניסויים עם מאפיין זה, ואולי לבסוף לא אשתמש בו כלל. מעבר לכך, גם בגישה המתמטית שהוצגה בפרק הקודם הוא לא נראה מבטיח במיוחד. נראה אם יפתיע בתרומתו בפרק הניסויים.

(בניסויים שביצעתי מאפיין זה היה אחד משלושת המאפיינים היעילים עבור מסווגים שקיבלו כקלט תמונה, אך הדיוק שלו היה נמוך משמעותית מדיוקים שהתקבלו ממסווג שהתבסס על ערכי החציונים של וקטור התכונות).

Mel spectogram



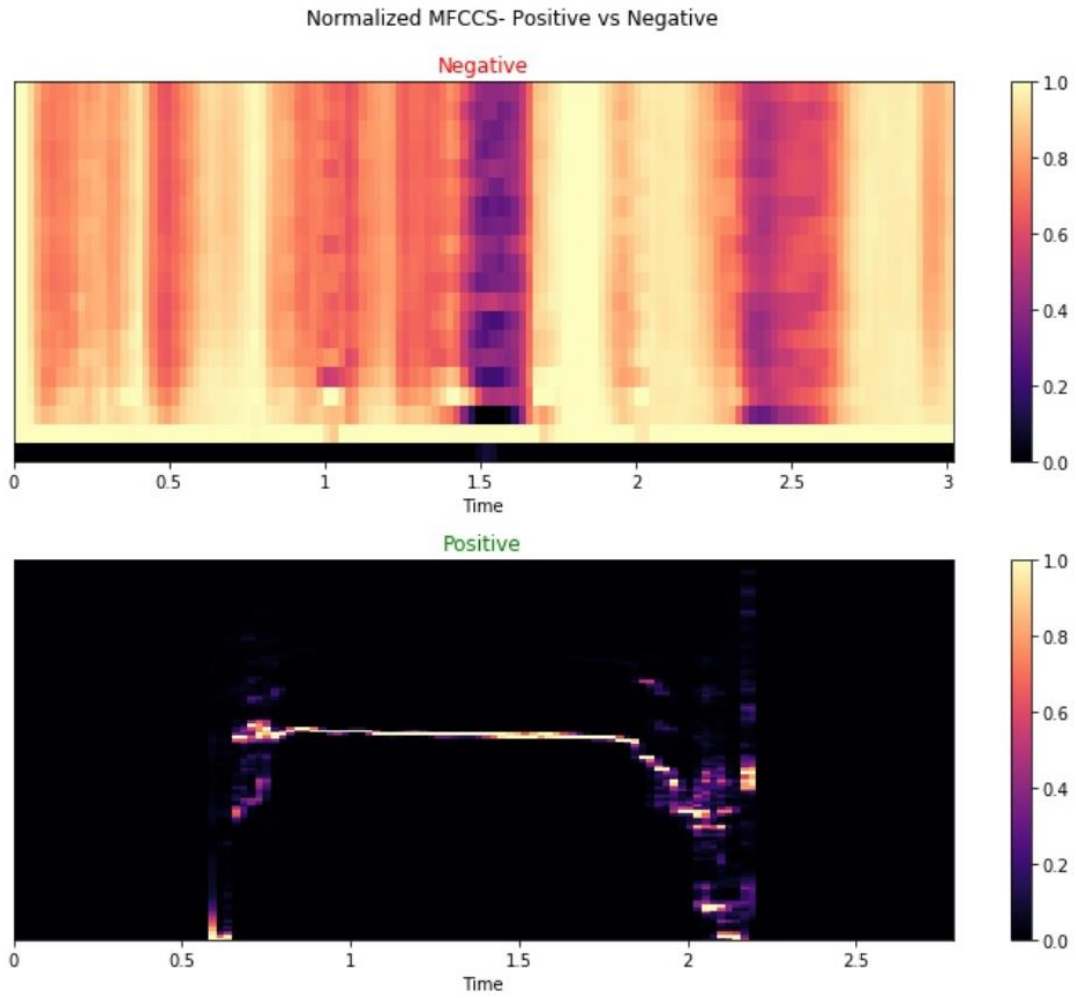
מסקנות :

מאפיין זה יחסית חופף למאפיין הבא (MFCC).

בחרתי להציגו כיוון שאיני רוצה לשלול מראש באיזה מאפיין להשתמש. ניתן לראות מהגרפים כי קיים דפוס בולט שמבדיל בין דגימה חיובית לשלילית. בדגימה החיובית קיים מרחב תדרים באיזור ה-2048 הרץ אשר נמשך כשנייה שלמה. במידה והמודל שלנו ידע לאפיין צבע, אורך וזווית של קו צהוב כמו הקו הבולט בדגימה החיובית אולי זה יעזור לו לנבא טוב יותר לאיזה קבוצה דגימה מסוימת תשתייך.

(בפרק הניסויים, כאשר מאפיין זה הוצג בתור תמונה, הוא הגיע לדיוק הגבוה ביותר מבין כלל המסווגים שאומנו).

MFCC



מסקנות :

מאפיין זה יחסית חופף למאפיין הקודם (Mel spectogram).

מאפיין זה הינו המאפיין הכי מומלץ מהמאפיינים שקראתי עליהם. הוא הופיע בכל אתר ובכל מדריך אשר הסביר על האופן שבו ניתן לרתום בינה מלאכותית למשימות שקשורות במודעות אקוסטית. לכן מאד חשוב לי לבדק את אופן תרומתו לתהליך הניבוי כשהוא בפורמט של תמונה.

ניתן לשים לב כי הדגימה החיובית שונה לחלוטין מהדגימה השלילית. מעניין אם גם בעבודה זו אסיק לאור הניסויים שאבצע כי הוא אחד המאפיינים הכי משמעותיים והכי תורמים לתהליך הניבוי. מהתמונות הנ"ל אני משער שהתשובה לכך תהיה חיובית.

(באופן מפתיע, מאפיין זה השיג דיוקים נמוכים מאד מבין כלל המסווגים שאומנו).

איסוף המידע הגולמי (scrape data)

כשביקשתי לבצע את עבודה זו, חשבתי שמאגרי המידע שכבר יש לי ושכבר אספתי בפרויקט קודם יספיקו, אך כאשר תוצאות הניסוי הראשון על קבוצת הווליידציה הראו כי שני המודלים קיבלו דיוק של 100% עם 50 סבבי אימון בלבד (50 epochs), הגעתי למסקנה שדרוש להשיג מאגר חדש שלא יהיה דומה כלל למאגרי האימון והווליידציה, ולכן התחלתי לחפש ב-YouTube כל מני שיחות אמיתיות למוקדי החירום אשר מכילים צעקות מצוקה.
כפי שפירטתי בפרק הנתונים הגולמיים על מאגר המידע, להלן הקבצים החדשים שהתווספו לפרויקט זה, ונועדו לאתגר את המסווגים שאימנתי:

הדגימות נלקחו מסירטוני YouTube של **תרחישים אמיתיים:**

- שיחת הטלפון למוקד 100 מאירוע חטיפת הנערים (לערך בויקיפדיה לחץ כאן)
- Shocking 911 Calls Detail Home Invasion Attack
- Top 15 Disturbing 911 Calls
- Woman secretly called 911
- Woman whispers for help in a chilling 911 call

ושל **כמה סצנות מסרט** שבו מישהי נחטפת ומתקשרת למשטרה לדווח על כך:

- (The Call, 2013)

מאגר זה בנוי מ-50 קבצי שמע סה"כ, ביחס של 1:1 לקבצים שמייצגים "צעקת מצוקה **אמיתית**", וקבצים שמייצגים "שיח טלפוני" שאינו צעקה. כל קובץ באורך של כ-3 שניות.

כיוון שעל המאגר הקודם שבניתי כבר פירטתי בעבודה הקודמת, אני שם את ההסבר עליו בפרק הנספחים.

יצירת מבנה הנתונים (generate data)

1. אלגוריתם :

יצירת מבנה הנתונים.

ייצור מבנה נתונים שכולל טבלה עם נתונים מספריים למען היבט ה-Math. בנוסף ייצור תמונת מאפיינים מכל קובץ שמע למען היבט ה-Image.

פרטים טכניים יבשים :

קובץ	הצהרה	קלט	פלט
Image_vs_Math.ipynb	def create_entire_dataset():	אין קלט	קובץ CSV עם מבנה הנתונים הרצוי ייפלט לנתיב היחסי (לדוגמא) csv\scream\train_test_data.csv בנוסף יהפוך כל קובץ שמע לתמונה עם הפיצ׳ר הרלוונטי

עיבוד מידע מקדים (pre-process)

אלגוריתם : (מתבצע ע״י פונקציות פנימיות בתוך הפונקציה שפורטה בסעיף הקודם).

<u>היבט ה-Math :</u>

ביצוע עיבוד מקדים לקבצי השמע טרם דליית הפיצ׳רים ע״י כך שהיא מקצה זמן מקסימלי לכל קובץ שמע (5 שניות כהגדרת ברירת מחדל), בודקת שהוא תקין, וקובעת קצב שבו יידגמו קבצי השמע (22050 הרץ).

<u>היבט ה-Image :</u>

הופכת כל קובץ שמע לקובץ תמונה עם המאפיינים הרלוונטיים ומנרמלת את הגרפים כך שעבור כולם טווח הערכים בציר ה-Y יהיה בין אפס לאחד.

