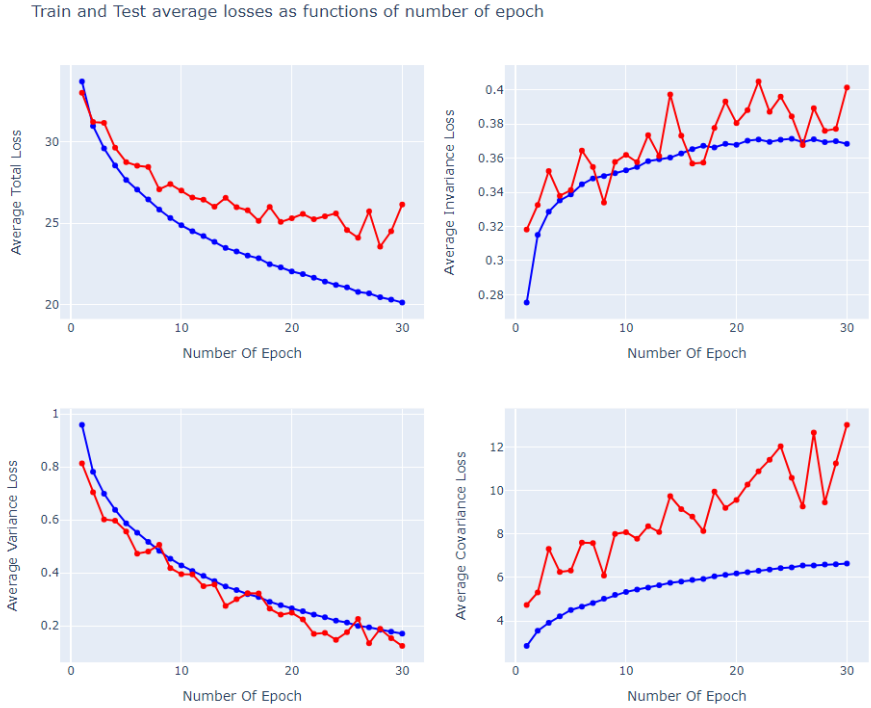
ML מתקדם – תרגיל בית 3
בר רוסו – 203765698

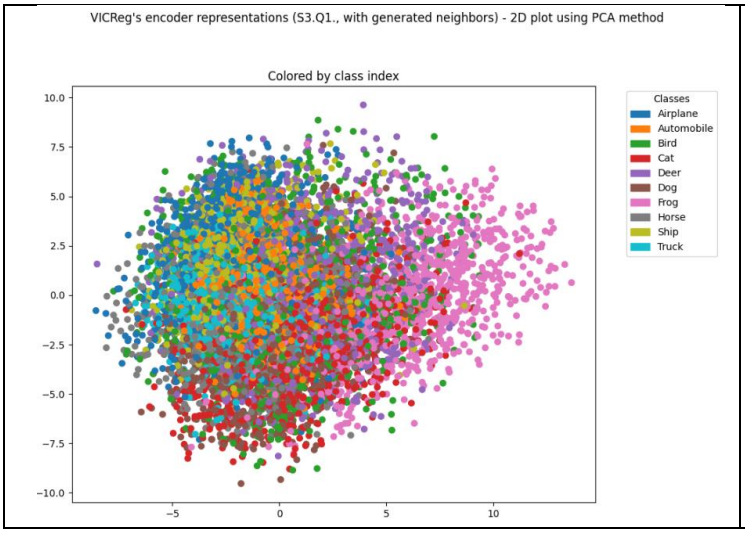
סעיף 3 - VICReg

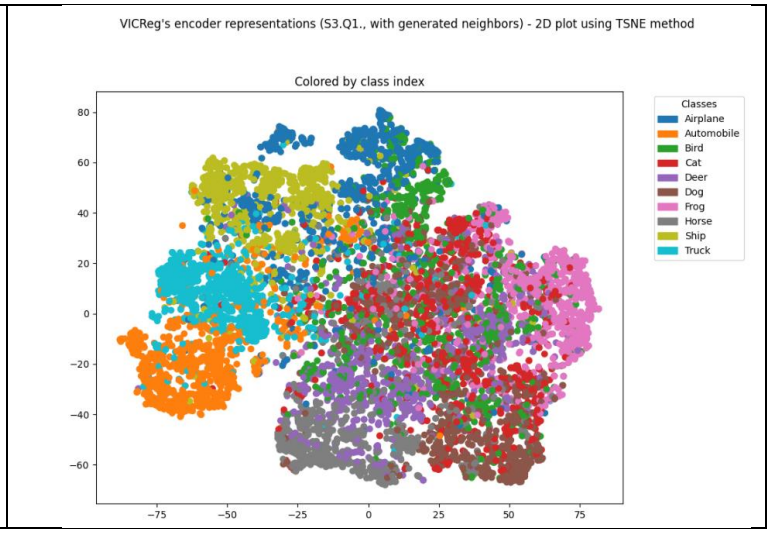
שאלה 1: סרטוט גרפים שמראים את הערך הממוצע של הlosses כפונקציה של מספר האפוקים.
חישבתי על 4 סוגי losses שונים: Total VICReg loss, Invariance Loss, Variance Loss and Covariance Loss.



