The Andrew and Erna Viterbi Faculty of ELECTRICAL & COMPUTER ENGINEERING

SIPL SIGNAL AND IMAGE PROCESSING LAB

TECHNION Israel Institute of Technology

דו"ח סיכום פרויקט: א'

זיהוי דיבור אוטומטי לקריאת תורה עם טעמי המקרא

ASR for Torah reading with cantillations

מבצעים:

אביב שם טוב **Aviv Shem-Tov**

אורי לוי **Ori Levi**

מנחים:

ד"ר אורן משלי **Oren Mishali**

נמרוד פלג **Nimrod Peleg**

סמסטר רישום: חורף תשפ"ד

תאריך הגשה: אוגוסט, 2024

תודות

ברצוננו להביע את הערכתנו העמוקה לאנשים הבאים, אשר תרמו משמעותית להצלחת הפרויקט:

ד"ר אורן משלי

תודתנו נתונה לד"ר משלי על הנחייתו המקצועית, תמיכתו המתמדת והכוונתו החכמה לאורך כל הדרך. הכלים שסיפק לנו למדידת הצלחת המודל היו בעלי ערך רב ותרמו באופן משמעותי לתוצאות הפרויקט.

מר נמרוד פלג

אנו מודים למר פלג על קבלת הפרויקט בהתלהבות ובפתיחות. עצותיו המועילות והטיפים שחלק עמנו לאורך התהליך היו נכס יקר ערך עבורנו.

מר יאיר משה

תודה מיוחדת למר משה על תרומתו המשמעותית לפרויקט. הטיפים שסיפק והתבנית שיצר לדו"ח המסכם היו לעזר רב בכתיבת דו"ח זה ותרמו רבות לאיכותו הסופית.

תרומתם של אנשים אלה הייתה מכרעת בהצלחת הפרויקט, ואנו אסירי תודה על מעורבותם ותמיכתם.

תוכן עניינים

רשימת איורים

רשימת טבלאות

1. תקציר

מטרת הפרויקט היא ליצור מודל "דיבור לטקסט" שמזהה קריאה בתורה עם טעמי המקרא ומתמלל את הפסוקים כולל הטעמים. מודל זה יאפשר זיהוי טעויות בקריאה והצעת תיקונים, ובכך ישפר את הקריאה בטעמי המקרא.
הפרויקט מתמקד בפיתוח מערכת שתוכל להאזין לקריאות תורה, לזהות את הטקסט, ולהפיק תמלול מדויק כולל טעמי המקרא. הטעמים הם סימנים מיוחדים שמלווים את הטקסט התנ"כי ומציינים את אופן ההגייה, הניגון והדגשת המילים.
לשם כך, נעשה שימוש בטכנולוגיות מתקדמות של למידת מכונה ורשתות נוירונים, תוך התאמה ספציפית לשפה העברית ולקריאה המסורתית של טעמי המקרא. המודל אומן על מאגרי נתונים של קריאות מוקלטות, ונעשו התאמות כדי להבטיח זיהוי מדויק של הטקסט והטעמים.
המודל הראה יכולת מבטיחה בזיהוי טעמי המקרא, עם דיוק של עד 51.7% במדד F1 על סט מבחן חיצוני. שימוש במודל Whisper של OpenAI ובאוגמנטציות ספציפיות לשמע הוביל לשיפור משמעותי בביצועים. אימון על מודל שאומן מראש על דאטה רב בעברית הראה שיפור ביכולת ההכללה על דוברים וסגנונות קריאה שונים, ומדגיש את החשיבות של שימוש במודלים חזקים ונתונים מגוונים.
מסקנות עיקריות הן שניתן לפתח מודל יעיל לזיהוי קריאה בתורה עם טעמי המקרא, אך הגדלת מאגרי הנתונים והוספת דאטה מגוון חיוניים לשיפור יכולות המודל. פיתוח שיטות יעילות לחיתוך והתאמה אוטומטית של קטעי אודיו ארוכים עשוי להוות מפתח להרחבת מאגר הנתונים ויאפשר אימון מודלים חזקים עוד יותר.

Abstract

This project aims to develop a speech-to-text model that recognizes Torah readings with cantillation marks (Trop) and transcribes the verses accurately, including the cantillations. This model will enable the detection of reading errors and suggest corrections, thereby improving the accuracy of Torah readings.
The project focuses on developing a system capable of listening to Torah readings, identifying the spoken text, and generating an accurate transcription, including the cantillation marks. These marks are special symbols accompanying the biblical text that indicate pronunciation, intonation, and word emphasis.
Advanced machine learning and neural network technologies are utilized, specifically adapted to the Hebrew language and the traditional reading of cantillations. The model is trained on datasets of recorded readings, with adjustments made to ensure precise recognition of both text and cantillations.

The model demonstrated promising ability in recognizing cantillation marks, achieving up to 51.7% F1 score on an external test set. Utilizing OpenAI's Whisper model and specific audio

augmentations led to significant performance improvements. Training on a model pre-trained on a large Hebrew dataset showed improvement in generalization ability across different speakers and reading styles, highlighting the importance of leveraging powerful models and diverse data.

Key conclusions are that developing an effective speech-to-text model for recognizing Torah readings with cantillation marks is feasible, but expanding and diversifying training datasets is essential for further improvement. Developing efficient methods for automatically segmenting and aligning long audio clips may be key to expanding the training dataset and enabling the training of even more powerful models.

2. רשימת קיצורים

ASR: Automatic Speech Recognition זיהוי דיבור אוטומטי

RNN: Recurrent Neural Network רשת נוירונים נשנית

CNN: Convolutional Neural Network רשת קונבולוציה

Log-mel Spectrum: ייצוג של אות הדיבור המבוסס על סקאלת ה-Mel

LLM: Large Language Model מודל שפה גדול

3. מבוא

קריאה בספר התורה היא אחד מהטקסים היהודיים הקדומים ביותר, וזה מאות רבות של שנים היא תופסת חלק מרכזי בתפילות ימות השנה ובפרט בתפילות השבתות, המועדים וראשי החודשים. קריאת התורה נערכת בבית הכנסת בתפילות שחרית של ימי שני וחמישי, שבת, ימים טובים, ראשי חודשים, חנוכה, פורים ותעניות, ובתפילות מנחה של שבת ושל תעניות.

הקריאה בספר נערכת מעל "במה" ייעודית במרכז בית הכנסת, לרוב מוגבהת מעט, שצורתה מיועדת לפתיחה נוחה של הספר לשם קריאה; בבתי כנסת במסורות עדות המזרח (יהודים שחיו בארצות מוסלמיות בעיקר מזרח התיכון וצפון אפריקה) המשטח העליון של הבימה מאוזן, שכן הם קוראים מתוך ספר תורה כשהוא מאוחסן בתיק המגן הקשיח שלו, המכונה "תיק" או "נרתיק", ואילו במרבית בתי הכנסת האשכנזיים (ארצות אירופה) המשטח מוטה ועליו מונח בשיפוע ספר התורה ללא נרתיק הבד הייעודי ("מעיל") שבו הוא מאוחסן רק בשעה שאין קוראים בו. מעניין לציין שיש גם קהילות ספרדיות ותימניות שבהן בעת הקריאה מטים את הספר עם תיקו הקשיח ומשעינים אותו על מעקה הבימה.

הקריאה בתורה נעשית במעמד של עשרה גברים יהודים לפחות ("מניין") והיא בלי ניקוד ובהגייה קפדנית ובנעימה מוטעמת ("טעמי המקרא") המשתנות לפי מסורת העדות השונות. בעבר היו מסמיכים לקריאת התורה גם את קריאת תרגום התורה לארמית (תרגום אונקלוס), כדי להקל על הבנת הנקרא בתקופה שבה השפה המדוברת הייתה ארמית (רמב"ם, הלכות תפילה יב, ו), אך מנהג זה בטל בימינו והשתמר רק בחלק מקהילות תימן (תחת השם "תרג'ום"), שבהן נהוג כי אחד הילדים חוזר על תרגום הפסוקים לאחר קריאתם.

הקטע שאותו קוראים בבית הכנסת מתחלק בין מספר מתפללים המכונים "קרואים" או "עולים לתורה" (וגם "עולים"). כיום מקובל במרבית בתי הכנסת האורתודוכסיים כי העולה לתורה יהא זכר יהודי מגיל שלוש עשרה. כיום, בשל ההקפדה על כך שהקריאה תהיה מדויקת וללא כל שגיאה ולו הקטנה ביותר מחד גיסא, והיעדרן של יכולות אלו אצל רוב העולים מאידך גיסא, נהוג להעמיד לצד העולה "בעל קורא" – אדם הבקיא בקריאה ודקדוקיה; בעל הקורא קורא בתורה בקול רם, והעולה קורא לעצמו בלחש. לעתים עומד לצד בעל הקורא "מכוון", שמרמז לבעל הקורא על הטעמים בקטע הקריאה.

לדקדוק של הקריאה בתורה חשיבות רבה מאוד ובעל הקורא (שקורא בתורה) אמור להתכונן אליה לפי ההלכה לפחות 3 פעמים כדי שיקרא אותה ברצף ובלי טעויות. כיום האופציות שעומדות בפני בעל הקורא להתכונן לקריאה בתורה הם למידה מבעל קורא מומחה שהיא בתשלום או מאינטרנט מאנשים פחות מקצועיים ובלי פידבק.

הפרויקט שלו מהווה פתרון ללימוד טעמי המקרא חינמי, זמין, נגיש ועם יכולת לתקן את הלומד טעמי המקרא. לשם כך הפרויקט מתבסס על מודלי דיבור-קריאה שיתנו את המענה לזהות את הקריאה עם הניגון ונתינת פידבק בו זמנית .

4. מבוא לטעמים

4.1 מבוא לטעמי המקרא

4.1.1 היסטוריה ומקור טעמי המקרא

טעמי המקרא, הם מערכת סימנים ניגוניים ותחביריים המשמשים בקריאת התנ"ך העברי. טעמי המקרה אינם כתובים בספרי התורה העשויים קלף, אלא בספרי התנ"ך הרגילים.

מקורם של הטעמים (הניגונים) עתיק יומין, עוד מימי האמוריים (המאות ה-2–5 לספירה) ישנם אזכורים על כך שהם כה עתיקים. אך סימני הטעמים כפי שאנו מכירים אותם כיום נחתמו סמוך למאה השמינית.

טעמי המקרא מתחלקות לשתי מערכות עיקריות - טעמי המקרא לספרי אמ"ת (איוב, משלי, תהלים) וטעמי המקרא לשאר 21 הספרים בתנ"ך.
בפרויקט שלנו אנו מתמקדים במערכת טעמי המקרא של 21 הספרים, משום שיש להם דרישה יותר גדולה (מכיל את כל התורה וכל נביאים) ומשום שניתן יהיה להשיג דאטה עבורם. בנוסף נוכל לציין כי יש קורלציה בין המערכות, ולכן בהינתן דאטה מתאים אמור להיות קל להתאים את המערכת.

