הפקולטה להנדסה

המעבדה לעיבוד אותות

Geometry-Aware Multi-Task Learning for Binaural Audio Generation from Video

במעבדה לעיבוד אותות

תמר מכאני

ליעוז ברמן

פרויקט שנה ד' לקראת תואר ראשון בהנדסה

מנחה: יוחאי ימיני

מנחה אקדמי: פרופ' שרון גנות

נובמבר 2023

תוכן עניינים

מונחים והגדרות: 3
הקדמה: 3
סקירה: 4
Backbone- המרה ממונו לסטריאו 4
Spatial coherence - קוהרנטיות בין הסאונד לווידאו 5
Geometric consistency - עקביות גיאומטרית 5
אתחול הפרמטרים : 6
Dataset 6
מבנה הרשתות 7
Visual Net 7
AudioNet 7
Fusion Net 7
Ap net 7
חלוקת ה- dataset 7
מבנה מסלול הTEST 7
קשיים מרכזיים בדרך 8
תוצאות האימון 9
מסקנות 10
הצעות לשיפור 10
נספחים 10

הפרויקט מבוסס על המאמר-

"Geometry-Aware Multi-Task Learning for Binaural Audio Generation from Video"

מאת – **Rishabh Garg, Ruohan Gao, Kristen Grauman**.

לכן, יש לציין כי בהסברים התיאורטיים אנחנו מסתמכים על המאמר.

מונחים והגדרות:

- מונו (mono) - מתייחס לאות סאונד יחיד (המוקלט ע"י מיקרופון יחיד)
- סאונד בינאורי (Binaural Audio) - מתייחס לאות הנקלט ע"י שתי האוזניים שלנו.
- פריים – תמונה יחידה בתוך רצף תמונות המרכיבות וידאו.

הקדמה:

לראייה ולשמיעה יש חלק גדול מהחוויה התפיסתית האנושית, שילוב הראיה והשמיעה יוצר תמונת מידע מלאה במרחב. הפרשי הצלילים המגיעים לכל אוזן וצורת האוזן החיצונית מספקים אפקטים מרחביים המשפיעים על תפיסת המרחב שלנו.

דבר נוסף המשפיע על תפיסת המרחב שלנו הם ההדהודים וההשתקפויות של הקול בסביבה המתבטאים כפונקציה של האקוסטיקה בחדר - הגיאומטריה שלו, החומרים מהם עשוי, התכולה שלו וכו'. (לדוגמה, אותו הקול ישמע לנו שונה במסדרון ארוך לעומת חדר גדול או חדר מלא ברהיטים לעומת חדר ריק)

אודיו בינאורי גורם למדיה להרגיש אמיתית וסוחפת יותר. עם זאת, איסוף נתוני אודיו בינאוריים הוא אתגר. כיום, אודיו בינאורי נאסף באמצעות מערך מיקרופונים או מתקן דמה מיוחד המחקה את האוזניים והראש האנושיים. לכן תהליך האיסוף פחות נגיש ויקר יותר בהשוואה לאודיו חד ערוצי סטנדרטי שנלכד בקלות מהמכשירים הניידים היום בכל מקום.