אינדקס: צבע כחול – training losses, צבע אדום test losses.

שאלה 2: PCA visualization VS T-SNE visualization on CIFAR10 test dataset





ניתן לראות ש T-SNE יש וויזואליזציה יותר טובה לעומת PCA, שכן יש הפרדה יותר טובה בין הclasses השונים.
ניתן להסביר זאת מכיוון ש:
PAC - זאת שיטה **לינארית** המטילה את הייצוגים השונים ממימד גבוהה לממדים נבחרים שביחס אליהם השונות של הייצוגים היא

הגבוהה ביותר. זאת אומרת שניתן לתפוס רק קשרים ליניאריים בין הייצוגים השונים, מה שיכול לפגוע בוויזואליזציה כפי שאנחנו רואים.

T-SNE - זאת שיטה **לא לינארית** שמאופטמת למצוא ייצוגים ממימד יותר נמוך על סמך קרבה של הייצוגים במימד הגבוהה – זה מאפשר לנו לתפוס קשרים יותר מסובכים ובהתאם אנחנו מקבלים וויזואליזציה יותר טובה.

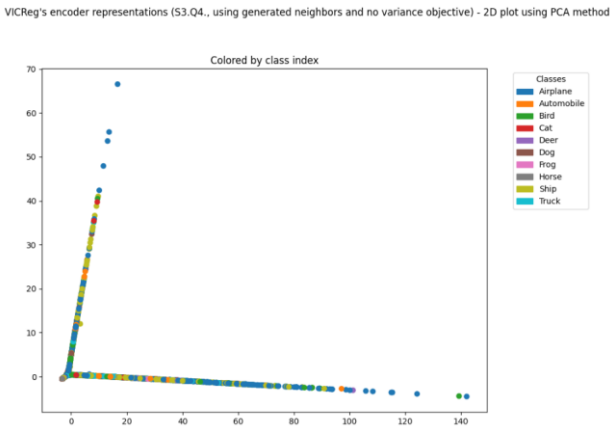
כמו כן, בהסתכלות יותר קרובה על המיפוי לדו מימד לפי T-SNE, ניתן לראות כי אנחנו מצליחים להפריד בצורה יותר ברורה בין מחלקות שונים של כלי רכב / משאיות / אוניות / מטוסים. לעומת זאת יש הפרדה פחות טובה עבור מחלקות של חיות (ציפור, חתול, צבי, כלב, צפרדע וסוס).

הסבר אפשרי לכך הוא שללא קשר לשיטה T-SNE בה אנו משתמשים, לencoder עצמו קשה יותר לתת ייצוגים משמעותיים לחיות שונות לעומת שאר המחלקות. יתכן שזה נובע מכך שמחלקות של חיות חולקות יותר מאפיינים משותפים (רגלים, עיניים, אוזניים וכו') לעומת שאר המחלקות שחולקות פחות מאפיינים משותפים (למשל תמונות של מטוסים ואוניות אינן דומות כלל).

שאלה 3: תוצאה של linear probing המבוסס על encoder משאלה 1:
Linear prober accuracy using VICReg's encoder representations (S3.Q1., with generated neighbors): **72.4000%**

שאלה 4: אימון VICReg ללא variance objective (כלומר μ=0):

א. PAC Visualization:



ב. Linear Probing Accuracy :
Linear prober accuracy using VICReg's encoder representations (S3.Q4., with generated neighbors and no variance objective): **15.2300%**

נשים לב שימוש בVICReg's encoder שאומן ללא Variance Objective מקבל שהPCA Visualization וה- Linear Probing Accuracy הרבה פחות טובים לעומת שימוש בVICReg's encoder משאלה 1 שכלל Variance Objective.

ניזכר שאנו משתמשים בVariance Objective על מנת לגרום לייצוג של כל מימד בווקטור הייצוג של תמונה להיות עם שונות מעל סף מסוים שאנו מגדירים מראש. לכן, אי שימוש בVariance Objective עלול לגרום לcollapse לייצוגים בתת מרחב במימד נמוך.
זו הסיבה שאנחנו מקבלים ב PCA visualization שהדגימות מתלכדות לשני קווים שכן זה מלמד אותנו שכל הייצוגים במימד הגבוהה לקו (תת מרחב ממימד 1) עם שני כיוונים עיקריים.

כמו כן collapse של הייצוגים למרחב ממימד נמוך גורם לכך שאנחנו מאבדים פיצ'רים חשובים שיכולים לתפוס מידע סמנטי על התמונות, זו הסיבה שאנחנו מקבלים דיוק פחות טוב משימוש בlinear probing.

