סקירה זו היא חלק מפינה קבועה בה אני סוקר מאמרים חשובים בתחום ה-ML/DL, וכותב גרסה פשוטה וברורה יותר שלהם בעברית. במידה ותרצו לקרוא את המאמרים הנוספים שסיכמתי, אתם מוזמנים לבדוק את העמוד שמרכז אותם תחת השם deepnightlearners.

לילה טוב חברים, היום אנחנו שוב בפינתנו deepnightlearners עם סקירה של מאמר בתחום הלמידה העמוקה. היום בחרתי לסקירה את המאמר שנקרא

Contrastive Learning Of Medical Visual Representations From Paired Images And Text

---

פינת הסוקר:

המלצת קריאה ממייק: חובה לעוסקים בתחום של צילום רפואי, לאחרים מומלץ מאוד.

בהירות כתיבה: גבוהה.

רמת היכרות עם כלים מתמטיים וטכניקות של ML/DL הנדרשים להבנת מאמר: היכרות עם טכניקות בסיסיות של למידת ייצוג (representation learning) .

יישומים פרקטיים אפשריים: שיפור איכות של pretraining של רשתות על דאטה מהדומיין הרפואי.

פרטי מאמר:

לינק למאמר: זמין כאן

לינק לקוד: לא רשמי 1, לא רשמי 2

פורסם בתאריך: 02.10.2020, בארקיב

הוצג בכנס: ICLR 2021

תחום מאמר:

- למידת ייצוג (representation learning) לצילומים רפואיים

כלים מתמטיים, מושגים וסימונים:

- Noise Contrastive Estimation - NCE
- Contrastive VIsual Representation Learning from Text - ConVIRT

תמצית מאמר:

המאמר מציע שיטה בשם (Contrastive VIsual Representation Learning from Text (ConVIRT, לבניית ייצוג במימד נמוך (לטנטי) של צילומים רפואיים תוך שימוש בגישה הנקראת (Noise Contrastive Estimation NCE). החידוש שמביא המאמר הוא שימוש ב-NCE לבניית ייצוגים עבור שני אופיינים של צילום רפואי **בו זמנית:** הראשון הוא ייצוג של צילום עצמו והשני הוא ייצוג של כותרת (תיאור) טקסטואלי של הצילום.

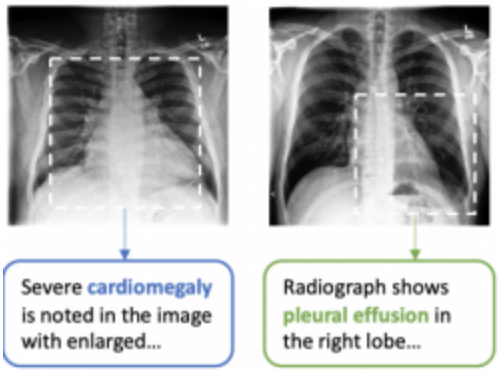


Figure 1: Two example chest radiograph images with different abnormality categories, along with sentences from their paired textual report and example views indicative of their characteristics.

רעיון בסיסי:

להבדיל מ-NCE המקורי המאמר מנסה לבנות ייצוגים של צילום ושל התיאור הטקסטואלי שלו כך שהייצוג של צילום יהיה "דומה (קרוב) יותר" (אחרי טרנספורמציה מסוימת) לייצוג של תיאור הצילום שהוא מופיע עליו מאשר לתיאור של כל צילום אחר. באופן משלים כל ייצוג של תיאור צריך להיות כמה שיותר ״קרוב״ לייצוג הצילום שהוא מתאר מאשר לייצוג של כל צילום אחר (זו הסיבה המחברים קוראים לגישה שלהם דו-כיוונית במאמר. לדעתם של המחברים גישה ״דו-כיוונית״ זו מאפשרת להגיע לייצוג צילום המכיל בתוכו תכונות "סמנטיות חזקות מהתיאור שלו".