יצירת וקטור התכונות (extract features)

למען ההגינות אציין כי פרק ספציפי זה נכתב כבר בעבודה קודמת שהגשתי, וכי בעבודה זו רק ערכתי את הכתוב בו. איני מעביר פרק זה לנספחים כיוון שהוא חשוב למי שלא קרא את העבודה הקודמת ומעוניין להבין על המאפיינים שהשתמשתי בהם.

להלן רשימה של הפיצ'רים שחברתי להשתמש בהם, כאשר מתחת לכל פיצ'ר מתומצת האופן שבו הוא עשוי לתרום לניבוי.

spectral centroid

תורם להבדלה בין בני אנוש לחיות, ולהבדלה בין בני אדם שונים. מראה את הטווח שבו נמצא מרכז המסה של הצליל.

zero-crossing rate

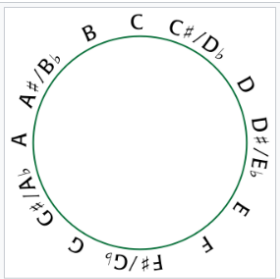
תורם בזיהוי "דיבור" של בני אנוש והבדלתו ממוזיקה\רעש רקע. מראה כמה פעמים עצמת הקול חוצה את קו האפס (קו האפס מייצג כי ה"חלקיקים" באוויר חזרו לנקודת המקור שלהם. כשאין דיבור כלל החלקיקים אינם זזים. כשההקלטה פעילה אך יש רק רעשי רקע אז קיימים מעברי אפס רבים שיש להתעלם מהם. (שנובעים מקפיצה משקט מוחלט לרעש הרקע).

roll-off frequency

סוג של "חתימה קולית". מייצג את התדר ש-85% מכלל התדרים באותו הרגע נמצאים בו או מתחתיו.

Chromagram

מאפיין זה מציג את העצמה הקיימת בכל סולם (במוזיקה קיימים סולמות, וניתן לחלקם ל-12 לפי סולם כרומטי. לחץ פה למידע נוסף).



Chromatic scale drawn as a circle

Mel-frequency cepstral coefficients (MFCCs)

תורם באבחנה בין צלילים של בני אנוש לצלילי רקע, ותורם באבחנה בין אותיות שונות שבני אדם מבטאים. מציג את המנעד הקולי. זהו בעצם ייצוג קצר-טווח של מנעד הצליל (ריכוז האנרגיה של הצליל).

root-mean-square (RMS)

תורם באבחנה בין קטעי שמע שונים. ערך שמייצג את חציון "שרש עצמת האנרגיה" של קטע השמע. מאפיין זה מסקרן ואיני יודע מראש אם הוא יועיל או לא. אראה זאת במהלך הניסויים.

mel-scaled spectrogram

תורם בדומה לנקודה 5, אך בייצוג מתמטי שונה. כלי נוסף שיעזור לנו להבדיל בין קטעי שמע שונים.

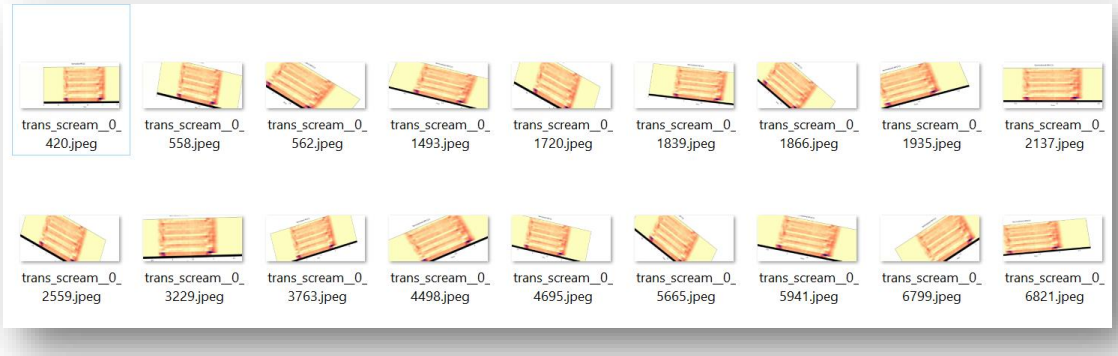
הרחבת מאגר הנתונים החזותי (Data-augmentation)

בהמשך ל-Future-work מהעבודה הקודמת שהגשתי, בה היה כתוב שכיוון מעניין יהיה להתעסק עם Data-augmentation, בעבודה זו החלטתי להשתמש בטכניקה זו כיוון שבאמת השתמשתי רק בכ-367 תמונות לכל סיווג.

השימוש בכלי זה בא לידי ביטוי דרך אובייקט מהספריה Keras, שנקרא ImageDataGenerator והוא נותן מענה לשתי בעיות עיקריות:

1. ניהול תהליך טעינת התמונות לזיכרון המחשב כך שבכל סבב אימון\וולידציה\מבחן תיטען כמות תמונות שהמחשב מסוגל להתמודד איתה
2. מניעת התאמת-יתר לקבוצת האימון ע"י שינויים עדינים שמתבצעים על התמונות, כך שלא קורה מצב שאותה תמונה מועברת פעמיים למודל.

התמונה הבאה באה להמחיש את האופן שבו האובייקט הנ"ל משנה תמונות ע"י סיבובן, הגדלתן, וכו'. בפועל השתמשתי רק באלמנט "ההגדלה" ולא באלמנטים האחרים שכן בניגוד לתמונות של חתולים למשל, בגרפים של קבצי שמע אין משמעות ל"הטיה הצידה" של הגרף וכו'.



אימון (train)

כיוון שעבודה זו הינה בעלת אוריינטציה מחקרית יותר, פירטתי על אופן האימון של המודלים השונים בפרק הניסויים, תחת "מתודולוגיית הניסויים".

בחירת פיצ'רים (feature selection)

בפרק הניסויים אציג ניסוי אשר משווה בין הדיוקים המתקבלים על קבוצת המבחן כאשר מייצגים את קבצי השמע בתור תמונה שמהווה גרף של תכונה מסוימת. היה מאפיין בודד שייצוגו בתור תמונה נתן את הדיוק הגבוה ביותר.

ניסויים

מתודולוגיית הניסויים:

כיוון שהמטרה שלי הייתה לבדק כיצד שינוי אופן ייצוג קבצי השמע משפיע על תוצאות דיוק הניבויים, הייתי צריך לקבע ככל הניתן (דיון על מתודולוגיה זו לאור התוצאות יתקיים בפרק הסיכום) את שאר "המשתנים" שקשורים למהלך הניסויים, לכן:

1. בכלל הניסויים, כל הדוגמאות של קבוצת המבחן, הוולידציה, והמבחן זהות. (רק אופן הייצוג שלהן משתנה- תמונה או ערכי חציונים מתמטיים).
2. כמות סבבי האימון מקובעת ושווה ל-50 עבור כלל המודלים שנבדקו.
3. כל הגרפים שחולצו מקבצי השמע מקובעים לאותו הגודל, ומנורמלים בציר ה-Y לטווח [0-1]

שחזור התוצאות:

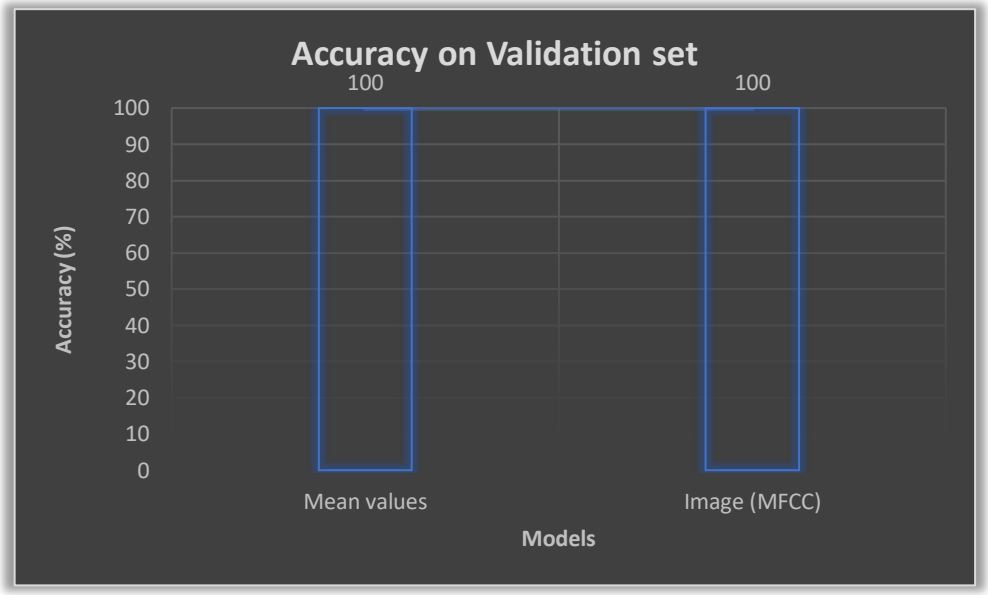
a. אנא עקוב ראשית אחרי ההוראות בפרק הנספחים תחת הכותרת "הקמת סביבת עבודה לפרויקט".
b. כל ניסוי אפשר למצוא בקובץ Image_vs_Math.ipynb, ע"י החיפוש exp_<number of experiment>. לדוגמא: כדי למצוא את הקוד של הניסוי הראשון, יש לחפש exp_1.
c. ניתן גם "לראות" את מהלך הניסויים ע"י חיפוש בקובץ הנ"ל, באמצעות: Experiment <number of experiment>. לדוגמא: Experiment 1. היתרון בחיפוש מסוג זה, הינו שכאשר ביצעתי את הניסויים, נתתי ל-Jupyter להריץ חלקים שונים מהניסוי בתאים שונים, מה שנתן לי לראות תוצאות מוחשיות של הגרפים והמידע בין שורות הקוד השונות. אני ממליץ על שיטה זו בבוא הקורא להבין לעומק את מהלך הניסויים.