4.1.2 תפקידם של טעמי המקרא בקריאה

טעמי המקרא ממלאים מספר תפקידים חשובים:

1. **ניגון**: הטעמים מנחים את הקורא כיצד "לנגן" את הטקסט.
2. **משמעות המילים**: דגש של מלרע ומלעיל, הבדל בדגש משנה את המשמעות, גם בדיבור כיום: "הוא אוכֹל אוֹכל". במשפט זה על אף שהמילים נשמעות אותו דבר מבחינת הצלילים, יש דגש והארכה של חלקים שונים במילה ורוב דוברי השפה "ירגישו" שמדובר במילה שונה.
3. **תחביר**: הטעמים מסמנים גם את המבנה התחבירי של המשפטים, איפה יש הפסקות ומה אורכן (מקביל ל",:." וכדומה)

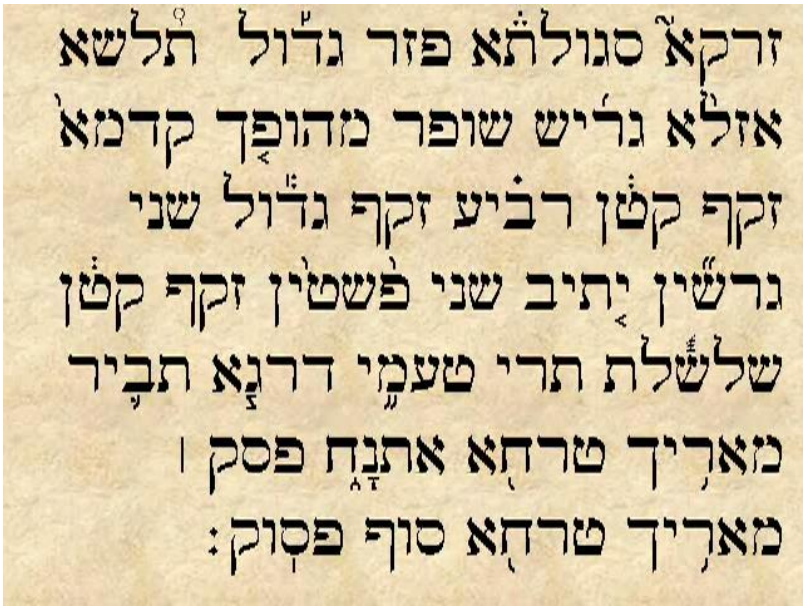
חשיבותם של הטעמים מתבטאת בכך שהם משמרים מסורת קריאה עתיקה ומסייעים בהבנה מדויקת של הטקסט המקראי.

4.1.3. פיצול לנוסחים שונים

טעמי המקרא התפתחו לאורך השנים במסורות שונות, מה שהוביל לפיצול לנוסחים מגוונים. הבנת הפיצול הזה חיונית לפיתוח מערכת מקיפה לזיהוי קריאה בטעמי המקרא, שכן סביר להניח כי על ידי למידה של נוסח מסוים לא נוכל לבצע חיזוי טוב לשאר הנוסחים.

4.1.4. דוגמאות של טעמי המקרא וקריאה בתורה

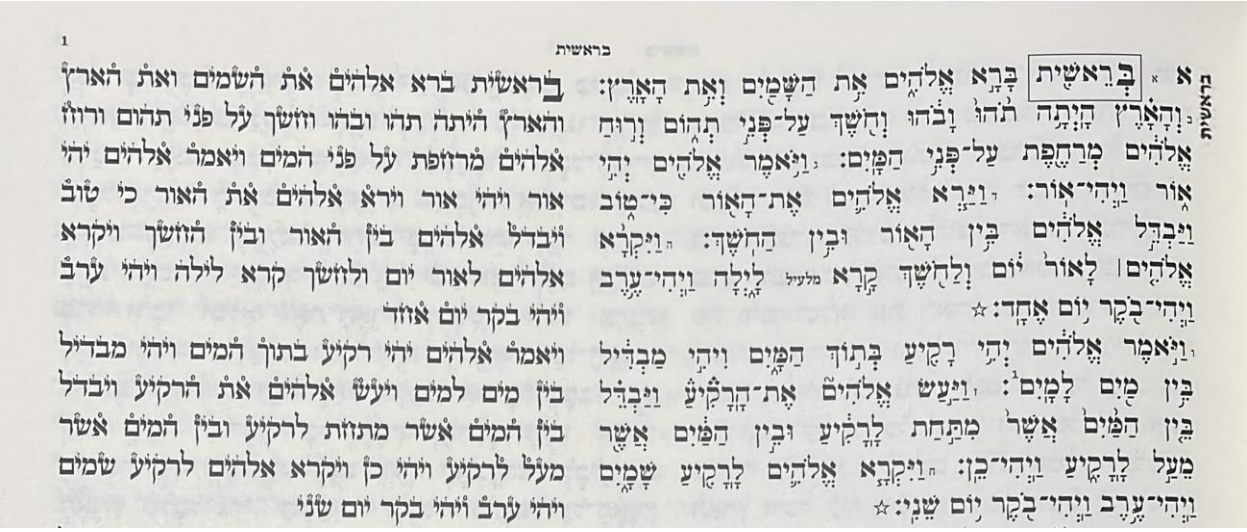
להלן רשימת טעמי המקרא כפי שמופיע בחומשים


זרקא סגולתא פזר גדול תלשא
אזלא גריש שופר מהופך קדמא
זקף קטן רביע זקף גדול שני
גרשין יתיב שני פשטין זקף קטן
שלשלת תרי טעמי דרגא תביר
מאריך טרחא אתנח פסק ׀
מאריך טרחא סוף פסוק׃

איור 1 – רשימת טעמי המקרא

וזה לקוח מספר לימוד טעמי המקרא שנקרא " תיקון קוראים"

מצד ימין זה המקרא עם ניקוד וטעמים לצורך למידה ומצד שמאל זה כמו שכתוב בספר תורה (בלי ניקוד וטעמים)


1 בראשית

בראשית א א בְּרֵאשִׁית בָּרָא אֱלֹהִים אֵת הַשָּׁמַיִם וְאֵת הָאָרֶץ׃ ב וְהָאָרֶץ הָיְתָה תֹהוּ וָבֹהוּ וְחֹשֶׁךְ עַל־פְּנֵי תְהוֹם וְרוּחַ
אֱלֹהִים מְרַחֶפֶת עַל־פְּנֵי הַמָּיִם׃ ג וַיֹּאמֶר אֱלֹהִים יְהִי
אוֹר וַיְהִי־אוֹר׃ ד וַיַּרְא אֱלֹהִים אֶת־הָאוֹר כִּי־טוֹב
וַיַּבְדֵּל אֱלֹהִים בֵּין הָאוֹר וּבֵין הַחֹשֶׁךְ׃ ה וַיִּקְרָא
אֱלֹהִים׀ לָאוֹר יוֹם וְלַחֹשֶׁךְ קָרָא מלעיל לָיְלָה וַיְהִי־עֶרֶב
וַיְהִי־בֹקֶר יוֹם אֶחָד׃ ס

ו וַיֹּאמֶר אֱלֹהִים יְהִי רָקִיעַ בְּתוֹךְ הַמָּיִם וִיהִי מַבְדִּיל
בֵּין מַיִם לָמָיִם׃ ז וַיַּעַשׂ אֱלֹהִים אֶת־הָרָקִיעַ וַיַּבְדֵּל
בֵּין הַמַּיִם אֲשֶׁר מִתַּחַת לָרָקִיעַ וּבֵין הַמַּיִם אֲשֶׁר
מֵעַל לָרָקִיעַ וַיְהִי־כֵן׃ ח וַיִּקְרָא אֱלֹהִים לָרָקִיעַ שָׁמָיִם
וַיְהִי־עֶרֶב וַיְהִי־בֹקֶר יוֹם שֵׁנִי׃ ס

בראשית ברא אלהים את השמים ואת הארץ׃
והארץ היתה תהו ובהו וחשך על פני תהום ורוח׃
אלהים מרחפת על פני המים׃ ויאמר אלהים יהי׃
אור ויהי אור׃ וירא אלהים את האור כי טוב׃
ויבדל אלהים בין האור ובין החשך׃ ויקרא׃
אלהים לאור יום ולחשך קרא לילה ויהי ערב׃
ויהי בקר יום אחד׃

ויאמר אלהים יהי רקיע בתוך המים ויהי מבדיל׃
בין מים למים׃ ויעש אלהים את הרקיע ויבדל׃
בין המים אשר מתחת לרקיע ובין המים אשר׃
מעל לרקיע ויהי כן׃ ויקרא אלהים לרקיע שמים׃
ויהי ערב ויהי בקר יום שני׃

איור 2 דוגמא מתוך ספר ללימוד טעמי המקרא

5. רקע לזיהוי דיבור

5.1. אותות דיבור

רקע לזיהוי דיבור כולל מספר מושגים בסיסיים באותות דיבור. פונמות (Phonemes) הן יחידות הצליל הקטנות ביותר בשפה שיכולות להבדיל בין משמעויות ומהוות את אבני הבניין הבסיסיות של מערכות זיהוי דיבור. פיץ' (Pitch) הוא התדירות היסודית של גל הקול, הנתפסת כגובה הצליל, ומשמש לזיהוי הנגנה (אינטונציה) ורגשות בדיבור. הוא מהווה מרכיב קריטי בזיהוי הטעמים, שכן הטעמים מתבטאים בין היתר כשינוי הפיץ'. פורמנטים (Formants) הם תדרי תהודה של מערכת הדיבור, המאפיינים בעיקר תנועות ומסייעים בזיהוי תנועות ומאפייני קול ייחודיים. עוצמת הקול (Intensity) מתייחסת לעוצמת הקול הנתפסת כרמת הקול ומשמשת לזיהוי הדגשים ורגשות בדיבור. הספקטרוגרמה (Spectrogram) היא ייצוג ויזואלי של הספקטרום של תדירויות כפונקציה של הזמן, מספקת מידע מקיף על מאפייני האות לאורך זמן ומאפשרת ניתוח מעמיק של המאפיינים האקוסטיים של הטעמים השונים במרחב התדר.

5.2. עקרונות בסיסיים בזיהוי דבור

עקרונות בסיסיים בזיהוי דיבור כוללים יצירת מאפיינים, מודל אקוסטי, מודל שפה ופענוח. יצירת מאפיינים מתבצעת באמצעות שיטות כמו Log-Mel Spectrum, Linear Predictive Coding (LPC), ומקדמי Mel-Frequency Cepstral (MFCC). LPC הוא טכניקה לייצוג הספקטרום של אות דיבור באמצעות מודל פרמטרי, המניח שכל דגימה של אות הדיבור יכולה להיות מנובאת כקומבינציה לינארית של מספר דגימות קודמות. Log-Mel Spectrum הוא ייצוג של אות הדיבור המבוסס על סקאלת ה-Mel, המדמה את תפיסת הצליל באוזן האנושית. MFCC הוא ייצוג של הספקטרום של הצליל בסקאלת ה-Mel, ומשמש כתכונות קלט נפוצות למודלים של זיהוי דיבור.

עבור קריאה בתורה, מקדמי MFCC או Log-Mel Spectrum עשויים להיות מתאימים יותר מ-LPC, שכן הם שומרים על יותר מידע רלוונטי לניגון והדגשים. מקדמי LPC אינם מתאימים שכן הם מכילים בעיקר מידע על צלילים של תנועות, ויש בהם איבוד מידע על הניגון והדגשים, שרלוונטי לנו. מקדמי MFCC שומרים על יותר מידע מ-LPC ומייצגים בצורה הדומה לתפיסה האנושית. Log-Mel Spectrum יכול להוות פתרון טוב לבעיה שלנו שכן הוא מייצג את הקלט בצורה טובה.