תקציר מאמר:

יצירת דאטהסטים מתויגים איכותיים בעולם צילומים רפואיים היא יקרה מאוד. רוב הדאטהסטים המתויגים הם לא גדולים שמקשה מאוד על אימון מודלים גדולים (הכוונה לרשתות נוירונים) בעלי יכולת הכללה טובה. מצד שני ניסיונות להשתמש בייצוגים מאומנים על דאטהסטים מדומיינים אחרים (כמודל pretrained וכיולו על דאטהסט קטן מהדומיין הרפואי לאחר מכן) בדרך כלל לא מובילים לייצוגים חזקים בדומיין הצילומים הרפואיים. הסיבה לכך היא הבדלים אינהרנטיים מהותיים בין האופיינים של תמונות טבעיות לבין צילומים רפואיים. מצד שני שימוש בגישות self-supervised לבניית ייצוגים בדומיין הרפואי נתקלים גם כן בקשיים עקב הבדלים ויזואליים די קטנים בין צילומים רפואיים מקלאסים (קטגוריות) שונים.

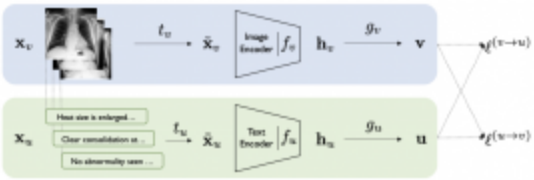


Figure 2: Overview of our ConVIRT framework. The blue and green shades represent the image and text encoding pipelines, respectively. Our method relies on maximizing the agreement between the true image-text representation pairs with bidirectional losses $\ell^{(v\to u)}$ and $\ell^{(u\to v)}$.

כדי לתת מענה לסוגייה זו, המאמר מציע לנצל דאטה טקסטואלי המופיע מעל צילומים רפואיים לבנייה של ייצוגים עשירים יותר. השיטה המוצעת במאמר - Contrastive VIsual Representation Learning from text ConViRT מנסה "לקרב" ייצוגים של צילום ותיאורו ו"להרחיק" עד כמה שניתן ייצוגים של זוגות אקראיים של (צילום, תיאור). בעצם ConViRT מהווה הרחבה "קרוס-דומיינית" (cross-domain) של NCE קלאסי. כלומר ConViRT בונה ייצוגים בשני דומיינים שונים בו זמנית להבדל מ-NCE שעושה זאת בדומיין אחד. למעשה הרחבה זו נותנת מענה לשוני ויזואלי קטן בין צילומים רפואיים מקטגוריות שונות, המקשה על שימוש ב-NCE סטנדרטי עבור דומיין זה. כפי שמקובל בדומיינים אחרים, מאמנים את ConViRT על כמה דאטהסטים גנריים מהדומיין הרפואי (pretrain) ולאחר מכן מכיילים את המודל למשימת downstream (פיין טיונינג).

הסבר קצר על NCE:

בשביל להבין איך עובד NCE בשני דומיינים בו זמנית, בואו ניזכר מה זה NCE קלאסי. הנחת היסוד ב-NCE אומרת כי ייצוג חזק בהכרח "מסוגל להפריד" בין זוג דוגמאות קרובות (דוגמאות קשורות או שתי אוגמנטציות של אותה דוגמא) לבין זוגות של דוגמאות רחוקות (כגון רנדומליות). כלומר דמיון במרחב המקורי בין דוגמאות צריך להיות מתורגם למרחב הייצוגים שלהם. כלומר ייצוגים של דוגמאות דומות צריכות להיות קרובות ואילו ייצוגים של דוגמאות לא דומות צריכות להיות רחוקים. בין השימושים של טכניקה זו ניתן למנות negative sampling שהשתמשו בו לבנייה של word2vec. ניתן להוכיח כי הקטנת ערך של פונקציית לוס (של דוגמא נתונה) עבור צורה מסוימת של NCE (הנקראת InfoNCE, שבה גם משתמשים במאמר זה) מובילה לעלייה במידע ההדדי בין הדוגמא במרחב המקורי לבין ייצוגה במרחב ממימד נמוך. עלייה זו כמובן מצביעה על אובדן פחות אינפורמציה בין דאטה מקורי לבין ייצוגה במימד נמוך כלומר הייצוג יהיה פחות לוסי ויכיל יותר מידע של הדוגמא. חשוב לציין שהאימון מתבצע במרחב הייצוג ולא במרחב המקורי כלומר הלוס מחושב על הייצוגים במרחב ממימד נמוך.