1. ניסוי ראשון

מטרה: לבחון כיצד ייצוג שונה של קבצי השמע ישפיע על תוצאות הדיוק של הניבויים.
המודל הראשון יהיה מודל שסופקו לו כלל החציונים המתמטיים של וקטור התכונות.
המודל השני יהיה מודל שסופק לו המאפיין MFCC בייצוג של גרף.

הצדקה: ההיפותזה שלי הייתה שתמונה תיתן דיוק גבוה יותר כי היא מראה מידע רב יותר. עם זאת, לא הייתי בטוח בכך, שכן המידע רב יותר רק במאפיין ספציפי, בעוד שאין כלל מידע מהמאפיינים האחרים, שהמודל שנשען על חציונים כן חשוף אליהם.
בחרתי במאפיין MFCC לניסוי הראשון כיוון שכאשר הסתכלתי על כלל הגרפים שבאפשרותי לחלץ, מאפיין זה נראה לי הכי מבטיח.

תיאור התוצאות:



ניתן לראות כי שני המודלים הצליחו להשיג רמת דיוק של 100% על קבצים שלא נחשפו למודל מעולם.

מסקנות מהניסוי:

1. התוצאות הינן זהות ומדויקות ללא טעות בודדת. דבר זה מאד מפתיע, שכן ציפיתי שלכל הפחות, אחד מהמודלים יטעה בסיווג כלשהו.
2. ככל הנראה, כיוון שקבצי השמע של קבוצת האימון והולידציה נלקחו מאותם האתרים, קיים ביניהם דמיון רב, ולכן המודלים מצליחים לנבא בצורה טובה גם על קבצים שלא ראו מעולם.
3. התוצאה הנ"ל אינה עוזרת להכריע האם קיים שוני בין המודלים, לכן בניסוי הבא אצטרך להעלות את רמת הקושי של דוגמאות קבוצת המבחן (ה- Test set). כלל הדוגמאות שהמודלים למעלה למדו ונבחנו עליהם נלקחו מהקלטות מהטלפון האישי שלי, ומהקלטות איכותיות ממגוון אתרים.
כדי להעלות את רמת הקושי אקח בניסוי הבא הקלטות מסרטים שונים, ובשפות שונות (עברית ואנגלית). פרטים נוספים בניסוי הבא.

2. ניסוי שני

מטרה: בעקבות הניסוי הקודם, בו שני המודלים הציגו רמת דיוק של 100%, הפעם ארצה להעלות את רמת הקושי ע"י ניבוי שיתבצע על קבוצת מבחן אשר שונה מהותית מקבוצת הלמידה והולידציה.

הצדקה: מטרת עבודה זו הינה להבין כיצד ייצוגים שונים של קבצי שמע משפיעים על דיוקי הניבוי.

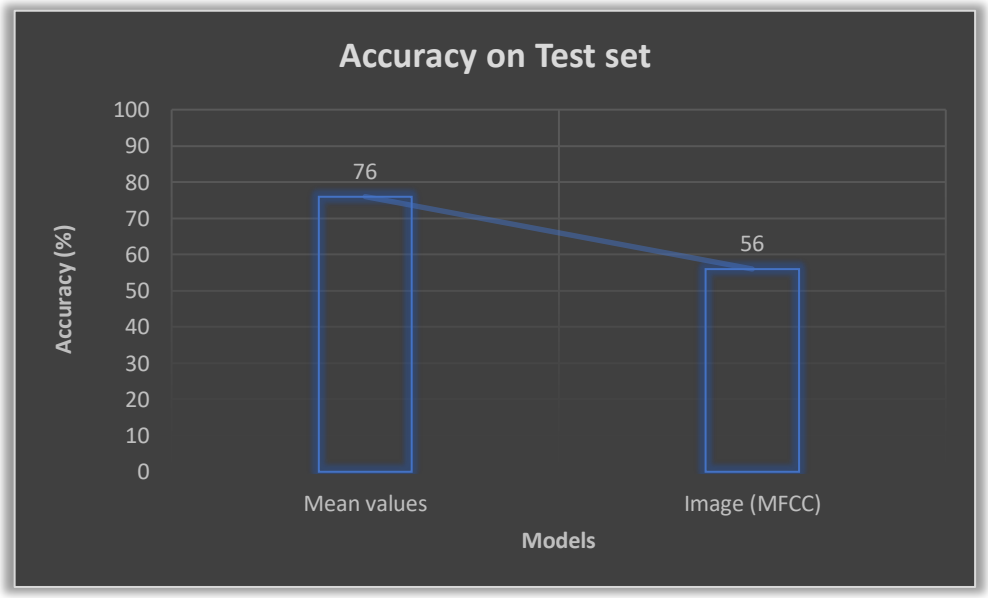
בניגוד לניסוי הקודם, בו הושג דיוק של 100% על קבוצת הוולידציה, הפעם אבחן על קבוצת דוגמאות שונה לחלוטין. קבוצת המבחן בניסוי זה תהיה קבוצה של 50 קבצי שמע באורך 3 שניות, ביחס של 1:1 לדגימות חיוביות ושליליות, כאשר הדגימות נלקחו מסירטוני Youtube של **תרחישים אמיתיים:**

a. שיחת הטלפון למוקד 100 מאירוע חטיפת הנערים (לערך בויקיפדיה לחץ כאן)
b. Shocking 911 Calls Detail Home Invasion Attack
c. Top 15 Disturbing 911 Calls
d. Woman secretly called 911
e. Woman whispers for help in a chilling 911 call

ושל **כמה סצנות מסרט** שבו מישהי נחטפת ומתקשרת למשטרה לדווח על כך:

f. (The Call, 2013)

תיאור התוצאות:



מסקנות מהניסוי:

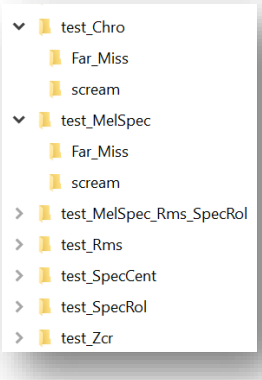
1. קבוצת המבחן שבחרתי לניסוי זה הוכיחה את עצמה- הפעם אף אחד מהמסווגים לא הצליח להגיע לרמת דיוק של 100%, לכן הפעם ניתן להסיק מסקנות לגבי השוני שבין איכות הניבויים של המסווגים.
2. הפעם ניתן לראות כי קיים הבדל ניכר בין המסווג שהתבסס על תמונה שניתן לה מאפיין בודד מסוג MFCC (רמת דיוק של 56%), לבין המסווג שהיה חשוף לערכי החציונים של כלל המאפיינים (76%). כלומר מסווג ה-Image תפקד בצורה פחות טובה.
3. גם הפעם, התוצאה מפתיעה שכן המאפיין MFCC נראה לי כמאפיין שהכי הבדיל בין הדוגמאות החיוביות לשליליות בפרק הנתונים הגולמיים על מאגר המידע.
4. המשפט "דברים אינם כפי שהם נראים" מתאים כאן, שכן המאפיין שנראה לי הכי מבטיח השיג דיוקים שקרובים ל-50%. ולכן בניסוי הבא אבדוק את שאר הגרפים שניתן לחלץ מקבצי השמע, ואראה כיצד הם ישפיעו על הדיוק שישיג מסווג ה-"Image".

3. ניסוי שלישי

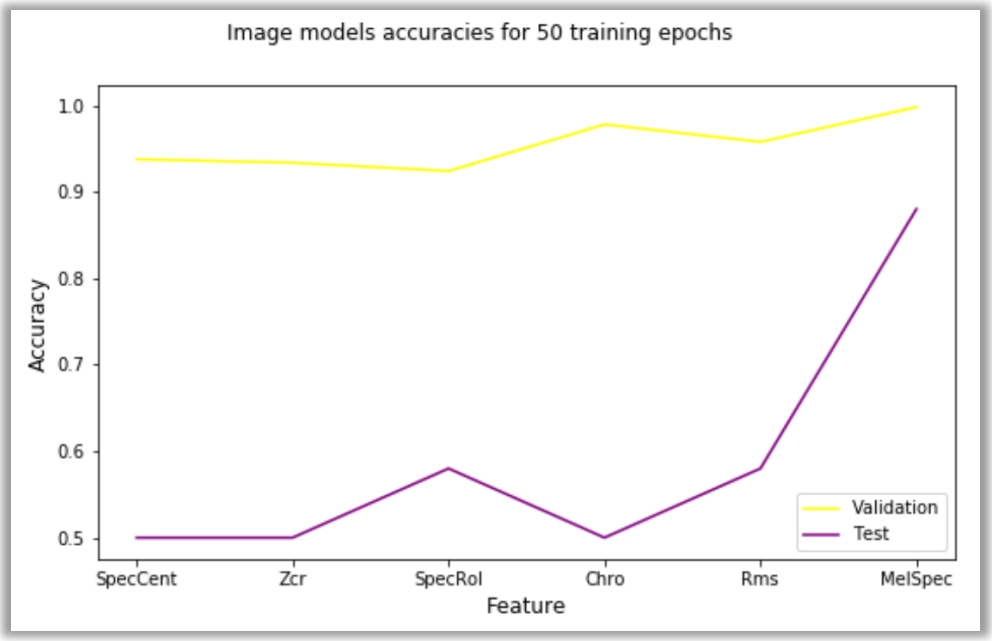
מטרה: לבחון כיצד ייצוג של קבצי השמע בתור גרפים (של מאפיינים שלא נבדקו עדיין), ישפיע על תוצאות הניבויים.