5.3. מודלים לזיהוי דיבור

מודלים שונים משמשים לזיהוי דיבור, כולל Hidden Markov Model (HMM), Connectionist Temporal Classification (CTC), ומודלים מבוססי End-to-End כמו Listen, Attend and Spell (LAS) ו-Wav2Vec.

HMM הוא מודל הסתברותי המניח שהדיבור הוא רצף של מצבים סמויים, כאשר כל מצב מייצר תצפית אקוסטית על פי התפלגות הסתברות. היתרונות של HMM כוללים גמישות ויכולת לטפל בשונות טמפורלית, אך הוא סובל ממגבלות כמו הנחת אי-תלות בין תצפיות עוקבות וקושי במידול תלויות ארוכות טווח. עבור טעמים שעשויים לפרוס צליל למשך זמן ארוך מאוד, הקושי יגדל בהרבה.

CTC היא פונקציית הפסד המאפשרת אימון end-to-end ללא צורך ביישור בין הצלילים לתווים עצמם, ומאפשר למודל ללמוד באופן אוטומטי את היישור בין האות לפלט.

מודלים מבוססי End-to-End כמו LAS משלבים encoder-decoder עם מנגנון attention ומאפשרים מיפוי ישיר מאות לטקסט ללא צורך במודלים נפרדים. Wav2Vec ומודלים מבוססי Self-Supervised מאפשרים למידה של ייצוגים אקוסטיים מנתונים לא מתויגים, מה שעשוי להביא לשיפור משמעותי בביצועים, במיוחד במצבים של מיעוט נתונים מתויגים.

ארכיטקטורות של מודלים מבוססי למידה עמוקה כוללות רשתות נוירונים קונבולוציוניות (CNN), רשתות נוירונים נשנות (RNN), רשתות LSTM (Long Short-Term Memory), Gated Recurrent Unit (GRU), ומודלי Transformer ו-CNN. Attention משמשות למיצוי תכונות אקוסטיות מספקטרוגרמות ומציגות יכולת ללמוד מאפיינים מקומיים ועמידות לשינויים קטנים באות. RNN מתאימות למידול רצפים טמפורליים ויש להן יכולת לזכור מידע לאורך זמן, אך סובלות מבעיית התאפסות הגרדיאנטים. LSTM מציעות פתרון לבעיית התאפסות הגרדיאנטים של RNN רגילות באמצעות מנגנון מבוסס שערים המאפשר למידה של תלויות ארוכות טווח. GRU היא פתרון נוסף לבעיית התאפסות הגרדיאנטים, יעיל יותר מבחינה חישובית ולעיתים בשימוש במודלים של ASR. מודלי Transformer ו-Attention מציעים מנגנון המאפשר למידה של קשרים בין חלקים שונים ברצף ברמה טובה מאוד, והם יעילים במיוחד בטיפול ברצפים ארוכים ובמשימות תרגום ושכתוב.

מודלים מבוססי **למידה עמוקה** כמו DeepSpeech, wav2vec, Whisper, ו-SeamlessM4T מציגים גישות מתקדמות לזיהוי דיבור ותרגום.

DeepSpeech הוא מנוע STT בקוד פתוח המשתמש בארכיטקטורת למידה עמוקה המבוססת על רשתות קונבולוציה ורשתות נשנות.

wav2vec היא גישת למידה בפיקוח עצמי לעיבוד דיבור, המכשירה מראש מודל על מערך נתונים גדול כדי ללמוד ייצוגים של שמע גולמי.

Whisper היא מערכת זיהוי דיבור המבוססת על ארכיטקטורת למידה עמוקה, שפותחה על ידי OpenAI ומיועדת למשימות זיהוי דיבור אוטומטי (ASR), תוך שימוש בשילוב של רשתות עצביות קונבולוציה ורשתות עצביות חוזרות.

מודל זה אומן בלמידה מונחית על מערך נתונים עצום (ביחס לשאר המודלים) המכיל 680,000 שעות של נתונים רב לשוניים מתויגים שנאספו מהאינטרנט.

SeamlessM4T היא מערכת זיהוי דיבור ומתרגמת רב לשונית המשלבת מודלים עצביים של רשתות קונבולוציה ורשתות חוזרות, תוך שימוש בארכיטקטורות חדשניות נוספות כגון Transformers.

5.4. אתגרים ייחודיים בזיהוי קריאה בטעמי המקרא

אתגרים ייחודיים בזיהוי קריאה בטעמי המקרא כוללים את מורכבות מערכת הטעמים, כולל מיקום הטעם במילה, ריבוי סימני טעמים ומשמעויותיהם המוסיקליות והתחביריות, ואתגר בזיהוי הבדלים עדינים בין טעמים דומים מאוד. בנוסף, ישנם הבדלים בין מסורות קריאה שונות, מה שמצריך מודלים חזקים וגמישים שיכולים להתאים למגוון רחב של נוסחים ולהתאים לנוסח ספציפי עם מעט דאטה. השונות היחסית גדולה בין הנוסחים השונים מציבה אתגר נוסף בפיתוח מערכת אחידה לזיהוי טעמי המקרא.

.

6. דאטה לאימון המודל

איסוף וארגון הדאטה היוו מרכיב קריטי בפרויקט שמטרתו זיהוי טעמים בקריאת תורה. המשימה המורכבת של המרת קריאה קולית לטקסט מנוקד עם טעמים העמידה בפנינו אתגר משמעותי: הצורך במקורות דאטה איכותיים המחולקים לקטעים קצרים (עד 30 שניות). לאחר חיפוש מקיף, זיהינו שלושה מקורות עיקריים שענו על דרישה זו: PocketTorah, Ben13, וספריא.

PocketTorah סיפק הקלטות בנוסח אשכנזי עם מבטא אמריקאי, מחולקות לקטעים קצרים המתאימים לדרישות המודל. Ben13 הציע את התורה כולה במספר נוסחים, כולם מוקראים על ידי אותו אדם, מה שאפשר עקביות בקריאה לצד גיוון בנוסחים. ספריא שימשה כמקור משלים לטקסט עבור הקלטות PocketTorah.

למרות שמקורות אלו אינם מגוונים באופן אידיאלי, הם סיפקו בסיס משמעותי לאימון המודל. כדי להרחיב את היריעה ולבחון את יכולת ההכללה של המודל, הוספנו דאטה מוגבל אך ייחודי מסרטון YouTube בנוסח ירושלמי ספרדי. דאטה זה, למרות היקפו המצומצם (13 דגימות של כ-30 שניות כל אחת), שימש לטסט חיצוני וסיפק תובנות חשובות על ביצועי המודל בסגנון קריאה שונה.

שילוב מקורות אלו אפשר לנו ליצור מאגר נתונים שמשלב בין עקביות לגיוון מסוים, המהווה בסיס לאימון מודל לזיהוי טעמי המקרא. עם זאת, האתגר של מציאת מקורות דאטה מגוונים יותר, תוך שמירה על הדרישה לקטעים קצרים, נותר משמעותי ומצביע על כיוון חשוב להמשך מחקר ופיתוח בתחום זה.

PocketTorah

קריאה בתורה בנוסח אשכנזי עם מבטא אמריקאי.

זהו דאטה שהשתמשנו בו בתחילת הפרויקט.

ספריא

מציעה API חינמי המאפשר למשוך את הטקסט המלא של התורה עם טעמים בפורמט,JSON מה שמאפשר גישה קלה ומדויקת לתוכן זה לצורכי פיתוח ומחקר.

ויקיטקסט (קרן ויקימדיה)

מכיל את כל התורה עם ניקוד וטעמים.

ben13

קריאה בתורה בכמה נוסחים עם מבטא ישראל.

הפסוקים באתר מחולקים לפי טעמי המקרא מה שמאפשר לעשות סיווג יותר טוב של המידע.

זהו הדאטה השני שאנחנו משתמשים בו.

הרב אבי זרקי- קריאת התורה פרשת "בראשית" מתוך יוטיוב

דאטה סט קטן (13 דוגמאות שיצרנו באופן ידני מתוך הסרטון)

7. תיאור הפתרון ותהליך העבודה

7.1 מודל

בתהליך פיתוח הפתרון שלנו, התמקדנו בבחירת המודל המתאים ביותר למשימה. תחילה, בחנו את מודל SeamlessM4T של פייסבוק. למרות שהמודל הפגין יכולת טובה בהבנת התוכן הכללי של הדיבור, גילינו כי לעיתים הוא התקשה בדיוק ברמת המילה הבודדת. לדוגמה, הביטוי "בראשית ברא" תורגם ל"בהתחלה ברא". מכיוון שהדיוק ברמת המילה הוא קריטי עבור משוב מדויק על ניגון הטעמים, הבנו שעלינו לחפש פתרון אלטרנטיבי.

לאור המגבלות שזיהינו ב-SeamlessM4T, החלטנו לעבור לשימוש במודל Whisper של OpenAI. הבחירה ב-Whisper נבעה מהיתרונות המשמעותיים שהוא מציע. ראשית, המודל אומן על כמות עצומה של נתונים - 680,000 שעות של מידע רב-לשוני ורב-משימתי שנאסף מהאינטרנט. היקף זה של נתוני אימון מקנה ל-Whisper יכולת הבנה מעמיקה של שפות ודיאלקטים מגוונים.

מבחינה טכנית, Whisper משלב רשתות עצביות קונבולוציה (CNNs) עם רשתות עצביות נשנות (RNNs), מה שמאפשר לו לעבד ולהבין קלט שמע בצורה מדויקת ביותר. שילוב זה של טכנולוגיות מתקדמות מביא לביצועים מרשימים בתחום זיהוי הדיבור האוטומטי (ASR), עם דגש על דיוק מרבי.

המעבר ל-Whisper אפשר לנו להשיג רמת דיוק גבוהה יותר ב-ASR, דבר שהוא קריטי למשימה שלנו. יכולת זו מאפשרת לנו לספק שירותים איכותיים ויעילים יותר למשתמשים שלנו, תוך שמירה על נאמנות לטקסט המקורי ולניגון של הדובר. בסופו של דבר, השימוש ב-Whisper מהווה צעד משמעותי קדימה ביכולתנו לספק משוב מדויק על ניגון הטעמים, ובכך לשפר את חווית המשתמש ואת יעילות המערכת כולה.

7.2 דאטה

בתהליך הכנת הנתונים למודל, ישנם שני שלבים עיקריים של קידוד: קידוד השמע וקידוד הטקסט. שני תהליכים אלו חיוניים להכנת הנתונים בפורמט המתאים למודל, ומבוצעים על ידי מבנה ייעודי המכונה "מעבד הדאטה לאימון".

בשלב קידוד השמע, המודל מתוכנן לקבל קלט שמע בפורמט ספציפי - log-mel spectrum באורך של 30 שניות.
כדי להתאים את הקלט למבנה זה, מתבצעים שני תהליכים עיקריים:
1. ריפוד הקלט באפסים בסופו, כדי להביא אותו לאורך הנדרש של 30 שניות.
2. המרת הקלט ל-log-mel spectrum.