אז איך נראית פונקציית לוס של NCE המקורי?

לדוגמא נתונה בונים זוג (חיובי) של דוגמאות דומות (קרובות) s_1 ו- s_2 (למשל אוגמנטציות של אותה תמונה). לאחר מכן בוחרים מספר דוגמאות רנדומליות ומרכיבים זוגות מ-s_1 ו-s_2 עם הדוגמאות האלו. פונקצית מטרה היא היחס של דמיון של הזוג הקרוב (חיובי) לסכום הדמיונות בין הדוגמאות רנדומליות (שליליות) והמטרה היא למקסם פונקציה זו. צריך לציין כי ככל יש מוסיפים יותר זוגות שליליים בפונקציות לוס של NCE הוא גבוה יותר, ניתן להשיג ערך גבוה יותר מידע הדדי בין דוגמא וייצוגה במימד נמוך באמצעות מזעור של פונקציית לוס (מקסום של פונקציית מטרה).

קרוס-דומיין NCE של ConViRt:

במקום לבנות זוגות מאותו דומיין ConViRT בונה זוגות מדומיינים שונים (מהדומיין של תמונות ומהדומיין הטקסטואלי). כלומר לוקחים צילום וחלק מהתיאור שלו ובונים מזה זוג חיובי. אחר כך בונים זוגות רנדומליים של צילומים והתיאורים שלהם. כמובן משתמשים באוגמנטציות של צילומים בשביל לבנות זוגות חיוביים. נגיד לוקחים

צילום, עושים לו crop ובונים זוג חיובי עם חלקים שונים של תיאורו (פשוט דוגמים משפטים מתיאור הצילום באופן רנדומלי).

אחרי שבונים את הזוגות החיוביים והשליליים מעבירים כל אחד מהם דרך הרשת המקודדת משלו (אחת לצילום והשנייה לטקסט). לאחר מכן בונים מיני-באטץ' המכיל זוג חיובי אחד והשאר הם זוגות שליליים. את הצילום מעבירים דרך המקודד שלו (המאמר השתמש ב-ResNet50) ואת הטקסט מעבירים דרך המקודד שלו (כמו שאתם יכולים לנחש זה לא אחר אלא BERT). לאחר מכן לוקחים את הפלטים של שני המקודדים האלו ו"מטילים" אותם על מרחב מאותו מימד כדי שניתן יהיה להשוותם (מעבירים את שניהם דרך רשת בעלת שתי שכבות - כמובן כל אחד מועבר דרך הרשת שלו). לאחר מכן מחשבים דמיון בין הפלטים באמצעות מרחק הקוסיין (cosine). בשלב האחרון מציבים את המרחקים האלו לשתי פונקציות לוס של InfoNCE: בראשנה המכנה מכיל את סכום אקספוננטים של כל המרחקים בין כל הזוגות המכילים את התיאור מהזוג החיובי וצילומים רנדומליים כאשר השני מכיל את המרחקים בין הצילום מהזוג החיובי לשאר התיאורים מהמיני-באטץ. המונה בשני האיברים הוא המרחק בין הייצוגים של הזוג החיובי. הלוס הסופי הינו סכום של שני הלוסים האלו.

הישגי מאמר:

המחברים ביצעו אימון pretrain של ConViRT על שני דאטהסטים: MIMIC-CXR, ועל הדאטהסט musculoskeletal מ-rhode island hospital. לאחר מכן הם כיילו את המודל המאומן לכמה סוגים של משימות:

Table 1: Results for the medical image classification tasks: (a) linear classification; (b) fine-tuning setting. All results are averaged over 5 independent models. Best results for each setting are in boldface. COVIDx 1% setting is omitted due to the scarcity of labels in COVIDx.