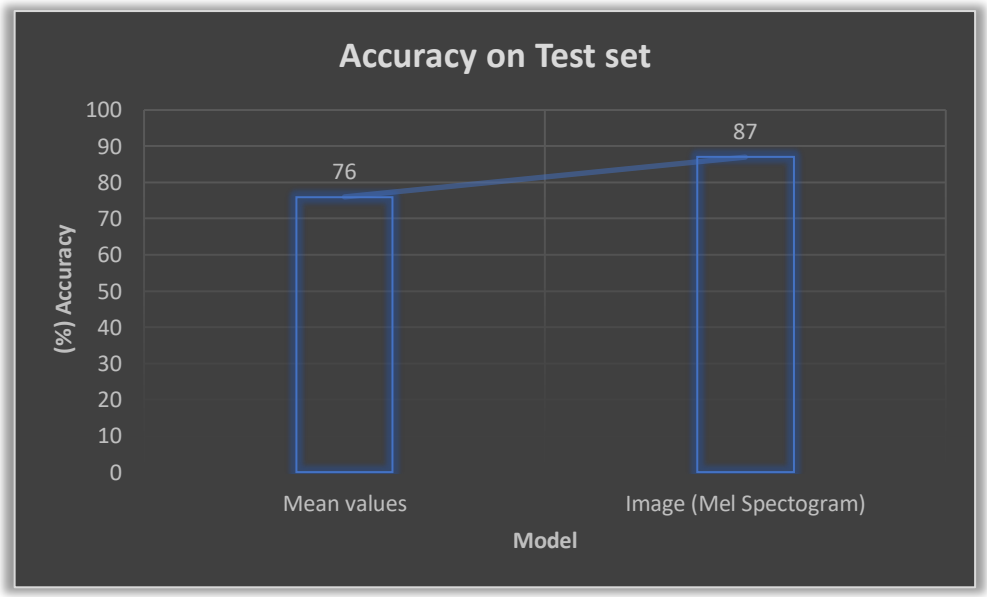
הצדקה: בניסוי הקודם המסווג "Mean values" קיבל דיוק גבוה יותר על קבוצת המבחן מאשר המסווג שלמד על תמונות."Image (MFCC)". עם זאת, אולי הדיוק הנמוך של המסווג שלמד על תמונות נבע מכך שבחרתי מאפיין שאינו מועיל למודל, ואם אבחר מאפיין אחר שאותו אציג בתור תמונה- אז אקבל דיוקים גבוהים יותר. לכן בניסוי הנוכחי אחלץ מתוך כל קובץ שמע מספר תמונות, כאשר כל תמונה הינה גרף של מאפיין ספציפי והאופן שבו הוא השתנה לכל משך ההקלטה.

כלומר ייבנו סה"כ שישה מסווגים שונים, אחד לכל סוג של מאפיין (גרף). להלן המחשה של עץ התיקיות שייבנה עבור כל קבוצת מבחן. test_<feature name>



תיאור התוצאות:





מסקנות מהניסוי:

1. המסווג שנבנה על תמונות מסוג MelSpec הביא לרמת הדיוק הכי גבוהה (87%). אדגיש כי כלל המסווגים בניסוי הנוכחי ובניסוי הקודם קיבלו 50 אפוקים (epochs) בזמן האימון. במקום השני נמצא כרגע המסווג המתמטי שנבנה על ערכי החציונים של כלל התכונות הנ"ל. עם רמת דיוק של (76%).

2. התוצאה מאד מעניינת שכן בפרק שבו הצגתי את ההבדלים בין החציונים המתמטיים של הדוגמאות החיוביות והשליליות, המאפיין MelSpec לא הראה מרחק מתמטי גדול בין הדגימות החיוביות לשליליות, ודווקא כאשר הוא מיוצג בתור תמונה, הוא נתן את הדיוק הטוב ביותר על קבוצת המבחן (וגם על קבוצת הולידציה).
3. מסקנה מעניינת נוספת, היא שבאמת רק מסווגי ה- Image(MFCC), Image (MelSpec) הם היחידים שהשיגו רמת דיוק של 100% על קבוצת הולידציה. לכן מעניין מדוע שניהם תפקדו בצורה כל כך שונה על קבוצת המבחן. (בניסוי הקודם המסווג MFCC השיג רמת דיוק של 56% בעוד שהמסווג MelSpec ניסוי זה השיג דיוק של 87%). יהיה מעניין לראות כיצד שילובים של תכונות שונות באותה תמונה ישפיע על תוצאות הניבויים.

4. <u>ניסוי רביעי</u>

<u>מטרה:</u> לבדק כיצד שילוב של מספר גרפים לתמונה אחת ישפיע על תוצאות הניבויים.

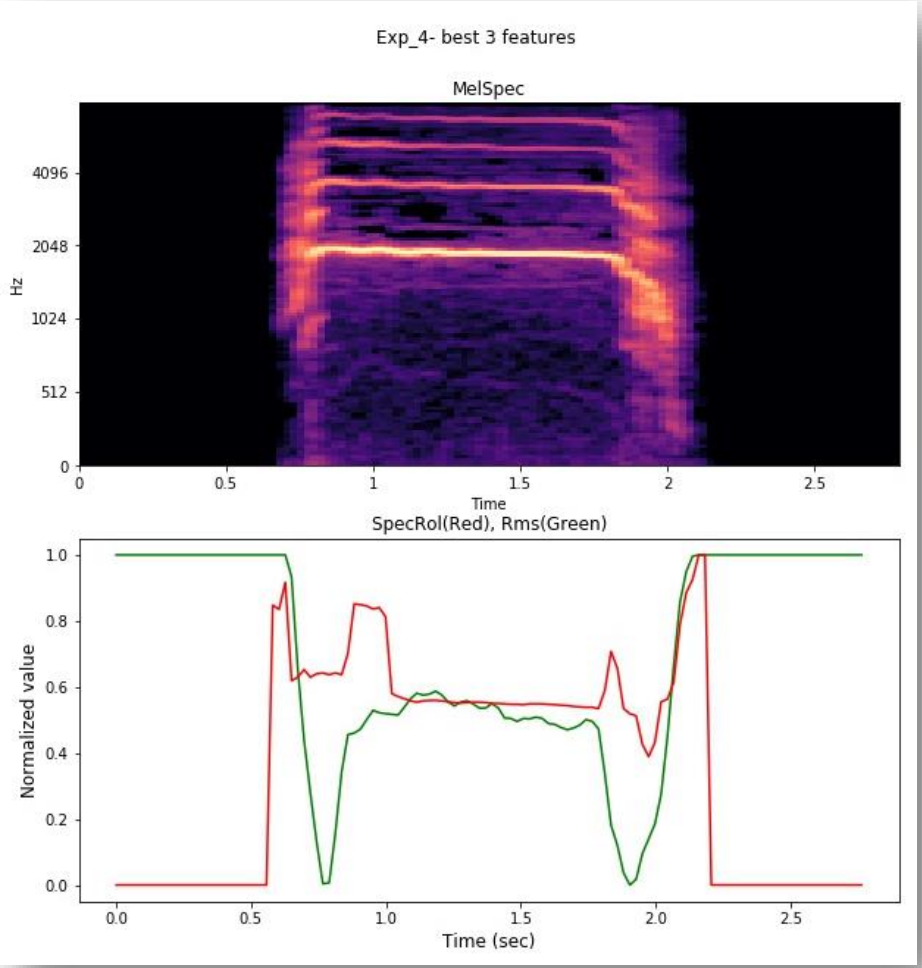
<u>הצדקה:</u> בניסוי הקודם, בפעם הראשונה הצלחתי לקבל דיוק גבוה יותר במודל שלמד על תמונות של תכונה בודדת (87%) מאשר במודל שלמד על ערכי החציונים של כלל התכונות (76%).

אף על פי כן, **רק מודל ספציפי** נתן את הדיוק הנ"ל (המודל שלמד על התכונה MelSpec), בעוד שכל שאר המודלים שנבנו על סמך תמונות השיגו דיוק נמוך משמעותית מהמודל "Mean values" (כולם השיגו באיזור ה-50%-60% דיוק על קבוצת המבחן).