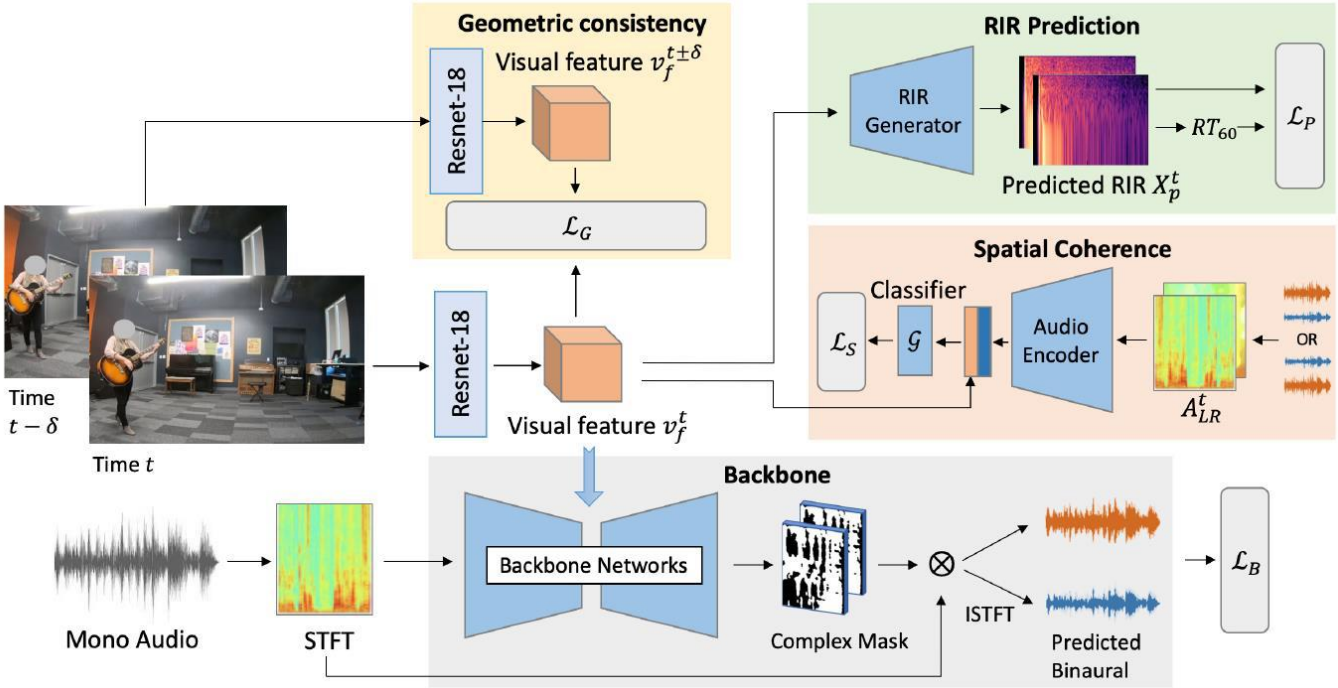
לאחרונה נעשו מחקרים במטרה לקבל אודיו בינאורי מתוך סרטוני וידאו ע"י שימוש במאפיינים חזותיים כלליים של הוידאו.

בשונה מהמאמרים הקודמים שם הלמידה התבססה על מאפיינים חזותיים כלליים, במאמר שעליו אנחנו מתבססים הלמידה נעשית על מידול של הסאונד במרחב(כלומר, במאמר עוסקים בפירוק התופעה המרחבית למרכיבים שלה והתבסוסת על המאפיינים החזותיים ללמוד מאפיינים אלה) ומתוך כך לקבל סאונד בינאורי מתוך הסאונד חד-ערוצי.

במאמר יש שימוש בגישת ריבוי משימות חדשנית להמרת סאונד במונו של סרטון לסאונד בנארי ע"י לימוד ייצוגים אודיו-ויזואליים שממנפים את המאפיינים הגיאומטריים של הסביבה ואת המידע המרחבי מהסרטונים.

סקירה:

על מנת שנוכל לייצא את האודיו בינאורי, בהינתן לנו אות אודיו במונו (אודיו בערוץ יחיד), אנחנו ניאלץ להתמודד עם בעיות שונות. נציג כל אחת מהבעיות ואת אופן ההתמודדות מול אותן הבעיות.
נציג תמונה להמחשת הבעיות ופתרונן:



כחלק מהבעיה אנחנו יודעים שניתן להיתקל ביותר מדובר אחד ולכן נרצה לזהות את מקור הסאונד. בנוסף, אנחנו רוצים לדעת את מיקומו של הדובר ביחס למיקרופון (לא רק ב 2 ממדים אלא ב 3 ממדים).
במאמר שלפיו עבדנו, פתרו את הבעיות הללו באמצעות 4 רשתות שונות:

- Backbone - המרה של הסאונד ממונו (mono) לסטריאו ע"י ניתוח מאפיינים כלליים של המרחב.
- Spatial coherence - קוהרנטיות בין הסאונד הבינאורי לבין הווידאו. כלומר, נרצה לדעת שאם אנחנו שומעים דובר בצד מסוים של החדר הוא אכן שם.
- Geometric consistency (עקביות גיאומטרית) – בהסתמך על ההנחה שיש עקביות של המאפיינים המרחביים של הווידאו עקביים נרצה לדעת שהיא אכן מתקיימת.
- RIR Pediction - תגובת התדר של החדר (אימון בשימוש מאפיין זה חלה מתקיים בdataset שנבנה מתוך סימולציה בו אנחנו יודעים נתונים רבים על החדר) .

