

Common thoracic lung diseases detections

מבוא: הצגת את הפרויקט, חשיבותו, מקורו ותיאור ה- output/input

הפרויקט מתעסק בלוקליזציה של בעיות נפוצות הנמצאות בריאות וקלסיפיקציה שלהם על תמונות X-rays של חזה. המטרה היא לסייע בהפחתת הלחץ על הרופאים ושיפור איכות האבחון. מקור ה- data מאתר kaggle, כאשר הוא נאסף מהפלטפורמה VinBigData. VinBigData היא חברה ששואפת לקדם מחקר ולחקור טכנולוגיות חדשות ורלוונטיות במיוחד.

Input: תמונות X-rays של חזה (תמונה 1).

Output: מספר bounding boxes שכוללות מחלה מסוימת וה- class של המחלה (תמונה 1).

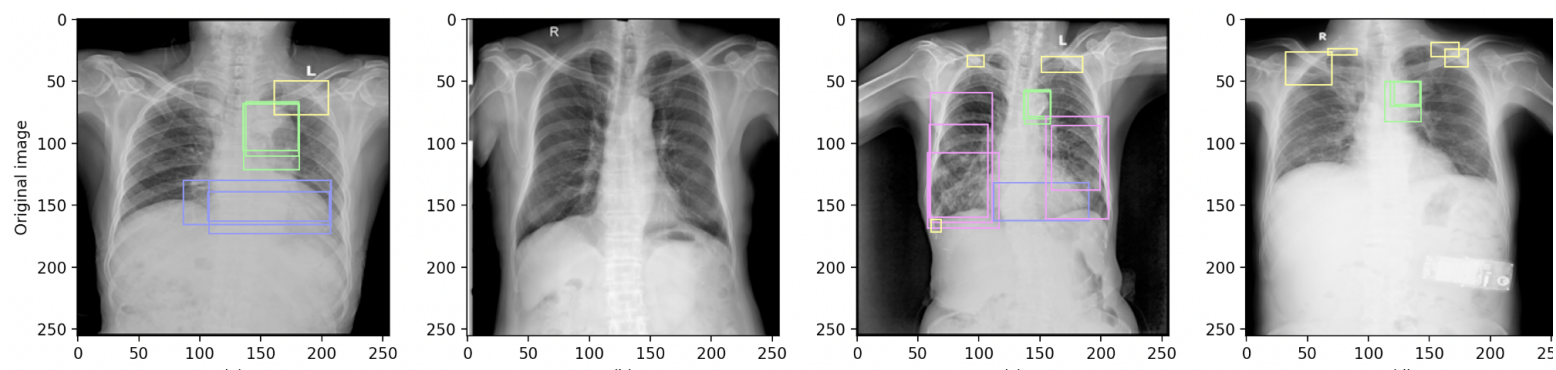
Metadata

במקור ה- dataset מורכב מ- 18,000 תמונות בפורמט DICOM. בעקבות הגודל של הקובץ (כ 200 Go) מצאתי את ה- dataset ברזולוציה נמוכה יותר (256X256) ובפורמט PNG כך שגודלו צומצם ל- 500 Mo. כל התמונות תויגו על ידי פאנל של רדיולוגים לנוכחות של 14 ממצאים רדיוגרפיים קריטיים:

1. Pneumothorax
2. Pulmonary fibrosis
3. Aortic enlargement
4. Atelectasis
5. Calcification
6. Cardiomegaly Consolidation
7. ILD
8. Infiltration
9. Lung Opacity
10. Nodule/Mass
11. Other lesion Pleural effusion
12. Pleural thickening
13. No finding

כדי להקל על סיבוכיות המודל ומהירות האימון, החלטתי להשתמש ב-4 המחלקות הגדולות ביותר: Aortic enlargement, Cardiomegaly, Pleural thickening, Pulmonary fibrosis ו- No finding (לא זוהה אובייקט בתמונה).