מעבד הדאטה מקבל שמע שנדגם בקצב של kHz16 ומבצע את ההמרה הנדרשת ל-log-mel spectrum, מה שמבטיח שהקלט יהיה בפורמט המתאים למודל.

בשלב קידוד הטקסט, התהליך מתמקד בהמרת הטקסט לייצוג שהמודל יכול לעבד. הפלט של המודל מיוצג על ידי טוקנים המוגדרים מראש במילון. לשם כך, מעבד הדאטה כולל טוקנייזר ייעודי. תפקידו של הטוקנייזר הוא דו-כיווני:

1. המרה מטוקנים לטקסט.
2. המרה מטקסט לטוקנים.

יכולת זו של הטוקנייזר מאפשרת גמישות בעיבוד הנתונים, הן בשלב ההכנה למודל והן בשלב הפענוח של תוצאות המודל.

שילוב של שני תהליכי הקידוד הללו - קידוד השמע וקידוד הטקסט - מבטיח שהנתונים יהיו מוכנים ומותאמים באופן מיטבי לעיבוד על ידי המודל. זה מאפשר למודל לבצע את משימת זיהוי הדיבור ביעילות ובדיוק מרביים, תוך ניצול מלא של יכולותיו המתקדמות בעיבוד שפה טבעית ואותות שמע.

7.2.1. דאטה – PocketTorah

בתהליך איסוף וארגון הנתונים לאימון המודל, התמקדנו במציאת מקור מתאים לקטעי שמע קצרים של קריאת התורה. לאחר חיפושים מקיפים, מצאנו כיPocketTorah מספק את הדאטה היחיד ברשת עם קטעים הקטנים מ-30 שניות, המתאימים לדרישות המודל שלנו.

מבנה הדאטה ב-PocketTorah מאורגן לפי עליות, כאשר כל פרשה מחולקת לשבע עליות. לכל עלייה קיימים שני קבצים עיקריים:

1. קובץ שמע: הכולל הקראה של העלייה בנוסח אשכנזי במבטא אמריקאי.
2. קובץ טקסט: המכיל רשימה של זמנים המציינים את תחילתה של כל מילה בהקראה.

לדוגמה:

פרשת יתרו עליה 7:

רשימת הזמנים:

2.17, 2.81, 4.56, 5.13, 5.34, 7.16, 7.78...

הטקסט המתאים (לא כלול בדאטה): וְכָל־הָעָם֮ רֹאִים֒ אֶת־הַקּוֹלֹ֣ת וְאֶת־הַלַּפִּידִ֗ם

בפועל המילה הנשמעת והזמן בו היא נשמעת:

וְכָל - [2.17, 2.81]

הָעָם֮ - [2.81, 4.56]

רֹאִים֒ - [4.56, 5.13]

אֶת - [5.13, 5.34]

הַקּוֹלֹת - [5.34, 7.16]
וכן הלאה

עם זאת, הדאטה מ- PocketTorah לא כלל את הטקסט עצמו של הקריאה, אלא רק את זמני המילים. לפיכך, נדרשנו להשיג את הטקסט המתאים בנפרד. ניסינו למשוך טקסט באופן ידני מויקיטקסט, אך התקשינו לאוטומט את התהליך עם חלוקה מדויקת לפי עליות, מה שהגביל את יכולתנו להשיג את כל הטקסט הנדרש.

כדי להתאים את הדאטה מ-PocketTorah לאימון המודל, יצרנו מבנה תיקיות מסודר:

- בתיקיית :"text" קבצי טקסט המכילים את הטקסט שהצלחנו למשוך מויקיטקסט.
- בתיקיית :"time" קבצי הזמנים המקוריים מ-.PocketTorah
- בתיקיית :"audio" קבצי השמע של ההקראות.

כל קובץ של עליה עם שם מתאים לדוגמה Noach-1.txt בהתאמה עם אותו שם בתיקיה time נמצאים הזמנים ועם אותו שם עם סיומת mp3 בתיקיה audio נמצאת ההקראה.

על בסיס מבנה זה, יצרנו דאטהסט המשלב כל מילה עם הטקסט המתאים לה, תוך שמירה על סדר המילים המקורי. בחרנו לשמור את הדאטה ברמת המילה הבודדת כדי לאפשר גמישות בחלוקת הדאטה לקטעים שונים, כולל קטעים חופפים. זה מאפשר לנו ליצור מגוון רחב יותר של דוגמאות אימון, כגון משפטים שונים המכילים חלקים משותפים.

לדוגמה בדאטה של האימון יהיה גם:
וְכָל־הָעָם֙ רֹאִ֣ים אֶת־הַקּוֹלֹ֔ת וְאֶת־הַלַּפִּידִ֖ם וְאֵ֣ת ק֣וֹל הַשֹּׁפָ֑ר
וגם:
וְאֶת־הַלַּפִּידִ֖ם וְאֵ֣ת ק֣וֹל הַשֹּׁפָ֑ר וְאֶת־הָהָ֣ר עָשֵׁ֔ן וַיַּ֥רְא הָעָ֖ם וַיָּנֻ֑עוּ וַיַּעַמְדֻ֖ מֵרָחֹֽק׃

למרות מאמצינו, ההתמודדות עם השגת הטקסט המלא באופן אוטומטי נותרה אתגר משמעותי. מגבלה זו הובילה אותנו לחזור ולחפש מקורות נוספים לדאטה, במטרה למצוא פתרון מקיף יותר שיכלול הן את השמע והן את הטקסט המתאים באופן מלא ומדויק.

המשך החיפוש אחר דאטה מתאים מדגיש את החשיבות של מקורות מידע איכותיים ומקיפים בפיתוח מודלים לעיבוד שפה טבעית, במיוחד כאשר מדובר בטקסטים מורכבים ומסורתיים כמו קריאת התורה.

7.2.2. דאטה – Ben13

במסגרת פיתוח הפרויקט, גילינו את אתר Ben13 המציע כמעט את כל התורה בארבעה נוסחים שונים, כולם מוקראים על ידי אותו אדם. פיתחנו שיטות scraping להשגת הדאטה של שמע וטקסט מהאתר. עם זאת, נתקלנו במגוון אתגרים בדאטה זה, כגון שמע שאינו תואם לטקסט, טקסט שאינו חלק מהתוכן המקורי (כמו מספרי פסוקים), וקבצי אודיו חסרים.

לניקוי הדאטה, פיתחנו שיטות לזיהוי מקרי קצה, כולל בדיקת אורך הטקסט והשמע, ובחינת הימצאות ניקוד וטעמים. לאחר ניקוי הדאטה, התאמנו את הדאטהסט כך שניתן יהיה לבחור אילו נוסחים ייכללו בו בזמן האימון.

בחלוקת הדאטה לסטים של אימון, ולידציה ומבחן, הקפדנו על שני עקרונות: שמירה על קרבה בין דוגמאות סמוכות לקבלת רצף הגיוני, והכללת כל הנוסחים של דוגמה מסוימת באותו סט, בשל הקורלציה בין הנוסחים והעובדה שמדובר באותו קורא. לשם כך, חילקנו את הנתונים לקבוצות של 50 דגימות כל אחת לפי הטקסט, וביצענו חלוקה של 80% אימון, 10% ולידציה, ו-10% מבחן, לאחר מכן החלוקה בכל הנוסחים הייתה לפי החלוקה של הטקסט.

התמודדנו עם מגבלת אורך השמע של המודל (30 שניות) על ידי חיבור קטעים קצרים סמוכים. בנוגע לטקסט, נתקלנו בהגבלה על אורך הטקסט בשל השימוש בשני טוקנים לכל תו ניקוד או טעם. פתרנו זאת על ידי הוספת טוקנים ייעודיים לניקוד ולטעמים, כך שכל טעם או ניקוד מיוצג על ידי טוקן אחד בלבד. החלטנו להסיר את הניקוד כדי להתמקד בטעמים ולהקל על המודל, לאחר שניסויים קצרים הראו כי למידת הניקוד מכבידה על המודל.

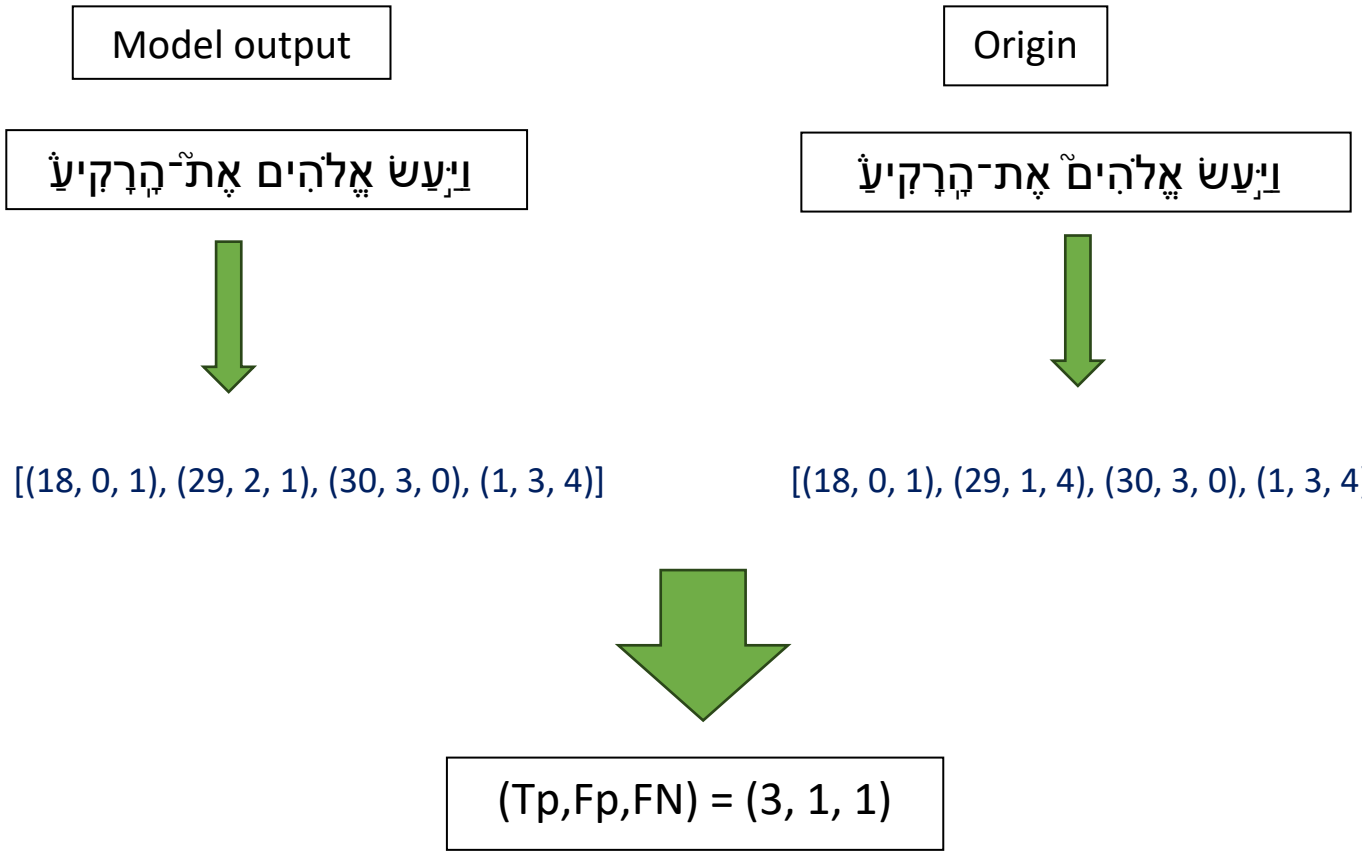
יצרנו אוגמנטציה ייחודית על ידי חיבור חלקים לא סמוכים, מה שאפשר יצירת כמות גדולה מאוד של דאטה לאימון. לדוגמה, יכולנו ליצור משפטים כמו "הוֹלִיד את־וכֹה תדבֹר: מעשֹׂך ובכֹל אֶיש על־וּברכך֑ והטֹוב קֹדש־הֶמה והי֑ה" עם אודיו מתאים.