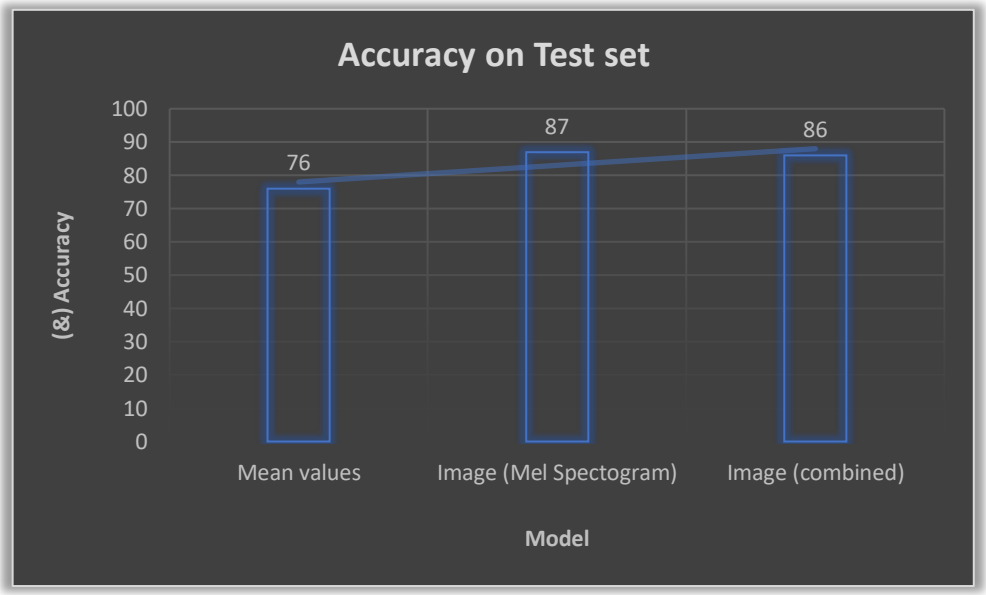
לכן בניסוי זה אבדוק האם שילוב של מספר גרפים בתמונה אחת, יעלה את רמת הדיוק של המסווג שנבנה על תמונות. לטובת הניסוי אקח את שלושת המאפיינים שנתנו את הדיוקים הגבוהים ביותר על קבוצת המבחן-

1. MelSpec – 87%
2. SpecRoll - 58%
3. Rms - 58%

ואבנה לכל קובץ שמע, קובץ תמונה שיכיל את שלושת המאפיינים הנ"ל. להלן דוגמא של קובץ שמע בייצוג זה (דוגמא עם סיווג Negative):



תיאור התוצאות:



מסקנות מהניסוי:

1. הוספת "מידע נוסף" לתמונות לא שיפרה את הדיוק מהניסוי הקודם.
   - התקבל דיוק של 86% על קבוצת המבחן, כלומר הפעם התוצאה ירדה ב-1% מהתוצאה הכי טובה שקיבלנו בניסוי הקודם. עם זאת, תוצאה זו גבוהה ב-10% מהתוצאה של המודל Mean values.
2. לאור התוצאות אני זונח את רעיון ה"שילוב מספר גרפים בתמונה אחת".
3. בעקבות זאת, בשלב זה קיימים שני כיוונים נוספים אשר נראים לי מעניינים לבדיקה:
   א. הראשון הינו התייחסות לכמות האפוקים (epochs) בתור פרמטר, ובעקבות זאת ניתן להשוות את דיוקי כל המודלים שבדקתי עד כה- עבור כמות רבה יותר של סבבי אימון (epochs).
   ב. השני, שימוש במודל מאומן של Keras שעליו אוסיף עוד שכבות אחרונות של משקולות שיעזרו למודל להחזיר ניבויים ספציפיים לבעיה שלי. כיוון זה נקרא Transfer-Learning, והוא מניח כי תכונות חשובות שונות בתמונות חוזרות על עצמן בבעיות שונות, ולכן קיים סיכוי טוב שמודל זה ישפר את הדיוק על קבוצת המבחן שלי.

אבחר לבדק את הכיוון השני- Transfer learning.

נימוק: את הכיוון הראשון כבר בדקתי בעבודות קודמות, ואני מנחש שהתוצאה תהיה שככל שמעלים את כמות סבבי הניסוי הדיוק ישתפר. בנוסף, במתודולוגית הניסויים שלי ציינתי כי אני רוצה לקבע כמה שיותר פרמטרים. אחד מהפרמטרים האלה היה "כמות סבבי האימון" אשר קיבעתי ל-50. לכן הליכה בכיוון הראשון תשנה את מתודולוגיית הניסויים שלי שכן לכל מודל אצטרך לאפשר לבצע כמות סבבי אימון עד להתכנסות מגמת השיפור שלו, ואין בכוונתי לשנות את המתודולוגיה שהצגתי בהתחלה. (על אף שזה כיוון נכון כאשר המטרה הינה למקסם את הדיוק של מודל מסוים, בעבודה זו הדגש ששמתי הינו על אופן ייצוג המידע והשפעתו על מודלים שונים).

לכן אני בוחר ללכת בכיוון שאיני מכיר, ושסיקרן אותי זה מכבר, ולנסות לראות מה יקרה כאשר משתמשים במודל שאומן בעבר על בעיה שונה, וכעת רק מוסיפים לו שיכבת משקולות אחרונה כדי למקדו על הבעיה שמוצגת בעבודה זו.

5. <u>ניסוי חמישי</u>

<u>מטרה</u>: לראות כיצד ישפיע שימוש במודל מאומן שאוסיף לו שכבת משקולות אחרונה, על דיוקי הניבוי של קבוצת המבחן שלי.

<u>הצדקה</u>: הרקע שלי בבניית מודלים מועט, ועל אף שמבחינה טכנית אני יכול לבנות מודלים, איני מבין (נכון לעת כתיבת שורות אלה) כיצד יש לבנות מודל כך שייתן את הדיוק הגבוה ביותר עבור זיהוי "צעקות מצוקה". בנוסף, כמות המידע שסיפקתי למודלים שלי אינה רבה (כמה מאות של תמונות), לכן שימוש במודל פופולרי מהספרייה Keras אשר אומן על כמות עצומה של תמונות וסבבי אימון נשמע כמו כיוון מבטיח.

המודל שאשתמש בו נקרא VGG16. זהו מודל שפותח באוקספורד במעבדה לגיאומטריה (Visual Geometry Group) והוא התפרסם בין היתר בגלל שהם שיתפו את המשקולות של המודל.

הסבר על האופן שבו אבצע את הניסוי:

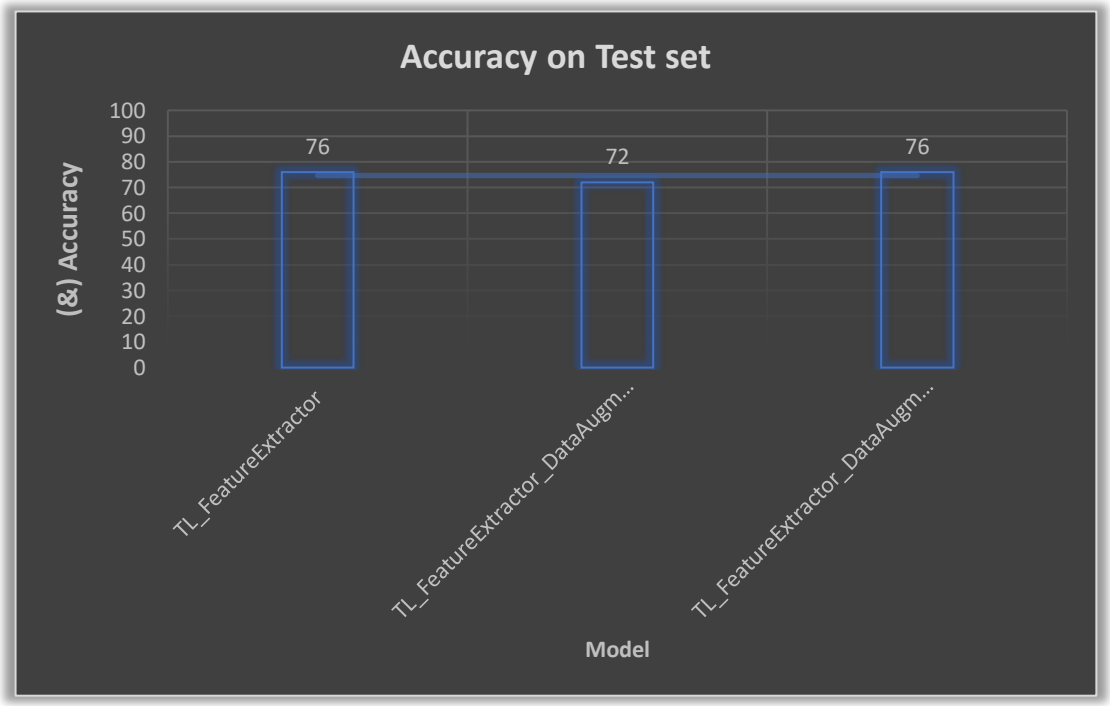
ראשית כל, אייצג בתור תמונה את המאפיין שנתן עד כה את הדיוק הכי גבוה בניסויים: MelSpec.

שנית, אבדוק שלושה תתי כיוונים אשר כולם נמצאים תחת מטריית ה-Transfer learning (לינק לכתבה מעולה על הנושא):

1. TL_as_FeatureExtractor- בשיטה בסיסית זו, באימון ובניבוי משתמשים במודל VGG16 בתור כלי לחילוץ תכונות בלבד. כלומר אריץ רק פעם אחת את סט האימון והולידציה שלי במודל זה, כך שאקבל וקטור של תכונות (ולא ניבוי של סיווג), ואת וקטור התכונות אתן כקלט למודל פשוט שתפקידו יהיה לנבא את הסיווג. כלומר רק המשקולות של המודל הפשוט ישתנו במהלך האימון.