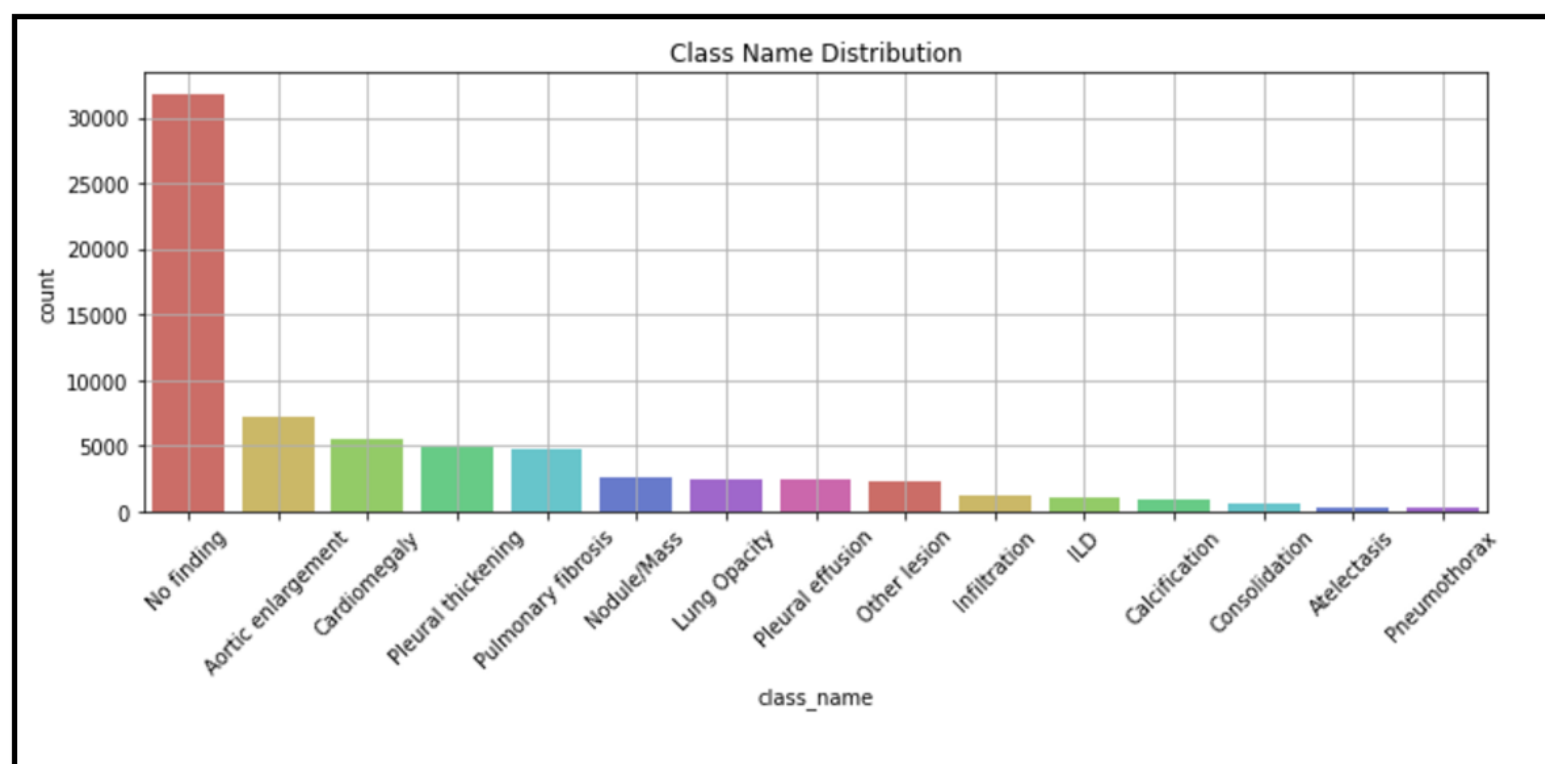
ה- metadata, הוא קובץ csv עם שורה אחת לכל אובייקט, כולל מחלקה קואורדינטות ל- bounding box. בחלק מהתמונות יש מספר עצמים כלומר יכול להיות שלאותה תמונה יש כמה ממצאים שונים. בנוסף זיהוי מחלה נעשתה על ידי שלושה רדיולוגים שונים ובמילים אחרות לכל תמונה יש לפחות 3 שורות.



תמונה 1: דוגמאות לתמונות בdataset עם bounding box שצויר בעזרת metadata. כאשר כל צבע מייצגת מחלה ספציפית.