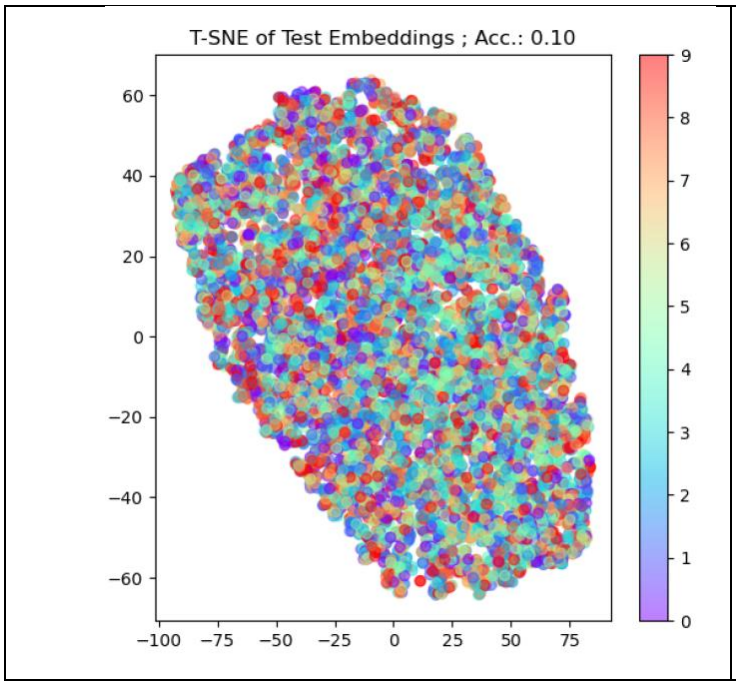
שאלה 5: שימוש ב VICReg שאומן ללא Generated Neighbors:

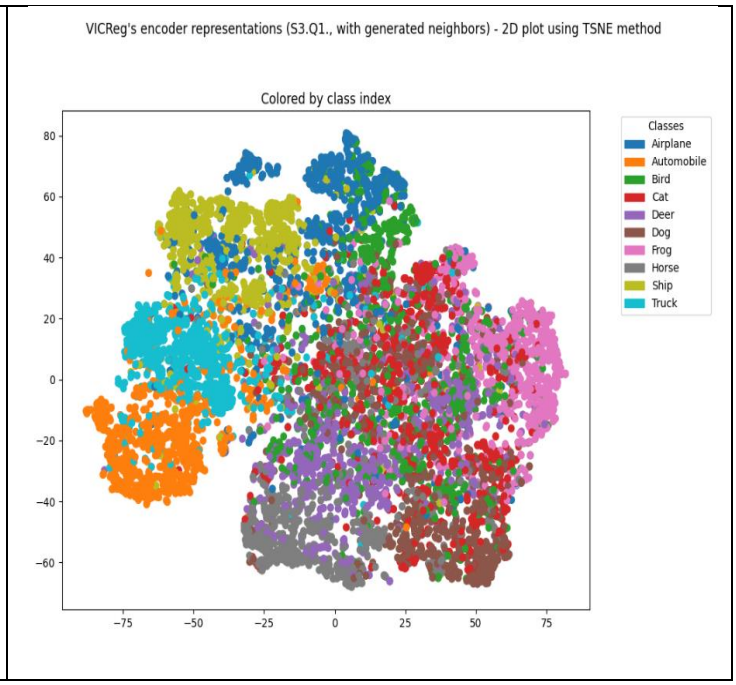
: Linear Probing Accuracy
Linear prober accuracy using VICReg's encoder representations (S3.Q5., without generated neighbors): 51.9000%

נשים לב שקיבלנו דיוק יותר נמוך בהשוואה לlinear probing עם encoder שאומן עם אוגמנטציות.
ניתן להסביר זאת מכיוון על ידי כך שכאשר אנחנו מוסיפים אוגמנטציות אנחנו גורמים למודל ללמוד ייצוגים יותר סמנטיים על האובייקט של התמונה. כתוצאה מכך אנחנו מקבלים ייצוגים יותר טובים שחסינים לשינויים של אותו אובייקט בתמונה.

לעומת זאת, ב VICReg משאלה זו, אנחנו בחרנו את השכנים הכי קרובים (לפי מטריקת l2) של הייצוג של כל תמונה לפי encoder משאלה 1. אומנם זה יותר זה יותר טוב מאשר לבחור את שכנים הכי קרובים לפי מטריקת l2 על התמונות עצמן, אך נשים שהencoder משלה אחת (שאומן עם אוגמנטציות) אינו לומד את הייצוגים באופן מושלם וגם לו יש טעויות.
למשל לפי ה T-SNE visualization של הencoder משאלה 1, אנחנו רואים שהוא מתקשה להבדיל בין חיות שונות, כתוצאה מכך יתכן שנבחר תמונה של כלב כשכן של תמונה של חתול (דבר שלא יתכן שיקרה בשימוש באוגמנטציות), מה שבסופו של דבר גורם ללימוד ייצוגים פחות טובים ולכן אנחנו גם מקבלים שהדיוק של הlinear probing לשאלה זו פחות טוב.