בסימולציה שלנו אימנו את המודל לעבוד על dataset בשם FairPlay המורכב מסרטונים אמיתיים ולכן אין לנו נתונים על מנת להמציא RIR. כיוון שכך, התרכזנו בשלוש הרשתות הראשונות. את יישום הפתרונות מימשנו באמצעות רשתות נוירונים.
פונקציית ה Loss משמשת כמדידה עד כמה המודל יוכל לחזות את התוצאה הצפויה, ולכן, המטרה הסופית שלנו היא מזעור של פונקציית ה Loss.

נפרט על פונקציית ה Loss של כל אחת מהרשתות שלנו, על פי המאמר:

Backbone- המרה ממונו לסטריאו

תחילה בנינו את הרשת המרכזית, הBackbone.
הרשת הזאת מסתמכת על מאמר קודם של הכותבים בשם:
"2.5d visual sound"
נסביר על הרעיון מאחורי הרשת:
בזמן האימון - אנחנו נמיר את הסאונד הבינאורי למונו ע"י מיצוע של האותות המרחביים - $a_M^t = \frac{a_L^t + a_R^t}{2}$

ובמקום ללמוד ישירות את כל אחד מן האותות אנחנו נלמד את השוני בין האותות . בצורה כזאת אנחנו נלמד מסיכה $\mathrm{M_D^t}$ שתעניק לנו את ספקטוגרמת השוני באופן הבא:

$$A_{D(pred)}^t = M_D^t\, A_M^t$$

כאשר A_M^t היא ספקטוגרמת האות במונו.

(הסיבה לחישוב ספקטוגרמת השוני היא שהיא מדויקת יותר במקרים העדינים ומונעת מקרים של חישוב אותו המוצא לשני הערוצים)

את ספקטוגרמת השוני האמתית נוכל למצוא ע"י התמרת STFT של $a_L^t - a_R^t$.

בנוסף, נחזה את המסכות M_R^t ו- $\mathrm{M_L^t}$ על מנת לשערך את הספקטוגרמות של האות.

פונקציית ה Loss:

$$\mathcal{L}_B = \left\|A_D^t - A_{D(pred)}^t\right\|_2^2 + \left\{ \left\|A_L^t - A_{L(pred)}^t\right\|_2^2 + \left\|A_R^t - A_{R(pred)}^t\right\|_2^2 \right\}$$

על מנת להימנע ממקרה של חיזוי קל בו אנחנו משערכים את אותה ספקטוגרמה עבור שני הערוצים אנחנו נשתמש בפונקציית הloss של הערוצים בספקטוגרמת ההפרש יחד עם סאונד המונו(בערוץ אחד נחבר ובערוץ השני נחסר)

<u>Spatial coherence - קוהרנטיות בין הסאונד לווידאו</u>

לאחר מכן התחשבנו ברשת Spatial coherence, ווידאנו כי הסאונד נקבע בצורה מרחבית בהתאם לווידאו.

בזמן האימון - אנחנו משתמשים במסווג g המשלב את הסאונד הבינאורי ביחד עם הפריימים המתאימים בווידאו. לאחר מכן אנחנו הופכים בחלק מן הדגימות את הסאונד הבינאורי כך שנקבל:

$$\mathrm{A_L^R} = \{\mathrm{A_R^t}, \mathrm{A_L^t}\}$$

אנחנו מחשבים את פונקציית ה- BCE Loss (Binary Cross Entropy) עבור הסיווג המתקבל יחד עם אינדיקטור ההפיכה $\hat{c}$. כלומר, פונקציית ה LOSS היא:

$$\mathcal{L}_G = BCE\left(g\left(A_{LR}^t, v_f^t\right), \hat{c}\right)$$

כאשר v_f^t מסמן את הפריים של הווידאו בזמן t .