ניסויים ראשונים הראו תוצאות מעודדות כאשר גם טעמים הוגדרו כתווים. עם זאת, המשכנו לחפש דרכים מותאמות יותר לבעיה הספציפית שלנו כדי לבדוק את המודל בצורה מדויקת יותר.

7.3. הגדרת מטריקה המתאימה לבעיה

המטרה שלנו היא שהמודל ידייק בטעמים עצמם, לא חשוב לנו אי דיוקים קלים בשאר הטקסט.
על מנת לאמת את המודל שלנו ,שמרנו את הטעמים במערך כך שלכל טעם יש את האינדקס שלו
והגדרנו את המדד הבא:
לכל טעם נגדיר את השלישייה הבאה :
(מיקום האות במילה, מיקום המילה במשפט, מיקום הטעם בטבלת הטעמים)

ניצור רשימה לכל קטע שיצא מהמודל שלנו ורשימה לפי המקור בפסוקים ונשווה ביניהם והמדדים כדלהלן:
TP – True Positive כמות השלישיות שזהות
FP – False Positive כמות השלישיות שהמודל נתן והם לא נכונות
FN – False Negative כמות השלישיות שהמודל החסיר
TN – True Negative כמות השלישיות האפשריות שהמודל לא כתב וזה נכון. לפי השלישייה שהגדרנו ,
המספר הזה יכול לשאוף לאינסוף ולכן הוא לא מייצג ולא נשתמש בו.
לדוגמא :



איור 3 דוגמא להגדרת מטריקה עבור המודל שלנו

לפי זה נגדיר 4 סוגי מטריקות שונות וכל אחת יכול להיות לפי Precision , Recall , F1 Score שהן:

1. TP FP FN Exact – בדיוק לפי מה שהוגדר לעיל.
2. TP FP FN with Letter Shift – הטעם ניחשב כ✓ גם אם מופיע באות אחת לפני או אחרי.
3. TP FP FN Word Level – הטעם ניחשב כ✓ גם אם מופיע באותה מילה(ולאוו דווקא באות הספיציפית)
4. TP FP FN with Word Shift – הטעם ניחשב כ✓ גם אם מופיע במילה אחת לפני או אחרי.

- TP FP FN Exact
- TP FP FN with Letter Shift
- TP FP FN Word Level
- TP FP FN with Word Shift

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

איור 4 המחשת המטריקות השונות

7.4. בניית בוט טלגרם

בתהליך פיתוח הפרויקט, יצרנו בוט טלגרם כדי לאפשר לאנשים להתנסות בקלות במודל שלנו. המשתמש שולח הודעה קולית או קובץ מוקלט מראש של קריאת קטע עם טעמים, והשרת מריץ את המודל על ההקלטה ומחזיר את הטקסט המוטעם למשתמש. הבוט אפשר לנו לקבל משוב מאנשים שונים על ביצועי המודל.

איור 5 דוגמא להרצת המודל על ידי בוט טלגרם

מהמשוב שקיבלנו, הבנו שהמודל לא הגיע לביצועים מרשימים עבור סגנונות קריאה שונים מאלו שהוא אומן עליהם. עם זאת, כאשר אנשים ניסו לדייק את סגנון הקריאה לזה של הקורא המקורי (בן13), שנחשב לאיטי יחסית, התוצאות היו טובות יותר. מכך הסקנו שתי מסקנות חשובות: ראשית, המודל מצליח להכליל בצורה טובה לקולות של אנשים שונים. שנית, ההגדרה של נוסח הקריאה אינה מספיקה, שכן כל נוסח מכיל מגוון סגנונות וקצבים שונים.

7.5. שיפור הפתרון

המשך הפיתוח של הפרויקט התמקד בשלושה היבטים עיקריים:

1. הדמייה של קריאה מהירה
2. הוספת אוגמנטציות מתקדמות לנתוני השמע
3. השילוב של שני מקורות הדאטה השונים שהיו ברשותנו.

כדי להתמודד עם הבעיה של הקריאה האיטית בדאטה המקורי, ניסינו לדמות קריאה מהירה יותר. ביצענו ניסיונות להאיץ את השמע תוך שמירה על התדר המקורי, אך גילינו שפתרון זה אינו מתאים לקריאה בטעמים בגלל הסלסולים הייחודיים שבה. האצת הקצב הגבירה את קצב הסלסולים באופן שאינו תואם לקריאה מהירה אמיתית.

לאור זאת, הבנו שעלינו להשקיע מאמצים נוספים בהרחבת מאגר הדאטה שלנו. חזרנו לדאטה של PocketTorah והחלטנו להשלים את קבצי הטקסט החסרים. לשם כך, השתמשנו ב- APIשל ספריא כדי למשוך את הטקסט המתאים. השתמשנו במודל שפה גדול (LLM) כדי ליצור 378 בקשות מותאמות ל- ,APIתוך התאמת הנוסח לזה של ספריא. ביצענו בדיקות על הדאטה החדש כדי לוודא את איכותו.

שמרנו כל קטע טקסט בקובץ נפרד, עם שם מתאים שהותאם לשמות של קבצי הזמנים והשמע הקיימים. גם כאן נעזרנו ב- LLMכדי לבצע את ההתאמה בצורה יעילה ומדויקת.
תהליך זה של הרחבת והעשרת הדאטה היה חיוני להמשך שיפור המודל שלנו, במיוחד לאור האתגרים שזיהינו בהתמודדות עם מגוון סגנונות וקצבי קריאה. השימוש בטכנולוגיות מתקדמות כמוLLM לצד עבודה מדוקדקת על הדאטה מדגיש את החשיבות של גישה רב-תחומית בפיתוח מודלים מורכבים כמו זה שלנו לזיהוי טעמי המקרא.

בתחום האוגמנטציות, הרחבנו משמעותית את מגוון הטכניקות שיישמנו, תוך שימוש בספריית audiomentations.
הטכניקות כללו:

1. הוספת רעש גאוסי (AddGaussianNoise): עם עוצמת רעש בין 0.001 ל-0.05 והסתברות הפעלה של 50%.
2. מתיחת זמן (TimeStretch): אפשרנו האטה עד 20% והאצה עד פי 4, עם הסתברות הפעלה של 100%.
3. שינוי גובה צליל (PitchShift): בטווח של 8 חצאי טון מעלה ומטה, עם הסתברות הפעלה של 100%.
4. הזזת שמע (Shift): עד 2 שניות, ללא גלישה של סוף הקטע לתחילתו.
5. סימולציית חדר (RoomSimulator): עם פרמטרים מגוונים לגודל חדר, זמן הדהוד ומרחק מיקרופון, והסתברות הפעלה של 50%.

אוגמנטציות אלו נועדו להגדיל את מגוון הדוגמאות ולשפר את יכולת ההכללה של המודל למצבי הקלטה וסגנונות קריאה שונים.

בנוסף לאוגמנטציות אלו, המשכנו עם האוגמנטציה הייחודית שפיתחנו קודם לכן, המשלבת מספר דוגמאות ליצירת דוגמה חדשה.

בהיבט של שילוב מקורות הדאטה, נדרשנו להתמודד עם האתגר של מבנים שונים מאוד בין שני המקורות. הדאטה מ-PocketTorah מחולק לקבצים קצרים עם מידע מדויק על זמני המילים, בעוד שהדאטה מ-Ben13 מכיל חלקים ארוכים יותר של קריאה רציפה.

כדי לשלב את שני מקורות הדאטה באופן יעיל, בחרנו בגישה פשוטה אך אפקטיבית. טענו כל אחד מהדאטהסטים בנפרד, ואז השתמשנו בפונקציה ConcatDataset של PyTorch לחיבור הדאטהסטים. גישה זו אפשרה לנו לשמור על המבנה הייחודי של כל מקור דאטה, תוך יצירת דאטהסט מאוחד לאימון המודל.

שילוב זה של אוגמנטציות מתקדמות יחד עם הרחבת מקורות הדאטה נועד לשפר משמעותית את יכולות המודל. הגישה המשולבת הזו מאפשרת למודל להיחשף למגוון רחב יותר של דוגמאות, סגנונות קריאה, ותנאי הקלטה, מה שצפוי לשפר את יכולת ההכללה שלו ואת דיוק הזיהוי של טעמי המקרא במגוון רחב של מצבים.

7.6. מבחן חיצוני

לצורך בדיקת המודלים לקחנו דאטה נוסף מהקלטת YouTube :

קישור: https://www.youtube.com/watch?v=DZYZndj2WsA

למרות היותו מוגבל בכמות (13 דגימות של כ-30 שניות כל אחת), מהווה צעד משמעותי בהערכת יכולת ההכללה של המודל. מדובר בקורא אחר המשתמש בנוסח ירושלמי ספרדי, אך בסגנון וקצב שונים מהנוסח הירושלמי של בן13. החלוקה הידנית של ההקלטה לקטעים והתאמת הטקסט באופן מדויק מבטיחה איכות גבוהה של הדאטה. למרות שכמות הדאטה קטנה מכדי לספק מדדים סטטיסטיים מדויקים, היא מאפשרת לקבל אינדיקציה ראשונית ליכולת ההכללה של המודל על קוראים וסגנונות שונים, לזהות פערים ספציפיים ביכולות המודל, ולהוות הוכחת קונספט חשובה ליכולתו להתמודד עם מגוון רחב של סגנונות קריאה. הוספת דאטה זה מספקת תובנות איכותניות חשובות על ביצועי המודל ויכולת ההכללה שלו, ומהווה צעד חשוב בתהליך הפיתוח והשיפור המתמשך של המודל לזיהוי טעמי המקרא. חשוב לציין כי למרות המגבלות הכמותיות, תוצאות הבדיקה על דאטה זה יכולות לכוון את המחקר העתידי, למשל בהצבעה על הצורך באיסוף דאטה נוסף ממגוון רחב יותר של קוראים וסגנונות, ובכך לתרום משמעותית לשיפור המודל בטווח הארוך.

7.7. המודלים שאימנו

כחלק מבחינת המודל אימנו מודלים עם היפר פרמטרים שונים, הטבלה הבאה מציגה את כל הקומבינציות שאומנו:

model	aug	data	nikud	rnd
Tiny	✓	Comb	×	×
Tiny	×	Comb	×	×
Base	✓	Comb	×	×
Small	✓	Comb	×	×
Small	×	Comb	×	×
Small	✓	**Ben13**	×	×
Small	✓	**Ben13**	×	✓
Small	✓	**PT**	×	✓
Small	✓	**PT**	×	×
Small	✓	**PT**	✓	×
Small	✓	**Yer**	×	×
Medium	✓	Comb	×	×
Large-v2	✓	Comb	×	×
v2-pd1-e1	✓	Comb	×	×

טבלה*1* רשימת המודלים שאימנו עליהם

כאשר:

Aug – האם בוצעו אוגמנטציות של אודיו

data – סוג הדאטה:

PocketTorah = PT, Ben13 = כל הנוסחים של בן13, Comb = כל הדאטה, Yer = נוסח ירושלים מתוך בן13

nikud – האם הטקסט כולל ניקוד

rnd – האם נעשה שימוש באוגמנטציה של חיבור חלקים אקראיים.