שאלה 6: T-SNE of Laplacian Eignemaps Representations VS T-SNE visualization of encoder from Q1



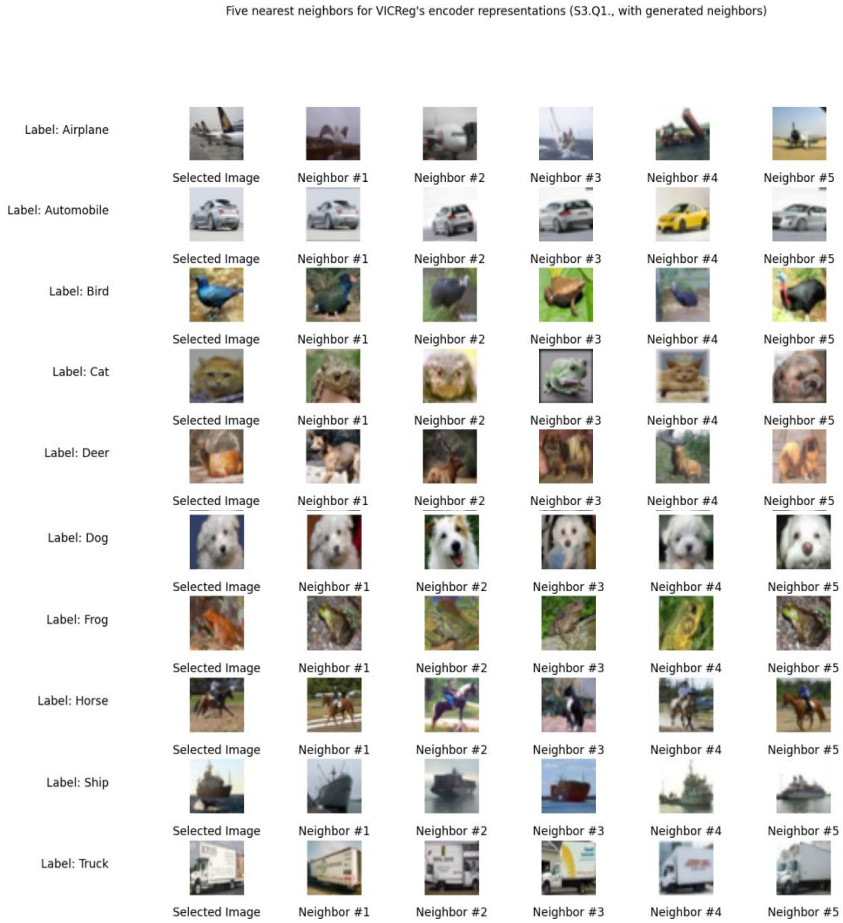


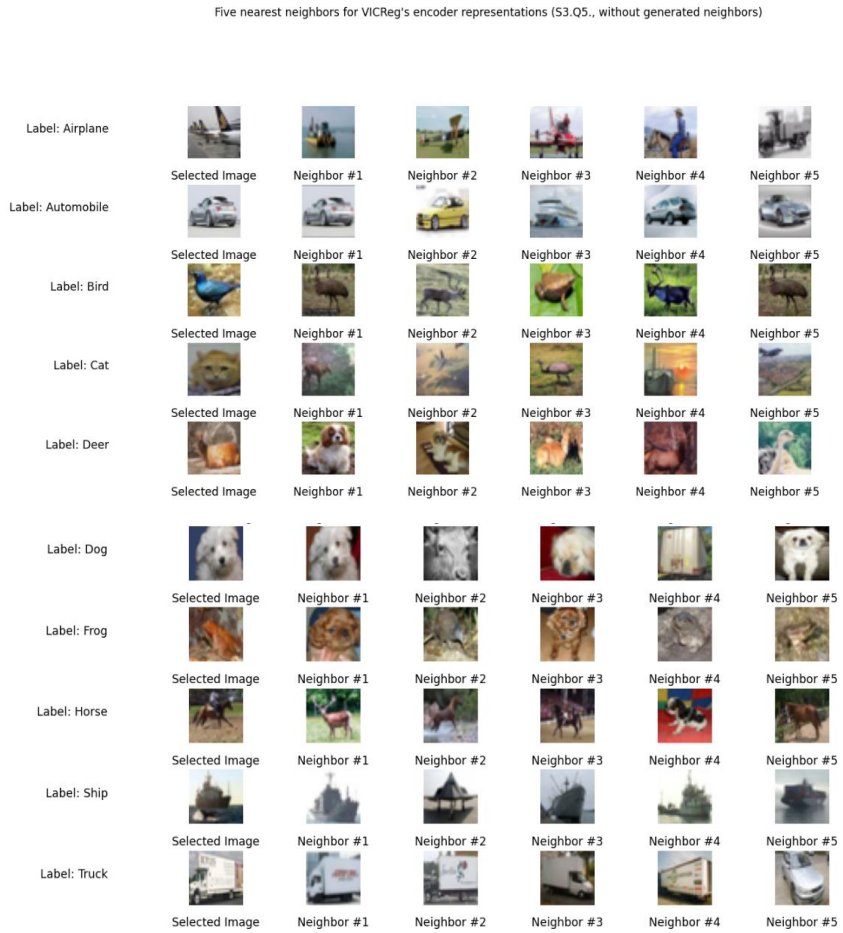
ניתן לראות ששימוש בVICReg מניב תוצאות הרבה יותר טובות לעומת שימוש ב Laplacian Eigenmaps, הן מבחינת הוויזואליזציה של T-SNE, והן מבחינת הדיוק של הLinear probing (72.4% לעומת 10%).