<u>Geometric consistency - עקביות גיאומטרית</u>

לאחר מכן יצרנו את הרשת של ה Geometric consistency. אנחנו רוצם לדעת שהווידאו (המורכב מרצף של תמונות) מכיל בתוכו עקביות גיאומטרית. כלומר, לא יכול להיות שאדם נע מצד אחד של הווידאו לצד השני במעבר בין שני פריימים.

לכן, אנחנו יכולים להסיק כי עבור - $-1 \le \delta \le 1$ ועבור $\alpha > 0$ אזי מתקיים כי $\left|v_f^t - v_f^{t\pm\delta}\right| < \alpha$. מכאן שפונקציית ה Loss בסעיף זה היא:

$$\mathcal{L}_S = \max\left(\|\mathrm{v_f^t} - \mathrm{v_f^{t\pm\delta}}\| - \alpha, 0\right)$$

כאשר מאפיין זה מבטיח לנו בעצם שהגיאומטריה של הוידאו בכלל ושל החדר בפרט אינה משתנה בצורה לא עקבית.

אם כן, פונקציית ה Loss הכוללת שלנו היא:

$$\mathcal{L} = \lambda_B \mathcal{L}_B + \lambda_S \mathcal{L}_S + \lambda_G \mathcal{L}_G$$

כאשר λ הם המשקלים השונים שיש לכל תכונה.

בזמן הריצה אנחנו זקוקים אך ורק לסאונד מונו ולווידאו הנלווה אליו.

אתחול הפרמטרים :

אתחלנו את הפרמטרים כמובא במאמר,
רשת הBackbone מבוססת על הרשתות המשמשות במאמר 2.5D Visual Sound.

רשת האודיו מורכבת מארכיטקטורה מסוג U-Net- רשת המורכבת משני חלקים - חלק מצמצם וחלק מרחיב.
החלק המצמצם בנוי מ5 שכבות קונבולוציה לדגימה מטה, שאחריה באה פונקציית אקטיבציה מסוג RELU.
החלק המרחיב של הרשת דומה לחלק המצמצם אך מטרתו היא הפוכה, להחזיר את התמונה המצומצמת לגודלה המקורי. באמצעות 5 שכבות upconvolution ברשת ה- upsampling.

דגמנו מחדש את כל האודיו ל-16kHz וכדי לאמן את ה- Backbone.
השתמשנו ב 0.63 שניות של קליפ לכל 10 שניות של אודיו מתאים.
הפריימים החזותיים נחתכים באקראי ל-448 × 224.
לצורך בדיקה, אנו משתמשים בחלון הזזה של 0.1 שניות כדי לחשב את האודיו הבינאורלי עבור כל השיטות.

אנו משתמשים בBatch size של 64.
ה-Learning rate הראשוני עבור רשתות האודיו הוא 0.001 ועבור שאר הרשתות הוא 0.0001 .
אימנו את מערך הנתונים של FAIR-Play עבור 1000 epochs.
ה-δ לבחירת המסגרת מוגדרת ל-שנייה אחת והלמדאות (λ) בשימוש נקבעות על סמך ביצועי ה- validation ל-$\lambda_G = 0.01$, $\lambda_S = 1$, $\lambda_B = 10$.

Dataset

ה dataset בו השתמשנו בעבודתינו נקרא FairPlay, הוא מתבסס על הdataset מהמאמר "2.5Visual Sound".
ה dataset משתמש בהקלטות אמיתיות של אנשים וכלי מוזיקה בחדרים קטנים.

את ה - dataset נחלק לשלושה:

1. Training Dataset: המידע שבפועל אנו משתמשים כדי לאמן את המודל ובאמצעותו מעדכנים את המשקולות של רשת הנוירונים.
2. Validation Dataset: מידע זה משמש להעריך את המודל שאמנו, הערכה זו מתבצעת לאחר כל epoch או כמות מסוימת של batch-ים. אנו משתמשים במידע זה בכדי לכוונן את ה- hyperparameter של המודל. נדגיש שהמודל רואה ומעביר את המידע הזה ברשת, אך לעולם אינו "לומד" ממנו (כלומר לא מעדכן את המשקולות). אנו משתמשים בתוצאות של הווליד‌ציה וכך יודעים לכוון את ה - hyperparameter בצורה נכונה יותר שתגרום לרשת להתכנס מהר יותר, ללמוד באופן מדויק ונכון ועוד.
3. Test Dataset: זה המידע שנכנס לרשת לאחר שהיא אומנה לחלוטין. מערך המידע הזה עוזר לנו לדעת האם המודל שלנו יפעל כראוי גם על מידע שהוא טרם "ראה".