8. תוצאות

לאורך התהליך סט הוולידציה שלנו היה מורכב מפיצול של הדאטה המקורי עליו אימנו את המודל.
לדוגמה אם אימנו רק על PocketTorah הוולידציה הכילה רק את סט הוולידציה של פקטתורה.
בסוף התהליך בחנו את המודלים עם הדאטה הנוסף שהוספנו.

8.1. ערכים בכל המבחנים (ברירת המחדל):

8.1.1. קבוע:

Weight-decay=0.05
BATCH_SIZE = 8
LR = 1e-5
WARMUP_STEPS = 500

8.1.2. ואם לא נאמר אחרת אז:

יש אוגמנטציות
הדאטה זה דאטה משולב בן13 + PocketTorah
אין ניקוד
לא בוצעה האוגמנטציה של שילוב חלקים באופן אקראי

8.2. תוצאות המבחן של כל המודלים שאימנו

כאמור אימנו מספר מודלים להלן טבלה המפרטת על כלל המודלים:

model	aug	data	nikud	rnd	f1	rcl	prc	wer	loss
Tiny	✓	Comb	×	×	0.221	0.257	0.199	76.0	3.05
Tiny	×	Comb	×	×	0.195	0.226	0.177	80.5	2.65
Base	✓	Comb	×	×	0.323	0.349	0.304	65.4	4.66
Small	✓	Comb	×	×	0.367	0.394	0.348	57.2	1.75
Small	×	Comb	×	×	0.303	0.348	0.275	67.8	1.95
Small	✓	**Ben13**	×	×	0.356	0.435	0.308	78.8	2.92
Small	✓	**Ben13**	×	✓	0.213	0.291	0.172	84.9	4.72
Small	✓	**PT**	×	✓	0.045	0.056	0.039	90.4	6.77
Small	✓	**PT**	×	×	0.082	0.097	0.074	89.4	3.65
Small	✓	**PT**	✓	×	0.080	0.122	0.061	92.1	19.44
Small	✓	**Yer**	×	×	0.260	0.261	0.258	79.1	3.47
Medium	✓	Comb	×	×	0.407	0.485	0.359	67.5	1.63
Large-v2	✓	Comb	×	×	0.448	0.510	0.408	63.4	1.31
v2-pd1-e1	✓	Comb	×	×	0.517	0.615	0.454	58.6	1.44

טבלה2 תוצאות המבחן של כל המודלים שאימנו

f1 – ציון F1 Exact שהוגדר מעלה

Rcl – ציון Recall Exact שהוגדר מעלה

Prc – ציון Precision Exact שהוגדר מעלה

Wer – Word Error Rate

Loss – cross-entropy loss

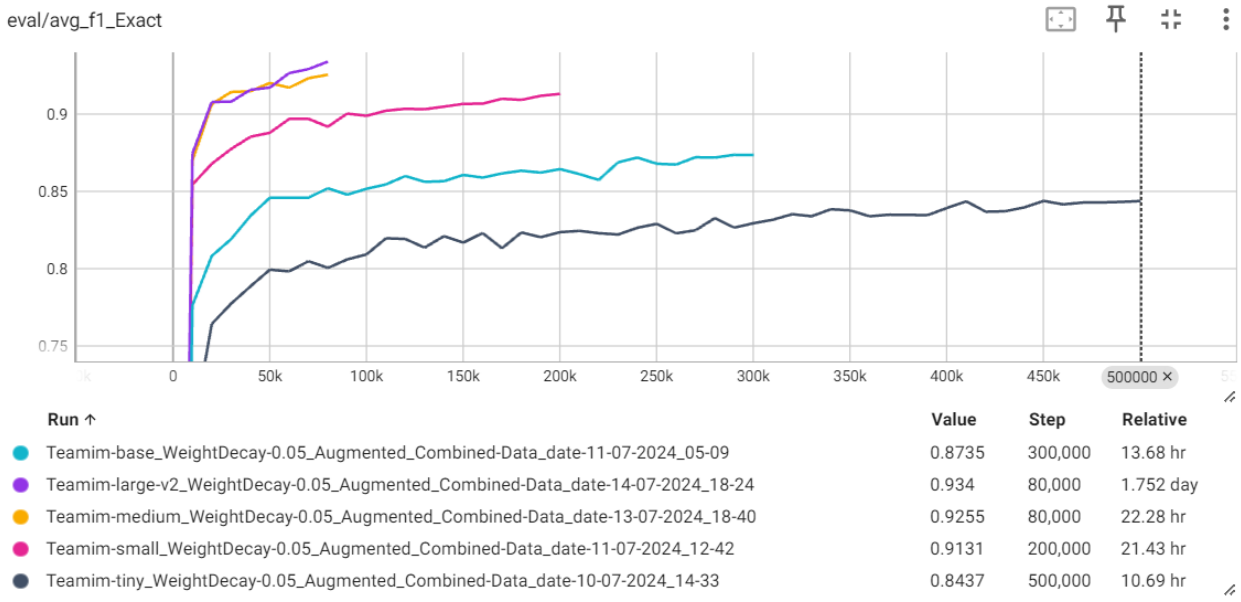
כעת נסתכל על מספר השוואות בין מודלים, הן על סט הוולידציה והן על סט המבחן.

8.3. תוצאות המודלים בגדלים השונים

ביצענו השוואה בין הגדלים השונים של המודלים, כל המודלים אומנו עם ערכי ברירת המחדל שלנו (ראו סעיף 8.1)

8.3.1. תוצאות על סט הוולידציה

להלן תוצאות הוולידציה כתלות במספר הצעדים:



איור 6 גרף של תוצאות המודלים בגדלים השונים מול סט הוולידציה

ניתן לראות כי אכן ככל שהמודל גדול יותר הוא מגיע לתוצאות טובות יותר בפחות צעדים, אך כל צעד שלו לוקח יותר זמן. למשל, ניתן לראות כי עבור מודל large הזמן שלקחו 80,000 הצעדים היה גדול מ-42 שעות.

8.3.2. תוצאות על סט המבחן החיצוני

להלן טבלה של התוצאות על סט המבחן החיצוני

model	f1	rcl	Prc	wer	loss
tiny	0.221	0.257	0.199	76.0	3.05
base	0.323	0.349	0.304	65.4	4.66
small	0.367	0.394	0.348	**57.2**	1.75
medium	0.407	0.485	0.359	67.5	1.63
large-v2	**0.448**	**0.510**	**0.408**	63.4	1.31

טבלה3 תוצאות המודלים בגדלים השונים מול סט המבחן החיצוני

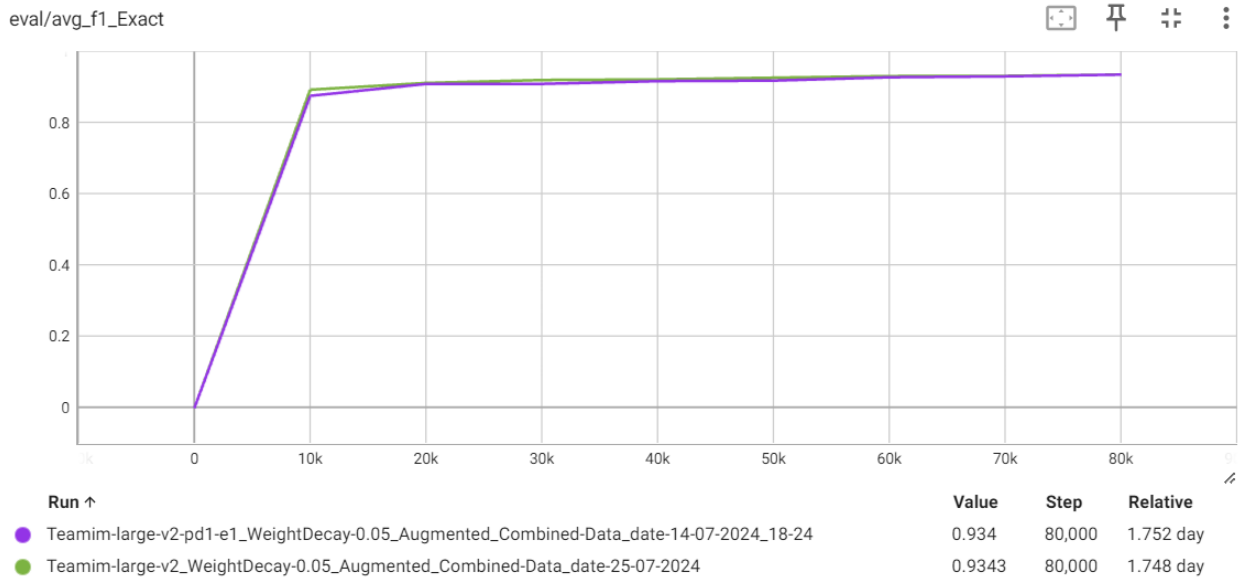
גם כאן כפי שניתן לראות התוצאות היו טובים יותר על המדד F1 Exact במטריקה שהגדרנו.

8.4. אימון על מודל שאומן מראש על דאטה רב בעברית לעומת המודל שלא אומן כך

במהלך העבודה על הפרויקט שוחרר מודל מבוסס whisper v2 שאומן על כמות גדולה מאוד של דאטה בעברית. החלטנו לבצע השוואה בין מודל זה למודל המקורי, אימנו את שניהם על אותו מספר פעמים.

8.4.1. תוצאות על סט הוולידציה

להלן תוצאות הוולידציה:



איור 7 גרף השוואה בין מודל שאומן על דאטה רב עברית לעומת אחד שלא אומן כך

כאשר מודל הבסיס הוא whisper large v2 ועבור המודל שאומן מראש השתמשנו במודל הפתוח שנחשב הכי טוב בעברית נכון לזמן כתיבת הדו"ח: ivrit-ai/whisper-v2-pd1-e1.

כפי שניתן לראות התוצאה זהה על סט הוולידציה (0.934) כלומר המודל שאומן מראש על דאטה בעברית לא נתן תוצאות טובות יותר על סט **הוולידציה**.

8.4.2. תוצאות על סט המבחן

להלן טבלה של התוצאות על סט המבחן החיצוני

model	aug	data	nikud	rnd	f1	rcl	prc	wer	loss
large-v2	✓	Comb	×	×	0.448	0.510	0.408	63.4	1.31
v2-pd1-e1	✓	Comb	×	×	**0.517**	**0.615**	**0.454**	**58.6**	1.44

טבלה*4* תוצאות המבחן - מודל רגיל לעומת מודל שגם אומן על דאטה רב עברית

ניתן לשים לב כי עבור סט המבחן קיבלנו שיפור יחסית משמעותי, בעיקר עבור ערך ה-recall כלומר המודל נותן יותר ערכים נכונים.