ניתן להסביר זאת מכיוון שבשיטה של Laplacian Eigenmaps אנחנו בוחרים את השכנים של כל תמונה בdataset לפי מדד מרחק **אוקלידי**, לעומת בVICReg אנחנו משתמשים באוגמנטציות. זהו מרכיב משמעותי ביותר עבור תמונות או כל אובייקט כלשהו ממימד גבוהה, כך שמרחק אוקלידי בין ייצוגים בraw dataset לא משקף את השוני הסמנטי של אותם אובייקטים שמיוצגים על ידי אותם ווקטורים. במקרה של תמונות יתכן שנבחר כשכנים שני תמונות שהערכים של הפיקסלים בניהם יחסית קרובים, אך האובייקטים המוצגים בהם שונים בתכלית.

זו הסיבה שב T-SNE visualization של ה Laplacian Eigenmaps אנחנו לא מקבלים הפרדה בין המחלקות השונות מכיוון שלא למדנו ייצוגים סימנתים על התמונות.

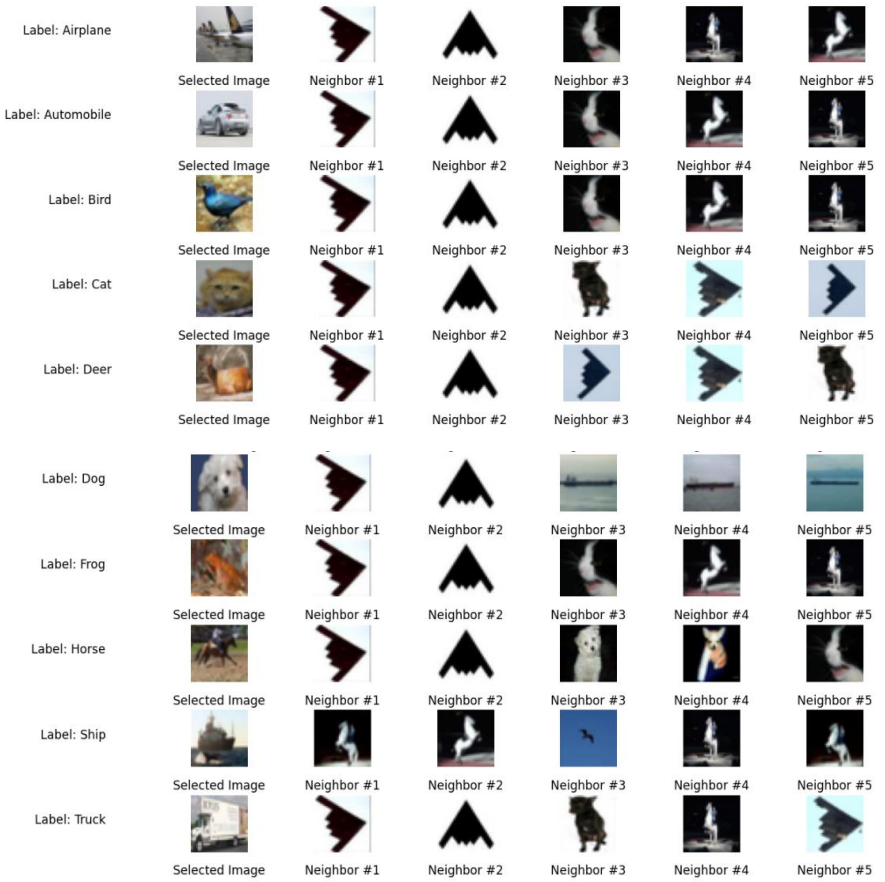
שאלה 7: 5 nearest and 5 distant images for each selected image according to encoder from Q1 and Q5