מבנה הרשתות

נסביר על מערך הרשתות השונות בהן השתמשנו:

Visual Net

הרשת מתבסס על ארכיטקטורת resnet-18 כאשר אנחנו מסירים את שתי השכבות האחרונות (שכבת poolingושכבת ה FCL האחרונה). מטרת הרשת היא להבין מאפיינים חזותיים כללים ע"י שימוש בפריימים.

AudioNet

רשת Unet המשמשת על מנת לחזות את הספקטוגרמה של הערוצים בצורה מתאימה.

Fusion Net

הרשת משמשת על מנת ללמוד את הספקטוגרמה של כל אחד מהערוצים על מנת לוודא שיש הפרדה ולא נלמד את אותה ספקטוגרמה לשני הערוצים.

Ap net

רשת המשמשת על מנת לדאוג להפרדה של מקורות שונים ובינוארליזציה של האות.

חלוקת ה - dataset

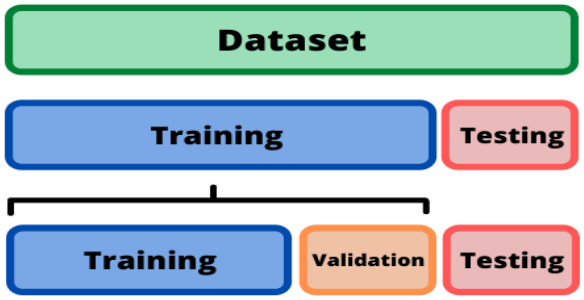
יחס חלוקת מערך מידע זה תלוי בעיקר בשני דברים. ראשית, גודל מערך המידע הכולל, שנית, המודל בפועל שנרצה לאמן, יש מודלים שזקוקים למערך מידע גדול כדי להתאמן עליהם.

מודלים עם מעט hyperparameter, קלים יותר לקביעת אותם פרמטרים כך שנוכל להקטין את גודל ה - validation dataset, אך אם למודל יש hyperparameter רבים, נרצה validation dataset גדול יותר (אם כי כדאי לשקול גם אימות צולב). אולם בסופו של דבר, קביעת חלוקת מערך הנתונים תלוי מקרה.

לרוב משתמשים בחלוקה הבאה: נפצל את מערך המידע שלנו ל- 2 - Train and Test. לאחר מכן, שומרים בצד את ה- test dataset, ובוחרים אחוז מסוים ממערך המידע שנשאר ל – training data ואת השאר ל – validation dataset

במקרה שלנו, חילקנו את ה dataset לשלושה חלקים כאשר:

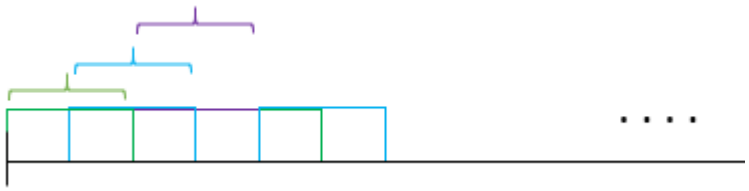
4. 70% – Train
5. 15% – Validation
6. 15% – Test



מבנה מסלול הTEST

הטסט פועל על Data של סרטונים באורך ___ שניות, בעלי קובץ אודיו מתאים אך מופרד.

על כל סרטון עברנו עם חלונות ברוחב ___ כך שכל חלון חופף חצי מהחלון הקודם, וביצענו מיצוע על כל חלון באופן הבא-

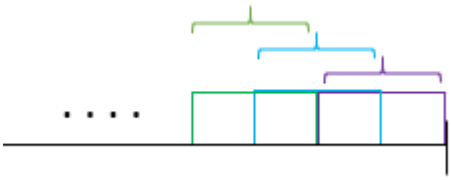


כפי שניתן להבין, על כל נקודת זמן בסרטון נעבור פעמיים מלבד חצי חלון הזמן הראשון ושלושת חלונות הזמן האחרונים. לכן, בנקודות אלה נטפל באופן פרטני.