כלומר על אף שהמודל לא הראה שיפור על דאטה שדומה מאוד לדאטה באימון, הוא כן הראה שיפור משמעותי על דאטה אחר.

המסקנה שלנו היא שהמודל הצליח להכליל בצורה יותר טובה עבור דוברי עברית, ולכן הניב תוצאות יותר טובות.

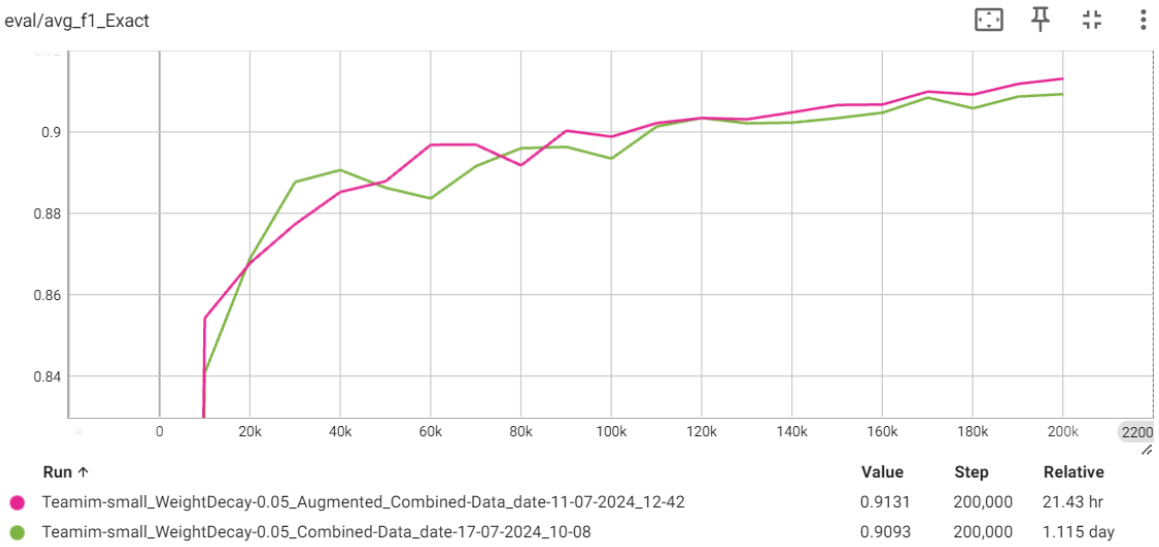
אנו ממליצים לנסות בפרויקטי המשך לבצע אימון המשלב דאטה ללא טעמים.

8.5. אימון עם או בלי אוגמנטציות:

ביצענו השוואה בין מודלים שאומנו עם אוגמנטציות למודלים שאומנו ללא אוגמנטציות. המודלים עליהם ביצענו את ההשוואה הם small (השלישי בגודלו), ו-tiny (הראשון).

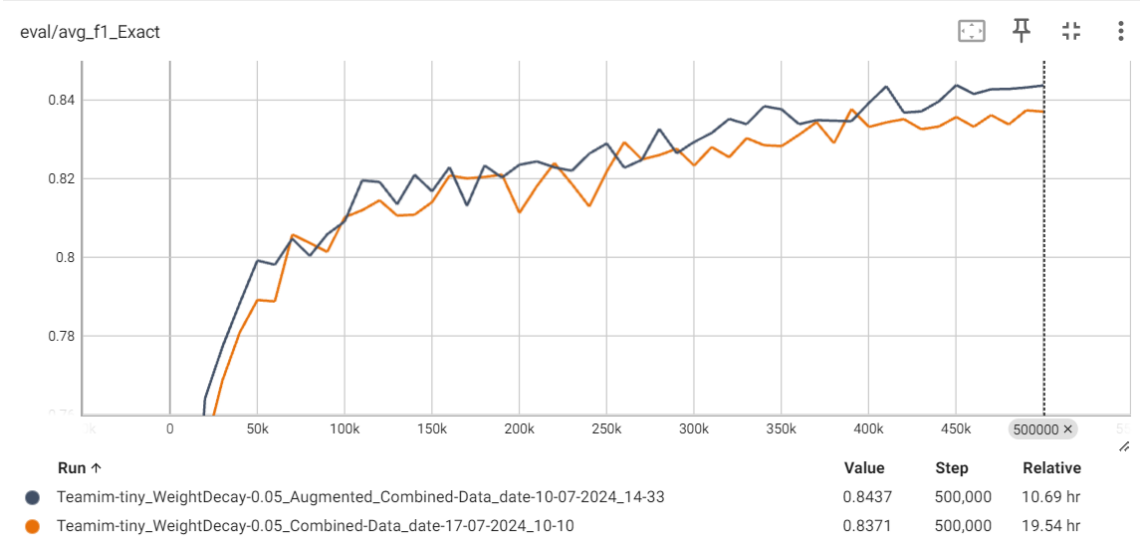
8.5.1 תוצאות על סט הוולידציה

להלן תוצאות הוולידציה על אימון מודל small:



איור 8 גרף השוואה בין מודלי *small* שאומנו עם אוגמנטציות למודלים שאומנו ללא אוגמנטציות

להלן תוצאות הוולידציה על אימון מודל tiny:



איור 9 גרף השוואה בין מודלי *tiny* שאומנו עם אוגמנטציות למודלים שאומנו ללא אוגמנטציות

ניתן לשים לב כי ישנו שיפור מסוים על סט הוולידציה כאשר משתמשים באוגמנטציות.

8.5.2. תוצאות על סט המבחן

להלן טבלה של התוצאות על סט המבחן החיצוני

model	aug	data	f1	rcl	prc	wer	loss
tiny	✓	Comb	**0.221**	**0.257**	**0.199**	**76.0**	3.05
tiny	✗	Comb	0.195	0.226	0.177	80.5	2.65
small	✓	Comb	**0.367**	**0.394**	**0.348**	**57.2**	1.75
small	✗	Comb	0.303	0.348	0.275	67.8	1.95

טבלה5 תוצאות על סט המבחן - בין מודלים שאומנו עם אוגמנטציות למודלים שאומנו ללא אוגמנטציות

כלומר יש שיפור גם עבור סט המבחן, כלומר יש **שיפור גם עבור סגנונות שונים** ממה שהמודל ראה.

בנוסף ניתן לראות כי השיפור על סט המבחן הוא גם על מדד ה-Precision וגם על מדד ה-Recall.

8.6. תוצאת האימון עם אוגמנטציה של חלקים אקראיים

כאמור משום שהדאטה שלנו מורכב מחלקים קצרים, יכולנו לבצע אוגמנטציה חיבור מספר חלקים ממקומות שונים ובכך לקבל כמות דאטה גדולה.

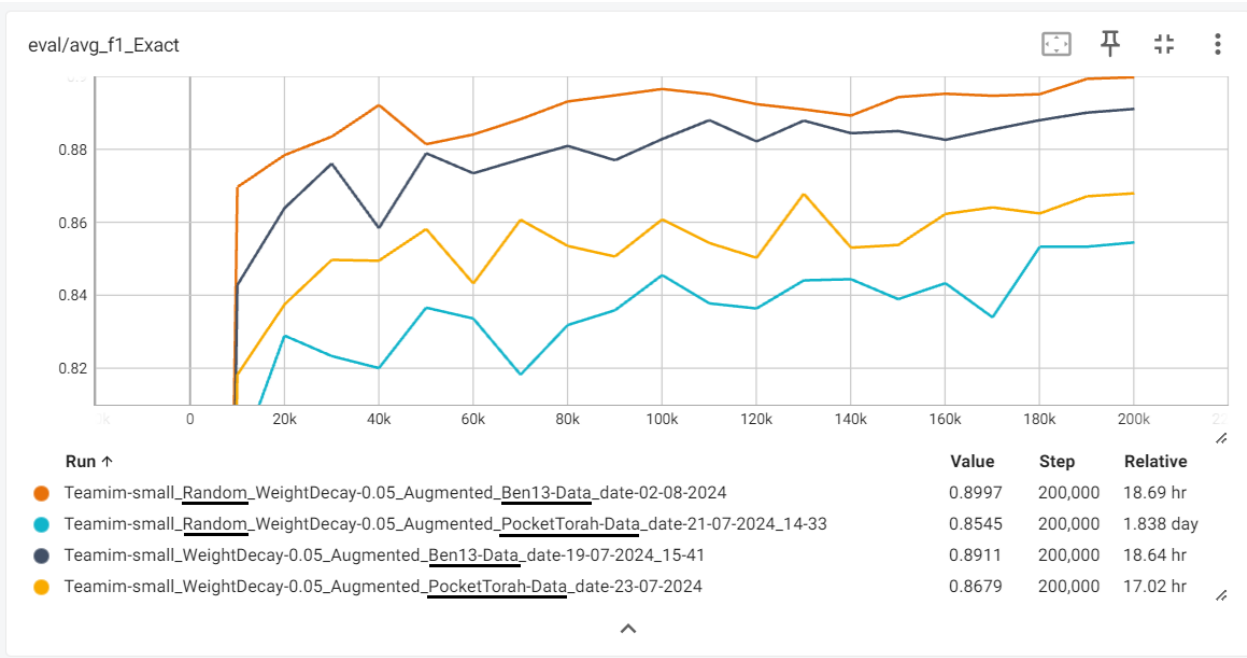
נבחין בין שני הסטים שיש לנו

PocketTorah – הדאטה מחולק למילים, אך ישנם רצפים של טעמים שאינם הגיוניים.

בן13 – הדאטה מחולק לפי רצפים בסייסים של טעמים, כלומר כל חלק מכיל מספר מילים קטן שמנוגן בצורה מסוימת. כלומר מתבסס על הרצפים הנכונים.

8.6.1. תוצאות על סט הוולידציה

להלן תוצאות הוולידציה:



איור 10 תוצאת האימון עם אוגמנטציה של חלקים אקראיים

כפי שניתן לשים לב האוגמנטציה הזו שיפרה את הביצועים על הדאטה של בן13 ופגעה בביצועים על הדאטה של PocketTorah, לדעתנו תוצאה זו מאוד הגיונית בהתחשב בכך שהדאטה של בן13 מחולק לרצפים הגיוניים של טעמים.

8.6.2. תוצאות על סט המבחן

להלן טבלה של התוצאות על סט המבחן החיצוני:

model	data	rnd	f1	rcl	prc	wer	loss
small	**Ben13**	×	**0.356**	**0.435**	**0.308**	**78.8**	2.92
small	**Ben13**	✓	0.213	0.291	0.172	84.9	4.72
small	**PT**	✓	0.045	0.056	0.039	90.4	6.77
small	**PT**	×	**0.082**	**0.097**	**0.074**	**89.4**	3.65

טבלה6 תוצאות מבחן - סטים שאומנו על דאטה אקראי

ניתן לראות כי על סט המבחן דווקא קיבלנו תוצאה פחות טובה כאשר השתמשנו בשיטה זו, אפילו כאשר עשינו זאת עם בן13

8.7. תוצאות האימון על סטים שונים של דאטה.