(a)

Method	RSNA (AUC) 1%	RSNA (AUC) 10%	RSNA (AUC) all	CheXpert (AUC) 1%	CheXpert (AUC) 10%	CheXpert (AUC) all	COVIDx (Accu.) 10%	COVIDx (Accu.) all	MURA (AUC) 1%	MURA (AUC) 10%	MURA (AUC) all
General initialization methods											
Random Init.	55.0	67.3	72.3	58.2	63.7	66.2	69.2	73.5	50.9	56.8	62.0
ImageNet Init.	82.8	85.4	86.9	75.7	79.7	81.0	83.7	88.6	63.8	74.1	79.0
In-domain initialization methods											
Caption-Transformer	84.8	87.5	89.5	77.2	82.6	83.9	80.0	89.0	66.5	76.3	81.8
Caption-LSTM	89.8	90.8	91.3	85.2	85.3	86.2	84.5	**91.7**	75.2	81.5	84.1
Contrastive-Binary	88.9	90.5	90.8	84.5	85.6	85.8	80.5	90.8	76.8	81.7	85.3
ConVIRT (Ours)	**90.7**	**91.7**	**92.1**	**85.9**	**86.8**	**87.3**	**85.9**	**91.7**	**81.2**	**85.1**	**87.6**

(b)

Method	RSNA (AUC) 1%	RSNA (AUC) 10%	RSNA (AUC) all	CheXpert (AUC) 1%	CheXpert (AUC) 10%	CheXpert (AUC) all	COVIDx (Accu.) 10%	COVIDx (Accu.) all	MURA (AUC) 1%	MURA (AUC) 10%	MURA (AUC) all
General initialization methods											
Random Init.	71.9	82.2	88.5	70.4	81.1	85.8	75.4	87.7	56.8	61.6	79.1
ImageNet Init.	83.1	87.3	90.8	80.1	84.8	87.6	84.4	90.3	72.1	81.8	87.0
In-domain initialization methods											
Caption-Transformer	86.3	89.2	92.1	81.5	86.4	**88.2**	88.3	92.3	75.2	83.2	87.6
Caption-LSTM	87.2	88.0	91.0	83.5	85.8	87.8	83.8	90.8	78.7	83.3	87.8
Contrastive-Binary	87.7	89.9	91.2	86.2	86.1	87.7	89.5	90.5	80.6	84.0	88.4
ConVIRT (Ours)	**88.8**	**91.5**	**92.7**	**87.0**	**88.1**	88.1	**90.3**	**92.4**	**81.3**	**86.5**	**89.0**

1. **סיווג צילום**

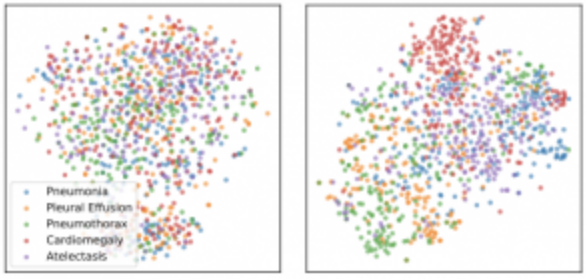
דאטהסטים: RSNA Pneumonia Detection, CheXpert, CovidX, MURA.

2. **מציאת צילום הדומה ביותר לצילום נתון (Zero-shot Image-image Retrieval).**

דאטהסטים: CheXpert 8×200 Retrieval Dataset.

3. מציאת צילום הכי דומה לתיאור נתון (דאטה סט כמו הקודם)

בכל המשימות האלו הצליחו המבחרים להשיג ביצועים טובים יותר עם השיטה המוצעת בהשוואה למגוון שיטות pretraining אחרות.



(a) ImageNet Pretraining (b) ConVIRT Pretraining

Figure 3: t-SNE visualizations of encoded image representations from different pretraining methods.

נ.ב. מאמר עם רעיון מגניב להתגבר על קושי בבניית ייצוגי צילומים בדומיין הרפואי. כתוב מאוד ברור ומפורט.
מומלץ

#deepnightlearners

הפוסט נכתב על ידי מיכאל (מייק) ארליכסון, PhD, Michael Erlihson.

מיכאל עובד בחברת הסייבר Salt Security בתור Principal Data Scientist. מיכאל חוקר ופועל בתחום הלמידה העמוקה, ולצד זאת מרצה ומנגיש את החומרים המדעיים לקהל הרחב.