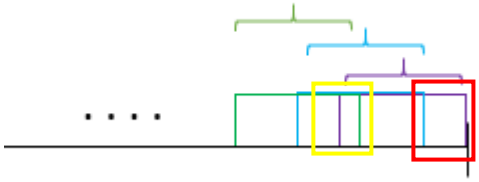
חצי חלון הזמן הראשון- יספר פעם אחת, מכיוון שהחלון הבא יכסה רק חצי מהחלון הראשון.

חלונות הזמן האחרונים- יש שני מקרים שונים שיכולים לקרות,

אורך הסרטון יתחלק ללא שארית באורך החלונות, כך שחצי החלון האחרון יספר פעם אחת, באופן הבא-



המקרה השני הוא שאורך הסרטון לא יתחלק במדויק באורך החלונות, ומקטע הזמן האחרון שישאר לנו יהיה קצר יותר מאורך חצי חלון, כמתואר בתמונה הבאה-



במצב זה נצטרך לחשב את שארית החלון האחרון (במסגרת האדומה) ולספור אותה פעם אחת, לחשב את האורך של חצי החלון פחות השארית (במסגרת הצהובה) ולספור אותו שלוש פעמים. את כל שאר המקטעים נספור כרגיל – פעמיים.

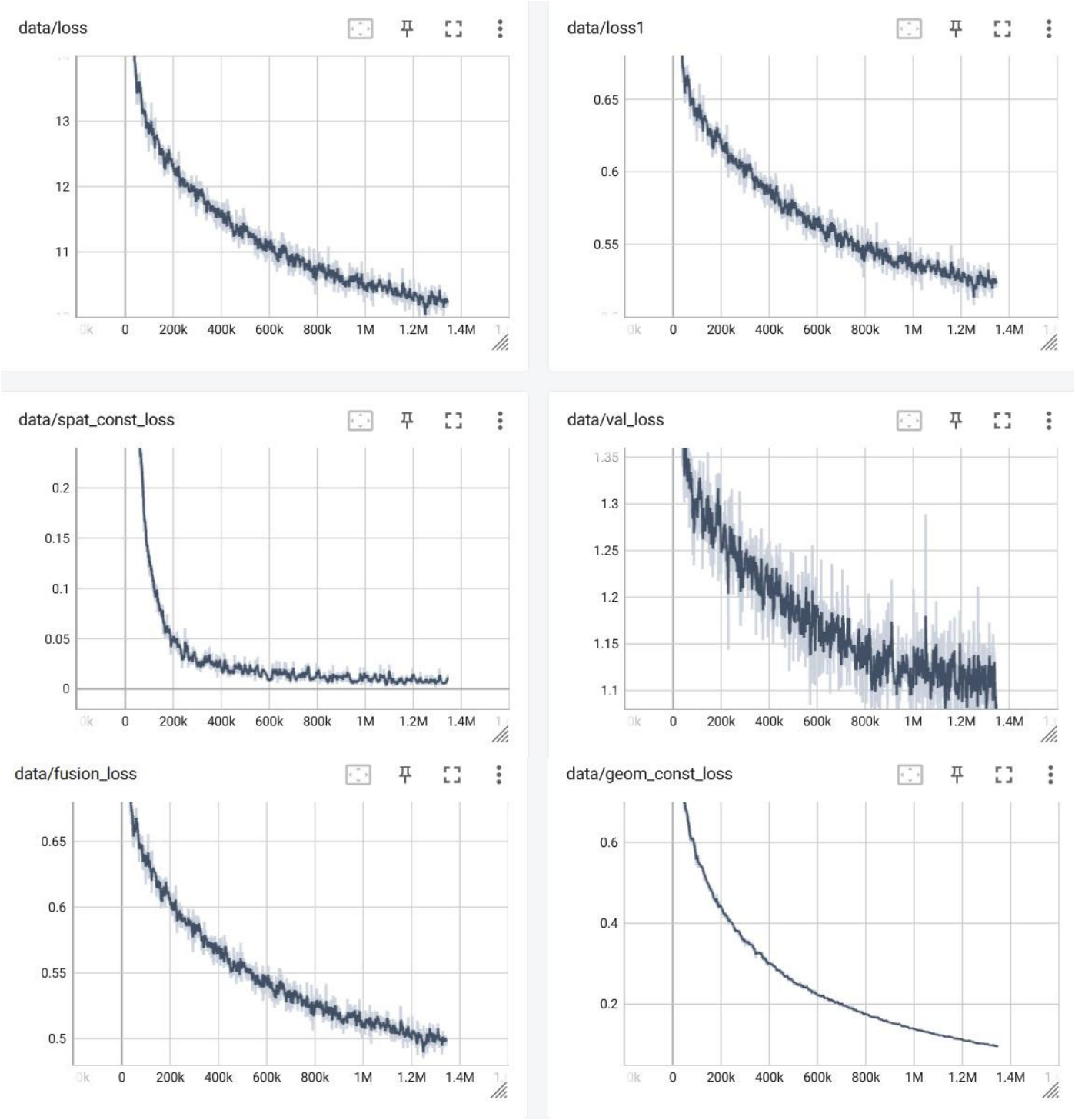
קשיים מרכזיים בדרך

לאחר שלמדנו את התיאוריה מהמאמר והתעמקנו בחומר הקורס Deep Learning, נתקלנו בקושי של להביא את התיאוריה לפרקטיקה.
לקח לנו זמן רב להבין איך עובדות הרשתות, איך לעבוד אל בשימוש עם pytorch, איך לעבוד בצורה יעילה עם GPU, איך מאמנים רשת ואיך עובדים עם Dataset כמו שלנו.
נדרשנו ללמוד ממקורות אינטרנטיים, לקרוא מאמרים נוספים(שהמאמר שלנו מסתמך עליהם), לראות דוגמאות של מקרים דומים באתרים כמו GitHub ודומיו ולהיעזר במנחה.