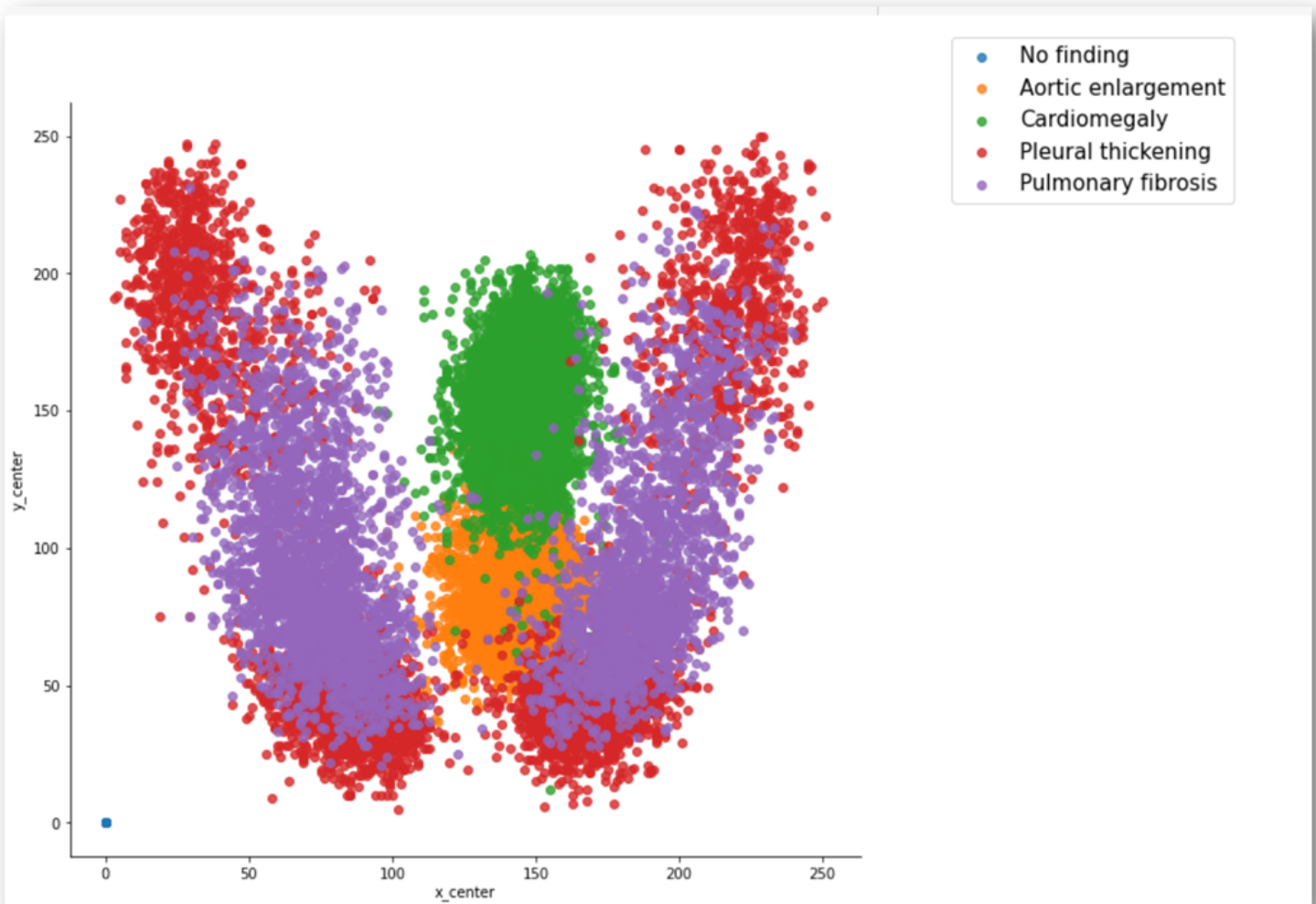
Exploratory Data Analysis

1. בוצע בדיקת פיזור מספר האובייקטים פר class. ניתן לראות בתמונה 2 שה-data לא מאוזן.



תמונה 2: פיזור מספר האובייקטים פר class

2. בדיקת פיזור מיקום של כל אנומליה על גבי התמונה. יש ממצאים שנמצאים רק באזורים מסוימים. בתמונה 3, אם נתמקד על ה-4 classes הכי שכיחים ניתן לראות שממצאים מסוימים נמצאים באזורים ספציפיים.



תמונה 3: פיזור מיקום של כל אנומליה של ה-4 classes הכי שכיחים.

3. האם יש מקרים בהם אחד רדיולוג מצא לפחות מחלה אחת בתמונה אבל רדיולוג אחר לא מצא שום מחלה באותה תמונה? לא נמצאו מצבים כאילו אבל חשוב לציין ש diagnostic של הרדיולוגים לא בהכרח זהה תמיד.

Pre-processing

1. החלפתי את הצגת ה-bounding boxes מפורמט (x_min, x_max, y_min, y_max) לפורמט (x_center, y_center, width, height) לטובת התאמה לארכיטקטורת המודל.
2. הפכתי את התמונות לgrayscale.

3. data augmentation בעזרת בהירות, סטורציה ו contrast כדי להתגבר על החוסר איזון של data.

Related Work

יש ארכיטקטורות נפוצות לפתירת בעיית object detection כגון YOLO, Faster-RCNN או SSD.