2. TL_as_FeatureExtractor_DataAugmentation- כנ"ל כמו בשיטה הבסיסית, כלומר גם הפעם רק המשקולות של המודל הפשוט ישתנו במהלך האימון, רק שכעת אוסיף את אלמנט ה"Data-augmentation", אשר קודם לא יכולתי להוסיף. (הפעם VGG16 ממש בנוי כחלק מהמודל הפשוט, רק ששכבותיו קפואות ואינן מושפעות במהלך האימון), המשמעות הינה הגדלת זמן האימון באופן משמעותי (50 סבבי אימון לקחו לי כ-6 שעות, כשבכל סבב היו כ-602 דוגמאות).
3. TL_as_FeatureExtractor_DataAugmentation_FineTuning- זוהי השיטה הכי מתקדמת, והיא כוללת "הפשרה" של שכבות מהמודל VGG16 כך שגם המשקולות שלהן נכנסות לתהליך האימון. המשמעות הינה הארכת זמני האימון, אך בתמורה הצפי הוא שנקבל דיוק גבוה יותר. (הלכתי לישון כשהאימון רץ, וכשקמתי האימון כבר הסתיים).

<u>תיאור התוצאות:</u>



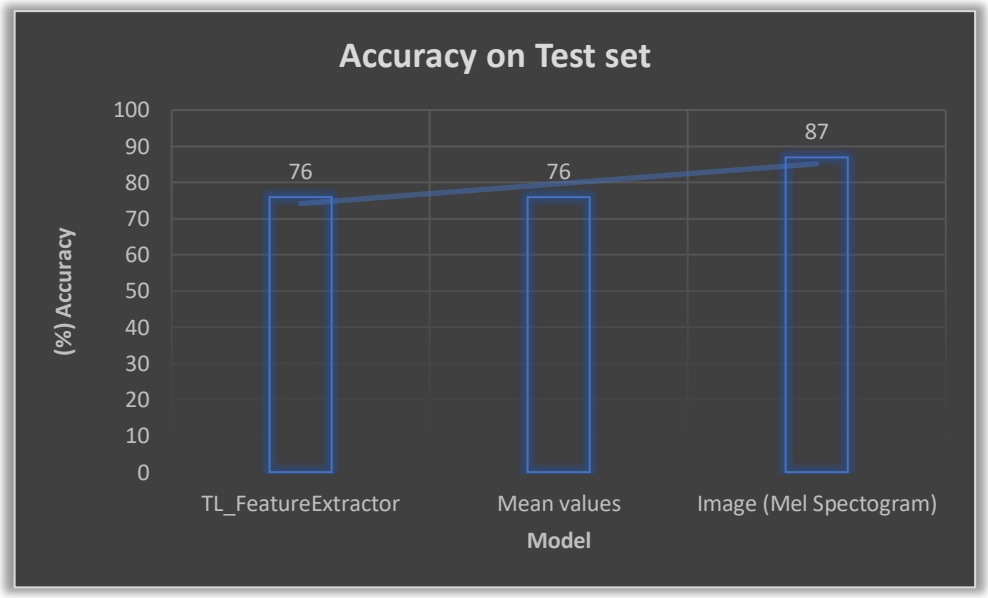
<u>מסקנות מהניסוי:</u>

1. שימוש בטכניקה Transfer-learning הנמיכה את תוצאות הניבויים ולא שיפרה אותו.
2. גם כאשר העליתי מתוך סקרנות את כמות סבבי האימון פי 2, עדיין תוצאות הניבויים לא השתפרו.
3. כלומר השקעתי יותר משאבים (זמן, זיכרון וכו'), וקיבלתי תוצאות פחות טובות. מודלים הרבה יותר פשוטים השיגו תוצאות טובות יותר. אדון בתוצאות המעניינות בפרק הסיכום.

סיכום

1. דיון בתוצאות

אציג את שלושת המודלים הפשוטים ביותר, שהשיגו את הדיוק הגבוה ביותר, ולאחר מכן אדון בתוצאות של פרק הניסויים:



כשהתחלתי את הניסויים, ההשערה היחידה הייתה שאופן הייצוג של קבצי השמע ישפיע על תוצאות הדיוק של הניבויים. כעת בסיכום, אני יכול לקבוע שבהחלט אופן הייצוג משפיע.

מה שהפתיע אותי, היה שדווקא המאפיין Mel Spectogram אשר לא בלט בהבדל של ערכי החציונים שלו בין הסיווגים השונים של הדגימות, נתן את הדיוק הכי גבוה ברגע שהוא יוצג בתור תמונה (גרף). כלומר האופן שבו הוא משתנה לאורך ההקלטות ניתן לזיהוי, ומהווה מרכיב חשוב בקביעת הסיווג.

נקודה נוספת שהפתיעה אותי, הייתה שהשימוש ב-Transfer-learning לא הביא לדיוק הכי גבוה. כלומר המודל שהשתמשתי בו לטובת טכניקה זו- מודל ה-VGG16, לא עזר למצוא מאפיינים נוספים בתמונות שלי. אחת ההשערות שלי היא שמודל זה בנוי לזיהוי של כ-1000 סיווגים שונים שביניהם; חתולים, כלבים וכו', אך אינו בנוי לדליית מאפיינים מגרפים אקוסטיים, ולכן הוא לא תרם להשגת הדיוק המקסימלי.

הנקודה האחרונה שאציין בדיון, היא שמתוצאות הניסויים, אני חושב שהוספת מידע נוסף לתמונה אינו שקול להוספת מאפיין נוסף לוקטור מאפיינים מתמטי. מחשבה זו נובעת מכך שבניסוי הרביעי הוספתי גרף (כלומר היו שני גרפים, במקום אחד, כאשר אחד הגרפים הציג מידע של שני מאפיינים ביחד), ועדיין הדבר הוריד ב-1% את תוצאות הדיוק מהניסוי שקדם לו, בו היה רק גרף בודד עם מאפיין בודד בכל תמונה. (אציין שוב נקודה זו ב- "Future work").

2. נקודות לשיפור

1. אני תוהה האם שינוי כמות סבבי האימון הייתה משנה את התוצאות. ההערכה שלי היא שלא, כיוון שלכל הפחות בניסוי אחד כשהעליתי את הסבבים הדבר לא השפיע על התוצאות היחסיות בין המודלים, אך עדיין אולי הייתי צריך לקבע את פרמטר סבבי האימון למספר גבוה יותר, ובעקבות כך לעבוד על מכונה מרוחקת חזקה במקום על המחשב הנייד שלי.
2. אולי הייתי צריך למצוא פיתרון לכך שלא כל קטעי השמע הינם באורך זהה. (כרגע קטעי השמע הינם באורך שבין 3-5 שניות). האתגר פה הוא שהפתרונות שנחשפתי אליהם אשר "דוחסים\מרחיבים" קטעי שמע מוסיפים גוון מטלי (metallic) לצליל, וחששתי שזה ישפיע על הניסויים.
3. השימוש ב- VGG16 בדיעבד לא היה בחירה מועילה לשיפור הדיוק. לא הצלחתי למצוא מודל שמסיק מתמונות על סיווג של מוזיקה, ולכן בחרתי ב-VGG16, אך אולי אם הייתי משקיע עוד זמן בחיפוש כן הייתי מוצא.
4. ההשוואה בין מודלים היא קשה, שכן ברגע שאני משנה ייצוג של קובץ שמע, בהכרח אני משנה גם את מבנה המודלים. כאשר הקלט הוא תמונה, ניתן להשתמש בשכבות כגון CONV2D, אך שכבות אלה אינן מועילות במיוחד במודלים שלא כוללים תמונות, לכן בסופו של דבר אני חושב שבהשוואה של מודלים שונים כן צריך לתת לכל מודל למצות את כמות סבבי האימון שהוא צריך כדי להגיע למיצוי הפוטנציאל שלו.

3. נקודות לשימור

1. המתודולוגיה הניסויית שבחרתי בה- קיבוע כמה שיותר "משתנים" כגון סבבי האימון, קבצי השמע, וכו' הוכיחה את עצמה ועזרה לי לצמצם כמה שיותר את ההבדלים שבין המודלים. על אף שכתבתי חלק מנקודה זו כנקודה לשיפור גם, אני חושב שהיא מתאימה לשתי הקטגוריות כי יש לה יתרונות וחסרונות, כאשר היתרון הוא שהיא עזרה להכריע מי המודל היעיל יותר בהתאם להתניות שהצגתי במתודולוגיה.
2. נקודות שצוינו בעבר כ-Future work באו לידי ביטוי בעבודה זו. הרצון להבין את השפעות השימוש ב-Transfer learning וב-Data augmentation התממש בעבודה זו, ותרם להבנת כלים נוספים בתחום הבינה.
3. הניסויים שביצעתי נבנו על סמך תוצאות של ניסויים קודמים. כלומר היה לי רעיון כללי על הניסויים שברצוני לבצע, אך ברגע שבניסוי הראשון קיבלתי תוצאה שסתרה את מה שחשבתי שאקבל, התאמתי את עצמי, השגתי קבוצת מבחן חדשה, ומשם והלאה בכל ניסוי הסקתי מסקנות, ובעקבות מסקנות אלו חשבתי על הניסויים הבאים.