דבר נוסף שהיה קשה לנו הוא עבודה מול conda ופתרון בעיות של קונפליקטים, נתקלנו במשך זמן רב בקונפליטים בין חבילות פייתון שונות וכאשר ניסינו להשתמש בconda על מנת להוריד את החבילות הרלוונטיות עבורנו ההורדה נכשלה כל פעם מחדש.
על מנת לפתור את הקושי המדובר נאלצנו לקרוא רבות באינטרנט ולנסות הרבה דרכים שונות.
לבסוף הצלחנו ע"י השוואה ידנית בין הגרסאות והדרישות של כל חבילה והתקנה ידנית של החבילות באמצעות conda.

תוצאות האימון

להלן תוצאות פונקציות ה- Loss שהוצאנו על פרמטרים שונים-



- Loss – הloss המשוקלל בהתאם למשקלים שהראינו לעיל.
- Loss1 – הLoss עבור הרשת הלומדת את המסיכה של ההפרש.
- Spat_const_loss – הLoss עבור הרשת הלומדת להתחשב במרחביות.
- Val_loss – הLoss עבור validation.
- Fusion_Loss – הloss עבור הרשת הלומדת את המסיכה של הערוצים(ימין ושמאל)
- Geometry_const_Loss – הloss עבור הרשת הלומדת עקביות של מאפיינים גיאומטריים.

מסקנות

בחנו את התוצאות שלנו ע"י ייצוא של סרטונים עם האודיו הבינאורי אותו קיבלנו כתוצאה מהרשת שלנו ושמיעת אותם הסרטונים.

ראינו(או יותר נכון, שמענו) כי הצלחנו לממש את מטרת המאמר ואכן הצלחנו להבדיל בין הערוצים ולהעניק מרחביות של השמע עבור המשתמש.

השוונו את הסרטונים שייצאנו אל מול הסרטונים של כותבי המאמר המקורי ונשמע שאין הבדל בצורת המרחביות. לכן, אפשר לסכם כי הצלחנו להפעיל את הרשת בצורה טובה מאוד.