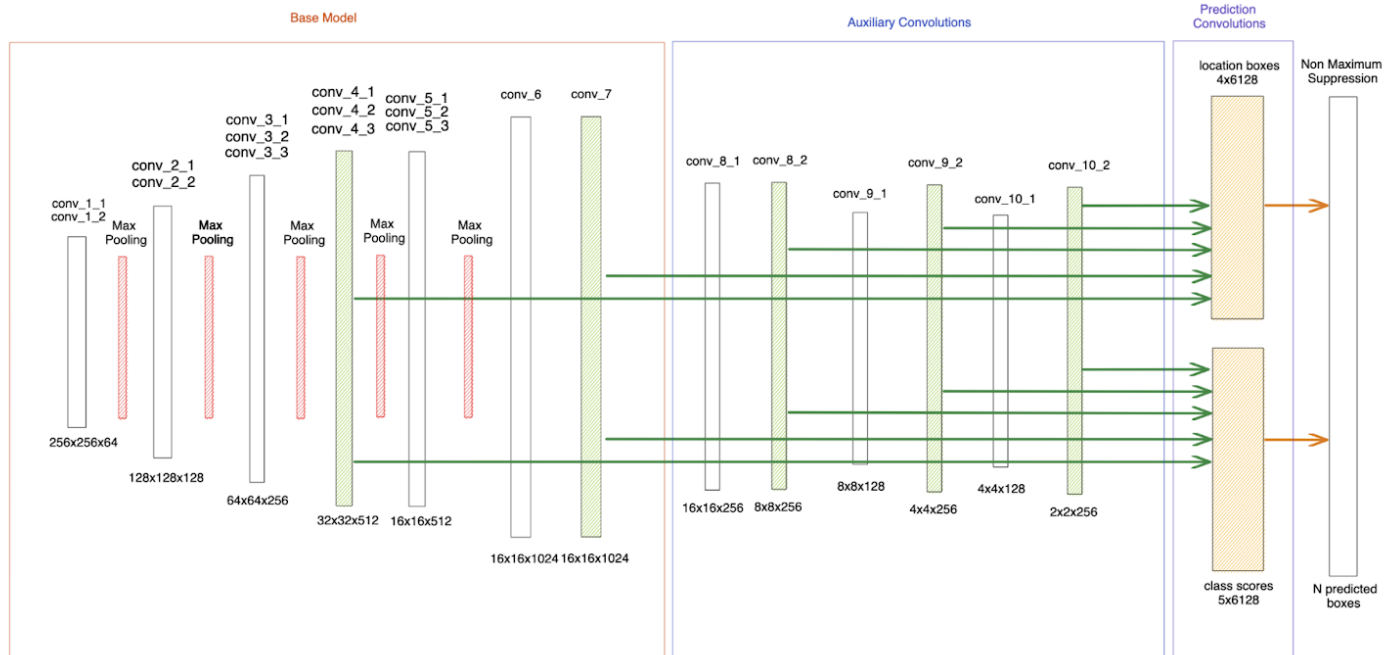
תמונה 4: גרף של accuracy vs speed של ארכיטקטורות שונות.

החלטתי להשתמש בSSD כי הוא משלב בין איכות למהירות גבוה במובן של זמן חישוב, כמו שמסבירים ב(1) הוא יותר מהיר מFaster-RCNN ויותר מדויק מYOLO. ניתן להמחיש את זה ע"י הגרף (תמונה 4).

בנוסף במאמר (2), ניסו לפתור בעיה דומה. הבעיה שהם מנסים לפתור היא זיהוי covid-19 בתמונות רנטגן חזה והשיגו תוצאות יפות עם שימוש של אלגוריתם ה SSD. הם השיגו accuracy של מעל 90% בממוצע. הם התבססו על המאמר (3) כדי לממש את האלגוריתם.

Model

המודל בנוי אך ורק מconvolution layers והוא בנוי משלושה חלקים עיקריים (תמונה 5):



תמונה 5: ארכיטקטורת המודל

1. החלק הראשון הוא מודל חסר רגיל בנוי בקונבולוציות. במאמר (3) השתמשו בvgg16 שבו יש fully connected layers והפכו את הfc layers לconvolution layers ע"י שינוי מימדים. אני ניסיתי לאמן מודל מ-0 והוא מעט שונה ממה שמוצג במאמר כמו שניתן לראות בתמונה 5.
 2. החלק השני הוא הוספה של עוד מספר convolution layers שמוסיפים לנו feature maps שהם יותר high level ושהולכים ומצטמצמים בגודלם.
 3. החלק השלישי הוא הניבוי, נבחר מספר features map מהlayers הקודמים ונלביש עליהם סט סופי של bounding boxes פוטנציאליים הנקראים priors, ונרצה לבוא לכל prior בכל feature map, את הפרש ה bounding box בינו לבין האובייקט האמיתי בתמונה ואת המחלקה של אותו אובייקט.
- בהשוואה למאמר (3), הקטנתי את כמות ה priors כדי להקל על הזמן חישוב בידיעה שזה כנראה ישפיע לרעה על הביצועים.