Five distant neighbors using VICReg's encoder representations (S3.Q1., with generated neighbors):s

Five distant neighbors for VICReg's encoder representations (S3.Q5., without generated neighbors)

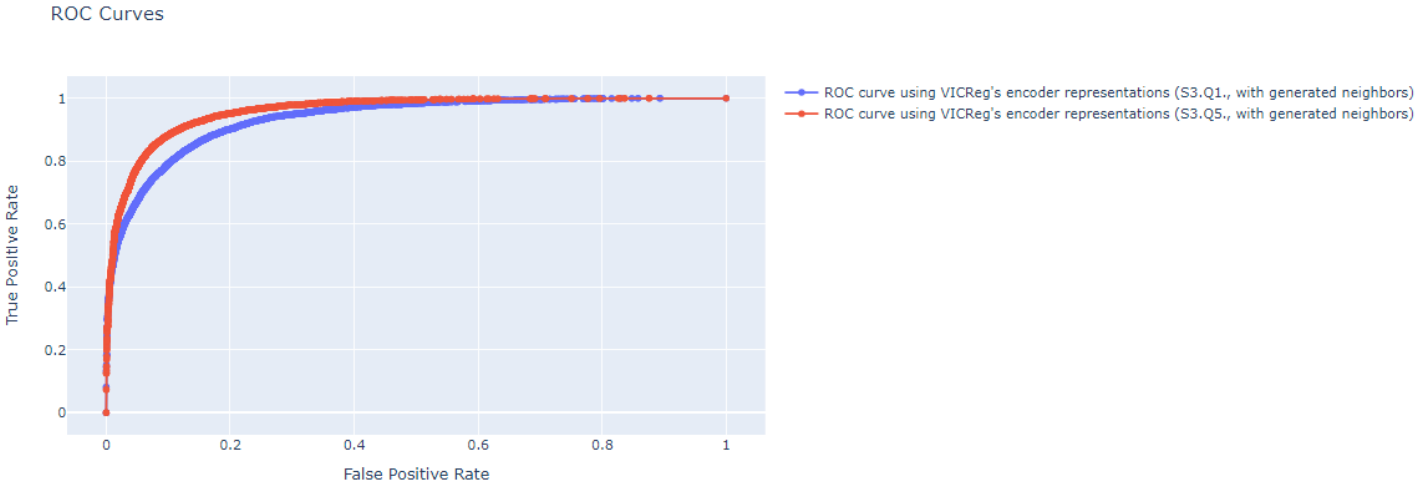


הסבר עבור שכנים הכי קרובים:
נשים לב שהencoder משאלה 1 מצליח לשמור יותר טוב על ייצוגים שקרובים סמנטית (שומר על האובייקט הנראה) לתמונת המקור, זאת לעומת הencoder משאלה 5, שבאופן כללי מתקשה יותר בתמונות ממחלקה של חיות.
ההסבר הוא דומה להסבר שניתן בשאלה 5 - בעוד שהencoder משאלה 1 מתאמן על אוגמנטציות שתמיד שומרות על הסמנטיות של התמונה (האובייקט הנראה), הencoder משאלה 5 מסתמך על הייצוגים של הencoder משאלה 1 כדי לקבוע את השכנים שלו לפי מטריקת מרחק L2.
מכאן, מכיוון שהencoder משאלה 1 מתקשה יותר להפריד בין תמונות ממחלקות של חיות (לפי ה T-SNE Visualization) כך נקבל שהencoder משאלה 5 יתקשה ללמוד ייצוגים סמנטיים ייחודיים לכל מחלקת חיות ולכן אנחנו רואים את הערבוב בתמונות של החיות חיות בשכנים הכי קרובים.

הסבר עבור שכנים הכי רחוקים:
שתיהן מציגות תמונות ממחלקות שונות מתמונות המקור (חוץ מאוניות עבור ה encoder משאלה 1 ומטוסים עבור הencoder משאלה 5).
בשימוש בencoder משאלה 5, אנחנו רואים כי מרבית התמונות הרחוקות מאופיינות סט של צבעים (שחור לבן) מה שמרמז שהייצוגים האלה תופסים היבטים חזותיים של התמונה.
לעומת זאת, בשימוש מencoder משאלה 1 ניתן לראות שהצבעים בתמונות המרוחקות מתפלגים באופן יותר אחיד, מה שלמלמד שהייצוגים האלה פחות תופסים היבטים חזותיים של התמונה.

סעיף 4.1 - Anomaly Detection

שאלה 2: ROC AUC Evaluation




```
AUC using VICReg's encoder representations (S3.Q1., with generated neighbors): 0.9348
AUC using VICReg's encoder representations (S3.Q5., with generated neighbors): 0.9590
```

נשים לב שדווקא בשימוש בencoder משאלה 5 קיבלנו תוצאות יותר טובות לעומת שימוש בencoder משאלה 1.