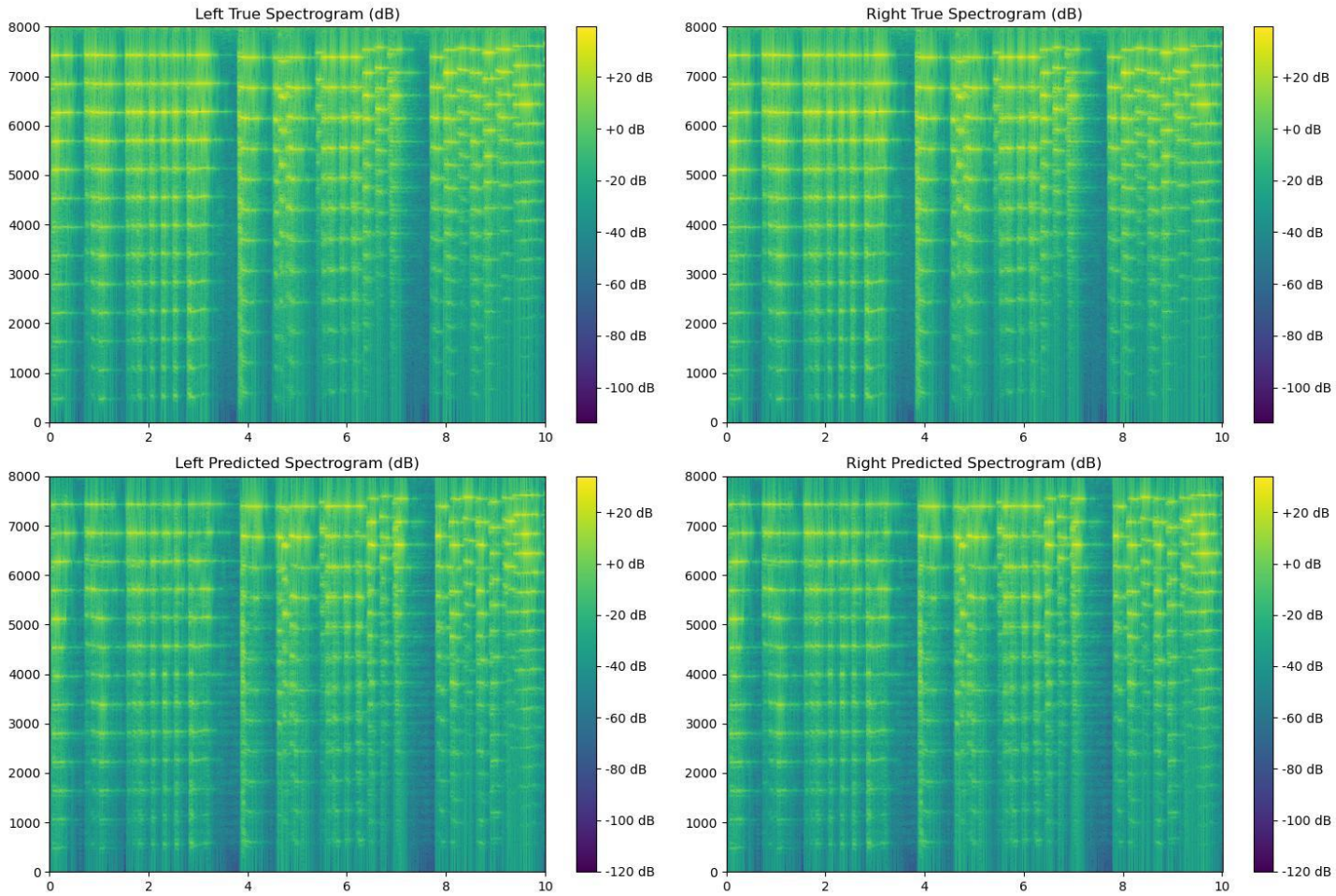
הצעות לשיפור

הסרטונים בהם השתמשנו כdataset הוקלטו בחדרים קטנים ולכן לא פשוט למוח להפריד בין הכיוונים השונים. לכן, לשיפור המחקר אולי היינו שוקלים להשתמש ב dataset שונה במרחב גדול יותר.

לשערך את תגובת התדר של החדר בעזרת האות בינאורי ובהתאם לבנות רשת שתלמד לחזות את תגובת התדר של החדר באמצעות אות מונו.

נספחים

נציג את הספקטוגרמות המתקבלות עבור אחד הסרטונים:



ניתן לראות כי התדרים בשני הערוצים די דומים (שהרי כיוון שהאודיו הוקלט בחדר קטן אותם התדרים מגיעים לשתי האוזניים בעוצמה מעט שונה).