Feature Map Layer	Feature Map Dimensions	אחוז מגודל התמונה	Ratios	Total Number of Priors on this Feature Map
conv4_3	32x32	0.1	1, 2, 0.5 + an extra prior	4096
conv7	16x16	0.2	1, 2, 3, 0.5 0.333 + an extra prior	1536
conv8_2	8x8	0.375	1, 2, 3, 0.5 0.333 + an extra prior	384
conv9_2	4x4	0.55	1, 2, 3, 0.5 0.333 + an extra prior	96
conv10_2	2x2	0.725	1, 2, 0.5 + an extra prior	16
סה"כ	—	—	—	6128 priors

טבלה 1: הגדרת priors לכל אחד מהlayers

הpriors מוגדרים לכל אחד מהlayers ב טבלה 1. ה"extra prior" ratio מתווסף לכל feature map והוא prior ריבועי בגודל הממוצע הגיאומטרי של scalen הנוכחי והscalen של feature map הבא (האחרון יוגדר כ1 כי אין לו הבא).

output של המודל יהיה מטריצה בגודל מספר priors על מספר המחלקות ומטריצה נוספת בגודל מספר priors על 4 (מספר נקודות של bounding box). משני המטריצות האלה, נסנן את הזיהויים הרלוונטיים מתוך 6128 priors של output.

Jaccard index מודד את אחוז החפיפה בין שני boxes. כלומר נחפש את החפיפה בין כל prior לכל אובייקט בתמונה וניצור מטריצה של מספר priors על מספר האובייקטים. לכל prior נתייחס לאובייקט עם אחוז החפיפה הגדול ביותר ונסמן הprior כהתאמה חיובית אם הוא גדול מ50% והתאמה שלילית אחרת.

אחרי אימון המודל, ייתכן שיהיו עדיין יותר התאמות מהצפוי בגלל התאמות כפולות, overlapping. נרצה להשתמש בjaccard index כדי לסנן את overlapping ולשמור את bounding box עם האחוז הגבוה ביותר. זה נקרא (Non-Maximum Suppression) NMS.

פונקצית המחיר (loss function)

פונקצית המחיר תהיה בעצם שילוב של 2 פונקציות מחיר, אחת למיקום ה bounding box ואחת על קלסיפיקצית המחלקה.

1. הפונקציה הראשונה היא הפרש בין הלוקליזציה של התאמות החיוביות לאובייקטים המתאימים להם בפועל ונתעלם מה התאמות השליליות
2. הפונקציה השנייה יהיה סכום הטעויות **cross entropy** (חיוביים ושליליים) לחלק למספר החיוביים.

נשים לב שמספר השליליים (אי מציאת אובייקט) יהיה מן הסתם הרבה יותר גדול מהחיוביים מה שיווצר הטעיה. כדי למנוע את זה נקח רק חלק מהאי התאמות ונשתמש בחלק עם הלוס היותר גדול, זה נקרא Hard Negative Mining. מספר השליליים שנשמור יהיה מוגדר ככפולה של מספר החיוביים (positive/negative ratio).

פונקצית המחיר הסופית תהיה סכום הפונקציות כפול מקדם אלפא שנגדיר 1 כמו במאמר (3)

$$L_{loc} = \frac{1}{n_{positives}} \left(\sum_{positives} Smooth L_1 Loss \right)$$

$$L_{conf} = \frac{1}{n_{positives}} \left(\sum_{positives} CE Loss + \sum_{hard negatives} CE Loss \right)$$

$$L = L_{conf} + \alpha \cdot L_{loc}$$

Analysis

שיטת מדידה

כדי למדוד ביצועי המודל, אשתמש במדדי AP וmAP

1. AP (Average Precision) הוא דרך לסכם את precision-recall curve לערך יחיד המייצג את הממוצע של כל precisions.
2. mAP (Mean Average Precision) הוא הממוצע של Average Precision עבור כל מחלקה.

ניסויים וביצועים

פרמטרים ותוצאות שהתקבלו של כל ה5 ניסויים מופיעים בטבלה 2.

1. Basic Model:

ראשית, הרצתי את המודל עם הפרמטרים שהשתמשו במאמר (3) עם learning rate התחלתי של 0.001 וללא שינוי learning rate במהלך האימון. אילו נתוני שלקחתי baselines.