יתכן שזה נובע מהאופי הוויזואלי של התמונות מ MNIST Dataset, שכן התפלגות הצבעים בכל התמונות דומה (שחור או לבן), בעוד שתמונות מ CIFAR10 פחות חולקות מאפיינים ויזואליים משותפים.
זה כנראה יוצר bias עבור התמונות מ MNIST, ומכיוון שה encoder משאלה 5 תופס יותר טוב בייצוגים שלו גם מאפיינים ויזואליים של התמונות, אז כנראה יותר קל לו להכריע אם תמונה שייכת לMNIST (קרי תמונה אנומלית) או לא.

באופן דומה, יתכן שאם במקום MNIST היינו בוחרים בdataset אחר של תמונות סמנטיות (למשל תמונות של בתים) שלא חולקים מאפיינים וויזואליים, אז ה encoder משאלה 1 היה מצליח יותר.

שאלה 3: Qualitative Evaluation - Ambiguity of Anomaly Detection

7 most anomalous images using VICReg's encoder representations (S3.Q1., with generated neighbors)

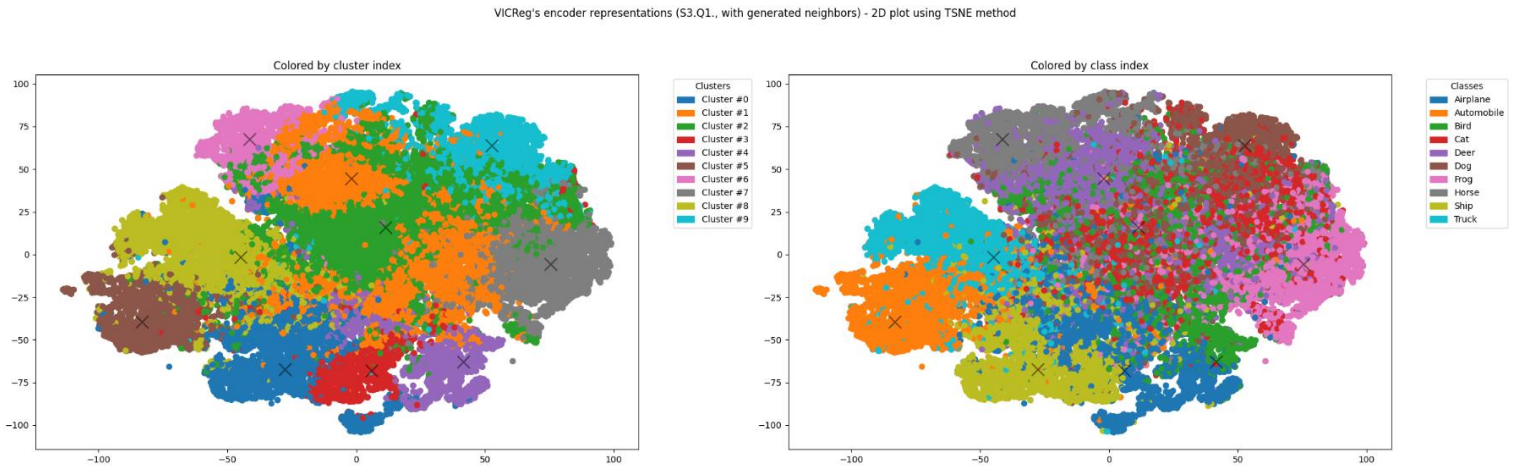
7 most anomalous images using VICReg's encoder representations (S3.Q5., without generated neighbors)