4. כיוונים להמשך המחקר

1. מציאת מודל שאומן על תמונות בבעיות סיווג אקוסטיות, ושימוש בו ב- Transfer learning.
2. ביצוע מחקר נוסף על הדרך שבה מבצעים Data-augmentation על קבצי שמע. הבנתי כיצד הדבר מתבצע על קבצי תמונה, אך לא על קבצי שמע.
3. העמקת ההבנה על חשיבות חלוקת קבצי השמע למקטעים קצרים, ובפרט חקירת השאלה "האם קבצי השמע צריכים להיות באורכים שווים?" ואם הם אינם באותו אורך, מה ניתן לעשות בנידון?
4. קיום ניסויים נוספים שבהם שמים בתמונה אחת "עוד גרפים" ובודקים כיצד הביצועים משתנים כאשר מספקים למודל כמות סבבי אימון גדולה.

5. סיכום

בפרויקט זה למדתי רבות על האופן שבו ייצוגים שונים משפיעים על איכות הניבויים של מסווגים שונים. פרויקט זה חיבר שני תחומים שהתנסיתי בהם בעבר ושרציתי להמשיך ולהעמיק את הבנתי בהם- שימוש בבינה מלאכותית לטובת בעיות סיווג של קבצי שמע, וסיווג של קבצי תמונה.

נהניתי מאד מביצוע הפרויקט, בין היתר, כיוון שכמות השאלות שהוא העלה שווה לפחות לכמות התשובות שקיבלתי בו, כאשר בסופו של דבר, המטרה היא לנתב את הידע הטכנולוגי לטובת כלים שיעזרו לנו לחיות בעולם טוב יותר.

לסיכום, אני מודה לפרופ' שאול מרקוביץ', על כך שהסכים להנחות אותי בביצוע הפרויקט, שבעצם מהווה את הקורס האחרון שלי בתואר הראשון.

- לב טוניק

נספחים

מקורות מידע

כתובת למקור	תיאור
https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5	כתבה מעולה על הדרך שבה ניתן להבין ביצועים של מודל מסוים
https://towardsdatascience.com/a-comprehensive-hands-on-guide-to-transfer-learning-with-real-world-applications-in-deep-learning-212bf3b2f27a	מקורות שלמדתי מהם כיצד לבצע Transfer learning
https://medium.com/datadriveninvestor/cnn-architecture-series-vgg-16-with-implementation-part-i-bca79e7db415	VGG16 Explanation השתמשתי במודל זה בתור מודל מאומן בניסוי החמישי
https://machinelearningmastery.com/how-to-configure-image-data-augmentation-when-training-deep-learning-neural-networks/	מקור זה עזר לי להבין ש**לא מתאים** להשתמש בפיצ'ר של "הזזת פיקסלים בצורה אנכית אלא רק אפקית
https://blog.keras.io/building-powerful-image-classification-models-using-very-little-data.html	המדריך שעבדתי איתו כדי להבין איך להעביר את התמונות שיצרתי למודל, וכיצד לבנות מודל שרלוונטי לתמונות
https://www.quora.com/What-are-the-advantages-of-using-spectrogram-vs-MFCC-as-feature-extraction-for-speech-recognition-using-deep-neural-network	מה ההבדל בין שימוש ב-MFCC לבין שימוש ב-Mel spectrogram (תשובה קצרה- מל מבצע log ו-MFCC לא), ועפ"י ההשערות פעולה זו מובילה לתוצאות טובות יותר בשימוש ברשת נוירונים
https://mc.ai/audio-data-analysis-using-deep-learning-part-1/ https://en.wikipedia.org/wiki/Chromatic_scale	הסבר על סוג המאפיין chroma_stft ומידע כללי על נושא זה מתחום תאוריית המוזיקה
https://www.asee.org/documents/zones/zone1/2008/student/ASEE12008_0044_paper.pdf https://wiki.aalto.fi/display/ITSP/Zero-crossing+rate	העמקה בסוג המאפיין Zero crossings והבנה שצריך לבחון אותו ביחד עם עצמת הקול
https://towardsdatascience.com/extract-features-of-music-75a3f9bc265d	כתבה שמסבירה על סוגי הפיצ'רים השונים שניתן לחלץ מקבצי שמע
https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd	כתבה שמסבירה טוב על MFCC
https://www.researchgate.net/post/What_is_Mfcc_and_how_to_know_which_part_of_signal_mfc_coefficient_are_important_to_train_a_neural_network_model	תשובה שעוזרת להבין כיצד להתייחס ל-MFCC בתהליך של בחירת פיצ'רים חשובים עבור מודל
https://www.youtube.com/watch?v=gjxrDf6Pp6M	Dark theme jupyter- credit to a productivity tool

https://stackoverflow.com/questions/58451650/pip-no-longer-working-after-update-error-module-object-is-not-callable	Great for our eyes…
https://scikit-learn.org/stable/about.html#citing-scikit-learn	Credit for this awesome python package for machine learning

אתגרים משניים שבחרתי להוסיף כנספח

1. השתמשתי ב Jupyter Notebook בתור ניסוי, והאמת שממש נהניתי. אני ממליץ.

הקמת סביבת עבודה לפרויקט

1. הפרויקט צריך להתבצע בסביבת עבודה של windows
2. את כל הספריות שעשיתי להן Import שמתי באחד התאים העליונים במחברת ה-Jupyter שבה ביצעתי את הפרויקט, לכן אני ממליץ להריץ את תא זה ע"י (shift+enter) ופשוט להתקין את כל מה שהמחשב לא יזהה.
3. רשימת כלל הספריות שהיו מותקנות אצלי במחשב בעת ביצוע הפרויקט נמצאת בקובץ libs.txt בתקיה הראשית של הפרויקט.
4. לטובת מיקוד, להלן שתיים מהספריות שנדרשתי להתקין והאופן שבו עשיתי זאת:

- conda install -c conda-forge pydub
- conda install -c conda-forge keras

הסבר על מבנה תיקיית הפרויקט

DA code

Name	
.ipynb_checkpoints	
csv	כלל קבצי ה-CSV, לטובת היבט ה-Math (וקטור ערכי החציונים של כלל התכונות)
data	כלל התמונות, לטובת היבט ה- Image (כל מאפיין בתת תקייה משל עצמו)
images	תמונות שהשתמשתי בהן לטובת המסמך המסכם
legacy	תמונות וקבצים שעזרו לי במהלך הפרויקט, ואינם רלוונטיים יותר
plots	גרפים שעזרו לי במהלך הפרויקט
preview	בתקיה זו שמרתי דוגמא של שימוש ב Data-Augmentation
source_files	הקבצים המקוריים של קבוצת המבחן כפי שהורדו מהאתר "יו-טיוב".
test_wav	תיקיה שבתחילת הפרויקט העברתי אליה את קבצי קבוצת המבחן בתור תמונות ושמע
train	התיקיה המקורית שבה היו קבצי השמע של קבוצת האימון והוולידציה
weights	תיקיה שבה שמרתי את משקולות המודלים השונים שבדקתי
youTube_Movies_after_split	קבצי השמע של קבוצת המבחן, לאחר חלוקתם לכ-3 שניות לכל קובץ
features_mean-Pos_vs_Neg.png	
Image_vs_Math.ipynb	הקובץ שבו בפועל ביצעתי את כל הפרויקט
libs.txt	כלל הספריות שהיו מותקנות במחשב שלי בעת ביצוע הפרויקט
scaler.pkl	אובייקט נירמול אשר נבנה על קבוצת האימון והופעל על קבוצות הוולידציה והמבחן של הייצוג המתמטי