2. Learning_rate_decay:

ניסוי שני היה לחלק את ה learning rate במאה כל 60 epochs. זה שיפר את הביצועי ה training set אך ה mAP של הטסט סט ירד.

3. More_Priors:

בניסוי הזה כדי לשפר את הרזולוציה ניסיתי להגדיל את מספר priors בfeature maps
היותר low level אך לא היה שיפור משמעותי אז לא שמרתי את השינוי כי הוא הגדיל את
זמן החישוב.

4. Without_empty_images:

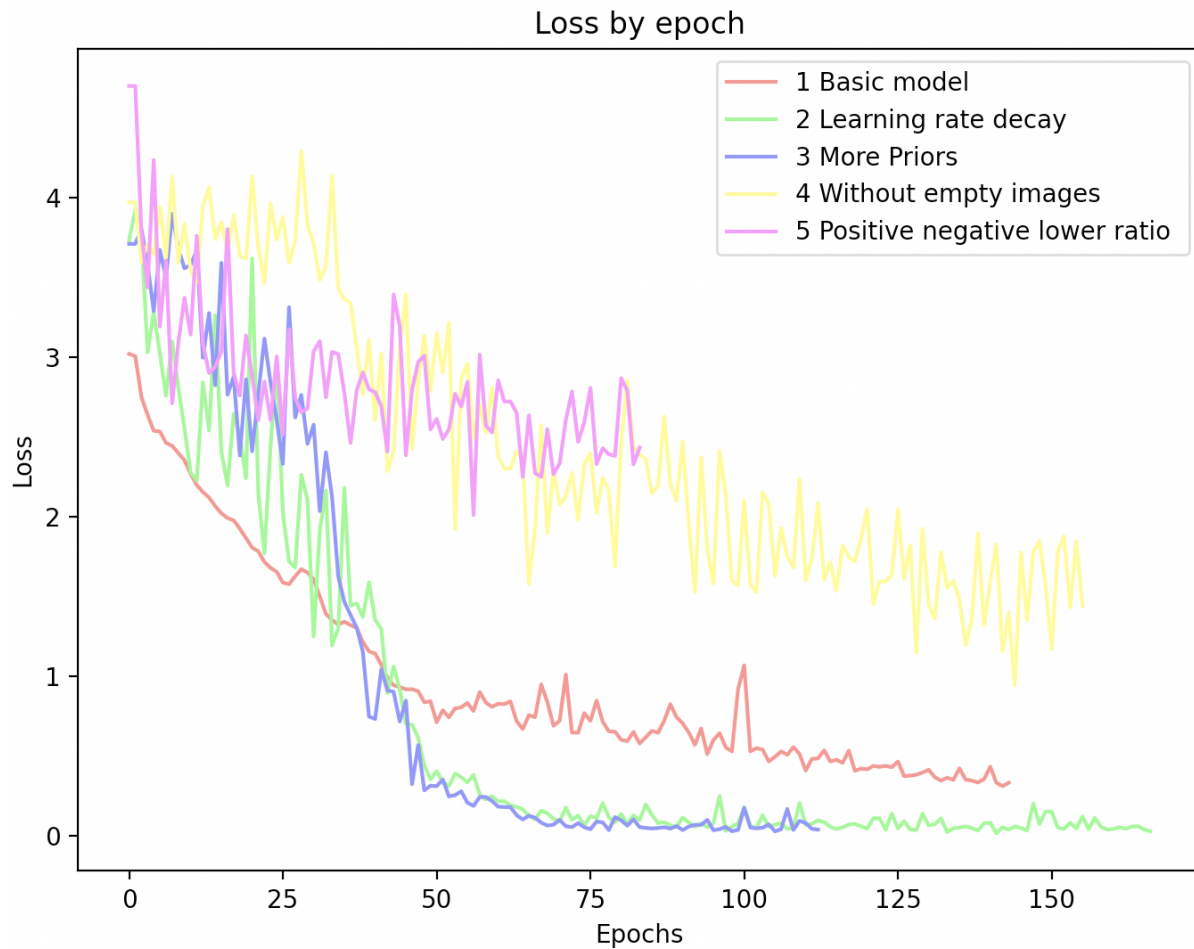
ניסוי של אי שימוש בתמונות ללא אובייקטים כדי להוריד את איזון בין המחלקות ולהקטין
את זמן החישוב. ניסיון הזה לא מוצלח כלל והביצועים ירדו משמעותית.