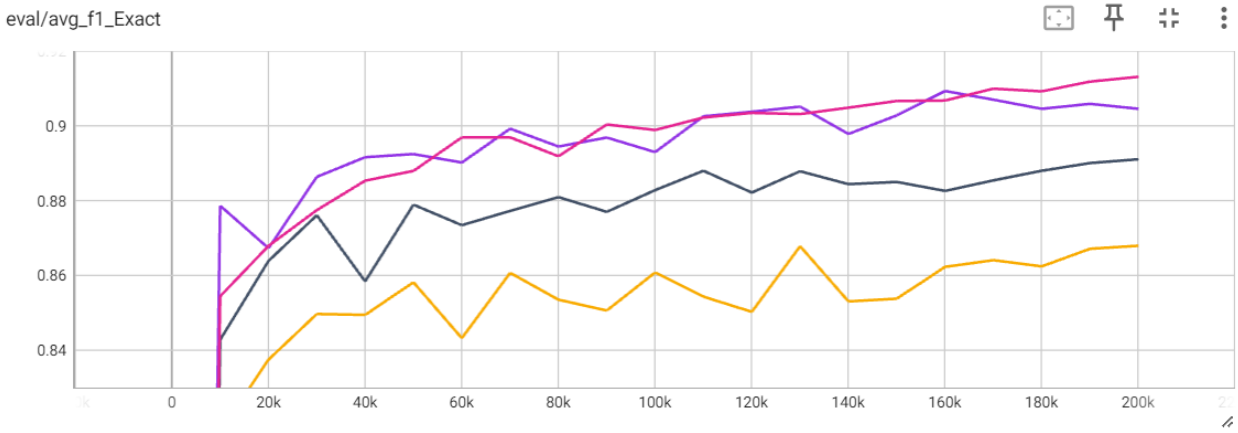
בנוסף ביצענו השוואה בין אימונים על מאגרי הנתונים השונים:

1) בן13 עם כל הנוסחים
2) בן13 בנוסח ירושלים
3) PocketTorah
4) כל הדאטה.

המטרה שלנו היא לבדוק את שיפור יכולת ההכללה של המודל כאשר מוסיפים דאטה, וכן למצוא איזה דאטה יותר קל ללמידה

8.7.1. תוצאות על סט הווליד‌ציה

להלן תוצאות הווליד‌ציה:



Run	Value	Step	Relative
Teamim-small_WeightDecay-0.05_Augmented_Ben13-Data_date-19-07-2024_15-41	0.8911	200,000	18.64 hr
Teamim-small_WeightDecay-0.05_Augmented_Ben13-Data_nusach-yerushalmi_date-24-07-2024	0.9045	200,000	17.18 hr
Teamim-small_WeightDecay-0.05_Augmented_Combined-Data_date-11-07-2024_12-42	0.9131	200,000	21.43 hr
Teamim-small_WeightDecay-0.05_Augmented_PocketTorah-Data_date-23-07-2024	0.8679	200,000	17.02 hr

איור 11 תוצאות האימון על סטים שונים של דאטה

ניתן לראות כי על סט הווליד‌ציה יש יתרון רב כמו כן על הדאטה של PocketTorah יש שגיאה גבוהה יותר מאשר על בן13 (כאשר כל אחד נבחן לפי סט ולידציה משלו). ההנחה שלנו היא כי זה נובע מכך שהדאטה של PocketTorah קשה יותר לחיזוי מכיוון שהבדלים בו בין הטעמים השונים יותר קטנים.

8.7.2. תוצאות על סט המבחן

להלן טבלה של התוצאות על סט המבחן החיצוני:

model	data	f1	rcl	prc	wer	loss
small	Comb	**0.367**	0.394	**0.348**	**57.2**	1.75
small	Ben13	0.356	**0.435**	0.308	78.8	2.92
small	PT	0.082	0.097	0.074	89.4	3.65
small	Yer	0.260	0.261	0.258	79.1	3.47

טבלה7 תוצאות מבחן - מודלים שאומנו על סטים שונים של דאטה

ניתן לראות כי על אף שהטסט שלנו הוא על נוסח ירושלמי, האימון על נוסח ירושלמי בלבד מתוך בן13, הניב תוצאה פחות טובה מאשר האימון על כל הנוסחים שנמצאים בדאטה מבן13, בנוסף נראה כי אפילו האימון המשולב עם PocketTorah נתן תוצאות יותר טובות, וזאת על אף שכחצי מהצעדים הוחלפו באימון על PocketTorah שהוא דאטה עם נוסח שונה לחלוטין).

תוצאה זו ממחישה לנו שהוספה של דאטה יניב **שיפור ביכולות המודל** גם עבור נוסחים **שונים לחלוטין**.

בנוסף ניתן לשים לב כי האימון על בן13 הניב תוצאות recall טובות למדי בעוד האימון המשולב העלה את הprecision.

תוצאה זו פותחת פתח למחקר עתידי עבור שיטות אימון וסידור הדאטה לשיפור הן של recall והן של precision.

9. סיכום

במסגרת פרויקט זה פותח, לראשונה, מודל "דיבור לטקסט" המותאם לזיהוי קריאה בתורה עם טעמי המקרא. המודל, המבוסס על טכנולוגיות מתקדמות של למידת מכונה ורשתות נוירונים, הראה יכולת מבטיחה בזיהוי הטעמים, תוך התמודדות עם מורכבותם והשונות במסורות הקריאה השונות.

במהלך הפרויקט התמודדנו עם מספר אתגרים ייחודיים, ביניהם מורכבות מערכת הטעמים, השונות במסורות הקריאה, והקושי בהשגת דאטה מתאים. כדי להתגבר על אתגרים אלו, פיתחנו מספר פתרונות חדשניים, ביניהם שימוש במודל Whisper של OpenAI, יישום אוגמנטציות ספציפיות לשמע, ושילוב של מקורות דאטה מגוונים.

המודל שהושג, לאחר אימון על מאגר נתונים משולב ושימוש באוגמנטציות, השיג דיוק של עד 51.7% במדד F1 על סט מבחן חיצוני. תוצאה זו מדגימה את הפוטנציאל הרב של המודל לשמש ככלי יעיל לזיהוי טעמי המקרא.

השוואה בין מודלים שונים הראתה כי גודל מאגר הנתונים ואיכותו מהווים גורמים מכריעים בהשגת דיוק גבוה. בנוסף, נמצא כי שימוש במודלים שאומנו מראש על דאטה רב בעברית משפר משמעותית את יכולת ההכללה של המודל על דוברים וסגנונות קריאה שונים.

למודל שפותח יש פוטנציאל רב לשמש ככלי עזר משמעותי עבור לימוד ותרגול קריאה בתורה, הן עבור לומדים עצמאיים והן עבור אנשים עם מוגבלויות. בנוסף, המודל עשוי לסייע בחקר ההיסטוריה של הקריאה בטעמים, ובהבנת התפתחותם של נוסחי הקריאה השונים.

אנו מאמינים כי המשך פיתוח המודל והרחבתו, תוך התמקדות בהרחבת מאגר הנתונים ויישום טכניקות מתקדמות כמו LoRA ו-One-shot learning, יוביל ליצירת כלי בעל ערך רב עבור לימוד ותרגול קריאה בתורה, ויסייע לשמר ולהנגיש את המסורת העתיקה של קריאת התורה לדורות הבאים.

10. הרחבות לעתיד – להמשך פרוייקט

ניתן להמשיך לעבוד על הפרויקט הזה מכיוונים רבים, להלן כמה כיוונים שעלו לנו:

1) פיתוח מערכת לחיתוך דוגמאות לצורך אימון, פתרון מצוין עבור הדאטה כי קיימים לא מעט נתונים המתאימים לאימון אך הם עם קטעי אודיו ארוכים מדי. כפתרון לבעיה זו ניתן להשתמש באחד מהפתרונות הבאים:

   א. תוכנה חצי-אוטומטית לחיתוך קטעי אודיו – המשתמש מקבל את הטקסט ושומע קטע אודיו שנחתך עבורו מראש ואז משייך טקסט לאודיו הזה.

   ב. מערכת אוטומטית לחיתוך קטעי אודיו, המבוססת על מודל שאומן כבר – ניתן להשתמש במודל שכבר אומן כדי ובכך להתאים את הדאטה מתוך הטקסט האמיתי, לפי הטקסט עם השגיאות.

   ג. פתרון היברידי המשלב חיתוך אוטומטי עם בקרה ידנית. (שילוב של שתי השיטות כדי להימנע מתקלות)

   נדגיש כי לשם כך נדרשת גם התאמת מבנה הנתונים של המודל לקבלת דאטה בתצורה של הדאטה החתוך. (מבנה הנתונים של הדאטה של המודל נבנה בצורה שיתאים למבנה שבו שמרנו כל דאטה שאספנו, ולא לצורה כללית)

2) התאמת מבנה הנתונים של המודל לקבלת מאגרי נתונים סטנדרטיים. פתרון לבעיה של התאמה מחודשת לכל דאטה.

3) אימון גם על נתונים ללא טעמים – עשוי לשפר את ההבנה של המודל, מה תפקיד הטעמים.

   ממה שמצאנו גם דאטה כזה של הקראה מהתורה ללא טעמים קיים בחלקים יותר ארוכים ועל כן יש לחלק אותו.

4) שימוש בטכניקת LoRA לאימון מהיר על מספר מצומצם של דוגמאות.

5) יישום גישת למידה מדוגמה בודדת(One-shot learning) בזמן החיזוי.

6) הוספת כניסה של טקסט ללא טעמים

7) התאמה של כל הקוד למודל whisper large v3 (המודל הזה יכול לשים לב יותר טוב לדקויות, כלומר עשויה להיות תוצאה טובה יותר עבור דאטה שהוא עם ניגון ולא רק מילים)

11.רשימת מקורות

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision." arXiv preprint arXiv:2212.04356 (2022). [Online]. Available: https://arxiv.org/abs/2212.04356

[2] I. Jordal, "audiomentations: A Python library for audio data augmentation," GitHub repository, 2019. [Online]. Available: https://github.com/iver56/audiomentations.

[3] Ivrit-AI, "whisper-v2-pd1-e1: A fine-tuned Whisper model for Hebrew speech recognition," Hugging Face, 2023. [Online]. Available: https://huggingface.co/ivrit-ai/whisper-v2-pd1-e1.

[4] "Signal and Image Processing Lab.," 2018. [Online]. Available: http://sipl.technion.ac.il/.

[5] R. Neiss, "PocketTorah: A mobile app for learning Torah with audio and text," GitHub repository, 2015. [Online]. Available: https://github.com/rneiss/PocketTorah.

[6] Sefaria, "Sefaria: A free living library of Jewish texts," GitHub repository. [Online]. Available: https://github.com/Sefaria.

[7] Wikisource, "Hebrew Wikisource: A free library of source texts," 2024. [Online]. Available: https://he.wikisource.org/.

[8] Ben13, A website for learning Torah reading for Bar Mitzvah [Online]. Available: https://www.ben13.co.il/.

[9] K. Anil, C. Bianchi, F. Dalle, and others, "Learning models of human behavior for interactive decision making," arXiv, 2023. [Online]. Available: https://arxiv.org/abs/2308.11596.

[10] A. Vaswani, N. Shazeer, N. Parmar, and others, "Attention is all you need," arXiv, 2017. [Online]. Available: https://arxiv.org/abs/1706.03762.

[11] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv, 2015. [Online]. Available: https://arxiv.org/abs/1508.01211.