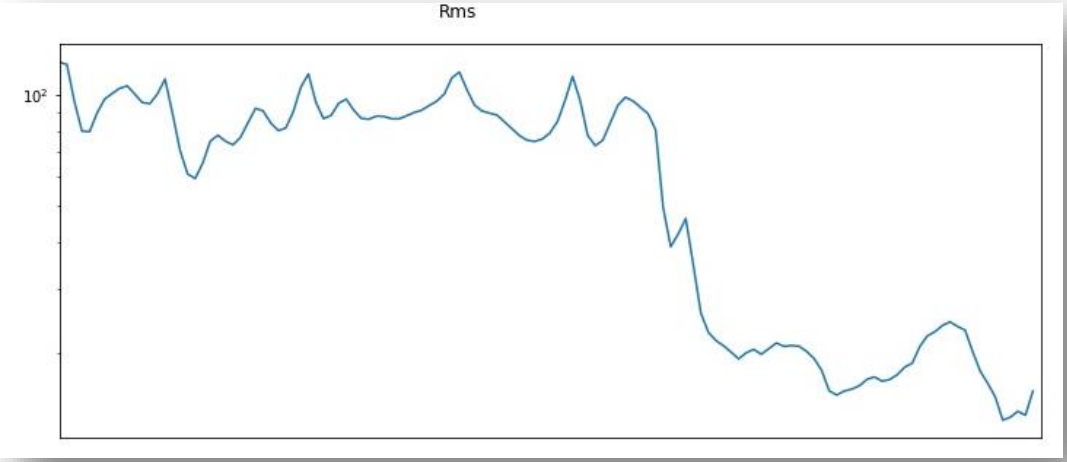
הבהרות נקודתיות

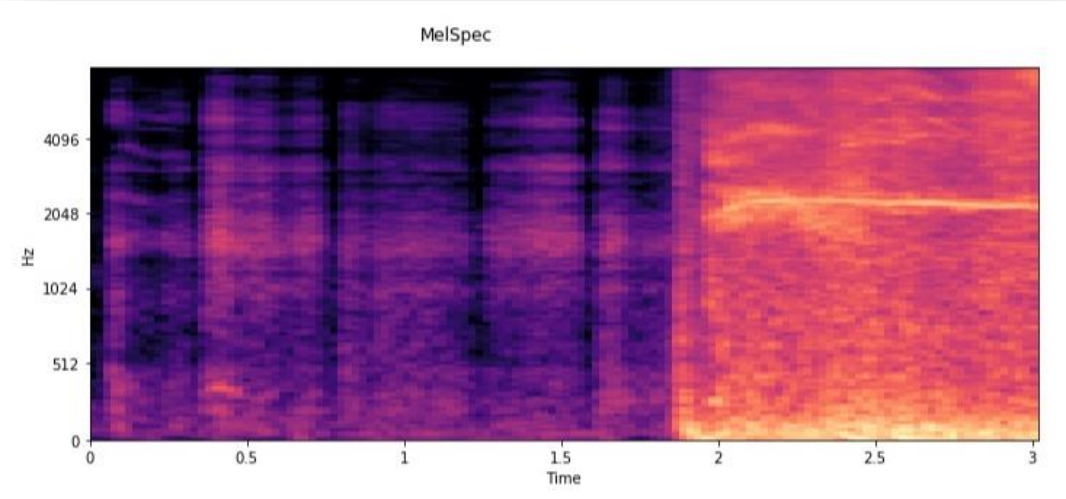
נירמול

בפרק הנתונים הגולמיים נרמלתי את כלל קבצי השמע כדי שאוכל להציגם בצורה אחידה בגרף ולהבליט איזה מאפיינים צפויים לעזור לי להבדיל בין Positives & Negatives. עם זאת בתהליך אימון המודל יש צורך בנירמול שנבנה רק על בסיס קבוצת האימון. הערה זו כתובה כדי להבהיר שבמהלך אימון המודלים, הנירמול בוצע רק על בסיס קבוצת האימון.

דוגמא לתמונות שהעברתי למודל

ב-2 הדוגמאות הבאות ניתן לשים לב כי קיים רק מידע ששייך לדגימה בודדת. שתי התמונות הבאות לקוחות מקבוצת המבחן מאותו הקובץ, אשר מייצג "צעקת מצוקה".





הסבר על מאגר המידע שבניתי בפרויקט הקודם (שימש בעבודה זו לקבוצת האימון והוולידציה)

1. משטרת ישראל - ניסיון שלא צלח לעת עתה.

בתחילה פנינו למשטרת ישראל כדי לקבל גישה לשיחות טלפון שנעשו למוקד החירום "100", אך המשטרה סירבה לספק לנו את ההקלטות. (הנימוק לסירוב והבקשה מצ"ב בנספחים, פירוט נוסף על מהות הבקשה יינתן בפרק הסיכום). לכן נאלצנו להסתמך על מקורות אחרים:

2. חברת גוגל (google) -בחרנו לבסוף שלא להשתמש במאגר זה.

לחברת גוגל קיים מאגר שנקרא AudioSet (לחץ פה להפניה למאגר). זוהי הייתה האופציה השנייה שחשבנו עליה, עם זאת בחרנו לבסוף שלא להשתמש במאגר זה עקב הסיבות הבאות:

1. אמינות- המאגר התבסס על סרטוני "יו-טיוב" (YouTube) שסווגו ע"י מכונות ונבדקו אקראית ע"י בני-אדם, כאשר הדיוק בסיווג שם עומד על כ-80% לפי מה שכתוב באתר. מבחינתנו דיוק זה אינו מספיק טוב למטרה רגישה כמו סיווג אנשים במצבי מצוקה.
2. איכות- כיוון שהמידע נלקח מ סרטוני "יו-טיוב" (YouTube), האיכות של מאפייני השמע שם הייתה בפורמט שאינו WAV ולכן היה שם אותנו במצב התחלתי פחות טוב מבחינת היכולת לבדל בין הקלטות שונות. (הסבר על חשיבות הפורמט תינתן בפסקאות הבאות).
3. פיצ'רים מוכנים מראש- המאגר ניתן להורדה בפורמט "טנזור-פלו רקורדס" (tensorFlow records) כאשר הפיצ'רים בו כבר חולצו מקטעי הוידאו. עצם המחשבה על כך שנשתמש בפיצ'רים מוכנים מראש, מקטעי וידאו ולא אודיו, באיכות שאינה האיכות הכי טובה שניתן למצוא, גרמה לנו סופית להבין שעדיף לנו לחפש מקורות אחרים. חשוב לנו לציין את קיומו של מאגר זה שכן ההתלבטות לגביו הייתה רצינית ושקולה ונמשכה זמן לא מועט.

3. מאגרי המידע שהשתמשנו בהם (לחץ על כל שם לטובת כניסה לאתר):

FSD
SHOCKWAVE-SOUND
Acoustic Event Dataset
FreeSound
soundbible
google AudioSet

בחרנו באתרים הנ"ל מהסיבות הבאות:

1. איכות- קבצי שמע בפורמט WAV.

פורמט זה הוא הפורמט הכי איכותי, כלומר הכי פחות מכווץ, אשר זמין במאגרי מידע קוליים ברשת. (לחץ כאן ללינק שמסביר את ההבדלים בין פורמטי wav ל-mp3)

2. מיקוד- קצר ולעניין.

הקבצים המוצעים הינם בין 1-5 שניות ברובם, כאשר קבצים ארוכים מכך נוחים לסינון אוטומטי כחלק מהממשק המוצע באתרים.

החשיבות בקבצים קצרים הינה קריטית לאלגוריתם שלנו. אנו הרי מחפשים מאפיינים קוליים ייחודיים, כאשר אנו רוצים לאמן מסווג שיזהה מאפיינים אלה. לכן ההשערה שלנו הינה שככל שנספק קבצים קצרים ואיכותיים יותר, המסווג שלנו יהיה יעיל יותר בזיהוי,

כאשר לקטעים אלה נוכל בהמשך להוסיף רעשי רקע שונים ע"י data-augmentation. (פירוט נוסף על כך בפרק הניסויים).

4. אופן איסוף המידע: (חצי אוטומטי)

Acoustic Event Dataset

הורדה פשוטה של מאגר שלם עם עשרות קבצים רלוונטיים ("צעקה") בלחיצת כפתור בודדת, רוב הקבצים הלא רלוונטיים נמחקו באופן ידני.

SHOCKWAVE-SOUND

הורדה של עשרות קבצים ע"י כלים להורדה אוטומטית כגון (לחץ על השם לכניסה ללינק) Simple mass downloader Batch Link Downloader. הכלים דורשים שהמשתמש ייכנס לדף האינטרנט הספציפי, ורק אז יוכל להוריד ממנו קבצים בפורמט מסוים.

FSD FreeSound

בתחילה ניסינו להשתמש ב-API להורדה מסיבית של קבצים איכותיים לפי סינון שקבענו. הדבר אפשרי טכנית, ובילינו מספר שעות בניסיון להשתמש ב-API, אך כשהתקלות הנפוצות הפנו אותנו ללמידת אלגוריתמי **OAuth2** (לינק לעמוד ההסבר שעקבנו אחריו), החלטנו שעדיף לא לבזבז זמן נוסף בהבנת פרוטוקולי הזדהות ופשוט להוריד קבצים ידנית. ההחלטה לא הייתה פשוטה שכן המשמעות שלה הייתה שבפועל בילינו שעות בהורדת קבצי שמע, עם זאת אנו שלמים עם ההחלטה שנלקחה בזמנה, שכן היא אפשרה לנו להתמקד בדברים יותר מהותיים מבחינתנו ולהתקדם לשלבים הבאים. יתרון נוסף הוא שהמאגר שלנו מאד איכותי שכן הוא וודא עד רמת הקובץ הבודד על ידינו. (פירוט נוסף על ראייתנו על דליית מידע מהאינטרנט בפרק הסיכום).

5. דגשים על לוגיקת איסוף המידע

בראייתנו, החשיבות **לאיכות** הקבצים הייתה קריטית- לכן נבחר הפורמט wav. בנוסף, הקבצים עצמם היו חייבים להיות כמה שיותר **ממוקדים** על המאפיינים הקוליים שחיפשנו. כך הגענו למסקנה שהקבצים עם הסיווגים החיוביים ("צעקה", "בכי" וכו') צריכים להיות באורך 3-5 שניות בממוצע. דגימות שליליות כמו רעשי רקע שונים ("רחוב, אוניברסיטה, כביש, חיות וכו') אשר אורך הקלטתם היה מעל לזמן זה, חולקו אוטומטית למקטעים של 5 שניות ע"י הכלי NCH (לחץ לכניסה לאתר הכלי) כדי ליצור אורכי מקטעים אשר תואמים לדגימות החיוביות. המטרה באחידות יחסית של זמני המקטעים נבעה מהרצון לשלול מקרים בהם המסווג יקבל מקטע ארוך יותר ויסיק על סמך כך שהקטע הינו בעל סיווג שלילי.