5. Positive_negative_lower_ratio:

ניסיון אחרון היה להקטין את negative/positive ratio מ3 ל2 שוב כדי לנסות להוריד את
האי שוויון בין המחלקות אך גם הפעם הניסיון לא הביא תוצאות טובות.

	epochs	mAP (train set)	loss	mAP (test set)
1_Basic_model	149	-	0.33	0.44
2_Learning_rate_decay	160	0.28	0.02	0.2528
3_More_Priors	110	0.4	0.036	0.219
4_Without_empty_images	150	-	1.43	0.26267
5_Positive_negative_lower_ratio	79	0.11	2.78	0.2476

טבלה 2: טבלה מסכמת של הביצועים הכוללת APs לכל מחלקה.



תמונה 6: גרף הוריאציות של lossn פר epoch בכל אחד מהביצועים.

כמו שניתן לראות בגרף של lossn פר epoch (תמונה 6), הוריאציות הטובות ביותר הן מספרי 2 ו-3 אך בtest set רואים ביצועים יותר טובים דווקא מ-1. זה יכול להעיד על overfitting במודלים 2 ו-3. למרות שהmAP של 1 טוב ביחס לוריאציות האחרות, ניתן לראות שכאשר בודקים כל class בנפרד (AP), יש מחלקות שהוא מזהה יותר טוב מאחרות.

- Aortic enlargement: 0.54,
- Cardiomegaly: 0.45,
- Pleural thickening: 0.25
- Pulmonary fibrosis: 0.51

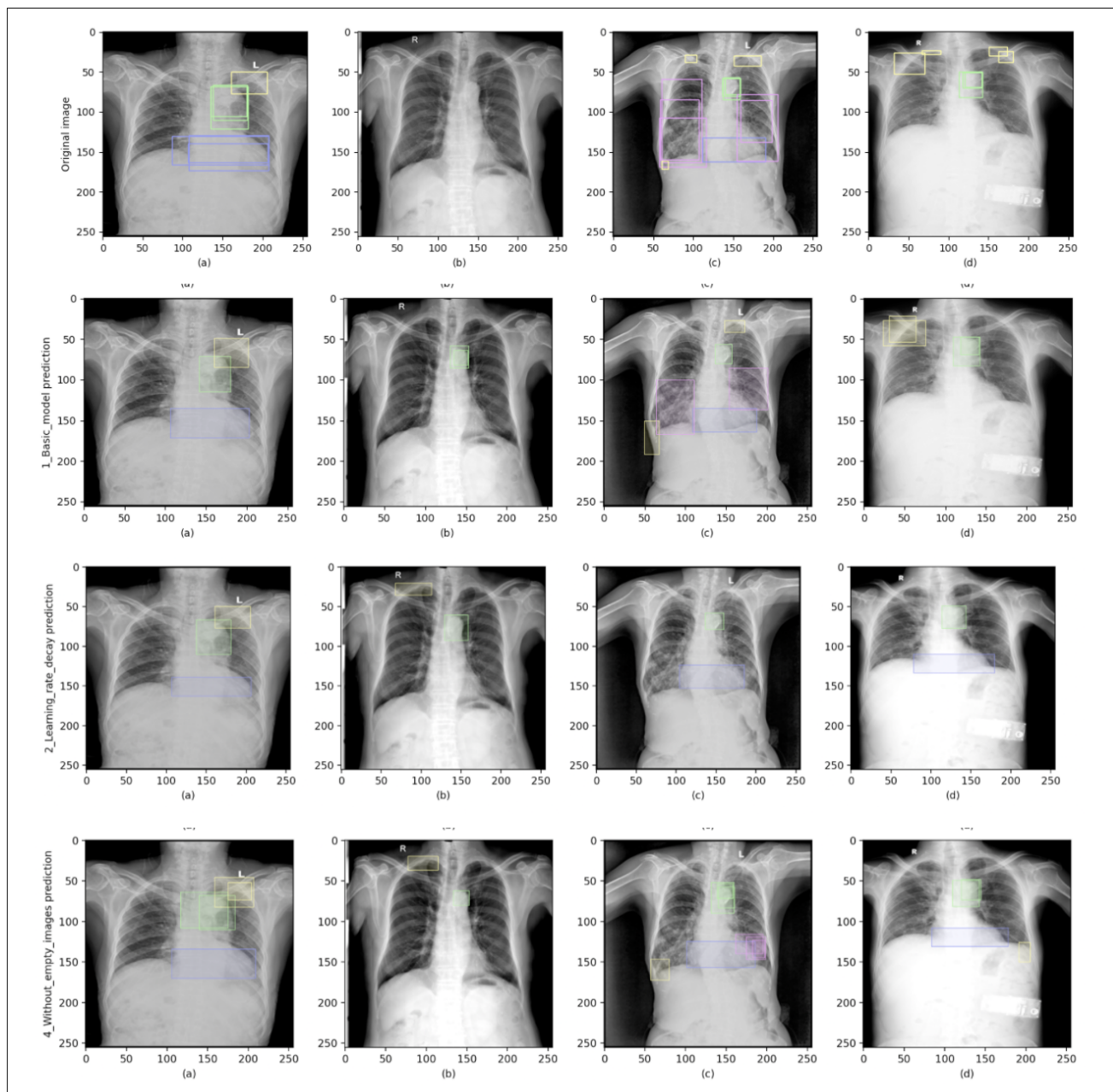
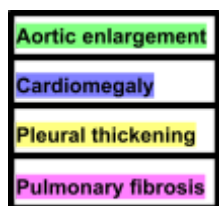
זה יכול לנבוע מכמות הדאטה על המחלקה הספציפית או קושי זיהוי כלשהוא בהבחנה של המחלה.