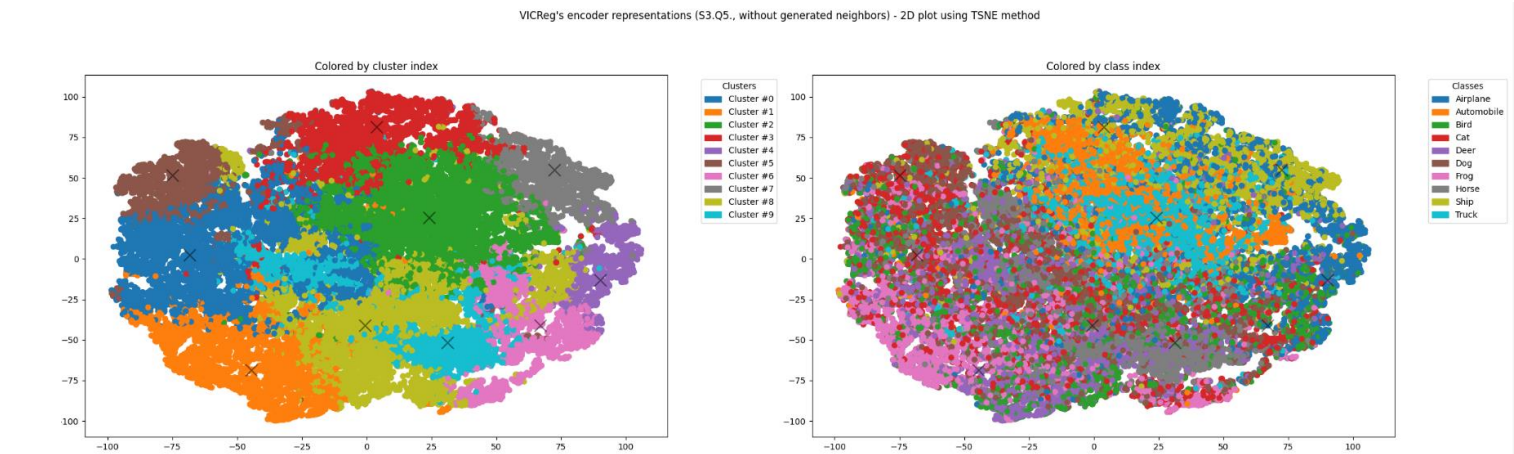
הסבר: יתכן שתמונות מMNIST עם הספרה 8 נמצאות כיותר חריגות לעומת הספרה 1, כי אולי הצורה של הספרה 1 יכולה להזכיר צורה (מלמעלה) של אוניה / רכב / משאית בעוד שהספרה 8 פחות דומה מבחינה סמנטית לתמונות של CFIAR10.
מצד שני, מכיוון שהencoder משאלה 5 נוטה גם להכליל בתוך הייצוגים שלו גם מידע ויזואלי על התמונות, אז זה מכריע את הכף לטובות תמונות עם ספרה 1.

עם זאת, בסה"כ נראה ששני הencoders מצליחים להבחין בתמונות מMINST בתור תמונות הכי אנומליות ולאור ההוצאות הצמודות שקיבלנו מהשאלה הקודמת נראה שאין יתכן ברור לאחד מהם על פני האחר עבור זיהוי אנומליות של תמונות מMNIST.

סעיף 4.2 - Clustering

שאלה 2: Visualizing the Clusters in 2D





הפרדה בין clusters שונים:
ניתן לראות ביצועים דומים עבור שני סוגי הencoders.

הפרדה בין מחלקות שונות:
ניתן לראות שהייצוגים של מחלקות שונות לפי ה encoder משאלה 1 על CIFAR train dataset פחות מעורבבים לעומת אותם ייצוגים לפי ה encoder משאלה 5.

הסבר: מכיוון שהencoder משאלה 1 אומן תחת VICReg עם אוגמנטציות, זה מאפשר לו ללמוד ייצוגים הרבה יותר סימנתים עבור כל תמונה, מה שאומר שייצוגים קרובים יהיו בהסתברות יותר גבוהה יהיו שייכים לאותה מחלקה.
זה מסביר מדוע אנחנו מקבלים שהייצוג לפי cluster מזכיר דומה הרבה יותר לייצוג לפי class אצל ה encoder משאלה 1 לעומת ה encoder משאלה 5.

שאלה 3: Quantitative Analysis

```
Silhouette score for VICReg's encoder representations (S3.Q1., with generated neighbors): 0.0884
Silhouette score for VICReg's encoder representations (S3.Q5., without generated neighbors): 0.1453
```

הערכים של Silhouette Score נעים בין מינוס 1 ל1, כאשר ככל שהscore יותר קרוב ל1 אז הclustering משעורך בצורה טובה.
אומנם קיבלנו תוצאות קצת טובות עבור clustering על ייצוגים של תמונות בעזרת ה encoder משאלה 5, נסים לב כי **ההפרש** בין שני הscores קטן מאוד באופן יחסי לטווח האפשרי (טווח 2), ואכן קיבלנו תוצאות דומות לפי Clusters T-NSE Visualization.

מצד שני, באופן יחסי אחד לשני, הscore משימוש בencoder משאלה 5 גדול כמעט פי לעומת הscore משימוש בencoder משאלה 1, אך לא ניתן לראות את היתרון הזה בא לידי ביטוי ב T-NSE Visualization של הclusters.
יתכן שבגלל שאנחנו ממפים את הdata ממימד גבוהה שלנו ל2D space, מה שגורם לנו לאבד מידע ייצוג, היתרון הזה נשמט בעת הוויזואליזציה של הclusters.

LLM Usage:

נעזרתי בChatGPT בשביל לכתוב פונקציות שעושות ויזואליזציה ומשתמשות בחבילות Matplotlib ו Plotly.
כלל הפונקציות האלה נמצאות תחת החלק של "**Auxiliary Functions**".

למעט הפונקציות שניתנו לנו **models.py** and **augmentations.py** ומופיעים בתחילת הקובץ, את כל שאר הפונקציות והמחלקות מימשתי בעצמי.