הבדלים בזיהוי בוריאציות שונות:

בתמונה 7 ניתן לראות תוצאה ויזואלית של הניבוי ע"י מודלים שונים ביחס לתמונה המקורית.

- 3 המודלים מצליחים לזהות היטב את תמונה (a) אבל מודל Without_empty_images(4) פחות חד ומכיל כמה זיהויים חופפים.
- תמונה (b) לא כוללת אנומליות בפועל אך המודלים זיהו false positives. ניתן לראות שמודל Basic Model(1) זיהה בטעות רק Aortic enlargement כאשר שני המודלים האחרים זיהו גם Pleural thickening.
- תמונה (c) קצת יותר מורכבת וכוללת מספר אובייקטים מכל מחלקה. גם כאן, מודל Basic Model(1) זיהה יפה את רוב האובייקטים כאשר מודל Learning_rate_decay(2) הצליח לזהות רק שני מחלקות. מודל Without_empty_images(4) הצליח לזהות את 4 המחלקות אך הוא פחות חד ממודל Basic Model(1) ופספס את 2 האובייקטים הקטנים למעלה (מודל Basic Model(1) פספס אחד).
- תמונה (d) כוללת pleural thickening קטנים באזור עצם הבריח וכמו בתמונה (c) הצד הימני פוספס ע"י מודל Basic Model(1) אך מודלים Learning_rate_decay(2) ו- Without_empty_images(4) בנוסף זיהו false positives.

לסיכום, גם בדיקה ויזואלית מראה שמודל Basic Model(1) הוא כראה הטוב מביניהם.



תמונה 7: דוגמאות של זיהוי על תמונות מהטסט סט בכל אחד מהניסיונות.

דיון ומסקנות

לסיכום, הושגו ביצועים טובים מבחינת מטריקות (AP, mAP) ותוצאות ויזואליות מהמודל הראשון. למרות כמה ניסיונות של שיפור המודל, לא הצלחתי להגיע לתוצאות טובות יותר על test set. בהינתן יותר זמן, הייתי רוצה לעשות יותר fine tuning ולהריץ את מודל בוריאציות שונות של thresholds מסוימים.

בנוסף, הייתי מנסה להחליף את החלק הראשון של האלגוריתם במודל חחכ קיים כמו VGG16 או Resnet50 ולהשתמש בtransfer learning כדי לקבל שכבות נמוכות כבר מאומנים היטב.

יהיה גם מעניין לממש אלגוריתמים נוספים כגון YOLO ו-Faster-RCNN.

אחד האתגרים העיקריים בפרויקט היה זמן חישוב. השתמשתי ב-GPU של google colab אך הוא מוגבל בזמן ועוצר אחרי זמן קצר, מה שממש הוריד את כמות הניסויים שהספקתי לנסות. הבעיה שניסיתי לפתור כאן לדעתי יותר מורכבת מהבעיות של מאמרי האלגוריתמים על זיהוי אובייקטים בתמונה במובן שבדרך כלל משתמשים בתמונות של חיי היום יום שכן אדם יכול לזהות בקלות. זיהוי "ידני" בתמונות רנטגן דורש מומחיות מסוימת וכך גם האלגוריתם צריך להיות יותר ספציפי. השגת יותר תמונות עם אנוטציות ידניות ישפיע על ביצועי המודל לטובה ללא ספק.

References

- (1) <https://cv-tricks.com/object-detection/faster-r-cnn-yolo-ssd/>
- (2) https://www.researchgate.net/publication/341116034_COVID-19_Detection_in_Chest_X-ray_Images_using_a_Deep_Learning_Approach
- (3) SSD: Single Shot MultiBox Detector - <https://arxiv.org/abs/1512.02325>
- (4) <https://towardsdatascience.com/breaking-down-mean-average-precision-map-ae462f623a52>